Proceedings of the Third International Workshop on Active Mining (AM-2004)

in Conjunction with the 18th Annual Conference of the Japanese Society for Artificial Intelligence, 2004

Organizers
Hiroshi Motoda
Takahira Yamaguchi
Shusaku Tsumoto
Masayuki Numao

Ishikawa Kousei Nenkin Kaikan, Kanazawa, JAPAN
June 1, 2004

ORGANIZERS

Hiroshi Motoda	Osaka University, Japan
Takahira Yamaguchi	Keio University, Japan
Shusaku Tsumoto	Shimane University, Japan
Masayuki Numao	Osaka University, Japan

PROGRAM COMMITTEE

Hiroki Arimura	Hokkaido University, Japan
Stephen D. Bay	Stanford University, U.S.A.
Hendrik Blockeel	Katholieke Universiteit Leuven, Belgium
Robert H.P. Engels	CognIT, Norway
Shoji Hirano	Shimane University, Japan
Tu Bao Ho	JAIST, Japan
Akihiro Inokuchi	IBM Japan, Japan
Hiroyuki Kawano	Kyoto University, Japan
Boonserm Kijsirikul	Chulalongkorn University, Thailand
Ross D. King	The University of Wales, Aberystwyth
Yasuhiko Kitamura	Kwansei Gakuin University, Japan
Ravi Kothari	IBM - India Research Lab, India
Marzena Kryszkiewicz	Warsaw University of Technology, Poland
Tsau Y. Lin	San Jose State University, U.S.A.
Huan Liu	Arizona State University, U.S.A.
Masayuki Numao	Osaka University, Japan
Miho Ohsaki	Doshisha University, Japan
Takashi Okada	Kwansei Gakuin University, Japan
Takashi Onoda	CRIEPI, Japan
Yukio Ohsawa	University of Tsukuba, Japan
Luc de Raedt	University of Freiburg, Germany
Henryk Rybinski	Warsaw University of Technology, Poland
Masashi Shimbo	NAIST, Japan
Einoshin Suzuki	Yokohama National University, Japan
Yoshimasa Takahashi	Toyohashi University of Technology, Japan
Masahiro Terabe	MRI, Japan
Ljupico Todorovski	Jozef Stefan Institute, Slovenia

Shusaku Tsumoto	Shimane University, Japan
Stefan Wrobel	Fraunhofer AIS and University Bonn, Germany
Seiji Yamada	NII, Japan
Takahira Yamaguchi	Keio University, Japan
Yiyu Yao	University of Regina, Canada
Kenichi Yoshida	University of Tsukuba, Japan
Tetsuya Yoshida	Hokkaido University, Japan

SPONSORS

Japanese Society for Artificial Intelligence
Active Mining Project (Grant-in-Aid for Scientific Research on Priority Areas, No.759)

Program and Table of Contents

June 1st (Tuesday)

Spiral Discovery of a Separate Prediction Model from Chronic Hepatitis Data

Masatoshi Jumi[1], Einoshin Suzuki[1], Muneaki Ohshima[2], Ning Zhong[2], Hideto Yokoi[3], and Katsuhiko Takabayashi[3]

1. Electrical and Computer Engineering, Yokohama National University, Japan
2. Faculty of Engineering, Maebashi Institute of Technology, Japan
3. Division of Medical Informatics, Chiba University Hospital, Japan
jumi@slab.dnj.ynu.ac.jp, suzuki@ynu.ac.jp, ohshima@wi-lab.com, zhong@maebashi-it.ac.jp, yokoih@telemed.ho.chiba-u.ac.jp, takaba@ho.chiba-u.ac.jp

Abstract. In this paper, we summarize our endeavor for spiral discovery of a separate prediction model from chronic hepatitis data. We have initially proposed various learning/discovery methods including time-series decision tree, PrototypeLines, and peculiarity-oriented mining method for mining the data. This experience has led us to believe that physicians tend to consider typical cases with the specific disease and rule out from the clearly exceptional cases. We have developed a spiral discovery system which learns a prediction model for each type of cases, and obtained promising results from experiments.

1 Introduction

Medical data are challenging to data miners since they show problems related to data quantity, data quality, and data form [5]. Chronic hepatitis data [1], which show high diversities in various aspects of cases, satisfy this nature and necessitate us to develop novel data mining methods. For the chronic hepatitis data, providers presented several objectives including progress of chronic hepatitis (difference between type B and type C) and effect of interferon therapy. Among them, LC (liver cirrhosis) prediction from blood test data can be regarded as one of the most important objectives since it can potentially substitute routine tests for a biopsy, which represents a highly invasive test.

We have been working on knowledge discovery from the chronic hepatitis data including the LC prediction. Our methods include a decision tree learner for time-series classification problem [5], a visualization method for medical test data [4], and a peculiarity-oriented mining method [6]. The methods have proved to be effective for various purposes including detection of exceptional cases.

From the endeavor, we have come to believe that physicians tend to start their differential diagnoses by distinguishing typical cases from apparently exceptional cases. In this paper, we propose a discovery method which distinguishes the two types of cases in its separate prediction model based on the abovementioned data mining methods.

2 Knowledge Discovery from the Chronic Hepatitis Data

2.1 Liver Cirrhosis Prediction

Chronic hepatitis represents a disease in which liver cells become inflamed and harmed by virus infection. In case the inflammation lasts a long period, the disease comes to an end which is called a liver cirrhosis (LC). During the process to an LC, the degree of fibrosis, which consists of five stages ranging from F0 (no fibrosis) to F4 (LC), represents an index of the progress. The degree of fibrosis can be inspected by biopsy which picks liver tissue by inserting an instrument directly into liver. A biopsy, however, cannot be frequently performed since it requires a short-term admission to a hospital and involves danger such as hemorrhage. Therefore, if we can predict the degree of fibrosis with a conventional medical test such as a blood test, it would be highly beneficial in medicine.

A time sequence $\boldsymbol{A}$ represents a list of values $\alpha_1, \alpha_2, \cdots, \alpha_I$ sorted in chronological order. 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. We assume that an attribute a_j represents a time-series attribute which takes a time sequence as its value[1]. The class attribute c represents a nominal attribute and its value is called a class. We show an example of a data set which consists of time-series attributes in Figure 1. The data set consists of examples 84, 85, 930, each of which is described with time-series attributes GPT, ALB, PLT, and a class.

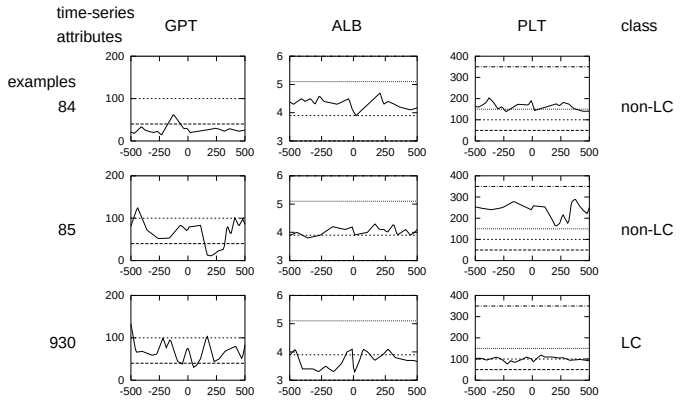


Fig. 1. Data set which consists of time-series attributes

In classification from time-series data, the objective represents induction of a classifier, which predicts the class of an example e, given a training data set D.

[1] Though it is straightforward to include other kinds of attributes in the definition, we exclude them for clarity.

In this paper, we assume liver cirrhosis (LC) and non-liver cirrhosis (non-LC) as classes.

2.2 Our Endeavor for the Chronic Hepatitis Data

Our time-series decision tree was proposed since physicians requested to use time sequences which exist in data in a classifier [5]. We applied the method to the LC prediction problem and the results attracted their interests. The physicians commented that the obtained decision tree is highly valid although we used medical knowledge only for selecting medical tests in the data [5]. Moreover, most of the cases who are mispredicted by the decision tree were recognized as exceptions by the physicians.

Our PrototypeLines represents a visualization method which is considered to enable discovery of interesting knowledge without extensive training [4]. Application of PrototypeLines to the chronic hepatitis data revealed interesting characteristics. A student in computer science discovered at least two exceptions which were both recognized as missing diseases in the original data by a domain expert. Moreover, a physician who was introduced PrototypeLines for the first time discovered an exception condition of a case after a 5-minute explanation.

Our peculiarity-oriented mining method obtains tendencies among examples with peculiar data [6]. The peculiar data are detected based on a distance measure, and the number of the peculiar data for an example is called the number of peculiar attributes. The method has successfully discovered exceptional cases in terms of interferon therapy.

3 Discovery of a Separate Prediction Model

3.1 Overall Architecture

Based on the motivation presented in Section 1, we propose a discovery method which learns a separate prediction model. Our method employs a naive Bayes learner and a 1-NN (1-nearest neighbor) classification method for typical cases and exceptional cases respectively, and refines the separate prediction model by a spiral interaction with a medical expert. The term "spiral" is employed since it represents iterative refinement of discovered knowledge. In the interaction, exceptional cases are identified with the data mining methods mentioned in Section 2.2 and a likelihood-based method to be proposed in the next Section.

Figure 2 shows the architecture of our method. In our method, detection of exceptional cases with our previous methods as well as update of a naive Bayes learner for typical cases are iterated in a spiral manner with interaction with an expert. A naive Bayes classifier predicts a class $\hat{c}_{NBayes,i}$ of an example e_i assuming that each attribute a_j is independent. Here v_{ij} represents the value for attribute a_j of p.

$$\hat{c}_{NBayes,i} = \mathrm{argmax}_c \Pr(c) \prod_{j=1}^{m} \Pr(a_j = v_{ij} \,|\, c) \tag{1}$$

As the result, we obtain a list of typical cases, conditional probabilities for typical cases, and a list of exceptional cases. Classification of a new case begins by detection of exceptional cases based on a 1-NN method, which will be presented in Section 3.3. If a case who is close to the new case is found, our method outputs the class of the former case. Otherwise, the class of the new case is predicted with a naive Bayes classifier.

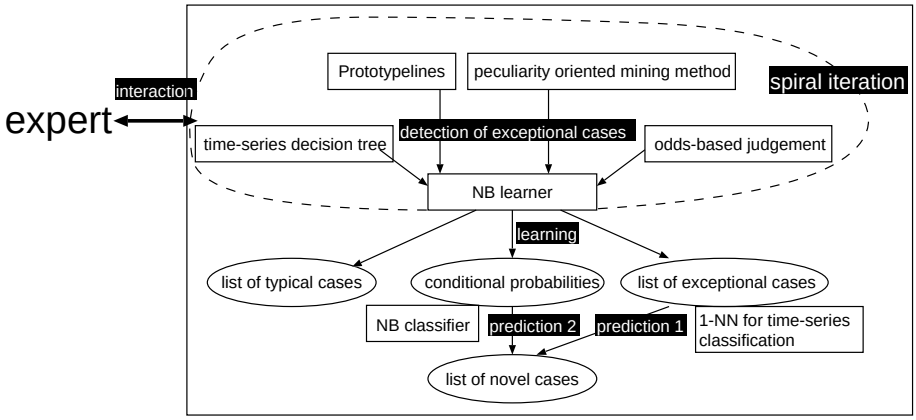


Fig. 2. Architecture of the proposed method

3.2 Likelihood-based Judgment for Exception Degrees

In a classification problem, an example can be intuitively regarded as "typical" or "atypical". The typicalness $\Phi(p)$, which is based on the right-hand side of Eq. (1), of a case p represents a degree to which p belongs to its class c compared with the other class $\overline{c}$. Here v_{ij} represents the value for attribute a_j of p.

$$\Phi(p) = \frac{\Pr(c)\prod_{j=1}^{m}\Pr(a_j = v_{ij} \mid c)}{\Pr(\overline{c})\prod_{j=1}^{m}\Pr(a_j = v_{ij} \mid \overline{c})} \tag{2}$$

The larger $\Phi(p)$ is, the more certain that p belongs to its class c.

If a naive Bayes classifier is relatively accurate for the class c, the typicalness $\Phi(p)$ tends to be large and vice versa. Thus the typicalness $\Phi(p)$ is relative since it depends on the preciseness of a naive Bayes classifier in terms of the class c. The degree how a naive Bayes classifier is precise for c can be measured by the degree of correct classification and the degree of incorrect classification. For c and $\overline{c}$, we represent the number of correctly-predicted examples by the naive Bayes method L and n respectively. Likewise, for c and $\overline{c}$, we represent the number of incorrectly-predicted examples by the naive Bayes method L_{e} and

n_{e} respectively. The preciseness $\Psi(\hat{c})$ of a naive Bayes classifier of the estimated class $\hat{c}$ represents the ratio of the precision for c and the precision for $\overline{c}$.

$$\Psi(\hat{c}) = \frac{(L+1)(n+n_{\mathrm{e}}+2)}{(L+L_{\mathrm{e}}+2)(n_{\mathrm{e}}+1)} \tag{3}$$

For our LC prediction problem, we define a degree of exception $E(p, \hat{c})$ for a case p and his/her estimated class $\hat{c}$ as follows in order to discriminate exceptional cases from typical cases.

$$E(p, \hat{c}) = \lceil -\log_{\Psi(\hat{c})} \Phi(p) \rceil \tag{4}$$

Intuitively, $E(p, \hat{c})$ represents an evaluation index which is equal to the number of upvaluated digits below the decimal point when we measure the typicalness $\Phi(p)$ of a case p in terms of the preciseness $\Psi(\hat{c})$ of a naive Bayes classifier of the estimated class $\hat{c}$. When prediction of the naive Bayes method is accurate, $\Psi(\hat{c})$ tends to be large, and the absolute value of $E(p, \hat{c})$ is relatively small even if the absolute value of $\Phi(p)$ is large. This fits our intuition that certain information rarely leads to an extreme degree of exception.

In each loop, cases whose degrees of exception are no less than a user-specified threshold are detected as exceptional cases. The loop continues until prediction accuracy estimated with 10-fold cross validation either decreases or reaches 100 %.

3.3 Similarity between a Pair of Cases Based on Dynamic Time Warping

Dynamic time warping (DTW) represents an index of dissimilarity between a pair of time sequences [3]. Unlike Euclidean distance, DTW can allow distortion along the time axis since a point in a time sequence can correspond to multiple points in the other sequence.

A 1-NN method based on DTW constantly showed high accuracy in our experiments for time-series classification including the LC prediction problem [5]. Due to this good result, we have chosen this method for detecting exceptional cases in test data. Following [5], the window width [3] of DTW was settled to 10 % of the length of the time sequence, and a dissimilarity measure $H(e_i, e_j)$ between a pair of examples e_i, e_j has been employed. This measure normalizes the results $G(e_i(a_k), e_j(a_k))$ of DTW for e_i and e_j in terms of a time-series attribute a_k with the maximum value $q(a_i)$, where $q(a_i) \geq \forall j \forall k G(e(a_j), e(a_k))$.

$$H(e_i, e_j) = \sum_{k=1}^{m} \frac{G(e_i(a_k), e_j(a_k))}{q(a_j)} \tag{5}$$

4 Experiments

4.1 Application of Conventional Methods

In the experiments, we used data from 180 days before the first biopsy to the day of the first biopsy following advice of medical experts. As the result, the num-

bers of LC and non-LC cases are 159 and 112 respectively. The blood tests that we use are GOT (glutamic oxaloacetic transaminase = AST (asparate aminotransferase)), GPT (glutamic pyruvic transaminase = ALT (alanine aminotransferase)), TTT (thymol turbidity test), ZTT (Zinc sulfate turbidity test), D-BIL (direct bilirubin), I-BIL (indirect bilirubin), T-BIL (total bilirubin), ALB (albumin), CHE (cholineesterase), TP (total protein), T-CHO (total cholesterol), WBC (white blood cell), PLT (platelet), and HGB (hemoglobin). In our experiment, we used 46 LC cases and 55 non-LC cases each of whom has test values for all of these blood tests.

For the classifier, we first averaged each time sequence then discretized each value as we show in Table 1, where we use for each attribute value U: extremely high, V: very high, H: high, N: normal, L: low, v: very low, and u: extremely low. Each conditional probability is estimated using Laplace correction in order to cope with the 0-occurrence problem [2]. A missing value is ignored both in estimating probabilities and in classifying an example.

Table 1. Blood tests in the chronic hepatitis data

attribute	intuitive explanation	discretization
GOT	amount of broken	N≤ 40< H≤ 100< V≤ 200< U
GPT	liver cells	N≤ 40< H≤ 100< V≤ 200< U
TTT	degree of	N≤ 5 < H≤ 10< V≤ 15< U
ZTT	immune activity	N≤ 12< H≤ 24< V≤ 36< U
D-BIL	disorder of	N≤ 0.3< H≤ 0.6< V≤ 0.9< U
I-BIL	bile excretion	N≤ 0.9< H≤ 1.8< V≤ 2.7< U
T-BIL		N≤ 1.2< H≤ 2.4< V≤ 3.6< U
ALB	decrease of	v≤ 3.0< L≤ 3.9< N≤ 5.1< H≤ 6.0<V
CHE	protein generation	v≤ 100< L≤180< N≤430< H≤510<V
TP		v≤ 5.5< L≤ 6.5< N≤ 8.2< H≤ 9.2<V
T-CHO		v≤ 90 < L≤125< N≤220< H≤255<V
WBC	good order	u≤ 2.0< v≤ 3.0< L≤ 4.0< N≤ 9.0<H
PLT	of a liver	u≤ 50 < v≤100< L≤150< N≤350<H
HGB	hemoglobin	L≤ 12< N ≤ 18< H

First, the medical expert investigated display result of PrototypeLines from 500 days before the first biopsy to 500 days after the first biopsy. As the result, cases 380, 336, 928, 251, 602, 913 are recognized as candidates of exceptional cases. Inspecting time sequences for these cases, he concluded that case 336 is recognized as an exception since his/her PLT is very low and his/her ALB, CHE, TP, WBC are low although s/he belongs to non-LC. Moreover a fatty liver is suspected since his/her GPT is much higher than GOT. Case 928 is suspected to show acute aggravation or an uninterpretable disease thus is removed.

Similarly, we applied our peculiarity-oriented mining method to the data, and showed cases with no less than four peculiar attributes to the medical expert. As the result, cases 611 and 903 show low ALB and CHE though their TP are high and T-CHO are normal. Since these four blood tests are closely related to each other and typically synchronize, these cases are recognized as exceptions. Case 916 showed a peculiar sequence for T-CHO compared with his/her ALB, CHE, TP thus is recognized as exception.

The above cases are considered to be peculiar in the whole data. In order to detect peculiar cases in LC cases or in non-LC cases, we applied this method to data with the corresponding class. From non-LC data, several cases are detected including the abovementioned cases 615 and 160, of which WBC and PLT only are low. At this moment, the medical expert decided to postpone decisions on "borderline cases" each of whom is similar to cases in the other class, and to remove only apparently abnormal cases. For instance, decision on case 596 is postponed although s/he looks very differently from the other cases who belong to 596's degree of index F2. Blood tests ALB, CHE, TP, T-CHO are closely related to each other as we described earlier, and a case who shows low values for one or two of them is called a monocytepenia or a bicytepenia respectively. While the medical expert was inspecting case 755, he discovered that such borderline cases can be classified into either bicytepenia or monocytepenia. In these results, cases 160, 918, 596, 755 are judged bicytepenia, and cases 615, 925, 596, 755 are judged monocytepenia. He inferred that removing these borderline cases might reduce predictive accuracy of similar kinds of cases, and this effect should be investigated.

Similarly, interesting cases are also detected from LC data. We confirmed removal of case 611, who was detected from the whole data, since s/he is suspected to suffer hemorrhage right after the biopsy. Case 903 was judged to stay in typical cases since s/he belongs to "partial LC" who shows partial aggravation of blood tests. We postponed removal of case 916, who was also detected from the whole data. Case 943 is judged to be removed since s/he is suspected to suffer from constitutional hyperbilirubinemia (as a complication) since his/her TTT and ZTT are high but no other results suggest LC.

The medical expert then inspected misclassified cases from the time-series decision tree. Case 737 shows good blood test results although s/he belongs to LC, thus is judged as a compensatory LC. Case 615 shows good blood test results except for T-CHO thus is misclassified. This is due to the nature of a decision tree which makes prediction based a relatively small number of attributes. Case 755 shows bad results although s/he belongs to non-LC. We have concluded that misclassified cases from the time-series decision tree represent borderline cases.

4.2 Application of the Proposed Method

Following the policy in the previous Section, we first defined exceptional cases as abnormal cases who are suspected to suffer from different diseases and tried to obtain the separate prediction model. After careful examination, cases 916, 928, 943 were first disregarded as exceptions. The naive Bayes learner was applied to

the remaining data, and the likelihood-based method in Section 3.2 was applied in order to detect candidates of exceptional cases, in which we used 3 as the value of the threshold. As the result, 14 cases showed exception degrees no less than 3. The cases identifiers are 203, 206, 236, 245, 256, 737 (LC, degree 5); 251, 758 (LC, degree 4); 184, 244, 553, 699, 913 (LC, degree 3); and 160 (non-LC, degree 3). The medical experts inspected their time sequences but none of them were recognized as abnormal thus the result at this moment was judged as final.

We then defined exceptional cases as borderline cases who resemble to cases in the other class. After careful examination, cases who are removed before the first spiral are grouped in terms of their causes. The following non-LC cases, though they belong to either F1 or F2, resemble to LC cases.
1. Blood cell decrease: 336, 160, 596, 755 (only PLT), 918, 926 (only PLT).
2. ALB decrease: 336, 596, 755, 918, 926.
3. CHE decrease: 336, 160, 596, 926.
4. T-CHO decrease: 615, 925.
The following LC cases, though they belong to F4, resemble to non-LC cases.
1. Blood cell non-decrease: 611, 903, 602 (only WBC), 203 (only WBC), 404, 737, 244 (only PLT).
2. ALB non-decrease: 737, 244, 913.
3. CHE non-decrease: 203, 737, 244, 913, 404.
4. T-CHO non-decrease: 611, 203, 244, 251, 404, 737, 903.

We define that a cycle consists of application of the naive Bayes method, detection of candidates of exceptional cases by the likelihood-based method, and determination of exceptional cases by the medical expert. Below we show exceptional cases detected in each cycle and their causes. All of them belong to LC.
1. Second cycle: 184 (ALB, CHE, T-CHO, WBC), 236 (ALB, CHE, T-CHO, WBC), 206 (ALB, CHE, T-CHO, WBC, PLT), 245 (ALB, T-CHO), 758 (ALB, CHE, TP, T-CHO), 553(ALB, CHE, TP, T-CHO, WBC).
2. Third cycle: 166 (ALB, T-CHO, CHE), 699 (ALB, CHE, TP, T-CHO, WBC, PLT).
3. Fourth cycle: 623 (ALB, CHE, T-CHO, WBC).

We show the final conditional probabilities of the naive Bayes classifier in Table 2. In the Table, for each blood test a and a category v, $\hat{\Pr}(a{=}v|\text{non-LC})$ $(n(a{=}v|\text{non-LC}))$ | $\hat{\Pr}(a{=}v|\text{LC})$ $(n(a{=}v|\text{LC}))$ are shown, where $n(\cdot)$ represents the corresponding number of examples in the data set. For instance, there are 6 non-LC cases and 1 LC cases for ZTT=N thus the probabilities are obtained using Laplace correction since there are 49 non-LC cases and 27 LC cases in the data set which corresponds to the Table. In the Table, we emphasize categories each of which shows more than 3 times of difference and no smaller than 10 % with boldface and with underline for non-LC predominant and LC predominant respectively.

According to the medical expert, this kind of conditional probabilities represent useful information in building a medical expert system. Compared with its first probabilities, the number of categories that are effective in discriminating

Table 2. Conditional probabilities (%) and numbers of examples for typical cases, where each categories shows $\hat{\Pr}(a{=}v|\text{non-LC})$ $(n(a{=}v|\text{non-LC}))$ | $\hat{\Pr}(a{=}v|\text{LC})$ $(n(a{=}v|\text{LC}))$

GOT	**N: 35.3(17) | 3.3(0)**	H: 39.2(19) | 43.3(12)	V: 21.6(10) | 46.7(13)	U: 3.9(1) | 6.7(1)
GPT	N: 15.7(7) | 6.7(1)	H: 51.0(25) | 23.3(6)	V: 21.6(10) | 60.0(17)	U: 11.8(5) | 10.0(2)
TTT	N: 39.2(19) | 26.7(7)	H: 27.5(13) | 46.7(13)	**V: 29.4(14) | 6.7(1)**	U: 3.9(1) | 20.0(5)
ZTT	N: 13.7(6) | 6.7(1)	H: 68.6(34) | 73.3(21)	V: 13.7(6) | 16.7(4)	U: 3.9(1) | 3.3(0)
D-BIL	N: 88.2(44) | 36.7(10)	H: 7.8(3) | 46.7(13)	V: 2.0(0) | 10.0(2)	U: 2.0(0) | 6.7(1)
I-BIL	N: 96.0(47) | 75.9(21)	H: 2.0(0) | 17.2(4)	V: 2.0(0) | 6.9(1)	
T-BIL	N: 96.0(47) | 69.0(19)	H: 2.0(0) | 24.1(6)	V: 2.0(0) | 6.9(1)	
ALB	L: 12.2(5) | 60.7(16)	N: 87.8(42) | 39.3(10)		
CHE	v: 2.0(0) | 10.0(2)	L: 2.0(0) | 53.3(15)	N: 92.2(46) | 33.3(9)	H: 3.9(1) | 3.3(0)
TP	L: 4.0(1) | 3.4(0)	N: 92.0(45) | 93.1(26)	H: 4.0(1) | 3.4(0)	
T-CHO	L: 2.0(0) | 16.7(4)	N: 88.0(43) | 76.7(22)	H: 6.0(2) | 3.3(0)	V: 4.0(1) | 3.3(0)
WBC	u: 1.9(0) | 6.5(1)	v: 1.9(0) | 6.5(1)	L: 9.6(4) | 12.9(3)	N: 82.7(42) | 71.0(21)
				H: 3.8(1) | 3.2(0)
PLT	u: 2.0(0) | 10.0(2)	v: 5.9(2) | 50.0(14)	L: 27.5(13) | 26.7(7)	**N: 64.7(32) | 13.3(3)**
HGB	L: 4.1(1) | 21.4(5)	N: 95.9(46) | 78.6(21)		

LC from non-LC cases increases. This signifies that the removed cases actually represent borderline cases. It has been also observed that the prediction accuracy of the naive Bayes method increases.

4.3 Analysis of Experimental Results

For the experimental results in the previous Section, the accuracy of the separate prediction model is 73.5 %. More precisely, the accuracies for exceptional cases and typical cases were 52.4 % and 79.2 % respectively. The overall accuracy is similar to that of a conventional naive Bayes classifier, but it should be noted that the predictive accuracy for LC cases is higher than that by a conventional naive Bayes classifier. The accuracies for LC and non-LC cases were 77.3 % and 70.4 % respectively mostly because the correctly predicted cases with the 1-NN method were all LC. Our separate prediction model, which first applies the 1-NN method for predicting the class of exceptional cases, regards LC cases important since it is adequate in predicting exceptional LC cases due to the nature of its dissimilarity measure[2]. This fits the nature of the LC prediction problem in which overlooking of LC cases costs more than misprediction of non-LC cases.

In data mining, discovered knowledge is typically more important than high accuracy. Our experiments have revealed that detection of exceptional LC cases could be done by searching for asynchronism in blood tests ALB, CHE, T-CHO, WBC, PLT; and HGB might be safely ignored. Although it is impossible to detect an exceptional LC case who shows good results for all blood tests, our 1-NN method could detect exceptional LC cases relatively accurately. A partial LC case who shows partial aggravation of blood tests was known among medical experts only empirically, but we have succeeded in detecting several of them.

[2] Exceptional LC cases are relatively stable in their time sequences and are more easily predicted with the 1-NN method

Such cases might suffer from genetic problems, and detailed inspection can be expected to reveal their true causes.

5 Conclusions

In this paper, we have described our endeavor with our separate prediction model. The motivation is based from our previous endeavor, from which we have come to believe that physicians tend to start by distinguishing typical cases from clearly exceptional cases in their diagnosis.

Among conventional prediction methods for LC cases, ALB such as Child Pugh classification is the most frequently used and PLT is also known a good indicator. Blood chemistry and Complete blood count tests such as CHE, T-CHO, WBC were known to decrease as liver cirrhosis progresses, but various factors have prohibited their quantitative evaluation. We obtained comments that our endeavor with the separate prediction model is expected to contribute to such analysis. Our future work concerns building an effective data mining method as well as contributing to analysis of the chronic hepatitis data.

References

1. P. Berka: ECML/PKDD 2002 Discovery Challenge, Download Data about Hepatitis, http://lisp.vse.cz/challenge/ecmlpkdd2002/ (current September 28th, 2002).
2. P. Domingos and M. Pazzani: On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, *Machine Learning*, Vol. 29, No. 2/3, pp. 103–130 (1997).
3. E. J. Keogh: Mining and Indexing Time Series Data, *Tutorial at the 2001 IEEE International Conference on Data Mining (ICDM)*, http://www.cs.ucr.edu/%7Eeamonn/tutorial_on_time_series.ppt (2001).
4. E. Suzuki, T. Watanabe, H. Yokoi, and K. Takabayashi: Detecting Interesting Exceptions from Medical Test Data with Visual Summarization, *Proc. Third IEEE International Conference on Data Mining (ICDM)*, pp. 315-322 (2003).
5. Y. Yamada, E. Suzuki, H. Yokoi, and K. Takabayashi: Decision-tree Induction from Time-series Data Based on a Standard-example Split Test, *Proc. Twentieth International Conference on Machine Learning (ICML)*, pp. 840-847 (erratum http://www.slab.dnj.ynu.ac.jp/erratumicml2003.pdf) (2003).
6. N. Zhong, Y. Y. Yao, and M. Ohshima: Peculiarity Oriented Multi-Database Mining, *IEEE Transaction on Knowledge and Data Engineering*, Vol. 15, No. 4, pp. 952-960 (2003).

Evaluation of Rule Interestingness Measures with a Clinical Dataset on Hepatitis

Miho Ohsaki[1], Shinya Kitaguchi[1], Kazuya Okamoto[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 empirically investigates the performance of conventional rule interestingness measures and discusses their availability for supporting KDD through human-system interaction in medical domain. We compared the evaluation results by a medical expert and those by selected measures for the rules discovered from a dataset on hepatitis. Recall, Jaccard, Kappa, CST, χ^2-M, and Peculiarity demonstrated the highest performance, and many measures showed a complementary trend under our experimental conditions. These results indicate that some measures can predict really interesting rules at a certain level and that their combinational use will be useful.

1 Introduction

Rule interestingness is one of active fields in Knowledge Discovery in Databases (KDD), and there have been many studies that formulated interestingness measures and evaluated rules with them instead of humans. However, many of them were individually proposed and not fully evaluated from the viewpoint of theoretical and practical validity. Although some latest studies made a survey on conventional interestingness measures and tried to categorize and analyze them theoretically [1–3], little attention has been given to their practical validity – whether they can contribute to find out really interesting rules.

