Proceedings

Second International Workshop on

Active Mining

(AM'03)

October 28, 2003

Maebashi TERRSA, Maebashi City, Japan

In conjunction with

14th International Symposium on

Methodologies for Intelligent Systems

Contents

A Fuzzy Set Approach to Query Syntax Analysis in Information Retrieval Systems

Dariusz J. Kogut

Institute of Computer Science, Warsaw University of Technology,
Nowowiejska 15/19, 00-665 Warsaw, Poland
Dariusz.Kogut@cern.ch

Abstract. This article investigates whether a fuzzy set approach to the natural language syntactic analysis can support information retrieval systems. It concentrates on a web search since the internet becomes a vast resource of information. In addition, this article presents a module of syntax analysis of TORCH project where the fuzzy set disambiguation has been implemented and tested.

1 Introduction

Some recent developments on information technology have concured to accelerate a research in the field of artificial intelligence [5]. The appeal of fantasizing about intelligent computers that understand human communication is practically unavoidable. However, the natural language research seems to be one of the hardest problems of artificial intelligence due to the complexity, irregularity and diversity of human languages [8].

This article investigates whether a fuzzy set approach to natural language processing can support the search engines in web universe. An overview of syntactic analysis based on fuzzy set disambiguation will be presented and may provide some insight for further inquiry.

2 Syntax Analysis in TORCH - a Fuzzy Set Approach

TORCH is an information retrieval system with a natural language interface. It has been designed and implemented by author in order to facilitate searching of physical data in CERN (European Organization for Nuclear Research, Geneva). TORCH relies on a textual database composed of web documents published in the internet and intranet networks. It became an add-on module in Infoseek Search Engine environment [6].

2.1 TORCH Architecture

The architecture of TORCH has been shown in figure 1. To begin with, a natural language query is captured by the TORCH user interface module and transmitted to the syntactic analysis section. After the analysis is completed, the query

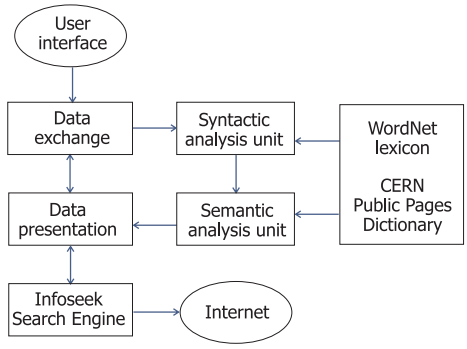


Fig. 1. TORCH Architecture

is being conveyed to the semantic analysis module [6], then interpreted and translated into a formal Infoseek Search Engine query.

Let us focus on the syntax analysis module of TORCH which is shown in figure 2. The syntax analysis unit has been based on a *stepping-up parsing* algorithm and equipped with a fuzzy logic engine that supports syntactic disambiguation. The syntactic dissection goes through several autonomous phases [6]. Once the natural language query is transformed into a *surface structure* [4], the preprocessor unit does the introductory work and query filtering.

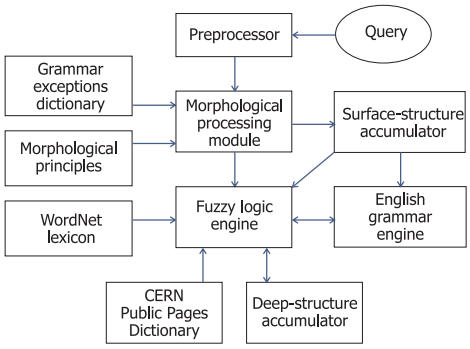


Fig. 2. Syntax Analysis Module

Next, the morphological transformation unit turns each word from a non-basic form into its canonical one [7]. WordNet lexicon and CERN Public Pages Dictionary provide the basic linguistic data and therefore play an important role in both syntactic and semantic analysis. WordNet Dictionary has been built on Princeton University [7], while CERN Public Pages Dictionary extends TORCH knowledge on advanced particle physics. At the end of dissection process, a fuzzy logic engine formulates the *part of speech membership functions* [6] which char-

acterize each word of the sentence. This innovatory approach allows to handle the most of syntax ambiguity cases which happen in English (tests proved that approx. 90% of word atoms are properly identified).

2.2 Fuzzy Set Approach

The fuzzy logic provides a rich and meaningful addition to standard logic and creates the opportunity for expressing those conditions which are inherently imprecisely defined [11]. The architecture of fuzzy logic engine has been shown in figure 3.

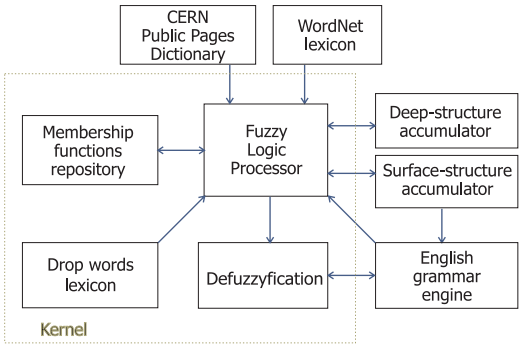


Fig. 3. Fuzzy Logic Engine

The engine solves several cases where syntactic ambiguation may occur. Therefore, it constructs and updates four descrete membership functions designated as Ψ_N, Ψ_V, Ψ_{Adj} and Ψ_{Adv}. At first stage, the linguistic data retrieved from WordNet lexicon are used to form the draft membership functions, as it is described below:

Let us assume that :

L_K - refers to the amount of categories that may be assigned to the word;
N_x - refers to the amount of meanings within the selected category;
Note that $x \in \mathcal{M}$, where $\mathcal{M} = \{N(oun), V(erb), Adj, Adv\}$;
N_Z - refers to the amount of meanings within the all categories;
$\boldsymbol{w}_{(n)}$ - refers to the n-vector (n-word) of the sentence;
ξ_K - is a heuristic factor that describes a category weight (e.g., $\xi_K = 0.5$)

The membership functions have been defined as follows:

$$N_Z[\boldsymbol{w}_{(n)}] = \sum_{x \in \mathcal{M}} N_x[\boldsymbol{w}_{(n)}]$$

$$\Psi_N[\boldsymbol{w}_{(n)}] = \frac{1-\xi_K}{N_Z[\boldsymbol{w}_{(n)}]} \cdot N_N[\boldsymbol{w}_{(n)}] + \frac{\xi_K}{L_K}$$

Similarly,

$$\Psi_V[\boldsymbol{w}_{(n)}] = \frac{1-\xi_K}{N_Z[\boldsymbol{w}_{(n)}]} \cdot N_V[\boldsymbol{w}_{(n)}] + \frac{\xi_K}{L_K}$$

$$\Psi_{Adj}[\boldsymbol{w}_{(n)}] = \frac{1-\xi_K}{N_Z[\boldsymbol{w}_{(n)}]} \cdot N_{Adj}[\boldsymbol{w}_{(n)}] + \frac{\xi_K}{L_K}$$

$$\Psi_{Adv}[\boldsymbol{w}_{(n)}] = \frac{1-\xi_K}{N_Z[\boldsymbol{w}_{(n)}]} \cdot N_{Adv}[\boldsymbol{w}_{(n)}] + \frac{\xi_K}{L_K}$$

In order to illustrate the algorithm, let us take an example query: *Why do the people drive on the right side of the road?* Figures 4 and 5 describe the draft Ψ_N, Ψ_V, Ψ_{ADJ} and Ψ_{ADV} functions.

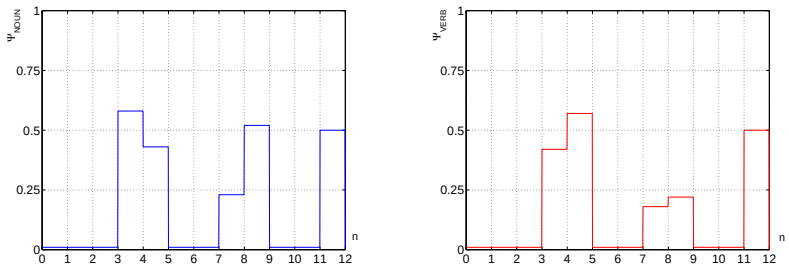


Fig. 4. Ψ_N and Ψ_V membership functions

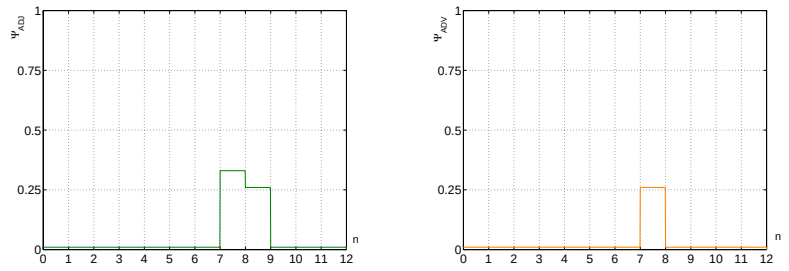


Fig. 5. Ψ_{ADJ} and Ψ_{ADV} membership functions

Certainly, each word must be considered as a specific part of speech and may belong to only one of the lingual categories in the context of a given sentence. Thus, the proper category must be assigned in a process of defuzzyfication, which may be described by the following steps and formulas ($\mathcal{K}_N, \mathcal{K}_V, \mathcal{K}_{Adj}$ and $\mathcal{K}_{Adv}$

represent the part of speech categories):

Step 1.: $\boldsymbol{w}_{(n)} \in \mathcal{K}_N \Leftrightarrow \Psi_N[\boldsymbol{w}_{(n)}] \geq \max(\Psi_{Adj}[\boldsymbol{w}_{(n)}], \Psi_V[\boldsymbol{w}_{(n)}], \Psi_{Adv}[\boldsymbol{w}_{(n)}])$

Step 2.: $\boldsymbol{w}_{(n)} \in \mathcal{K}_{Adj} \Leftrightarrow$
$\Psi_{Adj}[\boldsymbol{w}_{(n)}] \geq \max(\Psi_V[\boldsymbol{w}_{(n)}], \Psi_{Adv}[\boldsymbol{w}_{(n)}]) \wedge (\Psi_{Adj}[\boldsymbol{w}_{(n)}] > \Psi_N[\boldsymbol{w}_{(n)}])$

Step 3.: $\boldsymbol{w}_{(n)} \in \mathcal{K}_V \Leftrightarrow$
$\Psi_V[\boldsymbol{w}_{(n)}] > \max(\Psi_N[\boldsymbol{w}_{(n)}], \Psi_{Adj}[\boldsymbol{w}_{(n)}]) \wedge (\Psi_V[\boldsymbol{w}_{(n)}] \geq \Psi_{Adv}[\boldsymbol{w}_{(n)}])$

Step 4.: $\boldsymbol{w}_{(n)} \in \mathcal{K}_{Adv} \Leftrightarrow \Psi_{Adv}[\boldsymbol{w}_{(n)}] > \max(\Psi_N[\boldsymbol{w}_{(n)}], \Psi_{Adj}[\boldsymbol{w}_{(n)}], \Psi_V[\boldsymbol{w}_{(n)}])$

Unfortunately, the syntactic disambiguation based on lexicon data exclusively cannot handle all the cases. Therefore, a second stage of processing - based on any grammar principles - seems to be necessary [2] [3]. For that reason, TORCH employs its own English grammar engine [6] (shown in figure 6) with a set of simple grammar rules that can verify and approve the membership functions.

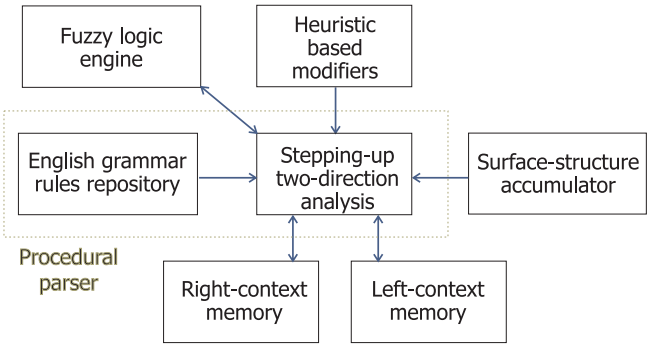


Fig. 6. English Grammar Engine

It utilizes *a procedural parsing*, so the rules are stored in a sort of static library [10] and may be exploited on demand. When the engine exploits English grammar rules upon the query and verifies the fuzzy logic membership functions, the deep structure of the question is constructed [1] [9] and the semantic analysis initialized [6].

The query syntax analysis of TORCH seems to be simple and efficient. A set of tests based on 20000-word samples of e-text books has been done, and the accuracy results (Acc-A when the English grammar engine was disable, and Acc-B with the grammar processing switched on) are shown in table 1.

Natural language clearly offers advantages in convenience and flexibility, but also involves challenges in query interpretation.

E-Text Books from Project Gutenberg	Acc-A	Acc-B
Cromwell by William Shakespeare	58%	84%
J.F. Kennedy's Inaugural Address (Jan 20, 1961)	67%	92%
The Poetics by Aristotle	61%	86%
An Account of the Antarctic Expedition by R. Amundsen	63%	89%
Andersen's Fairy Tales by H.Ch. Andersen	70%	94%

Table 1. The accuracy of TORCH syntax analysis

TORCH with its fuzzy set approach to the syntactic disambiguation attempts to step towards intelligent systems that one day might be able to understand human communication.

3 Acknowledgements

The author wish to thank the many individuals who have contributed to this article through generous sharing of their time and insights. TORCH would not have been possible without the cooperative effort and spirit kind people all over the world. Special thanks to Maria Dimou (CERN), professor Henryk Rybinski (Warsaw University of Technology), Robert Cailliau (CERN) and Michael Naumann (Infoseek).

References

1. Borsley, R.: Syntactic theory. A unified approach. Arnold Publishing Company, London (1990)
2. Brill, E.: Pattern-Based Disambiguation for Natural Language Processing. EMNLP/VLC, (2000)
3. Brill, E., Wu, J.: Classifier Combination For Improved Lexical Disambiguation. COLING/ACL, (1998)
4. Chomsky, N.: Aspects of the theory of syntax. MIT Press, Cambridge (1965)
5. Genesereth, M.,Nilsson, N.: Logical foundations of artificial intelligence. Morgan Kaufmann Publishers, Los Altos (1989)
6. Kogut, D. : TORCH system - the theory, architecture and implementation. http://home.elka.pw.edu.pl/∼dkogut/torch
7. Princeton University: WordNet Dictionary. http://www.cogsci.princeton.edu
8. Rama, D., Srinivasan, P.: An investigation of content representation using text grammars. ACM Transactions on Information Systems, Vol.11, ACM Press, (1993) 51–75
9. Riemsdijk, H., Williams, E.: Introduction to the theory of grammar. MIT Press, Cambridge (1986)
10. Roche, E., Schabes, Y.: Finite-state language processing. MIT Press, Cambridge (1997)
11. Zadeh, L., Kacprzyk, J.: Fuzzy logic for the management of uncertainty. Wiley, New York (1992)

Acquisition of Hypernyms and Hyponyms from the WWW

Ratanachai Sombatsrisomboon [*1] Yutaka Matsuo [*2] Mitsuru Ishizuka [*1]

[*1] Department of Information and Communication Engineering
School of Information Science and Technology
University of Tokyo

[*2] National Institute of Advanced Industrial Science and Technology (AIST)

ratchai@miv.t.u-tokyo.ac.jp

Abstract. Recently research in automatic ontology construction has become a hot topic, because of the vision that ontology will be the core component to realize the semantic web. This paper presents a method to automatically construct ontology by mining the web. We introduce an algorithm to automatically acquire hypernyms and hyponyms for any given lexical term using search engine and natural language processing techniques. First, query phrase is constructed using the frame *"X is a/an Y"*. Then corpora sentences is obtained from the result from search engine. Natural language processing techniques is then employed to discover hypernym/hyponym lexical terms. The methodologies proposed here in this paper can be used to automatically augment natural language ontology, such as WordNet, domain specific knowledge.

1. Introduction

Recently research in ontology has been given a lot of attention, because of the vision that ontology will be the core component to realize the next generation of the web, Semantic Web [1]. For example, natural language ontology (NLO), like WordNet [2], will be used to as background knowledge for a machine; domain ontology will be needed as specific knowledge about a domain when particular domain comes into discussion, and so on.

However, the task of ontology engineering has been very troublesome and time-consuming as it needs domain expert to manually define the domain's conceptualization, left alone maintaining and updating ontologies.

This paper presents an approach to automatically acquire hypernyms and hyponyms for any given lexical term. The methodologies proposed can be used to augment domain specific natural language ontology automatically or as a tool to help constructing the ontology manually.

2. Related Work

There has been studies on extracting lexical relation from natural language text. The work by Hearst [3] introduces an idea that hyponym relation can be extracted from free text by using predefined lexico-syntactic patterns, such as "NP_0 *such as* {NP_1, NP_2 …, *(and | or)*} NP_n" or "*NP* {,} *including* {*NP* ,}* {*or | and*} *NP*", and so on. For example, in the former pattern, the relations *hyponym*(NP_i, NP_0); *for all i from 1 to n* can be inferred. With these predefined patterns, hyponym relations can be obtained by running an acquisition algorithm using pattern matching through text corpus.

Step 3: Filtering Discovered Rules

We filter discovered rules by using the result of MEDLINE document retrieval. More precisely, based on a result of document retrieval, we rank discovered rules. How to rank discovered rules by using the result of document retrievals is a core method of discovered rule filtering.

We assume the number of MEDLINE documents hit by a set of keywords shows a trend of research activity related to the keywords, so we may say that the more the number of hits is, the more the rule that contains the keywords is commonly known in the research field. The published month or year of document may be another hint to rank rules. If many documents related to a rule are published recently, the rule may contain a hot topic in the field.

3 Two Approaches to Discovered Rule Filtering

How to filter discovered rules according to the search result of MEDLINE document retrieval is a most important issue of this work. We have two approaches; micro view approach and macro view approach, to realize discovered rule filtering.

3.1 Micro View Approach

In the micro view approach, we retrieve and show documents related to a discovered rule directly to the user.

By using the micro view approach, the can obtain not only novel rules discovered by a data mining system, but also documents related to the rules. By showing a rule and documents related to the rule at once, the user can get more insights on the rule and may have a chance to start a new data mining task. For the detail, please refer to [3].

However, it is actually difficult to retrieve appropriate documents rightly related a rule because of the low performance of information technique. Especially, when a rule is simple as it is composed of a small number of attributes, the IR system returns a noisy output, documents including a large number of unrelated ones. When a rule is complicated as it is composed of a large number of attributes, it returns few documents.

To see how a micro view approach works, we performed a preliminary experiment of discovered rule filtering. We used 20 rules obtained from the team in Shizuoka University and gathered documents related to the rules from the MEDLINE database. The result is shown in Table 1.

In this table, "ID" is the ID number of rule and "Keywords" are extracted from the rule and are submitted to the Pubmed. "No" shows the number of submitted keywords. "Hits" is the number of documents returned. "Ev" is the evaluation of rule by a medical doctor. He evaluated each rule, which was given in a form depicted in Fig. 1, and categorized into 2 classes; R (reasonable rules) and U (unreasonable rules).

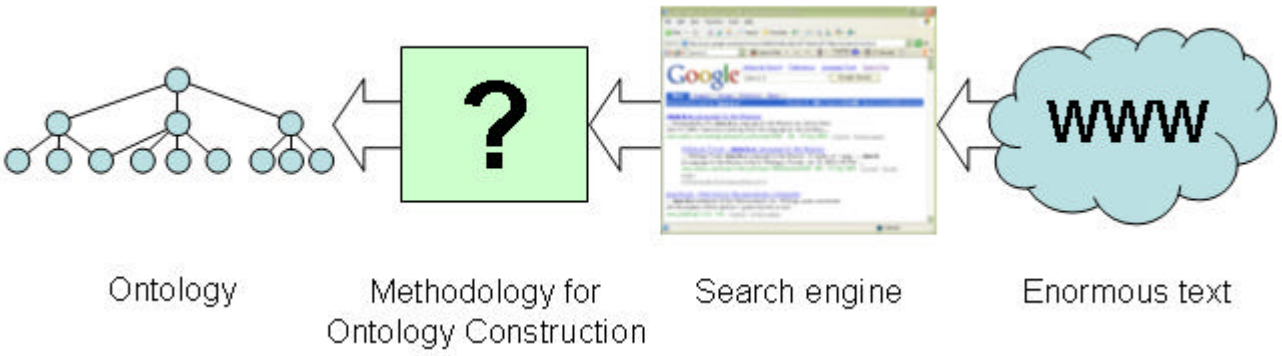


Fig. 1. New approach to Ontology Construction from the WWW using search engine

Another lexical relation acquisition using pattern matching proposed by Sundblad [4]. The approach in [4] extracts hyponym and meronym relation from question corpora. For example, *'Maebashi'* can be inferred as a location from a question like *"Where is Maebashi?"* However, question corpora are very limited in amount since question is less frequently used in normal text. Moreover, relations that can be acquired form question are very limited.

Both methodologies in both [3] and [4], acquires whatever relation found in the text corpora. Because matching pattern on large text corpora consumes a lot of machine processing power and time, therefore relation of specific interest of specific concept cannot be specifically inquired. The methodology to solve this problem is described in the next section.

3. Our Approach

Huge amount of information is currently available on the WWW and increasing at a very high rate. At the time of writing, there are more than 3 billion web pages on the web with more than one million web pages are added daily. Web users can get to these pages by querying specific term(s) to search engine. With index structure of currently available search engine, it is possible to form a more specific a query with phrasal expression, boolean operation, and etc. The result returned by search engine contains a URL and a sentence that the query appears with their context, which is called snippet. All of above inspired us the new approach in acquiring lexical relations, which will be introduced next.

The web is huge, and therefore we want to use it as our corpus to discover lexical relations, but neither obtaining all the text from the web, nor processing it is possible. Therefore we construct a small corpus according to query on-fly from putting sentences together. Corpora sentences are extracted from the query result's snippets, which means there is no interaction with web page's host, only interaction with search engine index is needed. However, to obtain corpora sentences that can be used to extracted lexical relation, we need to formulate a query according to the pattern that we will use to extract relation to a search engine. For example, if we will use *"X is a/an Y"* pattern to extract hyponym relation between *X* and *Y*, then we build up a query as the phrase *"X is a/an"* or *"is a/an Y"*.

After the corpora sentences have been obtained, pattern matching and natural language processing technique can be applied to discover lexical relation, and then statistical technique will be employed to guarantee the accuracy of the result. Finally, domain ontology can be constructed. (see figure 1).

4. Acquiring Hypernyms

From the definition of hyponymy relation written in [2] – a concept represented by the synset {*x*, *x*', …} is said to be a hyponym of the concept represented by the synset {*y*, *y*', …} if native speaker of English accept sentences constructed from such frames as "*An x is a (kind of) y*", we query to search engine using the query phrase "*X is a/an"* to acquire hypernyms for *X*, For example, we query to search engine with the phrase "*scripting language is a/an*" to find hypernyms for the "*scripting language*". For each result returned from search engine, the sentence that contains the query phrase is then extracted from snippet. Example of these sentences can be seen in figure 2.

From those sentences we then filter out undesirable sentences, e.g. sentences marked by * in figure 2, and extract lexical items that are conceivable as query term's hypernyms (The terms that appear in bold face in the figure). In filtering out the undesirable sentences, we remove the sentences that is the query phrase is not the start of the sentence or not preceded by conjunctions such as 'that', 'because', 'since', 'while', 'although', 'though', 'even if', and etc. For extracting the hypernyms term, we first tag all terms with POS and capture the first noun (or compound noun) in the noun phrase ignoring its descriptive adjective.

Extracted lexical items that represents the same concept are then grouped together according to synonymy (synsets) defined in WordNet. Finally, the concepts that occur more frequently proportion to the others are then suggested as hypernyms for the query. As an example, from figure 2., we can infer relation *hypernym("scripting language", "programming language")*, since 'programming language' is the lexical concept that appear most frequently.

A *scripting language is a* lightweight **programming language.**

Scripting language, is an interpreted **programming language** that …

I think that a *scripting language is a* very limited, **high-level language**.

Since a *scripting language is a* full function **programming language**

* The BASIC *scripting language is a* dream to work with.

* The *scripting language is a* bit tough for me, esp.…

Fig. 2. Example of sentences extracted from results obtained from a search engine for query phrase "scripting language is a/an".

5. Acquiring Hyponyms

The algorithm to acquire hyponyms is similar to acquiring hypernyms in section 4. It begins with formulate a query phrase, query to search engine, and then obtain the results which will be used as a corpus to extract hyponyms.

Same as the in acquiring hypernyms, we exploit the frame of "*An x is a (kind of) y*" to discover hyponyms. To discover hyponyms for lexical concept *Y*, we first construct query phrase, "*is a/an Y"*, and query to a search engine. As an example, the sentences extracted from returned result for the query "*is a scripting language"* are shown in figure 3.

XEXPR *is a scripting language* that uses XML as…

I know that **python** *is a scripting language*, but I'm not sure…

JavaScript *is a scripting language* used in Web pages, similar to…

Tell your friend that **C** *is a scripting language* too.

* What is *a scripting language*?

* A language is decided upon whether it *is a scripting language* by…

Fig. 3. Example of sentences extracted from results obtained from a search engine for query phrase "is a scripting language"

After we have extracted the sentence out of each snippet, and filter out the sentence like the one marked by * in figure 3., lexical items that comes right before the query phrase *"is a/an Y"* are spotted as candidate hyponyms of *Y*. (Shown as bold face in the figure.)

Subsequently, each candidate with small number of occurrence is then confirmed by acquiring its hypernyms and check if concept *Y* is in its top hypernyms acquired. If it is, then the candidate term is accepted as hyponym. For example, in the figure 3, there is a statement that 'C' is a scripting language, however, as you might know 'C' is actually not a scripting language (or if it is, definitely it is not well recognized as a scripting language and thus should there be no formal semantic relation between the terms). Therefore, 'C' will be rejected in the confirmation process since scripting language will not be in the top hypernyms of 'C'.

Finally, the candidate terms that have large number of occurrence and the candidate terms that pass the confirmation process are suggested as hyponyms for *Y*.

6. Examples

In this section, we report an experiment of our proposed algorithm for acquiring hypernyms and hyponyms. The system is implemented with Perl using Google Web APIs [5] as an interface with index of Google search engine[6]. The number of corpora sentences retrieved from the search engine ranges from zero to more than thousand sentences. We avoid a problem of reliability of information source by limit maximum number of sentences extracted from a particular domain to 2 sentences. The result of experiment for query terms that yield outputs are shown in figure 4 and 5.

In figure 4, given query terms as input, a list of their hypernyms can be derived. There are large amount of information regarding the first three query terms on the web, in which a lot of *"X is a/an NP"* sentence pattern can be extracted (number of corpora sentences extracted is written in brackets next to the query term), and thus yield very accurate results, as the system acquires 'programming language' and 'language' as hypernyms for query terms 'Java', 'Perl', and 'Python' with highest percentage relative to acquired hypernyms (number on the right of acquired hypernyms shows proportion of number of sentence the hypernym appears with number of total corpora sentences expressed as a percentage). For the query terms 'Maebashi' and 'Active Mining', which less information is available on the web (in English text, to be accurate), there is only a small number of corpora sentences can be extracted. Nevertheless, the result hypernyms yielded are still accurate as it can tell that 'Maebashi' is a city, however for 'Active Mining', 'new direction' is the result because 3 out of 5 corpora sentences are "*Active mining is a new direction in data mining/the knowledge discovery process*"

The result of applying hyponyms acquisition algorithm proposed in this paper can be seen in figure 5. The query terms are 'programming language', 'scripting language', and 'search engine'. Each of these lexical concepts represent a very large class with a lot of members, which a number of those members can be discovered as shown in the figure. The precision of the acquired hyponyms is quite satisfying as shown in the acquired result, however we do not show the recall measure here, and thus it will be our future work on result evaluation.

7. Discussion and Conclusions

We have introduced a new approach to automatically construct ontology using search engine, natural language processing techniques. Methodologies for acquiring hypernyms and hyponyms of a query term are described. With these two techniques, taxonomy of domain ontology as shown in figure 6 and alike can be automatically constructed in a very low cost with no domain-dependent knowledge is required. Finally, the domain specific ontology constructed can then be augmented to natural language ontology (NLO), e.g. WordNet.

QUERY TERM	ACQUIRED HYPERNYMS
JAVA (1011)	programming language(0.33), language(0.20), object-oriented language(0.06), interpreted language(0.05), trademark(0.05)
PERL (913)	programming language(0.23), language(0.23), interpreted language(0.15), scripting language(0.11), tool(0.03), acronym(0.03)
PYTHON (777)	programming language(0.37), language(0.20), scripting language(0.14), interpreted language(0.06), object-oriented language(0.03)
SEARCH ENGINE (84)	tool(0.13), index(0.07), site(0.06), database(0.06), program(0.6), searchable database(0.05)
SCRIPTING LANGUAGE (23)	programming language (0.35)
MAEBASHI (11)	City (0.63), international city (0.45)
ACTIVE MINING (5)	new direction (0.60)

Fig. 4. Hypernyms acquired for selected query terms

PROGRAMMING LANGUAGE : ABAP, ADA, APL ,AppleScript, awk, C, CAML, Cyclone, DarkBASIC, Eiffel, Erlang, Esterel, Expect, Forth, FORTRAN, Giotto, Icon, INTERCAL, Java, JavaScript, Kbasic, Liberty BASIC, Linder, Lisp, LOGO, Lua ,ML, Mobile BASIC, Modula-2, Nial, Nickle, Occam ,Pascal, Perl, PHP, Pike, PostScript, Prolog, Python, Quickbasic, Rexx, Smalltalk, SPL, ToonTalk, Turing, VBScript, VHDL, Visual DialogScript, XSLT

SCRIPTING LANGUAGE : AppleScript, AREXX, ASP, AWK, CCSH, CFML, CFSCRIPT, ColdFusion, Compaq Web , anguage, CorbaScript, DelphiWebScript, ECMAScript, Expect, FDL, ferrite, Glish, IDLScript, JavaScript, Jint, Jscript, KiXtart, ksh, Lingo, Lite, Lua, Lucretia, Miva Script, MML, Perl, Pfhortran, PHP, PHP3, Pnuts, Python, REXX, Ruby, RXML, STEP, Tcl, Tcl/Tk, UserTalk, VBScript, WebScript, WinBatch

SEARCH ENGINE: Aeiwi, AFSearch, AlltheWeb, AltaVista, antistudy.com, Ask Jeeves, ASPseek, Biolinks, Convonix, CPAN Search, Dataclarity, DAYPOP, FDSE, Feedster, FileDonkey, Fluffy Search, FreeFind, GamblingSeek, Google, HotBot, Inktomi, kids.net.au, Londinium.com, Mirago, mnoGoSearch, Northern Light, OttawaWEB, Overture, Phantis, PsychCrawler, Scirus, Search Europe, search4science, SearchNZ, Searchopolis.com, SiteSearch, SpeechBot, Teoma, Vivisimo, WebCrawler, WebWombat, XL Search, Yahoo!, Yahooligans!

Fig. 5. Hyponyms acquired for query term programming language, scripting language, and search engine

Moreover, this methodology requires only small amount of time (in the matter of seconds for hypernyms acquisition or minutes for hyponyms including confirmation process) to discover the query's subset/superset concepts, and thus domain specific lexical concept can be learned on-fly (given the term is described somewhere on the web with *"is a/an"* pattern, and has been indexed by the search engine employed in the system).

Although the work reported in this paper uses only the pattern *"NP is a/an NP"*, we can also use a set of lexicon syntactic patterns suggested by Hearst [3] to acquire hyponyms of a query term. For example, we can use a pattern "NP_0 *such as* {NP_1, NP_2 …, (and | or)} NP_n" to formulate the query to search engine as *"Y such as"* and derived corpora sentences to subsequently extract the hyponyms of query term *Y*. However, as a trade-off for the web's growth rate and its size, there is innumerable natural language text on the web that is highly informal, unstructured or unreliable. For that reason, using only pattern matching alone will result in low precision. To solve this problem, we treat terms extracted from pattern matching as candidate terms for hyponym, and then confirm if a candidate term is actually the query term's hyponym by acquiring their hypernyms using method described in section 4, and then check with their acquired hypernyms as suggested in section 5.

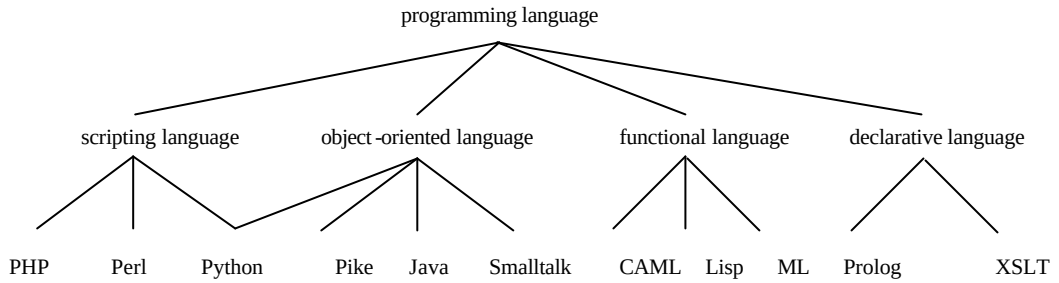


Fig. 6. Example of taxonomy built using proposed methodology

Our method works very well for specific terms. However, it often fail in acquiring hypernyms of general noun, such as 'student', 'animal', and etc. because descriptive sentence with a *"is a/an"* pattern is rarely appear in normal text. Nevertheless, acquiring hypernyms for general terms is hardly of any interest, since general term usually has already been defined in well-established machine understandable dictionary such as WordNet.

8. Future Work

Firstly, we need to formulate an evaluation method based on precision and recall of acquired hypernyms and hyponyms. Secondly, In addition to acquiring lexical term, we also want to acquire its meaning, and thus using corpora sentence's context to disambiguate word sense is our interest for future study. Lastly, apart from hyponymy or ISA relation, there are other semantic relations that we are interested to automate the acquisition process, such as meronymy or HASA relation, and non-taxonomic relation such as *'created by'*, *'produce'*, and etc.

References

1. T. Berners-Lee, J. Hendler, and O. Lassila, The semantic web. In *Scientific American*, May 2001.
2. G. Miller, R. Beckwith, C. Fellbaum, D Gross, and K. Miller. Five papers on wordnet. Technical report, Stanford University, 1993.
3. M. A. Hearst, Automatic Acquisition of Hyponyms from Large Text Corpora, in *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France.
4. H. Sundblad, Automatic Acquisition of Hyponyms and Meronyms from Question Corpora, in Proceedings of the Workshop on Natural Language Processing and Machine Learning for Ontology Engineering at ECAI'2002, Lyon, France.
5. Google Web APIs, http://www.google.com/apis/
6. Google, http://www.google.com
7. B. Omelayenko, Learning of ontologies for the Web: the analysis of existent approaches. In *Proceedings of the International Workshop on Web Dynamics*, 2001.
8. E. Agirre, O. Ansa, E. Hovy and D. Martinez, Enriching Very Large Ontologies Using the WWW, in *Proceedings of the Ontology Learning Workshop, ECAI*, Berlin, Germany, 2000.

Micro View and Macro View Approaches to Discovered Rule Filtering

Yasuhiko Kitamura[1], Akira Iida[2], Keunsik Park[3], Shoji Tatsumi[2]

[1] School of Science and Technology, Kwansei Gakuin University,
2-1 Gakuen, Sanda, Hyogo 669-1337, Japan
ykitamura@ksc.kwansei.ac.jp
http://ist.ksc.kwansei.ac.jp/~kitamura/index.htm
[2] Graduate School of Engineering, Osaka City University,
3-3-138, Sugimoto, Sumiyoshi-ku, Osaka, 558-8585
{iida, tatsumi}@kdel.info.eng.osaka-cu.ac.jp
[3] Graduate School of Medicine, Osaka City University,
1-4-3, Asahi-Machi, Abeno-ku, Osaka, 545-8585
kspark@msic.med.osaka-cu.ac.jp

Abstract. A data mining system can semi-automatically discover knowledge by mining a large volume of data, but the discovered knowledge is not always novel and may contain unreasonable facts. We try to develop a discovered rule filtering method to filter rules discovered by a data mining system to be novel and reasonable ones for the user by using information retrieval technique. In this method, we rank discovered rules according to the results of information retrieval from an information source on the Internet. In this paper, we show two approaches toward discovered rule filtering; micro view approach and macro view approach. The micro view approach tries to retrieve and show documents directly related to discovered rules. On the other hand, the macro view approach tries to show research activities related to discovered rules by using the results of information retrieval. We discuss advantages and disadvantages of micro view approach and possibilities of macro view approach by using an example of clinical data mining and MEDLINE document retrieval.

1 Introduction

The active mining [1] is a new approach to data mining, which tries to discover "high quality" knowledge that meets users' demand in an efficient manner by integrating information gathering, data mining, and user reaction technologies. This paper argues the discovered rule filtering method [3,4] that filters rules obtained by a data mining system based on documents retrieved from an information source on the Internet.

Data mining is an automated method to discover useful knowledge for users by analyzing a large volume of data mechanically. Generally speaking, conventional methods try to discover significant relations among attributes in the statistics sense from a large number of attributes contained in a given database, but if we pay attention to only statistically significant features, we often discover rules that have been known by the user. To cope with this problem, we are developing a discovered rule

filtering method that filters a large number of rules discovered by a data mining system to be novel to the user. To judge whether a rule is novel or not, we utilize information sources on the Internet and try to judge the novelty of rule according to the search result of document retrieval that relates to the discovered rule..

In this paper, we show the concept and the procedure of discovered rule filtering using an example of clinical data mining in Section 2. We then show two approaches toward discovered rule filtering; the micro view approach and the macro view approaches in Section 3. Finally we conclude this paper with our future work in Section 4.

2 Discovered Rule Filtering

As a target of data mining, we use a clinical examination database of hepatitis patients, which is offered by the Medical School of Chiba University, as a common database on which 10 research groups cooperatively work in our active mining project. Some groups have already discovered some sets of rules. For example, a group in Shizuoka University analyzed sequential trends between a set of blood test data (GPT), which represents a progress of hepatitis, and other test data and has already discovered a number of rules, as one of them is shown in Fig. 1.

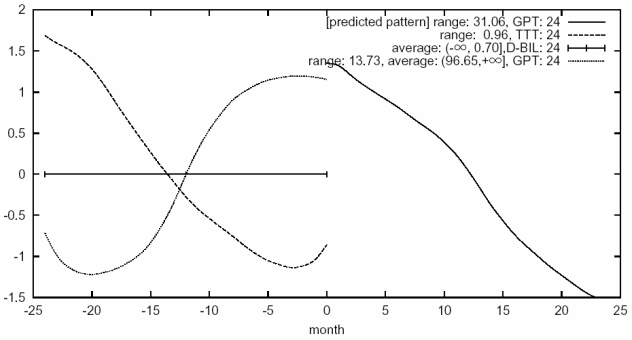


Fig. 1. An example of discovered rule.

This rule shows a relation among GPT (Glutamat-Pyruvat-Transaminase), TTT (Thymol Turbidity Test), and D-BIL (Direct Bilirubin) and means "If, for 24 months, D-BIL stays unchanged, TTT decreases, and GPT increases, then GPT decreases for 24 months." A data mining system can semi-automatically discover a large number of rules by analyzing a set of data given by the user. On the other hand, discovered rules may include ones that are known by the user. Just showing all of the discovered rules to the user may not be a good idea and may result in putting a burden on her. We need to develop a method to filter the discovered rules into a small set of unknown rules for her. To this end, in this paper, we try to utilize information retrieval technique from an information source on the Internet.

When a set of discovered rules are given from a data mining system, a discovered rule filtering system first retrieves information related to the rules from an information source on the Internet and then filter the rules based on the result of information retrieval. In our project, we aim at discovering rules from a hepatitis database, but it is not easy to gather information related to hepatitis from the Web by using a naïve search engine because the Web information sources generally contain a huge amount of various and noisy information. We instead use the MEDLINE (MEDlars on LINE) database as the target of retrieving information, which is a bibliographical database (including abstracts) that covers more than 4000 medical and biological journals that have been published in about 70 countries. It has already stored more than 11 million documents since 1966. PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) is a free MEDLINE search service on the Internet run by NCBI (National Center for Biotechnology Information). By using Pubmed, we can retrieve MEDLINE documents by submitting a set of keywords just like an ordinary search engine. In addition, we can retrieve documents according to the year of publication and/or a category of documents. These functions are not available in ordinary search engines.

A discovered rule filtering process takes the following steps.

Step 1: Extracting keywords from a discovered rule

At first, we find a set of proper keywords to retrieve MEDLINE documents that relate to a discovered rule. Such keywords are extracted from a discovered rule and the domain of data mining as follows.

- **Keywords related to attributes of a discovered rule.** These keywords represent attributes of a discovered rule. For example, keywords that can be acquired from a discovered rule shown in Fig. 1 are GPT, TTT, and D-BIL because they are explicitly shown in the rule. When abbreviations are not acceptable for Pubmed, they need to be translated into normal names. For example, TTT and GPT should be translated into “thymol turbidity test” and “glutamic pyruvic transaminase” respectively.
- **Keywords related to the domain.** These keywords represent the purpose or the domain of the data mining task. They should be included as common keywords. For hepatitis data mining, “hepatitis” is the domain keyword.

Step 2: Gathering MEDLINE documents efficiently

We then perform a sequence of MEDLINE document retrievals. For each of discovered rules, we submit the keywords obtained in Step 1 to the Pubmed system. However, redundant queries may be submitted when many of discovered rules are similar, in other words common attributes constitute many rules. The Pubmed is a popular system that is publicly available to a large number of researchers over the world, so it is required to reduce the load to the system. Actually, too many requests from a user lead to a temporal rejection of service to her. To reduce the number of submissions, we try to use a method that employs a graph representation to store the history of document retrievals. By referring to the graph, we can gather documents in an efficient way by reducing the number of meaningless or redundant keyword submissions.

Table 1. The preliminary experiment of discovered rule filtering.

ID	Ev.	Hits	No.	Keywords
1	R	6	4	hepatitis, gpt, t-cho, albumin
2	U	0	4	hepatitis b, gpt, t-cho, chyle
3	U	0	4	hepatitis c, gpt, lap, hemolysis
4	R	0	5	hepatitis, gpt, got, na, lap
5	R	0	6	hepatitis, gpt, got, ttt, cl, (female)
6	U	0	5	hepatitis, gpt, ldh, hemolysis, blood group a
7	R	7	4	hepatitis, gpt, alb, jaundice
8	R	9	3	hepatitis b, gpt, creatinine
10	R	0	4	hepatitis, ttt, t-bil, gpt
11	U	0	4	hepatitis, gpt, alpha globulin, beta globulin
13	U	8	4	hepatitis, hemolysis, gpt, (female)
14	U	0	4	hepatitis, gpt, ttt, d-bil
15	U	0	3	hepatitis, gpt, chyle
17	R	0	5	hepatitis, gpt, ttt, blood group o, (female)
18	R	2	3	hepatitis c, gpt, t-cho
19	R	0	6	hepatitis, gpt, che, ttt, ztt, (male)
20	R	0	5	hepatitis, gpt, lap, alb, interferon
22	U	0	7	hepatitis, gpt, ggtp, hemolysis, blood group a, (female), (age 45-64)
23	U	0	4	hepatitis b, gpt, got, i-bil
27	U	0	4	hepatitis, gpt, hemolysis, i-bil

As we can see, except Rule 13, rules with hits more than 0 are categorized in reasonable rules, but a number of reasonable rules hit no document. It seems that the number of submitted keywords affects the number of hits. In other words, if a rule is complex with many keywords, the number of hits tends to be few.

This result tells us that it is not easy to distinguish reasonable or known rules from unreasonable or gabage ones by using only the number of hits. It shows a limitation of macro view approach.

To cope with the problem, we need to improve the performance of micro view approach as follows.

(1) **Accurate document retrieval.** In our current implementation, we use only keywords related to attributes contained in a rule and those related to the domain, and the document retrieval is not accurate enough and often contains documents unrelated to the rule. To improve the accuracy, we need to add adequate keywords related to relations among attributes. These keywords represent relations among attributes that constitute a discovered rule. It is difficult to acquire such keywords directly from the rule because, in many cases, they are not explicitly represented in the rule. They need

to be included manually in advance. For example, in the hepatitis data mining, "periodicity" should be included when the periodicity of attribute value change is important.

(2) **Document analysis by applying natural language processing methods.** Another method is to refine the results by analyzing the documents using natural language processing technique. Generally speaking, information retrieval technique only retrieves documents that contain the given keyword(s) and does not care the context in which the keyword(s) appear. On the other hand, natural language processing technique can clarify the context and can refine the result obtained by information retrieval technique. For example, if a keyword is not found in the same sentence in which another keyword appears, we might conclude that the document does not argue a relation between the two keywords. We hence can improve the accuracy of discovered rule filtering by analyzing whether the given keywords are found in a same sentence. In addition, if we can analyze whether the sentence argues the conclusion of the document, we can further improve the accuracy of rule filtering.

3.2 Macro View Approach

In the macro view approach, we try to roughly observe the trend of relation among keywords. For example, the number of documents in which the keywords co-occur approximately shows the strength of relation among the keywords. We show two methods based on the macro view approach.

(1) Showing research activities based pair-wise keyword co-occurrence graph

We depicted a graph which shows a research activities related to a discovered rule by using the number of co-occurrences of every two keywords found in the rule. A node in the graph represents a keyword which specifies an attribute found in the rule and an edge represents the number of co-occurrences of two keywords connected by the edge.

Fig. 2 shows a graph concerning rule 1.

This graph shows that the number of co-occurrences of "albumin" and "gpt" is 150, that of "total cholesterol" and "albumin" is 16, and that of "gpt" and "total cholesterol" is 14. As shown in Table 1, the rule is judged as "reasonable" by a medical doctor and we can see each attribute is interconnected to other attributes strongly.

Fig. 3 shows research activities related to rule 2. This rule is judged as "unreasonable" and the number of hits is 0. Research activities look weak because only 14 documents related to "total cholesterol" and "gpt" are retrieved.

Fig. 4 shows research activities related to rule 4. This rule is judged as "reasonable", but the number of hits is 0. Contrasting with rule 2, research activities related to rule 4 look active because a number of documents are retrieved except documents related to "na" and "lap".

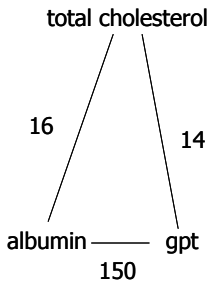


Fig. 2. Research activities related to rule 1.

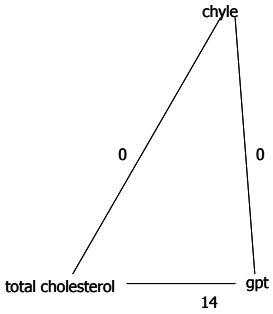


Fig. 3. Research activities related to rule 2.

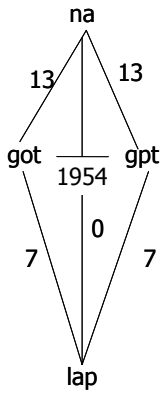


Fig. 4. Research activities related to rule 4.

As a conclusion, the graph shape of reasonable rules looks different from that of unreasonable rules. But, when given a graph, how to judge whether the rule is reasonable or not is our future work.

(2) The yearly trend of research activities

The MEDLINE database contains bibliographical information of bioscience articles, which includes the year of publication, and the Pubmed can retrieve the information according to the year of publication. By observing the yearly trend of co-occurrences, we can see the change of the research activity. For example, we can have the following interpretations as shown in Fig. 5.

(a) If the number of co-occurrences moves upward, the research topic related to the keywords is hot.

(b) If the number of co-occurrences moves downward, the research topic related to the keywords is terminating.

(c) If the number of co-occurrences keeps high, the research topic related to the keyword is commonly known.

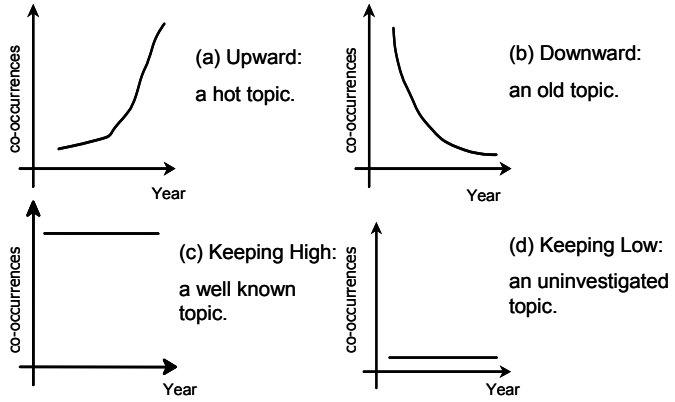


Fig. 5. Yearly Trends of co-occurrences.

(d) If the number of co-occurrences keeps low, the research topic related to the keyword is not known. Few researchers show interest in the topic.

To evaluate a feasibility of this method, we submitted 4 queries to the MEDLINE database and show the results in Fig. 6.

(a) "hcv, hepatitis"

The number of co-occurrences has been increasing since 1989. In 1989, we have an event of succeeding HCV cloning. HCV is a hot topic of hepatitis research.

(b) "smallpox, vaccine"

The number of co-occurrences has been decreasing. In 1980, the World Health Assembly announced that smallpox had been eradicated. Recently, we see the number turns to increasing because discussions about smallpox as a biochemical weapon arise

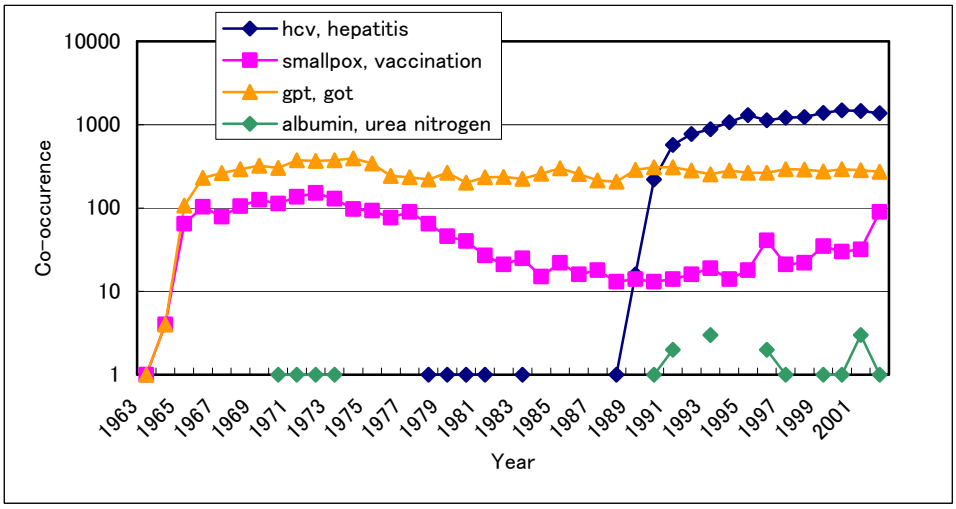


Fig. 6. The yearly trend of research activities.

(c) "gpt, got"

The number of co-occurrences stays high. GPT and GOT are well known blood test measure and they are used to diagnose hepatitis. The relation between GPT and GOT is well known in the medical domain.

(d) "albumin, urea nitrogen"

The number of co-occurrences stays low. The relation between albumin and urea nitrogen is seldom discussed.

From above results, the yearly trends well correspond with historical events in the medical domain, and can be a measure to know the research activities.

4 Summary

We discussed a discovered rule filtering method which filters rules discovered by a data mining system into novel ones by using the IR technique. We proposed two approaches toward discovered rule filtering; the micro view approach and the macro view approach and showed merits and demerits of micro view approach and possibilities of macro view approach.

Our future work is summarized as follows.

- We need to find a measure to distinguish reasonable rules from unreasonable one, which can be used in the macro view method. We also need to find a measure to know the novelty of rule.
- We need to improve the performance of micro view approach by adding keywords that represent relations among attributes and by using natural language processing techniques. The improvement of micro view approach can contribute the improvement of macro view approach.
- We need to implement the macro view method in a discovered rule filtering system and apply it to an application of hepatitis data mining.

Acknowledgement

This work is supported by a grant-in-aid for scientific research on priority area by the Japanese Ministry of Education, Science, Culture, Sports and Technology.

References

1. H. Motoda (Ed.), Active Mining: New Directions of Data Mining, IOS Press, Amsterdam, 2002.
2. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison Wesley, 1999.
3. Y. Kitamura, K. Park, A. Iida, and S. Tatsumi. Discovered Rule Filtering Using Information Retrieval Technique. Proceedings of International Workshop on Active Mining, pp. 80-84, 2002.
4. Y. Kitamura, A. Iida, K. Park, and S. Tatsumi, Discovered Rule Filtering System Using MEDLINE Information Retrieval, JSAI Technical Report, SIG-A2-KBS60/FAI52-J11, 2003.

Relevance Feedback Document Retrieval using Support Vector Machines

Takashi Onoda[1], Hiroshi Murata[1], and Seiji Yamada[2]

[1] Central Research Institute of Electric Power Industry, Communication & Information Laboratory, 2-11-1 Iwado Kita, Komae-shi, Tokyo 201-8511 JAPAN
{onoda, murata}@criepi.denken.or.jp
[2] National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 JAPAN
{seiji}@nii.ac.jp http://research.nii.ac.jp/~seiji/index-e.html

Abstract. We investigate the following data mining problems from the document retrieval: From a large data set of documents, we need to find documents that relate to human interest as few iterations of human testing or checking as possible. In each iteration a comparatively small batch of documents is evaluated for relating to the human interest. We apply active learning techniques based on Support Vector Machine for evaluating successive batches, which is called *relevance feedback*. Our proposed approach has been very useful for document retrieval with relevance feedback experimentally. In this paper, we adopt several representations of the Vector Space Model and several selecting rules of displayed documents at each iteration, and then show the comparison results of the effectiveness for the document retrieval in these several situations.

1 Introduction

As progression of the internet technology, accessible information by end users is explosively increasing. In this situation, we can now easily access a huge document database through the WWW. However it is hard for a user to retrieve relevant documents from which he/she can obtain useful information, and a lot of studies have been done in information retrieval, especially document retrieval [1]. Active works for such document retrieval have been reported in TREC(Text Retrieval Conference) [2] for English documents, IREX(Information Retrieval and Extraction Exercise) [3] and NTCIR(NII-NACSIS Test Collection for Information Retrieval System) [4] for Japanese documents.

In most frameworks for information retrieval, a Vector Space Model(which is called VSM) in which a document is described with a high-dimensional vector is used [5]. An information retrieval system using a vector space model computes the similarity between a query vector and document vectors by cosine of the two vectors and indicates a user a list of retrieved documents.

In general, since a user hardly describes a precise query in the first trial, interactive approach to modify the query vector by evaluation of the user on documents in a list of retrieved documents. This method is called *relevance*

feedback [6] and used widely in information retrieval systems. In this method, a user directly evaluates whether a document is relevant or irrelevant in a list of retrieved documents, and a system modifies the query vector using the user evaluation. A traditional way to modify a query vector is a simple learning rule to reduce the difference between the query vector and documents evaluated as relevant by a user.

In another approach, relevant and irrelevant document vectors are considered as positive and negative examples, and relevance feedback is transposed to a binary classification problem [7]. For the binary classification problem, Support Vector Machines(which are called SVMs) have shown the excellent ability. And some studies applied SVM to the text classification problems [8] and the information retrieval problems[9].

Recently, we have proposed a relevance feedback framework with SVM as *active learning* and shown the usefulness of our proposed method experimentally[10]. Now, we are interested in which is the most efficient representation for the document retrieval performance and the learning performance, boolean representation, TF representation or TFIDF representation, and what is the most useful selecting rule for displayed documents at each iteration. In this paper, we adopt several representations of the Vector Space Model and several selecting rules of displayed documents at each iteration, and then show the comparison results of the effectiveness for the document retrieval in these several situations.

In the remaining parts of this paper, we explain a SVM algorithm in the second section briefly. An active learning with SVM for the relevance feedback, and our adopted VSM representations and selecting displayed documents rules are described in the third section. In the fourth section, in order compare the effectiveness of our adopted representations and selecting rules, we show our experiments using a TREC data set of Los Angeles Times and discuss the experimental results. Eventually we conclude our work in the fifth section.

2 Support Vector Machines

Formally, the Support Vector Machine (SVM) [11] like any other classification method aims to estimate a classification function $f : \mathcal{X} \rightarrow \{\pm 1\}$ using labeled training data from $\mathcal{X} \times \{\pm 1\}$. Moreover this function f should even classify unseen examples correctly.