Therefore, this research aims to (1) systematically grasp the conventional interestingness measures, (2) compare them with real human interest through an experiment, and (3) discuss their performance to estimate real human interest and their utilization to support human-system interaction in KDD process. The experiment required the actual rules from a real dataset and their evaluation results by a human expert. We determined to set the domain of this research as medical data mining and to use the outcome of our previous research on hepatitis [4] because medical data mining is scientifically and socially important and especially needs human-system interaction support for enhancing rule quality.

In this paper, Section 2 introduces conventional interestingness measures and selects dozens of measures suitable to our purpose. Section 3 shows the experiment that evaluated the rules on hepatitis with the measures and compared the evaluation results by them with those by a medical expert. In addition, it discusses their performance to estimate real human interest, availability to support human-system interaction, and advanced utilization by combining them. Section 4 concludes the paper and comments on the future work.

2 Conventional Rule Interestingness Measures

The results of our and other researchers' surveys [1–3, 5] show that interestingness measures can be categorized with the several factors in Table 1. The subject to evaluate rules, a computer or human user, is the most important categorization factor. Interestingness measures by a computer and human user are called objective and subjective ones, respectively. There are more than forty objective measures at least. They estimate how a rule is mathematically meaningful based on the distribution structure of the instances related to the rule. They are mainly used to remove meaningless rules rather than to discover really interesting ones for a human user, since they do not include domain knowledge [6–14, 16–18]. In contrast, there are only a dozen of subjective measures. They estimate how a rule fits with a belief, a bias, or a rule template formulated beforehand by a human user. Although they are useful to discover really interesting rules to some extent due to their built-in domain knowledge, they depend on the precondition that a human user can clearly formulate his/her own interest and do not discover absolutely unexpected knowledge. Few subjective measures adaptively learn real human interest through human-system interaction.

Table 1. The factors to categorize interestingness measures.

Factors	Meaning	Subfactors
Subject	Who evaluates?	Computer / Human user
Object	What is evaluated?	Association rule / Classification rule
Unit	By how many objects?	A rule / A set of rules
Criterion	Based on what criterion?	Absolute criterion / Relative criterion
Theory	Based on what theory?	Number of instances / Probability / Statistics / Information / Distance of rules or attributes / Complexity of a rule

The conventional interestingness measures, not only objective but also subjective, do not directly reflect the interest that a human user really has. To avoid the confusion of real human interest, objective measure, and subjective measure, we clearly differentiate them. **Objective Measure**: The feature such as the correctness, uniqueness, and strength of a rule, calculated by the mathematical

analysis. It does not include human evaluation criteria. **Subjective Measure**: The similarity or difference between the information on interestingness given beforehand by a human user and those obtained from a rule. Although it includes human evaluation criteria in its initial state, the calculation of similarity or difference is mainly based on the mathematical analysis. **Real Human Interest**: The interest which a human user really feels for a rule in his/her mind. It is formed by the synthesis of cognition, domain knowledge, individual experiences, and the influences of the rules that he/she evaluated before.

This research specifically focuses on objective measures and investigates the relation between them and real human interest. We then explain the details of objective measures here. They can be categorized into some groups with the criterion and theory for evaluation. Although the criterion is absolute or relative as shown in Table 1, the majority of present objective measures are based on an absolute criterion. There are several kinds of criterion based on the following factors: Correctness – How many instances the antecedent and/or consequent of a rule support, or how strong their dependence is [6, 7, 13, 17], Generality – How similar the trend of a rule is to that of all data [11] or the other rules, Uniqueness – How different the trend of a rule is from that of all data [10, 14, 18] or the other rules [11, 13], and Information Richness – How much information a rule possesses [8]. These factors naturally prescribe the theory for evaluation and the interestingness calculation method based on the theory. The theory includes the number of instances [6], probability [12, 14], statistics [13, 17], information [7, 17], the distance of rules or attributes [10, 11, 18], and the complexity of a rule [8] (See Table 1). We selected the objective measures in Table 2 as many and various as possible. They were used in the experiment in Section 3. Note that many of them do not have the reference numbers of their original papers but those of survey papers in Table 2 to avoid too many literatures in this paper.

Now, we explain the motivation of this research in detail. Objective measures are useful to automatically remove obviously meaningless rules. However, some factors of evaluation criterion have contradiction to each other such as generality and uniqueness and may not match with or contradict to real human interest. In a sense, it may be proper not to investigate the relation between objective measures and real human interest, since their evaluation criterion does not include the knowledge on rule semantics and are obviously not the same of real human interest. However, our idea is that they may be useful to support the KDD through human-system interaction if they possess a certain level of performance to detect really interesting rules. In addition, they may offer a human user unexpected new viewpoints. Although the validity of objective measures has been theoretically proven and/or experimentally discussed using some benchmark data [1–3], very few attempts have been made to investigate their comparative performance and the relation between them and real human interest for a real application [5]. We think that our investigation is novel in this light.

Table 2. The objective measures of rule interestingness used in this research. **N:** Number of instances included in the antecedent and/or consequent of a rule. **P:** Probability of the antecedent and/or consequent of a rule. **S:** Statistical variable based on P. **I:** Information of the antecedent and/or consequent of a rule. **D:** Distance of a rule from the others based on rule attributes.

Measure Name (Abbreviation) [Reference Number of Literature]	Theory
Coverage [5]	**P**
Prevalence [5]	**P**
Precision [3, 5]	**P**
Recall [5]	**P**
Support [1, 3, 5]	**P**
Specificity [5]	**P**
Accuracy [5]	**P**
Lift [5]	**P**
Leverage [5]	**P**
Added Value (AV) [3]	**P**
Relative Risk (RR) [1]	**P**
Jaccard [3]	**P**
Certainty Factor (CF) [3]	**P**
Odds Ratio (OR) [3]	**P**
Yule's Q [3]	**P**
Yule's Y [3]	**P**
Kappa [3]	**P**
Klosgen's Interestingness (KI) [1, 3]	**P**
Brin's Interest (BI) [3]	**P**
Brin's Conviction (BC) [3]	**P**
Gray and Orlowska's Interestingness emphasizing Dependency (GOI-D) [1, 5, 12]	**P**
Gray and Orlowska's Interestingness emphasizing Generality (GOI-G) [1, 5, 12]	**P**
Credibility [5, 9]	**P, N**
Laplace Correction (LC) [3]	**P**
Collective Strength (CST) [3]	**P**
χ^2 Measure (χ^2-M) [5, 13]	**S**
Gini Index (Gini) [3]	**S**
Goodman and Kruskal's Interestingess (GKI) [3]	**S**
Normalized Mutual Information (NMI) [3]	**I**
J-Measure [1, 3, 5, 7]	**I**
Yao and Liu's Interestingness 1 based on one-way support (YLI1) [1]	**I**
Yao and Liu's Interestingness 2 based on two-way support (YLI2) [1]	**I**
Yao and Liu's Interestingness 3, the addition of YLI1 and YLI2 (YLI3) [1]	**I**
K-Measure (K-M) [5]	**I**
ϕ Coefficient (ϕ) [3]	**N**
Piatetsky-Shapiro's Interestingness (PSI) [3, 5, 6]	**N**
Cosine Similarity (CSI) [3]	**N**
Gago and Bento's Interestingness (GBI) [11]	**D**
Peculiarity [18]	**D**

3 Evaluation Experiment of Objective Measures

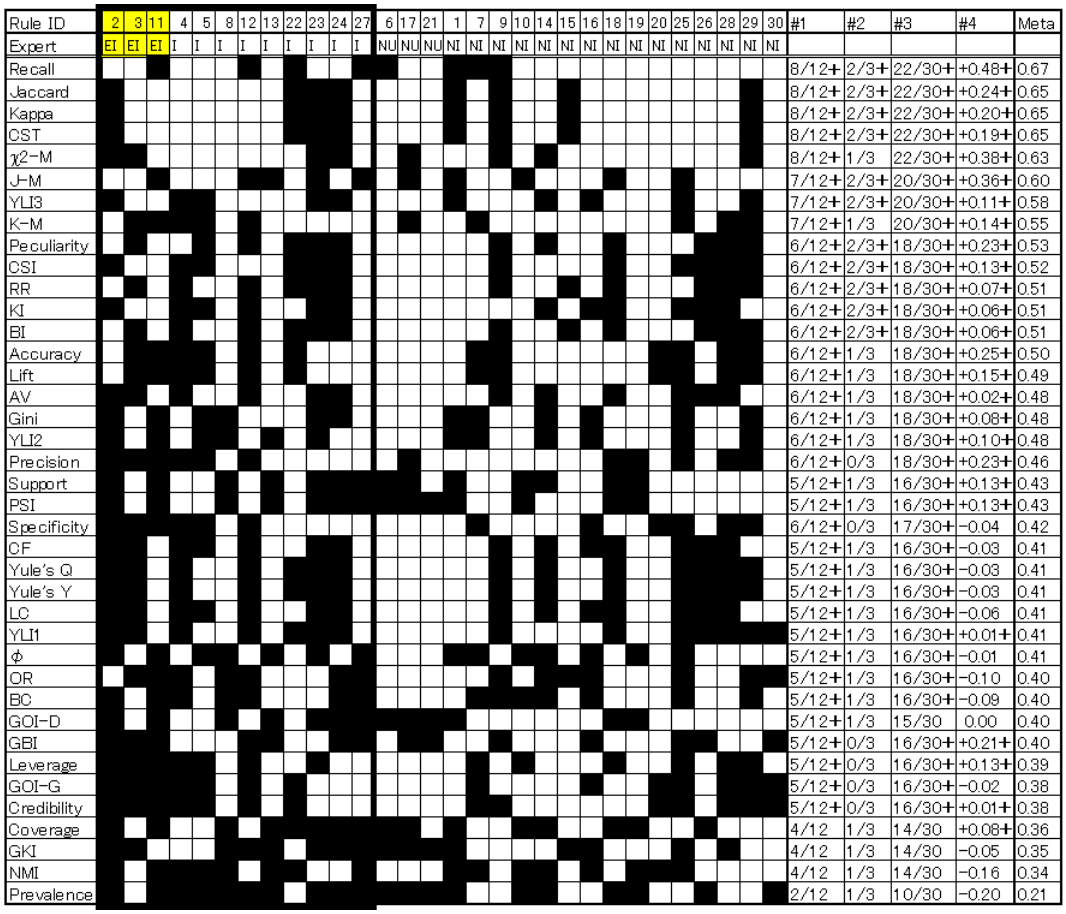
3.1 Experimental Conditions

This experiment examined the performance of objective measures to estimate real human interest by comparing the evaluation by them and a human user. Concretely speaking, the selected objective measures and a medical expert evaluated the same medical rules, and their evaluation values were qualitatively and quantitatively compared. We used the objective measures in Table 2 and the rules and their evaluation results by a medical expert in our previous research [4]. Here, we note the outline of the previous research. We had tried to extract graph-based rules consisting of the temporal patterns of medical test results from a clinical dataset on hepatitis. We had repeated the rule generation by our mining system and the rule evaluation by a medical expert twice. The several rules in the first mining inspired him to make a hypothesis, a seed of new medical knowledge, and the ones in the second mining supported him to verify the hypothesis. Finally, we obtained two sets of rules and their evaluation results.

The evaluation procedure by the medical expert was as follows: After each mining, he gave each rule the comment on it's medical interpretation and one of the rule quality labels, which were Especially-Interesting (**EI**), Interesting (**I**), Not-Understandable (**NU**), and Not-Interesting (**NI**). **EI** means that the rule was a key to generate or verify the hypothesis. Three and nine rules received **EI** and **I** in the first mining, respectively. Similarly, two and six rules did in the second mining. The evaluation procedure by the objective measures was designed as follows: All objective measures evaluated the same rules. We then sorted the rules in the descending order of the evaluation values and assigned the rule quality labels to them based on the number of rules with **EI** and **I** in the evaluation by the medical expert. GOI and Peculiarity have pre-set parameters. We call GOI with the dependency coefficient value at the double of the generality one GOI-D, and vice versa for GOI-G. We adopted the default value, 0.5, for the parameter α of Peculiarity.

3.2 Results and Discussion

Fig. 1 and 2 show the experimental results in the first and second mining, respectively. We analyzed the relation between the evaluation results by the medical expert and the objective measures qualitatively and quantitatively. As the qualitative analysis, we visualized their degree of agreement to easily grasp its trend. We colored the rules with perfect agreement white, probabilistic agreement gray, and disagreement black. A few objective measures output same evaluation values for too many rules. For example, although the eight rules were especially interesting (**EI**) or interesting (**I**) for the medical expert in second mining, the objective measure OR estimated 14 rules as **EI** or **I** ones (See Fig. 2). In that case, we colored such rules gray. The pattern of white (possibly also gray) and black cells for an objective measure describes how its evaluation matched with those by the medical expert. The more the number of white cells in the left-hand side, the better its performance to estimate real human interest.



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	Meta
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					
Recall																															8/12+	2/3+	22/30+	+0.48+	0.67
Jaccard																															8/12+	2/3+	22/30+	+0.24+	0.65
Kappa																															8/12+	2/3+	22/30+	+0.20+	0.65
CST																															8/12+	2/3+	22/30+	+0.19+	0.65
χ^2-M																															8/12+	1/3	22/30+	+0.38+	0.63
J-M																															7/12+	2/3+	20/30+	+0.36+	0.60
YLI3																															7/12+	2/3+	20/30+	+0.11+	0.58
K-M																															7/12+	1/3	20/30+	+0.14+	0.55
Peculiarity																															6/12+	2/3+	18/30+	+0.23+	0.53
CSI																															6/12+	2/3+	18/30+	+0.13+	0.52
RR																															6/12+	2/3+	18/30+	+0.07+	0.51
KI																															6/12+	2/3+	18/30+	+0.06+	0.51
BI																															6/12+	2/3+	18/30+	+0.06+	0.51
Accuracy																															6/12+	1/3	18/30+	+0.25+	0.50
Lift																															6/12+	1/3	18/30+	+0.15+	0.49
AV																															6/12+	1/3	18/30+	+0.02+	0.48
Gini																															6/12+	1/3	18/30+	+0.08+	0.48
YLI2																															6/12+	1/3	18/30+	+0.10+	0.48
Precision																															6/12+	0/3	18/30+	+0.23+	0.46
Support																															5/12+	1/3	16/30+	+0.13+	0.43
PSI																															5/12+	1/3	16/30+	+0.13+	0.43
Specificity																															6/12+	0/3	17/30+	−0.04	0.42
CF																															5/12+	1/3	16/30+	−0.03	0.41
Yule's Q																															5/12+	1/3	16/30+	−0.03	0.41
Yule's Y																															5/12+	1/3	16/30+	−0.03	0.41
LC																															5/12+	1/3	16/30+	−0.06	0.41
YLI1																															5/12+	1/3	16/30+	+0.01+	0.41
ϕ																															5/12+	1/3	16/30+	−0.01	0.41
OR																															5/12+	1/3	16/30+	−0.10	0.40
BC																															5/12+	1/3	16/30+	−0.09	0.40
GOI-D																															5/12+	1/3	15/30	0.00	0.40
GBI																															5/12+	0/3	16/30+	+0.21+	0.40
Leverage																															5/12+	0/3	16/30+	+0.13+	0.39
GOI-G																															5/12+	0/3	16/30+	−0.02	0.38
Credibility																															5/12+	0/3	16/30+	+0.01+	0.38
Coverage																															4/12	1/3	14/30	+0.08+	0.36
GKI																															4/12	1/3	14/30	−0.05	0.35
NMI																															4/12	1/3	14/30	−0.16	0.34
Prevalence																															2/12	1/3	10/30	−0.20	0.21

Fig. 1. The evaluation results by a medical expert and objective measures for the rules in first mining. Each column represents a rule, and each row represents the set of evaluation results by an objective measure. The rules are sorted in the descending order of the evaluation values given by the medical expert. The objective measures are sorted in the descending order of the meta criterion values. A square in the left-hand side surrounds the rules labeled with **EI** or **I** by the medical expert. White, gray, and black cells mean that the evaluation by an objective measure was perfectly, was probabilistically, and was not the same by the medical expert, respectively. The five columns in the right side show the performance on the four comprehensive criteria and the meta one. '+' means the value is greater than that of random selection.

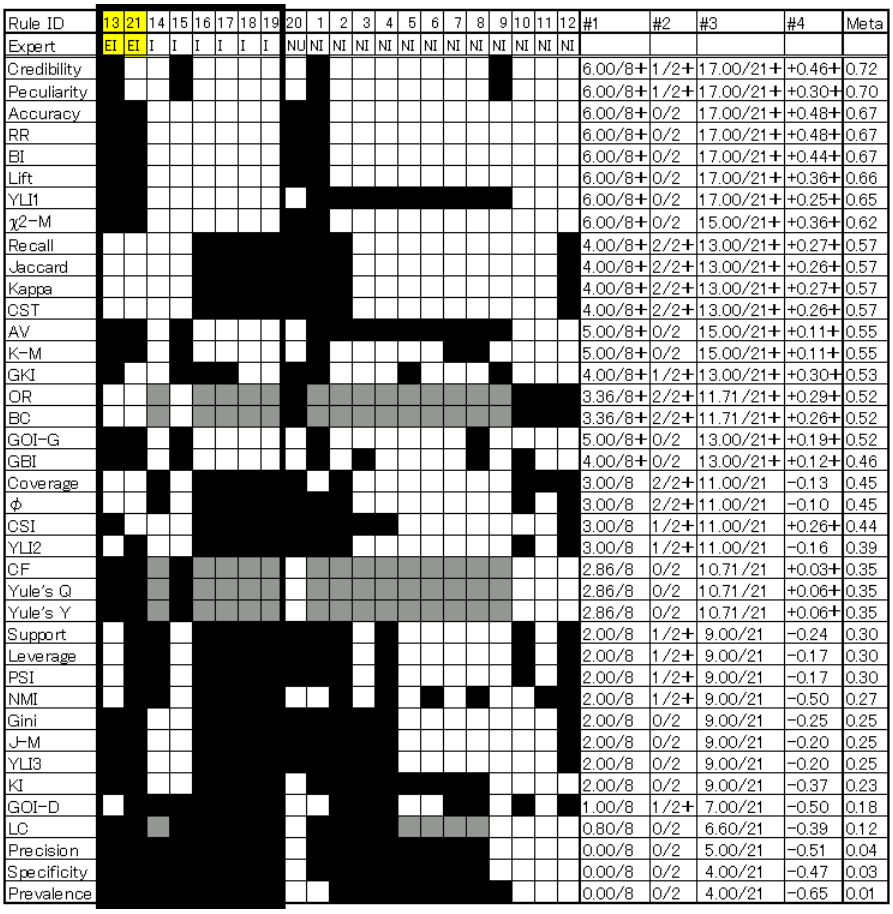


Fig. 2. The evaluation results by a medical expert and objective measures for the rules in second mining. See the caption of Fig. 1 for the details.

Table 3. The summary of the objective measures with the highest or the lowest performance in the first and second mining.

Top 3

Ranking	First Mining (Second Mining)	Second Mining (First Mining)
1	Recall(9)	Credibility(34)
2	Jaccard(9), Kappa(9), CST(9)	Peculiarity(9)
3	χ^2-M(8)	Accuracy(14), RR(11), BI(11)

Last 3

Ranking	First Mining (Second Mining)	Second Mining (First Mining)
37	GKI(15)	Precision(19)
38	NMI(30)	Specificity(22)
39	Prevalence(39)	Prevalence(39)

For the quantitative analysis, we defined four comprehensive criteria to evaluate the performance of an objective measure. #1: Performance on **I** (the number of rules labeled with **I** by the objective measure over that by the medical expert. Note that **I** includes **EI**). #2: Performance on **EI** (the number of rules labeled with **EI** by the objective measure over that by the medical expert). #3: Number-based performance on all evaluation (the number of rules with the same evaluation results by the objective measure and the medical expert over that of all rules). #4: Correlation-based performance on all evaluation (the correlation coefficient between the evaluation results by the objective measure and those by the medical expert). The values of these criteria are shown in the right side of Fig. 1 and 2. The symbol '+' besides a value means that the value is greater than that in case rules are randomly selected as **EI** or **I**. Therefore, an objective measure with '+' has higher performance than random selection does at least. To know the total performance, we defined the weighted average of the four criteria as a meta criterion; we assigned 0.4, 0.1, 0.4, and 0.1 to #1, #2, #3, and #4, respectively, according to their importance. The objective measures were sorted in the descending order of the values of meta criterion.

The results in the first mining in Fig. 1 show that Recall demonstrated the highest performance, Jaccard, Kappa, and CST did the second highest, and χ^2-M did the third highest. Prevalence demonstrated the lowest performance, NMI did the second lowest, and GKI did the third lowest. The results in the second mining in Fig. 2 show that Credibility demonstrated the highest performance, Peculiarity did the second highest, and Accuracy, RR, and BI did the third highest. Prevalence demonstrated the lowest performance, Specificity did the second lowest, and Precision did the third lowest. We summarized these objective measures in Table 3. As a whole, the following objective measures maintained their high performance through the first and second mining: Recall, Jaccard, Kappa, CST, χ^2-M, and Peculiarity. NMI and Prevalence maintained their low performance. Only Credibility changed its performance dramatically, and the other objective measures slightly changed their middle performance.

More than expected, some objective measures – Recall, Jaccard, Kappa, CST, χ^2-M, and Peculiarity – showed constantly high performance. They had comparatively many white cells and '+' for all comprehensive criteria. In addition, the mosaic-like patterns of white and black cells in Fig. 1 and 2 showed that the objective measures had almost complementary relationship for each other. These results and the comments on them by the medical expert imply that his interest consisted of not only the semantics in medical domain but also the statistical characteristics of rules. The combinational use of objective measures may be useful to reductively analyze such human interest and to recommend interesting rule candidates from various viewpoints through human-system interaction in medical KDD. One method to obtain the combination of objective measures is to formulate a function consisting of the summation of weighted outputs from different objective measures. Another method is to learn a decision tree using these outputs as attributes and the evaluation result by the medical expert as a class.

We think that although the experimental results are not enough to be generalized, they gave us two important implications: some objective measures will work at a certain level in spite of no consideration of domain semantics, and the combinational use of objective measures will help human-system interaction. We need more experiment with different medical datasets and experts. We also need to mathematically analyze objective measures to grasp the theoretical possibility and limitation of them [15]. These are our latest future work.

4 Conclusions and Future Work

This paper discussed how objective measures can contribute to detect interesting rules for a medical expert through the experiment using the rules on hepatitis. Recall, Jaccard, Kappa, CST, χ^2-M, and Peculiarity demonstrated good performance, and the objective measures used in this research had complementary relationship for each other. It was indicated that their combination will be useful to support human-system interaction. We have already started another experiment with a clinical dataset on meningoencephalitis and a different medical expert. We are also conducting the mathematical analysis of objective measures. We will consider the results of these empirical and theoretical investigations and will discuss the actual combinational use of objective measures.

Acknowledgments

We are grateful to three anonymous reviewers who substantially corrected and commented on this paper. This research was partially supported by the Grant-in-Aid for Scientific Research on the Priority Area (B),13131205, by the Ministry of Education, Science, and Culture for Japan.

References

1. Yao, Y. Y. Zhong, N.: An Analysis of Quantitative Measures Associated with Rules. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-1999 (1999) 479–488
2. Hilderman, R. J., Hamilton, H. J.: Knowledge Discovery and Measure of Interest. Kluwer Academic Publishers (2001)
3. Tan, P. N., Kumar V., Srivastava, J.: Selecting the Right Interestingness Measure for Association Patterns. Proceedings of International Conference on Knowledge Discovery and Data Mining KDD-2002 (2002) 32–41
4. 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 –. Proceedings of International Workshop on Active Mining AM-2002 in IEEE International Conference on Data Mining ICDM-2002 (2002) 97–102
5. Ohsaki, M., Sato, Y., Yokoi, H., Yamaguchi, T.: Investigation of Rule Interestingness in Medical Data Mining. Lecture Notes in Computer Science, Springer-Verlag (2004) will appear.

6. 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
7. Smyth, P., Goodman, R. M.: Rule Induction using Information Theory. in Piatetsky-Shapiro, G., Frawley, W. J. (eds.): Knowledge Discovery in Databases. AAAI/MIT Press (1991) 159–176
8. Hamilton, H. J., Fudger, D. F.: Estimating DBLearn's Potential for Knowledge Discovery in Databases. Computational Intelligence, 11, 2 (1995) 280–296
9. Hamilton, H. J., Shan, N., Ziarko, W.: Machine Learning of Credible Classifications. Proceedings of Australian Conference on Artificial Intelligence AI-1997 (1997) 330–339
10. Dong, G., Li, J.: Interestingness of Discovered Association Rules in Terms of Neighborhood-Based Unexpectedness. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-1998 (1998) 72–86
11. Gago, P., Bento, C.: A Metric for Selection of the Most Promising Rules. Proceedings of European Conference on the Principles of Data Mining and Knowledge Discovery PKDD-1998 (1998) 19–27
12. Gray, B., Orlowska, M. E.: CCAIIA: Clustering Categorical Attributes into Interesting Association Rules. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-1998 (1998) 132–143
13. Morimoto, Y., Fukuda, T., Matsuzawa, H., Tokuyama, T., Yoda, K.: Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases. Proceedings of International Conference on Very Large Databases VLDB-1998 (1998) 380–391
14. Freitas, A. A.: On Rule Interestingness Measures. Knowledge-Based Systems, 12, 5–6 (1999) 309–315
15. Lin, T. Y.: Attribute (Feature) Completion, – The Theory of Attributes from Data Mining Prospect, Proceedings of IEEE International Conference on Data Mining ICDM-2002 (2002) 282–289
16. Liu, H., Lu, H., Feng, L., Hussain, F.: Efficient Search of Reliable Exceptions. Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD-1999 (1999) 194–203
17. Jaroszewicz, S., Simovici, D. A.: A General Measure of Rule Interestingness. Proceedings of European Conference on Principles of Data Mining and Knowledge Discovery PKDD-2001 (2001) 253–265
18. Zhong, N., Yao, Y. Y., Ohshima, M.: Peculiarity Oriented Multi-Database Mining. IEEE Transaction on Knowledge and Data Engineering, 15, 4 (2003) 952–960

Process to Discovering Iron Decrease as Chance to Use Interferon to Hepatitis B

Yukio Ohsawa[1, 4] Hajime Fujie [2] Akio Saiura [3] Naoaki Okazaki,[4] and Naohiro Matsumura[5]

[1] Graduate School of Business Sciences, University of Tsukuba
osawa@gssm.otsuka.tsukuba.ac.jp
[2] Department of Gastroenterology, The University of Tokyo Hospital
[3] Department of Digestive Surgery, Cancer Institute Hospital, Tokyo
[4]Graduate School of Information Science and Technology, The University of Tokyo
[5] Faculty of Economics, Osaka University

Abstract Chance discovery is the process of human interaction with the environment for discovering events significant for making a decision. We executed the *double helix* process of chance discovery, on the blood-test data for hepatitis B, for obtaining scenarios telling when and how symptoms essential for treatment appear. In the double-helix process, the presented scenario maps are evaluated and fed back to the following cycles, in order to obtain novel and potentially useful knowledge for treatment. Due to the combination of the objective facts in the data and the subjective focus of the hepatologists' concerns emerging in this process, the relation between the changes of iron quantities due to iron-carrying proteins and the successful treatment of hepatitis B with interferon, has got clarified visually.

1. Introduction: Scenarios in the Basis of Critical Decisions

According to the definition of "chance" in [Ohsawa and McBurney 2003], i.e., an event or a situation significant for decision making, a chance occurs at the cross point of multiple scenarios because a decision is to select one scenario in the future. Here, a scenario is defined as a sequence of events to occur in a certain context. Generally speaking, a set of scenarios forms a basis of decision, in domains where the planning of event-sequence, rather than of a single event, affects the future profits significantly. For example, let us stand on the position of a surgeon looking at a time course or clinical course of symptoms observed in an individual patient. This surgeon should provide this patient with proper treatment at the right time. If he does so, the patient's disease may be cured. However, otherwise the patient's status might be worsened radically. This problem is to choose one from multiple scenarios. For example, suppose states4 and state5 in Eq. (1) mean two opposite situations.

Scenario1 = {state1 -> state2 -> state3 -> state4 (a normal condition)}. (**1**)

Scenario2 = {state 0 -> state2 -> state5 (a fatal condition)}.

Each event-sequence in Eq.(1) is called a *scenario* if the events in it share some common context. For example, Scenario1 is a scenario in the context of cure, and

Scenario2 is a scenario of the context of disease progress. The surgeon should choose an effective action at the time of state 2, in order to turn this patient to state3 and state4 rather than to state5, if possible. Such a state as state2, essential for making a decision, is a *chance* in this case.

Detecting an event at a crossover point among multiple scenarios, as state 2 above, and selecting the most valuable scenario at such a cross point means a chance discovery. Discovering a chance and taking it into consideration is required for making valuable useful scenarios, but proposing a number of scenarios even if some are useless is desired in advance for realizing chance discovery.

2. Scenario "Emergence" in the Mind of Experts

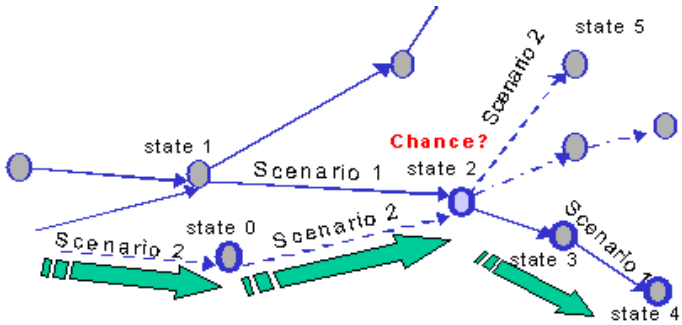


Fig. 1. A chance existing at the cross of scenarios. The scenario in the thick arrows emerged from Scenario1 and Scenario2.

Scenarios, presented from the viewpoint of each participant's environment, are bridged via ambiguous pieces of information about the different mental worlds they belong to. From these bridges, each participant recognizes situations or events which may work as "chances" to import others' scenarios to get combined with one's own. In the example of Eq.(1), a surgeon who almost gave up (paying attention only to Scenario2) may obtain a new hope in Scenario1 proposed by his colleague who noticed that state2 is common to both scenarios – only if it is still before or at the time of state2. Here, state2 is uncertain in that its future can potentially proceed into either of two directions, and this uncertainty can make state2 a chance, i.e., an opportunity not only a risk.

In this paper we applied a method for aiding the emergence of useful scenarios, by means of the interaction with real data using two tools of chance discovery, KeyGraph in [Ohsawa 2003b] and Polaris [Okazaki and Ohsawa 2003]. Here, KeyGraph with additional causal directions in the co-occurrence relations between the values of variables in blood-test data of hepatitis patients (let us call this a *scenario map*), and Polaris helps in dealing with data matching with the concern of experts, i.e. hepatologists here (we use "concern" for including its negative meaning of worrying, i.e., the concern with risks not only with opportunities, because a "chance" includes the meaning of a risk).