For SV learning machines that implement linear discriminant functions in feature spaces, the capacity limitation corresponds to finding a large margin separation between the two classes. The margin ϱ is the minimal distance of training points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_i, y_i), \mathbf{x}_i \in \mathbf{R}, y_i \in \{\pm 1\}$ to the separation surface, i.e. $\varrho = \min_{i=1,\ldots,\ell} \rho(\mathbf{z}_i, f)$, where $\mathbf{z}_i = (\mathbf{x}_i, y_i)$ and $\rho(\mathbf{z}_i, f) = y_i f(\mathbf{x}_i)$, and f is the linear discriminant function in some feature space

$$f(\mathbf{x}) = (\mathbf{w} \cdot \Phi(\mathbf{x})) + b = \sum_{i=1}^{\ell} \alpha_i y_i (\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})) + b, \tag{1}$$

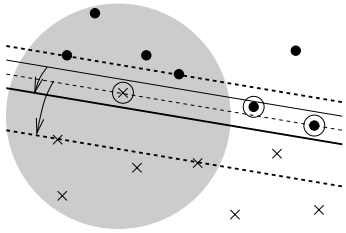


Fig. 1. A binary classification toy problem: This problem is to separate black circles from crosses. The shaded region consists of training examples, the other regions of test data. The training data can be separated with a margin indicated by the slim dashed line and the upper fat dashed line, implicating the slim solid line as discriminate function. Misclassifying one training example(a circled white circle) leads to a considerable extension(arrows) of the margin(fat dashed and solid lines) and this fat solid line can classify two test examples(circled black circles) correctly.

with $\mathbf{w}$ expressed as $\mathbf{w} = \sum_{i=1}^{\ell} \alpha_i y_i \Phi(\mathbf{x}_i)$. The quantity Φ denotes the mapping from input space $\mathcal{X}$ by explicitly transforming the data into a feature space $\mathcal{F}$ using $\Phi : \mathcal{X} \to \mathcal{F}$. (see Figure 1). SVM can do so implicitly. In order to train and classify, all that SVMs use are dot products of pairs of data points $\Phi(\mathbf{x}), \Phi(\mathbf{x}_i) \in \mathcal{F}$ in feature space (cf. Eq. (1)). Thus, we need only to supply a so-called kernel function that can compute these dot products. A kernel function k allows to implicitly define the feature space (Mercer's Theorem, e.g. [12]) via

$$k(\mathbf{x}, \mathbf{x}_i) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i)). \tag{2}$$

By using different kernel functions, the SVM algorithm can construct a variety of learning machines, some of which coincide with classical architectures:

Polynomial classifiers of degree d: $k(\mathbf{x}, \mathbf{x}_i) = (\kappa \cdot (\mathbf{x} \cdot \mathbf{x}_i) + \Theta)^d$, where κ, Θ, and d are appropriate constants.
Neural networks(sigmoidal): $k(\mathbf{x}, \mathbf{x}_i) = \tanh(\kappa \cdot (\mathbf{x} \cdot \mathbf{x}_i) + \Theta)$, where κ and Θ are appropriate constants.
Radial basis function classifiers: $k(\mathbf{x}, \mathbf{x}_i) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_i\|^2}{\sigma}\right)$, where σ is an appropriate constant.

Note that there is no need to use or know the form of Φ, because the mapping is never performed explicitly The introduction of Φ in the explanation above was for purely didactical and not algorithmical purposes. Therefore, we can computationally afford to work in implicitly very large (e.g. 10^{10}- dimensional) feature spaces. SVM can avoid overfitting by controlling the capacity and maximizing the margin. Simultaneously, SVMs learn which of the features implied by the kernel k are distinctive for the two classes, i.e. instead of finding well-suited features by ourselves (which can often be difficult), we can use the SVM to select them from an extremely rich feature space.

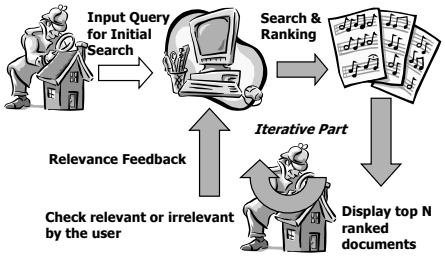


Fig. 2. Image of the relevance feedback documents retrieval: The gray arrow parts are made iteratively to retrieve useful documents for the user. This iteration is called feedback iteration in the information retrieval research area.

With respect to good generalization, it is often profitable to misclassify some outlying training data points in order to achieve a larger margin between the other training points (see Figure 1 for an example). This soft-margin strategy can also learn non-separable data. The trade-off between margin size and number of misclassified training points is then controlled by the regularization parameter C (softness of the margin). The following quadratic program (QP) (see e.g. [11, 13]):

$$\begin{array}{ll} \min \|\mathbf{w}\|^2 + C \sum_{i=1}^{\ell} \xi_i & \\ \text{s.t.} \quad \rho(\mathbf{z}_i, f) \geq 1 - \xi_i & \text{for all } 1 \leq i \leq \ell \\ \qquad\quad \xi_i \geq 0 & \text{for all } 1 \leq i \leq \ell \end{array} \tag{3}$$

leads to the SV soft-margin solution allowing for some errors.

In this paper, we use VSMs, which are high dimensional models, for the document retrieval. In this high dimension, it is easy to classify between relevant and irrelevant documents. Therefore, we generate the SV hard-margin solution by the following quadratic program.

$$\begin{array}{ll} \min & \|\mathbf{w}\|^2 \\ \text{s.t.} & \rho(\mathbf{z}_i, f) \geq 1 \text{ for all } 1 \leq i \leq \ell \end{array} \tag{4}$$

3 Active Learning with SVM in Document Retrieval

In this section, we describe the information retrieval system using relevance feedback with SVM from an active learning point of view, and several VSM representation of documents and several selecting rules, which determine displayed documents to a user for the relevance feedback.

3.1 Relevance Feedback Based on SVM

Fig. 2 shows the concept of the relevance feedback document retrieval. In Fig. 2,

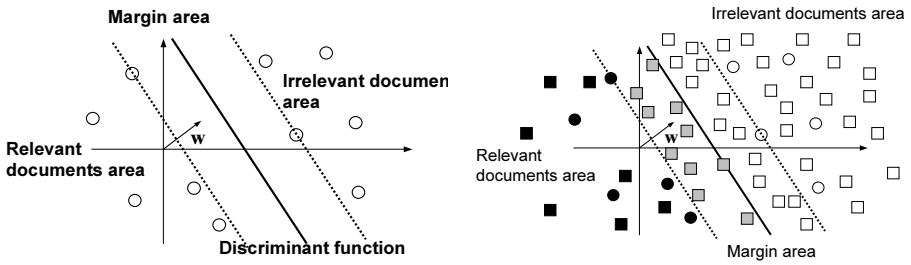


Fig. 3. The left side figure shows a discriminant function for classifying relevant or irrelevant documents: Circles denote documents which are checked relevant or irrelevant by a user. The solid line denotes a discriminant function. The margin area is between dotted lines. The right side figure shows displayed documents as the result of document retrieval: Boxes denote non-checked documents which are mapped into the feature space. Circles denotes checked documents which are mapped into the feature space. The system displays the documents which are represented by black circles and boxes as the result of document retrieval to a user.

the iterative procedure is the gray arrows parts. The SVMs have a great ability to discriminate even if the training data is small. Consequently, we have proposed to apply SVMs as the classifier in the relevance feedback method. The retrieval steps of proposed method perform as follows:

Step 1: **Preparation of documents for the first feedback:** The conventional information retrieval system based on vector space model displays the top N ranked documents along with a request query to the user. In our method, the top N ranked documents are selected by using cosine distance between the request query vector and each document vector for the first feedback iteration.

Step 2: **Judgment of documents:** The user then classifiers these N documents into relevant or irrelevant. The relevant documents and the irrelevant documents are labeled. For instance, the relevant documents have "+1" label and the irrelevant documents have "-1" label after the user's judgment.

Step 3: **Determination of the optimal hyperplane:** The optimal hyperplane for classifying relevant and irrelevant documents is determined by using a SVM which is learned by labeled documents(see Figure 3 left side).

Step 4: **Discrimination documents and information retrieval:** The documents, which are retrieved in the Step1, are mapped into the feature space. The SVM learned by the previous step classifies the documents as relevant or irrelevant. Then the system selects the documents based on the distance from the optimal hyper plane and the feature of the margin area. The detail of the selection rules are described in the next section. From the selected documents, the top N ranked documents, which are ranked using the distance from the optimal hyperplane, are shown to user as the information retrieval results of the system. If the number of feedback iterations is more than m,

then go to next step. Otherwise, return to Step 2. The m is a maximal number of feedback iterations and is given by the user or the system.

Step 5: **Display of the final retrieved documents:** The retrieved documents are ranked by the distance between the documents and the hyper-plane which is the discriminant function determined by SVM. The retrieved documents are displayed based on this ranking(see Figure 3 right side).

3.2 VSM Representations and Selection Rules of Displayed Documents

We discuss the issue of the term t_i in the document vector d_j. In the Information Retrieval research field, this term is called the term weighting, while in the machine learning research filed, this term is called the feature. t_i states something about word i in the document d_j. If this word is absent in the document d_j, t_i is zero. If the word is present in the document d_j, then there are several options. The first option is that this term just indicates whether this word i is present or not. This presentation is called boolean term weighting. The next option is that the term weight is a count of the number of times this word i occurs in this document d_j. This presentation is called the term frequency(TF). In the original Rocchio algorithm[6], each term TF is multiplied by a term $\log\left(\frac{N}{n_i}\right)$ where N is the total number of documents in the collection and n_i is the number of documents in which this word i occurs. This last term is called the inverse document frequency(IDF). This representation is called the term frequency-the inverse document frequency(TFIDF)[1]. The Rocchio algorithm is the original relevance feedback method. In this paper, we compare the effectiveness of the document retrieval and the learning performance among boolean term weighting, term frequency(TF) and term frequency inverse document frequency(TFIDF) representations for our relevance feedback based on SVM.

Next, we discuss two selection rules for displayed documents, which are used for the judgment by the user. In this paper, we compare the effectiveness of the document retrieval and the learning performance among the following three selection rules for displayed documents.

Rule 1: The retrieved documents are mapped into the feature space. The learned SVM classifies the documents as relevant or irrelevant. The documents, which are discriminated relevant and in the margin area of SVM are selected. From the selected documents, the top N ranked documents, which are ranked using the distance from the optimal hyperplane, are displayed to the user as the information retrieval results of the system(see Figure 4 left side). This rule should make the best learning performance from an active learning point of view.

Rule 2: The retrieved documents are mapped into the feature space. The learned SVM classifies the documents as relevant or irrelevant. The documents, which are on the optimal hyperplane or near the optimal hyperplane of SVM, are selected. The system chooses the N documents in these selected documents and displays to the user as the information retrieval results of the

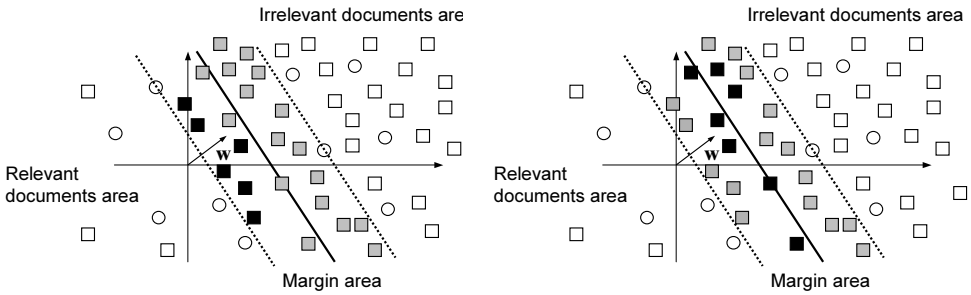


Fig. 4. Mapped non-checked documents into the feature space: Boxes denote non-checked documents which are mapped into the feature space. Circles denotes checked documents which are mapped into the feature space. Black and gray boxes are documents in the margin area. We show the documents which are represented by black boxes to a user for next iteration. These documents are in the margin area and near the relevant documents area(see the left side figure). These documents are near the optimal hyperplane(see the right side figure).

system(see Figure 4 right side). This rule is expected to achieve the most effective learning performance. This rule is our proposed one for the relevance feedback document retrieval.

4 Experiments

4.1 Experimental setting

In the reference [10], we already have shown that the utility of our interactive document retrieval with active learning of SVM is better than the Rocchio-based interactive document retrieval[6], which is conventional one. This paper presents the experiments for comparing the utility for the document retrieval among several VSM representations, and the effectiveness for the learning performance among the several selection rules, which choose the displayed documents to judge whether a document is relevant or irrelevant by the user. The document data set we used is a set of articles in the Los Angeles Times which is widely used in the document retrieval conference TREC [2]. The data set has about 130 thousands articles. The average number of words in a article is 526. This data set includes not only queries but also the relevant documents to each query. Thus we used the queries for experiments.

We adopted the boolean weighting, TF, and TFIDF as VSM representations. The detail of the boolean and TF weighting can be seen in the section 3. And the detail of the adopted TFIDF can be seen in the reference [10]. In our experiments, we used two selection rules to estimate the effectiveness for the learning performance. The detail of these selection rules can be seen in the section 3.

The size N of retrieved and displayed documents at each iteration in the section 3 was set as twenty. The feedback iterations m were 1, 2, and 3. In order

to investigate the influence of feedback iterations on accuracy of retrieval, we used plural feedback iterations. In our experiments, we used the linear kernel for SVM learning, and found a discriminant function for the SVM classifier in this feature space. The VSM of documents is high dimensional space. Therefore, in order to classify the labeled documents into relevant or irrelevant, we do not need to use the kernel trick and the regularization parameter C(see section 2). We used LibSVM [14] as SVM software in our experiment.

In general, retrieval accuracy significantly depends on the number of the feedback iterations. Thus we changed feedback iterations for 1, 2, 3 and investigated the accuracy for each iteration. We utilized *precision* and *recall* for evaluating the two information retrieval methods [15][16] and our approach.

4.2 Comparison of recall-precision performance curves among the boolean, TF and TFIDF weightings

In this section, we investigate the effectiveness for the document retrieval among the boolean, TF and TFIDF weightings, when the user judges the twenty higher ranked documents at each feedback iteration. In the first iteration, twenty higher ranked documents are retrieved using cosine distance between document vectors and a query vector in VSMs, which are represented by the boolean, TF and TFIDF weightings. The query vector is generated by a user's input of keywords. In the other iterations, the user does not need to input keywords for the information retrieval, and the user labels "+1" and "-1" as relevant and irrelevant documents respectively.

Figure 5 left side shows a recall-precision performance curve of our SVM based method for the boolean, TF and TFIDF weightings, after four feedback iterations. Our SVM based method adopts the selection rule 1. The thick solid line is the boolean weighting, the broken line is the TFIDF weighting, and the thin solid line is the TF weighting.

This figure shows that the retrieval effectiveness of the boolean representation is higher than that of the other two representations, i.e., TF and TFIDF representations. Consequently, in this experiment, we conclude that the boolean weighting is a useful VSM representation for our proposed relevant feedback technique to improve the performance of the document retrieval.

4.3 Comparison of recall-precision performance curves between the selection rule 1 and 2

Here, we investigate the effectiveness for the document retrieval between the selection rule 1 and 2, which are described in the section 3.

Figure 5 right side shows a recall-precision performance curves of the selection rule 1 and 2 for the boolean weightings, after four feedback iterations. The thick solid line is the selection rule 1, and the thin solid line is the selection rule 2. Table 1 gives the average number of relevant documents in the twenty displayed documents for the selection rule 1 and 2 as a function of the number of iterations.

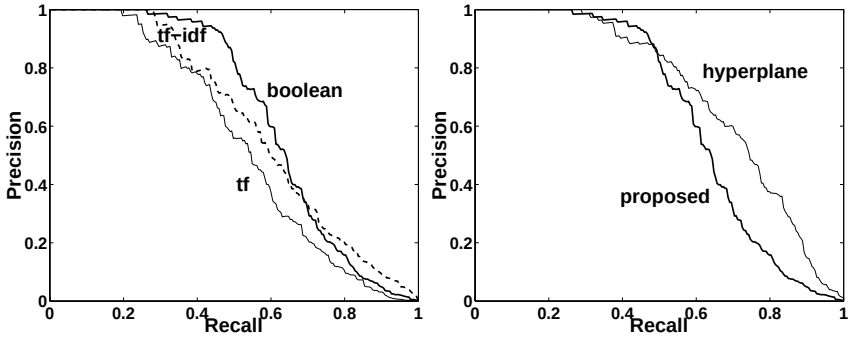


Fig. 5. The left side figure shows the retrieval effectiveness of SVM based feedback(using the selection rule 1) for the boolean, TF, and TFIDF representations: The lines show recall-precision performance curve by using twenty feedback documents on the set of articles in the Los Angeles Times after 3 feedback iterations. The wide solid line is the boolean representation, the broken line is TFIDF representation, and the solid line is TF representation. The right side figure shows the retrieval effectiveness of SVM based feedback for the selection rule 1 and 2: The lines show recall-precision performance curves by using twenty feedback documents on the set of articles in the Los Angeles Times after 3 feedback iterations. The thick solid line is the selection rule 1, and the thin solid line is the selection rule 2.

This figure shows that the precision-recall curve of the selection rule 1 is better than that of the selection rule 2. However, we can see from the table 1 that the average number of relevant documents in the twenty displayed documents for the selection rule 1 is higher than that of the selection rule 2 at each iteration. After all, the selection rule 2 is useful to totally put on the upper rank the documents, which relate to the user's interesting. When the selection rule 2 is adopted, the user have to see a lot of irrelevant documents at each iteration. The selection rule 1 is effective to immediately put on the upper rank the special documents, which relate to the user's interesting. When the selection rule 1 is adopted, the user do not need to see a lot of irrelevant documents at each iteration. However, it is hard for the rule 1 to immediately put on the upper rank all documents, which relate to the user's interesting. In the document retrieval, a user do not want to get all documents, which relate to the user's interest. The user wants to get some documents, which relate to the user's interest as soon as possible. Therefore, we conclude that the feature of the selection rule 1 is better than that of the selection rule 2 for the relevance feedback document retrieval.

5 Conclusion

In this paper, we adopt several representations of the Vector Space Model and several selecting rules of displayed documents at each iteration, and then show

Table 1. Average number of relevant documents in the twenty displayed documents for the selection rule 1 and 2 using the boolean representation

No. of feedback iterations	Ave. No. of relevant documents	
	selection rule 1	selection rule 2
1	11.750	7.125
2	9.125	7.750
3	8.875	7.375
4	8.875	5.375

the comparison results of the effectiveness for the document retrieval in these several situations.

In our experiments, when we adopt our proposed SVM based relevance feedback document retrieval, the boolean representation and the selection rule 1, where the documents that are discriminated relevant and in the margin area of SVM, are displayed to a user, show better performance of document retrieval. In future work, we will plan to analyze our experimental results theoretically.

References

1. Yates, R.B., Neto, B.R.: Modern Information Retrieval. Addison Wesley (1999)
2. TREC Web page: (http://trec.nist.gov/)
3. IREX: (http://cs.nyu.edu/cs/projects/proteus/irex/)
4. NTCIR: (http://www.rd.nacsis.ac.jp/˜ntcadm/)
5. Salton, G., McGill, J.: Introduction to modern information retrieval. McGraw-Hill (1983)
6. Salton, G., ed. In: Relevance feedback in information retrieval. Englewood Cliffs, N.J.: Prentice Hall (1971) 313–323
7. Okabe, M., Yamada, S.: Interactive document retrieval with relational learning. In: Proceedings of the 16th ACM Symposium on Applied Computing. (2001) 27–31
8. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. In: Journal of Machine Learning Research. Volume 2. (2001) 45–66
9. Drucker, H., Shahrary, B., Gibbon, D.C.: Relevance feedback using support vector machines. In: Proceedings of the Eighteenth International Conference on Machine Learning. (2001) 122–129
10. Onoda, T., Murata, H., Yamada, S.: Relevance feedback with active learning for document retrieval. In: Proc. of IJCNN2003. (2003) 1757–1762
11. Vapnik, V.: The Nature of Statistical Learning Theory. Springer (1995)
12. Boser, B., Guyon, I., Vapnik, V.: A training algorithm for optimal margin classifiers. In Haussler, D., ed.: 5th Annual ACM Workshop on COLT, Pittsburgh, PA, ACM Press (1992) 144–152
13. Schölkopf, B., Smola, A., Williamson, R., Bartlett, P.: New support vector algorithms. Neural Computaion **12** (2000) 1083 – 1121
14. Kernel-Machines: (http://www.kernel-machines.org/)
15. Lewis, D.: Evaluating text categorization. In: Proceedings of Speech and Natural Language Workshop. (1991) 312–318
16. Witten, I., Moffat, A., Bell, T.: Managing Gigabytes: Compressing and Indexing Documents and Images. Van Nostrand Reinhold, New York (1994)

Using sectioning information for text retrieval: a case study with the MEDLINE abstracts

Masashi Shimbo, Takahiro Yamasaki, and Yuji Matsumoto

Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{shimbo,takah-ya,matsu}@is.aist-nara.ac.jp

Abstract. We present an experimental text retrieval system to facilitate search in the MEDLINE database. A unique feature of the system is that the user can search not only throughout the whole abstract text, but also from the limited "sections" in the text. These "sections" are determined so as to reflect the structural role (background, objective, conclusions, etc.) of the constituent sentences. This feature of the system makes it easier to narrow down search results when adding extra keywords does not work, and also enables to rank search results according to the user's needs. The realization of the system requires each sentence in the MEDLINE abstracts be classified to one of the sections. For this purpose, we exploit the "structured" abstracts contained in the MEDLINE database in which sections are explicitly marked by the headings. These abstracts provide training data for constructing sentence classifiers that are used to section unstructured abstracts, in which explicit section headings are missing.

Key words: MEDLINE database, structured abstracts, information retrieval, text classification.

1 Introduction

With the rapid increase in the volume of scientific literature, demands are growing for systems with which the researchers can fin relevant pieces of literature with less effort. Online literature retrieval services, including PubMed [8] and CiteSeer [5], are increasing their popularity, as they permit the users to access the large corpora of abstracts or full papers.

PubMed facilitate retrieval of medical and biological papers by means of keyword-search in the MEDLINE abstracts [7]. It also provides a number of auxiliary ways to filter search results. For example, it is possible to perform search on a specific field such as titles and publication date. And most abstracted citations are given peer-reviewed annotation of topics and keywords chosen from controlled vocabulary, which can also be used to restrict search. All these facilities, however, rely on information external to the content of the abstract text. In this report, by contrast, we explore the use of information that is inherent in the abstract text, to make retrieval process more goal-oriented.

The information we exploit is the structure underlying the abstract texts. Our system allows search to be executed within restricted portions of the texts, where 'portions' are determined in accordance with the structural role of the sentences that constitute the text. We expect such a system to substantially reduce users' effort to narrow down search results, whose amount may be overwhelming if only one or two keywords are specified. Consider sentences in the abstracts were classified a priori by their roles, say, Background and Objectives, (Experimental) Methods, (Experimental) Results, and Conclusions.

Generally speaking, the goal of the user is closely related to some of these "sections," but not to the rest. For instance, if a clinician intends to find whether an effect of a chemical substance on a disease is known or not, she can ask the search engine for the passages in which the names of the substance and the disease co-occur, but only in the sentences from the Results and the Conclusions sections. Such a restriction is not easily attainable by simply adding extra query terms. Furthermore, it is often not immediately evident what extra keywords are effective for narrowing down the results. Specifying target sections may be helpful in such a case as well.

A problem in constructing such a system is how to deduce the sectioning of each abstract text in the large corpus of MEDLINE. Due to the size of the corpus, it is not viable to manually assign a section label to each sentence; we are hence forced to seek for a way to automate this process. In our previous work [11, 13], we have reported preliminary results concerning the use of text classification techniques for inferring the label of the sentences but with no emphasis on any specific application. This paper reports the extension of the work with more focus on its application to the search system for MEDLINE.

The main topics of the paper are (1) how to reduce reliance on human supervision in making training data for the sentence classifier, and (2) what classes, or sections, should be presented to the users to which they can restrict search. This decision must be made on account of the trade-off between usability and accuracy of sentence classification. Another topic also addressed is (3) what types of features are effective for classification.

2 Statistics on the MEDLINE abstracts

Our method of classifying sentences into sections relies on the "structured abstracts" contained in MEDLINE. Since these abstracts have explicit sections in its text, we use them as the training data for constructing the sentence classifiers for the rest of the abstracts. Because the quality of training data affects the performance of resulting classifiers, we first analyze and present some statistics on the structured abstracts as contained in MEDLINE, as well as their impact on the design of the system.

2.1 Structured abstracts

Since its proposal in 1987, a growing number of biological and medical journals have begun to adopt so-called "structured abstracts" [1]. These journals require authors to divide the abstract text into sections that reflect the structure of the text, such as BACKGROUND, OBJECTIVES, and CONCLUSIONS. The sectioning schemes are sometimes regulated by the journals, and sometimes left to the choice of the authors. As

Table 1. Ratio of structured and unstructured abstracts in MEDLINE 2002.

	# of abstracts /	%
Structured	374,585 /	6.0%
Unstructured	5,912,271 /	94.0%
Total	11,299,108 /	100.0%

Table 2. Frequency of individual sections in the structured abstracts in MEDLINE 2002.

Sections	# of abstracts	# of sentences
CONCLUSION(S)	352,153	246,607
RESULTS	324,479	1,378,785
METHODS	209,910	540,415
BACKGROUND	120,877	264,589
OBJECTIVE	165,972	166,890
⋮	⋮	⋮
Total		2,597,286

the sections in these structured abstracts are explicitly marked with a heading (usually written in all upper-case letters), this allows us to identify a heading as a category label for the sentences that follow. Unfortunately, the number of unstructured abstracts in the MEDLINE database far exceeds that of structured abstracts (Table 1). Tables 2 and 3 respectively show the frequencies of individual headings as well as sectioning schemes.

2.2 Section headings = categories?

The fact that unstructured abstracts form a majority leads to the idea of automatically labeling each sentences when the abstracts are unstructured. This labeling process can be formulated as a text categorization task if we fix a set of sections (categories) into which the sentences should be classified. The problem remains what categories, or sections, must be presented to the users to specify the portion of the abstract texts to which search should be restricted. It is natural to choose the categories from the section headings occurring in the structured abstracts, as it will allow us to use those abstract texts to train the sentence classifiers. However, there are more than 6,000 distinct headings in MEDLINE 2002.

To maintain usability, the number of sections offered to the user must be kept as small as possible, but not too small as to render the facility useless. But then, if we restrict the number of categories, how should a section in a structured abstract be treated when its heading does not match any of the categories presented to the users? If the selection of the category set were sensible, most sections translate into a selected category in a straightforward way, such as identifying “OBJECTIVES” and “PURPOSE” sections is generally admissible. But there are headings such as “BACKGROUND AND PURPOSES.” If BACKGROUND and PURPOSES were two distinct categories presented to the user, which we believe is a sensible decision, we would have to determine which of these two classes each sentence in the section belongs to. Therefore, at least some of the sentences in the structured abstracts must go through the same labeling

Table 3. Frequency of sectioning schemes (# of abstracts). Percentages show the frequency relative to the total number of structured abstracts. Schemes marked with '*' and '†' are used for the experiment in Section 3.3.

Rank	# /	%	Section sequence
1	61,603 /	16.6%	BACKGROUND / METHOD(S) / RESULTS / CONCLUSION(S)
*2	54,997 /	14.7%	OBJECTIVE / METHOD(S) / RESULTS / CONCLUSION(S)
*3	25,008 /	6.6%	PURPOSE / METHOD(S) / RESULTS / CONCLUSION(S)
4	11,412 /	3.0%	PURPOSE / MATERIALS AND METHOD(S) / RESULTS / CONCLUSION(S)
†5	8,706 /	2.3%	BACKGROUND / OBJECTIVE / METHOD(S) / RESULTS / CONCLUSION(S)
6	8,321 /	2.2%	OBJECTIVE / STUDY DESIGN / RESULTS / CONCLUSION(S)
7	7,833 /	2.1%	BACKGROUND / METHOD(S) AND RESULTS / CONCLUSION(S)
*8	7,074 /	1.9%	AIM(S) / METHOD(S) / RESULTS / CONCLUSION(S)
9	6,095 /	1.6%	PURPOSE / PATIENTS AND METHOD(S) / RESULTS / CONCLUSION(S)
10	4,087 /	1.1%	BACKGROUND AND PURPOSE / METHOD(S) / RESULTS / CONCLUSION(S)
⋮	⋮	⋮	⋮
Total	374,585 /	100.0%	

process we use for unstructured abstracts, namely, when sectioning does not coincide with the categories presented to the users.

Even when the section they belong to has a heading that seems straightforward to assign a class, there are cases in which we have to classify sentences in a structured abstract. The above mentioned OBJECTIVE (or PURPOSE) class is actually one such category that needs sentence-wise classification. Below, we will further analyze this case.

As Table 3 shows, the most frequent *sequences* of headings are (1) BACKGROUND, METHOD(S), RESULTS, and CONCLUSION(S), followed by (2) OBJECTIVE, METHOD(S), RESULTS, and CONCLUSION(S). Inspecting abstract texts that conform to formats (1) and (2), we found that in the BACKGROUND and OBJECTIVE sections, most of these texts actually contain both the sentences describing the research background, and those describing the research objectives.

We can verify this claim by computing Sibson's information radius (Jensen-Shannon divergence) [6] for each sections. Information radius D_{JS} between two probability distributions $p(x)$ and $q(x)$ is defined as follows, using Kullback-Leibler divergence D_{KL}.

$$
\begin{aligned}
D_{\mathrm{JS}}(p\|q) &= \frac{1}{2}\left[D_{\mathrm{KL}}\left(p \,\middle\|\, \frac{p+q}{2}\right) + D_{\mathrm{KL}}\left(q \,\middle\|\, \frac{p+q}{2}\right)\right] \\
&= \frac{1}{2}\left[\sum_x p(x)\log\frac{p(x)}{\big(p(x)+q(x)\big)/2} + \sum_x q(x)\log\frac{q(x)}{\big(p(x)+q(x)\big)/2}\right].
\end{aligned}
$$

Hence, information radius is a measure of dissimilarity between distributions. It is symmetric in p and q, and is always well-defined, which are not always the case with D_{KL}.

Table 4. Information radius between the sections.

(a) Word bigrams

Class	BACKGROUND	OBJECTIVE	METHODS	RESULTS	CONCLUSION(S)
BACKGROUND	0	**0.1809**	0.3064	0.3152	0.2023
OBJECTIVE	**0.1809**	0	0.2916	0.3256	0.2370
METHODS	0.3064	0.2916	0	0.2168	0.3201
RESULTS	0.3152	0.3256	0.2168	0	0.2703
CONCLUSIONS	0.2023	0.2370	0.3201	0.2703	0

(b) Word unigrams and bigrams

Class	BACKGROUND	OBJECTIVE	METHODS	RESULTS	CONCLUSION(S)
BACKGROUND	0	**0.1099**	0.2114	0.2171	0.1202
OBJECTIVE	**0.1099**	0	0.1965	0.2221	0.1465
METHODS	0.2114	0.1965	0	0.1397	0.2201
RESULTS	0.2171	0.2221	0.1397	0	0.1847
CONCLUSIONS	0.1202	0.1465	0.2201	0.1847	0

Table 4 shows that the sentences under the BACKGROUND and OBJECTIVE sections have similar distributions of word bigrams as well as the combination of words and word bigrams. Also note the smaller divergence between these classes (bold faced figures), compared with those for the other class pairs. The implication is that these two headings are not reliable as separate category labels.

3 Classifier design

3.1 The number and the types of categories

In our previous work [11, 13], we used five categories, BACKGROUND, OBJECTIVE, METHOD(S), RESULTS, and CONCLUSION(S), based on the frequency of individual headings (Table 2). We believe this to be still a reasonable choice considering the usability of the system and the ambiguity arising from limiting the number of classes. For example, the BACKGROUND and the OBJECTIVE section headings are unreliable to be taken as a category label for the sentences in the sections, as we mentioned earlier. Nevertheless, it is not acceptable to merge them as a single class, although it might the classification task much easier. Merging them would deteriorate the usefulness of the system, since they are quite different in their structural roles, which, as a result, have different utility according to search purposes.

3.2 Support Vector Machines and feature representation

Following our previous work, soft-margin Support Vector Machines (SVMs) [2, 12] were used as the classifier for each categories. We first construct SVM classifiers for each of the BACKGROUND, OBJECTIVE, METHODS, RESULTS, and CONCLUSIONS classes, using the one-versus-rest configuration. Since SVM is a binary classifier while our task involves five classes, we combine the results of these classifiers as follows: the class i assigned to a given test example x is the one the one represented by

the SVM whose value of $f_i(x)$ is the largest, where $f_i(x)$ is a decision function of SVM for the *i*-th class, i.e., the signed distance from the optimal hyperplane after the margin width is normalized to 1.

The basis of our feature representations is words and word bigrams. In the previous work [11, 13], we have used non-contiguous sequential word patterns as features. We use word bigrams here instead of sequential patterns due to a practical reason: the speed of the feature construction is prohibitive because of the large number of documents to process.

3.3 Contextual information

Since we are interested in labeling a *series* of sentences, it is expected that incorporating contextual information into the feature set will improve classification performance. For example, it is unlikely that experimental results (RESULTS) are presented before the description of experimental design (METHODS). Thus, knowing that preceding sentences have been labeled as METHODS conditions the probability of the present sentence being classified as RESULTS. And the sentences of the same class have high probability of appearing consecutively; we would not expect the authors to interleave sentences describing experimental results (RESULTS) with those in the CONCLUSIONS and OBJECTIVES classes.

Since it is not clear what kind of contextual information performs best, the following types of contextual representation were examined in an experiment (Section 4.1).

1. The class of the previous sentence.
2. The classes of the previous two sentences.
3. The class of the next sentence.
4. The classes of the next two sentence.
5. Relative location of the current sentence in the abstract text.
6. The word features of the previous sentence.
7. The word features of the next sentence.
8. The word features of the previous and the next sentences.
9. The number of previous sentences having the same class as the previous sentence, and its class.

4 Experiments

This section reports the results of preliminary experiments that we conducted to examine the performance of the classifiers used for labeling sentences.

4.1 Contextual information

In this experiment, structured abstracts from MEDLINE 2002 were used. The classes we considered (or, sections to which sentences are classified) are OBJECTIVE(S), METHOD(S), RESULT(S), and CONCLUSION(S). Note that this set does not coincide with the five classes we employed in the final system. According to Table 3, the

Table 5. Performance of context features

Features	Accuracy (%) sentence	abstract
(0) No context features	83.6	25.0
(1) The class of the previous sentence	88.9	48.9
(2) The classes of the previous two sentences	89.9	50.6
(3) The class of the next sentence	88.9	50.9
(4) The classes of the next two sentences	89.3	51.2
(5) Relative location of the current sentence	91.9	50.7
(6) The word features of the previous sentence	87.3	37.5
(7) The word features of the next sentence	88.1	39.0
(8) The word features of the previous and the next sentences	89.7	46.4
(9) The number of preceding sentences having the same class as the previous sentence, and its class	90.6	50.9

section sequence consisting of these sections are only second after the sequence BACKGROUND / METHOD(S) / RESULT(S) / CONCLUSION(S). However, identifying the sentences with headings PURPOSE(S) and AIM(S) with those with OBJECTIVE(S) makes the corresponding sectioning scheme the most frequent.

Hence, we collected structured abstracts whose heading sequences matches the following patterns:

1. OBJECTIVE(S) / METHOD(S) / RESULTS / CONCLUSION(S),
2. PURPOSE(S) / METHOD(S) / RESULTS / CONCLUSION(S),
3. AIM(S) / METHOD(S) / RESULTS / CONCLUSION(S).

We split each of these abstracts into sentences using UIUC Sentence Splitter [9], after removing all symbols and replacing every contiguous sequence of numbers with a single symbol '#'. After sentence splitting, we filtered out the abstracts that produced a sentence with less than three words, regarding it as a possible error in sentence splitting. This yielded a total of 82,936 abstracts.

To reduce the number of features, we only took into account word bigrams occurring in at least 0.05% of the sentences, which amounts to 9,078 bigrams. The number of (unigram) word features was 104,733.

We obtained 103,962 training examples (sentences) from 10,000 abstracts randomly sampled from the set of 82,936 structured abstracts described above, and 10,356 test examples (sentences) from 1,000 abstracts randomly sampled from the rest of the set.

The quadratic kernel is used with SVMs, and the optimal soft margin (or capacity) parameter C is sought for each of the SVMs using different context features. The results are listed in Table 5.

There were not much differences in the performance of contextual features as far as accuracy were measured on a per-sentence basis. All contextual features (1)–(9) obtained about 90% accuracy, which is an improvement of 4 to 8% over (0) when no context features were used. In contrast, the performance on a per-abstract basis, in which a classification of an abstract is judged to be correct only if all the constituent sentences are correctly classified, varied between 50% and 37.5%. The maximum performance of

51%, which is 25% improvement over the baseline (0) in which no context feature was used, was obtained for features (3), (4), and (5).

4.2 Separating 'Objectives' from 'Background'

The analysis in the Section 2.2 suggest that it is unreliable to use the headings BACKGROUND and OBJECTIVE(S) as the labels of the sentences in the sections, because the BACKGROUND section frequently contains sentences that should rather be classified as OBJECTIVES and vice versa. Yet, it is not acceptable to merge them as a single class, because they are quite different in their structural roles; doing so would severely impairs the utility of the system.

To resolve this situation, we construct an SVM classifier to distinguish between these classes again. To train this classifier, we use the sentences in the structured abstracts that contain both the BACKGROUND and the OBJECTIVES sections (such as in the scheme marked with a dagger in Table 3).

To assess the feasibility of this approach, we collected 11,898 abstracts that contain both the BACKGROUND and the OBJECTIVE(S) headings. The texts in this collection were preprocessed in an identical manner as the previous subsection, and the number of sentences in the BACKGROUND and the OBJECTIVES sections from this collection was 34,761. The classification of individual sentences with SVMs exhibited an F1-score of 96.4 (which factors into a precision of 95.6% and a recall of 97.2%), on average over 10-fold cross validation trials. The SVMs used quadratic kernels, and used the bag-of-words-and-word-bigrams features only. No context features were used.

5 A prototype implementation

Using the feature set described in Section 3.2 as well as the context feature (5) of Section 3.3, we constructed five SVM classifiers for each of the five sections, BACKGROUND, OBJECTIVES, METHODS, RESULTS, and CONCLUSIONS. With these SVMs, we labeled the sentences in the unstructured abstracts in MEDLINE 2003 whose publication year is 2001 and 2002. The same labeling process is applied to the sentences in structured abstracts as well, but only when their section heading when the correspondence to any of the above five sections is not evident. We also classified each sentence in the BACKGROUND and OBJECTIVE (and equivalent) sections into one of the BACKGROUND and OBJECTIVE classes using the classifier of Section 4.2, when the structured abstract contained only one of them.

We implemented an experimental search system for these labeled data using PHP on top of an Apache web server. The full-text retrieval engine Namazu was used as a back-end search engine. The screen shot for the service page is shown in Figure 1. The form on the page contains a field for entering query terms, a 'Go' button as well as radio buttons marked 'Any' and 'Select from' for choosing whether the keyword search should be performed on the whole abstract texts, or on limited sections. Plain keywords, phrases (specified by enclosing the phrase in braces), and boolean conjunction ('and'), disjunction ('or'), and negation ('not') are allowed for query field. If the user chooses 'Select from' button rather than 'Any,' the check boxes on its right are

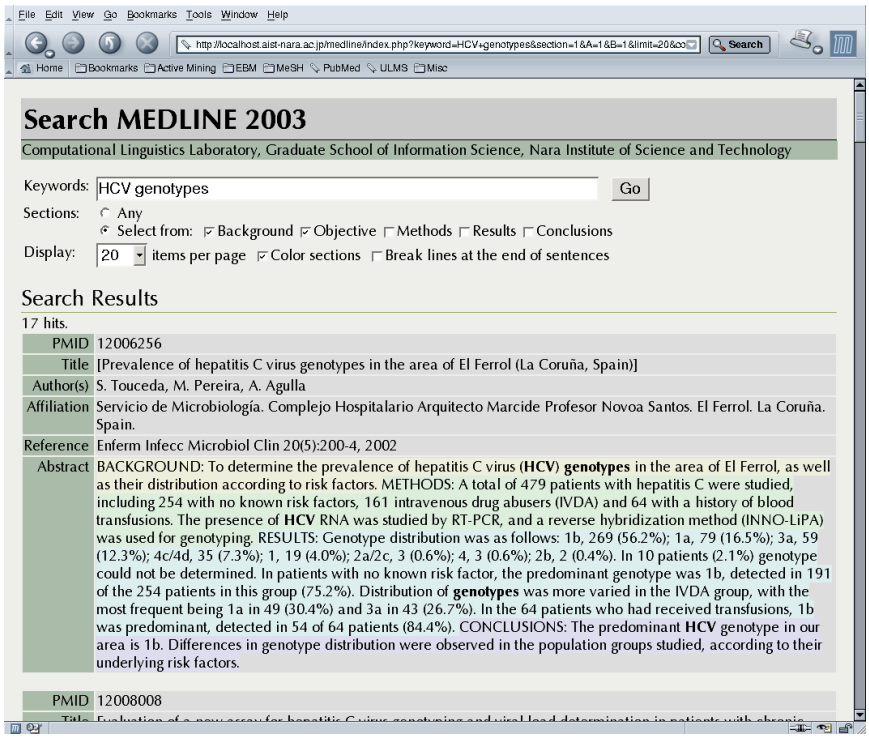


Fig. 1. A screen shot.

activated. These boxes corresponds to the five target sections, namely, ‘Background,’ ‘Objectives,’ ‘Methods,’ ‘Results,’ and ‘Conclusions.’

Matching query terms found in the abstract text are highlighted in bold face letters, and the sections (either deduced from headings or from the content of the sentence with automatic classifier) are shown in different background colors.

6 Conclusions and future work

We have reported the first step towards construction of a search system for the MEDLINE database that allows the users to exploit the underlying structure of the abstract text. The implemented system, however, is only experimental, and surely needs more elaboration.

First of all, the adequacy of five sections presented to the user needs evaluation. In particular, OBJECTIVE and CONCLUSIONS are different as they each describes what has been sought and what is really achieved, respectively, but they are the same in the sense that they provides a summary of what the paper deals with. They are not about the details of experiments, and not about what is done elsewhere. Thus grouping them into one class might be sufficient for most users.

We plan to incorporate re-ranking procedure of label sequences based on the overall consistency of the sequences. By 'consistency' here, we mean the constraint on the sequences such as it is unlikely that conclusions appear in the beginning of the text, and the same section seldom occur twice in a text. The similar lines of research [4, 10] have been reported recently in ML and NLP communities, in which the sequence of classification results is optimized over all possible sequences. We also plan to incorporate features that reflect cohesion or coherence between sentences [3].

Acknowledgment

This research was supported in part by MEXT under Grant-in-Aid for Scientific Research on Priority Areas (B) no. 759. The first author is also supported in part by MEXT under Grant-in-Aid for Young Scientists (B) no. 15700098.

References

[1] Ad Hoc Working Group for Critical Appraisal of Medical Literature. A proposal for more informative abstracts of clinical articles. *Annals of Internal Medicine*, 106(4):598–604, 1987.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[3] M. A. K. Halliday and Ruqaiya Hasan. *Cohesion in English*. Longman, London, 1976.

[4] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML-2001)*, pages 282–289. Morgan Kaufmann, 2001.

[5] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[6] L. Lee. Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Comutational Linguistics (ACL-99)*, pages 25–32, 1999.

[7] MEDLINE. http://www.nlm.nih.gov/databases/databases_medline.html, 2002–2003. U.S. National Library of Medicine.

[8] PubMed. http://www.ncbi.nlm.nih.gov/PubMed/, 2003. U.S. National Library of Medicine.

[9] Sentence splitter software. http://l2r.cs.uiuc.edu/~cogcomp/cc-software.htm, 2001. University of Illinois at Urbana-Champaign.

[10] Fei Sha and Fernando Pereira. Shallow parsing with conditional random fields. In *Proceedings of the Human Language Technology Conference North American Chapter of Association for Computational Linguistics (HLT-NAACL 2003)*, pages 213–220, Edmonton, Alberta, Canada, 2003. Association for Computational Linguistics.

[11] Masashi Shimbo, Takahiro Yamasaki, and Yuji Matsumoto. Automatic classification of sentences in the MEDLINE abstracts. In *Proceedings of the 6th Sanken (ISIR) International Symposium*, pages 135–138, Suita, Osaka, Japan, 2003.

[12] Vladimir Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[13] Takahiro Yamasaki, Masashi Shimbo, and Yuji Matsumoto. Automatic classification of sentences using sequential patterns. Technical Report of IEICE AI2002-83, The Institute of Electronics, Information and Communication Engineers, 2003. In Japanese.

RULE-BASED CHASE ALGORITHM FOR PARTIALLY INCOMPLETE INFORMATION SYSTEMS

AGNIESZKA DARDZIŃSKA-GLEBOCKA[+,*] *and* ZBIGNIEW W. RAŚ[*,°]
[*] University of North Carolina, Department of Computer Science
Charlotte, N.C. 28223, USA
[°] Polish Academy of Sciences, Institute of Computer Science
Ordona 21, 01-237 Warsaw, Poland
[+]Bialystok University of Technology, Department of Mathematics
15-351 Bialystok, Poland
ras@uncc.edu or adardzin@uncc.edu

Abstract. A rule-based chase algorithm used to discover values of incomplete attributes in a database is described in this paper. To begin the chase process, each attribute that contains unknown or partially incomplete values is set, one by one, as a decision attribute and all other attributes in a database are treated as condition attributes for that decision attribute. Now, assuming that d is a decision attribute, any object x such that $d(x) \neq NULL$ is used in the process of extracting rules describing d. In the next step, each incomplete database field in a column corresponding to attribute d is chased with respect to previously extracted rules describing d. All other incomplete attributes in a database are processed the same way.

1 Introduction

Common problems encountered by Query Answering Systems QAS, introduced by Ras[8], [9], either for Information Systems S or for Distributed Autonomous Information Systems $DAIS$ include the handling of incomplete attributes when answering a query. One plausible solution to answer a query involves the generation of rules describing all incomplete attributes used in a query and then chasing the unknown values in the local database with respect to the generated rules. These rules can be given by domain experts but also can be discovered locally or at remote sites of $DAIS$. Since all unknown values would not necessarily be found, the process is repeated on the enhanced database until all unknowns are found or no new information is generated. When the fixed point is reached by this process, QAS will run the original query against the enhanced database. The chase algorithm presented in [2] was based only on consistent set of rules. The notion of a tableaux system and the chase algorithm based on functional dependencies F is presented for instance in [1]. Chase algorithm based on F always terminates if applied to a finite tableaux system. Also, it was shown that, if one execution of the algorithm generates a tableaux system that satisfies F, then every execution of the algorithm generates the

same tableaux system. Using Chase algorithm for predicting what attribute value should replace an incomplete value has a clear advantage over many other methods for predicting incomplete values mainly because of the use of existing associations between values of attributes. To find these associations we can use either any association rule mining algorithm or any rule discovery algorithm like $LERS$ [5] or $Rosetta$ [11]. Unfortunately, these algorithms, including Chase algorithm presented by us in [3], do not handle partially incomplete data, where $a(x)$ is equal, for instance, to $\{(a_1, \frac{1}{4}), (a_2, \frac{1}{4}), (a_3, \frac{1}{2})\}$. Clearly, we assume here that a is an attribute, x is an object, and $\{a_1, a_2, a_3\} \subseteq V_a$. By V_a we mean the set of values of attribute a. The weights assigned to these 3 attribute values should be read as: the confidence that $a(x) = a_1$ is $\frac{1}{4}$, the confidence that $a(x) = a_2$ is $\frac{1}{4}$ and, the confidence that $a(x) = a_3$ is $\frac{1}{2}$.

In this paper we present a new chase algorithm (called **Chase2**) which can be used for chasing incomplete information systems with rules which do not have to be consistent (this assumption was required by **Chase1** algorithm [2]. Also, we show how to compute a confidence of inconsistent rules and how it can be normalized. These rules are used by **Chase2**.

2 Handling Incomplete Values using Chase Algorithms

There is a relationship between interpretation of queries and the way the incomplete information in an information system is seen. Assume, for example, that we are concerned with identifying all objects in the system satisfying a given description. For example an information system might contain information about students in a class and classify them using four attributes of *hair_color*, *eye_color*, *gender* and *size*. A simple query might be to find all students with *brown_hair* and *blue_eyes*. When the information system is incomplete, students having brown hair and unknown eye color can be handled by either including or excluding them from the answer to the query. In the first case we talk about optimistic approach to query interpretation while in the second case we talk about pessimistic approach. Another option to handle such a query is to discover rules for *eye_color* in terms of the attributes *hair_color*, *gender*, and *size*. Then, these rules can be applied to students with unknown *eye_color* to discover that color and possibly to identify more objects satisfying the query.

Consider that in our example one of the generated rules said:
$(hair\,brown) \wedge (size\,medium) \rightarrow (eye\,brown)$.

Thus, if one of the students having *brown_hair* and *medium_size* has no value for *eye_color*, then the student should not be included in the list of students with *brown_hair* and *blue_eyes*. Attributes *hair_color* and *size* are

classification attributes and eye *color* is the decision attribute.

Now, let us give example showing the relationship between incomplete information about objects in an information system and the way queries (attribute values) are interpreted. Namely, the confidence in object x that he has $brown\_color\_of\_hair$ is $\frac{1}{3}$ can be either written as $(brown, \frac{1}{3}) \in color\_of\_hair(x)$ or $(x, \frac{1}{3}) \in I(brown)$, where I is an interpretation of queries (the term *brown* is treated here as a query).
In [2] we presented **Chase1** strategy based on the assumption that only consistent subsets of rules extracted from an incomplete information system S can be used for replacing Null values by new less incomplete values in S. Clearly, rules discovered from S do not have to be consistent in S. Taking this fact into consideration, the Chase algorithm (**Chase2**) proposed in this paper has less restrictions and it allows chasing information system S with inconsistent rules as well.

Assume that $S = (X, A, V)$, where $V = \bigcup\{V_a : a \in A\}$ and each $a \in A$ is a partial function from X into $2^{V_a} - \{\emptyset\}$.

In the first step of Chase algorithms, we identify all incomplete attributes used in S. An attribute is incomplete if there is an object in S with incomplete information on this attribute. The values of all incomplete attributes in S are treated as concepts to be learned (in a form of rules) either only from S or from S and its remote sites (if S is a part of a distributed autonomous information system).

The second step of Chase algorithm is to extract rules describing these concepts. These rules are stored in a knowledge base D for S (see [8], [9], [10]). The algorithm **Chase1** presented in [3] assumes that all inconsistencies in D have to be repaired before they are used in the chase process. Rules describing attribute value v_a of attribute a are extracted from the subsystem $S_1 = (X_1, A, V)$ of S where $X_1 = \{x \in X : card(a(x)) = 1\}$. **Chase2** does not have such restrictions placed on D.

The final step of Chase algorithms is to replace incomplete information in S by less incomplete information provided by rules from D.

3 Partially Incomplete Information Systems

We say that $S = (X, A, V)$ is a partially incomplete information system of type λ, if S is an incomplete information system and the following three conditions hold:

- $a_S(x)$ is defined for any $x \in X$, $a \in A$,

- $(\forall x \in X)(\forall a \in A)[(a_S(x) = \{(a_i, p_i) : 1 \leq i \leq m\}) \rightarrow \sum p_i = 1]$,
- $(\forall x \in X)(\forall a \in A)[(a_S(x) = \{(a_i, p_i) : 1 \leq i \leq m\}) \rightarrow (\forall i)(p_i \geq \lambda)]$.

Now, given two partially incomplete information systems S_1, S_2 classifying the same sets of objects (let's say objects from X) using the same sets of attributes (let's say A), we assume that $a_{S_1}(x) = \{(a_{1,i}, p_{1,i}) : i \leq m_1\}$ and $a_{S_2}(x) = \{(a_{2,i}, p_{2,i}) : i \leq m_2\}$.

We say that containment relation Ψ holds between S_1 and S_2, if the following two conditions hold:

- $(\forall x \in X)(\forall a \in A)[card(a_{S_1}(x)) \geq card(a_{S_2}(x))]$,
- $(\forall x \in X)(\forall a \in A)[card[(a_{S_1}(x)) = card(a_{S_2}(x))] \rightarrow \sum_{i \neq j} \mid p_{2,i} - p_{2,j} \mid > \sum_{i \neq j} \mid p_{2,i} - p_{2,j} \mid]$.

If containment relation Ψ holds between systems S_1 and S_2, both of type λ, we say that information system S_1 was mapped onto S_2 by containment mapping Ψ and denote that fact as $\Psi(S_1) = S_2$ which means that $(\forall x \in X)(\forall a \in A)[\Psi(a_{S_1}(x)) = \Psi(a_{S_2}(x))]$. We also say that containment relation Ψ holds between $a_{S_1}(x)$ and $a_{S_2}(x)$, for any $x \in X$, $a \in A$.

So, the containment mapping Ψ for incomplete information systems does not increase the number of possible attribute values for a given object. If the number of possible values of a given attribute assigned to an object is not changed, then the average difference in confidence assigned to these attribute values has to increase.

Let us take an example of two systems S_1 , S_2 of type $\lambda = \frac{1}{4}$ (see Table 1 and Table 2).

It can be easily checked that values $a(x_3)$,$a(x_8)$,$b(x_2)$,$c(x_2)$,$c(x_7)$, $e(x_4)$ are different in S_1 than in S_2. In each of these six cases, the attribute value assigned to an object in S_2 is less general that in S_1. It means that $\Psi(S_1) = S_2$.

Assume now that $L(D) = \{(t \rightarrow v_c) \in D : c \in In(A)\}$ is the set of all rules extracted from S by **ERID**$(S, \lambda_1, \lambda_2)$, where λ_1, λ_2 are thresholds for minimum support and minimum confidence, correspondingly. **ERID** is the algorithm for discovering rules from incomplete information systems presented in [3]. Finally, let us assume that by $N_S(t)$ we mean the interpretation of term t in $S = (X, A, V)$ defined as:

- $N_S(v) = \{(x, p) : (v, p) \in a(x)\}$, for any $v \in V_a$,
- $N_S(t_1 + t_2) = N_S(t_1) \oplus N_S(t_2)$,

X	a	b	c	d	e
x_1	$(a_1, \frac{1}{3})$,$(a_2, \frac{2}{3})$	$(b_1, \frac{2}{3})$,$(b_2, \frac{1}{3})$	c_1	d_1	$(e_1, \frac{1}{2})$,$(e_2, \frac{1}{2})$
x_2	$(a_2, \frac{1}{4})$,$(a_3, \frac{3}{4})$	$(b_1, \frac{1}{3})$,$(b_2, \frac{2}{3})$		d_2	e_1
x_3		b_2	$(c_1, \frac{1}{2})$,$(c_3, \frac{1}{2})$	d_2	e_3
x_4	a_3		c_2	d_1	$(e_1, \frac{2}{3})$,$(e_2, \frac{1}{3})$
x_5	$(a_1, \frac{2}{3})$,$(a_2, \frac{1}{3})$	b_1	c_2		e_1
x_6	a_2	b_2	c_3	d_2	$(e_2, \frac{1}{3})$,$(e_3, \frac{2}{3})$
x_7	a_2	$(b_1, \frac{1}{4})$,$(b_2, \frac{3}{4})$	$(c_1, \frac{1}{3})$, $(c_2, \frac{2}{3})$	d_2	e_2
x_8		b_2	c_1	d_1	e_3

Table 1: System S_1

X	a	b	c	d	e
x_1	$(a_1, \frac{1}{3})$,$(a_2, \frac{2}{3})$	$(b_1, \frac{2}{3})$,$(b_2, \frac{1}{3})$	c_1	d_1	$(e_1, \frac{1}{2})$,$(e_2, \frac{1}{2})$
x_2	$(a_2, \frac{1}{4})$,$(a_3, \frac{3}{4})$	b_1	$(c_1, \frac{1}{3})$,$(c_2, \frac{2}{3})$	d_2	e_1
x_3	a_1	b_2	$(c_1, \frac{1}{2})$,$(c_3, \frac{1}{2})$	d_2	e_3
x_4	a_3		c_2	d_1	e_2
x_5	$(a_1, \frac{2}{3})$,$(a_2, \frac{1}{3})$	b_1	c_2		e_1
x_6	a_2	b_2	c_3	d_2	$(e_2, \frac{1}{3})$,$(e_3, \frac{2}{3})$
x_7	a_2	$(b_1, \frac{1}{4})$,$(b_2, \frac{3}{4})$	c_1	d_2	e_2
x_8	$(a_1, \frac{2}{3})$,$(a_2, \frac{1}{3})$	b_2	c_1	d_1	e_3

Table 2: System S_2

- $N_S(t_1 * t_2) = N_S(t_1) \otimes N_S(t_2)$

where, for any $N_S(t_1) = \{(x_i, p_i)\}_{i \in I}$, $N_S(t_2) = \{(x_j, q_j)\}_{j \in J}$ we have:

- $N_S(t_1) \oplus N_S(t_2) = \{(x_j, p_j)\}_{j \in J-I} \cup \{(x_i, p_i)\}_{i \in I-J} \cup \{(x_i, max(p_i, q_i))\}_{i \in I \cap J}$,
- $N_S(t_1) \otimes N_S(t_2) = \{(x_i, p_i \cdot q_i)\}_{i \in I \cap J}$

Algorithm **Chase2**, given below, converts information system S of type λ to a new more complete information system **Chase2**(S).

Algorithm Chase2$(S, In(A), L(D))$
Input
System $S = (X, A, V)$,
Set of incomplete attributes $In(A) = \{a_1, a_2, ..., a_k\}$,
Set of rules $L(D)$.
Output
System **Chase2**(S)
begin $j := 1$;
while $j \leq k$ **do**
begin
$S_j := S$;
for all $x \in X$ **do**
$p_j := 0$;
begin
$b_j(x) := \emptyset$;
$n_j := 0$;
for all $v \in V_{a_j}$ **do**
if $card(a_j(x)) \neq 1$ and $\{(t_i \rightarrow v) : i \in I\}$
is a maximal subset of rules from $L(D)$
such that $(x, p_i) \in N_{S_j}(t_i)$ **then**
if $\sum_{i \in I}[p_i \cdot conf(t_i \rightarrow v) \cdot sup(t_i \rightarrow v)] \geq \lambda$ **then**
begin
$b_j(x) := b_j(x) \cup \{(v, \sum_{i \in I}[p_i \cdot conf(t_i \rightarrow v) \cdot sup(t_i \rightarrow v)])\}$;
$n_j := n_j + \sum_{i \in I}[p_i \cdot conf(t_i \rightarrow v) \cdot sup(t_i \rightarrow v)]$;
end
$p_j := p_j + n_j$;
end
if $\Psi(a_j(x)) = [b_j(x)/p_j]$
(containment relation holds between $a_j(x)$, $[b_j(x)/p_j]$)
then $a_j(x) := [b_j(x)/p_j]$;
$j := j + 1$;

end
$S := \prod\{S_j : 1 \leq j \leq k\}$
(see the definition of $S := \prod\{S_j : 1 \leq j \leq k\}$ below)
Chase2$(S, In(A), L(D))$
end

Definitions:
To define S it is enough to assume that $a_S(x) =$(if $a = a_j$ then $a_{S_j}(x)$ for any attribute a and object x).
Also, if $b_j(x) = \{(v_i, p_i)\}_{i \in I}$, then $[b_j(x)/p]$ is defined as $\{(v_i, p_i/p)\}_{i \in I}$.

To explain the algorithm, we apply **Chase2** to the information system given in Table 3. We assume that $L(D)$ contains the following rules (listed with their support and confidence):
$r_1 = [a_1 \rightarrow e_3]$, $sup(r_1) = 1$, $conf(r_1) = 0.5$
$r_2 = [a_2 \rightarrow e_2]$, $sup(r_2) = \frac{5}{3}$, $conf(r_2) = 0.51$
$r_3 = [a_3 \rightarrow e_1]$, $sup(r_3) = \frac{17}{12}$, $conf(r_3) = 0.51$
$r_4 = [b_1 \rightarrow e_1]$, $sup(r_4) = 2$, $conf(r_4) = 0.72$
$r_5 = [b_2 \rightarrow e_3]$, $sup(r_5) = \frac{8}{3}$, $conf(r_5) = 0.51$
$r_6 = [c_2 \rightarrow e_1]$, $sup(r_6) = 2$, $conf(r_6) = 0.66$
$r_7 = [c_3 \rightarrow e_3]$, $sup(r_7) = \frac{7}{6}$, $conf(r_7) = 0.64$
$r_8 = [a_3 * c_1 \rightarrow e_3]$, $sup(r_8) = 1$, $conf(r_8) = 0.8$
$r_9 = [a_3 * d_1 \rightarrow e_3]$, $sup(r_9) = 1$, $conf(r_9) = 0.5$
$r_{10} = [c_1 * d_1 \rightarrow e_3]$, $sup(r_{10}) = 1$, $conf(r_{10}) = 0.5$

Only two values $e(x_1)$, $e(x_6)$ of the attribute e can be changed. The next section shows how to compute these two values and decide if the current attribute values assigned to objects x_1, x_6 can be replaced by them. Similar process is applied to all incomplete attributes in S. After all changes of all incomplete attributes are recorded, system S is replaced by $\Psi(S)$ and the whole process is recursively repeated till some fix point is reached.

Algorithm **Chase2** will try to replace the current value of $e(x_1)$ which is $\{(e_1, \frac{1}{2}), (e_2, \frac{1}{2})\}$ by a new value $e_{new}(x_1)$ initially denoted by $\{(e_1, ?), (e_2, ?), (e_3, ?)\}$. Because $\Psi(e(x_1)) = e_{new}(x_1)$, the value $e(x_1)$ will change.

To justify our claim, let us compute $e_{new}(x)$ for $x = x_1, x_4, x_6$:
For x_1:

$(e_3, \frac{1}{3} \cdot 1 \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{8}{3} \cdot \frac{51}{100} + 1 \cdot 1 \cdot \frac{1}{2}) = (e_3, 1.119)$;
$(e_2, \frac{2}{3} \cdot \frac{5}{3} \cdot \frac{51}{100}) = (e_2, 1.621)$; $(e_1, \frac{2}{3} \cdot 2 \cdot \frac{72}{100}) = (e_1, 0.96)$.

X	a	b	c	d	e
x_1	$(a_1, \frac{1}{3})$,$(a_2, \frac{2}{3})$	$(b_1, \frac{2}{3})$,$(b_2, \frac{1}{3})$	c_1	d_1	$(e_1, \frac{1}{2})$,$(e_2, \frac{1}{2})$
x_2	$(a_2, \frac{1}{4})$,$(a_3, \frac{3}{4})$	$(b_1, \frac{1}{3})$,$(b_2, \frac{2}{3})$		d_2	e_1
x_3	a_1	b_2	$(c_1, \frac{1}{2})$,$(c_3, \frac{1}{2})$	d_2	e_3
x_4	a_3		c_2	d_1	$(e_1, \frac{2}{3})$,$(e_2, \frac{1}{3})$
x_5	$(a_1, \frac{2}{3})$,$(a_2, \frac{1}{3})$	b_1	c_2		e_1
x_6	a_2	b_2	c_3	d_2	$(e_2, \frac{1}{3})$,$(e_3, \frac{2}{3})$
x_7	$(a_2$	$(b_1, \frac{1}{4})$,$(b_2, \frac{3}{4})$	$(c_1, \frac{1}{3}, (c_2, \frac{2}{3})$	d_2	e_2
x_8	a_3	b_2,	c_1	d_1	e_3

Table 3: System S

So, we have:

$$e_{new}(x_1) = \{(e_1, \tfrac{0.96}{0.96+1.621+1.119}), (e_2, \tfrac{1.621}{0.96+1.621+1.119}), (e_3, \tfrac{1.119}{0.96+1.621+1.119})\} = \{(e_1, 0.26), (e_2, 0.44), (e_3, 0.302)\}.$$

Because, the confidence assigned to e_1 is below the threshold λ, then only two values remain: $(e_2, 0.44), (e_3, 0.302)$. So the value of attribute e assigned to x_1 is $\{(e_2, 0.59), (e_3, 0.41)\}$.

For x_4:

$$(e_3, 1 \cdot 1 \cdot \tfrac{1}{2}) = (e_3, 0.5);\ (e_2, 0);\ (e_1, 1 \cdot \tfrac{17}{12} \cdot \tfrac{51}{100}) = (e_1, 0.7225).$$

So, we have:

$$e_{new}(x_4) = \{(e_1, \tfrac{0.7225}{0.5+0.7225}), (e_3, \tfrac{0.5}{0.5+0.7225})\} = \{(e_1, 0.59), (e_3, 0.41)\}$$

which means, the value of attribute e assigned to x_4 remains unchanged.

For x_6:

$$(e_3, \tfrac{8}{3} \cdot 1 \cdot \tfrac{51}{100} + 1 \cdot \tfrac{7}{6} \cdot \tfrac{64}{100}) = (e_3, 2.11);\ (e_2, 1 \cdot \tfrac{5}{3} \cdot \tfrac{51}{100}) = (e_2, 0.85);\ (e_1, 0).$$

So, we have:

X	a	b	c	d	e
x_1	$(a_1, \frac{1}{3})$,$(a_2, \frac{2}{3})$	$(b_1, \frac{2}{3})$,$(b_2, \frac{1}{3})$	c_1	d_1	$(e_2, 0.59)$,$(e_3, 0.41)$
x_2	$(a_2, \frac{1}{4})$,$(a_3, \frac{3}{4})$	$(b_1, \frac{1}{3})$,$(b_2, \frac{2}{3})$		d_2	e_1
x_3	a_1	b_2	$(c_1, \frac{1}{2})$,$(c_3, \frac{1}{2})$	d_2	e_3
x_4	a_3		c_2	d_1	$(e_1, \frac{2}{3})$,$(e_2, \frac{1}{3})$
x_5	$(a_1, \frac{2}{3})$,$(a_2, \frac{1}{3})$	b_1	c_2		e_1
x_6	a_2	b_2	c_3	d_2	e_3
x_7	a_2	$(b_1, \frac{1}{4})$,$(b_2, \frac{3}{4})$	$(c_1, \frac{1}{3}, (c_2, \frac{2}{3})$	d_2	e_2
x_8	a_3	b_2	c_1	d_1	e_3

Table 4: New System S

$$e_{new}(x_6) = \{(e_2, \tfrac{0.85}{2.11+0.85}), (e_3, \tfrac{0.85}{2.11+0.85})\} = \{(e_2, 0.29), (e_3, 0.713)\}.$$

Because, the confidence assigned to e_3 is below the threshold λ, then only one value remains: $(e_3, 0.713)$. So, the value of attribute e assigned to x_6 is e_3.

The new resulting information system is presented by Table 4.

This example shows that the values of attributes stored in the resulting table depend on the threshold λ. Smaller the threshold λ, the level of incompleteness of the system will get lower.

Initial testing performed on several incomplete tables of the size $50 \times 2,000$ with randomly generated data gave us quite promising results.

4 Conclusion

We expect much better results if a single information system is replaced by distributed autonomous information systems investigated by Ras in [8, 9, 10]. Our claim is justified by experimental results showing higher confidence in rules extracted through distributed data mining than in rules extracted through local mining.

References

1. Atzeni, P., DeAntonellis, V., "Relational database theory", in *The Benjamin Cummings Publishing Company*, 1992

2. Dardzinska, A., Ras, Z.W., "Chasing Unknown Values in Incomplete Information Systems", Proceedings of the ICDM'03 *Workshop on Foundation and New Directions in Data Mining*, Melbourne, Florida, 2003, will appear
3. Dardzinska, A., Ras, Z.W., "On Rules Discovery from Incomplete Information Systems", Proceedings of the ICDM'03 *Workshop on Foundation and New Directions in Data Mining*, Melbourne, Florida, 2003, will appear
4. Grzymala-Busse, J., "On the unknown attribute values in learning from examples", in *Proceedings of ISMIS'91, LNCS/LNAI, Springer-Verlag*, Vol. 542, 1991, 368-377
5. Grzymala-Busse, J. "A new version of the rule induction system LERS", in *Fundamenta Informaticae*, Vol. 31, No. 1, 1997, 27-39
6. Kodratoff, Y., Manago, M.V., Blythe, J. "Generalization and noise", in *Int. Journal Man-Machine Studies*, Vol. 27, 1987, 181-204
7. Kryszkiewicz, M., Rybinski, H., "Reducing information systems with uncertain attributes",in *Proceedings of ISMIS'96*, LNCS/LNAI, Springer-Verlag, Vol. 1079, 1996, 285-294
8. Ras, Z., "Resolving queries through cooperation in multi-agent systems", in *Rough Sets and Data Mining* (Eds. T.Y. Lin, N. Cercone, Kluwer Academic Publishers, 1997, 239-258
9. Ras, Z., Joshi, S., "Query approximate answering system for an incomplete DKBS", in *Fundamenta Informaticae Journal*, IOS Press, Vol. 30, No. 3/4, 1997, 313-324
10. Ras, Z., "Dictionaries in a distributed knowledge-based system", in *Proceedings of Concurrent Engineering: Research and Applications Conference*, Pittsburgh, August 29-31, 1994, Concurrent Technologies Corporation, 383-390
11. Skowron, A., "Boolean reasoning for decision rules generation", in *Methodologies for Intelligent Systems, Proceedings of the 7th International Symposium on Methodologies for Intelligent Systems*, (eds. J. Komorowski, Z. Ras), Lecture Notes in Artificial Intelligence, Springer Verlag, No. 689, 1993, 295-305

Data Mining Oriented CRM Systems Based on MUSASHI: C-MUSASHI *

Katsutoshi Yada[1], Yukinobu Hamuro[2], Naoki Katoh[3], Takashi Washio[4], Issey Fusamoto[1], Daisuke Fujishima[1] and Takaya Ikeda[1]

[1] Kansai University, 3-3-35, Yamate, Suita, Osaka 564-8680, Japan
[2] Osaka Sangyo University 3-1-1 Nakagaito, Daito, Osaka, 574-8530 Japan
[3] Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
[4] Osaka University 8-1 Mihogaoka, Ibaraki, Osaka, 567-0047 Japan

Abstract. MUSASHI is a set of commands that enables us to efficiently execute various types of data manipulations in a flexible manner, mainly aiming at data processing of huge amount of data required for data mining. Data format which MUSASHI can deal with is either an XML table written in XML or plain text file with table structure. In this paper we shall present data mining oriented CRM systems based on MUSASHI, which are integrated with the marketing tools and data mining technology. Everybody can construct useful CRM systems at extremely low cost by introducing MUSASHI.

1 Introduction

MUSASHI is a set of commands developed for the processing of a large amount of data required for data mining in business field and open-source software for achieving efficient processing of XML data [1] [2] [3] [4]. We have developed a data mining oriented CRM system that runs on MUSASHI by integrating several marketing tools and data mining technology. Discussing the cases regarding simple customer management we shall describe general outlines, components, and analytical tools for CRM system which we have developed.

With the progress of deflation in recent Japanese economy, retailers in Japan are now under severe pressure. Many of these enterprises are now trying to encompass and maintain loyal customers through the introduction of FSP [5][6]. FSP (Frequent Shoppers Program) is defined as one of the CRM systems to accomplish effective sales promotion by accumulating purchase history of the customers in its own database and by recognizing the nature and the behavior of the loyal customers. However, it is very rare that CMS system such as FSP has actually contributed to successful business activities of the enterprises in recent years.

* Research of this paper is partly supported by the Grant-in-Aid for Scientific Research on Priority Areas (2), RCSS fund by the Ministry of Education, Science, Sports and Culture of Japan, and the Kansai University Special Research fund, 2002.

There are several reasons why the existing CRM system cannot contribute to the acquisition of customers and to the attainment of competitive advantage in the business. First of all, the cost to construct CRM system is very high. In fact, some of the enterprises have actually spent a large amount of money merely for the construction of data warehouse to accumulate purchase history data of the customers and, as a result, no budget is left for carrying out customer analysis.

Secondly, it happens very often that data are actually accumulated while technique, software and human resources in their firms to analyze these data are in shortage, and the analysis of the customers is not in progress. Therefore, in many cases, the enterprises simply accumulate the data but do not carry out the analysis of the customers.

In this paper, we shall introduce a CRM system which can be constructed at very low cost by the use of the open-source software MUSASHI, which can be adopted freely even by a small enterprise. The components of the system comprise marketing tools and data mining technology, which we developed so far through joint research activities with various types of enterprises. Thus, it is possible to carry out the analysis of the customers without building up a new analytical system.

2 C-MUSASHI in Retailers

C-MUSASHI is defined as a CRM system that runs on MUSASHI, by which it is possible to process the purchase history of a large number of customers and to analyze consumer behavior in detail for an efficient customer control. Fig. 1 shows the positioning of C-MUSASHI in a system for daily operation of the retailers.

By using C-MUSASHI, everybody can build up a CRM system without introducing data warehouse through the processes given below. POS registers used in recent years output the data called electronic journal, in which all operation logs are recorded. Store controller collects the electronic journals from the registers in the stores and accumulates them. The electronic journal data is converted by "MUSASHI journal converter" to XML data with minimal data loss.

The framework of such system design provides two advantageous features. First, the loss of data is minimized. In the framework given above, no data will be lost basically. If new data is needed to discover new knowledge, the data can be provided from XML data if it can be obtained from the operation of POS registers.

Secondly, the burden of the system design can be extensively reduced. Because all of the data are accumulated, necessary data can be easily extracted later. Therefore, these schemes can flexibly cope with the changes in the system.

However, if all operation logs at the POS registers are accumulated on XML data, the amount of data may become enormous which in turn leads to the decrease of the processing speed. In this respect, we define a table-type data structure called XML table. A system is built up by combining XML data such as operation logs with XML data and XML table data. Thus, by properly using

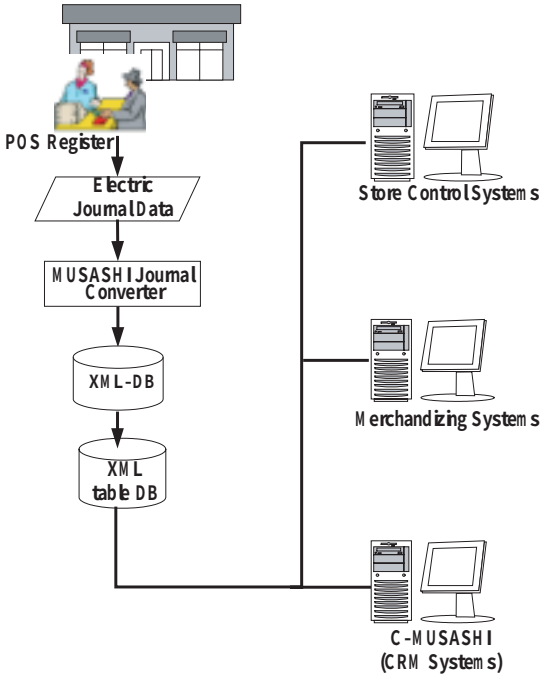


Fig. 1. C-MUSASHI system in the retailers.

pure XML and XML table depending on the purposes, MUSASHI is able to construct an efficient system with high degree of freedom.

Based on the purchase data of the customers thus accumulated, the following systems are built up: store management system for the basic data processing in stores such as accounting information, merchandising system for commodity control and price control, and C-MUSASHI, which will be explained in the succeeding sections.

3 Components of C-MUSASHI

Software tools of C-MUSASHI are categorized into two groups: those for basic customer analysis and CRM systems using data mining technique. The former provides basic information necessary for the implementation of CRM strategy. The latter is a system to discover new knowledge to carry out effective CRM by using data mining technique.

In this section, we shall introduce the tools for basic customer analysis. Basic tools in C-MUSASHI have several tools usually incorporated in general CRM systems: decil analysis, RFM analysis, customer attrition analysis, and LTV measurement. They are used for basic customer analysis. C-MUSASHI also has many other tools for customer analysis. We will present here only a part of them here.

3.1 Decil analysis

In decil analysis, based on the ranking of the customers derived from the amount of purchase, customers are divided into 10 groups with equal number of customers, and then basic indices such as average amount of purchase, number of visits to the store, and etc. are computed for each group [5][6]. From this report, it can be understood that all customers do not have equal value for the store, but only a small fraction of the customers contribute to most of the profits in the store.

3.2 RFM analysis

RFM analysis [6][7] is one of the tools most frequently used in the application purpose such as direct-mail marketing. The customers are classified according to three factors, i.e. recency of the last date of purchase, frequency of purchase, and monetary factor (purchase amount). Based on this classification, adequate sales promotion is executed for each customer group. For instance, in a supermarket, if a customer had the highest purchase frequency and the highest purchase amount, and did not visit to the store within one month, sufficient efforts must be made to bring back this customer from the stores of the competitors.

3.3 Customer attrition analysis

This analysis indicates what fraction of customers in a certain customer group would continuously visit the store in the next period (e.g. one month later) [7]. In other words, this is an analysis to indicate how many customers have gone away to the other stores. These numerical values are also used for the calculation of LTV as described below.

3.4 LTV (Life Time Value)

LTV is a net present value of the profit which an average customer in a certain customer group brings to a store (an enterprise) within a given period [7][8]. It is calculated from the data such as sales amount of the customer group, customer maintaining rate, and discount rate such as the rate of interest on a national bond. Long-term customer strategy should be set up based on LTV, and it is an important factor relating to CRM system. However, the component for calculation of LTV prepared in C-MUSASHI is very simple and it must be customized depending on enterprises to use it.

These four tools are minimally required as well as very important for CRM in business field. It is possible to set up various types of marketing strategies based on the results of analysis. However, they are general and conventional, and then do not necessarily bring new knowledge to support differentiation strategy of the enterprise.

4 CRM Systems Based on the data mining technique

In this section, CRM system based on the data mining technique will be presented, which discovers new knowledge useful for implementing effective CRM strategy from the purchase data of the customers. General CRM system commercially available simply comprises the processes of retrieval and aggregation for each customer group, and there are very few CRM systems in which analytical system that can deal with large-scale data equipped with data mining engine is available in actual business field. In this section, we explain our system that can discover useful customer knowledge by integrating the data mining technique with CRM system.

4.1 Systems structure of C-MUSASHI and four modules

Fig. 2 shows a structure of CRM system using C-MUSASHI. Customer purchase history data accumulated as XML table is preprocesses by a core system of MUSASHI. The preprocessed data is then provided as retail support information in two different ways.

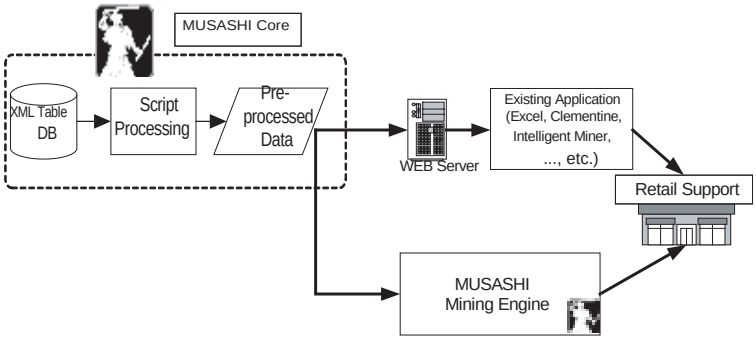


Fig. 2. System structure of C-MUSASHI.

In the first approach, the data preprocessed at the core of MUSASHI is received through WEB server. Then, the data is analyzed and provided to the retail stores by existing application software such as spread-sheet or data mining tools. In this case, C-MUSASHI is only in charge of preprocessing of a large amount of data.

In the second approach, the data is directly received from the core system of MUSASHI. Rules are extracted by the use of data mining engine in MUSASHI, and useful knowledge is obtained from them. In this case, C-MUSASHI carries out a series of processing to derive prediction model and useful knowledge. Whether one of these approaches or both should be adopted by the enterprise should be determined according to the existing analytical environment and daily business activities.

CRM system in C-MUSASHI which integrates the data mining technique consists of four modules corresponding to the life cycle of the customers [8][9][10]. Just as each product has its own life cycle, each customer has life cycle as a growth model. Fig. 3 shows the time series change of the amount of money used by a typical customer. Just like the life cycle of the product, it appears that customer life cycle has the stages of introduction, growth, maturation, and decline.

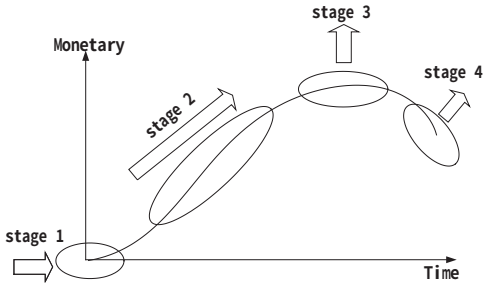


Fig. 3. Customer life cycle and four CRM modules.

It is not that all customers should be treated on equal basis. Among the customers, there are bargain hunters which purchase only the commodities at a discounted price and also loyal customers who give great contribution to the profit of the store. In the stage 1, the loyal customers are discriminated from among new customers. We developed an early discovery module to detect loyal customers in order to attract the customers who may bring higher profitability.

In the stage 2, an analysis is required to promote new customers and quasi-loyal customers to turn them to the loyal customers. For this purpose, we developed a decil switch analysis module. In the stage 3, merchandise assortment is set up to meet the requirements of the loyal customers. A basket analysis module was prepared for the loyal customers in order to attain higher satisfaction and to raise the sales for them. In the final stage 4, for the purpose of preventing the loyal customers from going away to the other competitive stores, analysis is performed using the customer attrition analysis module. Detailed description will be given below on these modules.

4.2 Early discovery modules to detect loyal customers from newcomers

This module is a system for constructing a predictive model to discover potential loyal customers from new customers within short time after the first visit to the store and to acquire knowledge to capture these customers [11][12]. The user can select the preferred range of the customer groups classified in Sections 3.1 and 3.2 and the period to be analyzed. The new customers are then classified into loyal customers and non-loyal ones.

The explanatory attributes are prepared from the purchase data during the specified period such as one-month or during the number of visits to the store from the first visit. Sales ratio of the product category for each customer (the ratio of sales amount of the product category to total sales amount) is generated in the module. When a model is built up by using the data mining engine of MUSASHI, a model for predicting loyal customers can be constructed from the above data joined together by using the model generating command "xtclassify".

As a result, the model tells us which category of purchasing features these new prospective loyal customers have. Such information provides valuable implication when loyal customers are obtained from the competitive stores or when it must be determined on which product category the emphasis should be put when a new store will be opened.

4.3 Decil switch analysis module

Decil switch analysis module is a system to find out what kind of changes of purchase behavior of each customer group based on decil give strong influence on the sales of the store. Given two periods, the following processing will be automatically started: The changes of purchase behavior of the customers during the two periods are calculated, and the customers are classified into 110 customer groups according to the decil value of both periods. For instance, the customers who was classified as decil 3 in the preceding period are to be classified as one of decil 1 through 10 or as the customers who did not visit the store in the subsequent period. For each of these customer groups, the difference between the purchase amount in the preceding period and that in the subsequent period is calculated, which makes it clear how the changes of sales amount of each of the customer groups give strong influence on total sales amount of the store.

Next, judging from the above numerical values (influence on total sales amount of the store), the user decides which of the following data he/she wants to see, e.g., decil switch of all customers, loyal customers of the store, or quasi-loyal customers. If the user wants to see the decil switch of quasi-loyal customers, sales ratio of each product category for each customer group in the preceding period is calculated, and a decision tree is generated, which shows the difference in the purchased categories between the quasi-loyal customers whose decil increased in the subsequent period (decil-up) and those whose decil value decreased (decil-down). Based on the rules obtained from the decision tree, the user can judge which product category should be recommended to quasi-loyal customers in order to increase the total sales of the store.

4.4 Basket analysis module of the loyal customer

For the purpose of increasing the sales amount of a store, the most important and also minimally required condition is to keep loyal customers exclusively for a store. In general, the loyal customers tends to continue to visit a particular store. As far as the merchandises and services to satisfy these customers are provided, it is easier to continuously keep these customers to the store than

to make efforts to acquire the new customers. This module is to find out the merchandises preferred by loyal customers according to the result of the basket analysis on their purchase data [13].

From the results obtained by this module, it is possible not only to find out which product category the loyal customer prefers, but also to extract the most frequently purchased merchandise and to indicate the product belonging to C rank in ABC analysis. In the store control practiced in the past, if sales amount of the products preferred by the loyal customers is not very large, then the product often tends to disappear from the sales counter. Based on such information extracted from this module, the store manager can display the particular merchandise on the sales counter which loyal customer prefers and can pay special attention so that the merchandise will not be out of stock.

4.5 Customer attrition analysis module

Customer attrition analysis module is a system for extracting the purchase behavior of the loyal customers who left the store and to provide information for effective sales promotion in order to regain such loyal customers. When the user defines the loyal customers, the group of the customers is extracted, who had been loyal customers continuously for the past four months and had gone thereafter to the other stores. Using the sales ratio of product category preceding the attrition of the customers as explanatory variable, a classification model of the customer group is generated. By elucidating which group of customers is more easily diverted to the other store and which category of products these customers had been purchasing, the store manager can obtain useful information on the improvement of merchandise lineup at the store to keep loyal customers.

4.6 The case of decil switch analysis module in a supermarket

Since we cannot discuss all of the cases of the above four modules in this paper, we will analyze the data of a large-scale supermarket and try to find out the possibility to promote quasi-loyal customers to loyal customers by using decil switch analysis module.

In the supermarket used in this research, the sales increased more than those of the other stores during the period we are concerned with. The purpose of the analysis is to elucidate the reason in terms of features of the purchase behavior of the customers.

First, two periods, i.e. April and May of 2003, were set up for analysis. Fig. 4 shows the changes of the purchase amounts in April and May of the customer groups classified according to the decil values of both periods. In the figure, the portion indicated by a circle shows that the sales for the quasi-loyal customers groups (the customers with decil 2 -4) in April increases in May. From the figure, it is clear that the sales increase of quasi-loyal customers makes great contribution to the increase of the total sales amount of the store.

Next, focusing on the quasi-loyal customers, decil switch analysis was carried out by using decision tree. In the rules obtained from the decision tree, we found

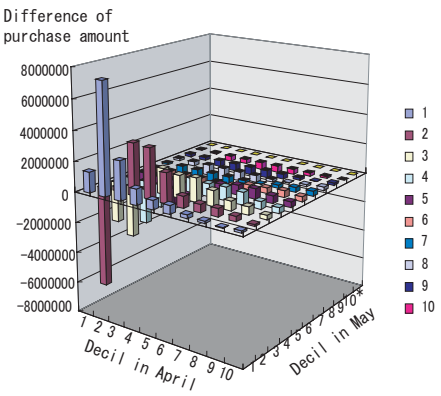


Fig. 4. Changes of sales amount for each decil switch group.

some interesting information. For instance, it was found that the customer who had purchased higher percentage of the product category such as milk, eggs, yoghurt, etc., which are easily perishable, shows high purchase amount in the subsequent period. Also, it was discovered that the customers who had been purchasing drugs such as medicine for colds or headache exhibited the increase in decil value in the subsequent period.

The store manager interpreted these rules as follows: If a customer is inclined to purchase daily foodstuffs at a store, total purchase amount of the customer including other categories can be maintained at high level. As a result, the customer may have a sense of comfort, relief and sympathy with the store and would be more likely to buy other goods relating to health such as drugs. Based on such information, the store manager is carrying out sales promotion to keep the store in such atmosphere as to give the customers a sense of comfort, relief and sympathy to the store.

5 Conclusion

In this paper, we have introduced a CRM system called C-MUSASHI which can be constructed at very low cost by the use of the open-source software MUSASHI. We have explained components and software tools of C-MUSASHI. In particular, C-MUSASHI contains several data mining tools which can be used to analyze purchase behavior of customers in order to increase the sales amount of a retail store. However, we could not explain the details of all of the modules in this paper. In some of the modules, sufficient analysis cannot be carried out in actual business. We will try to offer these modules to the public as soon as possible so that those who are concerned in business field would have an advantage to use the modules. In future, we will continue to make improvement

for the construction of effective CRM systems by incorporating the comments and advices from the experts in this field.

In C-MUSASHI, a typical decision tree tool and basket analysis tool were used as data mining technique. A number of useful data mining algorithms are now provided by the researchers. We will continuously try to utilize and incorporate these techniques into C-MUSASHI, and we will act as a bridge between the research results and actual business activities.

References

1. MUSASHI:Mining Utilities and System Architecture for Scalable processing of Historical data. URL: http://musashi.sourceforge.jp/
2. Hamuro, Y., Katoh, N. and Yada, K.: Data Mining oriented System for Business Applications. Lecture Notes in Artificial Intelligence 1532. Proceedings of First International Conference DS'98. (1998) 441–442
3. Hamuro, Y., Katoh, N., Matsuda, Y. and Yada, K.: Mining Pharmacy Data Helps to Make Profits. Data Mining and Knowledge Discovery. **2** (1998) 391–398
4. Hamuro, Y., Katoh, N. and Yada, K.: MUSASHI: Flexible and Efficient Data Preprocessing Tool for KDD based on XML. DCAP2002 Workshop held in conjunction with ICDM2002. (2002) 38–49
5. Hawkins, G. E.: Building the Customer Specific Retail Enterprise. Breezy Heights Publishing. (1999)
6. Woolf, B. P.: Customer Specific Marketing. Teal Books. (1993)
7. Hughes, A. M.: Strategic Database Marketing. The McGraw-Hill. (1994)
8. Blattberg, R. C., Getz, G., Thomas, J. S.: Customer Equity: Building and Managing Relationships as Valuable Assets. Harvard Business School Press. (2001)
9. Reed, T.: Measure Your Customer Lifecycle. DM News 21. **33** (1999) 23
10. Rongstad, N.: Find Out How to Stop Customers from Leaving. Target Marketing 22. **7** (1999) 28–29
11. Ip, E., Johnson, J., Yada, K., Hamuro, Y., Katoh, N. and Cheung, S.: A Neural Network Application to Identify High-Value Customer for a Large Retail Store in Japan. Neural Networks in Business: Techniques and Applications. Idea Group Publishing. (2002) 55–69
12. Ip, E., Yada, K., Hamuro, Y. and Katoh, N.: A Data Mining System for Managing Customer Relationship. Proceedings of the 2000 Americas Conference on Information Systems. (2000) 101–105
13. Fujisawa, K., Hamuro, Y., Katoh, N., Tokuyama, T. and Yada, K.: Approximation of Optimal Two-Dimensional Association Rules for Categorical Attributes Using Semidefinite Programming. Lecture Notes in Artificial Intelligence **1721**. Proceedings of First International Conference DS'99. (1999) 148–159

Integrated mining for cancer incidence factors from healthcare data

Xiaolong Zhang, Tetsuo Narita

School of Computer Science and Technology,
Wuhan University of Science and Technology, Wuhan 430081 P.R.China
xiaolong.zhang@mail.wust.edu.cn
CRM/BI Solutions, IBM Japan, Tokyo 103-8510 Japan
narita4@jp.ibm.com

Abstract. This paper describes how data mining is being used to identify primary factors of cancer incidences and living habits of cancer patients from a set of health and living habit questionnaires, which is helpful for cancer control and prevention. Decision tree, radial basis function and back propagation neural network have been employed in this case study. Decision tree classification uncovers the primary factors of cancer patients from rules. Radial basis function method has advantages in comparing the living habits between cancer patients and healthy people. Back propagation neural network contributes to elicit the important factors of cancer incidences. This case study provides a useful data mining template for characteristics identification in healthcare and other areas.

Key words: Knowledge acquisition, Data mining, Radial basis function, Decision tree, Back propagation network, Sensitivity analysis, Healthcare, Cancer control and prevention

1 Introduction

With the development of data mining approaches and techniques, the applications of data mining can be found in many organizations, such as banking, insurance, industries, and government. Large volumes of data are produced in every social organization, which can be from scientific research or business. For example, the human genome data is being created and collected at a tremendous rate, so the maximization of the value from this complex data is very necessary. Since the ever increasing data becomes more and more difficult to be analyzed with traditional data analysis methods. Data mining has earned an impressive reputation in data analysis and knowledge acqusition. Recently data mining methods have been applied to many other areas including banking, finance, insurance, retail, healthcare and pharmaceutical industries as well as gene analysis[1, 2].

Data mining methods [3–5] are used to extract valid, previously unknown, and ultimately comprehensible information from large data sources. The extracted information can be used to form a prediction or classification model,

identifying relations between database records. Data mining consists of a number of operations each of which is supported by a variety of techniques such as rule induction, neural networks, conceptual clustering, association discovery, etc. Usually, it is necessary to apply several mining methods to a data set to identify discovered rules and patterns with each other. Data mining methods allow us to apply multi-methods to broadly and deeply discover knowledge from data sets. For example, there is a special issue about the comparison and evaluation of KDD methods with common medical databases [6], where researchers employed several mining algorithms to discover rules from common medical databases.

This paper presents a mining process with a set of health and living habit questionnaire data. The task is to discover the primary factors and living habits of cancer patients via questionnaires. These rules or patterns are helpful for cancer control and cancer incidence prevention. Several mining methods have been used in the mining process. Decision tree, radial basis function (RBF) and back propagation neural network (BPN) are included. Decision tree method helps to generate important rules hidden behind the data, and facilitates useful rule interpretation. However, when the severity distribution of the objective variable (with its class values) exists, decision tree method is not effective, being highly skewed with generating a long thick tail tree (also pointed out by Epte et al. [7]). Moreover, decision tree rules do not give the information that can be used to compare living habits between healthy people and cancer patients. RBF, with its "divide and conquer" ability, performs well in predictions even though it is in presence of severity distribution and noisy data. With RBF, we can obtain rules and patterns of cancer patients and healthy people as well, which is useful for the living habit comparison. On the other hand, back propagation neural network can be used for both prediction and classification. In this study, BPN is used as neural classification and sensitivity analysis [8]. Both predication (for a numerical variable) and classification (for a categorical variable) of BPN have been applied in many areas (e.g., in genome analysis [9, 10]).

This paper describes a multi-strategy data mining application. For example, with BPN's sensitivity analysis, irrelevant and redundant variables are removed, which leads to generate a decision tree in a more understandable way. By means of the combined mining methods, rules, patterns and critical factors can be effectively discovered. Of course, mining with the questionnaire data should carefully investigate the data contents in the data processing, since such a data set contains not only noise but also missing values.

In the rest of this paper, there is the description of decision tree classification, RBF and BPN algorithms. There is about how to build data marts, how to use these algorithms to perform a serial of effective mining processes. Finally, there is the related works and conclusion.

2 Data mining approach

Generally, there may be several methods available to a mining problem. Considering the number of variables and records, as well as the quality of training

data, applying several mining methods on the given mining data is recommended. Since one mining algorithm may outperform another, more useful rules and patterns can be further revealed.

2.1 Classification method

Classification method is very important for data mining, which has been extensively used in many data mining applications. It analyzes the input data and generate models to describe the *class* using the attributes in the input data. The input data for classification consists of multiple attributes. Among these attributes, there should be a label tagged with *class*. Classification is usually performed through decision tree classifiers. A decision-tree classifier creates a decision-tree model with tree-building phase and tree-pruning phase (e.g., CART [11], C4.5 [12] and SPRINT [13]).

In the tree-building phase, a decision tree is grown by repeatedly partitioning the training data based on values of a selected attribute. Therefore the training set is split into two or more partitions according to the selected attribute. The tree-building process is repeated recursively until a stop criterion is met. Such a stop criterion may be all the examples in each partition has its class label or the depth of tree reaches a given value or other else.

After the tree-building phase, the tree-pruning phase is used to prune the generated tree with test data. Since the tree is built with training data, it may grow to be one that fits the noisy training data. Pruning tree phase removes the overfitting branches of the tree given the estimated error rate.

Decision tree has been used to build predictive models and discover understandable rules. In this paper, decision tree is applied to discover rules.

2.2 Radial basis function method

The radial basis function algorithm is used to learn from examples, usually used for prediction. RBF can be viewed as a feedforward neural network with only one hidden layer. The outputs from the hidden layer are not simply the product of the input data and a weight. All the input data to each neuron in the hidden layer are treated as a measure of distance which can be viewed as how far the data are from a center. The center is the position of the neuron in a spatial system. The transfer functions of the nodes are used to measure the influence that neurons have at the center. These transfer functions are usually radial spline, Guassian or power functions. For example, 2-dimension Gaussian radial basis function centered in t can be written as:

$$G(\|x-t\|^2) \equiv e^{\|x-t\|^2} = e^{(x-t_x)^2} e^{(y-t_y)^2}. \tag{1}$$

This function can be easily extended for dimensions higher than 2 (see [14]).

As a regularization network, RBF is equivalent to generalized splines. The architecture of backpropagation neural network consists of multilayer networks where one hidden layer and a set of adjustable parameters are configured. Their

Boolean version divides the input space into hyperspheres, each corresponding to a center. This center, also call it a radial unit, is active if the input vector is within a certain radius of its center and is otherwise inactive. With an arbitrary number of units, each network can approximate the other, since each can approximate continuous functions on a limited interval.

This property, we call it "divide and conquer", can be used to identify the primary factors of the related cancer patients within the questionnaires. With the "divide and conquer", RBF is able to deal with training data, of which the distributions of the training data are extremely severity.

The RBF is also effect on solving the problem in which the training data includes noise. Because the transfer function can be viewed as a linear combination of nonlinear basis functions which effectively change the weights of neurons in the hidden layer. In addition, the RBF allows its model to be generated in an efficient way.

2.3 Back propagation neural network

BPN can be used for both neural prediction and neural classification. BPN consists of one layer of input nodes, one layer of output nodes, and one or more hidden layers between the input and output layers. The input data (called vectors) x_i, each multiplied by a weight w_i and adjusted with a threshold θ, is mapped into the set of two binary values $R^n \rightarrow \{0, 1\}$. The threshold function is usually a sigmoid function

$$f(sum) = \frac{1}{1 + e^{-sum}} \tag{2}$$

where the sum is the weighted sum of the signals at the unit input. The unit outputs a real value between 0 and 1. For $sum = 0$, the output is 0.5; for large negative values of sum, the output converges to 0; and for large positive value sum, the output converges to 1. The weight adjustment procedure in backpropagation learning is explained as: (1) Determine the layers and the units in each layer of a neural net; (2) Set up initial weights $w1_{ij}$ and $w2_{ij}$; (3) Input a input vector to the input layer; (4) Propagate the input value from the input layer to the hidden layer, where the output value of the jth unit is calculated by the function $h_j = \frac{1}{1+e^{-\sum_i (w1_{ij} \cdot x_i)}}$. Propagate the value to the output layer, and the output value of the jth unit in this layer is defined by the function: $o_j = \frac{1}{1+e^{-\sum_i (w2_{ij} \cdot h_i)}}$; (5) Compare the outputs o_j with the given classification y_j, calculate the correction error $\delta 2_j = o_j(1-o_j)(y_j-o_j)$, and adjust the weight $w2_{ij}$ with the function: $w2_{ij}(t+1) = w2_{ij}(t) + \delta 2_j \cdot h_j \cdot \eta$, where $w2_{ij}(t)$ are the respective weights at time t, and η is a constant ($\eta \in (0, 1)$); (6) Calculate the correction error for the hidden layer by means of the formula $\delta 1_j = h_j(1 - h_j) \sum_i \delta 2_i \cdot w2_{ij}$, and adjust the weights $w1_{ij}(t)$ by: $w1_{ij}(t+1) = w1_{ij}(t) + \delta 1_j \cdot x_j \cdot \eta$; (7) Return to step 3, and repeat the process.

The weight adjustment is also an error adjustment and propagation process, where the errors (in the form of weights) are feedback to the hidden units. These

errors are normalized per unit fields in order to have a sum of all as 100%. This process is often considered as a sensitivity analysis process which shows the input field (variable) contributions to the classification for a *class label*. Therefore, omitting the variables that do not contribute will improve the training time and those variables are not included in the classification run. As we know, real data sets often include many irrelevant or redundant input fields. By examining the weight matrix of the trained neural network itself, the significance of inputs can be determined. A comparison is made by sensitivity analysis, where the sensitivity of outputs to input perturbation is used as a measure of the significance of inputs. Practically, in the decision tree classification, by making use of sensitivity analysis and removing the lowest contributed variables, understandable and clear decision trees are easily generated.

3 Application template overview

Our task for this study is to uncover the primary factors of cancer incidences and living habits of cancer patients, and further compare these factors and habits with healthy people.

The mining sources are from an investigation organization which collected data via questionnaires. The questionnaire data set consists of 250 attributes and is full of 47,000 records. These attributes are from 14 categories. These categories mainly include: personal information (the date of birth, living area, etc.); records of one's illnesses (apoplexy, high blood pressure, myocardial infarction, tuberculosis, cancer, etc.); the health statute of one's families (parents, brothers and sisters); the health status of one in the recent one year (bowels movement, sleeping, etc.); one's drinking activity (what kind of liquor, how often, how much, etc.); one's smoking records (when begun, how many cigarettes a day, etc.); one's eating habit (regular meal time, what kind of food, what kind of meat and fish, etc.); occupation (teacher, doctor, company staff, etc.), and so on. A questionnaire record may contain missing values of some attributes. For example, a female-oriented question is not answered by males. In other cases, missing data is due to lack of responses. There are several ways to deal with missing-data (see [15]). One effective way is EM (Expectation and Maximization) [16]. This method has been implemented in some analytical tool, such as Spss. We use Spss to deal with the missing values before mining operations.

Mining data marts are the data sources used for mining. Several data marts are created, which are used for decision tree, Rbf and Bpn mining. In fact, decision tree could not directly applied to the whole dataset. The decision tree mining operation is applied to the output from clustering operation, where the distribution for a selected objective variable is almost balanced. For example, to classify the cancer/non-cancer (with value 1/0) people, the variable "cancer flag" is selected as the class variable. For efficiently generating decision trees, removing irrelevant or redundant processes are performed by first applying the Bpn sensitivity analysis. In addition, the data marts and mining processes are also

repeated with BPN operation, for acquiring a list of significant cancer incidence factors.

When building a data mart for RBF analysis, the analysis object is to predict the probability of the cancer accident. The variable "cancer flag" is mapping to a new variable "probV26" with value of [0,1]. This new variable is used as a dependence variable, whose values (e.g., probability) will be predicted using the other variables (independence variables).

4 Mining process and mining result

IBM INTELLIGENT MINER FOR DATA (IM4D) [17] is used as the mining tool, since the combined mining processes can be performed with IM4D, which includes decision tree, RBF and BPN algorithms as well as clustering. In this application, we clean the data source with SPSS, build data marts for mining, and perform mining operations to acquire useful cancer incidence factors and related rules.

The mining process includes building data marts, creating mining bases, mining with decision trees, RBF and BPN. As we know, not all the questionnaire objects are cancer patients. In fact, only 561 persons were or are cancer patients in the given records. The missing values are allocated as *unknown* (more detail processes see [18]).

4.1 Mining with decision tree

Decision tree is performed based on the output data from a clustering model, where the class variable distribution can be well balanced. With carefully selected data, objective-oriented trees are generated. For example, a decision tree can be created for describing the male cancer patients or the female ones. In order to further investigate the cancer factors, decision trees are generated with selected variables from personal information, from illness category, health statute of family category, from drinking, smoking and eating activity, respectively.

The results of the decision tree mining are useful in understanding primary factors of cancer incidences. For instance, with respect to male cancer patients, there are tree rules generated like: (a) the cancer patients were alcohol drinkers, there were smokers in their family, and their smoking period was more than 9 years; (b) They were accepted surgery on the abdomen, 62 age over, suffering from apoplexy, experiencing blood transfusion; (c) They have no work now but are living with stresses, they ever had been teachers, or doing management work in company or government. From another decision tree, for both male and female cancer patients, there is a rule: among 122 cancer patients, 42 of them accepted surgery on the abdomen, contracted gallstone/gallbladder inflammation, and suffered from diabetes. Another rule denotes: among 372 cancer patients, 100 of female cancer patients' menses were stopped by surgery, their initial marriage ages were less than 26, and suffering from both constipation and diabetes.

4.2 Mining with RBF

As shown above, these rules from decision trees are useful for understanding the primary factors of cancers. But there is no information for comparison between cancer patients and healthy people.

With RBF, several predictive models have been generated. First, one predictive model is generated with the dependence variable "probV26" (mentioned before) and all independence variables. More predictive models have been built for male and female cancer patients. Second, in order to further discover the different living habits between the cancer patients and healthy people, the independence variables of predictive models are selected from only personal information and illness information, from only health statute of families, and from only drinking activity and smoking information. With these data marts RBF predictive models are built respectively.

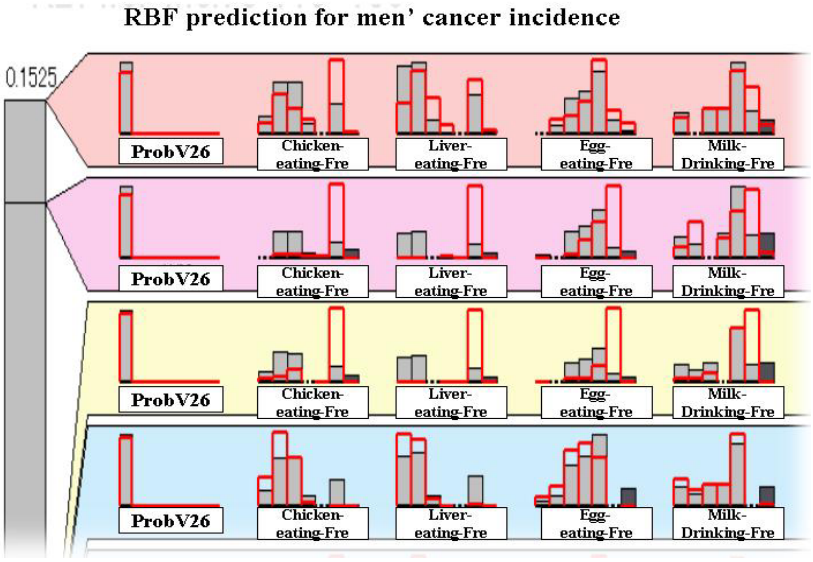
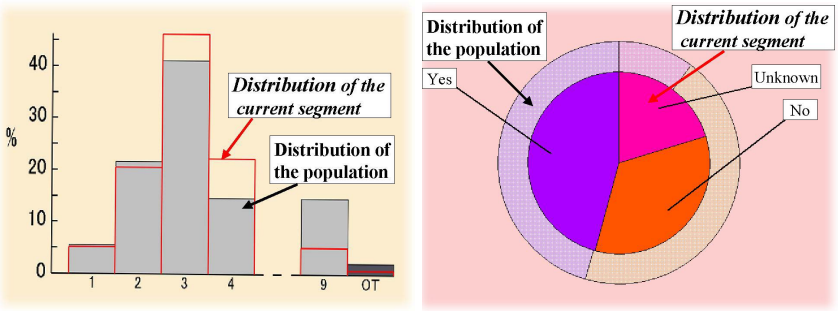


Fig. 1. The "divide and conquer" of RBF method

Fig. 1 shows the RBF predictive models which predict probability of cancer incidences, where RBF automatically builds small, local predictive models for different regions of data space. This "divide and conquer" strategy appears well for prediction and factor identifier. This chart indicates eating habits in terms of food-eating frequency. The segment with low probability of cancer incidence segment (at bottom) shows the chicken-eating frequency (the variable located in the second from the left) is higher than that of the segment (at top) with higher probability of cancer incidences. The detailed explanation of distribution



Histogram chart Pie chart

Fig. 2. Variable distribution in the population and current segment in RBF

Table 1. Comparison between female patients and healthy people in suffered illness

Illness	Female patients	Healthy women
Kidney	6.5%	3.5%
Womb illness	24.6%	14.2%
Blood transfusion	12.7%	7.1%
Diabetes	3.5%	1.2%

is described in Fig. 2, where for a numerical variable, the distributions in the population and in the current segment are indicated. Moreover, the percentage of each distribution can be given if necessary. With RBF predictive ability, the characteristic of every segment can be identified. By means of RBF, the living habits of cancer patients and healthy people are discovered.

Fig. 2 shows an example of distributions of a variable. The histogram chart is for a numerical variable in a segment of RBF models, where the distributions of population and current segment of the variable are described, respectively. For instance, the percentage of partition 4 in the current segment is higher than that of the population (the entire data set). The pie chart is for a categorical variable. The outside ring of the pie chart shows the distribution for the variable over the population. The inside ring shows the distribution for this variable in the current segment. In the pie chart of Fig. 1, the distribution of *No* in the current segment is less than that of the population.

The results of RBF mining processes are interesting. The results provide comparable information between cancer patients and healthy people in suffered illness. Some results from RBF are described in Table 1. In this table, there is a prediction case of cancer incidence of women, 6.5% of a high cancer incidence group has kidney illness, while the percentage for healthy people is 3.5%. For womb illness and blood transfusion, the figures for female cancer patients are higher than those of healthy women.

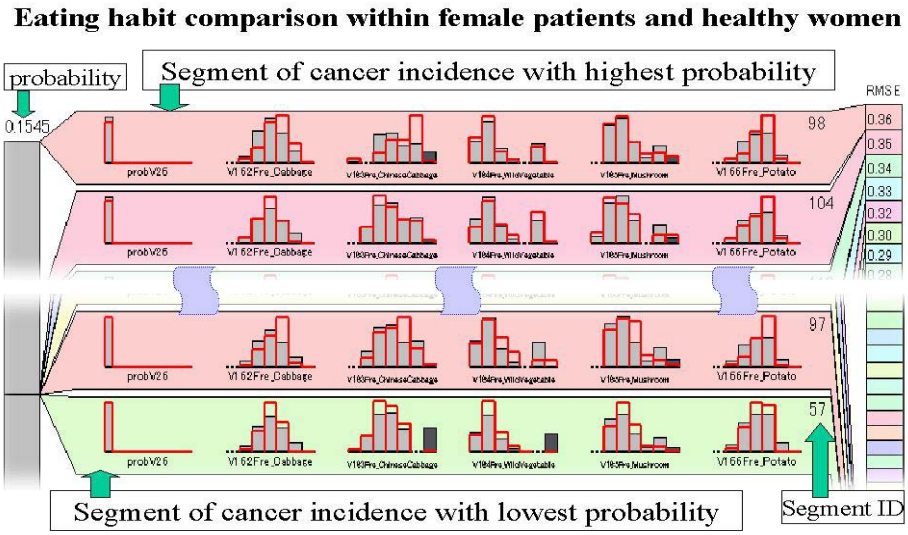


Fig. 3. Comparison of eating habits: female cancer incidences and healthy women

The eating habits between female cancer incidences and healthy women are compared. The results of RBF (described in Fig. 3) show the characteristics in the segment with highest cancer incidence probability (with 0.15 probability and segment ID 98) and that of lowest probability as well (with 0 probability and segment ID 57). By picking up the variables with obviously different distributions between these two segments, comparable results are obtained. Within eating habits, comparing regular breakfast habit category, 87% of female cancer patients have regular breakfast habit, which is lower than 93.3% of healthy women. Amongst meat-eating and chicken-eating 3-4 times per-week, the percentage figures are 12% and 8% of female cancer patients, 27.7% and 18% of healthy women, respectively. In one's personal life, 54.5% of female cancer patients are living in a state of stress, far higher than 15% the percentage of healthy women. For judgment on matters, 36% of female cancer patients gives out their judgments very quickly. This percentage is also higher than 13%, the percentage figure of healthy women.

4.3 Mining with back propagation neural network

This section describes the mining results with back propagation neural network. When the back propagation method is used for classification, it is also used as sensitivity analysis simultaneously. With sensitivity analysis, the important factors of cancer incidences can be acquired. In this case study, sensitivity analysis

has been performed with the data marts built for in decision tree classification. With BPN, mining run on these data sets tries to discover the relationship between the input variables and the class variable *cancer flag*.

Table 2. Most important factors of cancer incidences

Attribute related cancer	Sensitivity
blood transfusion	1.4
job-category (now)	1.3
menstruation	1.2
surgery experience or not	1.2
volume of meal	1
the longest occupation	0.8
lactation type (of woman)	0.8
site of uterine cancer check	0.8
site of breast cancer check	0.8
somking	0.7
site of stomach cancer check	0.7
biological mother suffered cancer	0.7
volume of alcolhol	0.6
hot favorance	0.6
rich oil favorance	0.6
biological father suffered cancer	0.6

Table 2 shows those variables with most contributions to the cancer incidence, which is the BPN mining results with a data mart where both male and female cancer patients' records are included. The important factors are those with high sensitivity. *blood transfusion* (sensitivity 1.4) and *job-category* (sensitivity 1.3) are important factors. For a female patient, *menstruation* (if her menstruation is regular or not) and *lactation type* (which type of lactation was applied for her children) are indicated as important factors. It also shows that *smoking* and *alcohol* are tightly related to cancer incidences, which is similar to the results acquired in the decision tree rules. The fact that one' biological father/mother has suffered cancer is also important. The factor *hot favorance* and *rich oil favorance* are also noticed.

As described above, decision tree classification, RBF and BPN mining processes generate useful rules, patterns and cancer incidence factors. By carefully prepared mining data marts and removing irrelevant or redundant input variables, decision tree rules show the primary factors of cancer patients. RBF predictive models reveal more details about comparison information between cancer patients and healthy people. BPN obviously reports the important cancer incidence factors. These results enable us to discover the primary factors and living habits of cancer patients, their different characteristics compared with healthy

people, further indicating what are the most related factors contributing to cancer incidences.