These tools aid in obtaining useful scenarios of a chrolonological course of hepatitis with/without treatment, reasonably restricted to understandable types of patients, from the complex data including the mixture of various scenarios. The scenarios obtained for

hepatitis were evaluated by two hepatologists, a surgeon and a physician, as useful in finding an exceptionally good chance to treat hepatitis patients. This is a strong advantage for medical experts, because it is very hard to observe a complete history of critical progress or of an exceptionally successful treatment.

3. Tools for Accelerating the Process of Chance Discovery

3.1 The Double Helix Model: The Process of Chance Discovery

In the studies on chance discovery, the discovery process has been supposed to follow the Double Helix (DH) model [Ohsawa 2003a] as in Fig. 2. The DH process starts from a state of user's mind concerned with catching a new chance. This *concern* is reflected to acquiring *object-data* to be analyzed by data-mining tools specifically designed for chance discovery. Looking at the visualized result of this analysis, possible scenarios and their values rise in each user's mind. Then users gather to be participants of a co-working group for chance discovery, sharing the same visual result. Then, words corresponding to the bridges among the contexts in the mind of participants are visualized in the next step, where visual data mining is applied to the *subject-data*, i.e., the text data recording the thoughts and opinions in the discussion. Via the participants' understanding of these bridges, the islands get connected and form novel scenarios. By this time, the participants may have discovered chances on the bridges, because each visualized island corresponds to a certain scenario familiar to some of the participants and a bridge means a cross-point of those familiar scenarios. Based on these chances, the user(s) may make actions, or simulate actions in a virtual environment, and may obtain concerns with new chances. Then, the helical process returns to the initial step of the next cycle.

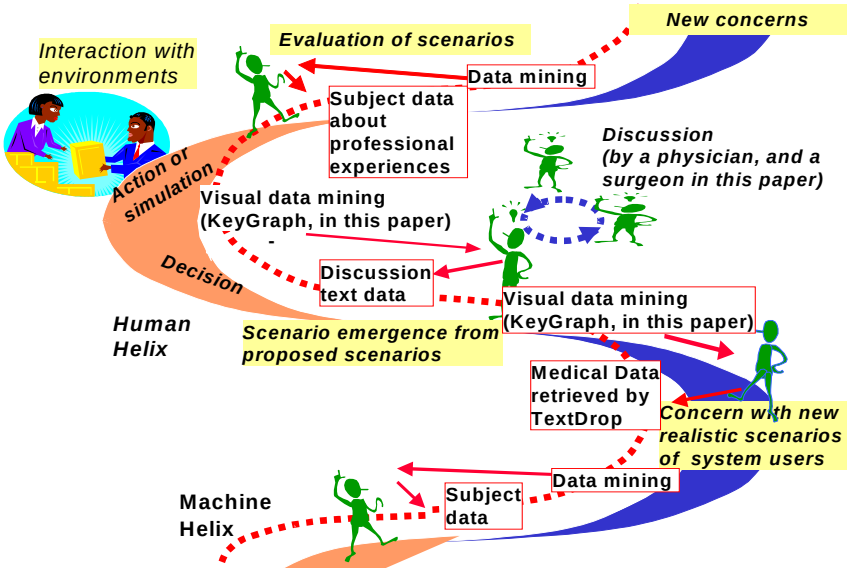


Fig.2. The DH Model: A process model of chance discovery.

In this paper, the DH process was applied to obtaining hepatitis scenarios. Users watched and discussed on KeyGraph [Ohsawa 2003b], working on Polaris an interactive visual interface to accelerate the DH process [Okazaki and Ohsawa 2003], applied to the subject-data and the object-data in the process. Users thought and talked about scenarios the diagram may imply, looking at the visual output of KeyGraph.

3.2 KeyGraph for Visualizing Scenario Map

KeyGraph is a tool for visualizing the map of event relations in the environment, in order to aid the process of chance discovery. If the environment represents a discussion room, an event may represent a word in by a participant. By visualizing the map where the words appear connected in a graph, one can see the overview of participants' interest. Suppose a text (string-sequence) *D* is given, describing an event-sequence sorted by time, with periods ('.') inserted at the parts corresponding to the moments of major changes. For example, let text *D* be:

D = "*Mr. A: Customers decreased in the market of general construction.*
Mr. B: Yes… Our company, making concrete and steel, is in this terrible trend.
Mr. C: This state of the market induces a further decrease of customers. Your company may have to introduce restructuring for satisfying customers.
Mr. B: Then the company can reduce the price of concrete and steel for construction.
M .D: But, it may also reduce the power of this company." (2)

In the case of a document as in Eq.(2), periods are put at the end of each sentence. However, a period can be put only at the end of each message if the user likes to. In the case of a sales (Position Of Sales: POS) data, periods are put in the end of each basket. *KeyGraph*, of the following steps, is applied to *D* ([Ohsawa 2003b] for details).

***KeyGraph-Step* 1:** Clusters of co-occurring frequent items (words in a document, or events in a sequence) are obtained as the bases, called *islands*. That is, items appearing many times in the data (e.g., the word "market" in Eq.(2)) are depicted with black nodes, and each pair of these items occurring often in the same sequence unit (a *sentence* in a document, a bought set of items in each basket in sales data, etc, between two delimiters) is linked to each other, e.g., `` market - customers - decrease" for Eq.(2) with a solid line. Each connected graph obtained here forms one island, implying the existence of a common context underlying the belonging items.

***KeyGraph-Step* 2:** Items which may not be so frequent as the black nodes in islands but co-occurring with multiple islands, e.g., ``restructuring" in Eq.(2), are obtained as *hubs*. A path of links connecting islands via hubs is called a *bridge*. If a hub is rarer than black nodes, it is colored in a different color (e.g. red or white). We can regard such a new hub as a candidate of *chance*, i.e., items significant (assertions in a document, or latent demand in a POS data) with respect to the structure of item-relations.

In the example of Fig. 3, the result of KeyGraph on Polaris, the island {customers} means the context asserting that the importance of customers is established, and the island of {steel, concrete, company} shows the established business context of the company. The bridge "restructuring" shows the company may introduce restructuring, e.g. firing employees, for winning the good feeling of customers. The word "restructuring" might be rare in the communication of the company staffs, but this expresses the concern of the employees.

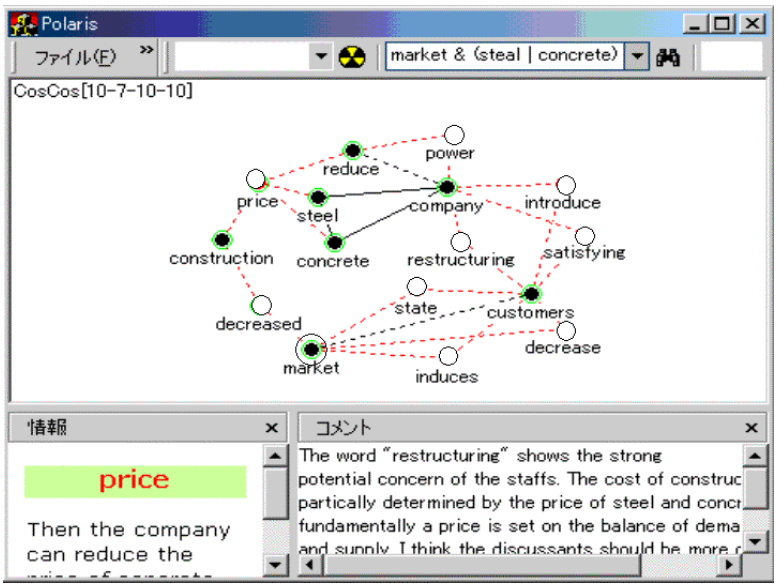


Fig.3. An example of *KeyGraph on Polaris*: Islands are obtained from *D* in Eq.(2), each including event-set {market}, {steel, concrete, company}, {customers} etc. The double-circled and the white nodes show frequent and rare words respectively, forming the hubs of bridges.

On Polaris, user can feedback new concern, i.e., the interest in new object-data by entering a new search query such as "market & (concrete OR steel)" in the tool bar shown in Fig.3, as mentioned in 3.3. In this case, the user is concerned with the market of concrete and steel, in the chat among Mr. A, Mr. B, Mr. C, and Mr. D. Further more, if user becomes concerned with a topic shown by a node in the graph, he/she can click on the node to read the sentences including words corresponding to the node (the lower left window in Fig.3). The set of extracted sentences becomes the input to KeyGraph, if user selects going to the next cycle of the DH process.

3.3 Retrieving Data Relevant to User's Concern

Polaris has a function for Boolean-selection of the part of data corresponding to users' concern described in a Boolean formula, e.g,

$$\textit{concern} = \text{"(product A | product B) \& product C \& !product D".} \qquad (3)$$

For this Boolean expression of user's concern, Polaris obtains a focused data made of baskets including product A or product B, and product C, but not including product D. This becomes a revised input to KeyGraph, at the step where user acquires a new concern and seeks a corresponding data, in the DH process. This is useful if the user can express his/her own concern in Boolean formula as in (2). The concern of a user/users might be more ambiguous, especially in the beginning of the DH process. In such a case, the user is supposed to enter the formula "as much as" specifically representing one's own concern. Having KeyGraph, query-search function, and the clickable nodes, Polaris aids the user to follow the procedure as listed below, a sped up DH process.

Step 1) Extract a part of the object-data on Polaris, corresponding to the user's concern with events or with the combination of events expressed in a Boolean formula.
Step 2) Apply KeyGraph to the data in Step 1) in order to visualize the map representing the relations between events, and attach causal arrows as much as possible, with the help of experts of the domain. Thus, a scenario map is produced (see 4.1).
Step 3) Manipulate KeyGraph in a group work with domain experts, as follows:
3-1) Move/remove nodes and links in KeyGraph, considering their importance.
3-2) Write down participants' comments about scenarios, proposed on KeyGraph.
Step 4) Read or visualize (with KeyGraph) the subject-data, i.e., the participants' comments obtained in 3-2), and extract noteworthy and realistic scenarios.
Step 5) Execute or simulate (draw concrete images of the future) the scenarios obtained in Step 4), and, based on this experience, refine the statement of the new concern in concrete words. Go to Step 1).

Table 1. The DH process aided by KeyGraph and Text Drop

4. Results for the Diagnosis Data of Hepatitis

4.1 The Hepatitis Data

The following shows the style of data obtained from blood-tests of hepatitis cases. Each event represents a pair, of a variable and its observed value. That is, an event put as "a_b" means the value of variable a was b. For example, T-CHO_high (T-CHO_low) means T-CHO (total cholesterol) was higher (lower) than the predetermined upper (lower) bound of normal range. Note that the lower (higher) bound of each variable was set higher (lower) than values defined in hospitals, in order to be sensitive to the moments the variable takes an unusual value. Each line delimited by '.' represents the sequence of blood-test results for one patient. As in Eq.(3), we regard one patient as a unit of co-occurrence of events. As a result, the scenario of a typical chrolonological course is expected to appear as a connected path in the scenario map obtained with KeyGraph.

Case1 = {event1, event2, ….., event m1}.
Case2 = {event 2, event 3, …., event m2}. (**3**)
Case3 = {event 1, event 5, …, event m3}.

For example, suppose we have the data in Table 2, where each event means a value of a certain attribute of blood, e.g. GPT_high means the status of a patient whose value of GPT exceeded its upper bound of normal range. Each period ('.') represents the end of one patient's case. If the doctor is interested in patients having experiences of both GTP_high and TP_low, then the doctor can enter "GTP_high & TP_low" to Polaris in Step 1) in Table 1 and get the italic lines as an input to KeyGraph in Step 2).

By applying KeyGraph to this data, the following components are obtained:

- *Islands of events*: A group of events co-occurring, i.e. all of which are experienced by many patients. The doctor is expected to know a patient's status corresponding to each island, because events in an island are frequent.
- *Bridges across islands*: A patient may switch from one island to another, in the progress of the disease and in the treatment.

GPT_high TP_low TP_low GPT_high TP_low GPT_high TP_low.
ALP_low F-ALB_low GOT_high GPT_high HBD_low LAP_high LDH_low TTT_high ZTT_high ALP_low CHE_high D-BIL_high F-ALB_low F-B_GL_low.
GOT_high GPT_high LAP_high LDH_low TTT_high ZTT_high F-ALB_low F-B_GL_low G_GL_high GOT_high GPT_high I-BIL_high LAP_high LDH_low TTT_high ZTT_high GOT_high GPT_high LAP_high LDH_low TP_low TTT_high ZTT_high B-type CAH2A
D-BIL_high F-CHO_high GOT_high GPT_high K_high LAP_high LDH_low T-CHO_high TP_low UN_high T-BIL_high ALP_high D-BIL_high GOT_high GPT_high I-BIL_high LDH_high T-BIL_high B-type CAH2B.

Table 2. An Example of Blood Test Data for KeyGraph

The data dealt with here was 771 cases, taken from 1981 through 2001. Fig. 4 is a KeyGraph obtained, for cases of progressive hepatitis B. The causal arrows in Step 2) of the DH Process, which does not appear in the original KeyGraph of Ohsawa (2003), depict approximate causations. If the direction from X to Y is apparent for an expert, the expert puts the arrow in the graph. If this direction is not apparent, the two results of KeyGraph are compared, one for the data retrieved for entry "X" with Polaris, and the other for the data retrieved for entry "Y." If the expert judges the former includes more causal events than the latter, X is regarded as a preceding event of Y in a scenario.

The order of causality and the order of occurrence time may be different. For example, the upper bound of ZTT may be set easier to exceed than that of G_GL, which makes ZTT_high appear before G_GL_high although ZTT_high is a result of G_GL_high. In such a case, we compare the results of KeyGraph, one for data including G_GL_high and the other for data including ZTT_high. If the latter includes F1, an early stage of fibrosis, and the former includes F2, a later stage, we understand G_GL_high was preceding ZTT_high. We call a KeyGraph with the arrows made in this way, a *scenario map*.

4.2 The Double Helix Process Executed for Hepatitis B

Two hepatologists, a surgeon and a physician joined the DH process in Table 1. In the preliminary cycle, we had understood the scenario map became a mixture of cases according to the subject-data (comments) from hepatologists. Thus we separated the data into each scenario, i.e., AH (acute hepatitis), CAH (chronic aggressive hepatitis) etc., by spotlighting events typical to each scenario using Polaris. Below we present how we reached a significant discovery in the continued process.

1) The obtained KeyGraph and how users understood it: For the cases of AH, the obtained scenario map matched with the background knowledge of the hepatologists. For CAH, some of their tacit experiences were externalized. For example, a quick sub-process from LDH_high to LDH_low (LDH: lactate dehydrogenase) shown in the scenario map had been sometimes observed in the introductory steps of fulminant hepatitis B, but has not been published because of the rareness.

2) Deepened concerns, the focused object data, and new scenarios: The results in 1) drove us to see more into the details of the progress of fibrosis denoted by F1, F2, F3, and F4 (or LC: liver cirrhosis). Fig. 4 is the scenario map for hepatitis B, in the spotlights of F1, F2, F3, and F4 (LC), i.e., for the data extracted for entry "type-B & (F1 | F2 | F3 | F4 | LC)"with Polaris. This result shows a novel connection among the basic state transitions in fibrosis listed from a. to e. below, useful for understanding the status of a patient at an arbitrary time.

a. A chronic active hepatitis sometimes turns into a severe progressive hepatitis and then to cirrhosis or cancer, in the case of hepatitis B.
b. The final states of critical cirrhosis co-occur with kidney troubles, and grows to malignant tumors (cancer) with the deficiencies of white blood cells.
c. Recovery is possible from the earlier steps of fibrosis.
d. LDH_low after high LDH_high can be a sign of fulminant hepatitis.
e. The low Fe (iron) level with cirrhosis can be a turning point to the recovery of liver.

In Fig.4, the appearance of FE_low, on the *only* bridge from cirrhosis to recovery, can be useful for finding the optimal timing to treat a patient of hepatitis B, and seems relevant to [Rubin et al 1995] having suggested iron reduction may improve the response of chronic hepatitis to interferon. However, Fig.4 does not include "interferon." A possible interpretation of this figure is that iron-reduction has been applied to patients. An iron reduction is to take blood out of the body, in order clean the iron pool stored too much in the liver. Hayashi et al (1995) also showed iron reduction improves the condition of a patience of hepatitis C. However, the problem is that doctors rarely use iron reduction for hepatitis B, so the FE_low in Fig.4 does not seem to mean iron reduction.

3) The subject-data from the discussions by users: The hepatologists made a discussion looking at Fig.4. From this discussion, the words were regarded as new subject-data and visualized with KeyGraph. From the result, shown in Fig.5, we can summarize the comments of hepatologists. First, the iron reduction for hepatitis of type B is not realistic. They explained transferrin, the protein denoted by F_B-GL in figures, carried iron to/from the liver and arranged the quantity of iron. Assuming iron reduction was not used, the discussion focused attention to "I-BIL_low" on the way to the "recovery" in Fig.4. This decrease in I-BIL (indirect bilirubin) seems to mean hem, a substance from hemoglobin to increase bilirubin, was carried back to the liver as in the normal status. This carrying is done by the protein F_A2-GL. Therefore, we can guess proteins such as F_A2-GL were working actively on the way to recovery.

The discussion then paid attention to "FE_high" near "FE_low" in Fig.4, and suggested it is possible that FE_low and FE_high means ASC (asymptomatic carriers). In the case of an ASC, a clear symptom of progress is hardly observed but the patient may suffer from sudden worsening, and sometimes the wrong progress and recovery interleaves to lead to a fatal condition. However, sometimes a patient of ASC recovers. Because of the missing symptoms in ASC, it is important to discover events significant for deciding to make treatment action, i.e. the chance to drive into the better scenario.

4) The effect of interferon, extracted as a result: Reflecting the new concern in 3), we extracted the blood-test data of patients whose history includes "type-B & FE_low & FE_high," and regarded each time of test as one sentence for KeyGraph. The result of KeyGraph for this extracted data is Fig.6. We find the cycle where FE_high and FE_low appear, and the events below occur in turns. That is, lipoprotein metabolism is weakened, with the decrease in hemoglobin (HBD_low). Then, iron and platelets in the blood decrease. Following this, kidney troubles and the decrease in amylase (AMY_low), imply liver cirrhosis. Finally, cancer and anomalous conditions of coagulation/fibrinolysis appear, and treatment operations or the existence of cancer is implied by the high value of amylase (AMY_high) [Miyagawa et al 1996, Chougle 1992]. Then, following the activation of F_A2-GL and F_B-GL, the iron in the blood increases, and the patient returns to the decrease in hemoglobin. This cyclic process means the scenario of progress

come after treatments. In Fig.6, we find the recovery of liver via FE_low, as was shown in Fig.4, has been achieved by the use of interferon. This corresponds to the result of [Rubin et al 1995], and we can conclude interferon works for hepatitis B, if used at the best timing i.e., at the time of iron decrease.

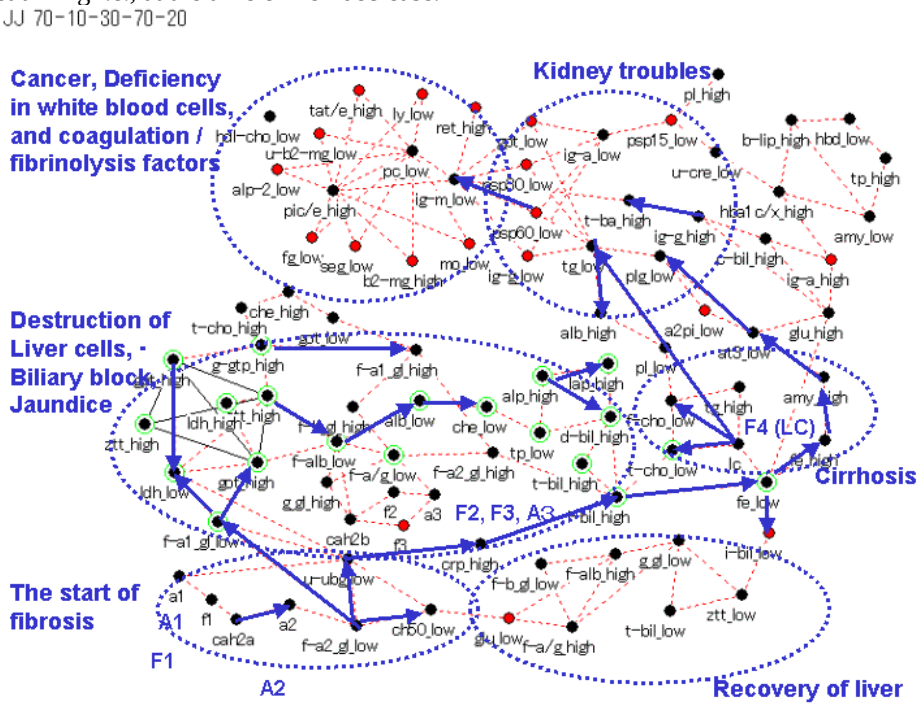


Fig. 4 The scenario map for hepatitis B, for the spotlights of F1, F2, F3, and F4 (LC).

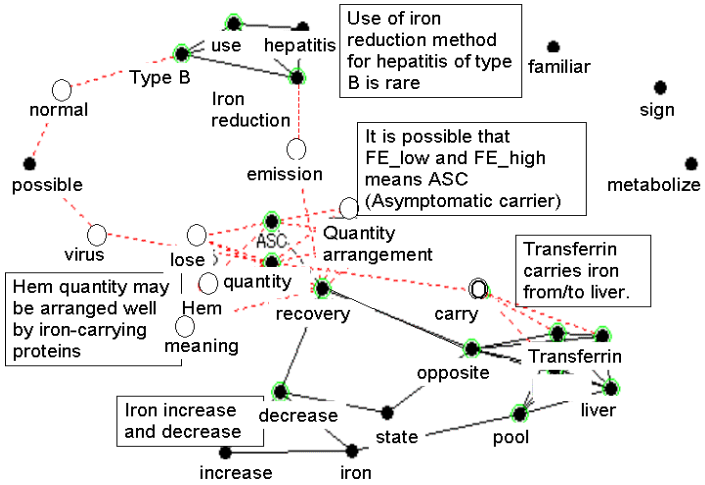


Fig.5 KeyGraph for the comments of hepatologists looking at Fig.6.

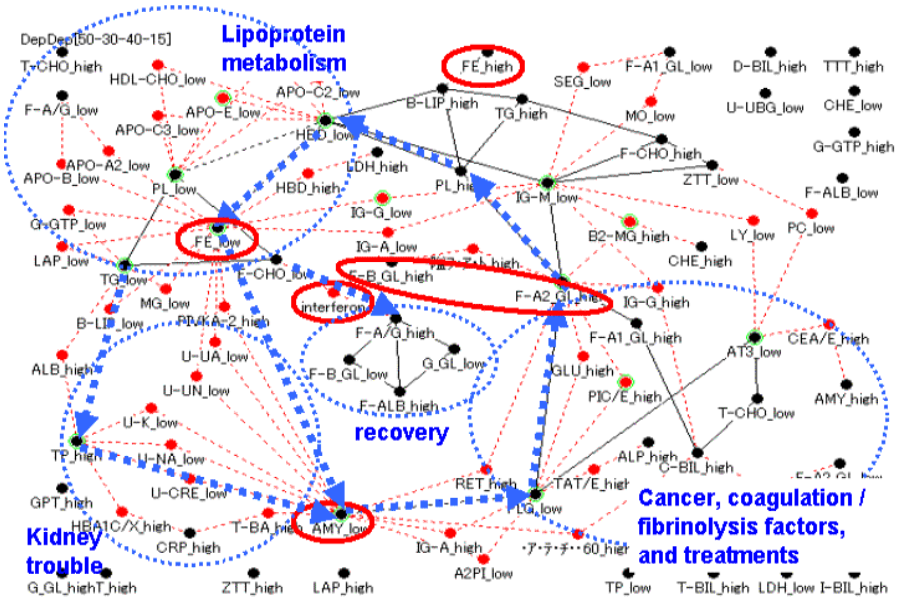


Fig.6. The latest scenario map for hepatitis B.

5. Conclusions

The scenarios of the progress of and the recovery form hepatitis B were discovered. The relevance of iron and the disease has been implied in previous studies, and this paper clarified that the decrease in iron means a good timing for using interferon to hepatitis B. This discovery is the effects of double helix (DH) process accelerated by using Polaris.

References

Chougle A; Hussain S; Singh PP; Shrimali R., 1992, Estimation of serum amylase levels in patients of cancer head and neck and cervix treated by radiotherapy, *Journal of Clinical Radiotherapy and Oncology*. 1992 Sept; 7(2): 24-26

Hayashi, H., T. Takikawa, N. Nishimura, M. Yano, T. Isomura, and N. Sakamoto. 1994. Improvement of serum aminotransferase levels after phlebotomy in patients with chronic active hepatitis C and excess hepatic iron, *American Journal of Gastroenterol.* 89: 986-988

Miyagawa S, Makuuchi M, Kawasaki S, Kakazu T, Hayashi K, and Kasai H., 1996, Serum Amylase elevation following hepatic resection in patients with chronic liver disease., *Am. J. Surg.* 1996 Feb;171(2):235-238

Ohsawa Y and McBurney P. eds, 2003, Chance Discovery, Springer Verlag

Ohsawa Y., 2003a, Modeling the Process of Chance Discovery, Ohsawa, Y. and McBurney eds, *Chance Discovery*, Springer Verlag pp.2—15 (2003)

Ohsawa Y, 2003b, KeyGraph: Visualized Structure Among Event Clusters, in Ohsawa Y and McBurney P. eds, 2003, *Chance Discovery*, Springer Verlag: 262-275

Okazaki N and Ohsawa Y, 2003, Polaris: An Integrated Data Miner for Chance Discovery, In *Proceedings of The Third International Workshop on Chance Discovery*, Crete, Greece

Rubin RB, Barton AL, Banner BF, Bonkovsky HL., 1995, Iron and chronic viral hepatitis: emerging evidence for an important interaction. in Digestive Diseases

Preliminary Analysis of Interferon Therapy by Graph-Based Induction

Tetsuya Yoshida **, Warodom Geamsakul*, Akira Mogi*,
Kouzou Ohara*, Hiroshi Motoda*, Takashi Washio*,
Hideto Yokoi**, and Katsuhiko Takabayashi**

* Institute of Scientific and Industrial Research, Osaka University, JAPAN
{yoshida,warodom,mogi,ohara,motoda,washio}@ar.sanken.osaka-u.ac.jp
** Division for Medical Informatics, Chiba University Hospital, JAPAN
yokoi@telemed.ho.chiba-u.ac.jp,takaba@ho.chiba-u.ac.jp

Abstract. A machine learning technique called Graph-Based Induction (GBI) extracts typical patterns from graph structured data by stepwise pair expansion (pairwise chunking). Because of its greedy search strategy, it is very efficient but suffers from incompleteness of search. GBI has been extended to Beam-wise GBI (B-GBI) by incorporating a beam search to improve its search capability without imposing much computational complexity. It has also been incorporated into Decision Tree Graph-Based Induction (DT-GBI) to extract substructures (discriminative patterns) to construct a decision tree for graph-structured data. We applied GBI (both B-GBI and DT-GBI) to analyze the effectiveness of interferon therapy in the hepatitis dataset provided by Chiba University Hospital. Response to interferon therapy in each patient is used as a class label and measurement patterns that are strongly correlated with the response were extracted in the experiments. In the first experiment, decision trees are constructed by DT-GBI for discriminating the patients from whom the hepatitis virus disappeared by interferon therapy and the patients from whom the virus continued to exist. In the second experiment, descriptive patterns were extracted by B-GBI and examples of extracted patterns are reported. The preliminary results of experiments are reported in this paper.

1 Introduction

Graph-Based Induction (GBI) [11] is a technique which was devised for the purpose of discovering typical patterns in a general graph structured data by recursively chunking two adjoining nodes. It can handle a graph structured data having loops (including self-loops) with colored/uncolored nodes and edges. GBI is very efficient because of its greedy search but suffers from incompleteness of search. GBI has been extended to Beam-wise GBI (B-GBI) to improve its search capability without imposing much computational complexity by incorporating a

* Current address: Gradudate School of Information Science and Technology, Hokkaido University, JAPAN

beam search [3]. It has also been incorporated into a method called Decision Tree Graph-Based Induction (DT-GBI), which constructs a classifier (decision tree) for graph-structured data [8]. A pair extracted by GBI, which consists of nodes and the edges between them, is treated as an attribute and the existence/non-existence of the pair in a graph is treated as its value for the graph. Although initial pairs consist of two nodes and the edge between them, attributes useful for classification task are gradually grown up into larger pair (subgraphs) by applying chunking recursively.

We have applied both B-GBI and DT-GBI for the analysis of hepatitis dataset provided by Chiba University Hospital [2, 9]. This paper reports yet another analysis of the hepatitis dataset by GBI (both B-GBI and DT-GBI) with respect to the effectiveness to interferon therapy. Response to interferon therapy is used as class label and two experiments were conducted for extracting discriminative patterns and descriptive patterns for interferon therapy using only the time sequence data of blood test and urinalysis. In the first experiment, decision trees are constructed by DT-GBI for discriminating the patients from whom the hepatitis virus disappeared by interferon therapy and the patients from whom the virus continued to exist. In the second experiment, descriptive patterns are extracted by B-GBI and examples of extracted patterns are reported. The preliminary results of experiments are reported in this paper.

There are some other analyses already conducted and reported on this dataset. [10] analyzed the data by constructing decision trees from time-series 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. [7] used multi-scale matching to compare time-series data and clustered them using rough set theory. [4] also clustered the time-series data of a certain time interval into several categories and used a standard classifier. These analyses examine the temporal correlation of each inspection separately and do not explicitly consider the relations among inspections. Thus, these approaches are not categorized to fall in structured data analysis. Furthermore, our previous analysis showed that the prediction accuracy of DT-GBI is well comparable to other approaches. For instance, the error rate of [10] in discriminating patients with cirrhosis from those with the other fibrosis stages is 11.8% whereas that of DT-GBI is 12.5%, and the error rate of the approach in [1] applied to discriminating patients with hepatitis type B from type C is 26.2%[1] whereas that of DT-GBI is 20.3%.

2 Graph-Based Induction Revisited

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, assumptions are made that typical patterns represent some concepts and "typicality" is characterized by the

[1] Personal communication.

GBI(G)
  Enumerate all the pairs P_{all} in G
  Select a subset P of pairs from P_{all} 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
    $P := P \cup \text{GBI}(G_c)$
  return P

Fig. 1. Algorithm of GBI

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. Because 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 Fig. 1.

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

Since the search in GBI is greedy and no backtracking is made, which patterns are 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 [3] to relax this problem, increase the search space, and extract more discriminative patterns while 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 Decision Tree by GBI (DT-GBI)

If pairs are expanded in a step-wise fashion by 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. 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. Chunking is applied for a specified number of times at each node of a decision tree and the chunked pairs grow up into larger nodes in size. Thus, although initial pairs consist of only two nodes and one edge between them, attributes useful for classification

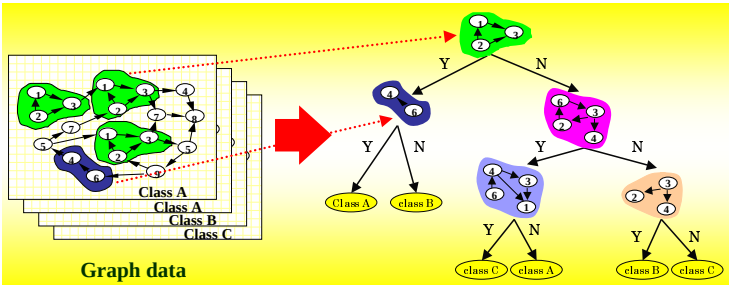


Fig. 2. Decision tree for classifying graph-structured data

task are gradually grown up into larger pairs (subgraphs) by applying chunking recursively. The above process is illustrated in Fig.2.

3 Preliminary Analysis of Interferon Therapy by GBI

An interferon is a medicine to deactivate and kill hepatitis virus and it is said that the smaller the amount of virus is, the more effective interferon therapy is. Unfortunately, the dataset provided by Chiba University Hospital does not contain the examination record for the amount of virus since it is expensive. However, experts (medical doctors) decide when to administer an interferon by estimating the amount of virus from the results of other pathological examinations. In the following experiments we hypothesized that the amount of virus in a patient was almost stable for a certain duration just before the interferon injection in the dataset. Response to interferon therapy was judged by a medical doctor for each patient, which was used as the class label for interferon therapy. The class labels specified by the doctor for interferon therapy are summarized in Table 1. Note that the following experiments were conducted for the patients with label R (38 patients) and N (56 patients). Medical records for other patients were not used.

Table 1. class label for interferon therapy

label	
R	virus disappeared (Response)
N	virus existed (Non-response)
?	no clue for virus activity
R?	R (not fully confirmed)
N?	N (not fully confirmed)
??	missing

3.1 Data Preprocessing

In phase 1, a new reduced data set is generated because the data of visit is not synchronized across different patients and the progress of hepatitis is considered slow. The data set provided is cleansed [2], and the numeric attributes are averaged over two-week interval and for some of them, standard deviations are calculated over six month interval and added as new attributes. Mathematical average is

[2] Letters and symbols such as H, L, +, or - are deleted from numeric attributes.

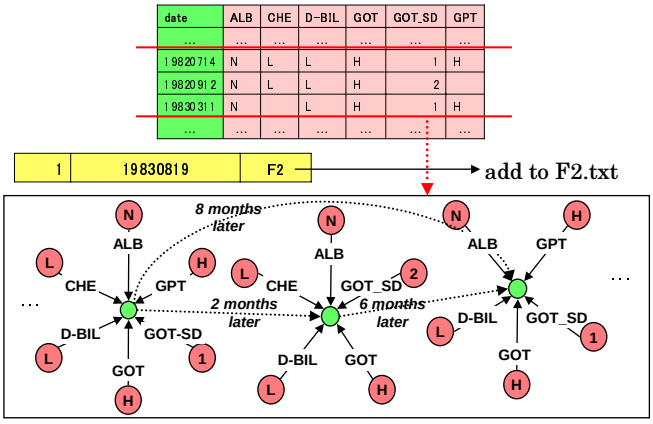


Fig. 3. An example of converted graph structured data

taken for numeric attributes and maximum frequent value is used for nominal attributes over the interval. Further, numerical values are discretized when the normal ranges are given. In case there are no data in the interval, these are treated as missing values and no attempt is made to estimate these values. At the end of this phase, reduced data is divided into several files so that each file contains the data of each patient.

In phase 2, data in the range of 90 days to 1 day before the administration of interferon were extracted for each patient. Furthermore, although original dataset contains hundreds of examinations, feature selection was conducted with the expert to reduce the number of attributes. Thus, we used the following 25 attributes: ALB, CHE, D-BIL, GOT, GOT_SD, GPT, GPT_SD, HCT, HGB, I-BIL, ICG-15, MCH, MCHC, MCV, PLT, PT, RBC, T-BIL, T-CHO, TP, TTT, TTT_SD, WBC, ZTT, and ZTT_SD.

In the last phase of data preparation, one patient record is mapped into one directed graph. An assumption is made that there is no direct correlation between two sets of pathological tests that are more than a predefined interval (here, 8 weeks) apart. Fig. 3 shows an example of converted graph structured data. In this figure, a star-shaped subgraph represents values of a set of pathological examination in the two-week interval. The center 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 (processed result) of the examination. The edge linking two hypothetical nodes represents time difference. Note that we hypothesized that each pathological condition in the extracted data could directly affect the pathological condition just before the administration. To represent this dependency, each subgraph was directly linked to the last subgraph in each patient. Table 2 shows the size of graphs after the data conversion.

Table 2. Size of graph structured data (interferon)

class label	R	N	Total
No. of graphs	38	56	94
Avg. No. of nodes	77	74	75
Max. No. of nodes	123	121	123
Min. No. of nodes	41	33	33

3.2 Analysis of Interferon Therapy by DT-GBI

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 [5] for finding discriminative patterns after chunking. A decision tree was constructed by applying chunking 20 times at every node of a decision tree (N_e=20). Pessimistic postpruning was conducted to construct a decision tree with higher prediction accuracy by setting the confidence level to 25% as in C4.5 [6].

Table 3. Error rate (%)

cycle of 10 CV	N_e=20
1	18.65
2	19.69
3	19.17
4	20.73
5	22.80
6	23.32
7	18.65
8	19.17
9	19.69
10	21.24
Average	22.60
Standard Deviation	1.57

We evaluated the prediction accuracy of decision trees constructed by DT-GBI by the average of 10 cycles of 10-fold cross-validation. Thus, 100 decision trees were constructed in total. In the first cycle of 10 fold cross validation, search beam width b was varied from 1 to 15. The prediction error rates reached the lowest level (18.75%) when $b = 3$ and remained the same thereafter. Thus, in the remaining nine cycles of 10-fold cross validation, we set the beam width to 3 when running DT-GBI. The results are summarized in Table 3 and the overall average error rate was 22.60%. Contingency tables for test data in cross validation are shown in Table 4 for the best and the worst trees out of the 100 decision trees constructed.

By regarding that class label R (Response) as positive and class label N (Non-response) as negative, decision trees constructed by DT-GBI mostly classified the patients with class label N as "N"correctly. This will contribute to reducing the fruitless interferon therapy of patients. On the other hand, some of the patients with class label R were also classified as "N". This may lead to miss the opportunity of curing patients with interferon therapy. From the above results, it was revealed that the decision trees constructed by DT-GBI tend to have more false negative for predicting the effectiveness of interferon therapy.

Table 4. Examples of contingency table

Actual Class	Predicted Class			
	Best		Worst	
	N	R	N	R
N	6	0	2	4
R	1	3	2	2

Two examples of decision trees selected out of 100 constructed are shown in Fig.4 and Fig.5. Fig.4 is selected from the decision trees with the best prediction accuracy and Fig.5 is selected from the ones with time-correlated patterns, prediction accuracy being about the average of 100. Patterns at the upper nodes in these trees are shown in Figs. 6, 7, 8, 9. Although the structure of decision tree in Fig.4 is simple, its prediction accuracy was actually good (error rate=10%). Furthermore, since the pattern shown in Fig.6 was used at the root node of many decision trees, it is considered sufficiently discriminative for classifying patients for whom interferon therapy was effective (with class label R). However, although

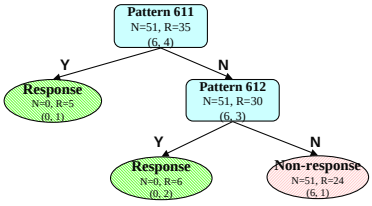


Fig. 4. Example of constructed decision tree

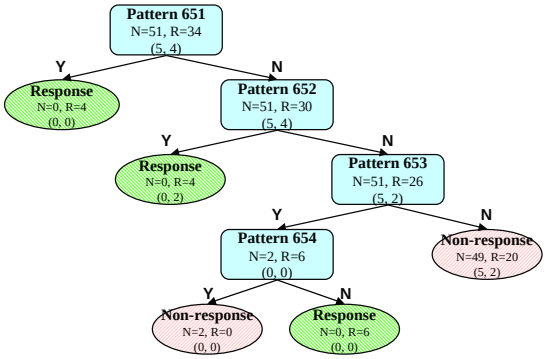


Fig. 5. Example of constructed decision tree

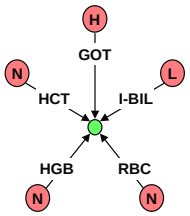


Fig. 6. pattern 611 (if exist, then R)

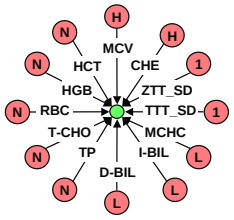


Fig. 7. pattern 612 (if exist, then R)

these patterns are discriminative in terms of information gain, the extracted decision trees were rather hard to interpret by a medical doctor because he could not see clear difference between the two groups of patterns each characterizing class label R and N. The interval used to calculate standard deviation to take fluctuation into account may be too long.

Unfortunately, only several patterns contain time interval edges such as the one shown in Fig. 8 and it was unable to investigate how the change or stability of blood test will affect the effectiveness of interferon therapy.

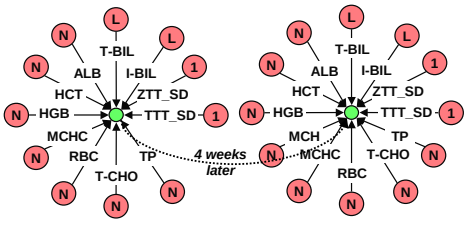


Fig. 8. pattern 651 (if exist, then R)

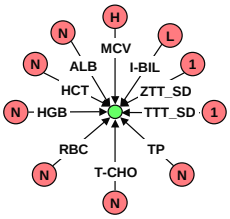


Fig. 9. pattern 652 (if exist, then R)

Table 5. Extracted patterns

	94 Graphs in Table 2				41 Graphs with patterns in Figs. 10,11,12			
	only R	only N	common	Total	only R	only N	common	Total
No. of patterns	2604	3468	5467	11539	1392	4587	3678	9657
Max. No. of nodes per pattern	35	38	38	38	39	42	39	42
Ave. No. of nodes per pattern	15.4	15.0	10.9	13.1	15.9	16.3	10.8	14.1

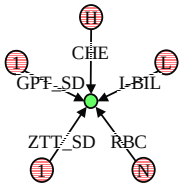


Fig. 10. pattern a1

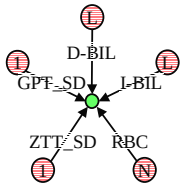


Fig. 11. pattern a2

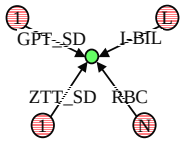


Fig. 12. pattern a3

3.3 Analysis of Interferon Therapy by B-GBI

We applied B-GBI to extract descriptive patterns from the patients analyzed in subsection 3.2. B-GBI was terminated when the support of all the extracted patterns [3] became less than 0.1. Beam width b was set to 3 as in subsection 3.2. The extracted patterns were divided into 3 groups: 1) patterns included only in the patients with class label R, 2) those with class label N, and 3) those with both label R N (these groups are called only R, only N, common, respectively). The number and size of the extracted patterns from the graphs in Table 2 are summarized in the left-hand side of Table 5.

The decision tree constructed by DT-GBI is rather unbalanced as shown in Fig. 5. This is because the patterns with large discriminative power (information gain) have a relatively small support. Small support means that these patterns are specific to some data and does not have sufficient generalization capability.

We, therefore, searched for the patterns by B-GBI in terms of not only the discriminative power but also the support. Patterns which were included both in the data with R and N were sorted out from the extracted patterns to focus on the patterns with a large support. The patterns were then sorted in descending order of information gain to reflect their discriminative power. Examples of extracted patterns with the largest information gain are shown in Fig. 10, Fig. 11 and Fig. 12. These patterns are included in 10 patients with label R and 31 patients with label N.

[3] The support of a pattern is defined as the number of graphs with the pattern divided by the total number of graphs.

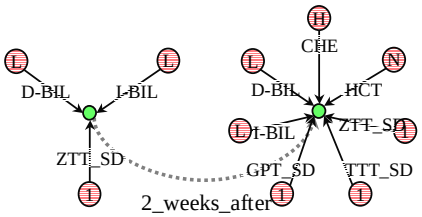


Fig. 13. pattern r1 (in 4 patients out of 10 with R)

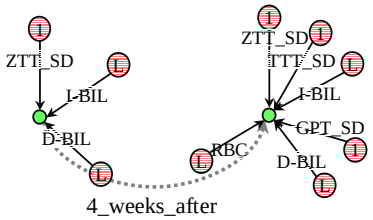


Fig. 14. pattern n9 (in 8 patients out of 31 with N)

The patients having these three patterns were further analyzed by B-GBI and the extracted patterns were also divided into 3 groups as before. The number and size of the extracted patterns from these graphs are summarized in the right-hand side of Table 5. Examples of extracted patterns with a large information gain are shown in Fig. 13 and Fig. 14. (actually the pattern in Fig. 13 had the largest information gain). Although it was difficult to extract patterns with time interval edges by DT-GBI, these patterns contain a time interval edge and still have sufficient discriminative power in the filtered data by the patterns in Fig. 10,11 and 12.

The same domain expert who evaluated the results by DT-GBI also evaluated the results by B-GBI. Unfortunately, many patterns appear in both class label R and N, as shown in the column "common" in Table 4, and most patterns were not judged as sufficiently characteristic. One encouraging comment is that the value of HGB might be some clue, because the results show that HBG is N (normal) in all the patterns with class label R but it is L (low) in patterns with class label N. Thus, investigating the effect of HGB is a future direction for the analysis of interferon therapy by B-GBI.

4 Conclusion

GBI extracts typical patterns from graph structured data by stepwise pair expansion (pairwise chunking) and its extensions (B-GBI and DT-GBI) have been applied for the analysis of hepatitis dataset provided by Chiba University Hospital. This paper reported yet another analysis of the dataset by B-GBI and DT-GBI with respect to the effectiveness to interferon therapy. Decision trees were constructed by DT-GBI for discriminating the patients for whom the hepatitis virus disappeared by interferon therapy from the patients for whom the virus continued to exist. In the second experiment, descriptive patterns are extracted by B-GBI and examples of extracted patterns are reported. Evaluation of the extracted patterns by a domain expert (medical doctor) suggested a next iteration of analysis. Immediate future work includes to 1) seek the appropriate duration of time correlation when converting to graph-structured data 2) continue the analysis of interferon therapy by discretizing the measurements in more reasonable way at preprocessing and representing the fluctuation of examination

values in a more approproate way, and 3) extract more time-correlated patterns by using some bias for the time interval edge.

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. Active mining from hepatitis data by beam-wise gbi. In *Working note of International Workshop on Active Mining (AM2002)*, pages 37–44, 2002.
3. 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.
4. M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi. A rule discovery support system for sequential medical data - in the case study of a chronic hepatitis dataset -. In *Working note of International Workshop on Active Mining (AM2002)*, pages 97–102, 2002.
5. J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.
6. J. R. Quinlan. *C4.5:Programs For Machine Learning*. Morgan Kaufmann Publishers, 1993.
7. 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.
8. 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.
9. G. Warodom, T. Yoshida, K. Ohara, H. Motoda, and T. Washio. Extracting diagnostic knowledge from hepatitis data by decision tree graph-based induction. Technical Report SIG-ICS-133, IPSJ, 2003.
10. 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.
11. K. Yoshida and H. Motoda. Clip : Concept learning from inference pattern. *Journal of Artificial Intelligence*, 75(1):63–92, 1995.

A Novel Hybrid Approach for Interestingness Analysis Of Classification Rules

Tolga Aydın[1], and Halil Altay Güvenir[1]

[1] Department of Computer Engineering, Bilkent University,
06800 Ankara, Turkey
{atolga, guvenir}@cs.bilkent.edu.tr

Abstract. Data mining is the efficient discovery of patterns in large databases, and classification rules are perhaps the most important type of patterns in data mining applications. However, the number of such classification rules is generally very big that selection of interesting ones among all discovered rules becomes an important task. In this paper, factors related to the interestingness of a rule are investigated and some new factors are proposed. Following this, an interactive rule interestingness-learning algorithm (IRIL) is developed to automatically label the classification rules either as "interesting" or "uninteresting" with limited user participation. In our study, VFP (Voting Feature Projections), a feature projection based incremental classification learning algorithm, is also developed in the framework of IRIL. The concept description learned by the VFP algorithm constitutes a novel hybrid approach for interestingness analysis of classification rules.

1 Introduction

Data mining is the efficient discovery of patterns, as opposed to data itself, in large databases [4]. Patterns in the data can be represented in many different forms, including classification rules, association rules, clusters, sequential patterns, time series, contingency tables, and others [5]. However, the number of discovered patterns is usually very big and the user analyzing the patterns is generally interested in a subset of them. Therefore, selection of interesting patterns is an important research topic.

In this paper, we concentrate on the patterns represented by the classification rules and develop an interactive rule interestingness-learning algorithm (IRIL) to automatically classify these rules as interesting or uninteresting, with limited user participation. In our study, VFP (Voting Feature Projections), a feature projection based incremental classification-learning algorithm, was also developed in the framework of IRIL. Being specific to our concerns, VFP takes the rule interestingness factors as features and is used to learn the rule interestingness concept and to classify the newly learned classification rules. The concept description learned by the VFP algorithm constitutes a novel hybrid approach for interestingness analysis of the classification rules.

Section 2 describes the interestingness issue of patterns. Section 3 is devoted to the knowledge representation used in our study. Section 4 and 5 are related to the

training and classifying phases of the VFP algorithm. IRIL is explained in the following section. Giving the experimental results in Section 7, paper is concluded.

2 Interestingness Issue of Patterns

The interestingness issue has been an important problem ever since the beginning of data mining research [1]. There are many factors contributing to the interestingness of a discovered pattern [1, 2, 3]. Some of them are coverage, confidence, completeness, action ability and unexpectedness. The first three factors are objective, action ability is subjective and unexpectedness is sometimes regarded as subjective [7, 8, 9] and sometimes as objective [10, 11]. Objective interestingness factors can be measured independently of the user and domain knowledge. However, subjective interestingness factors are not user and domain knowledge independent. The measurement of a subjective interestingness factor may vary among users analyzing a particular domain, may vary among different domains that a particular user is analyzing and may vary even for the same user analyzing the same domain at different times.

An objective interestingness measure is constructed by combining a proper subset of the objective interestingness factors in a suitable way. For example, objective interestingness factor x can be multiplied by the square of another objective interestingness factor y to obtain an objective interestingness measure of the form xy^2. It is also possible to use an objective interestingness factor x alone as an objective interestingness measure (e.g. *Confidence*). Discovered patterns having *Confidence* $\geq$ *threshold* are regarded as "interesting". Although the user determines the threshold, this is regarded as small user intervention and the interestingness measure is still assumed to be an objective one.

The existing subjective interestingness measures in the literature are constructed upon unexpectedness and action ability factors. Assuming the discovered pattern to be a set of rules induced from a domain, the user gives her knowledge about the domain in terms of fuzzy rules [9], general impressions [8] or rule templates [7]. The induced rules are then compared with user's existing domain knowledge to determine subjectively unexpected and/or actionable rules.

Both types of interestingness measures have some drawbacks. A particular objective interestingness measure is not sufficient by itself [9]. They are generally used as a filtering mechanism before applying a subjective measure. On the other hand, subjective measures are sometimes used without prior usage of an objective one. In the case of subjective interestingness measures, user may not be well in expressing her domain knowledge at the beginning of the interestingness analysis. It'd be better to automatically learn this knowledge based on her classification of some presented rules as "interesting" or "uninteresting". Another drawback of a subjective measure is that the induced rules are compared with the domain knowledge that addresses the unexpectedness and/or action ability issues. Interestingness is assumed to depend on these two issues. That is, if a rule is found to be unexpected, it is automatically regarded as an interesting rule. However, it would be better if we learned a concept description that dealt with the interestingness issue directly and if

we benefited from unexpectedness and action ability as two of the factors used to express the concept description. That is, interestingness of a pattern may depend on factors other than unexpectedness and action ability issues.

The idea of a concept description that is automatically determined and directly related with the interestingness issue motivated us to design IRIL algorithm. The concept description learned by the VFP algorithm, which was also developed in this framework, constitutes a novel hybrid approach for interestingness analysis of classification rules.

To ensure that the concept description is directly related to the rule interestingness issue, some existing and newly developed interestingness factors that have the capability to determine the interestingness of rules were used instead of the original attributes of the data set. Current implementation of IRIL does not incorporate unexpectedness and action ability factors, leading to no need for domain knowledge. Although the interestingness factors are all of type objective in the current version of IRIL, the thresholds of the objective factors are learned automatically rather than expressing them manually at the beginning. The values of these thresholds are based upon the user's classification results of some presented rules. So, although in the literature subjectivity is highly related to the domain knowledge, IRIL differs from them. IRIL's subjectivity is not related with the domain knowledge. IRIL makes use of objective factors (actually the current version makes use of only objective factors) but for each such a factor, it subjectively learns what ranges of factor values (what thresholds) lead to interesting or uninteresting rule classifications if only that factor is used for classification purposes. That is, IRIL presents a hybrid interestingness measure.

IRIL proceeds interactively. An input rule is labeled if the learned concept description can label the rule with high certainty. If the labeling or classification certainty factor is not of sufficient strength, user is asked to classify the rule manually. The user looks at the values of the interestingness factors and labels the rule accordingly. In IRIL, concept description is learned or updated incrementally by using the interestingness labels of the rules that are on demand given either as "interesting" or "uninteresting" by the user.

3 Knowledge Representation

The aim of the study presented in this paper is to label a set of classification rules as interesting or uninteresting. This labeling problem is modeled as a new classification problem and a *rule set* is produced for the given rules, which are previously learned by applying a rule induction algorithm on a data set. Each instance of the rule set is represented by a vector whose components are the interestingness label and the interestingness factor values having the potential to determine the interestingness of the corresponding rule.

The classification rules used in the study are probabilistic and have the following general structure:

If (A_1 *op* $value_1$) AND (A_2 *op* $value_2$) AND …AND (A_n *op* $value_n$) THEN
($Class_1$: $vote_1$, $Class_2$: $vote_2$,…,$Class_k$: $vote_k$)

In the above structure, A_i's are the features, $Class_i$'s are the classes and $op \in \{=, \neq, <, \leq, >, \geq\}$.

The instances corresponding to probabilistic classification rules have either "interesting" or "uninteresting" as the interestingness label, and the interestingness factors shown in Table 1. In this new classification problem, these factors are treated as determining features, and interestingness label is treated as the target feature of the rule set.

Table 1. Features of the rule set

Feature	**Short description and/or formula**
Major Class	$Class_i$ that takes the highest vote
Major Class Frequency	Ratio of the instances having $Class_i$ as the class label in the data set
Rule Size	Number of conditions in the antecedent part of the rule
Confidence with respect to Major Class	$\|Antecedent \ \& \ Class_i\| \ / \ \|Antecedent\|$
Coverage	$\|Antecedent\| \ / \ \|N\|$
Completeness with respect to Major Class	$\|Antecedent \ \& \ Class_i\| \ / \ \|Class_i\|$
Zero Voted Class Count	Number of classes given zero vote
Standard Deviation of Class Votes	Standard deviation of the votes of the classes
Major Class Vote	Maximum vote value distributed
Minor Class Vote	Minimum vote value distributed
Decisive	True if Std.Dev.of Class.Votes > s_{min}

Each feature carries information about a specific property of the corresponding rule. For example, if we let $Class_i$ to take the highest vote, it then becomes the *Major Class* of that classification rule. If we shorten the representation of any rule as "If *Antecedent* THEN $Class_i$" and assume the data set to consist of N instances, we can define *Confidence*, *Coverage* and *Completeness* as in Table 1. Furthermore, a rule is decisive if the standard deviation of the votes is greater than s_{min}, whose definition is given in the following equation:

$$s_{min} = \frac{1}{(Class\ Count - 1)\sqrt{Class\ Count}} \quad (1)$$

If a rule distributes its vote "1" evenly among all classes, then the standard deviation of the votes becomes zero and the rule becomes extremely indecisive. This is the worst vote distribution that can happen. The next worst vote distribution is obtained if exactly one class takes a zero vote, and the whole vote is distributed evenly among the remaining classes. The standard deviation of the votes that will occur in such a situation is called s_{min}.

4 Training in the VFP Algorithm

VFP (Voting Feature Projections) is a feature projection based classification-learning algorithm developed in our study. It is used to learn the rule interestingness concept and to classify the unlabeled rules in the context of modeling rule interestingness problem as a new classification problem.

The training phase of VFP, given in Figure 3, is achieved incrementally. On a nominal feature, concept description is shown as the set of points along with the numbers of instances of each class falling into those points. On the other hand, on a numeric feature, concept description is shown as the normal (gaussian) probability density functions for each possible class. Training can better be explained by looking at the sample data set in Figure 1, and the associated learned concept description in Figure 2.

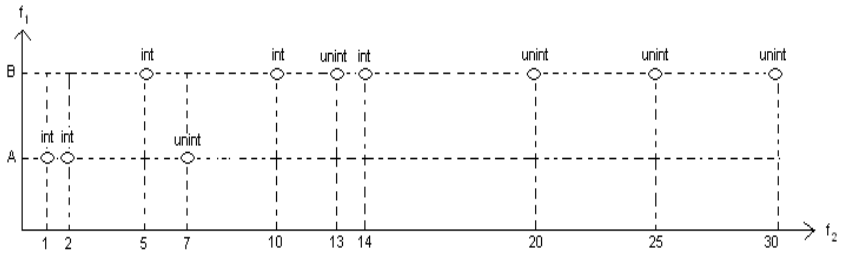


Fig. 1. Sample data set

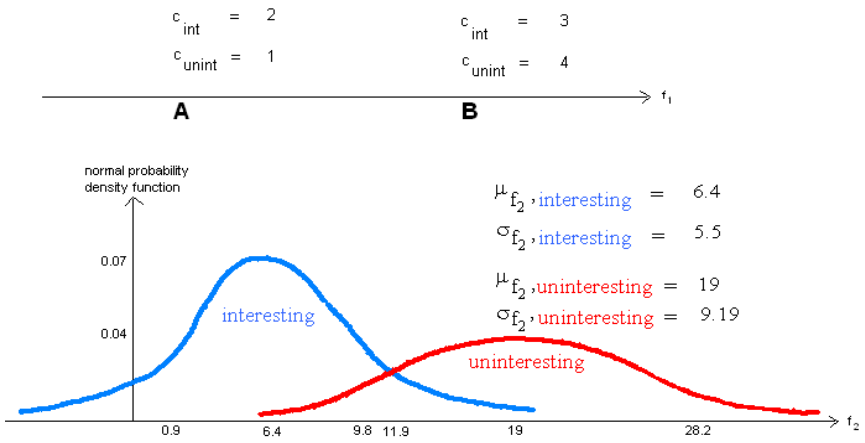


Fig. 2. Concept description learned for the sample data set

The example data set consists of 10 training instances, having nominal f_1 and numeric f_2 features. f_1 takes two values: "A" and "B", whereas f_2 takes some integer values. There are two possible classes: "interesting" and "uninteresting". f_2 is assumed to have gaussian probability density functions for both classes.

VFP$_{train}$ (*t*) /* *t*: newly added training instance */
begin
let *c* be the class of *t*
let *others* be the remaining classes other than *c*
if *training set* = {*t*}
for each class *s*
class_count[*s*] = 0

class_count[*c*]++

for each feature *f*
if *f* is nominal
p = *find_point*(*f*,t_f)
if such a *p* exists
/* if t_f value exists in the *training set* */
point_class_count [*f*,*p*,*c*] ++
else /* add new point for *f* */
add a new *p'* point
point_class_count [*f*,*p'*,*c*] = 1
point_class_count [*f*,*p'*,*others*] = 0
else if *f* is numeric
if *training set* = {*t*}
$\mu_{f,c} = t_f$, $\mu_{f,others} = 0$
$\mu^2_{f,c} = t_f^2$, $\mu^2_{f,others} = 0$
$\sigma_{f,c}$ = *Undefined*
norm_density_func.$_{f,c}$= *Undefined*
else
n = *class_count*[*c*]
$\mu_{f,c} = (\mu_{f,c} * (n-1) + t_f) / n$ /*update*/
$\mu^2_{f,c} = (\mu^2_{f,c} * (n-1) + t_f^2) / n$ /*update*/
$\sigma_{f,c} = \sqrt{\frac{n}{n-1}(\mu^2_{f,c} - (\mu_{f,c})^2)}$
norm_density_func.$_{f,c}$= $\frac{1}{\sigma_{f,c}\sqrt{2\pi}} e^{-\frac{(x-\mu_{f,c})^2}{2\sigma_{f,c}^2}}$

return { For numeric features: *norm_density_func.*$_{f,c}$(∀*f*, *c*) For nominal features: *point_class_count*[*f*, *p*, *c*] (∀*f*, *p*, *c*)
end.

Fig. 3. Incremental train in VFP

In Figure 3 for a nominal feature *f*, *find_point* (*f*, t_f) procedure tries to find (t_f), the new training instance's value at feature *f*, in the *f* projection. If t_f is found at a point *p*,

then *point_class_count* [*f*, *p*, *c*] is incremented, assuming that the training instance is of class *c*. If t_f is not found, then a new point *p'* is constructed and *point_class_count* [*f*, *p'*, *class*] is initialized to 1 for *class* = *c*, and to 0 for all other classes. In our study, features used in VFP are the interestingness factor values computed for the classification rules, and we have only "interesting" and "uninteresting" as the classes.

For a numeric feature *f*, if a new training instance *t* of class *c* is examined, we let the previous training instances of class *c* to construct a set *P* and let $\mu_{f,c}$ and $\sigma_{f,c}$ to be the mean and the standard deviation of the *f* feature projection values of the instances in *P*, respectively. The previous training instances' values on *f* need not be stored anywhere, so $\mu_{f,c}$ and $\sigma_{f,c}$ are updated incrementally. Updating $\sigma_{f,c}$ incrementally requires $\mu^2_{f,c}$ to be updated incrementally, as well.

5 Classification in the VFP Algorithm

Classification phase of VFP is shown in Figure 4. The query instance is projected on all features and each feature gives votes for each class. If a feature is not ready to classification process, it gives zero, otherwise gives normalized votes. Normalization ensures that each feature has the same weight in classifying the query instances. However, if a feature is not ready, it does not involve in the classification process, therefore need not give normalized votes. For a feature to be ready for the classification process, it should have at least two different values for each class.