5 Concluding remarks

This paper has described an integrated mining method that includes multiple algorithms, and mining operations for efficiently obtaining results.

In applying data mining to medical and healthcare area, the special issue in [6] describes details mining results from variety of mining methods, where common medical databases are presented. This is a very interesting way to identifying the mined results. The results obtained with proposed algorithms are interesting. However, for a specific mining propose, the researchers have not described which algorithm is effective and which result (or part of results) is more useful or accurate compared to that generated with other algorithms.

In applying decision tree to medical data, our earlier work [19] was implemented with clustering and classification, where comparison between cancer patients and healthy people could not be given. In addition, the critical cancer incidence factors were not acquired. This case study with RBF and BPN has successfully compared the difference between cancer patients and healthy people and the significant cancer incidence factors. With applying RBF in factor identification, a case study of semiconductor yield forecasting can be found in [20]. We believe that this application template can be also used in computational biology. The work of [9, 10] are applications of BPN in genome. However, these applications are single method mining runs. Therefore, the mining results can be improved with multistrategy data mining. The integration of decision tree, RBF and BPN to do mining is an effective way to discover rules, patterns and important factors.

Data mining is very a useful tool in the healthcare and medical area, as this study has demonstrated. Ideally, large amounts of data (e.g., the human genome data) are continuously collected. This data is then segmented, classified, and finally reduced to a predictive model. With an interpretable predictive model, significant factors of a predictive object can be uncovered. In some extent, this paper has given an application template which answers the questions of how to employ multistrategy data mining tools to discover quality rules and patterns.

References

1. S. L. Pomeroy et al. *Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature*, 405:436–442, 2002.
2. Y. Kawamura and X. Zhang and A. Konagaya. *Inference of genetic network in cluster level. 18th AI Symposium of Japanese Society for Artificial Intelligence, SIG-J-A301-12P, 2003.*
3. R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5:914–925, 1993.

4. M.S. Chen, J. Han, and P.S. Yu. Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8:866–883, 1996.
5. Xiaolong Zhang. *Knowledge Acqusition and Revision with First Order Logic Induction*. PhD Thesis, Tokyo Institute of Technology, 1998.
6. Special Issue. Comparison and evaluation of KDD methods with common medical databases. *Journal of Japanese Society for Artificial Intelligence*, 15:750–790, 2000.
7. C. Apte, E. Grossman, E. Pednault, B. Rosen, F. Tipu, and B. White. Probablistic estimation based data mining for discovering insurance risks. Technical Report IBM Research Report RC-21483, T. J. Watson Research Center, IBM Research Division, Yorktown Heights, NY 10598, 1999.
8. Gedeon TD. Data mining of inputs: analysing magnitude and functional measures. *Int. J. Neural Syst*, 8:209–217, 1997.
9. Wu Cathy and S. Shivakumar. Back-propagation and counter-propagation neural networks for phylogenetic classification of ribosomal RNA sequences. *Nucleic Acids Research*, 22:4291–4299, 1994.
10. Wu Cathy, M. Berry, S. Shivakumar, and J. McLarty. Neural networks for full-scale protein sequence classification: Sequence encoding with singular value decomposition. *Machine Learning*, 21:177–193, 1994.
11. L. Breiman, J. Friedman amd R. Olshen, and C. Stone. *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.
12. J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
13. J.C. Shafer, R. Agrawal, and M. Mehta. SPRINT: A scalable parallel classifier for data mining. In *Proc. of the 22th Int'l Conference on Very Large Databases*, Bombay, India, 1996.
14. T. Poggio and F. Girosi. Networks for approximation and learning. *Proceedings of the IEEE*, 78:1481–1497, 1990.
15. Roderick J. A. Little and Donald B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 1987.
16. Dempster A., Laird N., and Rubin D. Maximun likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. B*, 39:1–38, 1977.
17. IBM Intelligent Miner for Data. *Using the Intelligent Miner for Data*. IBM Corp., Third Edition, 1998.
18. P. Cabena et al. *Discovering data mining*. Prentice Hall PTR, 1998.
19. X. Zhang and T. Narita. Discovering the primary factors of cancer from health and living habit questionaires. In S. Arikawa and K. Furukawa, editors, *Discovery Science: Second International Conference (DS'99)*. LNAI 1721 Springer, 1999.
20. Ashok N. Srivastava. Data mining for semiconductor yield forecasting. *Future Fab International*, 1999.

Multi-Aspect Mining for Hepatitis Data Analysis

Muneaki Ohshima[1], Tomohiro Okuno[1], Ning Zhong[1], Hideto Yokoi[2]

[1] Dept. of Information Engineering, Maebashi Institute of Technology
460-1 Kamisadori-cho, Maebashi 371-0816, Japan
[2] School of Medicine, Chiba University
1-8-1 Inohana, Chuo-ku, Chiba 260-8677, Japan

Abstract. When therapy using IFN (interferon) medication for chronic hepatitis patients, various conceptual knowledge/rules will benefit for giving a treatment. In this paper, we describe an ongoing work on using various data mining agents including the GDT-RS inductive learning system for discovering classification rules, the LOI (learning with ordered information) for discovering important features, as well as the POM (peculiarity oriented mining) for finding peculiarity data/rules, in a discovery process with multi-phase such as pre-processing, rule mining, and post-processing, for multi-aspect analysis of the hepatitis data. Our methodology and experimental results show that the perspective of medical doctors will be changed from a single type of experimental data analysis towards a holistic view, by using our *multi-aspect mining* approach.

1 Introduction

Multi-aspect mining in a multi-phase KDD (Knowledge Discovery and Data Mining) process is an important methodology for knowledge discovery from real-life data [2, 7, 11, 12]. There are two main reasons why a multi-aspect mining approach needs to be used for the hepatitis data analysis.

The first reason is that we cannot expect to develop a single data mining algorithm for analyzing all main aspects of the hepatitis data towards a holistic view since complexity of the real-world applications. Hence, various data mining agents need to be cooperatively used in the multi-phase data mining process for performing multi-aspect analysis as well as multi-level conceptual abstraction and learning.

The other reason is that when performing multi-aspect analysis for complex problems such as the hepatitis data mining, a data mining task needs to be decomposed into sub-tasks. Thus these sub-tasks can be solved by using one or more data mining agents that are distributed over different computers. Thus the decomposition problem leads us to the problem of distributed cooperative system design.

More specifically, when therapy using IFN (interferon) medication for chronic hepatitis patients, various conceptual knowledge/rules will benefit for giving a treatment. The knowledge/rules, for instance, include (1) when the IFN should

be used for a patient so that he/she will be able to be cured, (2) what kinds of inspections are important for a diagnosis, and (3) whether some peculiar data/patterns exist or not.

In this paper, we describe an ongoing work on using various data mining agents including the GDT-RS inductive learning system for discovering classification rules [8, 13], the LOI (learning with ordered information) for discovering important features [4, 14], as well as the POM (peculiarity oriented mining) for finding peculiarity data/rules [15], for multi-aspect analysis of the hepatitis data so that such rules mentioned above can be discovered automatically.

We emphasize that both pre-processing/post-processing steps are important before/after using data mining agents. In particular, informed knowledge discovery, in general, uses background knowledge obtained from experts (e.g. medical doctors) about a domain (e.g. chronic hepatitis) to guide a discovery process with multi-phase such as pre-processing, rule mining, and post-processing, towards finding interesting and novel rules/features hidden in data. Background knowledge may be of several forms including rules already found, taxonomic relationships, causal preconditions, ordered information, and semantic categories.

In our experiments, the result of the blood test of the patients, who performed INF before starting medication, is first pre-treated. After that, the pre-processed data are used for each data mining agent, respectively. By using the GDT-RS, the rules with respect to know whether a medical treatment is effective or not, can be found. And, by using the LOI, what attributes affect the medical treatment of hepatitis C greatly can be investigated. Our methodology and experimental results show that the perspective of medical doctors will be changed from a single type of experimental data analysis towards a holistic view, by using our multi-aspect mining approach.

The rest of the paper is organized as follows. Section 2 describes how to pre-process the hepatitis data and decide the threshold values for condition attributes according to the background knowledge obtained from medical doctors. Section 3 gives the main results mined by using the GDT-RS and the post-processing. Section 4 discusses the analysis and evaluation of the results given in Section 3, based on a medical doctor's advice and comments. Then in Section 5, we extend our system by adding the LOI (learning with ordered information) and POM (peculiarity oriented mining) data mining agents for multi-aspect mining and analysis. Finally, Section 6 gives conclusions.

2 Pre-processing

2.1 Selection of Inspection Data and Class Determination

We use the following conditions to extract inspection data.

- Patients of chronic hepatitis type C who may be medicated with IFN.
- Patients with the data of judging the IFN effect by using whether a hepatitis virus exists or not.
- Patients with inspection data collected in one year before IFN is used.

Thus, 197 patients with 11 condition attributes as shown in Table 1 are selected and will be used in our data mining agents.

Table 1. Condition attributes

T-CHO	CHE	ALB	TP
T-BIL	D-BIL	I-BIL	PLT
WBC	HGB	GPT	

Furthermore, the decision attribute (i.e. classes) is selected according to a result of judging the IFN effect by using whether a hepatitis virus exists or not. Hence, the 197 extracted patients can be classified into 3 classes as shown in Table 2.

Table 2. The decision attribute (classes)

class	The condition of the patient after IFN	# of patients
R	Disappearance of the virus	58
N	Existence of virus	86
?	Reliability lack of data	53

2.2 Evaluation of Condition Attributes

As shown in Fig. 1, the condition attributes are evaluated as follows.

1. All the inspection values in one year before IFN is used for each patient are divided into two groups, the first half and the second half of the inspection values.
2. When the absolute value of the difference between average values of the first half and the second half of the inspection values exceeds the threshold, it is estimated as up or down. Otherwise, it is estimated as "–" (i.e. no change). Moreover, it is estimated as "?" in the case where not inspection data or only once (i.e. a patient is examined only once).

Furthermore, the threshold values can be decided as follows.

– **The threshold values for each attribute except GPT** are set up to 10% of the normal range of each inspection data. As the change of a hepatitis patient's GPT value will exceed the normal range greatly, the threshold value for the GPT needs to be calculated in a more complex method to be described below. The threshold values used for evaluating each condition attribute is shown in Table 3.

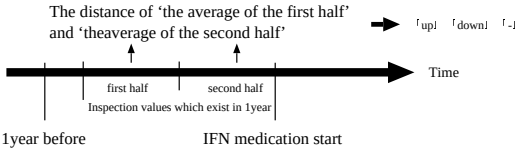


Fig. 1. The evaluation method of condition attributes

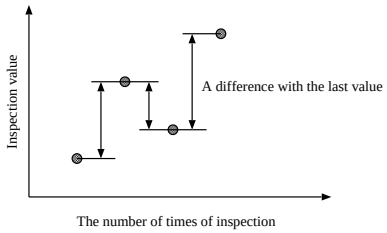


Fig. 2. Standard deviation of the difference of adjacent values

– **The threshold value for GPT** is calculated as follows. As shown in Fig. 2, the standard deviation of the difference of the adjacent test values of each hepatitis patient's GPT is first calculated, respectively; And then the standard deviation of such standard deviation is used as a threshold value.
In the GPT test, let m be the number of patients, $t_m(1 \leq m \leq M)$ the time of test (that may be different for each patient), $d_{mi}(1 \leq i \leq t_m - 1)$ the difference of adjacent test values. Thus, the threshold value of GPT can be calculated in the following equation.

$$\text{threshold of GPT} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(s_m - \overline{s})^2} \tag{1}$$

where $s_m(1 \leq m \leq M)$ is the standard deviation of the difference d_{mi} of the test value that is calculated for each patient, respectively, and $\overline{s}$ is the average value of s_m.

Finally, the threshold values are set up as shown in Table 3.

Table 3. The threshold values for evaluating condition attributes

T-CHO > 9.5	CHE > 25	ALB > 0.12	TP > 0.17
T-BIL > 0.1	D-BIL > 0.03	I-BIL > 0.07	PLT > 20
WBC > 0.5	HGB > 0.6	GPT > 54.56	

3 Main Results

3.1 Rule Discovery by GDT-RS

GDT-RS is a soft hybrid induction system for discovering classification rules from databases with uncertain and incomplete data [8, 13]. The system is based on a hybridization of the *Generalization Distribution Table (GDT)* and the *Rough Set* methodology. The main features of GDT-RS are the following:

- Biases for search control can be selected in a flexible way. Background knowledge can be used as a bias to control the initiation of GDT and in the rule discovery process.
- The rule discovery process is oriented toward inducing rules with high quality of classification of unseen instances. The rule uncertainty, including the ability to predict unseen instances, can be explicitly represented by the rule strength.
- A minimal set of rules with the minimal (semi-minimal) description length, having large strength, and covering of all instances can be generated.
- Interesting rules can be induced by selecting a discovery target and class transformation.

In the experimental results at the accuracy 60%, only the rules with which the number of condition attributes is less than or equal to three are extracted. This is because it will become unclear if the number of condition attributes increases. Tables 4 and 5 show such rules that are divided into classes R and N, respectively.

Table 4. Rules with respect to class R

rule-ID	rule & accuracy
001	GPT(up) & (10/16)=62%
002	T-CHO(down) $\wedge$ PLT(down) & (6/9)=66%
003	T-BIL(up) $\wedge$ GPT(down) & (3/4)=75%
004	TP(down) $\wedge$ GPT(down) & (3/4)=75%

3.2 Results of Post-processing

As a post-processing, we checked each discovered rule covers what patient(s) related data. Table 6 shows the results, where the *Positive* (or *Negative*) *ID* means that the patient is covered by a rule as a *positive* (or *negative*) instance. From this table, we can see it becomes clear that what patient group is covered by what rule. Hence it is useful for finding the main features of a patient group.

As an example of post-processing, Table 8 shows a part of result of the post-processing about class R. Here "+" and "–" denote the patient covered by a rule as a positive or negative instance, respectively. For example, *rule 001* is covered by the patient IDs: {158, 778, 801, 909, 923, 940, 942}.

Table 5. Rules with respect to class N

rule-ID	rule & accuracy
101	D-BIL(down) & (26/43)=60%
102	T-CHO(down) ∧ I-BIL(down) & (7/11)=63%
103	I-BIL(down) ∧ WBC(down) & (7/8)=87%
104	D-BIL(up) ∧ PLT(down) & (4/6)=66%
105	TP(up) ∧ I-BIL(down) & (5/6)=83%
106	TP(up) ∧ T-BIL(down) & (4/6)=66%
107	TP(up) ∧ PLT(down) & (4/5)=80%
108	CHE(up) ∧ T-BIL(down) & (2/4)=50%

Table 6. Patients covered by rules with respect to class R

rule-ID	Positive patient ID	Negative patient ID
001	158 351 534 547 778 801 909 923 940 942	35 188 273 452 623 712
002	91 351 650 703 732 913	169 712 952
003	431 592 700	122
004	37 71 730	122

4 Analyses and Evaluations

The results derived by the GDT-RS and post-processing have been evaluated by a medical doctor based on acceptability and novelty of each rule. The evaluations of the rules are divided into five stages: 1 is the lowest and 5 is the highest evaluation for acceptability and novelty of each rule.

4.1 Evaluation of Rules

From the viewpoint of the rules with a higher support (e.g. *rule-001* and *rule-101*), we observed that

- It will heal up in many cases if a patient is medicated with IFN at the time when GPT is going up (hepatitis is getting worse);
- It does not heal up in many cases even if a patient is medicated with IFN at the time when D-BIL is descending.

Furthermore, the following two points on the effect of IFN is understood clearly.

- It is relevant to different types of hepatitis viruses;
- It is hard to be effective when there are large amounts of hepatitis virus.

Hence, we can see that *rule-001* and *rule-101* do not conflict with the existing medicine knowledge.

Table 7. Patients covered by rules with respect to class N

rule-ID	Positive patient ID	Negative patient ID
101	2 104 125 182 184 191 203 208 239 290 546 439 493 498 529 578 585 634 652 653 669 715 719 743 750 756	37 71 133 169 180 206 248 276 413 593 610 683 702 713 732 771 948
102	2 239 563 634 652 653 952	169 413 650 732
103	2 138 208 432 578 653 736	413
104	187 260 289 712	703 778
105	72 182 219 546 920	35
106	72 182 219 546	180 610
107	104 182 260 535	180
108	210 634	180 683

From the two rules, the hypothesis: "IFN is more effective when the inflammation of hepatitis is stronger" can be formed. Based on this hypothesis, we can evaluate the rules discovered as follows.

- In class R, the acceptability of the rules with respect to aggravation of liver function will be good.
- In class N, the acceptability of the rules with respect to recovery of liver function will be good.

Hence, the evaluations as shown in Tables 9 and 10 can be obtained. In class N, we can see that the acceptability of some rules is 2. This is because both the recovery and aggravation of liver function are included in the premise of the rules.

4.2 Evaluation of Post-processing

In the discovered rules, we can see there is some relevance among the patients supported by bilirubin (T-BIL, D-BIL, I-BIL) in class N. From the relation denoted in T-BIL = D-BIL + I-BIL, it is clear that the rules with respect to bilirubin are relevant. Hence the rules are supporting the same patients group.

Moreover, in order to examine the hypothesis, "the medical background which a rule shows is not contradictory to a patient's condition", the discovered rules are categorized, based on liver function, into three categories: recovery, aggravation, or mixed recovery and aggravation, as shown in Table 11.

From Table 11, we observed that there are many rules with the same conditions in the rule group supported by a patients group, and it may conflict with

Table 8. Post-processing about class R

Patient ID	rule 001	rule 002	rule 003	rule 004
35	−			
37				+
71				+
78				
91		+		
122			−	−
158	+			
169		−		
⋮	⋮	⋮	⋮	⋮
778	+			
801	+			
909	+			
913		+		
923	+			
⋮	⋮	⋮	⋮	⋮
940	+			
942	+			
⋮	⋮	⋮	⋮	⋮

Table 9. Evaluation of rules with respect to class R

rule-ID	rule	accuracy	acceptability	novelty
001	GPT(up)	(10/16)=62%	4	5
002	T-CHO(down) ∧ PLT(down)	(6/9)=66%	3	5
003	T-BIL(up) ∧ GPT(down)	(3/4)=75%	4	5
004	TP(down) ∧ GPT(down)	(3/4)=75%	4	5

unknown medical background that is not represented in the conditions of the rules. However, it does not mean that the rules are incorrect. The reason may be that the rules cannot be simply categorized by recovery and aggravation.

For example, although it can show liver function aggravation, the lower values of WBC and ALB may not be the real reason of liver function aggravation. On other hand, since WBC and PLT are the same blood cell ingredient, and T-CHO and ALB are relevant to protein that makes liver, they may be relevant from this point of view. However, T-CHO and ALB do not only provide for liver, but also, for example, T-CHO is related to eating, and ALB is related to the kidney, respectively. Hence it cannot declare there is such correlation.

In summary, there is correlation if we are mentioning about mathematical relevance like BIL. However, it is difficult to find out correlation for others. We need the following methods to solve the issue.

- Finding out the rules which are significant from the statistical point of view, based on rough categorizing such as recovery and aggravation.
- Showing whether such rough categorizing is sufficient or not.

Table 10. Evaluation of rules with respect to class N

rule-ID	rule	accuracy	acceptability	novelty
101	D-BIL(down)	(26/43)=60%	4	5
102	T-CHO(down) $\wedge$ I-BIL(down)	(7/11)=63%	2	3
103	I-BIL(down) $\wedge$ WBC(down)	(7/8)=87%	2	3
104	D-BIL(up) $\wedge$ PLT(down)	(4/6)=66%	1	1
105	TP(up) $\wedge$ I-BIL(down)	(5/6)=83%	3	4
106	TP(up) $\wedge$ T-BIL(down)	(4/6)=66%	3	4
107	TP(up) $\wedge$ PLT(down)	(4/5)=80%	2	3
108	CHE(up) $\wedge$ T-BIL(down)	(2/4)=50%	3	4

Table 11. Category of discovered rules

	Recovery	Aggravation	Recovery and Aggravation
class R	rule 007 rule 008 rule 009 rule 011	rule 001 rule 002	rule 003 rule 004 rule 005 rule 006 rule 010
class N	rule 101 rule 105 rule 106 rule 108 rule 110	rule 104 rule 109	rule 102 rule 103 rule 107 rule 111

5 Multi-Aspect Analysis by LOI and POM

Based on the results stated above, we have been extending our system by adding the LOI (learning with ordered information) and POM (peculiarity oriented mining) data mining agents for multi-aspect mining and analysis.

5.1 Ordering Rule Mining

The LOI uses the background knowledge called *ordered relation* for discovering *ordering rules* and important attributes for an ordered decision class [4, 14]. For example, since the values for T-CHO are the larger and the better, the ordered relation can be set to $\infty \succ 0$, where "$\succ$" denotes a weak order. Furthermore, if a decision attribute has two classes: R (response) and N (no response), the ordered relation can be set to $R \succ N$.

Ordering rules can be generated from a binary information table by using GDT-RS. The rules can be used to predict the effect of IFN. Moreover, the most important attributes can be found when a rule with too many condition attributes.

Although they have been obtained in hepatitis data analysis, the results need to be evaluated and interpreted by medical doctors. The ongoing work will be reported in detail in our next papers.

5.2 Peculiarity Oriented Mining

Peculiarity represents a new interpretation of interestingness, an important notion long identified in data mining [3, 10, 15]. Peculiarity, unexpected relation-

ships/rules may be hidden in a relatively small number of data. *Peculiarity rules* are a typical regularity hidden in many scientific, statistical, medical, and transaction databases. They may be difficult to find by applying the standard association rule mining method [1], due to the requirement of large support. In contrast, the POM (peculiarity oriented mining) agent focuses on some interesting data (peculiar data) in order to find novel and interesting rules (peculiarity rules).

We have been applying the POM for hepatitis data analysis and had some preliminary result [6]. Currently, we are also working with Suzuki's group to integrate the Peculiarity Oriented Mining approach with the Exception Rules/Data Mining approach for discovering more refined LC (Liver Cirrhosis) and non-LC classification models. The ongoing work will be also reported in detail in our next papers.

6 Conclusions

We presented a multi-aspect mining approach in a multi-phase, multi-aspect hepatitis data analysis process. Both pre-processing and post-processing steps are important before/after using data mining agents. Informed knowledge discovery in real-life hepatitis data needs to use background knowledge obtained from medical docotors to guide the discovery process with multi-phase such as pre-processing, rule mining, and post-processing, towards finding interesting and novel rules/features hidden in data.

Our methodology and experimental results show that the perspective of medical doctors will be changed from a single type of experimental data analysis towards a holistic view, by using our multi-aspect mining approach in which various data mining agents are used in a distributed cooperative mode.

Acknowledgements

This work was supported by the grant-in-aid for scientific research on priority area "Active Mining" from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

1. Agrawal R. et al. "Fast Discovery of Association Rules", *Advances in Knowledge Discovery and Data Mining*, AAAI Press (1996) 307-328.
2. Fayyad, U.M., Piatetsky-Shapiro, G, and Smyth, P. "From Data Mining to Knowledge Discovery: an Overview", *Advances in Knowledge Discovery and Data Mining*, MIT Press (1996) 1-36.
3. Liu, B., Hsu W., Chen, S., and Ma, Y. "Analyzing the Subjective Interestingness of Association Rules", *IEEE Intelligent Systems*, Vol.15, No.5 (2000) 47-55.
4. Sai, Y., Yao, Y.Y., and Zhong, N. "Data Analysis and Mining in Ordered Information Tables", *Proc. 2001 IEEE International Conference on Data Mining (ICDM'01)*, IEEE Computer Society Press (2001) 497-504.

5. Suzuki, E. Undirected Discovery of Interesting Exception Rules, *International Journal of Pattern Recognition and Artificial Intelligence*, World Scientific, Vol.16, No.8 (2002) 1065-1086.
6. Yokoi, H., Hirano, S., Takabayashi, K., Tsumoto, S., and Satomura, Y. "Active Mining in Medicine: A Chronic Hepatitis Case – Towards Knowledge Discovery in Hospital Information Systems", *Journal of Japanese Society for Artificial Intelligence*, Vol.17, No.5 (2002) 622-628 (in Japanese).
7. Zhong, N. and Ohsuga, S. "Toward A Multi-Strategy and Cooperative Discovery System", *Proc. First Int. Conf. on Knowledge Discovery and Data Mining (KDD-95)*, AAAI Press (1995) 337-342.
8. Zhong, N., Dong, J.Z., and Ohsuga, S., "Rule Discovery in Medical Data by GDT-RS", Special Issue on Comparison and Evaluation of KDD Methods with Common Medical Datasets, *Journal of Japanese Society for Artificial Intelligence*, Vol.15, No.5 (2000) 774-781 (in Japanese).
9. Zhong, N. "Knowledge Discovery and Data Mining", *the Encyclopedia of Microcomputers*, Volume 27 (Supplement 6) Marcel Dekker (2001) 93-122.
10. Zhong, N., Yao, Y.Y., Ohshima, M., and Ohsuga, S. "Interestingness, Peculiarity, and Multi-Database Mining", *Proc. 2001 IEEE International Conference on Data Mining (ICDM'01)*, IEEE Computer Society Press (2001) 566-573.
11. Zhong, N. and Ohsuga, S. "Automatic Knowledge Discovery in Larger Scale Knowledge-Data Bases", C. Leondes (ed.) *The Handbook of Expert Systems*, Vol.4: Chapter 29, Academic Press (2001) 1015-1070.
12. Zhong, N., Liu, C., and Ohsuga, S. "Dynamically Organizing KDD Process", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 15, No. 3, World Scientific (2001) 451-473.
13. Zhong, N., Dong, J.Z., and Ohsuga, S. "Rule Discovery by Soft Induction Techniques", *Neurocomputing*, An International Journal, Vol. 36 (1-4) Elsevier (2001) 171-204.
14. Zhong, N., Yao, Y.Y., Dong, J.Z., Ohsuga, S., "Gastric Cancer Data Mining with Ordered Information", J.J. Alpigini et al (eds.) *Rough Sets and Current Trends in Computing*, LNAI 2475, Springer (2002) 467-478.
15. Zhong, N., Yao, Y.Y., Ohshima M., "Peculiarity Oriented Multidatabase Mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol.15, No.4 (2003) 952-960.

Investigation of Rule Interestingness in Medical Data Mining

Miho Ohsaki[1], Yoshinori Sato[1], Shinya Kitaguchi[1], Hideto Yokoi[2], and Takahira Yamaguchi[1]

[1] Shizuoka University, Faculty of Information
3-5-1 Johoku, Hamamatsu-shi, Shizuoka 432-8011, JAPAN
{miho, cs8037, cs9026, yamaguti}@cs.inf.shizuoka.ac.jp
http://www.cs.inf.shizuoka.ac.jp/~miho/index.html
http://panda.cs.inf.shizuoka.ac.jp/index.html
[2] Chiba University Hospital, Medical Informatics
1-8-1 Inohana, Chuo-ku, Chiba-shi, Chiba 260-0856, JAPAN
yokoih@telemed.ho.chiba-u.ac.jp

Abstract. This research experimentally investigates the performance of conventional rule interestingness measures and discusses their availability to supporting KDD through system-human interaction. We compared the evaluation results by a medical expert and that by several selected measures for the rules discovered from the medical test data on chronic hepatitis. The measure on antecedent-consequent dependency using all instances showed the highest performance, and ones on the both of dependency and generality the lowest under this experimental condition. The whole trend of the experimental results indicated that the measures detected really interesting rules at a certain level and offered us the rough guideline to apply them to system-human interaction.

1 Introduction

The concern with the contribution of data mining to Evidence-Based Medicine (EBM) has been growing for the last several years, and there have been many medical data mining studies [21, 27]. It is experientially known as the important factor influencing on discovered knowledge quality how to pre-/post-process real ill-defined clinical data and how to polish up rules through the interaction between a system and its human user. However, it has not been discussed enough [4].

Thus, to discuss the pre-/post-processing and the system-human interaction in medical domain, we have been conducting case studies using medical test data on chronic hepatitis. We estimated that the temporal patterns of medical test results would be useful for a medical expert to grasp diagnosis and predict prognosis. We then obtained the graph-based rules predicting the future pattern of GPT, one of major medical tests. We iterated the rule generation by our mining system and the rule evaluation by a medical expert two times [28].

As the results, we obtained the knowledge on the rule interestingness for a medical expert and learned the lessons on the pre-/post-processing and the system-human interaction in medical domain. We then focused on system-human interaction and proposed its concept model and its semi-automatic system framework [29]. These case studies made us recognize the significance of clarifying the rule interestingness really required by a human user and that of returning such information to a mining system.

Therefore, this research has the following two purposes: **(1)** investigating the conventional interestingness measures in Knowledge Discovery in Databases (KDD) and comparing them with the rule evaluation results by a medical expert, and **(2)** discussing whether they are available to support system-human interaction in medical domain.

In this paper, Section 2 introduces conventional interestingness measures and shows the selected several measures suitable to our purpose. Section 3 notes the experimental conditions and results to evaluate the rules on chronic hepatitis with the measures and to compare them with the evaluation results by a medical expert. In addition, it discusses the availability of the measures for system-human interaction support. Finally, Section 4 concludes the paper and comments on the future work.

2 Related Work

2.1 Outcome of Our Previous Research

We have conducted the case studies to discover the rules on diagnosis and prognosis from a chronic hepatitis data set. The set of the rule generation by our mining system and the rule evaluation by a medical expert was iterated two times and leaded us to discover the rules valued as interesting ones by the medical expert.

We used the data set of the medical test results on viral chronic hepatitis [15]. Before mining, we finely pre-processed it based on medical expert's advice since such a real medical data set is ill-defined and has many noises and missing values. We then extracted the representative temporal patterns from the data set by clustering and generated the rules consisting of the patterns and predicting the prognosis by a decision tree [5].

Figure 1 shows one of the rules, which the medical expert focused on, obtained in the first mining. It estimates the future trend of GPT, one of major medical tests to grasp chronic hepatitis symptom, in the future one year by using the change of several medical test results in the past two years. The medical expert commented on it as follows: the rule offers a hypothesis that GPT value changes with about a three-years cyclic, and the hypothesis is interesting since it differs from the conventional common sense of medical experts that GPT value basically decreases in a monotone.

We then improved our mining system, extended the observation term, and generated new rules. Figure 2 shows two of the rules, which the medical expert

valued, obtained in our second mining. The medical expert commented on them that they imply GPT value globally changes two times in the past five years and more strongly support the hypothesis of GPT's cyclic change.

The literature [28] explains the details of our previous research, namely the pre-processing methods, the developed system, and the process of rule generation and evaluation. Refer it if you need to know the details.

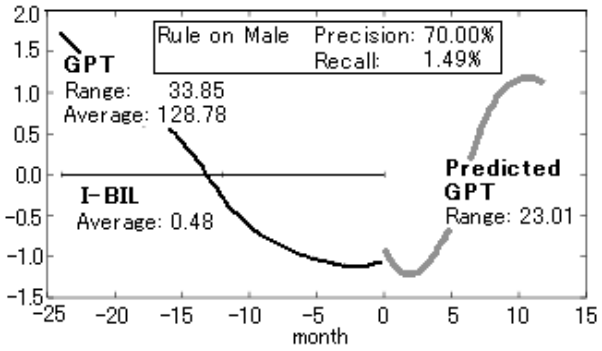


Fig. 1. Rule valued by a medical expert in the first mining.

As the next research, we tried to systematize the knowledge obtained in the previous research, on the pre-/post-processing suitable to medical data and the system-human interaction to polish up rules. Especially on system-human interaction, we formulated its concept model that describes the roles and the functions of a system and a human user (See Figure 3) and the framework to support semi-automatic system-human interaction based on the model (See Figure 4).

As shown in Figure 3, a mining system discovers the rules faithfully to the data and offers them to a medical expert as the materials for hypothesis generation and justification. While, the medical expert generates and justifies a hypothesis, a seed of new knowledge, by evaluating the rules based on his/her domain knowledge. A system to support such interaction requires the function to generate and present rules to a medical expert based on their validity at the viewpoints of objective data structure and subjective human evaluation criteria. The flow of "System Evaluation" and "Human Evaluation" in Figure 4 means that.

Note that the word 'justification' in this paper does not mean the highly reliable proof of a hypothesis by additional medical experiments under strictly controlled conditions. It means the additional information extraction to enhance the reliability of an initial hypothesis from the same data.

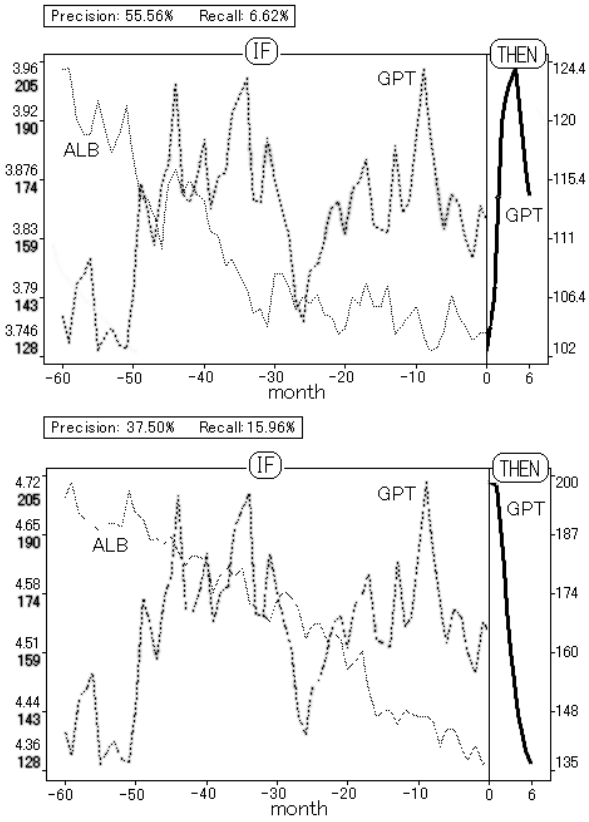


Fig. 2. Rules valued by a medical expert in the second mining.

The literature [29] explains the details of the concept model of system-human interaction and the framework for semi-automatic system-human interaction. Refer it to know the details.

These researches notified us that it is required for realizing the framework in Figure 4 to investigate the rule interestingness measures available for "System Evaluation" and the relation between "System Evaluation" and "Human Evaluation". Therefore, this research selects several conventional measures and compares the rule evaluation results by them with "Human Evaluation", namely that by a medical expert.

2.2 Rule Interestingness Measures

Rule interestingness is one of active research fields in KDD. There have been many studies to formulate interestingness measures and to evaluate rules with them instead of humans. Interestingness measures are categorized into objective

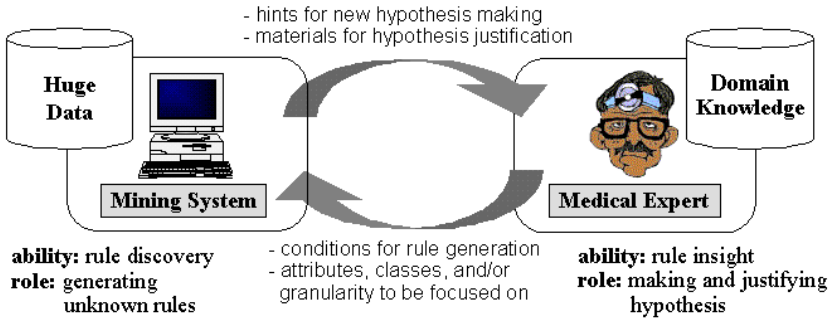


Fig. 3. Interaction model between a system and a human expert at a concept level.

and subjective ones. Objective measures mean how a rule is mathematically meaningful based on the distribution structure of the instances related to the rule. Subjective measures mean how a rule fit with a belief, a bias, or a rule template formulated beforehand by a human user [16].

Objective measures are mainly used to remove meaningless rules at the viewpoint of data structure rather than to discover really interesting ones for a human user, since they do not include domain knowledge [31, 34, 12, 6, 9, 26, 8, 25, 18, 36] On the other hand, subjective measures are available to discover really interesting rules to some extend due to their built-in domain knowledge. However, they depend on the precondition that a human user can clearly formulate his/her own interest [20, 19, 23, 24, 30, 32, 33] Few subjective measures adaptively learn real human interest through system-human interaction [10].

The conventional interestingness measures, not only objective ones but also subjective ones, do not directly reflect the interest that a human user really has. To avoid the confusion of real human interest and the interestingness measures, we define them as follows (Note that while we define "Real Human Interest" by ourselves, the definitions of the other terms are based on many conventional studies on interestingness measures):

Objective Measure: The feature such as correctness, uniqueness, etc. of a rule or a set of rules, mathematically calculated using data structure. It does not include human evaluation criteria. **Subjective Measure:** The similarity or the difference between the information on interestingness beforehand given by a human user and that obtained from a rule or a set of rules. Although it includes human evaluation criteria in its initial state, its calculation of similarity or difference is mainly based on data structure. **Real Human Interest:** The interest in a rule, which a human user really feels in his/her mind. It is formed from the synthesis of human natural cognition, individual domain knowledge and experiences, and the influences of the rules that the human user evaluated before.

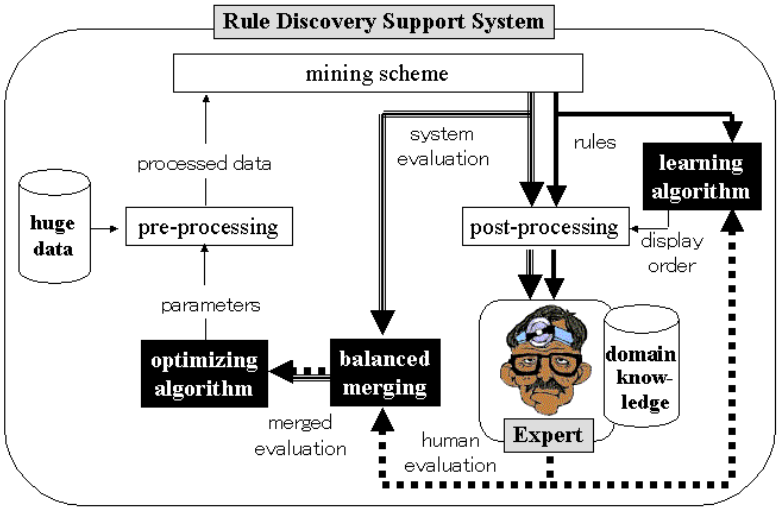


Fig. 4. Framework for semi-automatic system-human interaction.

Objective Measures Objective measures are the mathematical analysis results of data distribution structure. There are many objective measures, and they can be categorized into some groups using evaluation criterion, evaluation target, and theory for analysis. The evaluation target means whether an objective measure evaluates a rule or a set of rules. This research deals with not the objective measures for a set of rules [13, 16] but ones for a rule, because it focuses on the quality of each rule.

Table 1 shows some major objective measures. They assume one of the following evaluation criteria and examine how a rule matches with the criteria by calculating the instance distribution difference between the data and the rule or between the antecedent and the consequent of the rule.

Correctness: How many instances the antecedent and/or the consequent of a rule supports, or how strong their dependence is [31, 34, 26, 25, 18] **Information Plentifulness:** How much information a rule possesses [12]. **Generality:** How similar the trend of a rule is to that of all data [9]. **Uniqueness:** How different the trend of a rule is from that of all data [6, 8, 36] or the other rules [9, 26].

Although objective measures are useful to automatically remove obviously meaningless rules, some evaluation criteria have the contradiction to each other such as generality and uniqueness. In addition, the evaluation criterion of an objective measure may not match with or may contradict real human interest. For example, a rule with a plenty of information may be too complex for a human user to understand. Many of the objective measure proposers showed the validity of their measures with mathematical proofs or the experimental results using benchmark data. However, they hardly conducted the comparison between

Table 1. List of the objective measures of rule interestingness. The measures used in this research have the symbol '*'. The following symbols in the column 'Calc.' mean what is used to calculate the measure. **N:** Number of instances included in the antecedent and/or the consequent. **P:** Probability of the antecedent and/or the consequent. **S:** Statistical variable based on **P**. **I:** Information of the antecedent and/or the consequent. **D:** Distance of the rule from the other rules based on rule attributes. **C:** Complexity of the tree structure of the rule.

Name	Calc.	Evaluation Criterion
Rule Interest [31]	N	Dependency between the antecedent and the consequent
Support *	P	Generality of the rule
Precision (Confidence) *	P	Performance of the rule to predict the consequent
Recall *	P	Performance of the rule not to leak the consequent
Accuracy	P	Summation of the precision and its converse of contrapositive
Lift *	P	Dependency between the antecedent and the consequent
Leverage *	P	Dependency between the antecedent and the consequent
Reliable Exceptions [25]	P	Rule with small support and high precision
Gray and Orlowska's measure (GOI) * [11]	P	Multiplication of the support and the antecedent-consequent dependency
Surprisingness [8]	P	Rule occurring Sympson's paradox
χ^2 measure 1 * [26]	S	Dependency between the antecedent and the consequent
χ^2 measure 2 [26]	S	Similarity between two rules
J-Measure [34] *	I	Dependency between the antecedent and the consequent
General Measure [18]	S & I	Fusion of the χ^2 measure 1 and the information gain measure
Distance Metric [9]	D	Distance of the rule from the rule with the highest coverage
Dong and Li's measure [6]	D	Distance of the rule from the other rules
Peculiarity [36]	D	Distance of the attribute value from frequent attribute values
I-Measure [12]	C	Complexity of the rule

their measures and the other ones or the investigation of the relation between their measures and real human interest for a concrete application.

Subjective Measures Subjective measures are the similarity or difference between the information given by a human user and that given by a rule. There are several subjective measures and can be categorized into some groups with human evaluation criterion, method to give information from a human user to a mining system, theory for calculating the similarity or difference. The human evaluation criterion means what feature of a rule is interesting for a human user and, 'Unexpectedness' and 'Actionability' are popular as the human evaluation criterion. The method to give information means whether a human user give information before mining or through system-human interaction.

If we categorize subjective measures with human evaluation criterion, there are the following groups: the group based on the rule template or the mathematical expression of human cognition characteristics and/or domain knowledge [7, 20, 22], one expressing unexpectedness based on the difference between a rule

and a human belief [33, 19, 30, 23], one expressing actionability based on the similarity based on that [24].

Although many of subjective measures use the information given beforehand by a human user, a few subjective measures use that given through iterative system-human interaction in the mining process. The subjective measure using a rule template in [20] allows a human user to modify the rule template. The literature [10] developed a mining system that interactively learns real human interest.

Objective measures are mainly used to remove meaningless rules. While, subjective measures are available to the positive usage finding really interesting rules. However, the trade-off exists between the generality and the correctness of subjective measures and their applicability. There are two contrastive types of subjective measures: mathematically defining subjective interestingness at a high abstract level to secure the generality and the correctness, and finely implementing real human interest specific for a certain domain to secure the applicability. Therefore, a generic subjective measure existing between them is required at present. Other problem is how to adaptively reflect the change of real human interest caused by comparing and evaluating various rules.

Selection of Interestingness Measures This research experimentally investigates the availability of objective measures by comparing them with real human interest in medical domain. As mentioned in Section 2.2, the evaluation criteria of objective measures are obviously not the same of humans, since objective measures do not include the knowledge on rule semantics. However, they may be available to support the KDD through system-human interaction if they possess a certain level of performance to detect really interesting rules. That is the motivation of this research. The investigation of subjective measures will be our future work.

From the objective measures shown in Table 1, we selected the followings as the investigation targets: the most popular ones (Support, Precision, Recall, Lift, and Leverage), probability-based one (GOI [11]), statistics-based one (χ^2 measure 1 [26]), and information-based one (J-Measure [34]).

3 Comparison between Objective Measures and Real Human Interest

3.1 Experimental Conditions

In our previous researches, we repeated the data mining process two times using a dataset of chronic hepatitis and generated a set of rules for each mining (Refer Section 2.1). After each mining, a medical expert evaluated the rules and gave each rule one of the following rule quality labels: 'Especially-Interesting', 'Interesting', 'Not-Understandable', and 'Not-Interesting'. 'Especially-Interesting' means that the rule was a key factor to generate the hypothesis of GPT's cyclic change in the first mining or to justify it in the second mining. As the results,

we obtained 12 and 8 'Interesting' rules in the first and the second mining, respectively.

In this research, we applied the objective measures selected in Section 2.2 to the same rules and sorted them in the descending order of their evaluation values. We then regarded the rules from top to 12-th in the first mining and that to 8-th in the second mining as 'Interesting' ones judged by the objective measures.

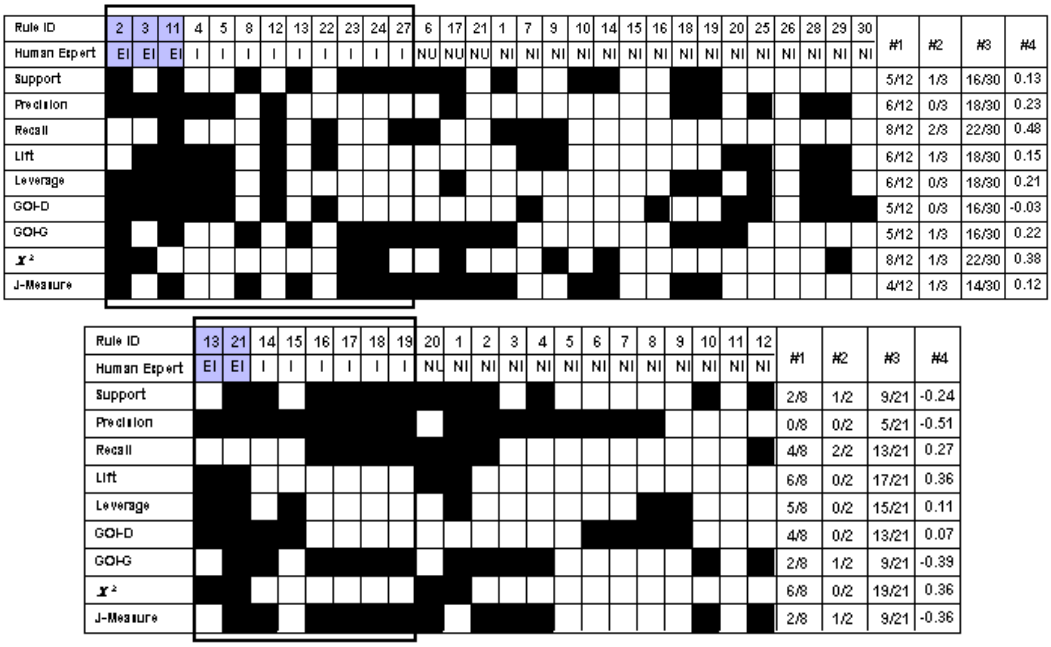
Note that there are two types of GOI [11], GOI-D (GOI emphasizing Dependency) and GOI-G (GOI emphasizing Generality). GOI is the multiplication of antecedent-consequent dependency and generality factors and possesses a parameter to balance them. Therefore, we used GOI-D in which the weight of the dependency factor was twice that of the generality one and GOI-G with the adverse condition.

3.2 Results and Discussion

The upper and the lower tables in Figure 5 show the evaluation results in the first and the second mining, respectively. The caption of Figure 5 explains the contents of these tables in detail. The tables describe how the evaluation results of an objective measure matches with that of the medical expert. The white cells in the square on the left side of a table mean the evaluation concordance on 'Interesting'. Similarly, that in the gray-colored columns means that on 'Especially-Interesting'. Therefore, the number of the former and the latter describes the detection performance of an objective measure on 'Interesting' and 'Especially-Interesting', respectively.

To grasp the whole trend of the experimental results, we define the comprehensive criteria to evaluate the objective measure's performance as follows: **#1** Performance on 'Interesting' (the number of 'Interesting' rules judged by an objective measure per that by the medical expert), **#2** Performance on 'Especially-Interesting' (the number of 'Especially-Interesting' rules judged by an objective measure per that by the medical expert), **#3** Count-based performance on all evaluation (the number of rules with the same evaluation results by an objective measure and the medical expert per that of all rules), and **#4** Correlation-based performance on all evaluation (the correlation coefficient between the evaluation results by an objective measure and that by the medical expert).

At first, we discuss on the results in each mining. As shown in the upper table of Figure 5, χ^2 measure 1 and Recall demonstrated the highest performance, and J-Measure, GOI-D, GOI-G, and Support the lowest in the first mining. While, in the lower table of Figure 5, χ^2 measure 1 and Lift demonstrated the highest performance, and J-Measure, GOI-D, and Support the lowest in the second mining. Although the objective measures with the highest performance failed to detect some of 'Especially-Interesting' and 'Interesting' rules, their availability to supporting system-human interaction was confirmed at a certain level.



Rule ID	2	3	11	4	5	8	12	13	22	23	24	27	6	17	21	1	7	9	10	14	15	16	18	19	20	25	26	28	29	30	#1	#2	#3	#4
Human Expert	EI	EI	EI	I	I	I	I	I	I	I	I	I	NU	NU	NU	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI				
Support	■		■			■		■		■	■	■	■	■		■			■	■			■	■							5/12	1/3	16/30	0.13
Precision	■	■	■	■	■		■						■	■									■	■		■		■	■		6/12	0/3	18/30	0.23
Recall			■				■		■			■	■			■	■	■													8/12	2/3	22/30	0.48
Lift		■	■	■	■		■		■							■	■	■							■	■		■	■		6/12	1/3	18/30	0.15
Leverage	■	■	■	■	■		■							■									■	■		■		■	■		6/12	0/3	18/30	0.21
GOI-D	■	■	■	■	■		■	■									■					■			■	■		■	■	■	5/12	0/3	16/30	-0.03
GOI-G		■				■		■		■	■	■	■	■	■	■							■	■	■						5/12	1/3	16/30	0.22
χ^2	■	■							■				■		■	■		■	■											■	8/12	1/3	22/30	0.38
J-Measure	■		■			■		■		■	■	■	■	■	■	■			■	■			■	■							4/12	1/3	14/30	0.12

Rule ID	13	21	14	15	16	17	18	19	20	1	2	3	4	5	6	7	8	9	10	11	12	#1	#2	#3	#4
Human Expert	EI	EI	I	I	I	I	I	I	NU	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI	NI				
Support		■		■		■	■	■	■	■	■		■						■		■	2/8	1/2	9/21	-0.24
Precision	■	■	■	■	■	■	■	■		■	■	■	■	■	■	■	■					0/8	0/2	5/21	-0.51
Recall						■	■	■	■	■	■										■	4/8	2/2	13/21	0.27
Lift	■	■							■	■												6/8	0/2	17/21	0.36
Leverage	■	■		■						■							■	■				5/8	0/2	15/21	0.11
GOI-D	■	■	■	■											■	■	■	■				4/8	0/2	13/21	0.07
GOI-G		■				■	■	■	■		■	■	■						■		■	2/8	1/2	9/21	-0.39
χ^2	■	■							■	■												6/8	0/2	19/21	0.36
J-Measure		■	■		■	■	■	■	■		■	■	■						■		■	2/8	1/2	9/21	-0.36

Fig. 5. Evaluation results by a medical expert and the selected objective measures, for the rules obtained in the first mining (the upper table) and that in the second one (the lower table). Each line means a set of evaluation results of the medical expert or the objective measure, and each column means each rule. The rules are sorted in the descending order of the evaluation values given by the medical expert. The rules judged 'Interesting' by the medical expert are surrounded by a square, and ones judged 'Especially-Interesting' are colored in gray. In the line of medical expert's evaluation, **EI:** 'Especially-Interesting', **I:** 'Interesting', **NU:** 'Not-Understandable', and **NI:** 'Not-Interesting'. In the lines of objective measure's evaluation, **white cell:** "Same as medical expert's evaluation", and **black cell:** "Different from that". In the four columns in the right side, **#1:** Performance on 'Interesting', **#2:** Performance on 'Especially-Interesting', **#3:** Count-based performance on all evaluation, and **#4:** Correlation-based performance on all evaluation.

Next, we discuss on the whole trend of the results through the first and the second mining. χ^2 measure 1 maintained the highest performance, and J-Measure, GOI-D, and Support the lowest. Although the performance of Recall and Lift slightly changed, there was no objective measure with dramatic performance change.

We then consider why such trend appeared comparing it with the analysis of medical expert's comments on evaluation. The analysis illustrated the following points of medical expert's observation: **(1)** the medical expert focused on the shape of temporal patterns in a rule rather than the rule performance to predict prognosis, **(2)** he evaluated a rule considering the reliability, the unexpectedness,

and the other factors, and **(3)** although the reliability was one of important evaluation factor, many reliable rules were not interesting due to their well-knownness.

The highest performance of χ^2 measure 1 may be caused by **(1)**. Only χ^2 measure 1 uses the instances for the all combination of supporting the antecedent, not supporting the antecedent, supporting the consequent, and not supporting the consequent [26]. Accordingly, it valued the rules in which the temporal patterns in the antecedent and that in the consequent were smoothly connected. This feature of χ^2 measure 1 possibly met the medical expert's needs. While, the lowest performance of Support seems to be deserved by considering **(3)**. The reason of the lowest performance of J-Measure and GOI-D can be estimated base on **(2)**. J-Measure and GOI-D consists of the generality and dependency factors [34, 11], and the balance of these factors was not the same in medical expert's mind in this experiment.

We then summarize the results and the discussions so far: χ^2 measure 1 [26] showed the highest performance, and while J-Measure [34], GOI-D [11], and Support the lowest under these experimental conditions. The objective measures used here possessed not enough but a certain level of performance to detect really interesting rules. The results indicated that the availability of the objective measures for supporting KDD through system-human interaction.

4 Conclusions and Future Work

This research discussed how objective measures can contribute to detect the interesting rules for a medical expert through the experiment using a real chronic hepatitis dataset. The objective measures used here possessed a certain level of detection performance, and then their availability for system-human interaction was indicated. In our future work, we will design the system-human interaction function using objective measures based on this research outcome and equip it to our rule discovery support system. In addition, we will continue the case studies on objective and subjective measures from the viewpoints of not only post-processing but also rule quality management.

References

1. Clarke, F., Ekeland, I.: Nonlinear oscillations and boundary-value problems for Hamiltonian systems. Arch. Rat. Mech. Anal. **78** (1982) 315–333
2. Michalek, R., Tarantello, G.: Subharmonic solutions with prescribed minimal period for nonautonomous Hamiltonian systems. J. Diff. Eq. **72** (1988) 28–55
3. Rabinowitz, P.: On subharmonic solutions of a Hamiltonian system. Comm. Pure Appl. Math. **33** (1980) 609–633
4. Cios, K. J., Moore, G. W.: Uniqueness of medical data mining. Artificial Intelligence in Medicine, **26 (1)**–**(2)** (2002) 1–24.
5. Das, G., King-Ip, L., Heikki, M., Renganathan, G., Smyth, P.: Rule Discovery from Time Series. In Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD'98, (1998) 16–22.

6. Dong, G., Li, J.: Interestingness of Discovered Association Rules in Terms of Neighborhood-Based Unexpectedness. In Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD'98 (1998) 72–86.
7. Matheus, C. J., Piatetsky-Shapiro, G.: Selecting and Reporting What Is Interesting: The KEFIR Application to Healthcare Data. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R.(eds.): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press (1995) 401–419.
8. Freitas, A. A.: On Rule Interestingness Measures. Knowledge-Based Systems, **12 (5)–(6)** (1999) 309–315.
9. Gago, P., Bento, C.: A Metric for Selection of the Most Promising Rules. In Proc. of Euro. Conf. on the Principles of Data Mining and Knowledge Discovery PKDD'98 (1998) 19–27.
10. Terano, T., Inada, M.: Data Mining from Clinical Data using Interactive Evolutionary Computation: In Ghosh, A., Tsutsui, S. (eds.): Advances in Evolutionary Computing. Springer (2003) 847–862.
11. Gray, B., Orlowska, M. E.: CCAIIA: Clustering Categorical Attributes into Interesting Association Rules. In Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD'98 (1998) 132–143.
12. Hamilton, H. J., Fudger, D. F.: Estimating DBLearn's Potential for Knowledge Discovery in Databases. Computational Intelligence **11 (2)** (1995) 280–296.
13. Hamilton, H. J., Shan, N., Ziarko, W.: Machine Learning of Credible Classifications. In Proc. of Australian Conf. on Artificial Intelligence AI'97 (1997) 330–339.
14. Hausdorf, C., Muller, C.: A Theory of Interestingness for Knowledge Discovery in Databases Exemplified in Medicine. In Proc. of Int. Workshop on Intelligent Data Analysis in Medicine and Pharmacology IDAMAP'96 (1996) 31–36.
15. Hepatitis Dataset for Discovery Challenge. In Web Page of Euro. Conf. on Principles and Practice of Knowledge Discovery in Databases PKDD'02 (2002) http://lisp.vse.cz/challenge/ecmlpkdd2002/index.html.
16. Hilderman R. J., Hamilton, H. J.: Knowledge Discovery and Measure of Interest. Kluwer Academic Publishers (2001).
17. Hogl, O., Stoyan, H., Stuhlinger, W.: "On Supporting Medical Quality with Intelligent Data Mining," Proc. of Hawaii Int. Conf. on System Sciences (HICSS'01), no. HCDAM03, 2001.
18. Jaroszewicz, S., Simovici, D. A.: A General Measure of Rule Interestingness. In Proc. of Euro. Conf. on Principles of Data Mining and Knowledge Discovery PKDD'01 (2001) 253–265.
19. Kamber, M., Shinghal, R.: Evaluating the Interestingness of Characteristic Rules. In Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD'96 (1996) 263–266.
20. Klementtinen, M., Mannila, H., Ronkainen, P., Toivone, H., Verkamo, A. I.: Finding Interesting Rules from Large Sets of Discovered Association Rules. In Proc. of Int. Conf. on Information and Knowledge Management CIKM'94 (1994) 401–407.
21. Lavrač, N.: Selected Techniques for Data Mining in Medicine. Artificial Intelligence in Medicine, **16** (1999) 3–23.
22. Liu, B., Hsu, W., Chen, S.: Using General Impressions to Analyze Discovered Classification Rules. In Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD'97 (1997) 31–36.
23. Liu, B., Hsu, W., Chen, S., Mia, Y.: Analyzing the Subjective Interestingness of Association Rules. Intelligent Systems, **15 (5)** (2000) 47–55.
24. Liu, B., Hsu, W., Mia, Y.: Identifying Non-Actionable Association Rules. In Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD'01 (2001) 329–334.