The classification starts by giving zero votes to classes on each feature projection. The features that are not ready do not participate in the classification process. The participating features are handled accordingly. For a nominal feature *f*, *find_point* (*f*, q_f) procedure is used to search whether q_f exists in the *f* projection. If q_f is found at a point *p*, feature *f* gives votes for each class as shown in the below equation, and then these votes are normalized to ensure equal voting power among features.

$$feature\_vote\ [f, c] = \frac{point\_class\_count\ [\ f, p, c\ ]}{class\_count\ [\ c\ ]} \tag{2}$$

In the above equation, we divide the number of class *c* instances on point *p* of feature projection *f* by the total number of class *c* instances to find the class conditional probability of falling into the *p* point. For a linear feature *f*, each class gets the vote given in equation 3. Normal probability density function values are used as the vote values. These votes are then normalized, too.

$$feature\_vote\ [f, c] = \lim_{\Delta x \to 0} \int_{q_f}^{q_f + \Delta x} \frac{1}{\sigma_{f,c}\sqrt{2\pi}} e^{-\frac{(q_f - \mu_{f,c})^2}{2\sigma_{f,c}^2}} dx \tag{3}$$

VFP$_{query}$(q) /* q: query instance*/
begin

for each feature f
for each class c
$feature\_vote[f,c] = 0$

if $feature\_ready\_for\_query\_process(f)$

if f is nominal
$p = find\_point(f, q_f)$
if such a p exists
/* if q_f value exists in the *training set* */
for each class c

$$feature\_vote\ [f,c] = \frac{point\_class\_count\ [f, p, c]}{class\_count\ [c]}$$

$normalize\_feature\_votes\ (f)$
/* such that $\sum_c feature\_vote[f,c] = 1$ */

else if f is numeric
for each class c

$$g = \frac{1}{\sigma_{f,c}\sqrt{2\pi}} e^{-\frac{(q_f - \mu_{f,c})^2}{2\sigma_{f,c}^2}}$$

$$feature\_vote\ [f,c] = \lim_{\Delta x \to 0} \int_{q_f}^{q_f + \Delta x} g\,dx$$

$normalize\_feature\_votes\ (f)$

for each class c

$$final\_vote\ [c] = \sum_{f=1}^{\#Features} feature\_vote\ [f, c]$$

for each class c
if $\min_{i=1}^{\#Classes} final\_vote[i] < final\_vote\ [c] = \max_{i=1}^{\#Classes} final\_vote[i]$
classify q as "c" with a certainty factor C_f
return C_f
else
$C_f = -1$
return C_f
end.

Fig. 4. Classification in VFP

Final vote for any class c is the sum of all votes given by the features. If there exists a class c that gets the highest vote and there also exists at least one other class that gets a lower vote than c, then class c is predicted to be the class of the query instance. The certainty factor of the classification (C_f) is computed as follows:

$$C_f = \frac{final\ vote[c]}{\sum_{i=1}^{\#Classes} final\ vote[i]} \quad (4)$$

If no prediction is made, certainty factor is taken as "–1" to indicate this situation.

6 IRIL Algorithm

IRIL algorithm, shown in Figure 6, needs two input parameters: R (The set of classification rules) and $MinC_t$ (Minimum Certainty Threshold). It tries to classify the rules in R. If $C_f \geq MinC_t$ for a query rule r, this rule is inserted into the successfully classified rules set (R_s). Otherwise, two situations are possible: either the concept description is not able to classify r ($C_f = -1$), or the concept description's classification (prediction of r's interestingness label) is not of sufficient strength. If $C_f < MinC_t$, rule r is presented, along with its computed eleven interestingness factor values such as *Coverage*, *Rule Size*, *Decisive* etc., to the user for classification. This rule or actually the instance holding the interestingness factor values and the recently determined interestingness label of this rule is then inserted into the training rule set R_t and the concept description is reconstructed incrementally.

All the rules in R are labeled either automatically by the classification algorithm, or manually by the user. User participation leads rule interestingness learning process to be an interactive one. When the number of instances in the training rule set increases, the concept description learned tends to be more powerful and reliable. When the labeling of the rules ends, the rules in R_s are relabeled by the latest version of the concept description. Because there may exist some rule r that was classified as "interesting" with a sufficient certainty factor by a weak version of the concept description, but now labeled as "interesting" or "uninteresting" with an insufficient certainty factor by the latest and the most reliable version of the concept description. Such rules called as R_{exc} are excluded from R_s and inserted into R. Therefore, we have $R = R_{exc}$ and $R_s = R_s - R_{exc}$. The cycle is repeated until R gets empty and IRIL concludes by presenting the labeled rules in R_s. It is guaranteed that the number of cycles is not infinite and R eventually gets empty. Proof is as follows:

At the end of any cycle, if $R_{exc} = \{\}$ then we are done. If $R_{exc} \neq \{\}$, then at least one rule will be classified by the user and then added into the R_t since the current version of the concept description could not classify the rules in R_{exc} with sufficient certainty. Unless $R_{exc} = \{\}$, at the end of each cycle R_t will expand by at least one element. Therefore, the cycle will be repeated at most $|R|$ times.

```
IRIL ( R, MinC_t)
begin
   R_t ←∅,    R_s ← ∅
   repeat
      for each rule r ∈R
         C_f ←VFP_query (r)
            if  C_f < MinC_t
               ask the user to classify r
               set C_f of  this classification to 1
               insert r into R_t
               VFP_train (r)
            else
               add r into R_s
            remove r from R

      for each rule r ∈ R_s
         C_f ←VFP_query (r)
            if  C_f  < MinC_t
               remove r from R_s
               add r into R
   until R is empty
   output rules in R_s
end.
```

Fig. 6. IRIL algorithm

7 Experimental Results

IRIL algorithm was tested to classify 184 classification rules induced from a financial distress domain using a benefit maximizing feature projection based rule learner proposed in [6]. The data set of the financial distress domain is a comprehensive set consisting of 25632 data instances and 164 determining features (159 numeric, 5 nominal). There are two classes: "Profit" and "Loss". The data set includes some financial information about 3000 companies collected during 10 years and the class feature states whether the company made a profit or loss on a particular year. Domain expert previously labeled all the 184 induced rules to make accuracy measurement possible. The expert labeled 50 rules (27.17%) as "interesting" and 134 rules (72.83%) as "uninteresting".

The results for $MinC_t = 60\%$ shows that the user classifies 54 rules with 100% certainty, and 130 rules are classified automatically with $C_f > MinC_t$. User participation is 29% in the classification process. While labeling the rules, user participation increases in proportion to the $MinC_t$ as expected. In the classification process, it is always desired that rules are generally classified automatically, and user participation is low.

Table 1. Results for IRIL

	$MinC_t$ 60%	$MinC_t$ 65%	$MinC_t$ 75%
Number of rules	184	184	184
Number of rules classified automatically with high certainty	130	108	90
Number of rules classified by user	54	76	94
User participation	29%	41%	51%
Overall Accuracy	80%	87.04%	90%
Accuracy among interesting rules	94.87%	95.45%	95.12%
Accuracy among uninteresting rules	73.63%	81.25%	85.71%

If we look at the accuracy results for $MinC_t$ = 60%, they are measured as 80%, 94.87% and 73.63% for the rules in R_s (overall accuracy), for the actually interesting rules in R_s (accuracy among interesting rules) and for the actually uninteresting rules in R_s (accuracy among uninteresting rules), respectively. It is important to keep the three accuracy values close to each other. For instance, if the above three accuracy values were 65%, 20% and 75%, respectively, we would easily claim that IRIL made biased classifications in favor of "uninteresting" class. Because, accuracy among uninteresting rules is too high, whereas accuracy among interesting rules is too low. Furthermore, user herself labels 134 of the rules (72.83%) as uninteresting, so we could label all the rules as "uninteresting" without using IRIL that would result in an accuracy value of 72.83%, which is very close to the overall accuracy of 65%. Fortunately, IRIL makes unbiased classifications since the three accuracy values are balanced. The accuracy values generally increase in proportion to the $MinC_t$. Because higher the $MinC_t$, higher the user participation is. And higher user participation leads to learn a more powerful and predictive concept description.

8 Conclusion

IRIL feature projection based, interactive rule interestingness learning algorithm was developed and gave promising experimental results. The concept description learned by the VFP algorithm, also developed in the framework of IRIL, constitutes a novel hybrid approach for interestingness analysis of classification rules. The concept description differs among the users analyzing the same domain. That is, IRIL determines the important rule interestingness factors for a given domain subjectively, by making use of objective factors.

As future work, other classification learning algorithms, which need not be feature projection based, can be used in the framework of IRIL. On the other hand, other objective and subjective interestingness factors, especially unexpectedness, may be used as the features of the rule sets.

References

1. Frawely, W.J., Piatetsky-Shapiro, G., and Matheus, C.J., "Knowledge discovery in databases: an overview" *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991, 1-27
2. Major, J.A., and Mangano, J.J., "Selecting among rules induced from a hurricane database" *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, 1993, 30-31
3. Piatetsky-Shapiro, G., and Matheus, C.J., "The interestingness of deviations" *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, 1994, 25-36
4. Fayyad, U., Shapiro, G., and Smyth, P., "From data mining to knowledge discovery in databases" *AI Magazine 17(3)*, 1996, 37-54
5. Hilderman, R.J., and Hamilton, H.J., "Knowledge discovery and interestingness measures: a survey" *Technical Report*, Department of Computer Science, University of Regina, 1999
6. Güvenir, H.A., "Benefit Maximization in Classification on Feature Projections" *Proceedings of the 3rd IASTED International Conference on Artificial Intelligence and Applications* (AIA'03), 2003, 424-429.
7. Klemettinen, M., Mannila, H., Ronkainen, P., Toivonen, H., and Verkamo, A.I., "Finding interesting rules from large sets of discovered association rules" *Proceedings of the 3rd Int. Conf. on Information and Knowledge Management*, 1994, 401-407.
8. Liu, B., Hsu, W., and Chen, S., "Using general impressions to analyze discovered classification rules" *Proceedings of the 3rd Int. Conf. on KDD*, 1997, 31-36.
9. Liu, B., and Hsu, W., "Post-analysis of learned rules", *AAAI*, 1996, 828-834.
10. Hussain, F., Liu, H., Suzuki, E., and Lu, H., "Exception rule mining with a relative interestingness measure" *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2000, 86-97.
11. Dong, G., and Li, J., "Interestingness of discovered association rules in terms of neighborhood-based unexpectedness" *Proceedings of the 2nd Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 1998, 72-86.

Preliminary Evaluations of Discovered Rule Filtering Methods

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

[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/
[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. Data mining systems semi-automatically discover knowledge by mining a large volume of data, but discovered knowledge is not always novel to users. We discuss a discovered rule filtering method to filter rules discovered by a data mining system into ones that are novel to the user by using information retrieval results from the Internet. We have two methods; the micro view and the macro view methods, to achieve discovered rule filtering. In the micro view method, we extract keywords from a discovered rule and rank the rule referring to the number of hits when the keywords are submitted to the MEDLINE database. In the macro view method, we first retrieve documents by submitting every pair of the extracted keywords and then form keyword clusters according to the results. We evaluated the methods by sending out a questionnaire to medical students. The evaluation indicates that the macro view method is promising as a discovered rule filtering method.

1 Introduction

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 a discovered rule filtering method [3,4,5] that filters a large number of rules obtained by a data mining system to be a small number of novel rules by using an information retrieval technique from the Internet.

Data mining is an automated method to discover useful knowledge by analyzing a large volume of data mechanically [6]. Generally speaking, conventional data mining methods try to discover significant patterns in the statistical sense from a large volume of raw data contained in a given database, but if a system pays attention to only statistically significant features, it may produce a large number of rules that have been known to users. To cope with this problem, we are developing a discovered rule fil-

tering method that filters a large number of rules discovered by a data mining system to be a small number of rules that is novel to the user. To judge whether a rule is novel or not, we utilize an information source on the Internet and judge the novelty of rule according to the number of retrieved documents that relate to the rule.

In this paper, we show two discovered rule filtering methods called the micro view method and the macro view method and evaluate the methods by conducting a questionnairing to medical students. We show the concept and the process of discovered rule filtering with an application example to clinical data mining in Section 2. We then show two discovered rule filtering methods; the micro view and the macro view methods in Section 3 and evaluate them in Section 4 by a questionnaire method. Finally we conclude this paper with our future work in Section 5.

2 Discovered Rule Filtering

The target of our active mining project is a clinical examination database of hepatitis patients, which is offered by the Medical School of Chiba University, on which 10 research groups cooperatively work as a common data source [7]. Several groups have already discovered some sets of rules. For example, Yamaguchi et al. in Shizuoka University analyzed sequential trends between GPT (Glutamic Pyruvic Transaminase), which represents the progress of hepatitis, and other blood test data, and has discovered a number of rules, as one of them is shown in Fig. 1 [8].

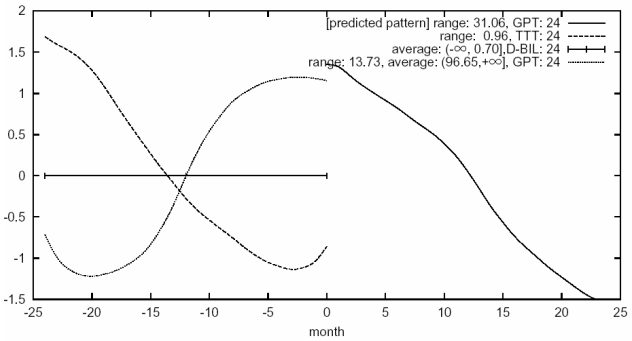


Fig. 1**.** An example of discovered rule [8].

This rule shows a relation among GPT, TTT (Thymol Turbidity Test), and D-BIL (Direct Bilirubin) and means "If, for the past 24 months, D-BIL stays unchanged, TTT decreases, and GPT increases, then GPT will decrease for the following 24 months." A data mining system can semi-automatically discover a large number of rules by analyzing a set of given data, but the discovered rules may include ones that have been known to users. Showing all the discovered rules to a user just results in putting a burden on her. We need to develop a method to filter the discovered rules into a small set of rules that is novel to her. To judge whether a rule is novel or not, we utilize an information source on the Internet and judge it according to the number of retrieved documents relate to the discovered rule.

When a set of discovered rules are given from a data mining system, the discovered rule filtering system first retrieves information related to the rules from the Internet and then filters the rules based on the result of information retrieval. In our project, we aim at discovering knowledge from a hepatitis clinical database, and it is not easy to gather information related to hepatitis from the Internet by using a naïve search engine because the Internet information sources generally contain a huge amount of various and noisy information. We instead use the MEDLINE (MEDlars on LINE) database as the source of retrieving information, which is the largest bibliographical database in the medical and biological domain. 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 the Pubmed, we can retrieve MEDLINE documents by submitting a set of keywords just like a normal search engine. In addition, we can retrieve documents according to the year of publication and/or the category of documents. These functions are not provided by normal search engines.

The discovered rule filtering process takes the following steps.

Step 1: Extracting keywords from a discovered rule

We need to 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 extracted from a discovered rule.** These keywords represent attributes in a discovered rule. For example, keywords that can be extracted directly from a discovered rule shown in Fig. 1 are GPT, TTT, and D-BIL because they are explicitly appeared in the rule. If any abbreviation is not acceptable for the Pubmed, it is translated into a normal name. For example, TTT and GPT are translated into "thymol turbidity test" and "glutamic pyruvic transaminase" respectively.
- **Keywords related to the mining domain.** These keywords represent the purpose or the domain of the data mining task. With keywords extracted form a rule, they should be submitted to the Pubmed as the common keywords to improve the quality of retrieved documents. For our hepatitis data mining, "hepatitis" is a domain keyword. The domain keywords are implicit keywords to be submitted to the Pubmed, and we do not explicitly indicate the keywords in the following discussion.

The rule shown in Fig.1 includes information not only about relations among attributes but also about how the attributes change, but it is difficult to represent the latter information in a sequence of keywords. This problem is left as our future work.

Step 2: Filtering Discovered Rules

We filter discovered rules by using the result of MEDLINE document retrieval. We have two methods called the micro view method and the macro view method to filter discovered rules. The details of the methods are discussed in the following section.

3. Two Methods for 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 methods; the micro view method and the macro view method, to filter discovered rules [5].

3.1 Micro View Method

The micro view method retrieves documents directly related to a discovered rule. It utilizes the result of retrieving documents not only to filter discovered rules, but also to show the documents to the user. By showing a rule and documents related to the rule together, the user may expand her insights on the rule and the data mining task [3]. Filtering rules by the micro view method is quite simple and is based on the following hypotheses.

[Hypotheses] (Micro View Method)
1. The number of documents related to a known rule is large.
2. The number of documents related to an unknown rule is small.
3. The number of documents related to a garbage rule is zero

We hypothesize that research activities on a known rule have been done a lot and that a large number of papers related to the rule have been published. On the other hand, those on an unknown rule have been done a little, and the number of papers related to the rule must be small. Nobody has interest in a garbage or nonsense rule, and the number of papers related to the rule must be zero.

As a strategy of discovered rule filtering, we filter out the garbage rules at the first stage. In the above hypotheses, the border between known rules and unknown one is vague, so we rank rules as the number of the related documents ascends.

However, the micro view method depends much on the performance of document retrieval, and it is actually difficult to retrieve appropriate documents rightly related a rule because of the low performance of keyword-based document retrieval technique. Generally speaking, when a rule is simple with a small number of attributes, the Pubmed system returns a large number of unrelated noisy documents. When a rule is complicated with a large number of attributes, it returns only few documents.

3.2 Macro View Method

The macro view method tries to roughly observe the research trend of discovered rules. Given a rule, it submits every pair of keywords extracted from the rule, not the whole sequence of the keywords, to the Pubmed system, and integrates the results in the form of keyword co-occurrence graph to judge the novelty of the rule.

Fig. 2, 3, and 4 show keyword co-occurrence graphs. In each graph, a node represents a keyword and the length of edge represents the inverse of the frequency of co-occurrences of keywords connected by the edge. The score attached to the edge repre-

sents the frequency of co-occurence. Hence, the more documents related to a pair keywords are retrieved from the Pubmed, the closer the keywords are located in the graph.

For example, Fig. 2 shows that the relation between any pair from ALB, GPT, and T-CHO is strong. Fig. 3 shows that the relation between T-CHO and GPT is strong, but that between chyle and either of T-CHO and GPT is rather weak. Fig. 4 shows that the relations among GPT, female, and G-GTP are strong, but the relation between hemolysis and G-GTP and those between "blood group a" and the other keywords are weak.

We then form clusters of keywords by using the Hierarchical Clustering Scheme [9]. As a strategy to form clusters, we adopt the complete linkage clustering method (CLINK). In the method, the distance between clusters A and B is defined as the longest among the distances of every pair of a keyword in cluster A and a keyword in cluster B. The method initially forms a cluster for each keyword. It then repeats to merge clusters within a threshold length into one or more clusters.

We can regard keywords in a cluster are strongly related and research activities concerning the keywords have been done much, so we have a hypothesis to filter rules in the macro view method as follows.

[Hypothesis] (Macro View Method)

1. The number of clusters concerning a known rule is 1.
2. The number of clusters concerning an unknown rule is 2.
3. The number of clusters concerning a garbage rule is more than 3.

A rule with only one cluster is regarded as a known rule because a large number of papers concerning every pair of keywords in the rule have been published. A rule with two clusters is regarded as an unknown rule. This is because research activities concerning keywords in each cluster have been done much, but those crossing the clusters have not been done. A rule with more than two clusters is regarded as a garbage rule. Such a rule is too complex to understand because the keywords are partitioned into many clusters and the rule consists of many unknown factors.

For example, if we set the threshold of CLINK to be 1 (the frequency of co-occurrences is 1), the rule in Fig. 2 is regarded as a known rule because all the keywords are merged into a single cluster. Keywords in Fig. 3 are merged into two clusters; one cluster consists of GPT and T-CHO and another consists of chyle only. Hence, the rule is judged to be unknown. Keyowrds in Fig. 4 are merged into 3 clusters as GPT, G-GTP, and female form a cluster and each of hemolysis and "blood group a" forms an individual cluster.

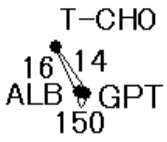


Fig. 2. The keyword co-occurrence graph of rule including GPT, ABL, and T-CHO.

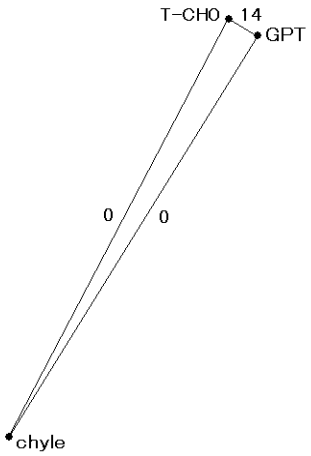


Fig. 3. The keyword co-occurrence graph of rule including GPT, T-CHO, and chyle.

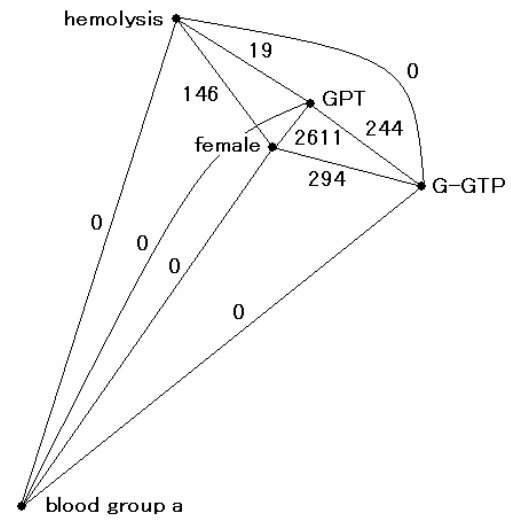


Fig. 4. The keyword co-occurrence graph of rule including GPT, G-GTP, hemolysis, female and "blood group a".

4 Evaluation of Discovered Rule Filtering Methods

4.1 Questionnaire

We performed an evaluation of discovered rule filtering methods by a questionnaire method. In the evaluation, we verified the hypotheses that are the basis of the

micro view and the macro view methods. We first made a questionnaire with two questions shown in Fig. 5. 20 items in Q1 are made from rules discovered by the data mining group in Shizuoka University [8] by extracting keywords from the rules. The reason why we do not show the discovered rules to the subjects is because we would like to evaluate the performance of our rule filtering method that submits only attribute keywords extracted from the rules. If we show the discovered rules directly to the subjects, the subjects judge them considering more than just the relation among attribute keywords, ex. how the attributes change, and that makes the evaluation not proper.

20 items in Q2 are randomly chosen from keywords in the discovered rules. The purpose of Q2 is to show that the number of retrieved documents correlates with the evaluation of medical students when we limit to the number of submitted keywords to be 2, as discussed in Section 4.2.

We sent out the questionnaire to 47 medical students in Osaka City University. The students were just before the state examination to be a medical doctor, so we suppose they are knowledgeable about the medical knowledge in text books.

Q1: How do you guess the result when you submit the following keywords to the Pubmed system? Choose one among A, B, and C.

A (Known): Documents about a fact that I know will be retrieved.
B (Unknown): Documents about a fact that I do not know will be retrieved.
C (Garbage): No document will be retrieved.

(1) [A B C] ALT and TTT
(2) [A B C] TTT, Direct-Bilirubin, and ALT
(3) [A B C] ALT, Total-Cholesterol, and Hepatitis C
(4) ….

Q2: Choose one among four choices about the relation between the following items.

A: The items have a strong relation with each other.
B: The items have a medium relation with each other.
C: The items have a weak relation with each other.
D: The items have no relation with each other.

(1)[A B C D] ALT, Total-Bilirubin
(2)[A B C D] ALT, Total-Cholesterol
(3)[A B C D] ALB, Total-Cholesterol
(4) …

Fig. 5. Questionnaire sent out to medical students.

4.2 Evaluation of the micro view method

We here evaluate the micro view method by analyzing the relation between the number of documents hit by the keywords and the ratio of choices in Q1 made by medical students. Fig. 6 shows the result. We plot the relation between the ratio of choice and the number of retrieved documents for each item in Q1. We also add regression lines to show the relation more clearly and assessed the significance by using the t-test method at the risk level of 5%, but we cannot find any significant relation.

The reason why the micro view method, in which all the keywords extracted from a rule are directly submitted to the Pubmed, does not work well is because the number of hits seems to depend much on the number of keywords submitted to the Pubmed. Generally speaking, the more the number of submitted keywords is, the less the number of retrieved documents is.

However, if we limit the number of submitted keywords to be 2, the method shows a better performance. Fig. 7 shows the relation between the number of documents and the evaluation of medical students obtained through Q2 in the questionnaire. The number of keywords used in Q2 is fixed to be 2. We put the score 3, 2, 1, and 0 to the choice A, B, C, and D respectively. The averaged score and the number of documents have a significant correlation since the correlation coefficient is 0.54. Hence, if we limit the number of submitted keywords to be 2, the result reflects the evaluation of medical students.

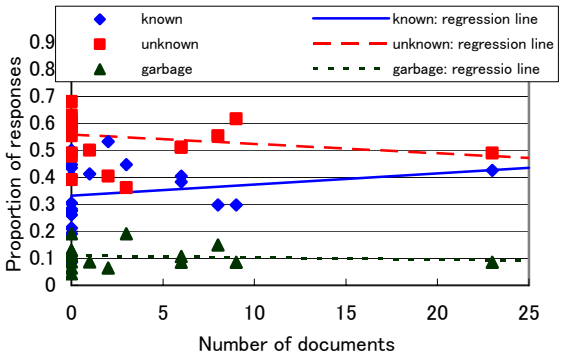


Fig. 6. The relation between the ratio of choice and the number of documents.

4.3 Evaluation of the macro view method

We verify the hypothesis of the macro view method by using the result of Q1 of the questionnaire. We show the relation between the number of clusters and the average ratio of choice in Fig. 8. The threshold of CLINK is 1. At the risk level of 5%, the graph shows two significant relations.

- As the number of clusters increases, the average ratio of “unknown” increases.
- As the number of clusters increases, the average ratio of “known” decreases.

The result does not show any significant relation about “garbage” choice because the number of students who chose “garbage” is relatively small to the other choices and does not depend on the number of clusters. We suppose the medical students hesitate to judge that a rule is just garbage.

The hypotheses of the macro view method are partly supported by this evaluation. The maximum number of clusters in this examination is 3. We still need to examine how medical students or experts judge rules with more than 4 clusters.

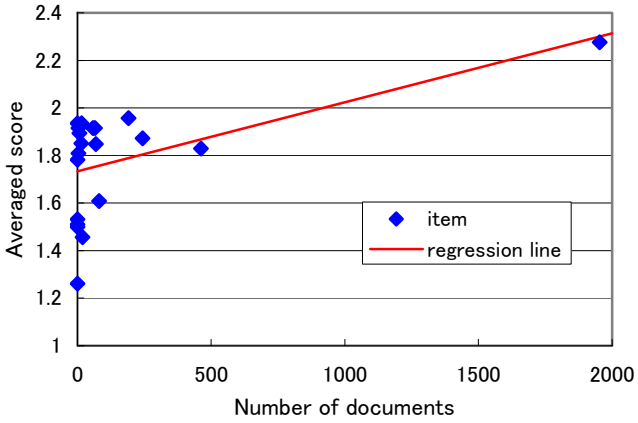


Fig. 7. The relation between the number of documents and the evaluation of medical students when the number of submitted keywords is 2.

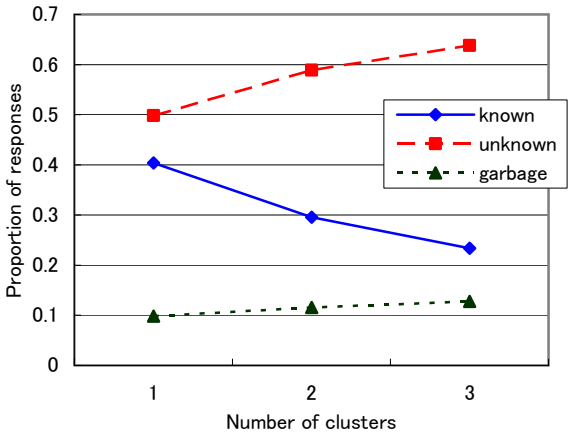


Fig. 8. The relation between the number of clusters and the evaluation of medical students.

5. Summary

We discussed two discovered rule filtering methods, the micro view and the macro view methods, which filters rules discovered by a data mining system into novel ones by using the information retrieval technique from the Internet. We evaluated the methods by using the questionnaire method. The result supports that the output of the macro view method reflects the evaluation of medical students.

Our future work is summarized as follows.

- We need to improve the performance of the information retrieval technique which is based on a naïve keyword search. We plan to improve the performance by applying natural language processing techniques [10].
- We apply the discovered rule filtering methods to a practical application domain such as hepatitis data mining, and evaluate its feasibility.

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.
5. Y. Kitamura, A. Iida, K. Park, and S. Tatsumi, Micro View and Macro View Approaches to Discovered Rule Filtering. Proceedings of 2nd International Workshop on Active Mining, pp.14-21, 2003.
6. U. M. Fayyad, G. P. Shapiro, P. Smyth, and R. Uthurusamy (Eds.), Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.
7. H. Yokoi, S. Hirano, K. Takabayashi, S. Tsumoto, Y. Satomura, Active Mining in Medicine: A Chronic Hepatitis Case – Towards Knowledge Discovery in Hospital Information Systems -, Journal of the Japanese Society for Artificial Intelligence, Vol.17, No.5, pp.622-628, 2002. (in Japanese)
8. M. Ohsaki, Y. Sato, H. Yokoi, and T. Yamaguchi, A Rule Discovery Support System for Sequential Medical Data – In the Case Study of a Chronic Hepatitis Dataset -, Proceedings of International Workshop on Active Mining, pp. 97-102, 2002.
9. S. C. Johnson, Hierarchical Clustering Schemes, Psychometrika, Vol.32, pp.241-254, 1967.
10. T. Yamasaki, M. Shimbo, and Y. Matsumoto: A MEDLINE document search system using section information, JSAI, SIG-KBS-A301-05, 2003.

Proposal of Relevance Feedback based on Interactive Keyword Map

Yasufumi Takama and Tomoki Kajinami

Tokyo Metropolitan Institute of Technology
6-6 Asahigaoka, Hino Tokyo 191-0065, JAPAN
ytakama@cc.tmit.ac.jp

Abstract. The relevance feedback based on a keyword map is proposed so that a Web interface can be more interactive. There exists vast amount of information in the Web, from which users usually gather information without definite information needs. Therefore, it is difficult for users to organize and understand what they have gathered from the Web. From this viewpoint, we have proposed the concept of RBA-based interaction, in which analysis operation aims to assist users in understanding the context of their web interaction. However, the interface currently developed focuses on the information flow from the interface to users. As the first step for realizing relevance feedback (RF) based on interactive keyword map, this paper proposes the algorithm for extracting the pair of keywords that reflects a user's interest from the keyword map. Experimental results are given for showing how the algorithm works on the keyword map that is modified by the user, and for discussing the difference between the RF based on keyword map and conventional RF methods.

1 Introduction

A Web interaction is defined as users' activities for viewing and collecting web pages with using search engines and Web browsers. There exists vast amount of information in the Web, from which users usually gather information without definite information needs. Therefore, it is difficult for users to organize and understand what they have gathered from the Web. We have proposed the concept of RBA-based interaction, in which analysis operation aims to assist users in understanding the context of their web interaction. The Web interface that supports RBA-based interaction employs both keyword map visualization and document clustering, which respectively present users the topic distribution and document clusters within gathered document set[13]. However, the interface currently focuses on the information flow from the interface to users. In this paper, the relevance feedback based on a keyword map is proposed so that the interface can be more interactive. As the first step for realizing keyword map-based RF, this paper proposes the algorithm for extracting the pair of keywords that reflects a user's interest from the keyword map. Experimental results are given for showing how the algorithm works on the keyword map that is modified by the

user, and for discussing the difference between the RF based on keyword map and conventional RF methods.