25. Liu, H., Lu, H., Feng, L., Hussain, F.: Efficient Search of Reliable Exceptions. In Proc. of Pacific-Asia Conf. on Knowledge Discovery and Data Mining PAKDD'99 (1999) 194–203.
26. Morimoto, Y., Fukuda, T., Matsuzawa, H., Tokuyama, T., Yoda, K.: Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases. In Proc. of Int. Conf. on Very Large Databases VLDB'98 (1998) 380–391.
27. Motoda, H. (eds.): Active Mining. ISO Press (2002).
28. Ohsaki, M., Sato, Y., Yokoi, H., Yamaguchi, T.: A Rule Discovery Support System for Sequential Medical Data, – In the Case Study of a Chronic Hepatitis Dataset –. In Proc. of Int. Workshop on Active Mining AM-2002 in IEEE Int. Conf. on Data Mining ICDM'02 (2002) 97–102.
29. Ohsaki, M., Sato, Y., Kitaguchi, S., Yokoi, H., Yamaguchi, T.: A Rule Discovery Support System for Sequential Medical Data, – In the Case Study of a Chronic Hepatitis Dataset –. Technical Report of the Institute of Electronics, Information, and Communication Engineers IEICE (2003) AI2002-81 (in Japanese).
30. Padmanabhan, B., Tuzhilin, A.: A Belief-Driven Method for Discovering Unexpected Patterns. In Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD'98, (1998) 94–100.
31. Piatetsky-Shapiro, G.: Discovery, Analysis and Presentation of Strong Rules. In Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 229–248.
32. Sahara, S.: On Incorporating Subjective Interestingness Into the Mining Process. In Proc. of IEEE Int. Conf. on Data Mining ICDM'02 (2002) 681–684.
33. Silberschatz, A., Tuzhilin, A.: On Subjective Measures of Interestingness in Knowledge Discovery. In Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD'95 (1995) 275–281.
34. Smyth, P., Goodman, R. M.: Rule Indcution using Information Theory. In Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 159–176.
35. Tan, P., Kumar, V., Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. In Proc. of Int. Conf. on Knowledge Discovery and Data Mining KDD'02 (2002) 32–41.
36. Zhong, N., Yao, Y. Y., Ohshima, M.: Peculiarity Oriented Multi-Database Mining. IEEE Trans. on Knowledge and Data Engineering **15 (4)** (2003) 952–960.

Experimental Evaluation of Time-series Decision Tree

Yuu Yamada Einoshin Suzuki
Electrical and Computer Engineering, Yokohama National University,
79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan.
yuu@slab.dnj.ynu.ac.jp, suzuki@ynu.ac.jp

Hideto Yokoi Katsuhiko Takabayashi
Division for Medical Informatics, Chiba-University Hospital,
1-8-1 Inohana, Chuo, Chiba, 260-8677, Japan
yokoih@telemed.ho.chiba-u.ac.jp, takaba@ho.chiba-u.ac.jp

Abstract

In this paper, we give experimental evaluation of our time-series decision tree induction method under various conditions. It has been empirically observed that the method induces accurate and comprehensive decision trees in time-series classification, which has gaining increasing attention due to its importance in various real-world applications. The evaluation has revealed several important findings including interaction between a split test and its goodness.

1 Introduction

Time-series data are employed in various domains including politics, economics, science, industry, agriculture, and medicine. Classification of time-series data is related to many promising application problems. For instance, an accurate classifier for liver cirrhosis from time-series data of medical tests might replace a biopsy which picks liver tissue by inserting an instrument directly into liver. Such a classifier is highly important since it would substantially reduce costs of both patients and hospitals.

Our time-series decision tree represents a novel classifier for time-series classification. Our learning method for the time-series decision tree has enabled us to discover a classifier which is highly appraised by domain experts [8]. In this paper, we perform extensive experiments based on advice from domain experts, and investigate on various characteristics of our time-series decision tree.

2 Time-series Decision Tree

2.1 Time-series Classification

A time sequence $\boldsymbol{A}$ represents a list of values $\alpha_1, \alpha_2, \cdots, \alpha_I$ sorted in chronological order. For simplicity, this paper assumes that the values are obtained or sampled with an equivalent interval (= 1).

A data set D consists of n examples $e_1,\ e_2,\ \cdots,\ e_n$, and each example e_i is described by m attributes $a_1,\ a_2,\ \cdots,\ a_m$ and a class attribute c. An attribute a_j can represent a time-series attribute which takes a time sequence as its value. The class attribute c represents a nominal attribute and its value is called a class. In time-series classification, the objective represents induction of a classifier, which predicts the class of an example e, given a training data set D.

2.2 Learning Time-series Decision Tree

Our time-series tree [8] has a time sequence which exists in data and an attribute in its internal node, and splits a set of examples according to the dissimilarity of their corresponding time sequences to the time sequence. The use of a time sequence which exists in data in its split node contributes to comprehensibility of the classifier, and each time sequence is obtained by exhaustive search. The dissimilarity measure is based on dynamic time warping (DTW) [6].

We call this split test a standard-example split test. A standard-example split test $\sigma(e, a, \theta)$ consists of a standard example e, an attribute a, and a threshold θ. Let a value of an example e in terms of a time-series attribute a be $e(a)$, then a standard-example split test divides a set of examples $e_1, e_2, \cdots, e_n$ to a set $S_1(e, a, \theta)$ of examples each of which $e_i(a)$ satisfies $G(e(a), e_i(a)) < \theta$ and the rest $S_2(e, a, \theta)$. We also call this split test a θ-guillotine cut.

As the goodness of a split test, we have selected gain ratio [7] since it is frequently used in decision-tree induction. Since at most $n - 1$ split points are inspected for an attribute in a θ-guillotine cut and we consider each example as a candidate of a standard example, it frequently happens that several split points exhibit the largest value of gain ratio. We assume that consideration on shapes of time sequences is essential in comprehensibility of a classifier, thus, in such a case, we define that the best split test exhibits the largest gap between the sets of time sequences in the child nodes. The gap $gap(e, a, \theta)$ of $\sigma(e, a, \theta)$ is equivalent to $G(e''(a), e(a)) - G(e'(a), e(a))$ where e' and e'' represent the example $e_i(a)$ in $S_1(e, a, \theta)$ with the largest $G(e(a), e_i(a))$ and the example $e_j(a)$ in $S_2(e, a, \theta)$ with the smallest $G(e(a), e_j(a))$ respectively. When several split tests exhibit the largest value of gain ratio, the split test with the largest $gap(e, a, \theta)$ among them is selected.

We have also proposed a cluster-example split test $\sigma'(e', e'', a)$ for comparison. A cluster-example split test divides a set of examples $e_1, e_2, \cdots, e_n$ into a set $U_1(e', e'', a)$ of examples each of which $e_i(a)$ satisfies $d(e'(a), e_i(a)) < d(e''(a), e_i(a))$ and the rest $U_2(e', e'', a)$. The goodness of a split test is equivalent to that of the standard-example split test without θ.

2.3 Experimental Results and Comments from Domain Experts

We have evaluated our method with Chronic hepatitis data [1], the Australian sign language data [4], and the EEG data [4]. As a result of pre-processing, we have obtained two data sets, which we call H1 and H2, from Chronic hepatitis data. Similarly, two data sets, which we call Sign and EEG, have been generated from the Australian sign language data and the EEG data respectively. The classification tasks in H1 and H2 are prediction of liver cirrhosis from medical tests data. We have employed time sequences each of which has more than 9 test values during a period of before 500 days and after 500 days of a biopsy. In both data sets, there are 30 examples of liver cirrhosis and 34 examples of the other class. Since the intervals of medical tests differ, we have employed liner interpolation between two adjacent values and transformed each time sequence to a time sequence of 101 values with a 10-day interval. In H1, one of us, who is a physician, suggested to use in classification 14 attributes (GOT, GPT, ZTT, TTT, T-BIL, I-BIL, D-BIL, T-CHO, TP, ALB, CHE, WBC, PLT, HGB) which are important in hepatitis. In H2, we have measured shifts for each of these attributes from its average value and employed these 14 attributes in addition to the original attributes. Experimental results have confirmed that our induction method constructs comprehensive and accurate decision trees.

We have prepared another data set, which we call H0, from the chronic hepatitis data by dealing with the first biopsies only. H0 consists of 51 examples (21 LC patients, and 30 non-LC patients) each of which is described with 14 attributes. Experimental results show that our time-series tree is promising for knowledge discovery. We show a time-series decision tree learned from H0 in figure 1.

We obtained the following comments from medical experts.

- The proposed learning method exhibits novelty and is highly interesting. The splits in the upper parts of the time-decision trees are valid, and the learning results are surprisingly well as a method which employs domain knowledge on attributes only.
- Medical test values which are measured after a biopsy are typically influenced by treatment such as interferon (IFN). It would be better to use only medical test values which were measured before a biopsy.
- 1000 days are long as a period of measurement since the number n of patients is small. It would be better to use shorter periods such as 365 days.
- The number of medical tests might be possibly reduced to 4 per year. Prediction from a smaller number of medical tests has a higher impact on clinical treatment.
- A medical expert is familiar with sensitivity, specificity, and an ROC curve as evaluation indices of a classifier. It causes more problems to overlook an LC patient than mistake a non-LC patient.

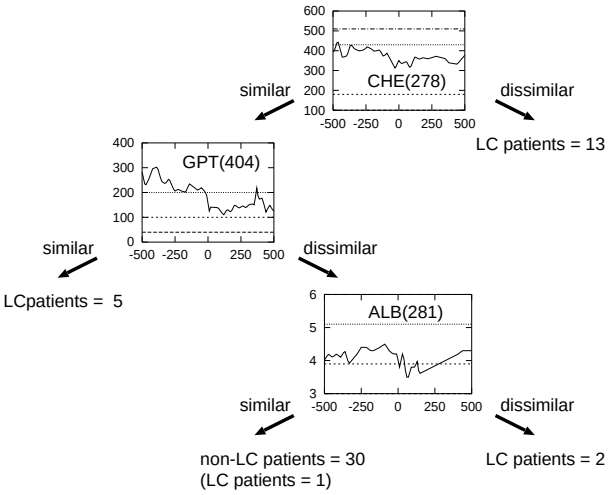


Fig. 1. Time-series tree learned from H0 (chronic hepatitis data of the first biopsies)

3 Experiments for Misclassification Costs

3.1 Conditions of Experiments

Based on a comment in the previous section, we evaluated our time-series decision tree without using medical test data after a biopsy. For a continuous attribute, C4.5 [7] employs a split test which verifies whether a value is greater than a threshold. This split test will be called an average-split test in this paper. We call our approach which employs both the standard-example split test and the average-split test a combined-split test. For the sake of comparison, we also employed the average-split test alone and a line-split test, which replaces a standard example by a line segment. A line segment in the latter method is obtained by discretizing test values by an equal-frequency method with $\alpha - 1$ bins, and connecting two points (l_1, p_1) and (l_2, p_2) where l_1 and l_2 represent the beginning and the end of a measurement respectively. Each of p_1 and p_2 represents one of the end values of discretized bins. For instance, it considers 25 line segments if $\alpha = 5$. The cluster-example split test was not employed since it exhibited poor performance in [8].

Table 1. Confusion matrix

	LC	non-LC
LC (Prediction)	TP	FP
non-LC (Prediction)	FN	TN

We show a confusion matrix in table 1. As the domain experts stated, it is important to decrease the number FN of overlooked LC patients than the number FP of mistaken non-LC patients. Therefore, we employ sensitivity, specificity, and (misclassification) cost in addition to predictive accuracy as evaluation indices. The added indices are considered to be important in the following order.

$$Cost = \frac{C\,FN + FP}{C(TP + FN) + (TN + FP)} \tag{1}$$

$$Sensitivity \text{ (True Positive Rate)} = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity \text{ (True Negative Rate)} = \frac{TN}{TN + FP} \tag{3}$$

where C represents a user-specified weight. We settled $C = 5$ throughout the experiments, and employed a leave-one-out method. Note that $Cost$ is normalized in order to facilitate comparison of experimental results from different data sets.

It is reported that Laplace correction is effective in decision tree induction for cost-sensitive classification [2]. We obtained the probability $\Pr(a)$ of a class a when there are $\nu(a)$ examples of a among ν examples as follows.

$$\Pr(a) = \frac{\nu(a) + l}{\nu + 2l} \tag{4}$$

where l represents a parameter of the Laplace correction. We settled $l = 1$ unless stated.

We modified data selection criteria in each series of experiments and prepared various data sets as shown in table 2. In a name of a data set, the first figure represents the selected period of measurement before a biopsy, the figure subsequent to a "p" represents the number of required medical tests, and the figure subsequent to an "i" represents the number of days of an interval in interpolation. Since we employed both B-type patients and C-type patients in all experiments, each name of a data set contains strings "BC". Since we had obtained novel data of biopsies after [8], we employed an integrated version in the experiments.

3.2 Experimental Results

Firstly, we modified the required number of medical tests to 6, 3, 2 under a 180-day period and a 5-day interpolation interval. We show the results in table 3. From the table, we see that the average-split test and the line-split test outperform other methods in cost for p2 and p6 respectively. For p3, the methods exhibit the same cost and outperform our standard-example split test. We believe that the poor performance of our method is due to lack of information on shapes of time sequences and the number of examples. We interpret the results that lack of the former information in p2 favors the average-split test, while lack of the latter information in p6 favors the line-split test. If simplicity of a classifier is also considered, the decision tree learned with the average-split test from p2 would be judged as the best.

Secondly, we modified the selected period to 90, 180, 270, 360 days under an interpolation interval 5 days and the number of required medical tests per 30 days 1. We show the results in tables 4 and 5. From table 4, we see that the average-split test and the line-split test almost always outperform our standard-example split test in cost though there is no clear winner between them. We again attribute these to lack of information on shapes of time sequences and the number of examples. Our standard-example split test performs relatively well for 90 and 180 and this would be due to their relatively large numbers of examples. If simplicity of a classifier is also considered, the decision tree learned with the line-split test from 180 would be judged as the best.

Thirdly, we modified the interpolation intervals to 2, 4, $\cdots$, 10 days under a 180-day period and the required number of medical tests 6. We show the results in tables 6 and 7. From the table 6, we see that our standard-example split test and the line-split test outperform the average-split test in cost though there is no clear winner between them. Since a 180 in tables 4 and 5 represents 180BCp6i5, it would be displayed as i5 in this table. Our poor performance of cost 0.35 for i5 shows that our method exhibits good performance for small and large intervals, and this fact requires further investigation. If simplicity of a classifier is also considered, the line-split test is judged as the best and we again attribute this to lack of information for our method.

Lastly, we modified the Laplace correction parameter l to $0, 1, \cdots, 5$ under a 180-day period, the required number of medical tests 6, and a 6-day interpolation interval. We show the results in table 8. From the table, we see that the Laplace correction increases cost for our standard-example split test and the line-split test contrary to our expectation. Even for the average-split test, the case without the Laplace correction (l = 0) rivals the best case with the Laplace correction (l = 1). The table shows that these come from the fact that the Laplace correction lowers sensitivity but this requires further investigation.

Table 2. Data sets employed in the experiments

experiments	data (# of non-LC patients : # of LC patients)
experiments for the number of medical tests	180BCp6i5 (68:23), 180BCp3i5 (133:40), 180BCp2i5 (149:42)
experiments for the selected period	90BCp3i5 (120:38), 180BCp6i5 (68:23), 270BCp9i5 (39:15), 360BCp12i5 (18:13)
experiments for the interpolation interval	180BCp6i2, 180BCp6i4, 180BCp6i6, 180BCp6i8, 180BCp6i10 (all 68:23)

Table 3. Results of experiments for test numbers, where data sets p6, p3, and p2 represent 180BCp6i5, 180BCp3i5, and 180BCp2i5 respectively

	accuracy (%)			size			cost			sensitivity			specificity		
method	p6	p3	p2	p6	p3	p2	p6	p3	p2	p6	p3	p2	p6	p3	p2
Combined	78.0	75.7	80.6	10.9	20.5	18.9	0.35	0.35	0.33	0.52	0.53	0.52	0.87	0.83	0.89
Average	83.5	82.1	87.4	3.2	24.7	7.4	0.39	0.27	0.27	0.39	0.63	0.57	0.99	0.88	0.96
Line	84.6	82.7	85.9	9.0	22.7	3.6	0.30	0.27	0.34	0.57	0.63	0.43	0.94	0.89	0.98

Table 4. Results for accuracy, size, and cost of experiments for periods, where data sets 90, 180, 270, and 360 represent 90BCp3i5, 180BCp6i5, 270BCp9i5, and 360BCp12i5 respectively

	accuracy (%)				size				cost			
method	90	180	270	360	90	180	270	360	90	180	270	360
Combined	77.8	78.0	64.8	45.2	19.5	10.9	8.5	5.5	0.36	0.35	0.52	0.69
Average	79.7	83.5	79.6	71.0	23.7	3.2	8.7	6.4	0.30	0.39	0.41	0.40
Line	77.2	84.6	74.1	48.4	18.7	9.0	8.7	6.5	0.41	0.30	0.40	0.58

Table 5. Results for sensitivity and specificity of experiments for periods

	sensitivity				specificity			
method	90	180	270	360	90	180	270	360
Combined	0.50	0.52	0.33	0.23	0.87	0.87	0.77	0.61
Average	0.61	0.39	0.40	0.54	0.86	0.99	0.95	0.83
Line	0.39	0.57	0.47	0.38	0.89	0.94	0.85	0.56

Table 6. Results for accuracy, size, and cost of experiments for intervals, where data sets i2, i4, i6, i8, and i10 represent 180BCp6i2, 180BCp6i4, 180BCp6i6, 180BCp6i8, and 180BCp6i10 respectively

	accuracy (%)					size					cost				
method	i2	i4	i6	i8	i10	i2	i4	i6	i8	i10	i2	i4	i6	i8	i10
Combined	85.7	85.7	82.4	81.3	82.4	10.9	10.9	12.4	12.3	12.4	0.29	0.31	0.33	0.33	0.33
Average	84.6	84.6	83.5	84.6	82.4	3.0	3.0	3.2	3.9	5.1	0.36	0.36	0.39	0.36	0.39
Line	85.7	83.5	83.5	84.6	79.1	9.0	9.0	8.9	9.1	11.2	0.29	0.32	0.32	0.30	0.32

Table 7. Results for sensitivity and specificity of experiments for intervals

	sensitivity					specificity				
method	i2	i4	i6	i8	i10	i2	i4	i6	i8	i10
Combined	0.57	0.52	0.52	0.52	0.52	0.96	0.97	0.93	0.91	0.93
Average	0.43	0.43	0.39	0.43	0.39	0.99	0.99	0.99	0.99	0.97
Line	0.57	0.52	0.52	0.57	0.57	0.96	0.94	0.94	0.94	0.87

Table 8. Results of experiments for Laplace correction values with 180BCp6i6, where methods C, A, and L represent Combined, Average, and Line respectively

	accuracy (%)			size			cost			sensitivity			specificity		
value	C	A	L	C	A	L	C	A	L	C	A	L	C	A	L
0	86.8	85.7	82.4	10.9	10.8	7.4	0.28	0.29	0.33	0.57	0.57	0.52	0.97	0.96	0.93
1	82.4	83.5	83.5	12.4	3.2	8.9	0.33	0.39	0.32	0.52	0.39	0.52	0.93	0.99	0.94
2	81.3	83.5	80.2	9.1	3.0	9.0	0.36	0.39	0.38	0.48	0.39	0.43	0.93	0.99	0.93
3	83.5	73.6	83.5	9.1	2.5	9.0	0.30	0.63	0.34	0.57	0.00	0.48	0.93	0.99	0.96
4	81.3	83.5	79.1	9.2	2.6	8.9	0.36	0.39	0.39	0.48	0.39	0.43	0.93	0.99	0.91
5	82.4	83.5	82.4	9.1	2.7	8.9	0.35	0.39	0.37	0.48	0.39	0.43	0.94	0.99	0.96

3.3 Analysis of Experiments

In the experiments of [8], we employed longer time sequences and a larger number of training examples than in this paper. It should be also noted that the class ratio in [8] was nearly equivalent. We believe that our time-series decision tree is adequate for this kind of classification problems. The classification problems in this paper, since they neglect medical tests data after a biopsy, exhibit opposite characteristics, favoring a robust method such as the average-split test. Though it is appropriate to neglect medical tests data after a biopsy from medical viewpoint, the effect is negative for our time-series decision tree.

The decision trees which were constructed using the split tests contain many LC-leaves, especially those with the combined-split test. Most of the leaves contain a small number of training examples, thus they rarely correspond to a test example. This observation led us to consider modifying tree-structures in order to decrease cost.

4 Experiments for Goodness of a Split Test

4.1 Motivations

Table 9. Two examples of a split test

Split test	Left	Right	gain	gain ratio
test 1	6 (0, 6)	113 (76, 37)	0.078	0.269
test 2	47 (42, 5)	72 (34, 38)	0.147	0.152

From the discussions in the previous section, we considered to use the medical tests data after a biopsy and to replace gain ratio by gain. The former was realized by using the data sets employed in [8]. For the latter, consider their characteristics as goodness of a split test with tests 1 and 2 in table 9. Tests 1 and 2 are selected with gain ratio and gain respectively. As stated in [7], gain ratio tends to select an unbalanced split test where a child node has an extremely small number of examples. We believe that example 1 corresponds to this case and decided to perform a systematic comparison of the two criteria.

4.2 Experiments

We have compared our standard-example split test, the cluster-example split test, the average-split test, a method by Geurts [3], and a method by Kadous [5]. We settled N_{max} = 5 in the method of Geurts, and the number of discretized bins 5 and the

number of clusters 5 in the method of Kadous. Experiments were performed with a leave-one-out method, and without the Laplace correction.

We show the results in table 10, and the decision trees learned from all data with the standard-example split test, the cluster-example split test, and the average-split test in figures 2, 3, and 4 respectively. The conditions are chosen so that each of them exhibits the lowest cost for the corresponding method.

From the table, we see that our standard-example split test performs better with gain ratio, and the cluster-example split test and the average-split test perform better with gain. We think that the former is due to affinity of gain ratio, which tends to select an unbalanced split, to our standard-example split test, which splits examples based on their similarities or dissimilarities to its standard example. Similarly, we think that the latter is due to affinity of gain, which is known to exhibit no such tendency, to the cluster-example split test and the average-split test, both of which consider characteristics of two children nodes in split. Actually, we have confirmed that a split test tends to produce a small-sized leaf with gain ratio while a split test tends to construct a relatively balanced split with gain.

Table 10. Experimental results with gain and gain ratio

		accuracy (%)		size		cost		sensitivity		specificity	
method	goodness	H1	H2	H1	H2	H1	H2	H1	H2	H1	H2
SE-split	gain	64.1	78.1	10.6	7.2	0.34	0.25	0.67	0.73	0.62	0.82
	gain ratio	79.7	85.9	9.0	7.1	0.24	0.18	0.73	0.80	0.85	0.91
CE-split	gain	81.2	76.6	9.0	8.7	0.20	0.23	0.80	0.77	0.82	0.76
	gain ratio	65.6	73.4	9.4	7.2	0.36	0.31	0.63	0.67	0.68	0.79
AV-split	gain	79.7	79.7	7.8	10.8	0.22	0.24	0.77	0.73	0.82	0.85
	gain ratio	73.4	70.3	10.9	11.4	0.31	0.39	0.67	0.57	0.79	0.82
Geurts	gain	68.8	70.3	10.1	9.7	0.28	0.32	0.73	0.67	0.65	0.74
	gain ratio	71.9	67.2	10.0	9.2	0.29	0.29	0.70	0.73	0.74	0.62
Kadous	gain	65.6	62.5	12.6	12.0	0.38	0.41	0.60	0.57	0.71	0.68
	gain ratio	71.9	65.6	8.8	13.2	0.29	0.27	0.70	0.77	0.74	0.56
1-NN		82.8	84.4	N/A	N/A	0.19	0.18	0.80	0.80	0.85	0.88

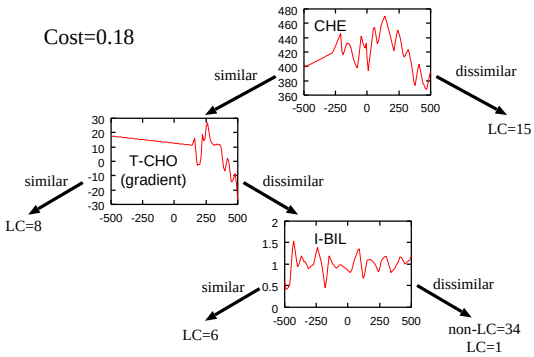


Fig. 2. SE-split decision tree (H2, GainRatio)

5 Conclusions

For our time-series decision tree, we investigated the case in which medical tests before a biopsy are neglected and the case in which goodness of a split test is altered. In the former case, our time-series decision tree is outperformed by simpler

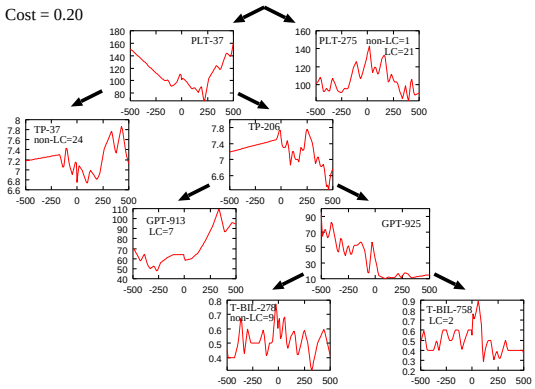


Fig. 3. CE-split decision tree (H1, Gain)

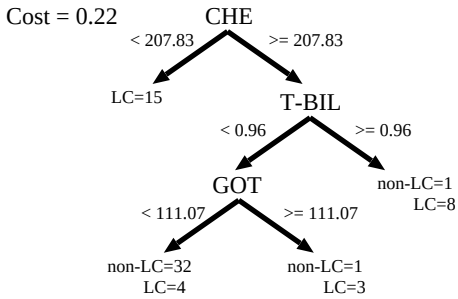


Fig. 4. AV-split decision tree (H1, Gain)

decision trees in misclassification cost due to lack of information on sequences and examples. In the latter case, our standard-example split test performs better with gain ratio, and the cluster-example split test and the average-split test perform better with gain probably due to affinities in each combination. We plan to extend our approach as both a cost-sensitive learner and a discovery method.

Acknowledgement

This work was partially supported by the grant-in-aid for scientific research on priority area “Active Mining” from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

1. P. Berka. ECML/PKDD 2002 discovery challenge, download data about hepatitis. http://lisp.vse.cz/challenge/ ecmlpkdd2002/, 2002. (current September 28th, 2002).
2. J. P. Bradford, C. Kunz, R. Kohavi, C. Brunk, and C. E. Brodley. Pruning decision trees with misclassification costs. In *Proc. Tenth European Conference on Machine Learning (ECML)*, pages 131–136. 1998.
3. P. Geurts. Pattern extraction for time series classification. In *Principles of Data Mining and Knowledge Discovery (PKDD), LNAI 2168*, pages 115–127. 2001.
4. S. Hettich and S. D. Bay. The UCI KDD archive. http://kdd.ics.uci.edu, 1999. Irvine, CA: University of California, Department of Information and Computer Science.
5. M. W. Kadous. Learning comprehensible descriptions of multivariate time series. In *Proc. Sixteenth International Conference on Machine Learning (ICML)*, pages 454–463. 1999.
6. E. J. Keogh. Mining and indexing time series data. http:// www.cs.ucr.edu/%7Eeamonn/tutorial_on_time_series.ppt, 2001. Tutorial at the 2001 IEEE International Conference on Data Mining (ICDM).
7. J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, Calif., 1993.
8. Y. Yamada, E. Suzuki, H. Yokoi, and K. Takabayashi. Decision-tree induction from time-series data based on a standard-example split test. In *Proc. Twentieth International Conference on Machine Learning (ICML)*, pages 840–847. 2003.

Extracting Diagnostic Knowledge from Hepatitis Dataset by Decision Tree Graph-Based Induction

Warodom Geamsakul, Tetsuya Yoshida, Kouzou Ohara,
Hiroshi Motoda, and Takashi Washio

Institute of Scientific and Industrial Research, Osaka University, JAPAN
{warodom, yoshida, ohara, motoda, washio}@ar.sanken.osaka-u.ac.jp

Abstract. Decision Tree Graph-Based Induction (DT-GBI) is a technique for constructing a decision tree from graph-structured data. In DT-GBI, substructures (discriminative patterns) are extracted by stepwise pair expansion (pair-wise chunking) and used as test attributes at nodes of a decision tree. We applied DT-GBI to a classification task of hepatitis dataset. In the first two experiments, a stage of fibrosis is used as a class and decision trees are constructed for discriminating patients with F4 (cirrhosis) and patients with the other stages, using only the time sequence data of blood inspection. In the third experiment, the types of hepatitis (B and C) are used as classes and decision trees are constructed as in the first experiment. The preliminary results of experiments, both constructed decision trees and their prediction accuracies, are reported in this paper.

Keywords: Data mining, graph-structured data,
Decision Tree Graph-Based Induction, hepatitis data analysis

Correspondent:
Warodom Geamsakul
Motoda lab., I.S.I.R., Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
warodom@ar.sanken.osaka-u.ac.jp

GBI(G)
  Enumerate all the pairs P_{all} in G
  Select a subset P of pairs from P_{all} (all the pairs in G) based on typicality criterion
  Select a pair from P_{all} based on chunking criterion
  Chunk the selected pair into one node c
  G_c := contracted graph of G
  while termination condition not reached
    $P := P \cup$ GBI(G_c)
  return P

Fig. 1. Algorithm of GBI

1 Introduction

Viral hepatitis is a very critical illness. If it is left without undergoing a suitable medical treatment, a patient may suffer from cirrhosis and fatal liver cancer. The progress speed of condition is slow and subjective symptoms are not noticed easily. Hence, in many cases, it has already become very severe when subjective symptoms are noticed. Although periodical inspection and proper treatment are important in order to prevent this situation, there are problems of expensive cost and physical burden on a patient. There is an alternative much cheaper method of inspection such as blood test. However, the amount of data becomes enormous since the progress speed of condition is slow.

The hepatitis dataset we are attempting to analyse is a real-world data provided by Chiba University Hospital. There are some other analyses already conducted and reported on this dataset. [8] analysed the data by constructing decision trees from time-seires data without discretizing numeric values. [1] proposed a method of temporal abstraction to handle time-series data, converted time phenomena to symbols and used a standard classifier. [6] used multiscale matching to compare time-series data and clustered them using rough set theory. [3] also clustered the time-series data of a certain time interval into several categories and used a standard classfier.

We have proposed a method called Decision Tree Graph-Based Induction (DT-GBI), which constructs a classifier (decision tree) for graph-structured data while simultaneously constructing attributes themselves for classification using GBI [7]. We conducted experiments to test our DT-GBI using this hepatitis data. The stages of fibrosis are used as classes in the first two experiments, and the types of hepatitis (B and C) are used as classes in the third experiment. The decision trees are constructed to discriminate between two groups of patients using no biopsy results but only the time sequence of blood inspection.

2 Decision Tree Graph-Based Induction

2.1 Graph-Based Induction (GBI)

GBI employs the idea of extracting typical patterns by stepwise pair expansion (we call this process "chunking"). In GBI, an assumption is made that typi-

```
DT-GBI(D)
  Create a node DT for D
  if termination condition reached
    return DT
  else
    P := GBI(D) (with the number of chunking specified)
    Select a pair p from P
    Divide D into D_y (with p) and D_n (without p)
    Chunk the pair p into one node c
    D_yc := contracted data of D_y
    for D_i := D_yc, D_n
      DT_i := DT-GBI(D_i)
      Augment DT by attaching DT_i as its child along yes(no) branch
  return DT
```

Fig. 2. Algorithm of DT-GBI

cal patterns represent some concepts and "typicality" is characterized by the pattern's frequency or the value of some evaluation function based on its frequency. Repeated chunking enables GBI to extract typical patterns of various sizes. The search is greedy and no backtracking is made. Becuase of this, some typical patterns that exist in the input graph may not be extracted. However, GBI's objective is not to find all typical patterns nor all frequent patterns, but to extract only meaningful typical patterns of certain sizes. The stepwise pair expansion algorithm is summarized in Figure 1.

2.2 Beam-wise Graph-Based Induction (B-GBI)

Since the search in GBI is greedy and no backtracking is made, which patterns extracted by GBI depends on which pair is selected for chunking. There can be many patterns which are not extracted by GBI. A beam search is incorporated to GBI, still, within the framework of greedy search [2] in order to relax this problem, increase the search space, and extract more discriminative patterns while still keeping the computational complexity within a tolerant level,. A certain fixed numbers of pairs ranked from the top are selected to be chunked individually in parallel. To prevent each branch growing exponentially, the total numbers of pairs to chunk (the beam width) is fixed at every time of chunking. Thus, at any iteration step, there is always a fixed number of chunking that is performed in parallel.

2.3 Feature Construction by B-GBI

If pairs are expanded in a step-wise fashion by B-GBI and discriminative ones are selected and further expanded while constructing a decision tree, discriminative patterns (subgraphs) can be constructed simultaneously while constructing a decision tree. The algorithm of DT-GBI is summarized in Figure 2. Here, we regard a substructure (subgraph) in a graph as an attribute so that graph-structured data can be represented with attribute-value pairs according to the

existence of particular subgraph. Since the values for an attribute are yes (this graph contains pair) and no (this graph does not contain pair), the constructed decision tree is represented as a binary tree. Every time when an attribute (pair) is selected to split the data, the pair is chunked into a larger node in size. Thus, although initial pairs consist of only two nodes and one link between them, attributes useful for classification task are gradually grown up into larger pairs (subgraphs) by applying chunking recursively.

2.4 Pruning Decision Tree

Recursively partitioning data until each subset in the partition contains data of a single class often results in overfitting to the training data and thus degrades the prediction accuracy, a pessimistic pruning used in C4.5 [5] is implemented by growing an overfitted tree first and then pruning it based on the confidence interval for binomial distribution.

3 Data Preprocessing

The dataset contains long time-series data (from 1982 to 2001) on laboratory examination of 771 patients of hepatitis B and C. The data can be broadly split into two categories. The first category includes administrative information such as patient's information (age and date of birth), pathological classification of the disease, date of biopsy, and result. The second category includes temporal record of blood test and urinalysis. It contains the result of 983 types of both in- and out-hospital examinations.

3.1 Cleansing In numeric attributes, letters and symbols such as H, L, or + are deleted. Values in nominal attributes are left as they are.

3.2 Conversion to table When converting the given data into an attribute-value table, both a patient ID (MID) and a date of inspection are used as search keys and an inspection item is defined as an attribute. Since it is not necessary that all patients must take all inspections, there are many missing values after this data conversion. No attempt is made to estimate these values. However, this situation is not an obstacle for extracting discriminative patterns as those missing values simply are not represented in graph-structured data.

3.3 Average and discretization This step is necessary due to the following two reasons: 1) obvious change in every time of inspection may not be found because the progress of hepatitis itself is slow, and 2) the date of visit is not synchronized across different patients. In this step, the numeric attributes are averaged and the most frequent value is used for nominal attributes over two-month interval. Further, for some inspections (GOT, GPT, TTT, and ZTT), standard deviations are calculated over six-month interval and added as new attributes. When we represent an inspection result as a node label, the number of node labels become too large and that causes the low efficiency of DT-GBI

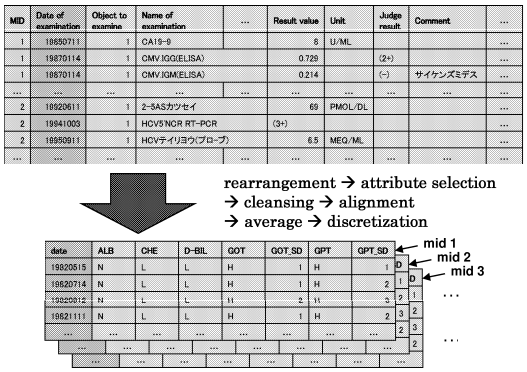


Fig. 3. An example of graph conversion in phase 1-2

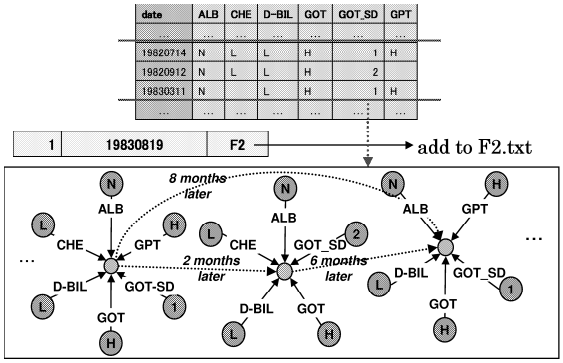


Fig. 4. An example of graph conversion in phase 3

as frequent and discriminative patterns cannot be extracted properly. Therefore, we reduced the number of node label by discretizing attribute values. For general numerical values, the normal ranges are specified and values are discretized into three ("L" for low, "N" for normal, and "H" for high). Based on the range, the standard deviations of GOT and GPT are discretized into five ("1" for the smallest deviation, "2","3","4","5" for the largest deviation), while the standard deviations of TTT and ZTT are discretized into three ("1" for the smallest deviation,"2","3" for the largest deviation).

Figure 3 illustrates the mentioned three steps of data conversion.

3.4 Limitation of data range Many patients in this dataset took biopsy only once. The longer the interval between the date of biopsy and the date of blood inspection is, the less reliable the correlation between them is. Besides, for a patient that biopsy was operated for several times, the treatment (*e.g.,* interferon medication) after a biopsy may influence the result of blood inspection and lower the reliability of data. We consider that the pathological conditions in the range from 500 days before to 500 days after the first biopsy remains the same. Therefore, we use only the data in this 1000-day range for the analysis.

As one record of data corresponds to two-month interval, the largest number of records in one patient is 17.

3.5 Conversion to graph structure When analysing data by DT-GBI, it is necessary to convert data to graph structure. One patient record is mapped into one directed graph. Assumption is made that there is no direct correlation between two sets of pathological tests that are more than a predefined interval (here, two years) apart. Hence, time correlation is considered only within this interval. Figure 4 shows an example of converted graph-structured data. In this figure, a star-shaped subgraph represents values of a set of pathological examination in a two-month interval. The centre node of the subgraph is a hypothetical node for the two-month interval. An edge pointing to a hypothetical node represents an examination. The node connected to the edge represents the value (preprocessed result) of the examination. The edge linking two hypothetical nodes represents time difference.

3.6 Class label setting In the first and second experiments, we set the result (progress of fibrosis) of the first biopsy as class. In the third experiment, we set the subtype (B or C) as class.

4 Preliminary Results

4.1 Initial Settings

To apply DT-GBI, we use two criteria for selecting pairs. One is frequency for selecting pairs to chunk, and the other is information gain [4] for finding discriminative patterns after chunking.

A decision tree is to be constructed in either of the following two ways: 1) apply chunking N_r=20 times at the root node and only once at the other nodes of a decision tree, 2) apply chunking N_e=20 times at every node of a decision tree. Decision tree pruning is conducted by postpruning: conduct pessimistic pruning by setting the confidence level to 25%. We evaluated the prediction accuracy of decision trees constructed by DT-GBI by the average of 10 cycles of 10-fold cross-validation. In the first cycle, search beam width is adjusted from 1 to 20. The narrowest beam width that brings to the lowest error rate is set as the beam width in the remaining 9 cycles.

4.2 Classifying Patients with Fibrosis Stages

Fibrosis stages are categorized into five stages: F0 (normal), F1, F2, F3, and F4 (severe = cirrhosis). We constructed decision trees which distinguish the patients at F4 stage from the patients at the other stages. In the following two experiments, we used 32 attributes. They are: ALB, CHE, D-BIL, GOT, GOT_SD, GPT, GPT_SD, HBC-AB, HBE-AB, HBE-AG, HBS-AB, HBS-AG, HCT, HCV-AB, HCV-RNA, HGB, I-BIL, ICG-15, MCH, MCHC, MCV, PLT, PT, RBC, T-BIL, T-CHO, TP, TTT, TTT_SD, WBC, ZTT, and ZTT_SD. Table 1 shows the size of graphs after the data conversion.

Table 1. Size of graphs (classified by fibrosis stage)

Stage	F0	F1	F2	F3	F4	All
No. of graphs	4	125	53	37	43	262
Avg. No. of node	303	304	308	293	300	303
Max. No. of node	349	441	420	414	429	441
Min. No. of node	254	152	184	182	162	152

Table 2. Average error rates (%) in exp. 1 and 2

	Experiment 1		Experiment 2	
cycle	N_r=20	N_e=20	N_r=20	N_e=20
1	14.81	11.11	27.78	25.00
2	13.89	11.11	26.85	25.93
3	15.74	12.03	25.00	19.44
4	16.67	15.74	27.78	26.68
5	16.67	12.96	25.00	22.22
6	15.74	14.81	23.15	21.30
7	12.96	9.26	29.63	25.93
8	17.59	15.74	25.93	22.22
9	12.96	11.11	27.78	21.30
10	12.96	11.1	27.78	25.00
average	**15.00**	**12.50**	**26.67**	**23.52**
SD	**1.65**	**2.12**	**1.80**	**2.39**

As shown in Table 1, the number of instances (graphs) in cirrhosis (F4) stage is 43 while the number of instances (graphs) in non-cirrhosis stages (F0+F1+F2+F3) is 219. Unbalance in the number of instances may cause a biased decision tree. In order to relax this problem, we limited the number of instances to the 2:3 (cirrhosis:non-cirrhosis) ratio which is the same as in [8]. Thus, we used all instances from F4 stage for cirrhosis class and select 65 instances from the other stages for non-cirrhosis class, 108 instances in all. How we selected these 108 instances is described later.

4.2.1 Experiment 1: F4 stage vs {F0+F1} stages

All 4 instances in F0 and 61 instances in F1 stage were used for non-cirrhosis class in this experiment. In the first cycle of 10-fold cross-validation, the beam width was varied from 1 to 15. The prediction error rates reached the lowest level when the width was 15 for both methods (1)N_r=20, 2)N_e=20). Thus, in the remaining nine cycles, we set the beam width to 15 when running DT-GBI.

The overall result is summarized in the left half of Table 2. The average error rate was 15.00% for 1) (N_r=20) and 12.50% for 2) (N_e=20). Figure 5 and Figure 6 show one of the decision trees each from the cycle with the lowest error rate (cycle 7) and from the cycle with the highest error rate (cycle 8) respectively. Comparing the both decision trees, there are three pairs of identical patterns appeared at the upper level of each tree.

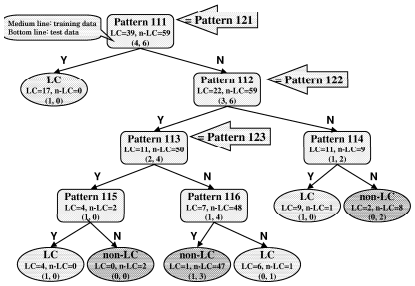


Fig. 5. One of trees from the best cycle in exp.1 (N_e=20)

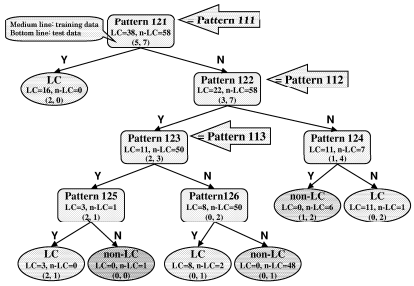


Fig. 6. One of trees from the worst cycle in exp.1 (N_e=20)

4.2.2 Experiment 2: F4 stage vs {F3+F2} stages

In this experiment, we used all instances in F3 and 28 instances in F2 stage for non-cirrhosis class. As in experiment 1, we performed 10 cycles of 10-fold cross-validation. The lowest prediction error rate was obtained in the first cycle when beam width was set to 14 for both 1) and 2). Thus, we set beam width to 14 when operating DT-GBI in the remaining nine cycles.

The overall result is summarized in the right half of Table 2. The average error rate was 26.67% for 1) (N_r=20) and 23.52% for 2) (N_e=20). Figure 9 and Figure 10 show examples of decision trees each from the cycle with the lowest error rate (cycle 3) and the cycle with the highest error rate (cycle 4) respectively. Comparing the both decision trees, there are two pairs of identical patterns appeared at the upper level of each tree.

4.2.3 Discussion

The average prediction error rate in the first experiment is better than that in the second experiment, as the difference in characteristics between data in F4 stage and data in {F0+F1} stages is intuitively larger than that between data in F4 stage and data in {F3+F2}. The averaged error rate of 12.50% in experiment 1 is fairly comparable to that of 11.8% obtained by the decision tree reported in [8].

Patterns shown in Figure 7, 8, 11, and 12 are sufficiently discriminative since all of them are used at the nodes in the upper level of all decision trees, meaning

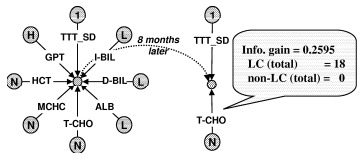


Fig. 7. Pattern 111 = Pattern 121, if exist then LC

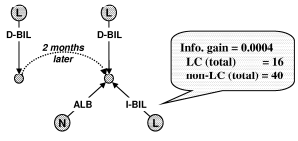


Fig. 8. Pattern 112 = Pattern 122

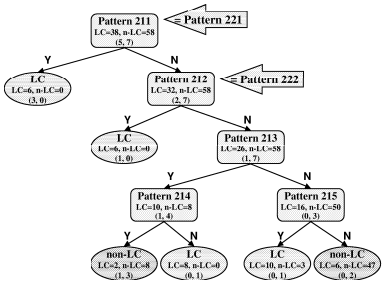


Fig. 9. One of trees from the best cycle in exp.2 (N_e=20)

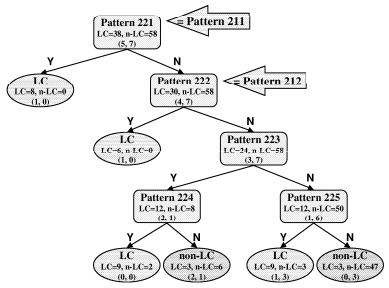


Fig. 10. One of trees from the worst cycle in exp.2 (N_e=20)

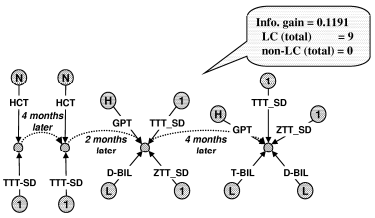


Fig. 11. Pattern 211 = Pattern 221, if exist then LC

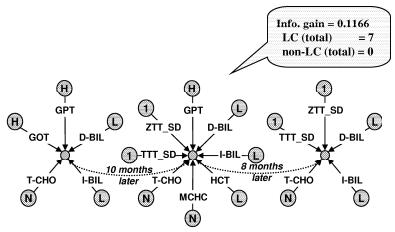


Fig. 12. Pattern 212 = Pattern 222

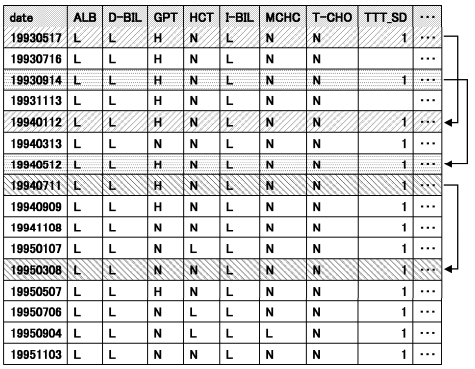
that the tree is stable. The certainty of these patterns is ensured as, for almost patients, they appear after the biopsy.

These patterns may appear only once or several times in one patient. Figure 13 shows the data of a patient for whom pattern 111 exists. As we did no attempt to estimate missing values, the pattern was not counted even if the value of only one attribute is missing. At data in the Figure 13, pattern 111 would have been counted four if the value of TTT_SD in the second line had been "1" instead of missing.

4.3 Classifying Patients with Types (B or C)

There are two types of hepatitis recorded in the dataset; B and C. We constructed decision trees which distinguish between patients of type B and type C. The attributes of antigen and antibody (HBC-AB, HBE-AB, HBE-AG, HBS-AB, HBS-AG, HCV-AB, HCV-RNA) were not included as they obviously indicate the type of hepatitis. Table 3 shows the size of graphs after the data conversion. To keep the number of instances at 2:3 ratio [8], we used all of 77 instances in type B as "Type B" class and 116 instances in type C as "Type C" class. Hence, there are 193 instances in all.

The lowest prediction error rates obtained in the first cycle (out of 10-cycles of 10 fold cross-validation) were obtained when beam width was set to 5. Thus, we set beam width to 5 when executing DT-GBI in the remaining nine cycles.



date	ALB	D-BIL	GPT	HCT	I-BIL	MCHC	T-CHO	TTT_SD	···
19930517	L	L	H	N	L	N	N	1	···
19930716	L	L	H	N	L	N	N		···
19930914	L	L	H	N	L	N	N	1	···
19931113	L	L	H	N	L	N	N		···
19940112	L	L	H	N	L	N	N	1	···
19940313	L	L	N	N	L	N	N	1	···
19940512	L	L	H	N	L	N	N	1	···
19940711	L	L	H	N	L	N	N	1	···
19940909	L	L	H	N	L	N	N	1	···
19941108	L	L	N	N	L	N	N	1	···
19950107	L	L	N	L	L	N	N	1	···
19950308	L	L	N	N	L	N	N	1	···
19950507	L	L	H	N	L	N	N	1	···
19950706	L	L	N	L	L	N	N	1	···
19950904	L	L	N	L	L	L	N	1	···
19951103	L	L	N	N	L	N	N	1	···

Fig. 13. Data of No.203 patient

Table 3. Size of graphs (classified by type)

Stage	Type B	Type C	Total
No. of graphs	77	185	262
Avg. No. of node	238	286	272
Max. No. of node	375	377	377
Min. No. of node	150	167	150

Table 4. Average error rates (%) in exp.3

	Experiment 3	
cycle	N_r=20	N_e=20
1	21.76	18.65
2	21.24	19.69
3	21.24	19.17
4	23.32	20.73
5	25.39	22.80
6	25.39	23.32
7	22.28	18.65
8	24.87	19.17
9	22.80	19.69
10	23.83	21.24
average	**23.21**	**20.31**
SD	**1.53**	**1.57**

The overall result is summarized in Table 4. The average error rate was 23.21% for 1) (N_r=20) and 20.31% for 2) (N_e=20). Figure 14 and Figure 15 show a sample of decision trees from the cycle with the lowest error rate (cycle 1) and the cycle with the highest error rate (cycle 6) respectively. Comparing the both decision trees, two patterns (shown in Figure 16 and 17) were identical and used at the upper level nodes. There patterns also appeared at almost all the decision trees and thus are considered sufficiently discriminative.

5 Conclusion

DT-GBI is a method of constructing attributes (substructures useful for classification task) on the fly while constructing a decision tree. This paper reports the preliminary results of analysing the hepatitis dataset from Chiba University Hospital by using DT-GBI. Decision trees were constructed to distinguish patients at the most severe stage of fibrosis and those at the other stages in the first two experiments, and decision trees distinguishing patients of type B and those of type C were constructed in the third experiment. We believe that the

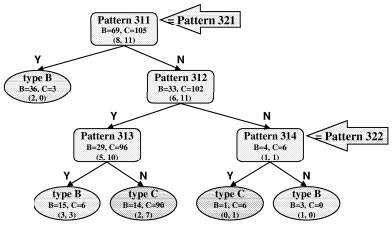


Fig. 14. One of trees from the best cycle in exp.3 (N_e=20)

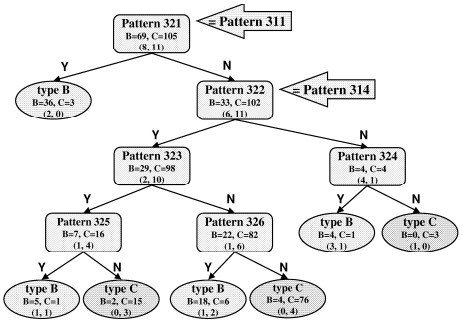


Fig. 15. One of trees from the worst cycle in exp.3 (N_e=20)

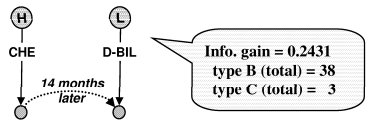


Fig. 16. Pattern 311 = Pattern 321, if exist then type B

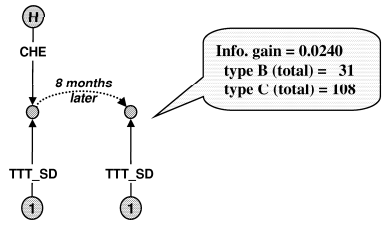


Fig. 17. Pattern 322 = Pattern 314

obtained prediction error rates are satisfactory in spite of the fact that many continuous attributes had to be discretized to keep the running time of DT-GBI within a reasonable amount.

The future work includes examining the effectiveness of DT-GBI against this hepatitis dataset with another way of preparing data, *e.g.,* randomly selecting instances from non-cirrhosis class both for training and testing in {cirrhosis vs non-cirrhosis} discrimination. Also, the validity of extracted patterns is to be evaluated and discussed by the domain experts (medical doctors).

Acknowledgment

This work was partially supported by the grant-in-aid for scientific research 1) on priority area "Realization of Active Mining in the Era of Information Flood" (No. 13131101, No. 13131206) and 2) No. 14780280 funded by the Japanese Ministry of Education, Culture, Sport, Science and Technology.

References

1. T. B. Ho, T. D. Nguyen, S. Kawasaki, S.Q. Le, D. D. Nguyen, H. Yokoi, and K. Takabayashi. Mining hepatitis data with temporal abstraction. In *Proc. of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 369–377, August 2003.

2. T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Knowledge discovery from structured data by beam-wise graph-based induction. In *Proc. of the 7th Pacific Rim International Confernce on Artificial Intelligence (Springer Verlag LNAI2417)*, pages 255–264, 2002.
3. M. Ohsaki, Y. Sato, S. Kitaguchi, and T. Yamaguchi. A rule discovery support system. In *Project "Realization of Active Mining in the Era of Information Flood" Report*, pages 147–152, March 2003.
4. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
5. J. R. Quinlan. *C4.5: Programs For Machine Learning*. Morgan Kaufmann Publishers, 1993.
6. S. Tsumoto, K. Takabayashi, M. Nagira, and S. Hirano. Trend-evaluation multiscale analysis of the hepatitis dataset. In *Project "Realization of Active Mining in the Era of Information Flood" Report*, pages 191–197, March 2003.
7. G. Warodom, T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Classifier construction by graph-based induction for graph-structured data. In *Proc. of the 7th Pacific-Asia Conference on Knowledge Discovery and Data Mining (Springer Verlag LNAI2637)*, pages 52–62, 2003.
8. Y. Yamada, E. Suzuki, H. Yokoi, and K. Takabayashi. Decision-tree induction from time-series data based on a standard-example split test. In *Proc. of the 12th International Conference on Machine Learning*, pages 840–847, August 2003.