2 Related Work

2.1 Concept of Retrieval, Browsing, Analysis (RBA)-based Interaction

One of the essential properties of our activities in the Web is that we do not always have the predetermined target topics while surfing on the Web. Therefore, not only submitting relevant queries, but also evaluating the relevance of web pages is difficult for us. Through the interaction with the Web, We find the topics of interest, and acquire the background knowledge about the topics, based on which the relevance of pages is evaluated.

Considering the commercial success of web search engines, it is rational that we assume the following steps for locating and gathering information in the Web:

Retrieval Obtains a set of pages by submitting tentative query to a search engine.

Browsing Starting from individual documents in the retrieved results, browses their neighboring pages and collect (save) the relevant ones.

We call the interaction based on these two steps RB-based interaction. It should be noted that a user cannot always evaluate the relevance of pages correctly, and the evaluation criteria frequently changes while interacting with the Web. In other words, the context that affects the evaluation criteria is composed of the pages that have been gathered so far. Therefore, we claim that the "analysis" step should be combined with RB-based interaction. We call the interaction based on these three steps RBA-based interaction. Although Gershon[4] has already denoted the importance of the analysis step, in which the properties within a single page is analyzed. Our focus is on analyzing the set of gathered documents.

From this viewpoint, conventional information visualization systems[1, 2, 5–7, 15–17] contribute for supporting RBA-based interaction to some extent. However, they put the analysis step inside retrieval and browsing steps. That is, the visualized space by browsing support systems is mainly used for users to browse the hyperspace. The space visualized by clustering-based information visualization systems helps user explore retrieved results. On the other hand, we have proposed to visualize the set of documents that is gathered as a result of the user's RB-based interaction[13].

Document clustering-based visualization is suitable for our aims, because it is assumed that a user usually gathers the pages of interest from various Web sites, and most documents have no direct hyperlink to others. In particular, this assumption becomes valid in retrieval step.

In order for users to understand context information from the visualized results, presenting only document clusters is not enough, but the relationship

among clusters should also be presented. The SOM-based visualization systems can satisfy this to some extent, but the obtained structure seems to be fixed, even if users can manipulate the visualized space with fisheye or fractal operation[16]. Furthermore, we think that the obtained document clusters should be presented to users as lists, because Web users are familiar with the document lists that are returned by most of search engines.

Therefore, we have proposed to visualize both of document and keyword space. Document clusters are presented to users as lists, while keyword space is visualized so that the relationship among document clusters can be reflected. For visualizing the keyword space, we employed the **keyword map** [12], on which the keywords extracted from documents are arranged so that the pair of keywords that frequently appears in the same documents can be arranged closely to each other.

The point is how to relate the keyword map with document space, and we have proposed a landmark-based approach, called plastic clustering method[11].

A prototype interface has been developed based on server-side programming technique. A user can interact with the Web with ordinary Web browsers as usual. The interface displays a small control panel on a separate browser window, which provide users with several assist for collecting pages as well as for analyzing the topic information over collected document set.

2.2 Relevance Feedback

Interaction should be bidirectional. That is, interactive interface should not only provide users with information in understandable manner, but also get their intentions and preferences. Relevance feedback (RF) is one of major approaches for implicitly obtaining the users' preferences.

Conventional RF algorithms[3, 8] modify a profile (query) vector based on user's judgment (relevant or irrelevant) on the retrieved documents. In this case, the user's intention is estimated indirectly from the document space. The FISH View system[10] extracts the user's viewpoint from the diagram, in which the user groups documents hierarchically. There also exists the system that supports the user's query generation by presenting the related keywords[9]. However, it is not RF approach in the sense the user has to generate the Boolean query manually.

As noted in section 2.1,it is rational that the interaction between humans and the Web involves existing search engines. Although the conventional relevance feedback technique is basically based on vector space model (VSM), it should be combined with widely-used search engines such as Google. That is, a query vector as a result of relevance feedback should be converted to a set of keywords, which can be submitted as a query to usual search engines. An easy solution for that is to select a couple of keywords that have higher weights in the query vector than others. However, the conversion from a query vector to a set of keywords is indirect approach for inferring a user's intention or preference, because a user has to select documents in spite of what he finally wants a set of query keywords.

In this paper, we propose keyword map-based RF, which infers the user's intention from the keyword space. This approach is more direct than the conventional RF algorithms applied to document space. It can also be said that keyword map-based RF is more flexible than conventional RF, as the latter cannot obtain arbitrary combination of keywords. Finally, keyword arrangement can reflect user's intention more implicitly than the diagram used by the FISH View system.

3 Relevance Feedback based on Keyword Map

3.1 Keyword Map Visualization System

A keyword map-based information visualization system is developed for visualizing the topic distribution within a document set[12]. The developed system called TMIT (Topic Map Idea Tool) employs the spring model[14] to arrange keywords on 2D space. Although a number of information visualization systems employ the 3D graphics, they seem to be suitable for the facilities such as museum, where visitors use the systems. We claim that the system that can be in daily use should be simple. Therefore, we employ the 2D graphics. The basic algorithm of TMIT is as follows.

1. Define the distance l_{ij} between keyword i and j based on their similarity $R_{ij}(\in [-1,1]^1)$ by Eq. (1) (m is positive constant).

$$l_{ij} = m\left(1 - R_{ij}\right). \tag{1}$$

2. The moving distance of keyword i in each step, $(\delta_{xi}, \delta_{yi})$, is calculated by Eq. (2).

$$(\delta_{xi}, \delta_{yi}) = \left(c\frac{\partial E}{\partial x_i}, c\frac{\partial E}{\partial y_i}\right), \tag{2}$$

$$E = \sum_i \sum_j \frac{1}{2} k_{ij} \left(d_{ij} - l_{ij}\right)^2, \tag{3}$$

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}. \tag{4}$$

3. In each step, the center of gravity is adjusted to the center of 2D space.

In addition to this basic algorithm, an arrangement priority based on **spring constant** is introduced [14][2]. It can be understood from Eq. (3) that the influence of strong spring (with large spring constant) is greater than that of weak ones. Here, the springs connecting to focused keywords (such as landmarks[11]) are given larger spring constant than others, so that they can have priority than other keywords in terms of arrangement.

[1] In the current keyword map system, $R_{ij} \in (0,1]$ when keywords co-occur within documents, and R_{ij}=–1 when they do not appear within a document.

[2] Another arrangement priority based on frictional force is also introduced for considering the topic stream, which is out of scope and omitted in this paper.

3.2 Keyword-Pair Extraction for Relevance Feedback

The keyword map system currently implemented considers the information flow from the system to a user. In this subsection, the information flow from a user to the system is considered.

When a keyword map is presented to a user, he usually finds the difference between the keyword arrangement on the map and his background knowledge. Therefore, he wants to modify the arrangement, as he likes. If the system can infer the user's intention from the keyword map modified by him, relevance feedback can be available.

Let us consider the following cases:

1. A user rearranges the keyword A close to keyword B, which were initially arranged far away from each other.
2. A user moves apart keyword A and B, which were initially arranged close to each other.

In the first case, the user estimates the relationship between keyword A and B closer than the initial keyword map. Therefore, collecting new document that contain both keywords should satisfy the user's interest. The latter case might be more complicated, and there will be several possibilities. For example, a user might simply want documents containing keyword A but B (i.e., a query might be "A AND NOT B".). As for another possibility, the user might want to divide the topic represented by keyword A and B into two detailed topics. In this case, finding new keywords that bridge keyword A and B will be useful for the user.

In this paper, we proposed a method for extracting such keyword pairs as discussed above, from a user's modification on a keyword map. Extracting such keyword pairs is expected to be a fundamental process for realizing keyword map-based RF.

In the following algorithm, an input data file for keyword map (KData) and the data file for keywords' coordinates in the modified map (XYData) are given. KData stores the similarity $S_{Ki}\,(= R_{lm} \in [-1, 1])$ for every keyword pair p_i (w_l, w_m), and XYData stores the coordinates (x_i, y_i) of every keyword w_i on the map after the user's modification.

1. Calculate similarities S_{Xi} for each keyword pair p_i, based on the distance d_i between the keywords, from XYData. The d_M is the maximum distance among all keyword pairs.

$$S_{Xi} = 1 - (d_i/d_M)\,. \tag{5}$$

2. Translate S_{Ki}s in KData into value within [0,1] by Eq. (6).

$$S'_{Ki} = \max(S_{Ki}, 0)\,. \tag{6}$$

3. For each keyword pair p_i, calculate the degree of "farness" (Far(S'_{Ki})) and "nearness" (Near(S'_{Ki})) in KData, and those (Far(S_{Xi}) and Near(S_{Xi})) in

XYData by Eq. (7) and (8), respectively.

$$\mathrm{Far}\,(x) = \max\left(-\frac{x}{t} + 1, 0\right), \tag{7}$$

$$\mathrm{Near}\,(x) = \max\left(\frac{x-t}{1-t}, 0\right). \tag{8}$$

4. Extract the keyword pairs having high values calculated by Eq. (9) as "Far2Near" pairs.

$$V_i^{F2N} = \max\left(\mathrm{Near}(S_{Xi}), \mathrm{Far}(S'_{Ki})\right) \ldots \mathrm{Near}(S_{Xi}), \mathrm{Far}(S'_{Ki}) > 0,$$
$$0 \ldots \text{otherwise}. \tag{9}$$

5. Extract the keyword pairs having high values calculated by Eq. (10) as "Near2Far" pairs.

$$V_i^{N2F} = \max\left(\mathrm{Near}(S'_{Ki}), \mathrm{Far}(S_{Xi})\right) \ldots \mathrm{Near}(S'_{Ki}), \mathrm{Far}(S_{Xi}) > 0,$$
$$0 \ldots \text{otherwise}. \tag{10}$$

4 Experiments on Keyword Map-based Relevance Feedback

The experiments on keyword map-based relevance feedback are performed with using the prototype interface shown in Section 2.1, combined with the algorithm described in Section 3.2. Currently, the algorithm has not been yet implemented inside the prototype interface, and it is difficult to perform experiments with test subjects. Therefore, the section shows the examples how the proposed algorithm works on a keyword map actually generated from the retrieval results of existing search engine. Furthermore, the advantage of the keyword map-based RF against conventional RF is also discussed based on the examples. It should also be noted that the experiments are performed on Japanese Web pages, and results are translated from Japanese into English hereafter.

When applying the algorithm to a keyword map, the parameter t in Eq. (7) and (8) should be given. In the experiments, R_{lm} is given based on Jaccard coefficient regarding co-occurrence of keywords l and m within a document set. Therefore, $t = 0.25$ for farness and $t = 0.5$ for nearness are used for the KData, because two keywords are assumed to be highly related if R_{lm} exceeds 0.5. The value for farness is determined empirically, so that the number of Far2Near keyword pairs can be limited. As for the XYData, $t = 0.5$ for farness and $t = 0.9$ for nearness are used, because we assume that users will arrange the keywords that they want to discriminate with more than half distance of the max distance among keywords. The value for nearness is determined empirically, in order to reduce the number of Far2Near keyword pairs.

A query "Kanazawa[3] AND Sightseeing (Kanko)" is submitted to the Google, and top 10 pages are collected as an initial page set, from retrieved result. Figure1

[3] Kanazawa is the name of the city where AM2004 workshop is held.

shows the keyword map that the prototype interface generates from the initial page set.

In Fig. 1, the word “Noto” and “Hot Spring” (Onsen) are close, from which “Gourmet” is far away. In fact, “Noto” and “Hot Spring” appear in the same set of documents, and “Gourmet” co-occurs with those keywords in only a single document.

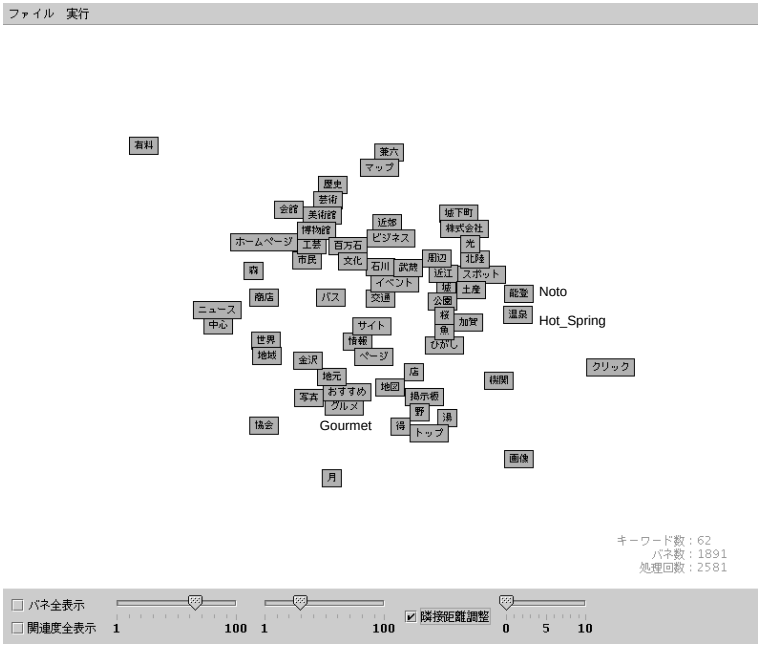


Fig. 1. Keyword Map Generated from Initial Document Set

The keyword map system used in the experiments is improved to be interactive, so that the relevance feedback can be available. In particular, a user can drag any keyword to arbitrary position, and other keywords are arranged automatically based on spring model. After the modification, the system can output the XYData as noted in Section 3.2.

Figure 2 and 3 show the maps that are modified by a user from the initial arrangement. In Fig. 2, “Noto” and “Gourmet” are close, and “Hot Spring” is far from those keywords. On the other hand, “Hot Spring” and “Gourmet” are close, and “Noto” is far from those in Fig. 3.

From Fig. 2, three pairs are extracted as Far2Near, whereas 5 pairs are extracted as Near2Far with the proposed algorithm. Among them, “Gourmet AND Noto” can obtain the highest Far2Near score, and “Hot Spring AND Noto” can obtain the highest Near2Far score.

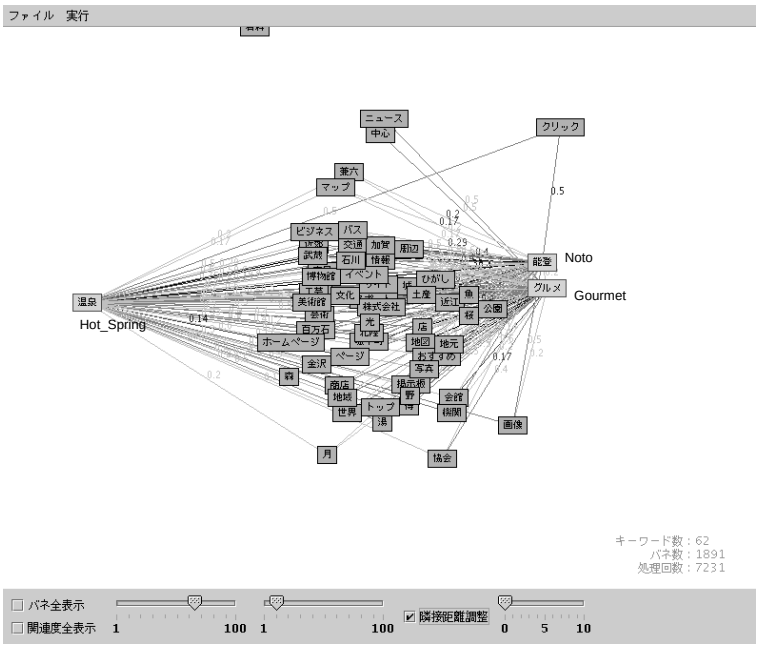


Fig. 2. Keyword Map Edited by User (1)

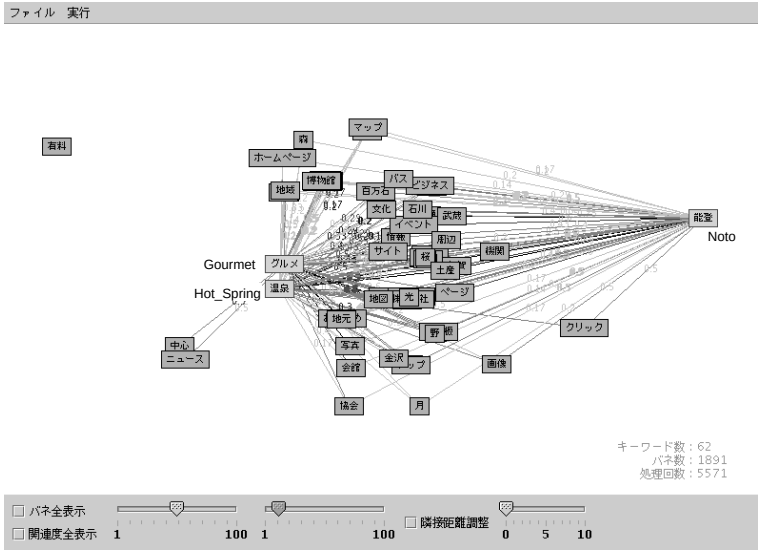


Fig. 3. Keyword Map Edited by User (2)

On the other hand, from Fig. 3, 19 Far2Near pairs and a single Near2Far pair, "Hot Spring AND Noto" are respectively extracted. Among Far2Near pairs, "Gourmet AND Hot Spring" has the highest score. These results show that the proposed algorithm can extract appropriate keyword pairs based on the difference between the initial data set and modified keyword map.

Furthermore, the difference between the result of Fig. 2 and that of Fig. 3 clearly shows the advantage of keyword map-based RF against conventional RF. As noted above, "Noto" and "Hot Spring" appear within the same documents, d_1, d_2, and d_3 among 10 documents. On the other hand, "Gourmet" is contained in 3 documents, d_1, d_5, d_6. Therefore, only a single document, d_1, contains all three keywords.

When Rocchio-based RF[3] with TFIDF weighting (i.e. conventional RF) are performed with d_1, d_2, d_3, d_5, d_6 as positive examples and other 5 documents as negative ones, those three keywords obtain high positive weights. Whereas, when only d_1 is given as positive, "Noto" and "Gourmet" obtain positive weights, and "Hot Spring" obtain negative weight, which corresponds to the result of Fig. 2. This result can be the query "Gourmet AND Noto AND NOT Hot Spring".

Then, how to obtain the query "Gourmet AND Hot Spring AND NOT Noto" with conventional RF? It seems difficult to obtain such a query, because "Noto" and "Hot Spring" appears within the same set of documents. In this example, such a query cannot be obtained by removing either d_2 or d_3 from positive document set. This result show the keyword map-based RF is more flexible than conventional RF when used in combination with existing search engines.

5 Conclusion

The relevance feedback based on interactive keyword map system is proposed. For the first step towards realizing keyword map-based RF, an algorithm is proposed for extracting the pair of keywords that reflects a user's interest from the keyword map modified by the user. Experimental results show how the algorithm works on the keyword map that is modified by the user, and discuss the difference between the RF based on keyword map and conventional RF methods.

We have already proposed the concept of Retrieval, Browsing, and Analysis (RBA)-based interaction. The prototype interface employs the keyword map visualization system so that users can easily understand the context of their interaction with the Web. Combination of the prototype interface with the algorithm proposed in this paper will realize bidirectional web interaction between users and the Web.

References

1. Ackerman, M. et al., "Learning Probabilistic User Profiles," AI Magazine, Vol. 18, No. 2, pp. 47—56, 1997.
2. Armstrong, R., Freitag, D., Joachims, T., Mitchell, T., "WebWatcher: A Learning Apprentice for the World Wide Web," AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, 1995.

3. Baeza-Yates, R. and Ribeiro-Neto, B., "5. Query Operations" in Modern Information Retrieval, Addison Wasley, 1999.
4. Gershon, N., LeVasseur, J., Winstead, J., Croall, J., Pernick, A., Ruh, W., "Case Study: Visualizing Internet Resources," Proc. Information Visualization (INFOVIS'95), pp. 122—128, 1995.
5. Hearst, M. A. and Pedersen, J. O., "Reexaming the Cluster Hypothesis: Scatter/Gather on Retrieval Results," Proc. Of 19th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96), pp. 76-84, 1996.
6. Lieberman, H., "Letizia: An Agent That Assists Web Browsing," Proc. 14th Int'l Joint Conf. on Artificial Intelligence (IJCAI95), pp. 924—929, 1995.
7. Mukherjea, S., Hara, Y., "Visualizing World-Wide Web Search Engine Results," Int'l Conf. on Information Visualization, p.400—405, 1999.
8. Onoda, T., Murata, H. and Yamada, S., "Document Retrieval based on Relevance Feedback with Active Learning," SIG-KBS-A301 (JSAI), pp.13-18, 2003.
9. Sunayama, W., Ohsawa, Y. and Yachida, M. "A Search Interface with Supplying Search Keywords by Using Structure of User Interest," J. of Japan Society for Artificial Intelligence, Vol. 15, No. 6, pp. 1117-1124, 2000.
10. Takama, Y. and Ishizuka, M., "FISH VIEW System: A Document Ordering Support System Employing Concept-structure-based Viewpoint Extraction," J. of Information Processing Society of Japan, Vol. 41, No. 7, pp.1976-1986, 2000.
11. Takama, Y. and Hirota, K., "Web Information Visualization Method Employing Immune Network Model for Finding Topic Stream from Document-Set Sequence," J. of New Generation Computing, Vol. 21, No. 1, pp. 49-59, 2003.
12. Takama, Y. and Tetsuya, H., "Application of Immune Network Metaphor to Keyword Map-based Topic Stream Visualization," Proc. 2003 IEEE Int'l Symp. on Computational Intelligence in Robotics and Automation (CIRA2003), pp. 770-775, 2003.
13. Y. Takama, "Intelligent Interface based on Retrieval, Browsing, Analysis Operations," *4th International Conference on Intelligent Technologies (InTech'03)*, pp. 806–811, 2003.
14. Takasugi, K. and Kunifuji, S., "A Thinking Support System for Idea Inspiration Using Spring Model," J. of Japanese Society for Artificial Intelligence, Vol. 14, No. 3, pp. 495—503, 1999 (written in Japanese).
15. Teraoka, T. and Maruyama, M., "Research Report: Adaptive Information Visualization Based on the User's Multiple Viewpoints –Interactive 3D Visualization of the WWW –," Proc. IEEE Symposium on Information Visualization (InfoVis'97), pp. 25—28, 1997.
16. Yang, C. C., Chen, H., Hong, K., "Internet Browsing: Visualizing Category Map by Fisheye and Fractal Views," Proc. Int'l Conf. On Information Technology: Coding and Computing (ITCC'02), pp. 34—39, 2002.
17. Zamir, O. and Etzioni, O., "Grouper: A Dynamic Clustering Interface to Web Search Results," Proc. 8th International WWW Conference, 1999.

A Correlation-Based Approach to Attribute Selection in Chemical Graph Mining

Takashi OKADA

Department of Informatics, Kwansei Gakuin University 2-1 Gakuen, Sanda-shi, Hyogo, 669-1337 Japan
E-mail: okada-office@ksc.kwansei.ac.jp,

Abstract. Data mining often encounters a problem with a huge number of descriptive features. Authors have analyzed structure activity data of dopamine antagonists, where we have to select useful features out of numerous fragments extracted from chemical structures. Correlation coefficients among categorical attributes are used to select attributes. Rules obtained by the cascade model were evaluated from chemists' point of view, and the importance of attribute selection was confirmed.

1 Introduction

One of the challenging problems in data mining is to cope with vast amount of attributes. A typical example is to find important genes from millions of single nucleotide polymorphisms (SNPs) that explain the cause of some disease. Authors have analyzed the structure-activity relationships using linear fragments derived from chemical graphs. The number of meaningful fragments was about 2000-3000. Its number is fewer than that of SNPs problem, but we cannot reach valuable knowledge unless we overcome this problem.

Association rule mining is a method that succeeded to solve the numerous attributes problem [1]. It can detect frequent itemsets in a customer's basket selected from thousands of items in a supermarket. However, its success depends on the sparseness of the data. That is, the method think of a few items in a basket, and it does not take into account the items that do not appear in the basket. When we treat a dense dataset, there appears a huge number of itemsets resulting in the combinatorial explosion of the itemset lattice. The cascade model developed by the author constructs the itemset lattice, too [2, 3]. It can handle a dense dataset, as it detects a useful rule from a single link located at the shallow region of the lattice. But, the number of attributes is limited to 100-150, and some improvement is necessary in order to treat a dataset with numerous attributes.

In the regression analysis we usually employ attribute selection procedures in order to avoid the over-fitting and the instability of the model arising from the collinearity among attributes. Attribute selection was also found useful in the decision tree approach, when a dataset contains more than dozens of attributes [4].

This paper reports an attempt to introduce attribute selection to the mining of SAR from chemical graphs. The next section briefly describes the overview of the analysis, the basic introduction to the mining method as well as the problems encountered. The selection of categorical attributes is done by using correlation coefficients among them, the definition of which is given in Section 3. Results of application to the chemical graph mining are shown in Section 4, where the quality of resulting rules is judged from chemist's point of view.

2 Mining Chemical Graphs by the Cascade Model

2.1 Overview of the Dopamine D2 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. 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.

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. The problem challenged in this paper is to discover the structural characteristics responsible for D2 antagonist activity, which is known to be the hardest problem among 4 antagonists.

Figure 1 shows the brief scheme of the analysis. All structural formulae of chemicals are stored in the SDF format file. First, four physicochemical properties: HOMO, LUMO, Dipole and LogP, are calculated by MM-AM1-Geo and ClogP programs. We also extract many linear fragments contained in chemical graphs, and the presence/absence of these fragments in a chemical structure is used as other type of attribute. Linear fragments are expressed by constituent elements and bond types like "c3H:c3---C4H-N3", and they are used as attribute names. Details of linear fragments generation is to be published.

Obviously, the number of all possible fragments is too large. Therefore, the length of linear fragments was limited to be shorter than 8, and one of the terminal atoms of a fragment was restricted to be a heteroatom or a carbon constituting a double or triple bond. Then, we got 8041 fragments, which was too many to be analyzed by the current implementation of the cascade model. Therefore, we selected 114 fragments, of which ratio of appearances in compounds is in the range: 15%-85%.

Application of the cascade model to the table generates rules that characterize the structures of D2 antagonists. Resulting rules are interpreted and evaluated by chemists.

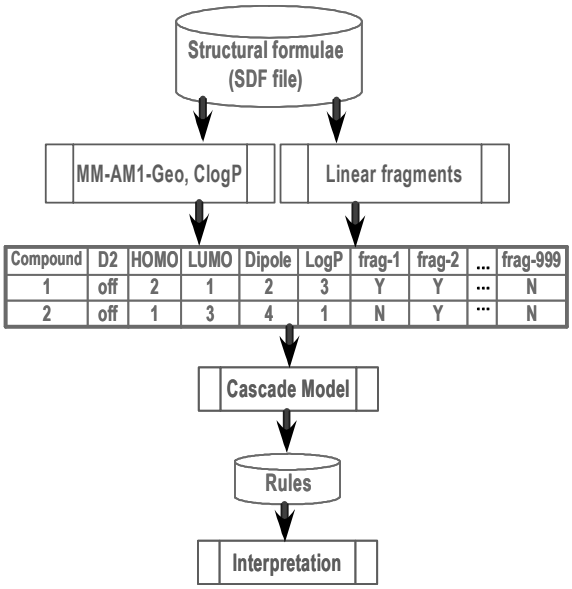


Fig. 1. Flow of chemical graph mining by the cascade model.

2.2 The Cascade Model and the Datascape Survey

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)
```

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*.

Recently, new facilities for *datascape survey* are introduced in order to reduce the number of rules [5], and to denote the details of data distribution specified by a rule [6]. Interpretation of rules became easier by these functions.

2.3 Attribute Selection Problem

There is no reason to justify the selection of 114 fragments appearing in 15%-85% of compounds. In fact, chemists could notice other important fragments contributing to the D2 activity by browsing structural formulae. However, if we use more attributes, the combinatorial explosion in the lattice size prohibits the analysis. The past experience suggested that the upper limit in attribute numbers was 100-150.

On the other hand, analysts often encountered a pair of fragments with the same number of supporting compounds like O1=S4-c3:c3H and S4-c3:c3H. The latter support must always be larger or equal to that of the former, since the latter is a substructure of the former. Then, the equality of these supports means that they appear exactly in the same compounds, and the selection of both fragments as attributes is redundant. That is, the correlation coefficient between these two attributes is 1.0.

Omission of an attribute from such pairs is expected to enable the analysis using more attributes with lower supports. Furthermore, attribute pairs do not need to be completely correlated. We can omit an attribute if it is in a highly correlated pair. Therefore, we decided to introduce a correlation coefficient between attributes, and to use it as a criterion to omit/keep the attribute.

3 A Correlation Coefficient between Categorical Variables

Correlation coefficient is a well-known concept in numerical attributes. Recently, we introduced a generalized covariance using vector expression for value differences [7]. Uniform treatment of covariance became possible among numerical and categorical variables. Here, we briefly mention a special case to define a correlation coefficient between a pair of binary attributes.

Gini successfully defined the variance of categorical data [8]. He first showed that the following equality holds for the variance of a numerical variable x_i.

$$V_{ii} = \left(\sum_a \left(x_{ia} - \bar{x}_i \right)^2 \right) \Big/ n = \frac{1}{2n^2} \sum_a \sum_b \left(x_{ia} - x_{ib} \right)^2 \quad , \tag{1}$$

where V_{ii} is the variance of the i-th variable, x_{ia} is the value of x_i for the a-th instance, and n is the number of instances.

Then, he introduced the distance definition (2) into value differences of (1), and got the categorical variance expression (3), which is well known as Gini-index.

$$x_{ia} - x_{ib} \begin{cases} = 1 & \text{if } x_{ia} \neq x_{ib} \\ = 0 & \text{if } x_{ia} = x_{ib} \end{cases} , \tag{2}$$

$$V_{ii} = \frac{1}{2n^2} \sum_a \sum_b (x_{ia} - x_{ib})^2 = \frac{1}{2} \left(1 - \sum_r p_i(r)^2 \right) \quad . \tag{3}$$

Extension of the above definition to the covariance fails, if we simply change $(x_{ia} - x_{ib})^2$ to $(x_{ia} - x_{ib})(x_{ja} - x_{jb})$. We employed a vector expression, $\overrightarrow{x_{ia}x_{ib}}$, instead of the distance, $x_{ia} - x_{ib}$, in the variance definition [7]. Our proposal for V_{ij} definition was the maximum value of $Q_{ij}(\boldsymbol{L})$ while changing $\boldsymbol{L}$, and $Q_{ij}(\boldsymbol{L})$ was defined by the subsequent formula,

$$V_{ij} = \max\left(Q_{ij}(L)\right) \quad , \tag{4}$$

$$Q_{ij}(L) = \frac{1}{2n^2}\sum_a\sum_b \left\langle \overrightarrow{x_{ia}x_{ib}} \middle| L \middle| \overrightarrow{x_{ja}x_{jb}} \right\rangle \quad . \tag{5}$$

Here, $\boldsymbol{L}$ is an orthonormal transformation applicable to the value space. The bracket notation, $\langle \boldsymbol{e}|\boldsymbol{L}|\boldsymbol{f}\rangle$, is evaluated as the scalar product of two vectors $\boldsymbol{e}$ and $\boldsymbol{Lf}$ (or $\boldsymbol{L}^{-1}\boldsymbol{e}$ and $\boldsymbol{f}$). If the lengths of the two vectors, $\boldsymbol{e}$ and $\boldsymbol{f}$, are not equal, zeros are first padded to the vector of the shorter length.

		x_j: u	v	
x_i	r	n_{ru}	n_{rv}	$n_{r\cdot}$
	s	n_{su}	n_{sv}	$n_{s\cdot}$
		$n_{\cdot u}$	$n_{\cdot v}$	n

We apply this definition to the simplest 2 x 2 contingency table shown at the left, where $n_{r\cdot}$ *and* $n_{\cdot u}$ show marginal distributions. Straightforward application of (5) to this table gives the following expressions for V_{ii}, V_{jj} and V_{ij}, and a correlation coefficient R_{ij} is given by (9).

$$V_{ii} = n_{r\cdot}n_{s\cdot}/n^2 = \tfrac{1}{2}\left(1 - (n_{r\cdot}/n)^2 - (n_{s\cdot}/n)^2\right) \quad . \tag{6}$$

$$V_{jj} = n_{\cdot u}n_{\cdot v}/n^2 = \tfrac{1}{2}\left(1 - (n_{\cdot u}/n)^2 - (n_{\cdot v}/n)^2\right) \quad . \tag{7}$$

$$V_{ij} = \frac{\left|n_{ru}n_{sv} - n_{rv}n_{su}\right|}{n^2} \quad . \tag{8}$$

$$R_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}} \quad . \tag{9}$$

The numerator in (8) is the critical term used to represent the extent of dependency between two variables. In fact, the correlation coefficient is 1.0 (0.0) for completely dependent (independent) data, respectively.

4 Results and Discussion

We applied the attribute selection scheme to the dopamine D2 antagonist problem. All generated fragments reached 8041 kinds. First, we selected a fragment as an attribute, if the probability of its appearance satisfied the following condition,

$$edge < P(\text{fragment}) < 1.0 - edge \quad . \tag{10}$$

When *edge* was set to 0.01, 0.02, 0.03, 0.05, 0.10 and 0.15, the number of selected fragments were 1698, 1056, 730, 377, 176, and 114, respectively. We employed the presence/absence of these fragments as the initial attribute set $\{x\}$.

4.1 Attribute Selection using Correlation Coefficients

The procedure of the attribute selection is as follows.

1. Calculate correlation coefficients among all attribute pairs: x_i and x_j, and put the pair into a list: *pairs*, when it satisfies the condition: $R_{ij} > min\text{-}R_{ij}$.
2. Sort *pairs* in the descending order of R_{ij}.
3. Pop *pairs,* and get a pair: x_i and x_j.
4. Omit an attribute (x_i or x_j) from $\{x\}$, if both attributes are members of $\{x\}$.
5. Repeat steps 3 and 4 until every pair in *pairs* is examined.

When we omit an attribute from a correlated attribute pair at step 4, a longer fragment is kept in the attribute set. It is because an analyst usually gets more ideas from the longer attribute name, when it appears in a rule. However, analysts have to consult a list of omitted attributes so that some important features are lost.

Figure 2 shows the numbers of selected attributes in log scale for 6 *edge* values, changing $min\text{-}R_{ij}$ value to 1.0, 0.99, 0.97, 0.95, 0.90, 0.85, 0.80, 0.75, and 0.70. Here, no attribute selection is carried out at $min\text{-}R_{ij}$ = 1.0, and attributes in perfectly correlated pair are omitted at $min\text{-}R_{ij}$ = 0.99.

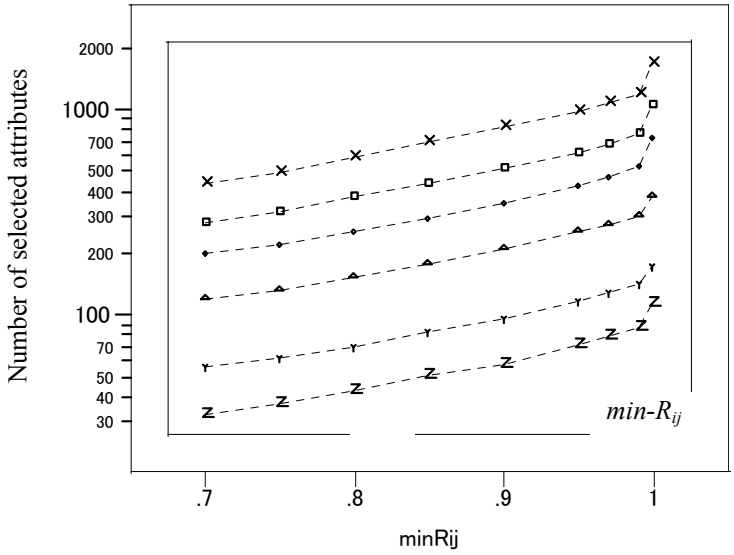


Fig. 2. Numbers of selected attributes changing $min\text{-}R_{ij}$ value. 6 lines from upper to lower show the results using *edge* = 0.01, 0.02, 0.03, 0.05, 0.10 and 0.15, respectively.

Shapes of plots in the figure do not depend on *edge* values. Another interesting point is steep slopes found at the right end of the plots. It means about 20-30% of attributes in the initial attribute sets are completely correlated in the chemical graph mining using linear fragments. Roughly speaking, about half attributes are omitted when we employ $min\text{-}R_{ij}$ = 0.90. Therefore, we can conclude that attribute selection using correlation coefficients works well in reducing the number of attributes.

4.2 Effects to Lattice Size

Lattice expansion in the cascade model is controlled by a parameter, *thres*. The smaller *thres* value we use, the larger number of nodes in the lattice we examine. Ordinary values of *thres* in this application have been in the range: 0.15-0.2, and the number of nodes in the lattice was 5,000 – 30,000 using 100-150 attributes. We examined the lattice size changing *edge* and *min-*R_{ij} values, for three *thres* values: 0.15, 0.175 and 0.20.

Figure 3 shows rough contour maps of the number of nodes (*#nodes*) in the lattice, where y-axis is *min-*R_{ij} and x-axes are number of selected attributes (*#attributes*) in (A) and *edge* in (B), respectively. The calculated points are shown by '+' in the figure, but the points resulted in the combinatorial explosion of the lattice are not depicted. The lowest contour lines (*#nodes* = 3000) are indicated by arrows.

Contour lines in (A) are all more or less parallel to y-axis in large *min-*R_{ij} values, but they trailed to the bottom right corner in small *min-*R_{ij} values. This fact shows that the lattice size does not arise sharply when we use more uncorrelated attributes. In fact, we could employ 400-500 attributes selected from more than 1000 attributes.

The contour lines in Figure 3 (B) are drawn from the upper right to the bottom left corner. The meaning of this fact can be seen by the inspection of two ⟠ points and two + points near the gray contour in the top right map. The data for these four points are summarized in Table 1.

Table 1. Calaculated results for 4 points near a gray contour (*thres*=0.15)

Point	*edge*	*min-*R_{ij}	*#attributes*	*#nodes*	*#detected*	*#rules*	*score*
P1	0.02	0.70	287	4992	23	6 (3)	3
P2	0.05	0.80	155	5983	39	8 (4)	3
P3	0.10	0.90	130	5223	72	9 (4)	2
P4	0.15	0.99	88	6265	97	14 (5)	2

This table shows that similar number of nodes emerge from a wide range of *#attributes* (88 – 287). That is, correlated attributes grow the lattice size, while the uncorrelated ones depress it. As a result, attribute selection using low *min-*R_{ij} affects greatly in the reduction of lattice size.

4.3 Evaluation of Rules

A matter of great importance is not the number of selected attributes nor the lattice size, but the quality of rules. *#detected* column in Table 1 shows the number of detected links with large *BSS* values, where the optimization to a rule starts. *#rules* column denotes the number of resulting rules. Also shown in parentheses are the numbers of principal rules after rules organization. These numbers tend to increase as we use higher *min-*R_{ij} values. Appearance of many rules does not always lead to good knowledge discovery. For example, there exist many highly correlated attributes in the calculation at P4, which might be a cause of redundant rules. In fact, the increase in the number of principal rules is very limited.

Here, we introduce an evaluation scheme of rules. Analysts noticed three important substructures relevant to the D2 antagonist activity, browsing various rules. They

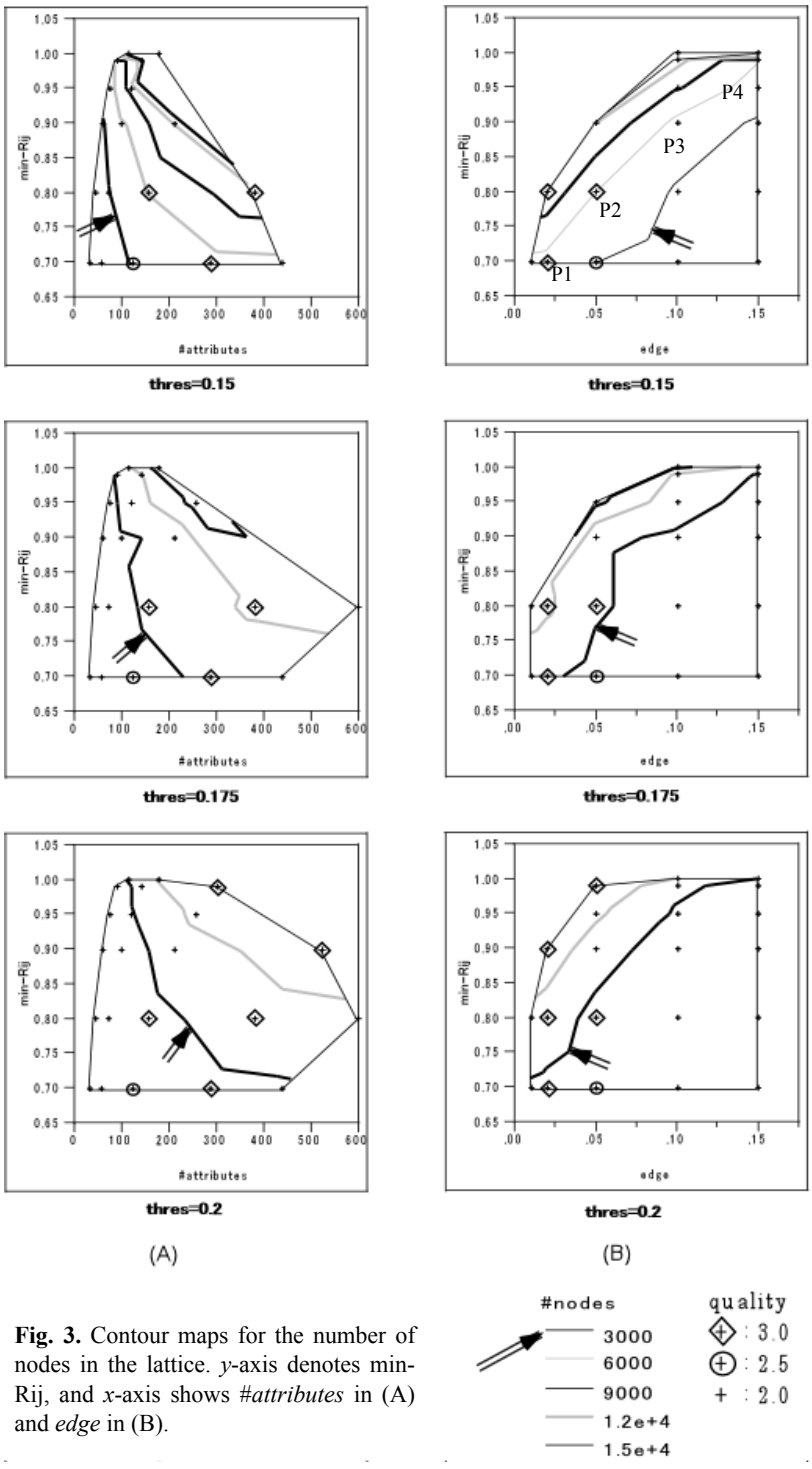


Fig. 3. Contour maps for the number of nodes in the lattice. *y*-axis denotes min-Rij, and *x*-axis shows *#attributes* in (A) and *edge* in (B).

were aromatic ether, tertiary amine separated from an aromatic ring by 3 single bonds, and CO group bonded to tertiary amine. The appearances of these features in rules were used to judge their quality. That is, we search three features in the main conditions of principal rules, and the numbers of found features were employed as the *score* of rules. When a feature appears only in a relative rule, we count it as 0.5. Note that the appearance of a feature is counted only once. Therefore, the highest *score* of resulting rules is 3. This evaluation scheme is rough, as the true mechanisms for the appearance of D2 antagonist activity are not known yet. But we can expect that this *score* will be a guide to judge the quality of rules.

The last column of Table 1 shows this *score* for four calculations. We can notice that the number of rules has no meaning from this viewpoint. The next problem is to find adequate values leading to good rules for the three parameters: *edge*, *min*-R_{ij} and *thres*,.

The calculated points with *score*=3, 2.5, 2 are shown by ⊕, ⊕and +, respectively in Figure 3. The distribution of high score points in Figure 3 (A) indicates that neither the number of attributes nor the size of the lattice have direct relationships to the *score* of rules. On the other hand, Figure 3 (B) shows that high score rules result from calculations at *edge* = 0.02 and 0.05. The attribute selection using *min*-R_{ij} = 0.8 seems to give good rules always.

The suggested plan of mining is to employ relatively lower *edge* values, followed by the selection of attributes using *min*-$R_{ij} \cong$ 0.8. The effect of *thres* value is limited, as far as the objective of mining is to grasp rough characteristics of chemical graphs.

5 Concluding Remarks

Attribute selection scheme introduced in this paper is essentially a method to cope with collinearity among explanation attributes. Lots of researches have been done to solve this problem in the regression analysis. They include various attribute selection schemes, canonical regression method and partial least squares.

Among mining methods for categorical data, reduct concept in the rough set solved this problem clearly [9]. However, the implementation cannot treat thousands of attributes. Another approach from the mining community is the closed itemset concept in the association rule mining [10, 11]. It is used to compute long frequent itemsets fast, and it is also applied in the filtering of rules to omit redundant ones. However, this method is useful only when a pair of attributes correlates completely. Even if the correlation coefficient is larger than 0.99, the method cannot be applied to data with a noise instance.

The cascade model has also encountered the collinearity problem. It first showed the collateral correlations in a rule, which have been used to detect an attribute highly correlated to the main condition. This information helps an analyst to interpret rules. Correlated attributes leads to the generation of a pair of rules, covers of which overlap considerably. This problem was solved by the reorganization of rules into a principal rule and its relative rules. The attribute selection introduced in this paper has been shown to be useful in the reduction of the lattice size. Moreover, the omission of a correlated attribute cuts self-evident collateral correlations, and it also reduces the

number of relative rules. Therefore, the load of an analyst has been reduced. All these functions work for partially correlated attributes, and it offers a superior framework than the closed itemset. However, the rule organization scheme in the cascade model has rooms to be improved, as the numbers of principal rules in Table 1 should reflect the scores given by an expert's inspection.

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 author wishes to thank Dr. Masumi Yamakawa and Dr. Hirotaka Niitsuma of Kwansei Gakuin University for their valuable discussions.

References

1 Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. Proc. VLDB (1994) 487-499
2 Okada, T.: Rule Induction in Cascade Model based on Sum of Squares Decomposition. Principles of Data Mining and Knowledge Discovery (Proc. PKDD'99), LNAI 1704, Springer-Verlag (1999) 468-475
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 Liu, H., Motoda, H. (ed.): Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers (1998)
5 Okada, T.: Datascape Survey using the Cascade Model. In: Satoh, K. et al. (eds.) Discovery Science 2002. LNCS 2534, Springer-Verlag (2002) 233-246
6 Okada, T.: Topographical Expression of a Rule for Active Mining. In: Motoda, H. (ed.) Active Mining. IOS Press, (2002) 247-257
7 Okada, T.: A Note on Covariances for Categorical Data. In: Leung, K.S. et al (eds.) Intelligent Data Engineering and Automated Learning - IDEAL 2000. LNCS 1983, Springer-Verlag (2000) 150-157
8 Gini, C.W.: Variability and Mutability, contribution to the study of statistical distributions and relations, Studi Economico-Giuridici della R. Universita de Cagliari (1912). Reviewed in: Light, R.J., Margolin, B.H.: An Analysis of Variance for Categorical Data. J. Amer. Stat. Assoc. 66 (1971) 534-544
9 Pawlak Z.: Rough sets: Theoretical aspects of reasoning about data. Dordrecht: Kluwer (1991)
10 Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient Mining of Association Rules Using Closed Itemset Lattices. Information Systems, 24 (1) (1999) 25-46
11 Zaki, M.J., Hsiao, C.J.: CHARM: An efficient algorithm for closed itemset mining. In: Proc. SDM'02, SIAM (2002) 457-473

Combining Partial Rules and Winnow Algorithm: Results on Classification of Dopamine Antagonist Molecules

Sukree Sinthupinyo[1], Cholwich Nattee[1], Masayuki Numao[1], Takashi Okada[2], and Boonserm Kijsirikul[3]

[1] Department of Architecture for Intelligence, 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
[3] Department of Computer Engineering, Chulalongkorn University
boonserm.k@chula.ac.th

Abstract. In this paper, we propose an approach which can be combined with the rules of Inductive Logic Programming to classify multiclass data. This approach is based on the idea that if a whole rule cannot be applied to an example, some partial matches of the rule can be useful. The most suitable class should be the class whose important partial matches cover the example more than those from other classes. Hence, the partial matches of the rule, called *partial rules*, are first extracted from the original rules. Then, we utilize the idea of Winnow algorithm to assign the weight to each partial rule. Finally, the partial rules and the weights are combined and used to classify new examples. The weights of partial rules also show another aspect of the knowledge which can be discovered from the data set. In the experiments, we apply our approach to a multiclass real-world problem, classification of dopamine antagonist molecules. The experimental results show that the proposed method gives the improvement over the original rules and yields 88.58% accuracy by running 10-fold cross validation.

1 Introduction

In recent years, Inductive Logic Programming (ILP) has been widely applied to various real-world applications [1, 2]. Standard ILP are usually two-class classifier (positive and negative classes). A test example which matches with some rules is classified as positive class, while the example which does not match with any rule are classified as negative class. This causes some troubles when we need to use ILP in multiclass problems. In such problems, when a test example does not match with any rule or matches with some rules from more than two classes, we cannot determine which class is most suitable for the example.

In this paper, we propose an approach which can help ILP in multiclass problems. Our approach is based on the idea that if a whole rule cannot be applied to an example, some parts of rule may match with that example. Thus, we can make use of these matches to determine the class for the example. The most suitable class should be the class whose the number of important matches is higher than those of other classes. Thus, in our approach, we first extract some part of rule which will be used as *partial rule*. Then, all partial rules are given the importance in term of weights using Winnow-based approach [3]. Finally, the partial rules and the weights are combined and used to classify new examples. Moreover, the weights assigned to the partial rules also show another aspect of the characteristic of data set that is very useful in knowledge discovery fashion.

We apply our approach to a real-world problem, classification of dopamine antagonist molecules. Dopamine antagonist molecule is a kind of molecules which can block the binding between dopamines and dopamine receptors in the signal transfer process in human brain. The excessive levels of the dopamine have been implicated in schizophrenia. Hence, in the medical treatment of schizophrenic patients, the dopamine antagonist molecules are used to decrease the signal transfer level which can limit the effect of the high density of dopamines. The knowledge discovered from this domain may be useful for schizophrenic drug development.

The paper is organized as follows. In the next section, we present a concept of ILP and the obstacles when ILP is applied to multiclass problems. The partial rules extraction strategy and the weights adjustment are expressed in Section 3 and Section 4, respectively. The details of the experiments are presented in Section 5. The paper ends with the conclusion in Section 6.

2 Using ILP in Multiclass Problems

ILP is the Machine Learning technique which is originally proposed as a two-class classifier. ILP aims to construct a rule set that covers all positive examples and none of the negatives. The output of ILP is the first-order rules which will be used to classify new examples. This causes some troubles when we need to use ILP in multiclass problems, i.e. (1) how to construct the rule for each class, and (2) how to select the class for each example. In the former case, as mention earlier, ILP systems search for the rules which cover positive examples, however, in multiclass problems, we need to construct the rules for each class. Hence, the additional technieques must be used to help ILP to construct the rules, such as one-against-all, round robin rule learning [4], and loss-based decoding [5]. Nevertheless, in this work, we emphasize on the latter case. Thus, we employ the common method, one-against-all, to construct the rules for each class.

In the one-against-all algorithm, a k-class problem is reduced to k two-class problems. To generate the rules for class i, the training examples are organized by using the training examples of class i as positive examples and using the training examples of class j where $j = 1, ..., k$ and $j \neq i$ as negative examples. For example, our data set contains 4 classes, i.e. D1, D2, D3, and D4 (as will be

described in Section 5). We use the training examples of class D1 as the positive examples and use those of classes D2, D3, and D4 as the negatives for learning rules of class D1. Using this strategy, the obtained rules are unordered.

The problem of class selection arises when an example does not exactly match with any rule or matches with some rules from more than two classes, especially in case of unordered rules. In case of ordered rules, the class selection is not complicated, the example which does not exactly match with any rule can be classified as the default class, while the example which matches with multiple rules from different classes can be classified as the class of the higher order rule. However, in case of unordered rules, as constructed in this work, ILP's rules alone cannot select the class of the example which does not match with any rule or matches with multiple rules from different classes. Hence, we propose an approach which is based on the idea that if the whole rule cannot be applied to the example, we can utilize some partial matches of the rule to determine the most appropriate class.

In our experiments, we employ an ILP system, Aleph [6], to construct the rules for each class. The rule construction of Aleph starts with building the most specific clause, called *bottom clause*. Then, to seek for the best generalised clause, Aleph provides many search algorithms which users can select the most suitable one for their domain. In our experiments, we selected the randomized search method using an altered form of the GSAT algorithm [7] that was originally proposed for solving propositional satisfiability problems. The GSAT algorithm provided by Aleph is modified to suit the clause searching process in ILP fashion.

3 Partial Rules

As described in the previous section, our approach is based on the idea that some partial matches in the rule can be used to classify the unclassifiable examples. Hence, several parts of a rule or *partial rules* are first extracted from the original rules. Then, they are used to classify unseen examples collaboratively. The following describes our partial rule extraction algorithm.

A *partial rule* is *a rule whose body contains a valid sequence of the literals, from the body of the original rule, which starts with the literal consuming the input variables in the head of the rule.* The partial rule extraction algorithm is based on the idea of the newly introduced variables, similar idea as the feature extraction in BANNAR [8]. As shown in Fig. 1, each input variable in a literal is introduced as a new variable in some preceding literals. Thus, we group the literal which consumes the new variable and the literal which introduces that variable into the same sequence. For example, in Fig. 1, the new variable `D` introduced in literal `link(A, B, C, D)` is used as the input variable of literal `D=2.7`. Thus, we group these two literals into the same sequence, `link(A, B, C, D), D=2.7`.

Our extraction procedure starts with an empty sequence, and uses variables in the head of the rule as new variables. Then, the literal which consumes the new variables as input variables is gradually added to the sequence. The new vari-

ables introduced in the newly added literal are again used as the new variables for searching other literals to be added. The search stops when the newly added literal introduces no new variable or cannot find any literal which consumes the new variables in this newly added literal. Finally, we make all possible combinations of the two sequences which have the common variables not occurring in the head. The partial rule extraction algorithm is shown in Fig. 2.

For example, we can extract the following partial rules from the rule shown in Fig. 1.

```
molecule(A) :- atm(A, B, C, D, E, F), C=n.
molecule(A) :- atm(A, B, C, D, E, F), E=2.8.
molecule(A) :- atm(A, B, C, D, E, F), bond(A, G, B, H, I, J),
               gteq(J, 1.5).
molecule(A) :- atm(A, B, C, D, E, F), C=n, E=2.8.
molecule(A) :- atm(A, B, C, D, E, F), C=n, bond(A, G, B, H,
               I, J), gteq(J, 1.5).
molecule(A) :- atm(A, B, C, D, E, F), E=2.8, bond(A, G, B, H,
               I, J), gteq(J, 1.5).
molecule(A) :- atm(A, B, C, D, E, F), C=n, E=2.8, bond(A, G,
               B, H, I, J), gteq(J, 1.5).
```

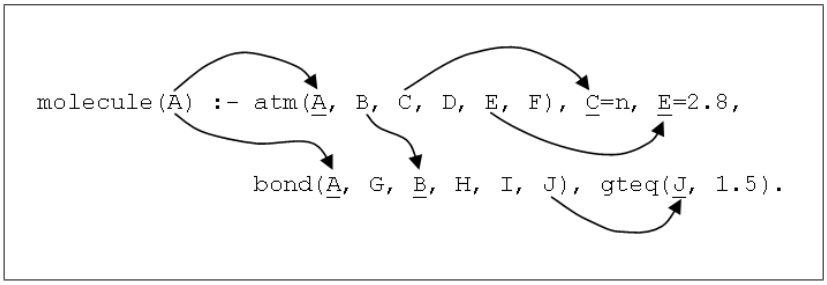


Fig. 1. New variables consumption. The underlined characters show the input variables of literals

4 Weights Adjustment Using Winnow-Based Approach

As described earlier, our approach is based on the idea that some partial matches can be used to classified new examples. Thus, we extract the partial rules from the original rules and use them collaboratively for classifying examples. The idea of using many partial rules to classify an example is that the partial rules are assigned the importance in form of the weights of each class and all applicable partial rules are combined with their weights for determining the class of the

```
function ExtractPartialRule(rule)
returns a sequence list
  inputs:    rule as original rule
  variables: literals, remained_literals as sequence of literals
             output, combined, partial_rules as sequence list
             literal as literal
  output ← empty list
  literals ← the body of rule
  for each literal, in literals, which consumes the variables in the head of
  rule as input
    remained_literals ← remove literal from literals
    partial_rules ← SearchPartialRule(literal, literal, remained_literals,
                    output)
    output ← add partial_rules to output
  combined ← make all possible combination of sequence of literals in
             output, which have common variables that do not occur
             in the head of the rule
  output ← add combined to output
  remove redundant sequence of literals from output
  return output


function SearchPartialRule(input_literal, partial_rule, literals, output)
returns a sequence list
  inputs:    input_literal as literal
             partial_rule, literals as sequence of literals
             partial_rules as sequence list
  variables: literal as literal
             remained_literals, new_partial_rule as sequence of literals
             unfinish as sequence of literals
             new_partial_rules as sequence list
             found as boolean, initially false
  new_partial_rule ← add input_literal to partial_rule
  if no new argument in input_literal then
    partial_rules ← add new_partial_rule to partial_rules
    output ← add partial_rules to output
    return output
  for each literal, in literals, which consumes the new variables
  in input_literal as input
    unfinish ← add literal to new_partial_rule
    remained_literals ← remove literal from literals
    partial_rules ← SearchPartialRule(literal, unfinish,
                    remained_literals, output)
    output ← add partial_rules to output
    found ← true
  if found then
    partial_rules ← add new_partial_rule to partial_rules
    output ← add partial_rules to output
  return output
```

Fig. 2. Partial Rule Extraction Algorithm

example. When we need to classify an example, we determine the summation of the weight of each class of all partial rules which match with the example. The class which has the highest summation of weights is selected as the class of the example.

The Winnow algorithm [3] is originally proposed as a linear threshold algorithm. For an input vector x, weight vector w, promotion factor $\alpha > 1$, and threshold $\theta > 0$, the algorithm predicts 1 if $w \cdot x \geq \theta$. Intuitively speaking, the Winnow algorithm activates the output if the input x is high enough. If $w \cdot x$ is too low, the weight vector w is increased by updating $w_i \leftarrow \alpha^{x_i} w_i$. On the other hand, if $w \cdot x$ is too high, the weight vector w is decreased by updating $w_i \leftarrow \alpha^{-x_i} w_i$. However, in our approach, the concept of prediction scheme is different. In our class prediction, instead of comparing the summation to the threshold we need only the highest summation of the weight from each class, so that we can make use of the Winnow algorithm by employing the following strategy.

Given a problem with n partial rules, m classes, and promotion factor α. P is a vector of length n, where element p_i of P is a partial rule. W_i is a vector of length m, where element $w_{i,j}$ of W_i is the weight of class j of partial rule p_i. V is a summation vector of length m, where v_i of V is the summation of the weights of class i. The weight vector W_i are updated by using the following procedure.

- Initialize all $w_{i,j} = 1$
- Until termination condition is met, Do
  - For each training example e, Do
    - Initialize all $v_i = 0$ and c as the class of e
    - For all partial rules p_i which match with e, add corresponding W_i to V,

$$V = V + W_i$$

    - Let v_k be the maximum element in V, predict the example e as class k
    - If $c = k$, no update is required; otherwise the weight w_i corresponding to p_i which matches with e is updated by,

$$w_{i,j} = \begin{cases} \alpha w_{i,j} & \text{if } j = k, \\ \alpha^{-1} w_{i,j} & \text{if } j = c. \end{cases}$$

Each partial rule is weighed by a weight vector of which elements are for each class and we classify an example as the class which has the highest summation of the weights of the applicable partial rules. When an example is incorrectly classified, the output class is different from the target class. This means the summation of the weight of the output class of all applicable partial rules is higher than that of the target class. Thus, we decrease the weights of the output class of all applicable partial rules by using Winnow algorithm's weight updating equation, $w_{i,j} = \alpha^{-1} w_{i,j}$ and increase the weights of the target class of all applicable partial rules by using promotion factor, $w_{i,j} = \alpha w_{i,j}$.

To classify an unseen example e, we use the following strategy.

– Initialize all $v_i = 0$
– For all partial rules p_i which match with e, add corresponding W_i to V,

$$V = V + W_i$$

– Let v_k be the maximum element in V, classify the example e as class k

5 Experiments

The data set used in the experiments contained 1366 molecules of dopamine antagonist molecules of 4 classes, D1, D2, D3, and D4. Information of the molecules was originally described in term of the position in three dimension space of atoms, types of atoms, types of bonds, and dopamine antagonist activity of molecules. However, the position in three dimension space was not useful for discriminating examples because a molecule could rotate or move to other positions in the space. Hence, we converted the position of atoms to the relation between atoms and bonds. We instead represented the information of atoms, bonds, and distances between atoms in term of 3 predicates, `atm/6`, `bond/6`, and `link/4`, respectively. The details of these three predicates are described below:

– `atm(A, B, C, D, E, F)` represents that the atom `B` is in molecule `A`, is type `C`, forms a bond with oxygen atom if `D` is 1, otherwise it does not link to any oxygen atom, has distance `E` to the nearest oxygen atom, and has distance `F` to the nearest nitrogen atom.
– `bond(A, B, C, D, E, F)` represents that the bond `B` is in molecule `A`, has atoms `C` and `D` on each end, is type `E`, and has length `F`.
– `link(A, B, C, D)` represents that in the molecule `A`, the distance between atoms `B` and `C` is `D`.

The following is an example of rules obtained from the experiments.