Discovery of Temporal Relationships using Graph Structures

Ryutaro Ichise[1] and Masayuki Numao[2]

[1] Intelligent Systems Research Division
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda, Tokyo 101-8430, Japan
ichise@nii.ac.jp

[2] The Institute of Scientific and Industrial Research
Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
numao@ai.sanken.osaka-u.ac.jp

Abstract. In managing medical data, handling time-series data, which contain irregularities, presents the greatest difficulty. In the present paper, we propose a first-order rule discovery method for handling such data. The present method is an attempt to use graph structure to represent time-series data and reduce the graph using specified rules for inducing hypothesis. In order to evaluate the proposed method, we conducted experiments using real-world medical data.

1 Introduction

Hospital information systems that store medical data are very popular, especially in large hospitals. Such systems hold patient medical records, laboratory data, and other types of information, and the knowledge extracted from such medical data can assist physicians in formulating treatment strategies. However, the volume of data is too large to allow efficient manual extraction of data. Therefore, physicians must rely on computers to extract relevant knowledge.

Medical data has three notable features [14]; namely, the number of records increases each time a patient visits a hospital; values are often missing, usually because patients do not always undergo all examinations; and the data include time-series attributes with irregular time intervals. To handle medical data, a mining system must have functions that accommodate these features. Methods for mining data include K-NN, decision trees, neural nets, association rules, and genetic algorithms [1]. However, these methods are unsuitable for medical data, in view of the inclusion of multiple relationships and time relationships with irregular intervals.

Inductive Logic Programming (ILP) [4] is an effective method for handling multiple relationships, because it uses horn clauses that constitute a subset of first order logic. However, ILP is difficult to apply to data of large volume, in view of computational cost. We propose a new graph-based algorithm for inducing

Table 1. Example medical data.

ID	Examination Date	GOT	GPT	WBC	RNP	SM
14872	19831212	30	18			
14872	19840123	30	16			
14872	19840319	27	17	4.9		
14872	19840417	29	19	18.1		
14872	...					
5482128	19960516	18	11	9.1	-	-
5482128	19960703	25	23	9.6		
5482779	19980526	52	59	3.6	4	-
5482779	19980811			4		
5482779	...					

horn clauses for representing temporal relations from data in the manner of ILP systems. The method can reduce computational cost of exploring in hypothesis space. We apply this system to a medical data mining task and demonstrate the performance in identifying temporal knowledge in the data.

This paper is organized as follows. Section 2 characterizes the medical data with some examples. Section 3 describes related work in time-series data and medical data. Section 4 presents new temporal relationship mining algorithms and mechanisms. Section 5 applies the algorithms to real-world medical data to demonstrate our algorithm's performance, and Section 6 discusses our experimental result and methods. Finally, in Section 7 we present our conclusions.

2 Medical Data

As described above, the sample medical data shown here have three notable features. Table 1 shows an example laboratory examination data set including seven attributes. The first attribute, ID, means personal identification. The second is Examination Date, which is the date the patient consults a physician. The remaining attributes designate results of laboratory tests.

The first feature shows that the data contain a large number of records. The volume of data in this table increases quickly, because new records having numerous attributes are added every time a patient undergoes an examination.

The second feature is that many values are missing from the data. Table 1 shows that many values are absent from the attributes that indicate the results of laboratory examinations. Since this table is an extract from medical data, the number of missing values is quite low. However, in the actual data set this number is far higher. That is, most of the data are missing values, because each patient undergoes only some tests during the course of one examination. In addition, Table 1 does not contain data when laboratory tests have not been conducted. This means that the data during the period 1983/12/13 to 1984/01/22 for patient ID 14872 can also be considered missing values.

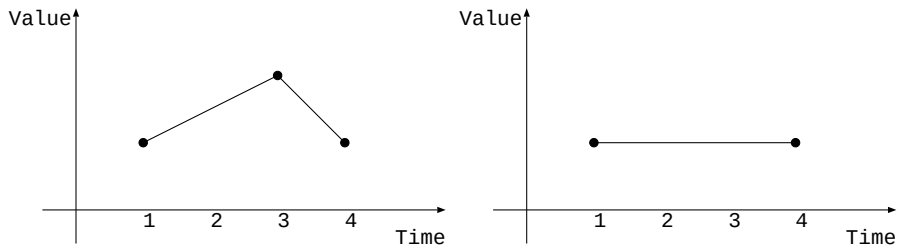


Fig. 1. Problem of graph similarity.

The other notable feature of the medical data is that it contains time-series attributes. When a table does not have these attributes, then the data contain only a relationship between ID and examination results. Under these circumstances, the data can be subjected to decision tree learning or any other propositional learning method. However, relationships between examination test dates are also included; that is, multiple relationships.

3 Related Work

These kinds of data can be handled by any of numerous approaches. We summarize related work for treating such data from two points of view: time-series data and medical data.

3.1 Time-Series Data

One approach is to treat the data described in Section 2 as time-series data. When we plot each data point, we can obtain a graph similar to stock market chart, and can apply a mining method to such data. Mining methods include the window sliding approach [3] and dynamic time warping [7]. Those methods can identify similar graphs. However, the methods assume that the time-series data are collected continuously. This assumption is not valid for medical data, because of the missing values. Let us illustrate the problem by way of example. When we look at the plots of two patient data sets in Figure 1, the graph shapes are not the same in. Therefore, those methods do not consider the two graphs to be similar graphs. However, if we consider that the second data set to have a missing value at time 3, these two graphs can be considered to be the same. Hence, this type of method is not robust for missing values and is not directly applicable to the medical data described in Section 2.

3.2 Medical Data

Many systems for finding useful knowledge from medical data have been developed [8]. However, not many systems for treating temporal medical data have

been developed. Active mining projects [9] progressing in Japan are now being developed in order to obtain knowledge from such data. The temporal description is usually converted into attribute features by some special function or dynamic time warping method. Subsequently, the attribute feature in a standard machine learning method such as decision tree [15] or clustering [2] is used. Since these methods do not treat the data directly, the obtained data can be biased by summarization of the temporal data.

Another approach incorporates InfoZoom [13], which is a tool for visualization of medical data in which the temporal data are shown to a physician, and the physician tries to find knowledge from medical data interactively. This tool lends useful support for the physician, but does not induce knowledge by itself.

4 Temporal Relationship Mining

4.1 Approach

An important consideration for obtaining knowledge from medical data is to have a knowledge representation scheme that can handle the features described in Section 2 One such scheme is Inductive Logic Programming (ILP)[4], because it uses horn clauses, which can represent such complicated and multi-relationship data [6]. Since ILP framework is based on the proof of logic, existing values are processed and missing values are ignored for inducing knowledge. Therefore, ILP constitutes one solution for the second problem inherent to medical data described in Section 2. Horn clause representation permits multiple relations, such as time-series relation and attributes relations. It can also be a solution to the third problem inherent to medical data. However, the ILP system does not provide a good solution to the first problem, because its computational cost is much higher than that of other machine learning methods. In this section, we propose a new algorithm for solving the problem by using graphs.

4.2 Temporal Predicate

Data mining of medical data requires a temporal predicate, which can represent irregular intervals for the treatment of temporal knowledge. We employ a predicate similar to one proposed by Rodríguez et al. [12]. The predicate has five arguments and is represented as follows:

$$blood\_test(ID, Test, Value, BeginningDate, EndingDate)$$

The arguments denote the patient ID, kind of laboratory test, value of the test, beginning date of the period being considered, and ending date of the period being considered, respectively. This predicate returns true if all tests conducted within the period have a designated value. For example, the following predicate

Table 2. The external loop algorithm.

E^+ is a set of positive examples, R is a set of discovered rules.
1. If $E^+ = \phi$, return R
2. Construct clause H by using the internal loop algorithm
3. Let $R = R \cup H$
4. Goto 1

Table 3. The internal loop algorithm.

H is a hypothesis.
1. Generate H, which contains only head
2. Use refinement operator to generate literal candidate
3. Select the best literal L according to MDL criteria
4. Add L as a body of literal H
5. If H qualified criteria, return H, otherwise goto 2

is true if patient ID 618 had at least one GOT test from Oct. 10th 1982 to Nov. 5th 1983, and all tests during this period yield very high values.

$$blood\_test(618, got, veryhigh, 19821010, 19831105)$$

This predicate is a good example for representing temporal knowledge in medical data, because it can represent the predisposition within a certain period, regardless of test intervals. Moreover, it can handle missing values without affecting the truth value. This naturally implies that our approach is a good solution for two of the problems inherent to medical data (e.g., multiple relationships and time relationships with irregular intervals) described in Section 2

4.3 Rule Induction Algorithm

In this paper, we utilize a top-down ILP algorithm similar to FOIL[11]. We can divide this algorithm into two parts. One part is an external loop for covering algorithm[10]. This algorithm is used for deleting from a positive example set examples that are covered by a generated hypothesis, and is shown in Table 2. The second part of the algorithm is an internal loop for generating a hypothesis. The algorithm is shown in Table 3. Initially, the algorithm creates the most general hypothesis. Subsequently, it generates literal candidates by using a refinement operator discussed in the following section. Next, the algorithm chooses the most promising literal according to MDL criteria, and adds it to the body of the hypothesis. If the MDL cannot be increased by adding a literal, the algorithm returns the hypothesis.

Table 4. Example data.

Id	Attribute	Value	Date
23	got	vh	80
31	got	vh	72
31	got	vh	84
35	got	vh	74

4.4 Refinement

In our method, the search space for the hypothesis is constructed by combinations of predicates described in Section 4.2. Suppose that the number of the kinds of tests is N_a, the number of test domains is N_v, and the number of date possibilities is N_d. Then, the number of candidate literals is $N_a \times N_v \times N_d^2/2$. As we described in Section 2, because medical data consist of a great number of records, the computational cost for handling medical data is also great. However, medical data have many missing values and consequently, often consist of sparse data. When we make use of this fact, we can reduce the search space and computational cost.

To create candidate literals which are used for refinement, we propose employing graphs created from temporal medical data. The purpose of this literal creation is to find literals which cover many positive examples. In order to find them, a graph is created from positive examples. The nodes in the graph are defined by each medical data record and the node has four labels; i.e. patient ID, laboratory test name, laboratory test value, and date test conducted. Arcs are created for each node. Suppose that two nodes represented by $n(Id_0, Att_0, Val_0, Dat_0)$ and $n(Id_1, Att_1, Val_1, Dat_1)$ exist. The arc is created if all the following conditions hold:

- $Id_0 \neq Id_1$
- $Att_0 = Att_1$
- $Val_0 = Val_1$
- For all $D\{D \geq Dat_0 \wedge D \leq Dat_1\}$,
  if $n(Id_0, Att_0, Val, D)$ exists, $Val = Val_0$
  and
  if $n(Id_1, Att_1, Val, D)$ exists, $Val = Val_1$

For example, supposing that we have data shown in Table 4, we can obtain the graph shown in Figure 2.

After constructing the graph, the arcs are deleted by the following reduction rules:

- The arc $n_0 - n_1$ is deleted if a node n_2 which is connected to both n_0 and n_1 exists, and
  - Dat_2 for n_2 is greater than both Dat_0 and Dat_1 or
  - Dat_2 for n_2 is smaller than both Dat_0 and Dat_1

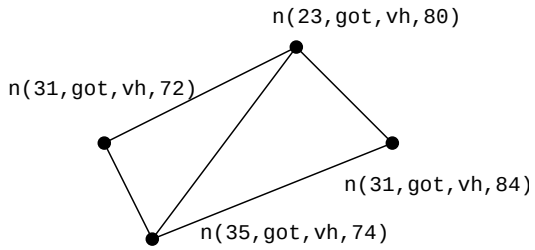


Fig. 2. Example graph.

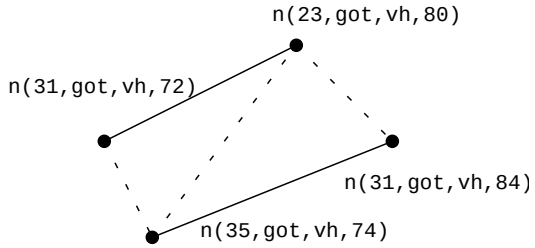


Fig. 3. Graph after deletion.

After deleting all arcs for which the above conditions hold, we can obtain the maximum period which contains positive examples. Then we pick up the remaining arcs and set the node date as *BeginingDate* and *EndingDate*. After applying the deletion rules for the graph shown in Figure 2, we obtain the graph shown in Figure 3. Then the final literal candidate for refinement is $blood\_test(Id, got, veryhigh, 72, 80)$ and $blood\_test(Id, got, veryhigh, 74, 84)$, which in this case covers all three patients.

5 Experiment

5.1 Experimental Settings

In order to evaluate the proposed algorithm, we conducted experiments on real medical data donated from Chiba University Hospital. These medical data contains data of hepatitis patients, and the physician requires us to find an effective timing for starting interferon therapy. Interferon is a kind of medicine for reducing the hepatitis virus. It has great effect for some patients; however, some patients exhibit no effect, and some patients exhibit deteriorated condition. Further, the medicine is expensive and has side effects. Therefore, physicians wants to know the effectiveness of Interferon before starting the therapy. According to our consulting physician, the effectiveness of the therapy could be changed by the patient's condition and could be affected by the timing for starting it.

We input the data of patients whose response is complete as positive examples, and the data of the remaining patients as negative examples. Complete response is judged by virus tests and under advice from our consulting physician. The number of positive and negative examples are 57 and 86, respectively. GOT, GPT, TTT, ZTT, T-BIL, ALB, CHE, TP, T-CHO, WBC, and PLT, which are attributes of the blood test, were used in this experiment. Each attribute value was discretized by the criteria suggested by the physician. We treat the starting date for interferon therapy as base date, in order to align data from different patients. According to the physician, small changes in blood test results can be ignored. Therefore, we consider the predicate blood_test to be true if the percentage p, which is set by parameter, of the blood tests in the time period show the specified value.

5.2 Result

Since not all results can be explained, because of space limitations, we introduce only three of the rules obtained by our system.

```
inf_effect(Id):-
        blood_test(Id,wbc,low,149,210). (1)
```

This rule is obtained when the percentage parameter p is set at 1.0. Among the 143 patients, 13 satisfied the antecedent of this rule, and interferon therapy was effective in 11 of these. The rule held for 19.3 percent of the effective patients and had 84.6 percent accuracy. The blood test data for effective patients are shown in Figure 4. The number of the graph line represents patient ID, and the title of the graph represents test-name/period/value[low:high]/parameter p, respectively.

```
inf_effect(Id):-
        blood_test(Id,tp,high,105,219). (2)
```

This rule is obtained when the percentage parameter p is set at 1.0. Among the 143 patients, 7 satisfied the antecedent of this rule, and interferon therapy was effective for 6 of these. The rule held for 10.5 percent of the effective patients and had 85.7 percent accuracy. The blood test data for effective patients are shown in Figure 5.

```
inf_effect(Id):-
        blood_test(Id,wbc,normal,85,133),
        blood_test(Id,tbil,veryhigh,43,98). (3)
```

This rule is obtained when the percentage parameter p is set at 0.8. Among the 143 patients, 5 satisfied the antecedent of this rule, and interferon therapy was effective in all of these. The rule held for 8.8

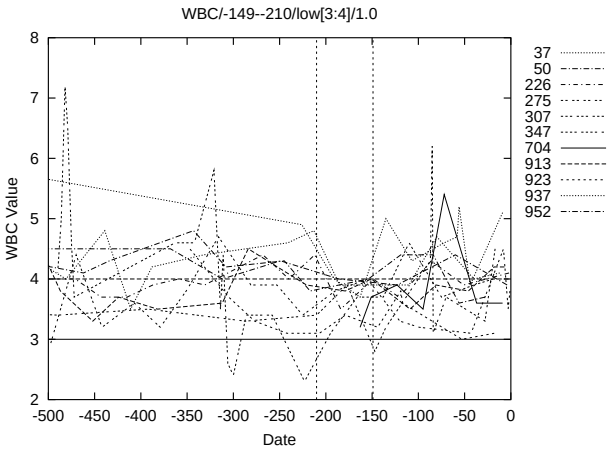


Fig. 4. Blood test graph for rule (1).

percent of the effective patients and had 100 percent accuracy. The blood test data for effective patients are shown in Figure 6. The patients who satisfied both test in rule (3) are positive patients. This means that we have to view both graphs in Figure 6 simultaneously.

6 Discussion

The results of our experiment demonstrate that our method successfully induces rules with temporal relationships in positive examples. For example, in Figure 4, the value range of WBC for the patients is wide except for the period between 149 and 210, but during that period, patients exhibiting interferon effect have the same value. This implies that our system can discover temporal knowledge within the positive examples.

We showed these results to our consulting physician. He stated that if the rule specified about half a year before therapy starting day, causality between the phenomena and the result would be hard to imagine. It was necessary to find a connection between them during the period. In relation to the third rule, he also commented that the hypothesis implies that temporary deterioration in a patient's condition would indicate the desirability to start interferon therapy with complete response.

The current system utilizes a cover set algorithm to induce knowledge. This method starts from finding the largest group in positive examples, then progresses to find smaller groups. According to our consulting physician, the patients could be divided into groups, even within the interferon effective patients. One method for identifying such groups is the subgroups discovery method [5].

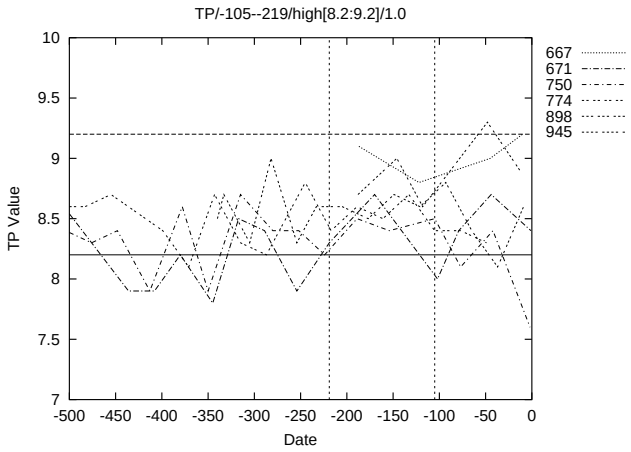


Fig. 5. Blood test graph for rule (2).

When it is used in place of the cover set algorithm, this method could assist the physician.

In its present form, our method uses only the predicate defined in Section 4.2. When we use this predicate with the same rule and different time periods, we can represent movement of blood test values. However, this induction is somewhat difficult for our system, because each literal is treated in each refinement step separately. This is a current limitation for representing the movement of blood tests. Rodrígues et al. [12] also propose other types of temporal literals. As we mentioned previously, the hypothesis space constructed by the temporal literal requires a high computational cost for searching, and only a limited hypothesis space is explored. In this paper, we propose inducing the literals efficiently by using graph representation of hypothesis space. We believe that we can extend this approach to other types of temporal literals.

7 Conclusion

In this paper, we propose a new data mining algorithm. The performance of the algorithm was tested experimentally by use of real-world medical data. The experimental results show that this algorithm can induce knowledge about temporal relationships from medical data. The temporal knowledge is hard to obtain by existing methods, such as a decision tree. Furthermore, physicians have shown interest in the rules induced by our algorithm.

Although our results are encouraging, several areas of research remain to be explored. As shown in Section 5.2, our system induces hypothesis regardless of the causality. We must bias the induction date period to suit the knowledge of

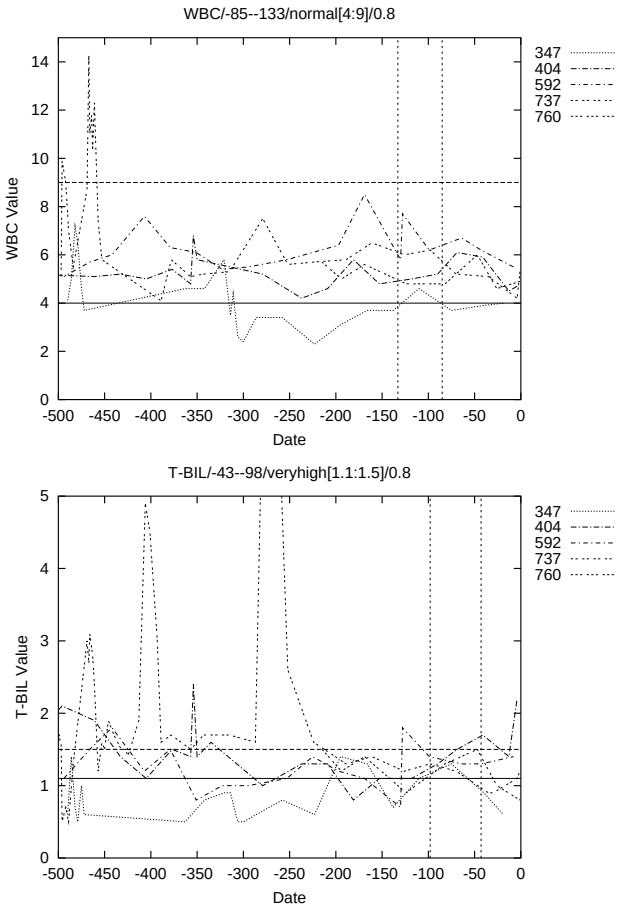


Fig. 6. Blood test graph for rule (3).

our consulting physicians. In addition, our algorithm must be subjected to experiments with different settings. We plan to apply this algorithm to other domains of medical data and also apply it to non-medical, temporal data. Extensions to treating numerical values also must be investigated. Our current method require attributes in discrete values. We plan to investigate these points in our future work.

Acknowledgments

We are grateful to Hideto Yokoi for fruitful discussions.

References

1. Adriaans, P., & Zantinge, D. (1996). *Data Mining*. London: Addison Wesley.
2. Baxter, R., Williams, G., & He, H. (2001). Feature Selection for Temporal Health Records. *Lecture Notes in Computer Science*. 2035, 198–209.
3. Das, D., Lin, K., Mannila, H., Renganathan, G. & Smyth, P. (1998). Rule Discovery from Time Series. *In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 16–22).
4. Džeroski, S. & Lavrač, N. (2001). *Relational Data Mining*. Berlin: Springer.
5. Gamberger, D., Lavrač, N., & Krstačić, G. (2003). Active subgroup mining: a case study in coronary heart disease risk group detection, *Artificial Intelligence in Medicine*, 28, 27–57.
6. Ichise, R., & Numao, M. (2001). Learning first-order rules to handle medical data. *NII Journal*, 2, 9–14.
7. Keogh, E., & Pazzani, M. (2000). Scaling up Dynamic Time Warping for Datamining Applications, *In the Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining* (pp. 285–289)
8. Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective, *Artificial Intelligence in Medicine*, 23, 89–109.
9. Motoda, H. editor. (2002) Active mining: new directions of data mining. In: Frontiers in artificial intelligence and applications, 79. IOS Press.
10. Muggleton, S., & Firth, J. (2001). Relational rule induction with CProgol4.4: a tutorial introduction, *Relational Data Mining* (pp. 160–188).
11. Quinlan, J. R. (1990). Learning logical definitions from relation. *Machine Learning*, 5, 3, 239–266.
12. Rodríguez, J. J., Alonso, C. J., & Boström, H. (2000). Learning First Order Logic Time Series Classifiers: Rules and Boosting. *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases* (pp. 299–308).
13. Spenke, M. (2001). Visualization and interactive analysis of blood parameters with InfoZoom. *Artificial Intelligence in Medicine*, 22, 159–172.
14. Tsumoto, S. (1999). Rule Discovery in Large Time-Series Medical Databases. *Proceedings of Principles of Data Mining and Knowledge Discovery: Third European Conference* (pp. 23–31).
15. Yamada, Y., Suzuki, E., Yokoi, H., & Takabayashi, K. (2003) Classification by Time-series Decision Tree, *Proceedings of the 17th Annual Conference of the Japanese Society for Artificial Intelligence*, in Japanese, 1F5-06.

A Scenario Development on Hepatitis B and C

Yukio Ohawa[1] Naoaki Okazaki[2] Naohiro Matsumura[2] Akio Saiura [3], and Hajime Fujie [4]

[1] Graduate School of Business Sciences, University of Tsukuba: 112-0012, Japan
osawa@gssm.otsuka.tsukuba.ac.jp
[2] Faculty of Engineering, The University of Tokyo 113-8656, Japan
{matumura,okazaki}@kc.t.u-tokyo.ac.jp
[3] Cancer Institute Hospital, Tokyo, 170-8455
[4] Department of Gastroenterology, The University of Tokyo Hospital, 113-8655

Abstract. Obtaining scenarios has been a major approach for decisions in domains where a sequence of events and actions in the future is significant. Chance discovery, in which events with significance for making a decision, can be regarded as the emergence of a scenario, with extracting chances i.e. essential events in the turning points of valuable scenarios by means of the interaction with the environment. In this paper, we apply a method of chance discovery to the data of diagnosis of hepatitis patients, for obtaining scenarios of how the most essential symptoms appear in the patients of hepatitis of type B and C. In the process of discovery, the results are evaluated to be novel and potentially useful, under the mixture of objective facts and the subjective widening and narrowing of the surgeon's concerns.

1. Introduction: Scenarios in the Basis of Critical Decisions

Scenarios have been a significant basis of decisions, in domains where a sequence of events in the future becomes significant. Let us stand on the position of a surgeon, for example, looking at the time series of symptoms during the progress of an individual patient's disease. The surgeon should make an appropriate action for curing this patient, at an appropriate time. If he does the patient's disease may be cured efficiently, but otherwise the patient's health condition might turn radically into a worse state. The situation of this surgeon can be described as a choice from two sequences, for example,

Sequence 1) state 1 -> state 2 -> state 3 -> state 4.
Sequence 2) state 1 -> state 2 -> state 5. **(1)**

Suppose that state 4 and 5 mean two opposite situations, i.e., one where the disease is cured and a fatal situation. The surgeon should choose a effective action at the time state 2 appears, for this patient to shift to state 4. This kind of state, which is essential for decision has been called a chance [1].

On the other hand, the event-sequence in (1) has been called a *scenario* in cases where considering a sequence is essential for decision. All in all, the discovery of a chance is quite relevant to obtaining valuable scenarios. Thus scenarios are tools for chance discovery, and also the purpose of chance discovery. That is, detecting the branching events between multiple scenarios between the two scenarios, as state 2 in (1), means the chance discovery for the surgeon. This chance is regarded as valuable,

only if the purpose is achievable, i.e. if the scenario including state 2 is valuable, because a scenario is easier to understand than an event shown alone. Suppose you are the patient and told just that you have a polyp in the stomach, it would be hard to decide to cut it or to do nothing to leave it in the current position. On the other hand, suppose the doctor tells that you are in the branch of two scenarios – in one, it will turn larger and worse. In the other, the polyp is cut away, you will be cured. Normally, you will prefer the latter choice.

2. Scenario "Emergence" in the Mind of Experts

In the term "scenario development", a scenario sounds like something to be "developed" by human(s) who consciously rules the process of making a scenario. However, scenario really "emerges" by partially unconscious interaction of human(s) and the environment. For example, a *scenario workshop* starts from scenarios preset by writers, then experts in the corresponding domain discusses to improve the scenarios [2]. It is usual that the discussants write down their opinions during the workshop, but rarely they notice why those opinions came out and the why the workshop selected the scenarios obtained finally. In the very origin of aiding creation, the KJ method begins from cards on which the initial ideas are written and arranged in the 2D-space by co-working discussants. The new combination of proposed scenarios may help the emergence of a new valuable scenario. In the design process, ambiguous information can trigger creations [3].

The common points among "experts" in scenario workshops, "combination" of ideas in KJ method, and the "ambiguity" in the information to a designer is that multiple scenarios in the interaction of subjects with their own environments are bridged via the links between the contexts in the mental world they attend. From these bridges, they unconsciously introduce some situations or events which may work as "chances" to import others' scenarios. In the example of (1), a surgeon who almost gave up because he imagined scenario 2, may obtain a new hope in scenario 1 proposed by his colleague, by noticing that state 2 is shared by the two scenarios.

In this paper, we show a method of aiding scenario emergence, by means of visual interaction with real data using two tools KeyGraph and Text Drop. KeyGraph, with an additional function to show causal directions in the relations between events (let us call this *scenario map*), visualizes the complex relations among values of variables in diagnosis data of hepatitis, and Text Drop helps in extracting the part of data corresponding to the interest of an expert, a surgeon here.

These tools help in picking essential scenarios of specific types of patience from the complex diagram of their mixture, i.e. KeyGraph These results are evaluated by the surgeon as useful in the decision of curing hepatitis B and C. This evaluation is subjective in the sense that too small number of patients were observed to follow the entire scenarios obtained, to evaluate the scenarios quantitatively. Further more, we should say the evaluation is made in the process of discovery, merging the subjective interest of the expert and the objective within the process of chance discovery. Rather than calling data-mining, this is a self-mining of the subject, where the quality of the self's experience affects much on the result.

3. Mining the Scenarios of Hepatitis Cases

3.1 The Double Helical Process of Chance Discovery

In the most recent state of art, the process of chance discovery is supposed to follow the Double Helix (DH) model [1].

DH starts from a state of mind concerned with winning a new chance, and this *concern* (ambiguous interest) is reflected to acquiring data to be analyzed by a data-mining tool, specifically designed for chance discovery, for making a new decision. Looking at the result of this analysis, possible scenarios and their values may become clarified in the user' mind. If multiple users co-works sharing the same data-mining result, the effect of scenario emergence might help in mining valuable scenarios by bridging the various experience of participants to form a novel scenario. Based on the chances discovered here, the user(s) make actions or simulate actions in a virtual (imagined /computed) environment, and obtains renewed concerns with chances – the helical process returns to the initial step. DH is embodied in this paper, in the application to hepatitis diagnosis. The user sees and manipulates KeyGraph, thinking and talking about the scenarios the diagram may imply. Here, "manipulate" means to cut, move, and unify nodes/links in KeyGraph - it enforces the bridges between multiple experiences to be combined in scenario emergence. In other words, manipulation urges user to ask "why is this node here?" and to virtually experience alternative scenarios s/he did not think of.

3.2 KeyGraph and Text Drop for Accelerating DH

In the case of marketing for textile products, Nittobo Inc. made a success in selling a new product with adding a new value represented by a scenario in the life of people who may buy the product. They visualized the map of the market by means of KeyGraph [1,4], shown as a diagram of co-occurrences between products in the basket data of buyers of textiles. In this map, their marketing researchers found a valuable new scenario in the life people who may go buying textiles across a wide range in the market. They successfully found essential new products in valuable scenarios they found. This realized a sales hit of the new textile.

However, it was not efficient to follow the process of DH, using KeyGraph solely. A critical defect of this method was that user could not extract the interesting part of the data easily, when s/he has got a new concern with chances. For example, they may become interested in a customer who buys product A or product B, who also buys product C, but does never touch product D. Then, the user desires to look *easily* into such customers deeply to take advantage of chances in the submarket formed by such kind of customers they came to be interested in. Text Drop is a simple tool for Boolean-selection of the part of data corresponding to users' interest which can be described in a Boolean formula, e.g.

$$boolean = \text{"(product A | product B) \& product C \& !product D"} \quad (2)$$

Then Text Drop obtains a new data, made of baskets including product A or product B, and product C, but not including product D. Its simple interface is useful in the case where the user can express his/her own interest in Boolean formula as in (2). The interest of user might be more ambiguous, especially in the beginning of the process of chance discovery. In such a case, the user is supposed to enter the formula "as much as precisely" reflecting one's own interest. Having KeyGraph, Text Drop, and the freedom to use these on necessity, the user can follow the procedure below to realize a DH process.

[DH Process supported by KeyGraph and Text Drop]

1) Obtain a data of baskets, reflecting user's interest
2) Apply KeyGraph to the data to visualize the map representing the relations, or the causal order of occurrences if possible, among items in the baskets.
3) Manipulate KeyGraph as follows:
   3-1) Move nodes and links to the positions in the 2D output of KeyGraph, or remove nodes and links which are apparently meaningless in the target domain.
   3-2) Write down scenarios, imaginable on KeyGraph
4) Read or visualize the comments of experts in 3-2), and become aware of interesting items in the data for user him/herself.
5) Enter interesting items or their combination in Boolean formula, into Text Drop. The data of baskets, reflecting user's new interest is obtained. Return to Step 1).

4. Results for the Diagnosis Data of Hepatitis

4.1 The Hepatitis Data

The following shows the style of data in the case of the diagnosis of hepatitis. Each item represents the pair, of a variable and its observed value. That is, an item put as "a_b" means a piece of data where the value of variable a is b. For example, T-CHO_high (T-CHO_low) means T-CHO (total cholesterol) is higher (lower) than a predetermined threshold. Each line in the data represents the sequence of diagnosis results for one patient. See description (3).

Patient 1) item1, item2, …..., item m1.
Patient 2) item 2, item 3, …., item m2. (**3**)
Patient 3) item 1, item 5, …, item m3.

As in (3), we can regard one patient as a unit of co-occurrence of items. That is, there are various cases of patients and the sequence of one patient's diagnosis items means his/her scenario of wandering in the map of the various symptoms. By applying KeyGraph to the data in (3), we can obtain the following components:

- *Islands of items*: A group of items co-occurring frequently, i.e. occurred to many same patients or many same lines in (3). The doctor can be expected to know what kind of patient each island corresponds to.
- *Bridges across islands*: A patient may switch from one island to another, in the progress of the disease or its cure.

Figure 2 is the KeyGraph obtained first, for cases of hepatitis B. The arrows, which does not appear in the original KeyGraph, depict approximate causations, i.e., X -> Y means that if event X appears in the scenario, then the patient tended to experience Y also. That is, each line in (3) is a set of observations of a certain patient from his/her certain situation in the disease progress or cure. On this specific feature of diagnosis data, the relative strength of the statement "if event X appears, then event Y follows" in comparison with its inverse, say "if event Y appears, then event X follows" means X is likely to be tied to the cause of Y. Thus, even if there are relations where the order of causality and time are opposite (i.e. if X is the cause of Y but was observed after Y due to the delay of observation or the setting of strict threshold for being detected), we can express approximate scenarios by giving an arrow from X to Y, just with comparing the two results of KeyGraph, one for data including X and the other for one including Y. That is, if the former includes more causal events than the latter

and if the latter includes more consequent events than the former, X is expected to proceed Y in the scenario. In a case the order of causality and time are opposite, we may interpret that Y appeared before X only because the threshold of Y was set easy to exceed, e.g., the upper threshold of ZTT may be set low and easy to exceed than that of G_GL, which makes ZTT appear before G_GL even if ZTT is a result of G_GL. Let us call a KeyGraph with these arrows a *scenario map*.

4.2 Results for Hepatitis B

An initial KeyGraph was shown to a surgeon (see acknowledgement) in Step 2), and was manipulated to form a scenario map in step 3). In the manipulation, the surgeon grouped the nodes in the circles in the graph, got rid of unessential nodes from the figure, and unified redundant nodes as "jaundice" and "T-BIL_high" (high total bilirubin), and the necessary process for making a scenario map in 4.1. Figure 1 was the result of this manipulation. Simultaneously, we wrote down what the doctor has been teaching us about hepatitis looking at the KeyGraph, and we applied KeyGraph to the memo. According to its result, two of the most significant terms were "mixture" and factors relevant to jaundice. An important lesson here was that KeyGraph depicted a mixture of various scenarios. Some of the scenarios were common sense for the surgeon, about the progress of hepatitis B, e.g.,

(Scenario B1) Transition from CPH (chronic persistent hepatitis) to CAH (chronic active hepatitis).

(Scenario B2) Decrease in blood platelets (PT) and hemoglobin (HDB), leading to jaundice i.e., increase in the T-BIL. Considering D-BIL increases more keenly than I-BIL, this is from the activation of lien, due to the critical illness of liver.

(Scenario B3) Biliary blocks accelerate jaundices

Although scenarios for the cure of hepatitis B were not observed apparently, we could see that a quick sub-process from LDH_high to LDH_low (LDH: lactate dehydrogenase) is a significant bridge from a light hepatitis to a critical state of liver as the high value of T-BIL.

According to the surgeon, a sudden change in the value of LDH is sometimes observed in the introductory steps of ferminate hepatitis in the real treatment, but the quick change has been regarded no more than an ambiguous information for treatment. However, the result of KeyGraph for the sub-data extracted from various aspects showed LDH plays a significant role in the scenario of progress of hepatitis, e.g., Fig.2. Figure 2 shows the scenario map, for cases including "IG-M_high", considering that IG-M (immunoglobulin M) is activated in the beginning of infection with virus. For the reason that the data of symptoms is taken from a certain time to the termination of the individual patient's data, the data including IG-M_high was regarded as a summary of the overall scenario of infectious disease.

Note that Fig.2 is still a simple summary of mixed scenarios, just showing 30 nodes where 50 to 60 is the typical range. Having obtained the new concern, i.e. what the scenario can be like if it includes the change in the value of LDH, i.e. decrease shortly after increase, we obtained the result in Figure 3 for data of hepatitis B including "LDH_high" and "LDH_low". This figure shows the change in the value of LDH triggers a shift from the initial state stared by biliary-relevant enzymes as G-GTP and ALP, to a critical state in the large circle where T-BIL, I-BIL, and D-BIL are high and CHE (choline esterase) decreases. According to the surgeon, he has been tacitly aware

of this position of the change in LDH, in the real experiences. This was a useful, but not published piece of knowledge for detecting a sign of critical changes in the liver.

4.3 Results for Hepatitis C

In cases of hepatitis C, as in Fig.4, we find a mixture of a number of scenarios, e.g.,

(Scenario C1) Transition from CPH to CAH,
(Scenario C2) Transition to critical states, e.g. cancer, jaundice.

These common-sense scenarios are quite similar to the scenarios in the cases of hepatitis B, but we also find "interferon" and an ambiguous region in the top (in the dotted circle) of figure 4. That is, GOT and GPT can be low both after the fatal progress of heavy hepatitis and if the disease is cured. The latter case is rare because GOT and GPT are expected to normally take "normal" value, i.e., between the lower and the upper threshold, rather than being "low" i.e. lower than the lower threshold.

Thus we saw the results of renewed scenario map for cases including both GOT_low and GPT_low. We still find a complex mixture of scenarios, and find some events looking like a better state in the region without arrows in figure 5.

Fig.5 seems to be separated roughly into good and bad liver states. We assumed this represents the shift from a bad liver to a mixture of good and bad states due to the treatment by interferon. This suggested that the data with "GOT_low & GPT_low & !interferon" (i.e. GOT and GOT both became low at least once, and interferon has never been used) may separate the two areas, one of critical scenario and the other not so severe. In the result of Fig.6, we find two clusters:

(Upper cluster) The scenario of fat liver, to be cured, *not requiring interferon*. This cluster does not mean to turn to a critical state. The item F-B_G1_high and F-A1_Gl_high are the turning points from bad to a better state.
(Lower cluster) Critical scenario *beyond the effect of interferon*, i.e., of high bilirubins. F-A1_Gl_low and F-A1_Gl_gl_low are on the switch to critical states.

Here we can suppose that globulins as F-A1_Gl, F-A12_Gl, and F-B_GL are relevant to the role of interferon, in curing hepatitis C, and this hypothesis also matches with Fig. 4 where interferon is linked in the branch of F-A1_Gl _high/low and F-A1_Gl _high/low. Finally, under this new concern with globulins, we obtained the scenario map as in Fig.7, showing that a case treated with interferon and had a switch from F-A1_GL_high to F-A1_GL_low had a significant turning point to be cured on the recovery of TP (total protein) quantity, matching with the result in Fig.5 where TP_low, TP_high, and globulins are in the intersection of critical and easier scenarios. The decrease in GOT and GPT, and then in ZTT followed this, matching with the results in [5]. HCV (virus of hepatitis C) also decreased.

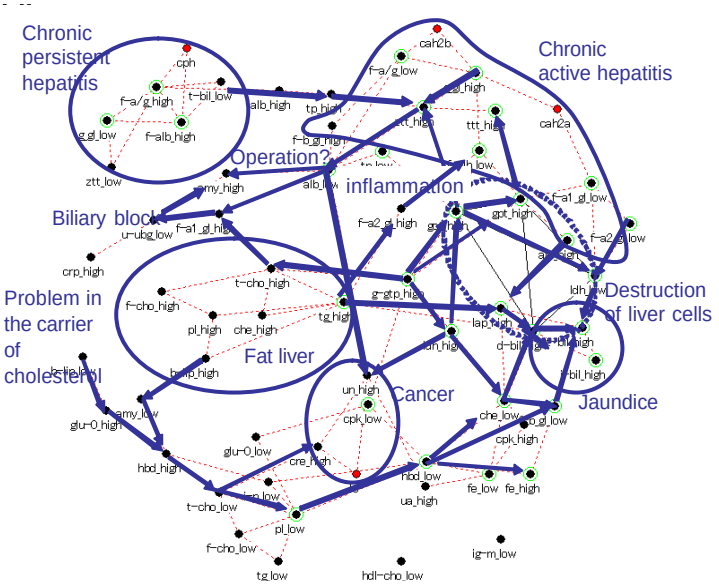


Fig. 1. The scenario map, after manipulation on Fig.2. The surgeon acquired concerns with biliary relevant enzymes, G_GTP, LDH etc., especially LDH he has been feeling important in his own experiences. *Note: in all figures hereafter, the arrows were drawn reflecting the causations as in 4.1, and round frames and their interpretations were drawn manually.*

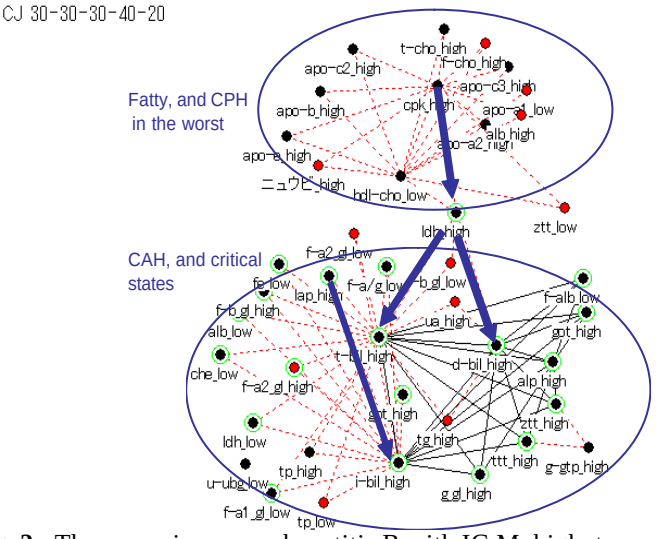


Fig. 2. The scenario map or hepatitis B with IG-M_high, to see scenarios in the early of disease where biliary enzymes as LDH are affected. A temporary increase in LDH is found to be a bridge from a persistent to an active states. This deepened the concern with LDH, and embodied the concern to an interest in entering "LDH_high & LDH_low" to Text Drop.

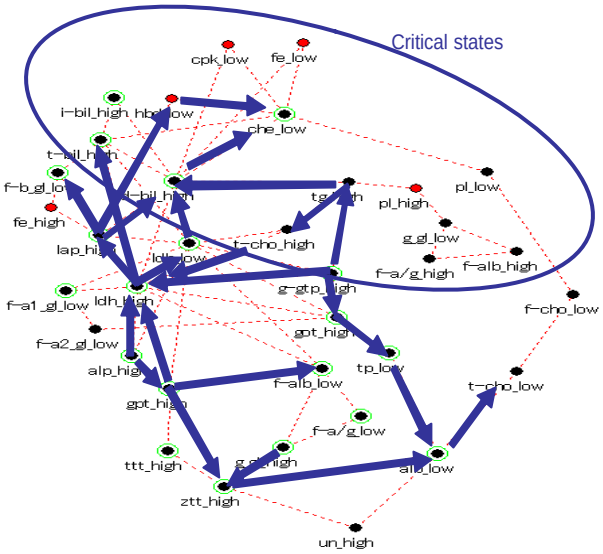


Fig. 3. The scenario map for hepatitis B, with "LDH_high & LDH_low." This "tacitly matches with experiences and potentially useful, although not published ever" according to the surgeon.

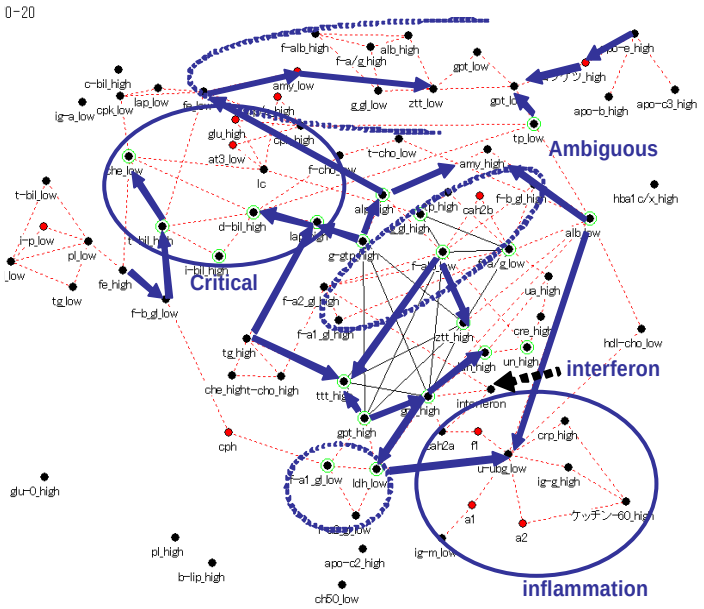


Fig. 4. For cases of hepatitis C. The ambiguity in interpreting GOT_low and GPT_low in the dotted frame at the top caused a new concern.

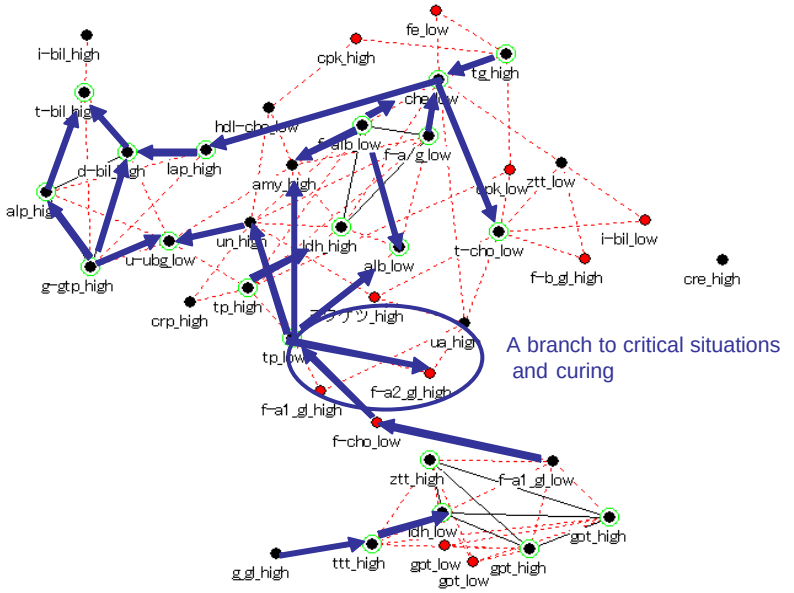


Fig. 5. Scenarios for cases including GOT_low and GPT_low

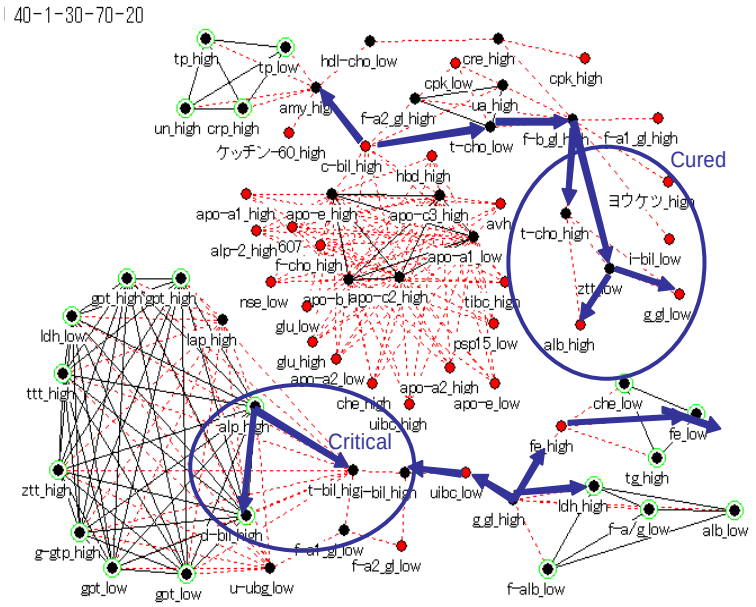


Fig. 6. Hepatitis C without interferon, with GOT_low & GPT_low. Making scenario maps including GOT_low, GPT_low, and additional conditions, we clarified the significance of proteins, e.g. F_B_GL.

5. Discussion: Subjective but Trustworthy

We obtained some qualitative new findings. That is, HDL is relevant to the turning point for hepatitis B, to shift to a critical situation of the liver. And, the effect of interferon is relevant to the change in the quantity of protein (e.g. F-A1_GL and F-B_GL), and this effect appears beginning with the recovery of TP. The latter result is apparently significant from Fig.7, but it is still an open problem how interferon affects such proteins as a globulins. All in all, "reasonable, and sometimes novel and useful" scenarios of state-transitions were obtained according to the surgeon.

Although not covered above, we also found other scenarios approaching to significant situations. For example, reader can find the increase in AMY (amylase) at the arrow terminals in Fig.1, Fig.4, Fig.5, and Fig.6, which seems to be relevant to surgical operations. This corresponds to reference [6]. Also, the decrease in GOT because of the exhaustion of liver cells, "bridge" effect of F-B_GL in the cure by interferon, etc. were visualized, not shown in this paper. These are the effects of the double helix process we developed, which can be summarized very shortly as a concern-based focusing of target data.

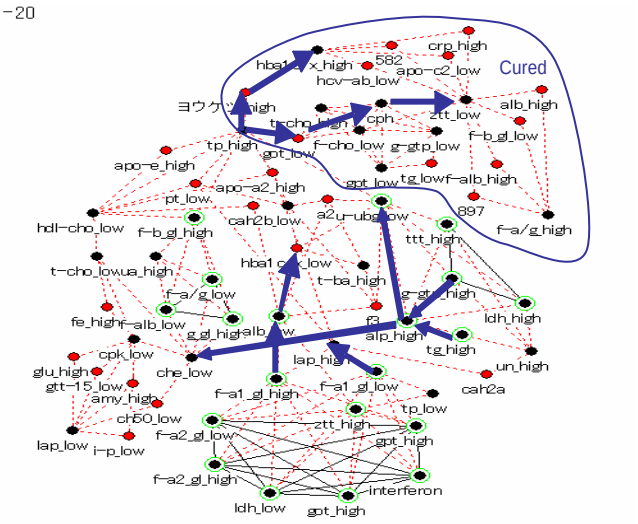


Fig. 7. Cases with F-A2_GL_high & F_A2_GL_low, and interferon. The effect of interferon is clearly extracted at the top of the map.

Apparently, subjective interests of the surgeon influences the results to be obtained in this manner, but the results are objectively trustworthy because they just show the summary of facts, i.e. objective facts selected on the surgeon's subjective experiences. As a positive face of this combination of subjectivity and objectivity we discovered a broad range of decision-oriented knowledge, i.e. scenarios, because subjectivity has both effects of widening and narrowing the target environment. Here, widening was aided by the projection of the scenario maps to one's own experiences in various real-world contexts. The narrowing, i.e. data-focusing, was aided by Text Drop. Like a

human wandering in an unknown island, a doctor wondering which way the patient can go will be helped by scenario map, a map with directed landmarks especially if the symptom is ambiguous or novel.

6. Conclusions

Scenario emergence and chance discovery are useful concepts for decisions in the real world, where events occur dynamically and one is required to make a decision promptly at the time a chance, i.e. significant event occurred. Here we showed an introductory application of scenario emergence with discovering the triggering events of essential scenarios, in the area of hepatitis treatment. Some interesting results were obtained according to the surgeon, but it is meaningless to say "interesting" – the next step we are to work on clarifying what a human means by "interesting" via the further runs on the DH process, which can be regarded as a process of human's self-discovery.

Acknowledgment The study has been conducted by Scientific Research on Priority Area "Active Mining." We appreciate Chiba University Hospital for serving us with the priceless data, under the convenient contract of its use for research.

References

[1] Ohsawa Y, McBurney P eds, Chance Discovery, Springer Verlag (2003)

[2]The Danish Board of Technology, European Participatory Technology Assessment: Participatory Methods in Technology Assessment and Technology Decision-Making,. www.tekno.dk/europta

[3] Gaver WW, Beaver J, and Benford S., Ambiguity as a Resource for Design, in CHI 2003 (2003)

[4] Masaki Usui and Yukio Ohsawa: Chance Discovery in Textile Market by Group Meeting with Touchable Key Graph, On-line Proceedings of Social Intelligence Design International Concerence (2003) .

[5] Tsumoto S., et al, Trend-evaluating Multiscale Analysis of the Hepatitis Dataset, Annual Report of Active Mining, Scientific Research on Priority Areas, 191-198 (2003)

[6] Miyagawa S, et al, Serum Amylase elevation following hepatic resection in patients with chronic liver disease. American J. Surg. 171(2), 235-238 (1996)

Empirical Comparison of Clustering Methods for Long Time-Series Databases

Shoji Hirano and Shusaku Tsumoto

Department of Medical Informatics, Shimane University School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
hirano@ieee.org, tsumoto@computer.org

Abstract. This paper presents a comparative study of methods for clustering long-term temporal data. We split a clustering procedure into two processes: similarity computation and grouping. As similarity computation methods, we employed dynamic time warping (DTW) and multiscale matching. As grouping methods, we employed conventional agglomerative hierarchical clustering (AHC) and rough sets-based clustering (RC). Using various combinations of these methods, we performed clustering experiments of the hepatitis data set and evaluated validity of the results. The results suggested that (1) complete-linkage (CL) criterion outperformed average-linkage (AL) criterion in terms of the interpret-ability of a dendrogram and clustering results, (2) combination of DTW and CL-AHC constantly produced interpretable results, (3) combination of DTW and RC would be used to find the core sequences of the clusters, (4) multiscale matching may suffer from the treatment of 'no-match' pairs, however, the problem may be eluded by using RC as a subsequent grouping method.

1 Introduction

Clustering of time-series data [1] has been receiving considerable interests as a promising method for discovering interesting features shared commonly by a set of sequences. One of the most important issue in time-series clustering is determination of (dis-)similarity between the sequences. Basically, the similarity of two sequences is calculated by accumulating distances of two data points that are located at the same time position, because such a distance-based similarity has preferable mathematical properties that extend the choice of grouping algorithms. However instead, this method requires that the lengths of all sequences be the same. Additionally, it cannot compare structural similarity of the sequences; for example, if two sequences contain the same number of peaks, but at slightly different phases, their 'difference' is emphasized rather than their structural similarity [2].

These drawbacks are serious in the analysis of time-series data collected over long time. The long time-series data have the following features. First, the lengths and sampling intervals of the data are not uniform. Starting point of data acquisition would be several years ago or even a few decades ago. Arrengement of

the data should be performed, however, shortening a time-series may cause the loss of precious information. Second, long-time series contains both long-term and short-term events, and their lengths and phases are not the same. Additionally, the sampling interval of the data would be variant due to the change of acquisition strategy over long time.

Some methods are considered to be applicable for clustering long time series. For example, dynamic time warping (DTW) [3] can be used to compare the two sequences of different lengths since it seeks the closest pairs of points allowing one-to-many point matching. This feature also enable us to capture similar events that have time shifts. Another approach, multiscale structure matching [6][5], can also be used to do this work, since it compares two sequences according to the similarity of partial segments derived based on the inflection points of the original sequences. However, there are few studies that empirically evaluate usefulness of these methods on real-world long time-series data sets.

This paper reports the results of empirical comparison of similarity measures and grouping methods on the hepatitis data set [7]. The hepatitis dataset is the unique, long time-series medical dataset that involves the following features: irregular sequence length, irregular sampling interval and co-existence of clinically interesting events that have various length (for example acute events and chronic events). We split a clustering procedure into two processes: similarity computation and grouping. For similarity computation, we employed DTW and multiscale matching. For grouping, we employed conventional agglomerative hierarchical clustering [8] and rough sets-based clustering [9], focusing that these methods can be used as un-supervised methods and are suitable for handling relative similarity induced by multiscale matching. For every combination of the similarity computation methods and grouping methods, we performed clustering experiments and evaluated validity of the results.

2 Materials

We employed the chronic hepatitis dataset [7], which were provided as a common dataset for ECML/PKDD Discovery Challenge 2002 and 2003. The dataset contained long time-series data on laboratory examinations, which were collected at Chiba University Hospital in Japan. The subjects were 771 patients of hepatitis B and C who took examinations between 1982 and 2001. We manually removed sequences for 268 patients because biopsy information was not provided for them and thus their virus types were not clearly specified. According to the biopsy information, the expected constitution of the remaining 503 patients were, B / C-noIFN / C-IFN = 206 / 100 / 197. However, due to existence of missing examinations, the numbers of available sequences could be less than 503.

The dataset contained the total of 983 laboratory examinations. However, in order to simplify our experiments, we selected 13 items from blood tests relevant to the liver function: ALB, ALP, G-GL, G-GTP, GOT, GPT, HGB, LDH, PLT, RBC, T-BIL, T-CHO and TTT. Details of each examination are available at the URL [7].

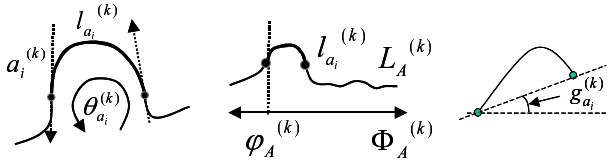


Fig. 1. Segment difference.

Each sequence originally had different sampling intervals from one day to one year. From preliminary analysis we found that the most frequently appeared interval was one week; this means that most of the patients took examinations on a fixed day of a week. According to this observation, we determined resampling interval to seven days. A simple summary showing the number of data points after resampling is as follows (item=ALB, $n = 499$) : mean=456.87, sd=300, maximum=1080, minimum=7. Note that one point equals to one week; therefore, 456.87 points equals to 456.87 weeks, namely, about 8.8 years.

3 Methods

We have implemented algorithms of symmetrical time warping describe briefly in [2] and one-dimensional multiscale matching described in [4]. We modified segment difference in multiscale matching as follows.

$$d(a_i^{(k)}, b_j^{(h)}) = \max(\theta, l, \phi, g), \tag{1}$$

where θ, l, ϕ, g respectively represent differences on rotation angle, length, phase and gradient of segments $a_i^{(k)}$ and $b_j^{(h)}$ at scales k and h. These differences are defined as follows:

$$\theta(a_i^{(k)}, b_j^{(h)}) = \mid \theta_{a_i}^{(k)} - \theta_{b_j}^{(h)} \mid /2\pi, \tag{2}$$

$$l(a_i^{(k)}, b_j^{(h)}) = \left| \frac{l_{a_i}^{(k)}}{L_A^{(k)}} - \frac{l_{b_j}^{(h)}}{L_B^{(h)}} \right|, \tag{3}$$

$$\phi(a_i^{(k)}, b_j^{(h)}) = \left| \frac{\phi_{a_i}^{(k)}}{\Phi_A^{(k)}} - \frac{\phi_{b_j}^{(h)}}{\Phi_B^{(h)}} \right|, \tag{4}$$

$$g(a_i^{(k)}, b_j^{(h)}) = \mid g_{a_i}^{(k)} - g_{b_j}^{(h)} \mid . \tag{5}$$

Figure 1 provides an illustrative explanation of these terms. Multiscale matching usually suffers from the shrinkage of curves at high scales caused by excessive smoothing with a Gaussian kernel. On one-dimensional time-series data, shrinkage makes all sequences flat at high scales. In order to elude this problem, we applied shrinkage correction proposed by Lowe [10].

Table 1. Comparison of the number of generated clusters. Each item represents clusters for Hepatitis B / C-noIFN / C-IFN cases.

Exam	Number of Instances	Number of Generated Clusters					
		DTW			Multiscale Matching		
		AL-AHC	CL-AHC	RC	AL-AHC	CL-AHC	RC
ALB	204 / 99 / 196	8 / 3 / 3	10 / 6 / 5	38 / 22 / 32	19 / 11 / 12	22 / 21 / 27	6 / 14 / 31
ALP	204 / 99 / 196	6 / 4 / 6	7 / 7 / 10	21 / 12 / 29	10 / 18 / 14	32 / 16 / 14	36 / 12 / 46
G-GL	204 / 97 / 195	2 / 2 / 5	2 / 2 / 11	1 / 1 / 21	15 / 16 / 194	16 / 24 / 194	24 / 3 / 49
G-GTP	204 / 99 / 196	2 / 4 / 11	2 / 6 / 7	1 / 17 / 4	38 / 14 / 194	65 / 14 / 19	35 / 8 / 51
GOT	204 / 99 / 196	8 / 10 / 25	8 / 4 / 7	50 / 18 / 60	19 / 12 / 24	35 / 19 / 19	13 / 14 / 15
GPT	204 / 99 / 196	3 / 17 / 7	7 / 4 / 7	55 / 29 / 51	23 / 30 / 8	24 / 16 / 16	11 / 7 / 25
HGB	204 / 99 / 196	3 / 4 / 13	2 / 3 / 9	1 / 16 / 37	43 / 15 / 15	55 / 19 / 22	1 / 12 / 78
LDH	204 / 99 / 196	7 / 7 / 9	15 / 10 / 8	15 / 15 / 15	20 / 25 / 195	24 / 9 / 195	32 / 16 / 18
PLT	203 / 99 / 196	2 / 13 / 9	2 / 7 / 6	1 / 15 / 19	33 / 5 / 12	34 / 15 / 17	1 / 11 / 25
RBC	204 / 99 / 196	3 / 4 / 6	3 / 4 / 7	1 / 14 / 26	32 / 16 / 13	40 / 23 / 17	1 / 6 / 17
T-BIL	204 / 99 / 196	6 / 5 / 5	9 / 5 / 4	203 / 20 / 30	17 / 25 / 6	20 / 30 / 195	11 / 23 / 48
T-CHO	204 / 99 / 196	2 / 2 / 7	5 / 2 / 5	20 / 1 / 27	12 / 13 / 13	17 / 23 / 19	12 / 5 / 23
TTT	204 / 99 / 196	7 / 2 / 5	8 / 2 / 6	25 / 1 / 32	29 / 10 / 6	39 / 16 / 16	25 / 16 / 23

We also implemented two clustering algorithms, agglomerative hierarchical clustering (AHC) in [8] and rough sets-based clustering (RC) in [9]. For AHC we employed two linkage criteria, average-linkage AHC (CL-AHC) and complete-linkage AHC (AL-AHC).

In the experiments, we investigated the usefulness of various combinations of similarity calculation methods and grouping methods in terms of the interpretability of the clustering results. Procedures of data preparation were as follows. First, we selected one examination, for example ALB, and split the corresponding sequences into three subsets, B, C-noIFN and C-IFN, according to the virus type and administration of interferon therapy. Next, for each of the three subgroups, we computed dissimilarity of each pair of sequences by using DTW. After repeating the same process with multiscale matching, we obtained 2×3 sets of dissimilarities: one obtained by DTW, and another obtained by multiscale matching.

Then we applied grouping methods AL-AHC, CL-AHC and RC to each of the three dissimilarity sets obtained by DTW. This yielded $3 \times 3 = 9$ sets of clusters. After applying the same process to the sets obtained by multiscale-matching, we obtain the total of 18 sets of clusters.

The above process is repeated with the remaining 12 examination items. Consequently, we constructed 12×18 clustering results. Note that in this experiments we did not perform cross-examination comparison, for example comparison of an ALB sequence with a GPT sequence.

We used the following parameters for rough clustering: $\sigma = 5.0$, $T_h = 0.3$. In AHC, cluster linkage was terminated when increase of dissimilarity firstly exceeded mean+SD of the set of all increase values.

4 Results

Table 1 provides the numbers of generated clusters for each combination. Let us explain the table using the raw whose first column is marked ALB. The second column "Number of Instances" represents the number of patients who took the ALB examination. Its value 204/99/196 represents that 204 patients of Hepatitis B, 99 patients of Hepatitis C (who did not take IFN therapy) and 196 patients of Hepatitis C (who took IFN therapy) took this examination. Since one patient has one time-series examination result, the number of patients corresponds to the number of sequences. The third column shows the number of generated clusters. Using DTW and AL-AHC, 204 hepatitis B sequences were grouped into 8 clusters. 99 C-noIFN sequences were grouped into 3 clusters, as well as 196 C-IFN sequences.

4.1 DTW and AHCs

Let us first investigate the case of DTW-AHC. Comparison of DTW-AL-AHC and DTW-CL-AHC implies that the results can be different if we use different linkage criterion. Figure 2 left image shows a dendrogram generated from the GTP sequences of type B hepatitis patients using DTW-AL-AHC. It can be observed that the dendrogram of AL-AHC has an ill-formed structure like 'chaining', which is usually observed with single-linkage AHC. For such an ill-formed structure, it is difficult to find a good point to terminate merging of the clusters. In this case, the method produced three clusters containing 193, 9 and 1 sequences respectively. Figure 3 left image shows a part of the sequences grouped into the largest cluster. Almost all types of sequences were included in this cluster and thus no interesting information was obtained.

On the contrary, the dendrogram of CL-AHC shown in the right of Figure 2 demonstrates a well formed hierarchies of the sequences. With this dendrogram the method produced 7 clusters containing 27, 21, 52, 57, 43, 2, and 1 sequences. Figure 3 right image and Figure 4 show examples of the sequences grouped into the first three clusters respectively. One can observe interesting features for each cluster. The first cluster contains sequences that involve continuous vibration of the GPT values. These patterns may imply that the virus continues to attack the patient's body periodically. The second cluster contains very short, meaningless sequences, which may represent the cases that patients stop or cancel receiving the treatment quickly. The third cluster contains another interesting pattern: vibrations followed by the flat, low values. This case may represent the cases that the patients were cured by some treatments, or naturally.

4.2 DTW and RC

For the same data, rough set-based clustering method produced 55 clusters. Fifty five clusters were too many for 204 objects, however, 41 of 55 clusters contained less than 3 sequences, and furthermore, 31 of them contained only one sequence. This was because of the rough set-based clustering tends to produce independent,

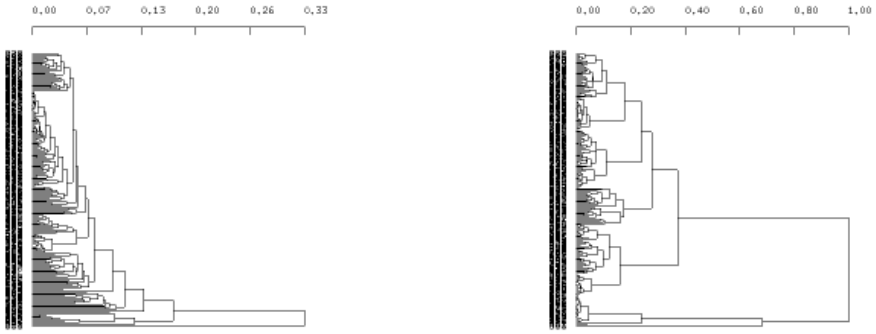


Fig. 2. Dendrograms for DTW-AHC-B. Left: AL-AHC. Right: CL-AHC.

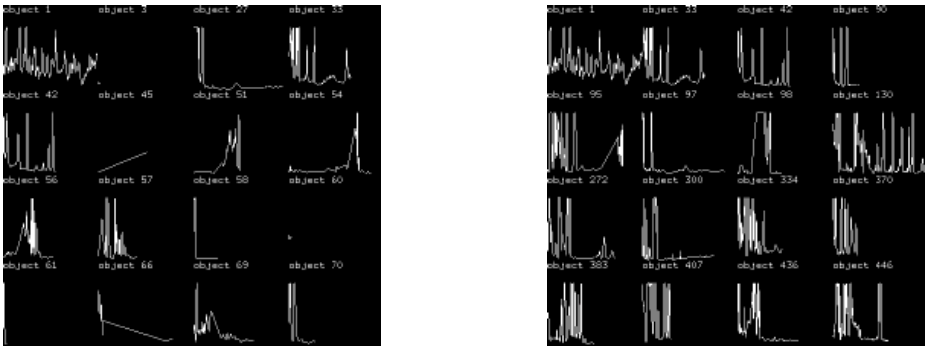


Fig. 3. Examples of the clusters. Left: AL-AHC. Right: CL-AHC.

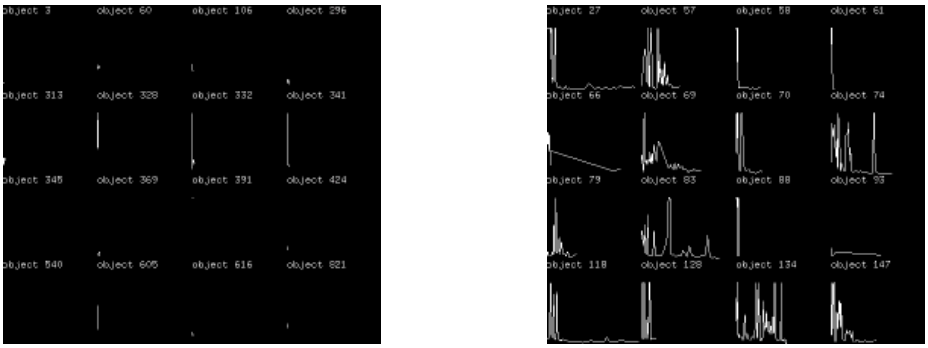


Fig. 4. Other examples of the clusters obtained by CL-AHC. Left: the second cluster containing 21 sequences. Right: the third cluster containing 52 sequences.

small clusters for objects being intermediate of the large clusters. Ignoring small ones, we found 14 clusters containing 53, 16, 10, 9, 6 ... objects. The largest

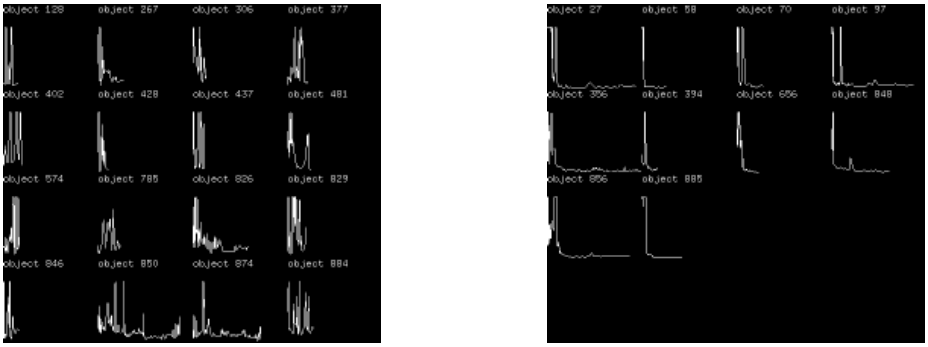


Fig. 5. Examples of the clusters obtained by RC. Left: the second cluster containing 16 sequences. Right: the third cluster containing 10 sequences.

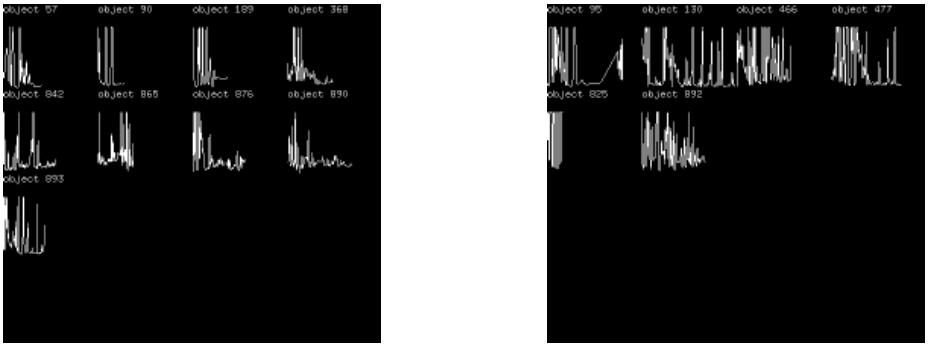


Fig. 6. Other examples of the clusters obtained by RC. Left: the fourth cluster containing 9 sequences. Right: the fifth cluster containing 6 objects.

cluster contained short sequences quite similarly to the case of CL-AHC. Figure 5 and 6 show examples of sequences for the 2nd, 3rd, 4th and 5th clusters. Because this method evaluates the indiscernibility degree of objects, each of the generated clusters contains strongly similar sets of sequences. Although populations in the clusters are not so large, one can clearly observe the representative of the interesting patterns described previously at CL-AHC.

4.3 Multiscale Matching and AHCs

Comparison of Multiscale Matching-AHC pairs with DTW-AHC pairs shows that Multiscale Matching's dissimilarities resulted in producing the larger number of clusters than DTW's dissimilarities.

One of the important issues in multiscale matching is treatment of 'no-match' sequences. Theoretically, any pairs of sequences can be matched because a sequence will become single segment at enough high scales. However, this is not a realistic approach because the use of many scales results in the unacceptable

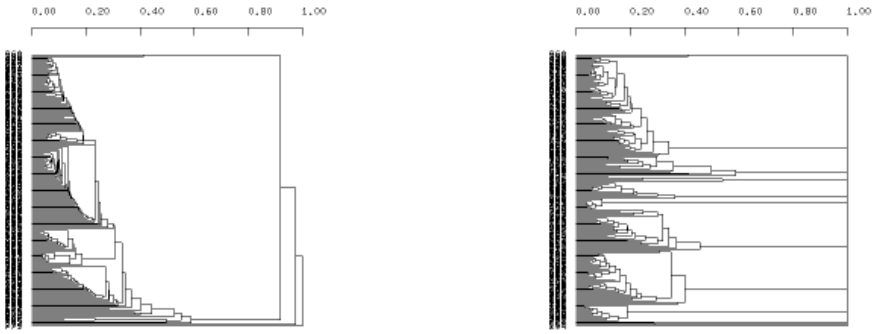


Fig. 7. Dendrograms for MSMmatch-AHC-C-IFN. Left: AL-AHC. Right: CL-AHC.

increase of computational time. If the upper bound of the scales is too low, the method may possibly fail to find the appropriate pairs of subsequences. For example, suppose we have two sequences, one is a short sequence containing only one segment and another is a long sequence containing hundreds of segments. The segments of the latter sequence will not be integrated into one segment until the scale becomes considerably high. If the range of scales we use does not cover such a high scale, the two sequences will never be matched. In this case, the method should return infinite dissimilarity, or a special number that identifies the failed matching.

This property prevents AHCs from working correctly. CL-AHC will never merge two clusters if any pair of 'no-match' sequences exist between them. AL-AHC fails to calculate average dissimilarity between two clusters. Figure 7 provides dendrograms for GPT sequences of Hepatitis C (with IFN) patients obtained by using multiscale matching and AHCs. In this experiment, we let the dissimilarity of 'no-match' pairs the same as the most dissimilar 'matched' pairs in order to elude computational problems. The dendrogram of AL-AHC is compressed to the small-dissimilarity side because there are several pairs that have excessively large dissimilarities. The dendrogram of CL-AHC demonstrates that the 'no-match' pairs will not be merged until the end of the merging process.

For AL-AHC, the method produced 8 clusters. However, similarly to the previous case, most of the sequences (182/196) were included in the same cluster. As shown in Figure 8 left image, no interesting information was found in the cluster. For CL-AHC, the method produced 16 clusters containing 71, 39, 29, ... sequences. Figure 8 right image and Figure 9 provide examples of the sequences grouped into the three primary clusters, respectively. Similar sequences were found in the clusters, however, obviously dissimilar sequences were also observed in their clusters.

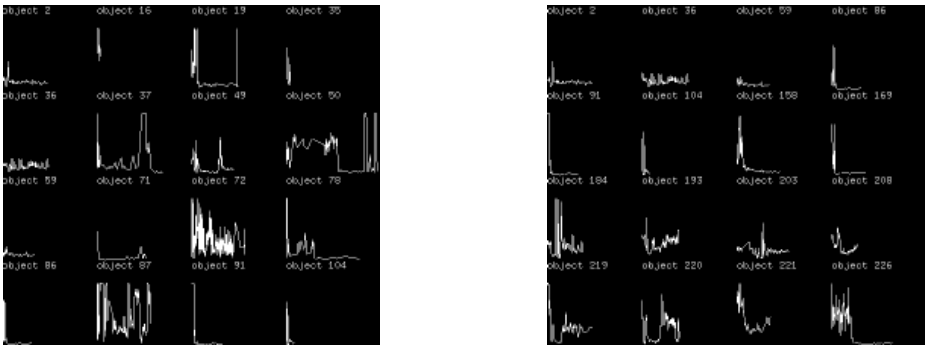


Fig. 8. Examples of the sequences clusters obtained by AHCs. Left: AL-AHC. The first cluster containing 182 sequences. Right: CL-AHC. the first cluster containing 71 sequences.

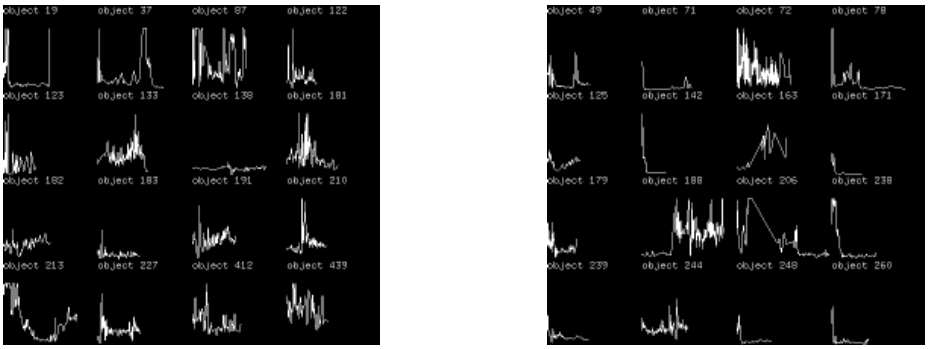


Fig. 9. Other examples of the clusters obtained by CL-AHC. Left: the second cluster containing 39 sequences. Right: the third cluster containing 29 sequences.

4.4 Multiscale Matching and RC

Rough set-based clustering method produced 25 clusters containing 80, 60, 18, 6 ... sequences. Figures 10 and 11 represent examples of the sequences grouped into the four primary clusters. It can be observed that the sequences were properly clustered into the three major patterns: continuous vibration, flat after vibration, and short. This should result from the ability of the clustering method for handling relative proximity.

5 Conclusions

In this paper we have reported a comparative study of clustering methods for long time-series data analysis. Although the subjects for comparison were limited, the results suggested that (1) complete-linkage criterion outperforms average-linkage criterion in terms of the interpret-ability of a dendrogram and

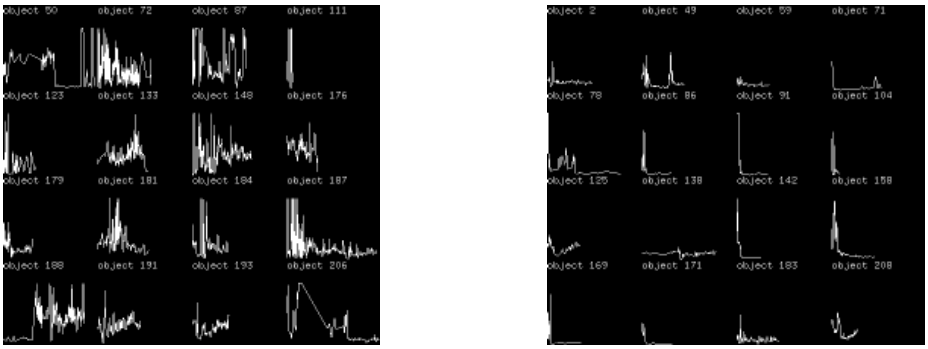


Fig. 10. Examples of the clusters obtained by RC. Left: the second cluster containing 16 sequences. Right: the third cluster containing 10 sequences.

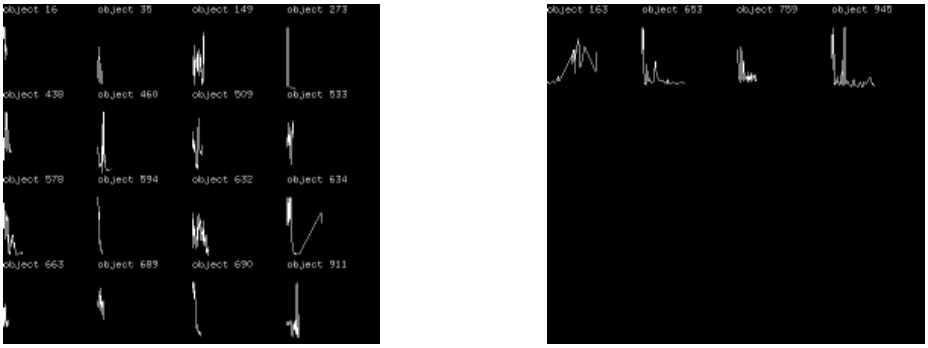


Fig. 11. Other examples of the clusters obtained by RC. Left: the fourth cluster containing 9 sequences. Right: the fifth cluster containing 6 objects.

clustering results, (2) combination of DTW and CL-AHC constantly produced interpretable results, (3) combination of DTW and RC would be used to find core sequences of the clusters. Multiscale matching may suffer from the problem of 'no-match' pairs, however, the problem may be eluded by using RC as a subsequent grouping method.

Acknowledgments

This work was supported in part by the Grant-in-Aid for Scientific Research on Priority Area (B)(No.759) "Implementation of Active Mining in the Era of Information Flood" by the Ministry of Education, Culture, Science and Technology of Japan.

References

1. E. Keogh (2001): Mining and Indexing Time Series Data. Tutorial at the 2001 IEEE International Conference on Data Mining.
2. Chu, S., Keogh, E., Hart, D., Pazzani, M. (2002). Iterative Deepening Dynamic Time Warping for Time Series. In proceedings of the second SIAM International Conference on Data Mining.
3. D. J. Berndt and J. Clifford (1994): Using dynamic time warping to find patterns in time series. Proceedings of AAAI Workshop on Knowledge Discovery in Databases: 359-370.
4. S. Hirano and S. Tsumoto (2002): Mining Similar Temporal Patterns in Long Time-series Data and Its Application to Medicine. Proceedings of the IEEE 2002 International Conference on Data Mining: pp. 219–226.
5. N. Ueda and S. Suzuki (1990): A Matching Algorithm of Deformed Planar Curves Using Multiscale Convex/Concave Structures. IEICE Transactions on Information and Systems, J73-D-II(7): 992–1000.
6. F. Mokhtarian and A. K. Mackworth (1986): Scale-based Description and Recognition of planar Curves and Two Dimensional Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(1): 24-43
7. URL: http://lisp.vse.cz/challenge/ecmlpkdd2003/
8. B. S. Everitt, S. Landau, and M. Leese (2001): Cluster Analysis Fourth Edition. Arnold Publishers.
9. S. Hirano and S. Tsumoto (2003): An Indiscernibility-based Clustering Method with Iterative Refinement of Equivalence Relations. Journal of Advanced Computational Intelligence and Intelligent Informatics, (in press).
10. Lowe, D.G (1980): Organization of Smooth Image Curves at Multiple Scales. International Journal of Computer Vision, 3:119–130.

Classification of Pharmacological Activity of Drugs Using Support Vector Machine

Yoshimasa TAKAHASHI, Katsumi NISHIKOORI, Satoshi FUJISHIMA

Department of Knowledge-based Information Engineering,
Toyohashi University of Technology
1-1 Hibarigaoka Tempaku-cho, Toyohashi, 441-8580 Japan
{taka, katsumi, fujisima}@mis.tutkie.tut.ac.jp

Abstract. In the present work, we investigated an applicability of Support Vector Machine (SVM) for classification of pharmacological activities of drugs. The numerical description of chemical structure of each drug was based the Topological Fragment Spectra (TFS) which was proposed by the authors. Dopamine antagonists of 1,227 that interact with different type of receptors (D1, D2, D3 and D4) were used for training the SVM. For a prediction set of 137 drugs that were not included in the training set, the obtained SVM classified 123 (89.8 %) drugs of them into their own activity classes correctly. The comparison of the usage of SVM and that of artificial neural network will be discussed too.

1 Introduction

For a half century, a lot of effort has been devoted to develop new drugs. It is true that such new drugs allow us to have better life. However, serious side effects of the drugs often have been reported and those raise a social problem. The aim of this research project is in establishing a basis of computer-aided risk report for chemicals on the basis of pattern recognition techniques and chemical similarity analysis.

The authors [1] proposed the Topological Fragment Spectral (TFS) method for a numerical vector representation of the topological structure profile of a molecule. The TFS provides us a useful tool for the evaluation of structural similarity between molecules. In our preceding work [2], we reported that an artificial neural network approach combined with input signals of the TFS allowed us to successfully classify the type of activities for dopamine receptor antagonists, and it could be applied to the prediction of active class of unknown compounds. And we also suggested that similar structure searching on the basis of the TFS representation of molecules could provide us a chance discovery for some new insight or knowledge from a huge amount of data. It is clear that these approaches are quite useful for data mining and data discovery problems too.

On the other hand, in the past few years, support vector machines (SVM) have brought us a great interest in the area of machine learning due to its superior generalization ability in a wide variety of learning problems [3-5]. Support vector machine is

originated from perceptron theory, but some classical problems such as multiple local minima, curse of dimensionality and over-fitting in artificial neural network little occur in this method. Here we investigate the utility of support vector machine combined with the TFS representation of chemical structures in classification of pharmacological activity of drugs.

2 Methods

2.1 Numerical representation of chemical structure

In the present work, to describe structural information of drugs, Topological Fragment Spectra (TFS) method [1] was employed. The TFS is based on enumeration of all the possible substructures from a chemical structure and numerical characterization of them. A chemical structure can be regarded as a graph in terms of graph theory. For graph representation of chemical structures, hydrogen suppressed graph is often used.

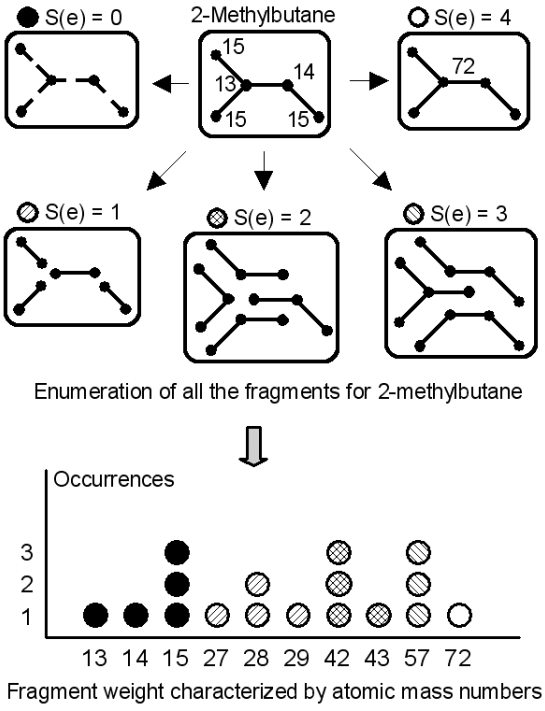


Fig. 1. A schematic flow of TFS generation. S(e) is the number of edges (bonds) of the fragments to be generated.

To get a TFS representation of a chemical structure, all the possible subgraphs with the specified number of edges are enumerated. Subsequently, every subgraph is characterized with a numerical quantity. For the characterization of a subgraph we used the overall sum of the mass numbers of the atoms corresponding to the vertexes of the subgraphs. In this characterization process, suppressed hydrogen atoms are taken into account as augmented atoms. The histogram is defined as a TFS that is obtained from the frequency distribution of a set of individually characterized subgraphs (i.e. substructures or structural fragments) according to the value of their characterization index.

The TFS generated along with this manner is a digital representation of topological structural profile of a drug molecule. This is very similar to that of mass spectra of chemicals. A schematic flow of the TFS creation is shown in Fig. 1.

The computational time required for the exhaustive enumeration of all possible substructures is often very large especially for the molecules that involve highly fused rings. To avoid such a problem the use of subspectrum was employed for the present work, in which each spectrum could be described with structural fragments up to a specified size in the number of edges (bonds).

2.2 Support Vector Machine

Support vector machine has been focused as a powerful nonlinear classifier due to introducing kernel function trick in the last decade [5]. The basic idea of SVM is described as follows: it maps the input vectors **x** into a higher dimensional feature space **z** through some nonlinear mapping chosen in advance. In this space, an optimal discriminant surface with maximum margin is constructed (Fig.2). Given a training dataset represented by $\mathbf{X}(\mathbf{x}_1,...,\mathbf{x}_i,...,\mathbf{x}_n)$, $\mathbf{x}_i$ that are linearly separable with class labels $y_i \in \{-1,1\}, i = 1,...,n$, the discriminant function can be described as the following equation.

$$f(\mathbf{x}_i) = (\mathbf{w} \cdot \mathbf{x}_i) + b \tag{1}$$

Where **w** is a weight vector, b is a bias. The discriminant surface can be described as $f(\mathbf{x}_i) = 0$. The plane with maximum margin can be determined by minimizing $L(\mathbf{w}) = \|\mathbf{w}\|^2$,

$$\|\mathbf{w}\|^2 = \mathbf{w} \cdot \mathbf{w} = \sum_{l=1}^{d} w_l^2 \tag{2}$$

with constraints,

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \quad (i = 1,...,n) \tag{3}$$

The decision function is represented as $f(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b)$ for classification, where *sign* is a simple sign function which returns 1 for positive argument and -1 for a negative argument. This basic concept would be generalized to a linearly inseparable case by introducing slack variables ξ_i and minimizing the following quantity,

$$\frac{1}{2}\mathbf{w}\cdot\mathbf{w}+C\sum_{i=1}^{n}\xi_i \tag{4}$$

which is subject to the constraints $y_i=(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1-\xi_i$ and $\xi_i\geq 0$. This primal information of the optimization problem reduces to the previous one for separable data when the constant C is enough large. This quadratic optimization problem with the constraints can be reformulated using Lagrangian multipliers α.

$$W(\alpha)=\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j \mathbf{x}_i\cdot\mathbf{x}_j-\sum_{i=1}^{n}\alpha_i \tag{5}$$

with the constraints $0\leq\alpha_i\leq C$ and $\sum_{i=1}^{n}\alpha_i y_i=0$.

Since the training points $\mathbf{x}_i$ do appear in the final solution only via dot products, this formulation can be extended to general nonlinear functions by using the concepts of nonlinear mappings and kernels [6]. Given a mapping, $\mathbf{x}\rightarrow\phi(\mathbf{x})$, the dot product in the final space can be replaced by a kernel function.

$$f(\mathbf{x})=g(\phi(\mathbf{x}))=\sum_{i=1}^{n}\alpha_i y_i K(\mathbf{x},\mathbf{x}_i)+b \tag{6}$$

Here we used radial basis function as the kernel function for mapping the data into a higher dimensional space.

$$K(\mathbf{x},\mathbf{x}')=\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{\sigma^2}\right) \tag{7}$$

Basically, the SVM is a binary classifier. For classification problem of three or more categorical data, plural discrimination functions are required for the current multi categorical classification. In this work, one-against-the-rest approach was used for the case. The TFS were used as input feature vectors to the SVM. All the SVM analyses were carried out using a computer program developed by the authors according to Platt's algorithm [7].

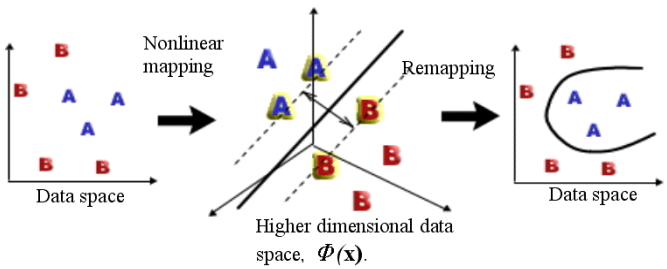


Fig. 2. Illustrative scheme of nonlinear separation of two classes by the SVM with a kernel trick.

2.3 Data set

In this work we employed 1,364 dopamine antagonists that interact with four different types of receptors (D1, D2, D3 and D4). The data are a subset of MDDR [6] database. The data set was divided into two groups; training set and prediction set. The two include 1,228 compounds and 136 compounds respectively.

3 Results and Discussion

3.1 Classification and Prediction of pharmacological activity of dopamine receptor antagonists by SVM

The applicability of the SVM combined with the TFS representation of chemical structures was validated in discriminating active classes of pharmaceutical drugs. Here, Dopamine antagonists of 1,227 that interact with different type of receptors (D1, D2, D3 and D4) were used for training the SVM with their TFS to classify the type of activity. The SVM model was able to learn to classify all the compounds into their own active classes correctly. Then, the trained SVM model was used for the prediction of activity for unknown compounds. For 137 separately prepared in advance, the activity classes of 123 compounds (89.8%) were correctly predicted. The results are summarized in Table 1. In the comparison among the results for four classes it is shown that the prediction result for D4 receptor antagonists is better than those of other classes in both cases of the training and the prediction. It would be considered that the SVM model have got a set of well defined support vectors from the training set because the number of samples is considerably larger than those of the others. These results show that the SVM provides us a very powerful tool for the classification and prediction of pharmaceutical drug activities, and that the TFS representation should be suitable as input signal to SVM in the case.

Table 1. Results of SVM analyses for 1364 dopamine antagonists

	Training		Prediction	
Class	Data	%Correct	Data	%Correct
All	**1227**	**100**	**123/137**	**89.8**
D1	155	100	15/18	83.3
D2	356	100	31/39	79.5
D3	216	100	22/24	91.7
D4	500	100	55/56	98.2

3.2 Comparison between SVM and ANN

In the preceding work [2], the authors reported that an artificial neural network based on the TFS gives us a successful tool for the discrimination of active classes of drugs. To evaluate the better performance of the SVM approach for the current problem, here, we tried to compare the results by SVM with those by artificial neural network (ANN). The data set of 1364 drugs used in the above section was employed for the analysis as well. Ten-fold cross validation technique was used for the computational trial. The results were summarized in Table 2.

Table 2. Comparison between SVM and ANN by ten-fold cross validation test.

	SVM		ANN	
Active Class	Training %Correct	Prediction %Correct	Training %Correct	Prediction %Correct
All	**100**	**90.6**	**87.5**	**81.1**
D1	100	87.5	76.0	70.7
D2	100	86.1	80.7	69.9
D3	100	88.3	90.9	85.8
D4	100	95.5	94.5	90.5

The table shows that the results obtained by SVM were better ANN for every case in this trial. These results show that the TFS-based support vector machine could give us more successful results than TFS-base artificial neural network for the current problem.

4 Conclusions and Future Work

In this work, we investigated the usage of support vector machine for classification of pharmacological activities of drugs. The Topological Fragment Spectra (TFS) method was used for the numerical description of chemical structure information of each drug. It is concluded that the TFS-based support vector machine can be successfully applied for the prediction of type of activities of drug molecules. However, because many instances are required to establish predictive risk assessment and risk report of drugs and chemicals, further works would be still required to test the usage of the TFS-based support vector machine with much more various kinds of drugs. In addition, it would also be interesting to identify the support vectors chosen in the training phase and analyze their structural features from the view point of structure-activity relationships of the drug molecules.

The authors thank Prof. Takashi Okada and Dr. Masumi Yamakawa of Kwansei Gakuin University for their valuable discussion and comments. This work was supported by Grant-In-Aid for Scientific Research on Priority Areas (B) 13131210, Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Y. Takahashi, H. Ohoka, and Y. Ishiyama, Structural Similarity Analysis Based on Topological Fragment Spectra, *In* "*Advances in Molecular Similarity"*, **2**, (Eds. R. Carbo & P. Mezey), JAI Press, Greenwich, CT, (1998) 93-104
2. Y. Takahashi, S. Fujishima and K. Yokoe: Chemical Data Mining Based on Structural Similarity, International Workshop on Active Mining, The 2002 IEEE International Conference on Data Mining, Maebashi (2002) 132-135
3. V.N. Vapnik : The Nature of Statistical Learning Theory, Springer, (1995)
4. C. J. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery **2**, (1998) 121-167
5. S. W. Lee and A. Verri, Eds, Support Vector Machines 2002, LNCS 2388, (2002)
6. MDL Drug Data Report, MDL, ver. 2001.1, (2001)
7. J. C. Platt : Sequential Minimal Optimization： A Fast Algorithm for Training Support Vector Machines, Microsoft Research Tech. Report MSR-TR-98-14, Microsoft Research, 1998

Mining Chemical Compound Structure Data Using Inductive Logic Programming

Cholwich Nattee[1], Sukree Sinthupinyo[1], Masayuki Numao[1], and Takashi Okada[2]

[1] The Institute of Scientific and Industrial Research, Osaka University, 8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan
{cholwich,sukree,numao}@ai.sanken.osaka-u.ac.jp
[2] Center for Information & Media Studies, Kwansei Gakuin University
okada@kwansei.ac.jp

Abstract. We propose an approach for making FOIL better handling multiple-instance data. This learning problem arises when trying to generate hypotheses from examples in the form of positive and negative bags. Each bag contains one or more instances and a bag is labelled as positive when there is at least one positive instance, otherwise, it is labelled as negative. Several algorithms have been proposed for learning in this framework. However all of them can only handle data in the attribute-value form which has limitations in knowledge representation. Therefore, it is difficult to handle examples consisting of structures among components, such as chemical compounds data. In this paper, the Diverse Density, a measure for multiple-instance data, is applied to adapt the heuristic function in FOIL in order to improve learning accuracy in multiple-instance data. We conducts the experiments on real-world data related to chemical compound data analysis in order to show the improvement.

1 Introduction

Multiple-instance (MI) learning [1] is a framework extended from supervised learning in the case that training examples cannot be labelled completely. Training examples are grouped into labelled bags marked as positive if there is at least one instance known to be positive. Otherwise, they are marked as negative. The MI learning framework was originally motivated by the drug activity prediction problem which aims to determine whether aromatic drug molecules bind strongly to a target protein. As a lock and a key, the shape of molecules is the most important factor for determining this binding. The molecules can nevertheless adapt their shapes widely. Then, each shape is represented as an instance and the positive bags are the molecules with at least one shape binding well. On the other hand, the negative bags contains molecules whose shapes did not bind at all. Dietterich et al. formalised this framework and proposed the axis-parallel rectangles algorithm [1]. After this work, many approaches have been proposed for MI learning [2–4], they nevertheless aim for handling only data in

the attribute-value form where an instance is represented as a fixed-length vector inheriting a limitation that complicated relations among instances become difficult to be denoted, for example, representing chemical compound structures by describing atoms and bonds among atoms.

Inductive Logic Programming (ILP) has introduced more expressive first-order representation to supervised learning. ILP has been successfully applied to many applications and the first-order logic can also be represented the MI data well. However, in order to make the ILP systems able to generate more accurate hypotheses, the distance among instances would be useful because in MI data the positive instances cannot be specified exactly, thus the distance between positive instances and negative instances plays an important role in this determination. If there is an area that many instances from various positive bags locating together and that area is far from the instances from negative bags, the target concept would come from the instances in that area.

This paper presents the extension of FOIL [5] using the Diverse Density (DD) [2], a measure for evaluating MI data. Applying this measure will make FOIL more precisely identify the positive instances and generate more suitable hypotheses from training data. In this research, we focus on applying this approach to predict the characteristics of chemical compound. Each compound (or molecule) is represented using properties of atom and bonds between atoms.

The paper is organised as follows. The next section described the background of DD and FOIL which are the bases of our approach. Then the modification of FOIL algorithm is considered. We evaluate the proposed algorithm with chemical compound data. Finally the conclusion and our research direction are given.

2 Background

2.1 Diverse Density

The Diverse Density (DD) algorithm aims to measure a point in an n-dimensional feature space to be a positive instance. The DD at point p in the feature space shows both how many *different* positive bags have an instance near p, and how *far* the negative instances are from p. Thus, the DD is high in the area where instances from various positive bags are located together. It can be calculated as

$$DD(x) = \prod_i (1 - \prod_j (1 - exp(-\|B_{ij}^+ - x\|^2))) \cdot \prod_i \prod_j (1 - exp(-\|B_{ij}^- - x\|^2)) \quad (1)$$

where x is a point in the feature space and B_{ij} represents the j^{th} instance of the i^{th} bag in training examples. For the distance, the Euclidean distance is adopted then

$$\|B_{ij} - x\|^2 = \sum_k (B_{ijk} - x_k)^2 \tag{2}$$

In the previous approaches, several searching techniques are proposed for determining the value of features or the area in the feature space that maximises DD. In this paper, the DD is however applied in the heuristic function in order to evaluate each instance from the positive bags with the value between 0 and 1.

2.2 FOIL

The learning process in FOIL starts with a training set (examples) containing all positive and negative examples, constructs a function-free Horn clause (a hypothesis) to cover some of the positive examples, and removes the covered examples from the training set. Then it continues with the search for the next clause. When the clauses covering all the positive examples have been found, they are reviewed to eliminate any redundant clauses and re-ordered so that all recursive clauses come after the non-recursive ones.

FOIL uses a heuristic function based on the information theory for assessing the usefulness of a literal. It seems to provide effective guidance for clause construction. The purpose of this heuristic function is to characterise a subset of the positive examples. From the partial developing clause $R(V_1, V_2, \ldots, V_k) \leftarrow L_1, L_2, \ldots, L_{m-1}$, the training examples covered by this clause are denoted as T_i. The information required for T_i is given by

$$I(T_i) = -\log_2(|T_i^+|/(|T_i^+| + |T_i^-|)) \tag{3}$$

If a literal L_m is selected and yields a new set T_{i+1}, then the similar formula is given as

$$I(T_{i+1}) = -\log_2(|T_{i+1}^+|/(|T_{i+1}^+| + |T_{i+1}^-|)) \tag{4}$$

From above, a heuristic used in FOIL is calculated an amount of information gained when applying a literal L_m;

$$Gain(L_i) = T_i^{++} \times (I(T_{i+1}) - I(T_i)) \tag{5}$$

T_i^{++} in this equation is the positive examples extended in T_{i+1}.

This heuristic function is used over every candidate literal and the literal with a largest value is selected. The algorithm will continue until the generated clauses cover all positive examples.

3 Our Approach

The essential difference between the MI problem and the classical classification problem is in the positive examples. In the classical problem, positive and negative examples are precisely separated, where in the MI problem, positive

instances cannot be specified exactly since positive bags only contain at least one positive instance. FOIL nevertheless evaluates and selects the best literal based on a number of positive and negative instances covered and uncovered. The negative examples can exactly be obtained from negative bags but for the positive examples, they are mixed in the positive bags together with the negative ones. Thus, if the MI data are applied to the original algorithm, it would be more difficult to get the correct concept since the positive examples contain a lot of noises. In order to handle these data, most of MI learning algorithms assume the area in the feature space where instances from different positive bags locating together as the target concept and this is formalised into the measure in DD.

The basic idea of our approach is to evaluate instances from the positive bags by using DD to show the strength of the instance to be positive using a value between 0 and 1. We then modified the heuristic function in FOIL to use the sum of DD values covered instead of the number of positive examples. For negative examples, as they are exactly labelled, we then use the number of negative examples in the same manner as the original function. Therefore, $|T_i^+|$ in formula (3) is changed to the sum of DD of positive examples, but $|T_i^-|$ still remains the same as in the original approach. The modified heurisfic function can be written as follows.

$$DD_s(T) = \sum_{T_i \in T} DD(T_i) \tag{6}$$

$$I(T_i) = -\log_2(DD_s(T_i^+)/(DD_s(T_i^+) + |T_i^-|)) \tag{7}$$

$$Gain(L_i) = DD_s(T_i^{++}) \times (I(T_{i+1}) - I(T_i)) \tag{8}$$

3.1 DD computation

In order to compute DD, the features describing each instance are necessary so that the instances can be separated. However, the first-order representation is so flexible that the feature can be described in several ways using one or more predicates. Therefore, predicates representing each instance have to be specified first.

In this research, the distance between predicates is calculated from the difference between each parameter in the predicates, then, these difference values are combined to the distance by using the Euclidean distance. For example, in the chemical compound data, we treat each atom in a molecule as an instance. The atom may be defined as *atm(compoundid, atomid, elementtype, atomtype, charge)*. The distance between two atoms can be computed by using the difference between parameters. However, a parameter may be discrete or continuous value. In case of continuous value, the difference is computed by substraction.

$$\Delta p = |p_1 - p_2| \tag{9}$$

In case of discrete value, the difference value will be 0 if they are the same value. Otherwise, it will be 1.

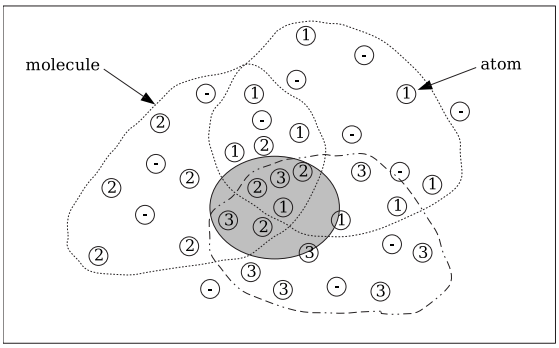


Fig. 1. An example of problem domain for MI data (a molecule represents a bag and an atom is an instance in a bag.)

$$\Delta p = \begin{cases} 0 & \text{if } p_1 = p_2, \\ 1 & \text{otherwise.} \end{cases} \tag{10}$$

Then, the distance between two predicates can be calculated in the same way as formula 2 as

$$\|P_1 - P_2\|^2 = \sum_{p_{1i} \in P_1, p_{2i} \in P_2} (\Delta p)^2 \tag{11}$$

For example, the distance between *atm(m1, a1_1, c, 20, 0.1)* and *atm(m1, a1_2, o, 15, 0.2)* will be calculated from the difference between 'c' and 'o', '20' and '15' (these values are treated as discrete because it is the atom type), and '0.1' and '0.2' that is $1^2 + 1^2 + 0.1^2 = 1.01$. Figure 1 shows an example of problem domain for MI data that a molecule represents a bag and an atom is an instance in a bag. The DD approach tries to evaluate the area that instances from various positive bags locating together and is far from negative instances. From the figure, the shaded area has high DD value. For the chemical compound prediction, this area shows the characteristics of atoms that are common among the positive molecules.

3.2 The Algorithm

From the proposed approach, we examined the heuristic calculation in order to suit the MI data. We then considered modifying the algorithm.

Figure 2 shows the main algorithm used in the proposed system. This algorithm starts by initialising the set *Theory* to null, and the set *Remaining* to the set of positive examples. The algorithm loops to find rules and add each rule found to *Theory* until all positive examples are covered. It can be seen that this main algorithm is the same as FOIL. We modified the heuristic calculation which is in subroutine $FindBestRule$.

- $Theory \leftarrow \emptyset$
- $Remaining \leftarrow Positive(Examples)$
- While not $StopCriterion(Examples, Remaining)$
  - $Rule \leftarrow$ **FindBestRule**$(Examples, Remaining)$
  - $Theory \leftarrow Theory \cup Rule$
  - $Covered \leftarrow Cover(Remaining, Rule)$
  - $Remaining \leftarrow Remaining - Covered$

Fig. 2. The main algorithm.

Subroutine *FindBestRule* is shown in figure 3. As explained above, the DD is applied for calculating a heuristic function. Another problem can nevertheless be considered in this learning approach. When using the DD in counting the number of positive examples covered, there are many cases that the heuristic value may not increase during the searching process (the information gained equals to 0) because there are usually few true positive instances in one positive bag; hence, most of instances from positive bags have the DD value close to 0. This situation makes it difficult to find the best rules using only the hill-climbing approach as in FOIL since there are various candidates with the same heuristic value, aimlessly selecting the candidate may lead to the wrong direction. In order to avoid this problem, the beam search is applied to the proposed system so that the algorithm maintains a set of good candidates instead of selecting of the best candidate at that time. This searching method makes the algorithm able to backtrack to the right direction and finally get to the goal.

FindBestRule(**Examples**, **Remaining**)

- Initialise $Beam$ with an empty rule, R as

$$R(V_1, V_2, V_3, \ldots, V_k) \leftarrow$$

- $R \leftarrow BestClauseInBeam(Beam)$
- While $Cover(R, Negative(Examples))$
  - $Candidates \leftarrow SelectCandidate(Examples, R)$
  - For each C in $Candidates$
    - GenerateTuple($Examples, Tuples$)
    - If C contains new relation Then re-calculate DD.
    - Calculate *heuristic value* for $Tuples$ and attach to C.
  - $UpdateBeam(Candidates, Beam)$.
  - $R \leftarrow BestClauseInBeam(Beam)$

Fig. 3. The algorithm for finding the best literals

Approach	Accuracy
Proposed method	0.82
Progol	0.76[7]
FOIL	0.61[7]

Table 1. Accuracy on the mutagenesis dataset comparing to the other ILP systems.

4 Experiments and Discussion

We conduct experiments on datasets related to chemical structures and activity. The objective of these dataset is to predict characteristics or properties of the chemical molecules which consist of several atoms and bond between atoms. Therefore, the first-order logic would be more suitable for representing this kind of data since it is able to denote relations among atoms comprehensibly. This learning problem can also be considered as multiple-instance problem because each molecule may consist of a lot of atoms but only some connected atoms may effect on the characteristics or properties of the molecule. Therefore, we treat a molecule as a bag that consists of instances as atoms.

4.1 Mutagenesis Prediction Problem

The problem aims to discover rules for testing mutagenicity in nitroaromatic compounds which are often known to be carcinogenic and also cause damage to DNA. These compounds occur in automobile exhaust fumes and are also common intermediates used in chemical industry. The training examples are represented in form of atom and bond structure. 230 compounds were obtained from the standard molecular modelling package QUANTA [6].

- *bond(compound, atom1, atom2, bondtype)*, showing that there is a bond of *bondtype* betweem the atom *atom1* and *atom2* in the *compound*.
- *atom(compound, atom, element, atomtype, charge)*, showing that in the *compound* there is the *atom* that has element *element* of *atomtype* and partial charge *charge*

We conduct the experiment on this dataset and evaluate the results with 10-fold cross validation comparing to the existing ILP systems (FOIL and Progol). Table 1 shows the experimental results on this dataset. Examples of rules generated from the proposed system is shown in figure 4.

4.2 Dopamine Antagonists Activity

This is another dataset that we conducted the experiment on. We used the MDDR database of MDL Inc. This dataset contains 1,364 molecules that describe dopamine antagonist activity with atoms and bonds structure in the similar manner to the mutagenesis dataset in the previous experiment. Dopamine

- active(A) ← atm(A,B,C,D,E), D=95.
  The molecule contains an atom whose type is 95.
- active(A) ← atm(A,B,C,D,E), D=27, E<0.
  The molecule contains an atom whose type is 27 and charge is less than 0.
- active(A) ← atm(A,B,C,D,E), E>=0.816, E<0.823, atm(A,F,G,H,I), I>=0.817.
  The molecule contains two atoms. One has charge value between 0.816 and 0.823. Another one has charge value greater than 0.817.
- active(A) ← atm(A,B,C,D,E), D=27, atm(A,F,G,H,I), H=27, bond(A,B,F,J).
  The molecule contains two atoms. Both atoms are the same type which is 27 and there is a bond between them.

Fig. 4. Examples of rules generated from the proposed system on the mutagenesis dataset.

is a neurotransmitter in the brain that neural signals are transmitted via the interaction between dopamine and proteins known as dopamine receptors. An antagonists are a chemical compound that binds to a receptor, but does not function as a neurotransmitter. It blocks the function of the dopamine molecule. Antagonists for these receptors might be useful for developing schizophrenia drugs. There are four antagonist activities (D1, D2, D3, and D4). In this experiment, we aim to predict these activities by using the background knowledge consisting of two kinds of predicate.

- *bond(compound, bond, atom1, atom2, bondtype, length)*, showing that there is a *bond* of *bondtype* between the atom *atom1* and *atom2* in the *compound* with *length*.
- *atom(compound, atom, element)*, showing that in the *compound* there is the *atom* of *element*.

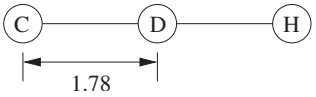
In this dataset, each atom is described by only its element. This would not be enough for computing distances for the DD. Therefore, due to consulting with the domain expert, the number of bonds linked to the atom and the average length of bonds are added for the distance computing. Hence, each atom is represented by three features, element type, number of bond linked, and average length of bonds.

Moreover, as the proposed method can handle only two-class data (only positive or negative), but there are four classes for the dopamine antagonist compounds. Then, hypotheses for each class are learned by one-against-the-rest technique, for instance, learning class D1 by using D1 as positive examples and D2,D3,D4 as negative examples. The example of rules for D1 activity are shown in figure 5.

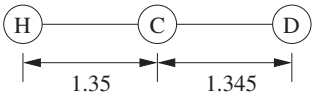
4.3 Discussion

From the experiments, we found that the proposed method generates more accurate rules when comparing to Progol and FOIL. Example of rules in figure

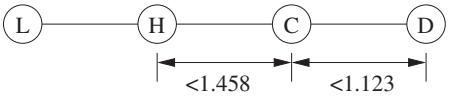
– d1(A) ← bond(A,B,C,D,E,F), F=1.45, atom(A,C,G), G=c.
This rule shows that a compound is class D1 if it contains a bond of length 1.45 and one of atom linked with this bond is a carbon atom.
– d1(A) ← bond(A,B,C,D,E,F), F=1.78, bond(A,G,H,D,I,J).
This rule shows that a compound is class D1 if it contains two bonds which link three atoms together as shown in the figure below where each node represents an atom and each edge represents a bond.



– d1(A) ← bond(A,B,C,D,E,F), F=1.345, bond(A,G,H,C,I,J), J=1.35.
This rule shows that a compound is class D1 if it contains two bonds which link three atoms together as shown in the figure below.



– d1(A) ← bond(A,B,C,D,E,F), F<1.123, bond(A,G,C,H,I,J), J<1.458, bond(A,K,H,L,M,N).
This rule shows that a compound is class D1 if it contains three bonds and the relation between bonds and their length is shown in the figure below.



These rules are needed to be evaluated by the domain expert so that the predicate or background knowledge can be improved in order to discover the interesting knowledge.

Fig. 5. Example of rules generated on the dopamine antagonist analysis.

4 also shows the benefit of the proposed method which produces hypotheses in the first-order represenation, for instance, rule (3) and (4) consist of two atoms or a bond between atoms. These kinds of rule cannot be represented using the propositional logic. Moreover, when considering the knowledge discovery, only properties of one atom may not be good enough for describing the characteristic of molecule. Therefore, in this classification the first-order logic would be more suitable than the propositional logic. We will also try to improve the heuristic function or the search technique in order to generate hypotheses that incorporate a group of atoms and bonds between atoms.

5 Conclusions

We have presented the extension of FOIL for better handling multiple-instance data by using Diverse Density to evaluate tuples from positive bags. This evaluation is similar to setting the instances with different sets of feature which is actually the benefit of using the first-order representation. The experimental results show that our approach learns from the real-world problem better than Progol and FOIL.

In the current system, only atoms are considered as instances in bags. However, each bond itself may also be considered as an instance. Therefore, we will try to improve the proposed system in order to handle various kinds of instances in one bag in the future works. We also plan to improve the background knowledge due to the discussion with the domain expert.

References

1. Dietterich, T.G., Lathrop, R.H., Lazano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence **89** (1997) 31–71
2. Maron, O., Lazano-Pérez, T.: A framework for multiple-instance learning. Neural Information Processing Systems 10 (1998) Available at ftp://ftp.ai.mit.edu/pub/users/oded/papers/NIPS97.ps.Z.
3. Chevaleyre, Y., Zucker, J.D.: A framework for learning rules from multiple instance data. In: Proceedings of the 12th European Conference on Machine Learning, Freiburg, Germany (2001) 49–60
4. Zhang, Q., Goldman, S.A.: Em-dd: An improved multiple-instance learning technique. In Dietterich, T.G., Becker, S., Ghahramani, Z., eds.: Advances in Neural Information Processing Systems 14, Cambridge, MA, MIT Press (2002)
5. Quinlan, J.R.: Learning logical definitions from relations. Machine Learning **5** (1990) 239–266
6. Srinivasan, A., Muggleton, S., King, R.D., Sternberg, M.: Mutagenesis: ILP experiments in a non-determinate biological domain. In Wrobel, S., ed.: Proceedings of the 4th International Workshop on Inductive Logic Programming. Volume 237., Gesellschaft für Mathematik und Datenverarbeitung MBH (1994) 217–232
7. Chevaleyre, Y., Bredeche, N., Zucker, J.D.: Learning rules from multiple instance data : Issues and algorithms. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU02). (2002)

Development of a 3D Motif Dictionary System for Protein Structure Mining

Hiroaki Kato, Hiroyuki Miyata, Naohiro Uchimura,
Yoshimasa Takahashi, and Hidetsugu Abe

Department of Knowledge-based Information Engineering,
Toyohashi University of Technology,
1-1 Hibarigaoka Tempaku-cho, Toyohashi, 441-8580 Japan
{hiro, miyata, utimura}@cilab.tutkie.tut.ac.jp
taka@mis.tutkie.tut.ac.jp, abe@cilab.tutkie.tut.ac.jp

Abstract. This paper describes a three-dimensional (3D) protein motif dictionary system that is closely related to the PROSITE sequence motif database. Because there are many different 3D motif patterns but having a particular PROSITE sequence pattern, we have investigated the approaches for quantitative comparison and clustering such 3D structure segments. For a pair of 3D structure segments, the dissimilarity value is defined as the root mean squares of inter-residue distances. A conformational pattern clustering is employed for grouping the 3D patterns on the basis of the dissimilarity matrix. Some additional knowledge information described in PROSITE are also referred to refine the result. The 3D motif dictionary was constructed using all the data set of PDB. The graphical user interface for using the dictionary is also developed. The usefulness of the additional approach for the 3D motif dictionary is also discussed with an illustrative example.

1 Introduction

With the rapidly increasing number of proteins of which three-dimensional (3D) structures are identified, the protein structure database is one of the key elements in many attempts being made to derive the knowledge of structure-function relationships of proteins [1]. However, it is almost impossible to search manually 3D local structural features called motifs within protein full structures because of increasing number and their structural complexity. For the reason, computerized methods are required for a systematic searching of the 3D features of proteins in such a database. For knowledge discovery based on 3D structural feature analysis of proteins, we have investigated to construct a dictionary of 3D partial structures that are conformed to the motifs in PROSITE [2], and systematic extensive analysis of 3D protein structures based on the 3D motif dictionary established.

As shown in Table 1, each sequence motif pattern in PROSITE database is described with regular expression. In our preceding work, using the sequence motif patterns, the corresponding sites of them were extensively explored on the 3D structures of proteins taken from the Protein Data Bank (PDB [3]) database. The segments

found by the searching are collected for constructing a 3D motif database [4]. However, the results of structural feature analysis showed that a particular PROSITE sequence motif pattern corresponds to many different 3D protein segments.

In this paper, we have investigated the approaches for quantitative comparison and clustering such 3D structure segments. A dictionary that contains the representative 3D geometrical pattern for each sequence motif is constructed using all the data sets in PDB (Fig.1). The WWW-based user interface programs are also developed for managing and using the motif dictionary.

Table 1. Example of sequence motif representation in PROSITE [2]

Motif	Pattern
Kringle	[FY]-C-R-N-P-[DNR].
Zinc fingerC2H2	C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H.
EF-hand	D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW].

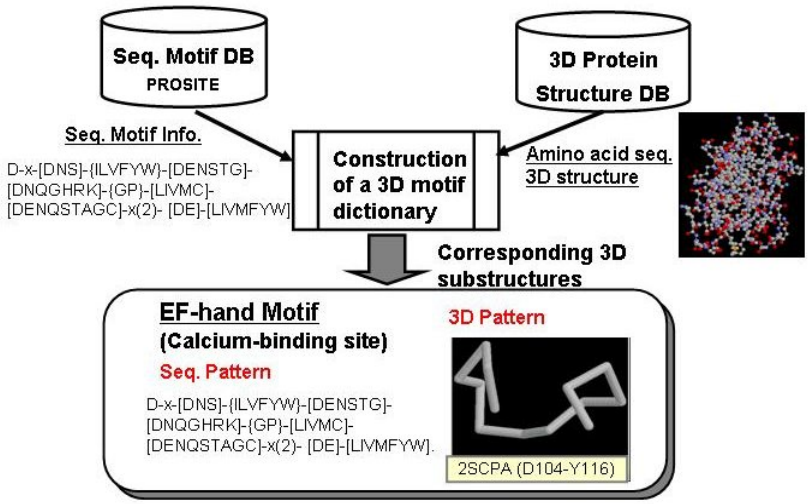


Fig. 1. Basic concept of 3D motif dictionary of proteins in the present work

2 Methods

2.1 Comparison and Clustering of 3D Geometric Patterns

In the present work, the 3D structure segment of a protein is represented by a set of amino acid residues, and their 3D coordinates are approximated at those of α-carbon atoms of the residues. For every pair of 3D segments, a value of dissimilarity is computed. The root-mean-square value of Euclidean distances between the corresponding α-carbons of segments to be compared is defined as a measurement of dissimilarity.

$$RMSD(A,B) = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(d_A(i,j) - d_B(i,j))^2}{n^2}} \quad \textbf{(1)}$$

Where n is the number of amino acid residues of the 3D structure segment, $d_A(i, j)$ and $d_B(i, j)$ are the Euclidean distances between i-th and j-th residues of segment A and B, respectively.

A dissimilarity matrix is made for the 3D segments that have a particular common sequence motif pattern. Using the matrix, the segments are clustered into several structural clusters with a certain threshold value. Then, for each cluster, a representative 3D feature pattern was determined on the basis of minimum variance of the distances between the representative and others. Here, all possible cluster patterns can be enumerated by using every element of the matrix as the threshold value. We defined a difference of the threshold values for a pair of adjacent cluster patterns as the priority value for the cluster (Fig.2).

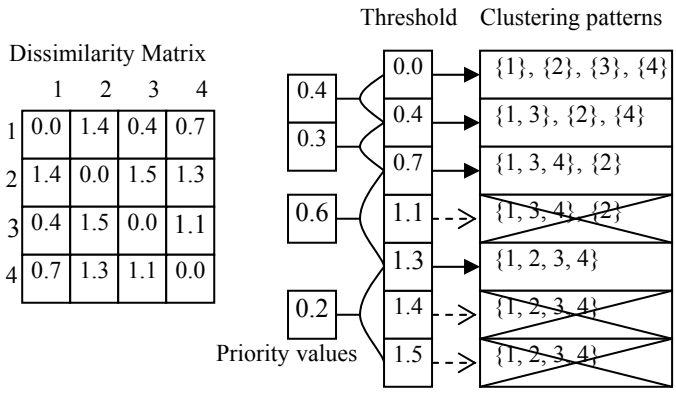


Fig. 2. Enumerated cluster patterns and their priority value. The patterns which have identical clustering result to the neighbors were removed for the candidate

2.2 Refinement of the Cluster Patterns

Some additional knowledge information described in PROSITE are also used to refine the result of cluster patterns. Especially, the DR (Database Reference) records contain the pointers to entries of a protein sequence database, SWISS-PROT [5], and flags that indicate the said motif is found or not. For example, the flag "T" in DR record means for a true positive entry, and "F" for a false positive: a sequence which does not belong to the set in consideration, respectively [2]. If this information is available, we have referred it and filtered the cluster patterns by using the cross reference table of SWISS-PROT to PDB [5]. That is, the cluster patterns which satisfied the following condition are selected: all 3D segments corresponding to the sequence with flag "T" ascribed to one cluster, and the segment with flag "F" to another cluster.

3 Results and Discussion

3.1 Construction of 3D Motif Dictionary

We have prepared a target database that contains 25,980 protein structures (chains) taken from PDB Rel.102. For 1,331 sequence patterns that are available on the PROSITE Rel.17.01, the 3D features are extensively explored on the target database. As the result, 907 patterns were found in the target database, and they were automatically clustered on the basis of dissimilarity of their 3D geometrical patterns mentioned above. If DR records in PROSITE are not available, or there are two or more candidate cluster patterns remained in the filtering procedure, the pattern which has the best priority value is temporarily selected. The cluster which contained "true" motif segments, or the largest cluster if there is no such information, is assumed as the major pattern for this motif. Then, a representative 3D segment pattern for major cluster is also automatically registered in 3D motif dictionary.

3.2 Graphical User Interface for Using 3D Motif Dictionary

For every PROSITE motif, 3D structure files for the corresponding motif segments, a list of enumerated cluster pattern, a representative 3D pattern, and other related information are registered in the dictionary. We also developed the graphical user interface program for this motif dictionary in the present work. It is implemented as a CGI program using Perl language. For displaying 3D structural information of proteins, MDL's Chime plug-in [6] is required, and RasMol scripts are used for instruction of the view models. The user can easily browse a list of motif with the representative 3D pattern, a table of other corresponding segments for each motif, and alternative cluster patterns. In the database administrator's mode, you may choice more reasonable cluster pattern if necessary. Then, the corresponding representative 3D pattern is automatically modified. For each 3D segment stored in this dictionary, a

whole protein structure and the corresponding motif sites are also interactively displayed (Fig. 3).

3.3 3D Pattern Searching Using the Dictionary

Alternatively, in the previous work, the authors reported a computer program for 3D structural feature searching, which allows us to identify all occurrences of a user-defined 3D query pattern consisting not only of chain-based peptide segments but

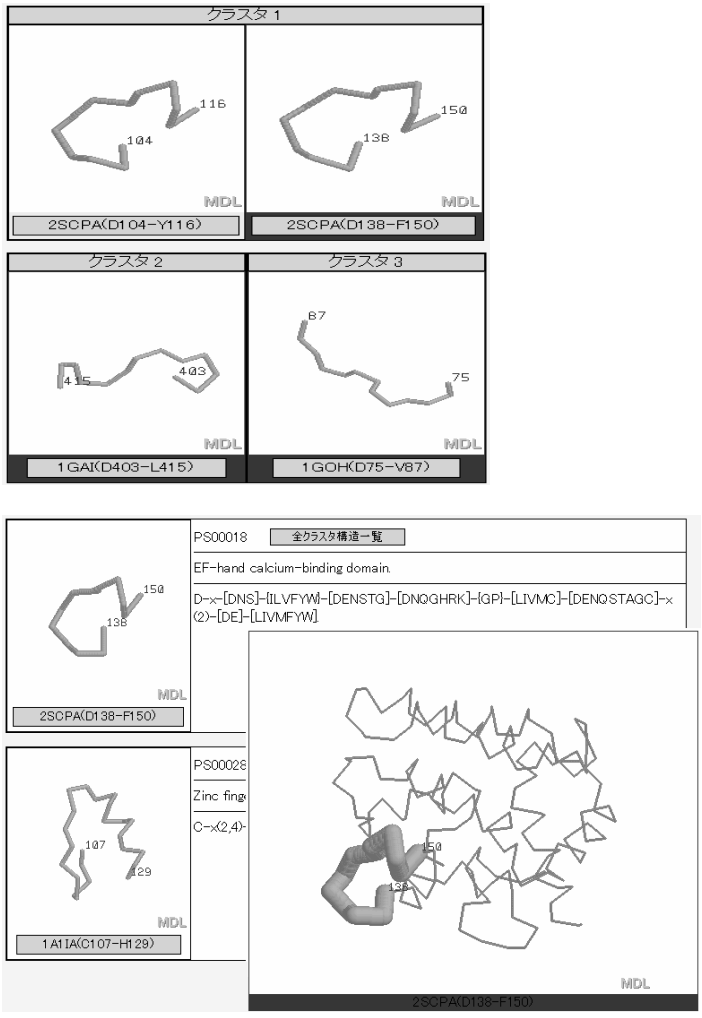


Fig. 3. Some snapshots of WWW-based user interface for 3D motif dictionary

also of a set of disconnected amino acid residues [7]. More extensive analyses of 3D structural features of proteins can be also executed by using our program with the representative patterns registered in the 3D motif dictionary. For example, 3D pattern searching was carried out for the representative pattern of EF-hand motif (PS00018). Several sites that are different from the sequence search were identified. The pattern of 1B47A (D229-F241) shows one of them. It was realized that the site is true for the EF-hand but that has a different residue in the sequence pattern reported in the PROSITE [8]. The result suggests that the present approach is quite useful for 3D structural feature analysis of proteins.

4 Conclusion and Future Works

We have developed 3D protein motif dictionary system based on PROSITE database. A representative 3D geometrical pattern for each sequence motif was automatically identified and stored into the dictionary with other structural information. The authors now are doing investigation on the development of filtering tool to get alternative features of protein too. We believe that the dictionary described here will be more and more important in post genomic research to understand structure-function relationships of proteins.

Acknowledgement

This work was supported by a Grant-in-Aid for Scientific Research on Priority Areas 'Active Mining', from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

References

1. Branden, C., Tooze, J.: Introduction to Protein Structure, Garland Publishing, New York (1991).
2. Bairoch, A.: PROSITE: a Dictionary of Sites and Patterns in Proteins, *Nucleic Acids Res.*, **19** (1991) 2241-2245
3. Helen, M., et al.: The Protein Data Bank, *Nucleic Acids Res.*, **28** (2000) 235-242
4. Kato, H. et al.: Construction of a Three- Dimensional Motif Dictionary for Protein Structural Data Mining, *Trans. Jpn. Soc. Artificial Intelligence*, **17** (2002) 608-613
5. Boeckmann, B., et al.: The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res.*, **31** (2003) 365-370
6. MDL Inc., http://www.mdli.com/
7. Kato, H., Takahashi, Y.: Automated Identification of Three-Dimensional Common Structural Features of Proteins, *J. Chem. Software*, **7** (2001) 161-170
8. Kretsinger, R. H.: Structure and Evolution of Calcium-Modulated Proteins, *CRC Crit. Rev. Biochem.*, **8** (1980) 119-174

Spiral Mining using Attributes from 3D Molecular Structures

Takashi OKADA, Masumi YAMAKAWA and Hirotaka NIITSUMA

Department of Informatics, Kwansei Gakuin University 2-1 Gakuen, Sanda-shi, Hyogo, 669-1337 Japan
E-mail: {okada-office, abz81166}@ksc.kwansei.ac.jp,

Abstract. Active responses from experts play an essential role in the knowledge discovery of SAR (structure activity relationships) from drug data. Experts often think of hypotheses, and they want to reflect these ideas to the attribute generation and selection processes. Authors have analyzed SAR of dopamine antagonists using the cascade model. In this paper, we generated attributes indicating the presence of hydrogen-bonded fragments from 3D coordinates of molecules, which were suggested by experts. The selection of attributes by experts has been shown to be useful in obtaining rules that lead to valuable knowledge.

1 Introduction

The importance of SAR (structure-activity relationship) studies relating chemical structures and biological activity is well recognized. Early studies used statistical techniques, and concentrated on establishing quantitative structure activity relationships involving compounds sharing a common skeleton. However, it is more natural to treat a variety of structures together, and to identify the characteristic substructures responsible for a given biological activity. Recent innovations in high throughput screening technology have produced vast amounts of SAR data, and the demand for a new data mining method to facilitate drug development has increased.

The author has already analyzed SAR's [1, 2] using the cascade model that we developed [3]. Later, I pointed out the importance of the "datascape survey" in the mining process in order to obtain valuable knowledge. We added several functions to the mining software of the cascade model (DISCAS) to facilitate the datascape survey [4, 5].

This new method was recently used in a preliminary study of the SAR for the antagonist activity of dopamine receptors [6]. The resulting interpretations of the rules were highly regarded by experts of drug design. However, the interpretation process of rules employed the visual inspection of supporting chemical structures as an essential step, and a user had to be very careful so that he/she did not miss characteristic substructures.

Fruitful mining results will never be obtained unless an active user's response is not expected. This paper reports an attempt to reflect expert's ideas to the attribute generation and selection processes. Attributes indicating the hydrogen-bonded

fragments are created using 3D coordinates of molecular structures. The attribute selection process provides a framework for active user's responses. The next section briefly describes the aims of mining as well as the basic introduction to the mining method employed. The attribute generation and selection methods are described in Section 3. Typical rules and their interpretations are discussed in Section 4.

2 Aims and Basic Methods

2.1 Aims and Data Source for the Dopamine Antagonists Analysis

Dopamine is a neurotransmitter in the brain. Neural signals are transmitted via the interaction between dopamine and proteins known as dopamine receptors. There are five different receptor proteins, D1 – D5, each of which has a different biological function. Their amino acid sequences are known, but their 3D structures are not yet established.

Certain chemicals act as antagonists for these receptors. An antagonist binds to a receptor, but does not function as a neurotransmitter. Therefore, it blocks the function of the dopamine molecule. Antagonists for these receptors might be used to treat schizophrenic patients. The structural characterization of these antagonists is an important problem in developing new schizophrenia drugs.

We used the MDDR database of MDL Inc. as the data source. It contains 1,349 records that describe dopamine (D1, D2, D3, and D4) antagonist activity. Some of the compounds affected multiple receptors. The problem is to discover the structural characteristics responsible for each type of antagonist activity.

2.2 The Cascade Model

The cascade model can be considered an extension of association rule mining [3]. The method creates an itemset lattice in which an [attribute: value] pair is used as an item to constitute itemsets. Links in the lattice are selected and interpreted as rules. That is, we observe the distribution of the RHS (right hand side) attribute values along all links, and if a distinct change in the distribution appears along some link, then we focus on the two terminal nodes of the link. Consider that the itemset at the upper end of a link is [A: y] and item [B: n] is added along the link. If a marked activity change occurs along this link, we can write the rule:

```
Cases: 200 ==> 50 BSS=12.5
IF [B: n] added on [A: y]
THEN [Activity]:    .80 .20  ==>  .30 .70 (y n)
THEN [C]:           .50 .50  ==>  .94 .06 (y n)
Ridge [A: n]:   .70 .30/100  ==>  .70 .30/50 (y n)
```

where the added item [B: n] is the main condition of the rule, and the items at the upper end of the link ([A: y]) are considered preconditions. The main condition

changes the ratio of the active compounds from 0.8 to 0.3, while the number of supporting instances decreases from 200 to 50. *BSS* means the between-groups sum of squares, which is derived from the decomposition of the sum of squares for a categorical variable. Its value can be used as a measure of the strength of a rule. The second "THEN" clause indicates that the distribution of the values of attribute [C] also changes sharply with the application of the main condition. This description is called the *collateral correlation*.

2.3 Functions for the Datascape Survey

New facilities introduced to DISCAS (mining software for the cascade model) consist of three points. Decreasing the number of resulting rules is the main subject of the first two points [4]. A rule candidate link found in the lattice is first greedily optimized in order to give the rule with the local maximum *BSS* value, changing the main and preconditions. Let us consider two candidate links, (M added on P) and (M added on P'). Here, their main conditions, M, are the same. If the difference between preconditions P and P' is the presence/absence of one precondition clause, the rules starting from these links converge on the same rule expression, and it is useful for decreasing the number of resulting rules.

The second point is the facility to organize rules into principal and relative rules. In the association rule system, a pair of rules, R and R', are always considered independent entities, even if their supporting instances overlap completely. We think that these rules show two different aspects of a single phenomenon. Therefore, a group of rules sharing a considerable amount of supporting instances are expressed as a principal rule with the largest *BSS* value and its relative rules. This function is useful for decreasing the number of principal rules to be inspected, and to indicate the relationships among rules.

The last point is to provide ridge information of a rule [5]. The last line of the aforementioned rule shows ridge information. This example describes [A: n], the ridge region detected, and the change in the distribution of "Activity" in this region. Compared to the large change in the activity distribution for the instances with [A: y], the distribution does not change on this ridge. This means that the *BSS* value decreases sharply if we expand the rule region to include this ridge region. This ridge information is expected to guide the survey of the datascape.

3 Attribute Generation and Selection

3.1 Basic Scheme

We used two kinds of explanation attributes generated from the structural formulae of chemical compounds. The first group consists of four physicochemical estimates: the HOMO and LUMO energy levels, the dipole moment, and LogP. The first three values were estimated by the molecular mechanics and molecular orbital calculations

using MM-AM1-Geo method provided by *Cache*. LogP values were calculated by ClogP program in *Chemoffice*.

The other group is the presence/absence of various structural fragments. Obviously, the number of all possible fragments is too large. We generated linear fragments with lengths shorter than 10. One of the terminal atoms of a fragment was restricted to be a heteroatom or a carbon constituting a double or triple bond.

Linear fragments were expressed by constituent elements and bond types. The number of coordinating atoms and the presence/absence of attached hydrogens are added to the terminal and its adjacent atoms. C3H:C3-C-N-C3=O1 is a sample expression, where "C3H" means a three-coordinated carbon atom with at least one hydrogen atom attached, and ":" denotes an aromatic bond.

Number of fragments generated from dopamine antagonist data was more than 120,000, but most of them are useless as they appear only a few times among 1349 molecules. On the other hand, the upper limit of the attributes is about 150 in the current implementation of DISCAS. Therefore, we selected 73 fragments, of which probability of appearance is in the range: 0.15 – 0.85.

3.2 Hydrogen-bonded Fragments

When we visualize chemical structures that satisfy rule conditions, we sometimes see a group of compounds that might be characterized by an intramolecular hydrogen-bond, XH…Y, where X and Y are usually oxygen or nitrogen. However, the fragment generation scheme above-mentioned utilizes only the graph topology of the structure, and we cannot recognize the hydrogen-bond.

The results of MM-AM1-Geo calculation used for the estimation of physicochemical properties provide 3D coordinates of atoms. Therefore, we can detect the existence of a hydrogen-bond using 3D coordinates. We judged the existence of a hydrogen-bond, XH…Y, when the following conditions were satisfied.

1. Atom X is O, N, S or 4 coordinated C with at least one hydrogen atom.
2. Atom Y is O, N, S, F, Cl or Br.
3. The distance between X and Y is less than 3.7 Å if Y is O, N or F; and it is less than 4.2 Å otherwise.
4. Structural formula does not contain fragments X-Y or X-Z-Y, where any bond type will do.

When these conditions are satisfied, we generate fragments: Xh.Y, V-Xh.Y, Xh.Y-W, and V-Xh.Y-W, where "h." denotes a hydrogen-bond, and neighboring atoms V and W are included. Other notations follow the basic scheme.

Application to the dopamine antagonists dataset resulted in 431 fragments, but the probability of the most frequent fragment was less than 0.1. Therefore, all hydrogen-bonded fragments are not employed in the standard mining process.

3.3 Attributes Selection and Spiral Mining

When experts think of a characteristic substructure for the appearance of some biological activity, it can be expressed by few linear fragments. Some fragments might lead to clear and strong rules even if its probability of appearance in the data set is out of the specified range: 0.15 – 0.85.

We provided a mechanism to add specified fragments as the attribute used in the mining. Consequently, a user can repeat the following steps, in order to discover better characteristic substructures.

1. Prepare fragment attributes by the basic scheme.
2. Compute rules.
3. Read resulting rules and make hypotheses by the language of chemistry.
4. Confirm the hypothesis by browsing supporting structural formulae.
5. If one notices a characteristic fragment that does not appear in the rule, add the fragment as an attribute. Go to step 2.
6. Repeat until satisfactory results are obtained.

Since experts can put his/her ideas in the mining process, adding fragments and reading rules are now an interesting exploration. This spiral mining process is not limited to the incorporation of hydrogen-bonded fragments, but it is applicable to all kinds of fragments.

4 Results and Discussion

For the calculations with the DISCAS ver.3 software the parameters were set at *minsup*=0.01, *thres*=0.1, *thr-BSS*=0.01, *min-rlv*=0.7. These parameters are defined elsewhere [3, 4, 5]. We added 32 fragments after reading rules and inspecting chemical structures. They consist of 27 hydrogen-bonded fragments with probability_of _appearance > 0.02, and 5 fragments (N3-C3:C3-O2, N3H-C3:C3-O2, N3-C3:C3-O2H, N3H-C3:C3-O2H, O1=S4). The inspection process is not complete yet, but we can depict two examples that show the effectiveness of the current method.

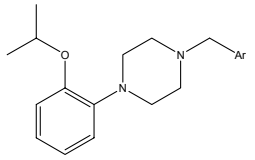
4.1 The D2 Antagonist Activity

The analysis of this activity was hard in the former study. That is, the strongest rule indicating active compounds takes the following form when we use the basic scheme for the attribute selection.

```
IF [C4H-C4H-O2: y]  added on   [ ]
THEN D2AN:       0.32 0.68 ==> 0.62 0.38(on off)
THEN C3-O2:      0.42 0.58 ==> 0.89 0.11(y n)
THEN C3H:C3-O2:  0.33 0.67 ==> 0.70 0.30(y)
Ridge [C3H:C3H:C:C3-N3: y] D2AN: 0.49 0.51/ 246 --> 0.92 0.08 / 71
```

There appear no preconditions, and the main condition shows that an oxygen atom bonded to alkyl carbon is important. However, this finding is so different from the common sense of experts, and it will never be accepted as a useful suggestion. In fact, the ratio of active compounds is only 62%. Collateral correlations suggest that the oxygen atom constitutes aromatic ethers, and the ridge information indicates the relevance of aromatic amines. But, it has been difficult even for a expert to make a reasonable hypothesis.

Experts found a group of compounds sharing the skeleton shown below, when they browse the supporting structures. So, they added fragments consisting of two aromatic carbons bonded to N3 and O2. This addition did not change the strongest rule, but there appeared a new relative rule shown below.




```
IF [N3-C3:C3-O2: y]  added on   [ ]
THEN D2AN:       0.31 0.69 ==> 0.83 0.17(on off)
THEN HOMO:  0.16 0.51 0.33 ==> 0.00 0.19 0.81 (low medium high)
THEN C3H:C3-N-C-C4H-N3: 0.24 0.76 ==> 0.83 0.17(y n)
```

This rule has a higher accuracy and it explains about 20% of active compounds. The tendency observed in HOMO value also gives us a useful insight. However, the collateral correlation on the last line shows that most compounds supporting this rule have the skeleton show above. Therefore, we cannot exclude the possibility that other part of this skeleton is responsible for the activity. Further inspection is necessary to reach satisfactory hypotheses.

4.2 The D3 Antagonist Activity

The analysis of this activity is also complex, because there appear more than 10 principal rules. We suggested 5 characteristic substructures in the former study. The strongest rule leading to this activity has the following form.

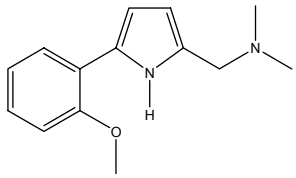
```
IF [O1: y] added on [C3-N3: n] [C3=O1: n]
THEN D3AN:       0.79 0.21 ==> 0.06 0.94(off on)
THEN C3:N2:      0.33 0.67 ==> 0.01 0.99(y n)
THEN N3H:        0.58 0.42 ==> 0.91 0.09(y n)
THEN O2:         0.45 0.55 ==> 0.76 0.24(y n)
THEN N3Hh.O2:    0.09 0.91 ==> 0.54 0.46(y n)
```

The main condition of this principal rule is an oxygen atom, and the preconditions employed are the absence of two short fragments. Therefore, its interpretation is very difficult. After the inclusion of hydrogen-bonded fragments, there appeared the last line in the collateral correlations, where a steep increase of N3Hh.O2 is observed. The relevance of this fragment was confirmed by the appearance of a relative rule shown below.

```
IF [N3Hh.O2: y] added on   [C3-N3H: n]
THEN D3AN: 0.88 0.12 ==> 0.05 0.95(off on)
```

In fact, this rule accounts for about half of the active compounds supported by the principal rule. Visual inspection of the supporting structures has shown that the following skeleton leads to this activity. We have to note that a hydrogen-bond itself is not responsible for the activity.

Another advantage of adding this hydrogen-bonded fragment is illustrated by the



next principal rule.

```
IF [C4H-N3H-C3=O1: y]
added on [LUMO: 0 - 2] [HOMO: 0 - 1] [N3Hh.O2: n]
THEN D3AN: 0.77 0.23 ==> 0.27 0.73(off on)
```

Here, the absence of the fragment appears as a precondition. Since the compounds with this skeleton are excluded, most compounds at the upper node are inactive, giving a higher *BSS* value to the rule. After we recognize the above skeleton as active compounds, this type of rule expression is easy to understand the existence of distinct lead structures for the activity.

5 Conclusion

The proposed mining process has succeeded to evoke active responses from experts. They can put their ideas in the mining task, and step up the mining spiral by themselves. Now, experts can make rough hypotheses by reading rules. Browsing the supporting structures is still necessary. However, they do not need to be nervous when they inspect structures, since they can add many fragments and judge their importance by the resulting rules.

We have to make a note on the interpretation of resulting rules after adding attributes. If an expert adds some fragments that were found in the preliminary step of analysis, they will often appear as a rule condition. But the appearance of a rule is no guarantee of truth. The fragment might be a part of large skeleton shared by the supporting structures. He/she has to check the collateral correlations carefully. Visual inspection of structures is also necessary before he/she reaches final hypotheses.

The comprehensive analysis of ligands for dopamine receptor proteins are now under progress using the proposed system. They include not only discriminations of antagonists, but also those among agonists. Also under investigation are factors that distinguish antagonists and agonists. The results will be a model work in the field of qualitative SAR analysis.

Acknowledgements

The authors wish to thank Ms. Naomi Kamiguchi of the Takeda Chemical Industries for her valuable comments.

References

1 Okada, T.: Discovery of Structure Activity Relationships using the Cascade Model: the Mutagenicity of Aromatic Nitro Compounds. J. Computer Aided Chemistry 2 (2001) 79-86
2 Okada, T.: Characteristic Substructures and Properties in Chemical Carcinogens Studied by the Cascade Model. Bioinformatics 19 (2003) 1208-1215
3 Okada, T.: Efficient Detection of Local Interactions in the Cascade Model. In: Terano, T. et al (eds.) Knowledge Discovery and Data Mining PAKDD-2000. LNAI 1805, Springer-Verlag (2000) 193-203.
4 Okada, T.: Datascape Survey using the Cascade Model. In: Satoh, K. et al. (eds.) Discovery Science 2002. LNCS 2534, Springer-Verlag (2002) 233-246
5 Okada, T.: Topographical Expression of a Rule for Active Mining. In: Motoda, H. (ed.) Active Mining. IOS Press, (2002) 247-257
6 Okada, T., Yamakawa, M.: Mining Characteristics of Lead Compounds using the Cascade Model (in Japanese). Proc. 30th Symposium on Structure-Activity Relationships (2002) 49-52

Architecture of Spatial Data Warehouse for Traffic Management

Hiroyuki Kawano[1] and Eiji Hato[2]

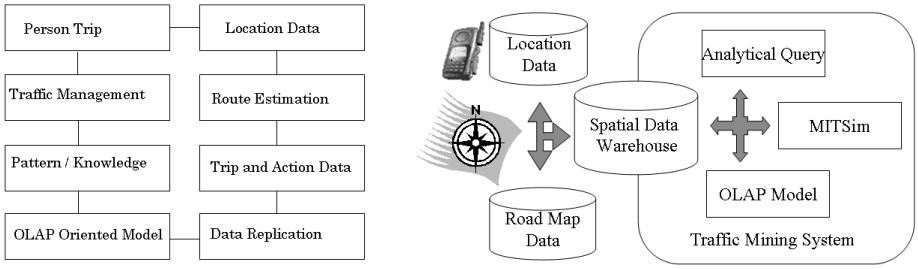
[1] Department of Systems Science, Graduate School of Informatics, Kyoto University, Yoshida Hommachi, Sakyo-ku, Kyoto, 606–8501 Japan
[2] Urban Environmental Engineering, Faculty of Engineering, Ehime University, Bunkyo-cho 3, Matsuyama-shi, 790–8577 Japan

Abstract. Recently, huge volume of spatial and geographic data are stored into the database systems by using GIS technologies and various location services. Based on the long term observation of person trip data, we can derive patterns of person trip data and discover trends of actions by executing spatial mining queries effectively. In order to improve the quality of traffic management, traffic planning, marketing and so on, we outline the architecture of spatial data warehouses for traffic data analysis. Firstly, we discuss some fundamental problems in order to achieve a data warehouse for person trip data analysis. From the view point of traffic engineering, we introduce our proposed route estimation method and advanced spatial queries based on the techniques of temporal spatial indices. Secondly, in order to analyze the characteristics of person trip sequences, we propose the OLAP (On-Line Analytical Processing) oriented spatial temporal data model. Our $\sum$-tree data structure is based on the techniques of data cube and 3DR-tree index. Finally, we evaluate the performance of our proposed architectures and temporal spatial data model by observing actual positioning data, which is collected by location service in Japan.

1 Introduction

In recent years, application systems of GIS (Geographic Information System) and spatial databases are becoming popular[13], and we have to operate and analyze the huge volume of spatial temporal data in location service systems based on GPS (Global Positioning System) and PHS (Personal Handy-phone System).

Furthermore, in the research fields of data mining, a lot of spatial data mining algorithms to derive patterns and discover knowledge in the huge volume of databases are proposed[8, 6, 1, 9]. For example, in order to make clusters effectively, clustering algorithms make full use of the spatial characteristics, such as density, continuity and so on. In our previous researches[3], we also proposed effective clustering algorithms based on the spatial index technologies[11, 12], such as R-Tree, R*-Tree, PR-Quadtree and others.



(a) Analytical processes for spatial temporal data.

(b) Spatial data warehouse and traffic mining system.

Fig. 1. Processes and Architecture for Person Trip Data Analysis.

Therefore, in this paper, we outline an architecture of spatial data warehouse for traffic management/control, traffic planning and marketing, in order to analyze the characteristics of traffic flows based on location services.

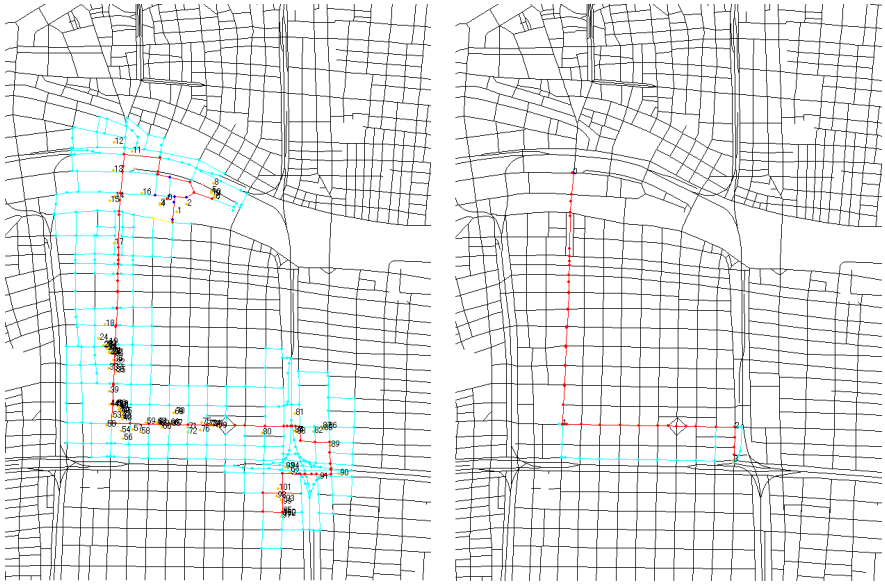
In Section 2, we point out several fundamental problems[3], which must be solved before constructing the spatial data warehouses. In Section 3, in order to derive trip routes from actual discrete positioning data effectively, we introduce the route estimation algorithm[4]. In Section 4, in order to execute large-scale traffic simulation, we apply data replication algorithm and estimate more accurate traffic parameters. We also discuss the performance of our proposed method by using actual positioning data by PHS location service in Sapporo city. In Section 5, from the view points of spatial data mining, we discuss typical OLAP queries for traffic mining, and we propose spatial temporal data structures. Finally, we conclude in Section 6.

2 Spatial Data warehouse for Traffic Management

In order to construct data warehouse for traffic management systems, firstly we have to integrate the technologies of GIS, spatial database and location services. In this section, we point out several problems of geographical information systems and location services.

2.1 Architecture of spatial data warehouse for traffic management

We introduce the analytical processes for person trip data in Fig.1(a) and we describe the data warehouses and related processing modules for person trip data analysis in Fig.1(b). Firstly, by using location services, we observe data of "*person trip*" and translate from raw data to "*location data*". We estimate a trip route by using discrete location data and map data for a specific person trip. The



(a) Positioning sequences and exact route.

(b) Primary estimated route and alternative route.

Fig. 2. Comparison between exact route and estimated route.

sequence of "*trip and action data*" is derived from data of "*route estimation*" and action attribute values. If we have a sufficient volume of person trip data, it may be possible to analyze "*OLAP oriented model*" directly, and we can derive "*pattern and knowledge*" for "*traffic management*" from data. However, due to the problems of privacy and location observing costs, it is very difficult to capture sufficient numbers of person trip data in a target area. Therefore, we apply "*data replication*" algorithm to original observing trip data, then we execute simulation tools (MITSim) by using cloning data of actual person trip.

2.2 Basic problems of GIS and location service

Integration of different types of GIS:

There are so many geographic information systems, which are composed of spatial databases and advanced analytical tools. However, it is not so easy to integrate different types of information systems in order to develop the spatial data warehouse for traffic management. Because, we have to integrate various spatial data with different formats with different accuracy.

Clearing warehouses of geographic data:

Clearing warehouses and common spatial data formats, which are sometimes described in XML, are becoming very useful in order to exchange different types of spatial and non-spatial attribute values. However, at present, it is very hard to integrate schemata, attributes and many other characteristic values. For example, in order to display the position on a map, we have to handle several different spatial coordinates such as WGS-84, ITRF(International Terrestrial Reference Frame) and others.

Advanced trip monitoring systems:

We evaluated the measurement errors by using mobile GPS terminals[2]. But it is not sufficient to determine the location by using PHS type location services. The major error factor of the PHS service seems to be caused by the reflection of buildings and constructions. Of course, the error is becoming smaller by GPS and pseudolites gradually. Therefore, we have to pay attention to advanced location services[$Note1$]. [3]

3 Estimation of Person Trip Route

In this section, we focus on the estimation of person trip route, and we examine the accuracy and validity of our proposed algorithm by using actual PHS location service.

After integrating the numerical map, various geographical and spatial attributes, such as nodes of crossroads, road arcs, directions and so on, spatial objects are stored into the spatial data warehouse. In this experiment, we utilized a numerical map of a restricted area. We also analyzed the sequences of positioning data during 30 minutes with 15 seconds interval of location service.

Fig.2 (a) and (b) show the comparison with the actual sequences of PHS positioning data and the estimated person trip routes that are derived by our Quad-tree based estimation algorithm[4].

In Fig.2 (a), the numbers show the sequences of positioning data provided by PHS location service, pale -blue- lines show the roads that should be searched, and the red -or black- line with points shows the exact person trip route. In Fig.2 (b), a line with points shows the primary person trip route that is estimated by out algorithm, and several pale lines mean alternative rational person trip route. In this case, the primary route is entirely consistent with the exact route. But the both of estimated routes are also rational, which satisfy traffic legal restrictions. The alternative routes may be rational within the error range of PHS location service. In several experiments, almost of all routes could be specified fast and correctly.

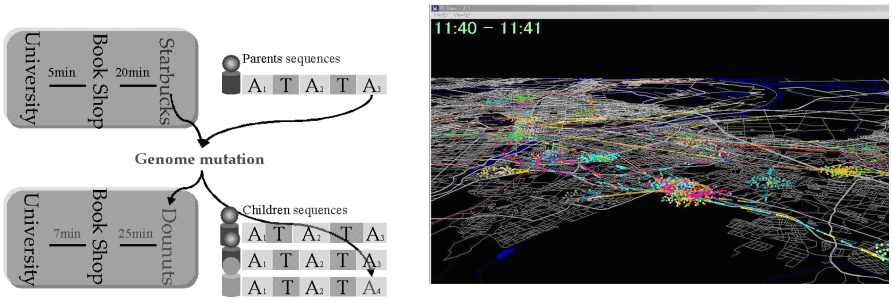
By using our proposed system, it is possible to collect sufficient volume of person trip routes into the spatial data warehouse in order to analyze the traffic congestion and patterns for statistical or mining objectives.

[3] One of important URLs is http://www.fcc.gov/911/.

4 Replication of Person Trip for Large-scale Simulation

There are several major analytic methods for traffic congestion, such as stochastic user equilibrium assignment model, numerical computing by simulation model and others. However, it is so hard to consider interfering constraints between trip parameters. Generally speaking, a sequence of person trips and actions strongly dependents each others, but almost of all analytical models don't care those interference of sequential actions.

On the other hand, by using advanced location service, it is to easy to capture the long-term sequence of detail trips and actions. Therefore, we propose our continuous person trip model with interfering actions, an array of observable trip-action is given by sequences of $OT_i = [A_1TA_2TA_3TA_4 \cdots]$ in Fig.3 (a). A person trip is presented by a sequence of actions A_i and trip time T.



(a) Replication of spatial information data.

(b) Visualization of MITSim results.

Fig. 3. OLAP Oriented Spatial Data Structure.

As we mentioned in previous section, due to the problems of privacy and observing cost, it is very difficult to collect sufficient numbers of person trip data. In order to execute MITSim as a large-scale simulation in Fig.3 (b), we need much more volume of observing data or their replications as initial settings. Therefore, with preserving the order of actions, we replicate and clone the array of trip-action data with different distribution of moving speed and staying time. Easily speaking, the sequence of $OT_i = [A_1TA_2TA_3TA_4]$ is preserved, we produce some mutants of a trip-action sequence presented in Fig.3 (a).

In our experiment, we collected the sequences of 99 person trip data, who had trip-actions from Sapporo city to Sapporo dome on November 24 in 2001, by observing PHS mobile terminals. We visualized sequences of person trips with 5 minutes interval time in Fig.4, we recognized that moving objects were concentrated into specific area, "*Sapporo dome*". Therefore, the replicants and

mutants of these sequences are also concentrated in the specific area, when we execute cloning process of trip-action data.

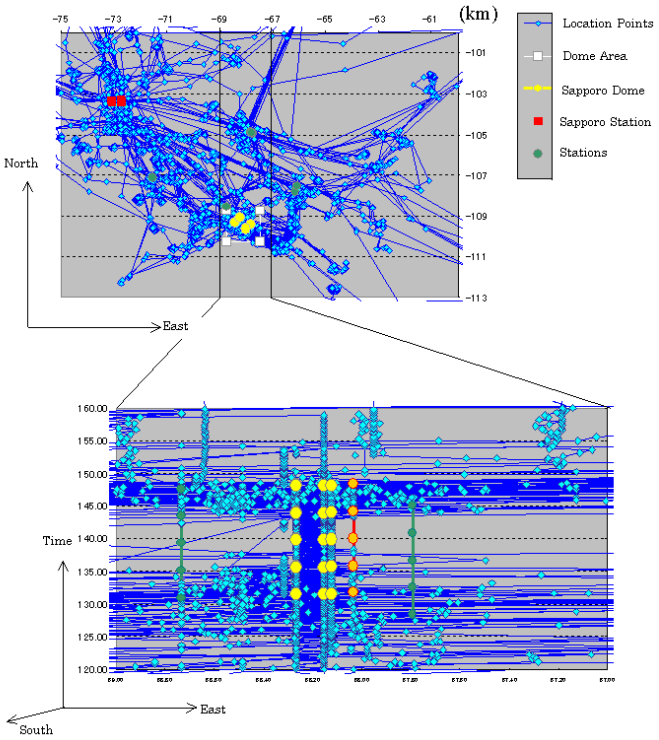


Fig. 4. Visualization of Person Trip Data.

Next, we executed MITSim as large-scale simulation presented in Fig.5. Based on the experimental results, it is possible to estimate the peak value of congesting situation correctly. There is a little bit delay of trips between the actual observation and our simulation results.

Next, it is important to analyze and discover the primary factors of this congesting situation in order to have traffic management and control.

5 Analytical Queries in Spatial Data Warehouse

In this section, we focus on the problems of OLAP (On-Line Analytical Processing) in a spatial data warehouse for traffic management and control. We discuss important analytical queries from the view point of traffic engineering. We also propose OLAP oriented spatial data structure in order to derive characteristics of moving objects in the data warehouse effectively.

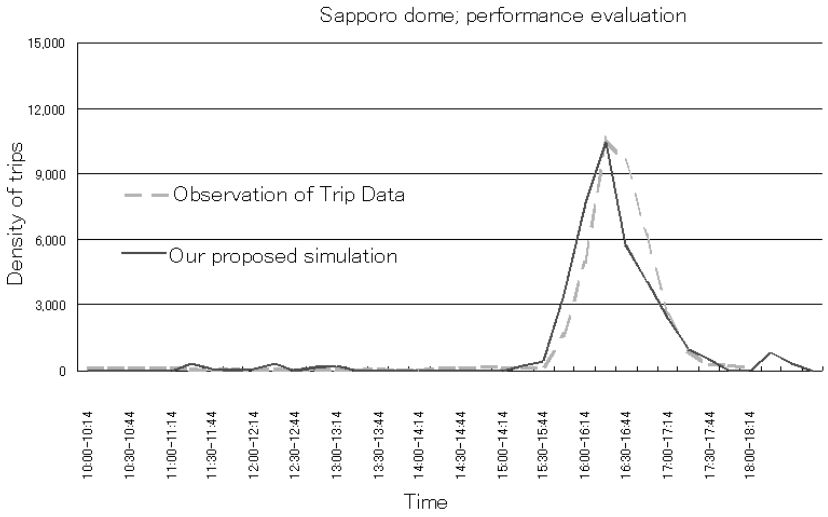


Fig. 5. Comparison of observation and simulation results.

5.1 OLAP queries for traffic management

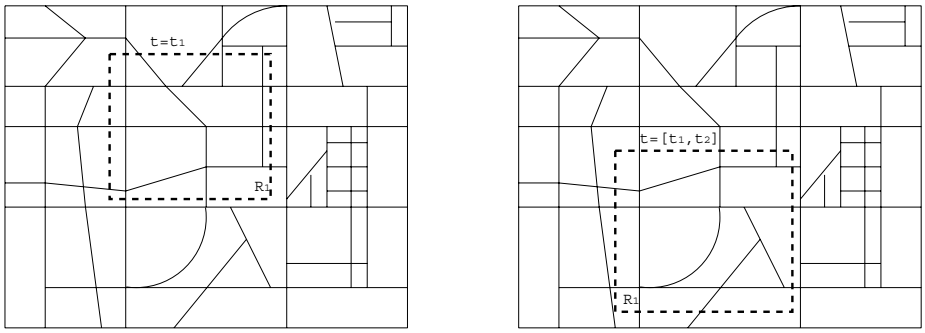
Firstly, we have to consider temporal spatial index such as TPR-tree (Time Parameterized R-tree)[10] in order to handle moving objects dynamically[7]. TPR-tree is an extension of R-tree index, and moving objects are stored in nodes of TPR-tree index by using the time function.

After we store a huge number of trip-action sequences into a spatial data warehouse, we need to execute temporal spatial queries[10] in order to discover characteristics of traffic flows from the view point of traffic management and control. Here, we use definitions of time series, t_1, $t_2(t_1 < t_2)$, and regions, R_1, R_2, and the following temporal spatial queries are important for our analysis.

1. **Timeslice query** $Q_{ts} = (R_1, t_1)$: At time point t, objects are searched for in a region R_1 in Fig.6 (a).
   (ex.) Based on the results of a query, we can calculate typical traffic flow parameters, such as traffic density, average traffic velocity and others.
2. **Window query** $Q_{win} = (R_1, t_1, t_2)$: Fig.6 (b) shows that moving objects are searched for in the regionR_1 from t_1 to t_2.
   (ex.) By using the results of window queries, we can calculate time average velocity which has rather stable property in traffic analysis.

5.2 $\sum$-tree for person trip data analysis

In order to execute analytical queries and mining processes, we proposed our spatial temporal data structure, which is based on the technologies of spatial



(a) Timeslice query. (b) Window query.

Fig. 6. Spatial Temporal Queries for Traffic Analysis.

temporal indices and data cube. Our proposed $\sum$-tree data structure in Fig.7 has a hierarchical tree structure with having total number of objects and sum of objects' speed stored in lower nodes[5].

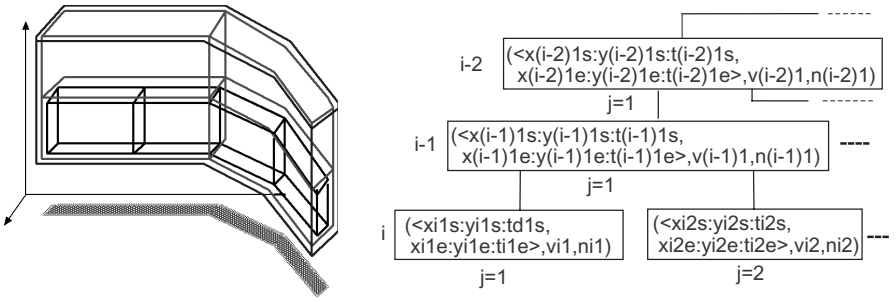
For instance, we consider spatial temporal data $(x_1, y_1, t_1), (x_2, y_2, t_2), \cdots, (x_n, y_n, t_n)$ with total number of objects n. This hierarchy is based on the spatial constraints with moving speed and direction of objects. Nodes $L_1, L_2, \cdots, L_N$ are constructed hierarchically by using spatial constraints of objects, such as similar vectors with moving speed and direction, and objects are stored in same segments of roads.

$$\begin{cases} (x_1, y_1, t_1), (x_2, y_2, t_2), \cdots, (x_i, y_i, t_i) \subset L_1 \\ (x_{i+1}, y_{i+1}, t_{i+1}), \cdots, (x_j, y_j, t_j) \subset L_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad : \\ (x_{k+1}, y_{k+1}, t_{k+1}), \cdots, (x_n, y_n, t_n) \subset L_N \end{cases}$$

Furthermore, considering additional speed attribute values (x_1, y_1, t_1, v_1), $(x_2, y_2, t_2, v_2), \cdots, (x_n, y_n, t_n, v_n)$ in a leaf node, sum of objects' speed V_{L_1}, $V_{L_2}, \cdots, V_{L_N}$ are also stored in upper nodes.

$$V_{L_1} = \sum_{l=1}^{i} v_l,\ V_{L_2} = \sum_{l=i+1}^{j} v_l,\ \cdots, V_{L_N} = \sum_{l=k+1}^{n} v_l$$

For example, in our proposed data structure of Fig.7 (b), we define the specific spatial node $(< x_{ijs} : y_{ijs} : t_{ijs}, x_{ije} : y_{ije} : t_{ije} >)$ based on two different tips of nodes $(x_{ijs}, y_{ijs}, t_{ijs})$ and $(x_{ije}, y_{ije}, t_{ije})$. We also store area of nodes $(x, y, t) \subset L_u$, sum of traffic parameters, such as total speed of objects V_{L_u} and total number of objects N_{L_u}. We can store those values into nodes recursively. We name this spatial temporal data structure as $\sum$-tree, in the specific nodes



(a) Data partitions for trip data analysis.

(b) Objects stored in $\sum$-tree.

Fig. 7. OLAP Orinted Spatial Data Structure.

including moving objects, it is possible to calculate summing-up and average values by using this structure with small computing cost effectively[5].

6 Conclusion

In this paper, we outlined the framework of our proposed spatial data warehouse for traffic management/control and discussed typical queries for traffic data analysis and mining. At present, in order to analyze the traffic flow and discover complex patterns from huge volume of person trip-action sequences, we need to execute our proposed cloning process in our architecture. In near future, if we can collect all actual positioning data in a region, we may omit the replication and simulation modules in our proposed system.

Acknowledgment
A fundamental part of this work is supported by the grant of Mazda Foundation. And a part of this paper is dependent on the research supported by Grant-in Aid for Scientific Research of the Ministry of Education, Science, Sports and Culture.

References

1. Ester, M., Frommelt, A., Kriegel, H.-P., and Sander, J., "Algorithms for Characterization and Trend Detection in Spatial Databases", Proc. of the fourth ACM SIGKDD International Conference (KDD-98), pp.44-50, 1998.
2. Hanshin Expressway Public Corporation, "Technical Trends of Mobile Location Systems", Technical Report of Mobile Location Service, February, 2000. (In Japanese)

3. Ito, Y., Kawano, H., Hasegawa, T., "Index based Clustering Discovery query for Spatial Data Mining," Proc. of the 10th Annual Conference of JSAI, pp.231–234, 1996. (In Japanese).
4. Kawano, H., "Architecture of Trip Database Systems: Spatial Index and Route Estimation Algorithm," XIV International Conf. of Systems Science, Vol. III, pp.110-117, Poland, 2001.
5. Kawano, H., "On-line Analytical Processing for person trip database," The sixteenth triennial conference of the International Federation of Operational Research Societies, p.118, 2002.
6. Koperski, K. and Han, J., "Discovery of Spatial Association Rules in Geographic Information Databases," Proc. 4th International Symposium SSD '95, pp. 275–289, 1995.
7. Minami, T., Tanabe, J. and Kawano, H., "Management of Moving Objects in Spatial Database: Architecture and Performance," Technical Report of IPSJ, Vol.2001, No.70, DBS-125, pp.225–232 (2001). (in Japanese)
8. Ng, R.T. and Han, J., "Efficient and Effective Clustering Methods for Spatial Data Mining," Proc. 20th VLDB, pp. 144–155, 1994.
9. Rogers, S., Langley, P., and Wilson, C., "Mining GPS Data to Augment Road Models," Proc. of the fifth ACM SIGKDD International Conference (KDD-99), pp.104–113, 1999.
10. Saltenis, S., Jensen, C. S., Leutenegger, S. T., and Mario A. Lopez, "Indexing the Positions of Continuously Moving Objects," Proc. of the 2000 ACM SIGMOD International Conference on Management of Data, pp.331-342, USA, 2000.
11. Samet, H., "Spatial Data structures," Modern Database Systems, (W. Kim, ed.), ACM Press, New York, pp. 361–385, 1995.
12. Samet, H., "The Design and Analysis of Spatial Data structures," Addison-Wesley, Reading, Mass. New York, 1995.
13. Shekhar, S. and Chawla, S., "Spatial Databases, A tour," Prentice Hall, 2003.

Title Index

Author Index