```
molecule(A,d1) :- link(A, B, C, D), bond(A, E, F, C, G, H), D=6.9,
    H=1.4, bond(A, I, J, F, K, H), bond(A, L, M, J, G, H),
    bond(A, N, B, O, G, P).
[Positive cover = 53 Negative cover = 5]

molecule(A,d2) :- atm(A, B, C, D, E, F), C=n, bond(A, G, B, H, I,
    J), gteq(J, 1.5), atm(A, L, M, D, N, O), N=5.1, O=1.5.
[Positive cover = 42 Negative cover = 1]

molecule(A,d3) :- link(A, B, C, D), D=4.1, atm(A, B, E, F, G, H),
    H=4.1, bond(A, I, B, J, K, L), bond(A, M, C, N, K, L).
[Positive cover = 56 Negative cover = 1]

molecule(A,d4) :- link(A, B, C, D), D=4.4, bond(A, E, C, F, G, H),
    bond(A, I, F, J, G, H), bond(A, K, L, C, G, H), bond(A, M, J,
    N, G, H).
[Positive cover = 130 Negative cover = 8]
```

We compared our approach with other two approaches, i.e. Majority Class [9, 10] and Decision Tree Learning. As described in Section 2, ILP's rules alone cannot classify the examples which match with multiple rules from different classes or do not match with any rule, so that we make the rules be fairly compared to other methods by using the Majority Class in such cases.

In the Majority Class method, we selected the class which had the maximum number of examples in training set as the default class. An example which matched with only rule(s) from one class was classified as that class, while an example which could not match with any rule was classified as the default class. In case of the examples which matched with the rules from two or more classes, we selected the class of which the matched rules covered maximum number of examples.

Another method compared in our experiment is the Decision Tree Learning (DTL) algorithm. DTL is a well-known propositional Machine Learning technique which employs the Information Theory to guide in searching for the best theories. The decision tree learner used in our experiments is C4.5 system [11]. In our experiments, the truth values obtained by comparing the partial rules with examples could be considered as the attributes of examples. From this point of view, we could apply C4.5 which is a propositional learner to this domain by using the truth value of each partial rule as the attribute of examples.

We ran 10-fold cross validation experiment using three methods, the original ILP system with the Majority Class method (ILP+Majority Class), Partial Rules and DTL (PR+DTL), and our approach, Partial Rules and Winnow algorithm (PR+Winnow).

The accuracy shown in Table 1 was separately evaluated when the rules were used as in two-class fashion. The covered examples were classified as positive, while the uncovered examples were classified as negative. The accuracy of each class was obtained from the test set consisting of only the examples from the test set of that class. The accuracy in Table 1 shows the accuracy of the rules from each class. The average accuracy of all classes is 76.42%. Furthermore, this percentage of accuracy also shows the coverage ratio on the test set.

Table 2 shows the accuracy of each approach in classifying test examples in multi-class fashion. The accuracy of ILP+Majority Class approach is 79.21%. This shows that the only the Majority Class method can slightly improve the accuracy of the original rules. The accuracy of PR+DTL is 85.72%, higher than ILP+Majority with 99.5% confidence level using the standard paired t-test method. The accuracy of PR+Winnow is 88.58%, higher than ILP+Majority and PR+DTL methods with 99.5% and 99.0% confidence level respectively using the same comparing method.

An example of some partial rules which are highly weighed is shown below.

```
molecule(A) :- atm(A, E, F, G, H, I), bond(A, N, E, O, P, M),
               atm(A, O, F, G, Q, R), H=2.4.
               [0.1999, 33.5451, 0.2812, 0.5303]
[The original rule is
molecule(A) :- link(A, B, C, D), atm(A, E, F, G, H, I), D=5.6,
```

Table 1. The accuracy of the output rules used to classify only the positive examples of each class

Class	Accuracy (%)
D1	77.42
D2	70.30
D3	74.80
D4	83.16
Average	76.42

Table 2. The accuracy of the compared methods.

Method	Accuracy (%)
ILP+Majority Class	79.11±4.37
PR+DTL	85.71±3.41
PR+Winnow	88.65±3.85

```
               H=2.4, gteq(I, 3.8), bond(A, J, B, K, L, M),
               bond(A, N, E, O, P, M), atm(A, O, F, G, Q, R),
               lteq(Q, 2.9), lteq(M, 1.4), bond(A, S, C, T, P, U).]

molecule(A) :- atm(A, B, C, D, E, F), bond(A, G, B, H, I, J),
               atm(A, H, K, D, L, M), L=6.5.
               [0.9524, 0.1566, 35.2224, 0.1904]
[The original rule is
molecule(A) :- atm(A, B, C, D, E, F), bond(A, G, B, H, I, J),
               atm(A, H, K, D, L, M), link(A, H, N, O),
               atm(A, N, K, D, L, M), atm(A, H, K, D, L, M),
               L=6.5, F=1.3.]
```

The weights in the above example show another advantage of our approach. We can see that when an example matches with these highly weighed partial rules, the example has the high probability of being classified as the class whose weighted is very high. This provides us some knowledge which can be discovered from the dataset, different from the original rules which sometimes are too specific and not useful. Our approach can seek for some pieces of knowledge which are more important than the others in the original rule. For example, the second partial rule in the above example shows that if an unseen example matches with this partial rule, that example has the high probability of being classified as class 3 which is very highly weighed.

6 Conclusion

We have proposed an approach that can improve the accuracy of the ILP's rules, especially in case of the unseen examples which match with no rule or multiple

rules from two or more classes. Our method is based on the idea that the unequal important partial rule matching with an example can be useful for classifying the example. The partial rules are extracted from the original rules and are assigned the importance in term of the weights obtained from Winnow-based approach. The experimental results on classifying the activity of the dopamine antagonist molecules show that our approach was successfully applied to such domain by yielding 88.58% accuracy. Furthermore, the accuracy obtained from the experiments also shows that using only the matching of the partial rules and an attribute learner C4.5 improved the accuracy over using the original rules with the majority class method, and the accuracy was further much improved when using the proposed method. Furthermore, the weights of the partial rules also show some pieces of knowledge which are previously hidden in the original rules.

References

1. Enot, D.P., King, R.D.: Application of Inductive Logic Programming to Structure-based Drug Design. In Lavrac, N., Gamberger, D., Todorovski, L., Blockeel, H., eds.: Proc. 7th European Conf. On Principles and Practice of Knowledge Discovery in Databases., Springer-Verlag (2003)
2. Quiniou, R., Cordier, M.O., Carrault, G., Wang, F.: Application of ILP to cardiac arrhythmia characterization for chronicle recognition. Lecture Notes in Computer Science **2157** (2001) 220–227
3. Littlestone, N.: Learning quickly when irrelevant attributes abound: A new linearthreshold algorithm. Machine Learning **2** (1988) 285–318
4. Fürnkranz, J.: Round robin rule learning. In Brodley, C.E., Danyluk, A.P., eds.: Proceedings of the 18th International Conference on Machine Learning (ICML-01), Williamstown, MA, Morgan Kaufmann Publishers (2001) 146–153
5. Allwein, E.L., Schapire, R.E., Singer, Y.: Reducing multiclass to binary: A unifying approach for margin classifiers. In: Proc. 17th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2000) 9–16
6. Srinivasan, A.: The Aleph Manual (2001)
7. Selman, B., Levesque, H.J., Mitchell, D.: A New Method for Solving Hard Satisfiability Problems. In: Proc. 10th National Conference on Artificial Intelligence, AAAI Press (1992) 440–446
8. Kijsirikul, B., Sinthupinyo, S., Chongkasemwongse, K.: Approximate match of rules using backpropagation neural networks. Machine Learning **44** (2001) 273–299
9. Clark, P., Boswell, R.: Rule induction with CN2: Some recent improvements. In: Proc. Fifth European Working Session on Learning, Berlin, Springer (1991) 151–163
10. Laer, W.V., Raedt, L.D., Dzeroski, S.: On multi-class problems and discretization in inductive logic programming. In: International Syposium on Methodologies for Intelligent Systems. (1997) 277–286
11. Quinlan, J.: C4.5: Programs for machine learning. Morgan Kaufmann Publishers (1993)

Identification of Activity Classes of Drugs under Existing Noise Compounds by ANN and SVM

Yoshimasa Takahashi[1], Satoshi Fujishima[1], Katsumi Nishikoori[1], Hiroaki Kato[1], and Takashi Okada[2]

[1] Department of Knowledge-based Information Engineering, Toyohashi University of Technology, 1-1Hibarigaoka, Tempaku-cho, Toyohashi 441-8580 Japan
[2] Department of Informatics, School of Science and Technology, Kwansei Gakuin University, 2-1 Gakuen Sanda 669-1337, Japan
{taka, fujisima, katsumi}@mis.tutkie.tut.ac.jp
hiro@cilab.tutkie.tut.ac.jp, okada-office@ksc.kwansei.ac.jp

Abstract. This paper describes classification and prediction for pharmacologically active classes of drugs under the presence of noise chemical compounds. Dopamine D1 receptor agonists, antagonists and other drugs were used for the work. Each drug molecule was characterized with Topological Fragment Spectra (TFS) reported by the authors. TFS-based artificial neural network (TFS/ANN) and support vector machine (TFS/SVM) were employed and evaluated for their classification abilities. It was concluded that the TFS/SVM works better than TFS/ANN in both of the training and the prediction.

1 Introduction

“Similarity" is very important concept in solving problems in science. This is true in chemistry. Especially structural similarity provides us a lot of information on structure-activity and structure-property problems [1,2]. And it is still under active development in the area of drug design, for the selection of candidate analogs as new chemicals and for the estimation of molecular properties [3-5]. The basic idea behind them is that structurally similar compounds are likely to possess similar molecular properties and similar biological activities. Most of the approaches for the evaluation are based on finding particular functional atoms or atomic groups defined in advance. However, the result of such a structural similarity analysis depends on the set of substructures defined as descriptors [6]. In the former work, the authors proposed Topological Fragment Spectral (TFS) method as a tool for the description of the topological structure profile of a molecule [7]. The TFS representation method doesn’t require any kind of a priori substructure definition like a dictionary file of substructures to be searched. The method is also useful for the similar structure searching on chemical structure databases [8] and visualization of similar structure data space [9].

The aim of our current research project is to establish a basis of computer-aided risk assessment for chemicals on the basis of chemical similarity analysis and machine-learning techniques. In our preceding works [10], we reported that an artificial

neural network (ANN) approach combined with the TFS as input signals to ANN allowed us to successfully classify the type of activities for dopamine receptor antagonists that interact with four different types of dopamine receptors, and it could be applied to the prediction of active class of unknown compounds. It was also shown that support vector machine (SVM) works for this problem much better [11]. Those were the results obtained with a set of chemicals that belong to any of typical activity classes without noise compounds. However, in practical, for risk estimation of drugs such as a side effect, we have to treat a lot of chemicals that belong to particular active classes and much more chemicals that never belong to any of them of our interest. In the present work, we investigated to identify pharmacological activity of drugs under the condition with many of inactive compounds that are regarded as noise data.

2 Data Set and Methods

2.1 Data Sets

In this work we employed 232 drugs that interact with dopamine D1 receptor. They were taken from MDDR [12] which is a structure database of investigative new drugs. Sixty-three of them are the agonists and 169 are the antagonists. In addition, 696 compounds were randomly chosen from the MDDR database excepting the dopamine D1 receptor actives. They were used as noise data for this work. Three data sets that have different sizes of noise data (50% for trial set 1, 100% for trial set 2, and 300% for trial set 3 in noise rate against 232 drugs of the interests) were prepared for the following analyses. Each trial set was divided into two groups; a training set and a prediction set. The former includes 90% of the data and 10% for the latter.

2.2 Numerical Representation of Structural Features of Chemicals

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.

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. An illustrative scheme of TFS creation from a chemical structure is given in Fig.

1. Another example of TFS of promazine characterized according to the present method is shown in Fig. 2. The TFS can be regarded as a function of chemical structure, i.e. *TFS= f(chemical structure)*.

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. 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).

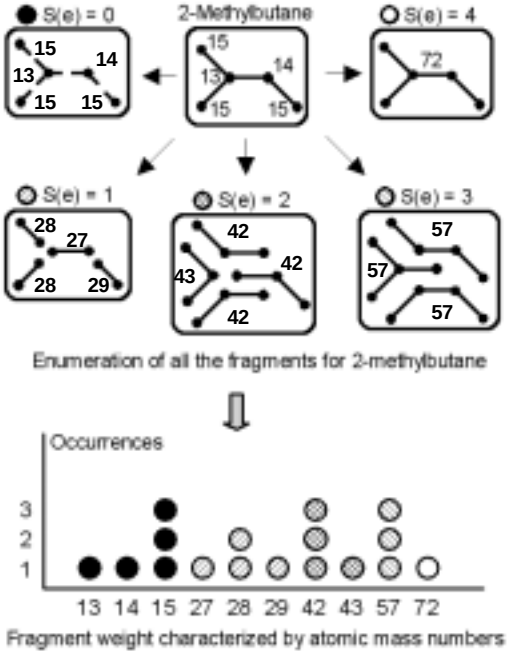


Fig. 1. A schematic flow of TFS creation. S(e) is the number of edges (bonds) of fragments to be generated.

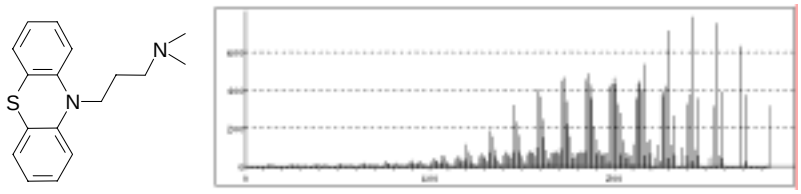


Fig. 2. TFS of promazine characterized by the sum of atomic mass numbers for each fragment.

Obviously, the fragment spectrum obtained by these methods can be described as a kind of multidimensional pattern vector. However, the number of dimensions of the TFS pattern description vector depends on individual chemical structures. The different dimensionalities of the spectra to be compared are adjusted as follows,

If $X_i = (x_{i1}, x_{i2}, ..., x_{iq})$ and $X_j = (x_{j1}, x_{j2}, ..., x_{jq}, x_{j(q+1)}, ..., x_{jp})$ (q<p), then X_i is redefined as $X_i = (x_{i1}, x_{i2}, ..., x_{iq}, x_{i(q+1)}, ..., x_{ip})$ (**1**)

$$\text{here, } x_{i(q+1)} = x_{i(q+2)} = ... = x_{ip} = 0$$

Where, x_{ik} is the intensity value of peak *k* of TFS for *i*-th molecule, and x_{jk} is that of peak *k* of TFS for the *j*-th molecule that have the highest value of the characterization index (in this work, the highest fragment mass number).

According to this manner, TFS for every trial set were generated. The dimensions of the TFS for trial set 1, trial set 2 and trial set 3 were 168, 166, and 186 respectively. For the prediction, each TFS is adjusted by padding with 0 or by cutting the higher mass region off to have the same dimensionality as that of the training set when a prediction sample is submitted.

2.3 Neural Network

Discrimination of pharmacological activity classes of chemicals was investigated using artificial neural network (ANN). Three-layer learning network with a complete connection among layers was used. The TFS was submitted to the ANN as input signals for the input neurons. Training of the ANN was carried out by error back propagation method. All the neural network analyses were carried out using a computer program, NNQSAR, developed by the authors [13]. For the present work, the number of neurons in the hidden layer of the TFS/ANN model was set two to avoid explicit over-fitting because of a large number of input neurons for accepting a TFS pattern.

2.4 Support Vector Machine (SVM)

The SVM implements the following basic idea: it maps the input vectors **x** into a higher dimensional feature space **z** through some nonlinear mapping, chosen a priori. In this space, an optimal discriminant surface with maximum margin is constructed. 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.

$$f(\mathbf{x}_i) = (\mathbf{w} \cdot \mathbf{x}_i) + b \quad (\mathbf{2})$$

Where $\mathbf{w}$ is a weight vector, b is a bias. The discriminant surface can be represented as $f(\mathbf{x}_i)=0$. The maximum margin can be obtained by minimizing square of the norm of weight vector $\mathbf{w}$,

$$\|\mathbf{w}\|^2 = \mathbf{w}\cdot\mathbf{w} = \sum_{l=1}^{d} w_l^2 \tag{3}$$

with the constraints $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \quad (i = 1,...,n)$

The decision function is described as $S(\mathbf{x}) = sign(\mathbf{w}\cdot\mathbf{x}+b)$ for classification, where *sign* is a sign function that returns 1 for positive value and -1 for negative value. This basic idea can be extended to a linearly inseparable case by introducing slack variables ξ_i and minimizing the following quantity,

$$\frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_{i=1}^{n}\xi_i \tag{4}$$

with the constraints $y_i = (\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1-\xi_i$ and $\xi_i \geq 0$.

This optimization problem reduces to the previous one for separable data when constant *C* is large enough. This quadratic optimization problem with constraints can be reformulated by introducing 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 [14]. 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 TFS-based SVM analyses were carried out using a computer program developed by the authors according to Platt's algorithm [15]. In the present work, C=100 and σ=40

were used for the training. The values of these parameters were determined by trial and error to give the best prediction.

3 Results and Discussion

3.1 Classification and Prediction by TFS/ANN

The classification and prediction abilities of the TFS/ANN were investigated for dopamine D1 receptor agonists (63 compounds), antagonists (169 compounds) and noise compounds. Three data sets that have different sizes of noise data (50%, 100% and 300% in noise rate) were used for the present computational experiments. Three neurons were set for the output layer of the TFS/ANN. Every dataset was divided into two groups, 90% of the data for a training set and 10% of them for a prediction set. Every trial was carried out with a single set that consists of the training set and the prediction set. The results for the trials are summarized in Table 1.

Table 1. Results of the training by TFS/ANN

%Noise	Training (%)			
	ALL	Agonists	Antagonists	Noise
50%	**279 / 313 (89.1)**	39 / 56 (69.6)	147 / 152 (96.7)	93 / 105 (88.6)
100%	**374 / 417 (89.7)**	38 / 56 (67.9)	137 / 152 (90.1)	199 / 209 (95.2)
300%	**762 / 835 (91.3)**	27 / 56 (48.2)	126 / 152 (82.9)	609 / 627 (97.1)

Table 2. Results of the prediction by TFS/ANN

%Noise	Prediction (%)			
	ALL	Agonists	Antagonists	Noise
50%	**30 / 35 (85.7)**	5 / 7 (71.4)	16 / 17 (94.1)	9 / 11 (81.1)
100%	**40 / 47 (85.1)**	5 / 7 (71.4)	15 / 17 (88.2)	20 / 23 (87.0)
300%	**81 / 93 (87.1)**	2 / 7 (28.6)	11 / 17 (64.7)	68 / 69 (98.6)

The TFS/ANN models classified 89.1%-91.3% of the drugs into their own classes correctly. However, the details of the results show that the recognition rate for individual class differs from each other. This matter is typical for the data set with 300%

noise. It is considered that the larger number of samples for each class give us the better recognition rate in the training. The matter is true in the prediction by the models obtained. The prediction results are summarized in Table 2. Both Table 1 and Table 2 show that the results for agonists are poorer than those for other classes in both cases of training and prediction. It is considered that the TFS/ANN model couldn't learn very much for the training set because the number of samples is relatively smaller than those of the other sets. The present results suggest that the training results with artificial neural network considerably depend on the sample size in each class.

3.2 Classification and Prediction by TFS/SVM

Next, we investigated classification and prediction abilities of the TFS/SVM for the same data sets used in the previous section. The results for these three training sets are shown in Table 3. The TFS/SVM models classified 99.4%-99.9% of the drugs into their own classes correctly in total. Then, the details of the classifications show that the recognition rate for individual class is highly good for every class regardless of the sample size of individual classes. The matter is still true even for the data set with 300% noise.

Table 3. Results of the training by TFS/SVM

%Noise	Training (%)			
	ALL	Agonists	Antagonists	Noise
50%	**311 / 313 (99.4)**	55 / 56 (98.2)	151 / 152 (99.3)	105 / 105 (100)
100%	**416 / 417 (99.8)**	56 / 56 (100)	151 / 152 (99.3)	209 / 209 (100)
300%	**834 / 835 (99.9)**	56 / 56 (100)	151 / 152 (99.3)	627 / 627 (100)

Table 4. Results of the prediction by TFS/SVM

%Noise	Prediction (%)			
	ALL	Agonists	Antagonists	Noise
50%	**32 / 35 (91.4)**	7 / 7 (100)	16 / 17 (94.1)	9 / 11 (81.1)
100%	**44 / 47 (93.6)**	7 / 7 (100)	15 / 17 (88.2)	22 / 23 (95.7)
300%	**91 / 93 (97.8)**	7 / 7 (100)	15 / 17 (88.2)	69 / 69 (100)

Then, the TFS/SVM models employed in the prediction for the prediction set. The results are summarized in Table 4. These results show that the TFS/SVM works better in the prediction too. The total prediction rates for the data sets with 50% noise, 100% noise and 300% noise are 91.4%, 93.6% and 97.8 % respectively. The results for individual classes also are good and stable for all the classes.

It is concluded that the TFS/SVM works better in the training and it would be stable for the prediction even in the case with diverse size of samples for classes to be analyzed.

4 Conclusions

Classification and prediction for pharmacologically active classes of drugs under the presence of noise chemical compounds were investigated with TFS-based artificial neural network (TFS/ANN) and TFS-based support vector machine (TFS/SVM). The results suggest that the training with TFS/ANN considerably depends on the sample size in each class. Thus the prediction ability tends to be less for the activity class that has smaller size of samples than others. On the other hand, the TFS/SVM works better than TFS/ANN in both of the training and the prediction. However, because many instances are required for predictive risk assessment and risk report, more large number of pharmacological activity classes should be treated in further works.

Acknowledgement

This work was supported by Grant-In-Aid for Scientific Research on Priority Areas (B) 13131210.

References

1. M. A. Johnson, G. M. Maggiora, (Eds), *"Concepts and Applications of Molecular Similarity"*, Wiley, New York, 1990.
2. Y. Takahashi, Identification of Structural Similarity of Organic Molecules. *Topics in Current Chemistry*, **174** (1995) 105-133.
3. M. Rarey, M. Stahl, Similarity Searching in Large Combinatorial Chemistry Spaces, *J. Comput. -Aied Mol. Des.*, **15** (2001) 497-520.
4. J. W. Raymond, P. Willett, Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases, *J. Comput. -Aied Mol. Des.*, **16** (2002) 59-71.
5. D. Wilton, P. Willett, Comparison of Ranking Methods for Virtual Screening in Discovery Program, *J. Chem. Inf. Comput. Sci.*, **43** (2003) 469-474.
6. Flower, D. On the Properties of Bit String-Based measures of Chemical Similarity, *J. Chem. Inf. Comput. Sci.*, **38** (1998) 379-386.

7. 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
8. Y. Takahashi, S. Fujishima, H. Kato, Chemical Data Mining Based on Structural Similarity, *J. Comput. Chem. Jpn.*, **2** (2003) 119-126
9. Y. Takahashi, M. Konji, S. Fujishima, MolSpace: A Computer Desktop Tool for Visualization of Massive Molecular Data, *J. Mol. Graph. Model.*, **21** (2003) 333-339
10. S. Fujishima, Y. Takahashi, Classification of Pharmacological Activity of Drugs using TFS-Based Artificial Neural Network, *J. Chem. Inf. Comput. Sci.*, in press.
11. Y. Takahashi, K. Nishikoori, S. Fujishima: Classification of Pharmacological Activity of Drugs Using Support Vector Machine, *Second International Workshop on Active Mining*, (2003) 152-158
12. MDL Drug Data Report, MDL, Ver. 2001.1, (2001).
13. H. Ando and Y. Takahashi, Artificial Neural Network Tool (NNQSAR) for Structure-Activity Studies, *Proceedings of the 24th Symposium on Chemical Information Sciences* (2000) 117-118.
14. S. W. Lee and A. Verri, Eds, Support Vector Machines 2002, LNCS 2388, (2002).
15. 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 of Three-Dimensional Structural Fragments in Drug Molecules

Hiroaki Kato, Takashi Koshika, 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, kosika}@cilab.tutkie.tut.ac.jp
taka@mis.tutkie.tut.ac.jp, abe@cilab.tutkie.tut.ac.jp

Abstract. It is well known that the structural formula of an organic molecule has rich information related to various physicochemical properties and biological activities of it. In the preceding works, we developed a computer program, named COMPASS, for automated identification of 3D maximal common substructures among molecules by a graph theoretical approach. In the present work, we have developed a software tool for exhaustive seeking of common 3D structural fragments among molecules based on the COMPASS algorithm. It can enumerate all the fragments which contain more than the specified number of atoms, and search for those fragments which appear at specified frequency or more in the given set of molecules. The search experiment was carried out with a data set of the dopamine D1 agonist molecules extracted from MDDR-3D. The present tool successfully found several fragments that are seemed to be characteristic for a particular activity class.

1 Introduction

An understanding of the structural features of drug molecules is necessary for solving many problems in chemistry. In particular, a substructural analysis or a functional group analysis is essential for structure-activity (or property) studies and rational molecular design based on them. Especially, it is well known that molecular properties, including biological functions, relay not only on atom connectivity or the topological level but also on the three-dimensional (3D) geometrical arrangement of the atoms [1]. Such a process involving a structural feature analysis could be done manually for a small set of molecules. However, the work is quite tedious and time consuming for a large set of molecules, even if it is handled in a topological level. For this reason, computerized methods are required for a systematic mining of the 3D features of molecules in such a database. For knowledge discovery based on 3D structural feature analysis of organic molecules, we have developed a computer program, named COMPASS (COMmon geometric PAttern Search System) [2]. Here, each molecule is treated as a set of points that correspond to its constituent atoms in the 3D space. The set of points is described by a matrix representation, of which each element involves the inter-atomic distance within the molecule. Thus, we can repre-

sent the structural information of a molecule, including its 3D geometry, with a weighted graph of which the nodes and edges correspond to atoms and the interatomic distances between them, respectively. On the basis of this, a maximal common subgraph matching algorithm can be used for the searching of the geometric patterns which are common in the molecules [3]. However, the COMPASS was designed to search for only the fragments such as "the greatest common divisor". Thus, it is difficult to obtain meaningful results when some of the molecules to be analyzed have quite different chemical structures from the others [4]. On the other hand, only the maximal common substructural feature(s) are not always important for the structure-activity problem. To the problem too, our previous system could not find any other smaller fragments excepting the maximal common subgraph(s) for the searching.

In this paper, we have developed a fragment search system for more flexible structural feature analysis based on COMPASS algorithm. The system allows us to enumerate all the fragments which contain more than the specified number of atoms, and search for the fragments which appear with a specified value of the frequency or more for the given set of molecules (Fig. 1).

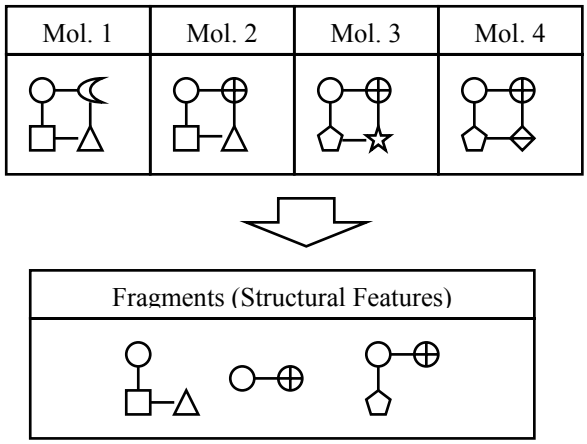


Fig. 1. Basic concept of the fragments search in the present work

There is a variety of different techniques in data mining included inductive logic programming or inductive databases. For example, De Raedt et al. proposed the level-wise version space algorithm that forms the basis of the inductive query and database system MolFea (Molecular Feature Miner), but it was restricted to linear molecular fragments as patterns [5]. AGM (Apriori-based Graph Mining) approach was proposed by Inokuchi et al. to overcome there limitations [6]. They also applied to the 3D graph structured data using multiple labeled edges [7]. However, their approach was depended on the information of the original edge labels (i.e. bond types in a molecular structure) and the virtual links (path lengths, or topological distances between the non-adjacency atoms) to mine the frequent 3D subgraphs. In the present work, only the 3D geometric pattern of atoms can be used for the primary information on mining, although atomic type, bond type, and/or other additional environment

information of molecules are also available for the filtering the search results. Some reduced representation of 3D molecular structure will be also introduced for more efficient analysis.

2 Methods

2.1 User Defined Parameters for Fragment Searching

The user can specify the searching conditions [8] as the following:

(1) The threshold value for the frequency of appearance (or minimum support [6]); The fragment patterns with the specified value or more for the given set of molecules are considered as the candidates for the larger fragments. The input value is specified in percentage (%).
(2) The minimum number of atoms for the fragments to be explored.
(3) The allowance at testing the equivalency for the inter-atomic distances; The inter atomic distances are regarded as equivalent when the difference is smaller than the allowance. The input value is specified in Å.

In addition to these conditions, some other optional conditions to be matched can be specified to use more detailed information such as atomic type, atomic charge, and/or enantiomer geometries.

2.2 Basic Algorithm for Finding Frequent Fragmental Patterns

At first, all distinct pairs of atoms (size-2 fragments) of the first molecule in the data set are generated with their inter-atomic distances. Exhaustive 3D substructure search for the data set are performed using each fragment in the above-generated set of size-2 fragments as a query, and those fragments which satisfy user-defined search conditions are only survived. For reserving the information whether a size-2 fragment satisfies the predefined conditions, a candidate matrix whose elements correspond to the fragments is prepared. When the fragment consisted of i-th and j-th atoms of the molecule satisfies the conditions, the element c(i, j) of the matrix sets to 1, otherwise sets to 0. It is obvious that a larger fragment made by extending the size-2 fragment which does not satisfied the conditions never satisfies them, too. Therefore, this matrix can be used for pruning the unnecessary search procedures.

Next, the fragments of size N+1 are generated from a size-N fragment with the candidate matrix for the reference molecule. For each fragment generated, a strict check is made using the result for size-N fragments. Then 3D substructure search is done to test the conditions for survival. For efficient exploration, the database searching is terminated when it is revealed that the fragment in question will or will not be satisfied the given conditions. The procedures are continued until the extension of the fragment is no longer possible.

Finally, all information about fragments of which size is equal or more the user-defined threshold value is put into the fragment dictionary. To avoid redundancy, those fragments which became origins of larger fragments are omitted. Then, all entries of the dictionary are sorted according to their frequencies in the data set. When two or more fragments are identical within the predefined threshold value, the fragments with lower frequency is omitted from the dictionary. If they have same frequency, the fragment originated from the parent molecule with smallest registry number is selected. Furthermore, if the parent molecule is same, i.e., the plural similar fragments are generated from the identical molecule, the fragment which consisted of the smaller indexes of atoms (in the dictionary order) is also prioritized. For example, in Fig. 2, the first fragment (2.5 Å) which matched in three molecules is selected to the representative geometric pattern of them. In addition to this, the fourth fragment (3.3 Å) is also registered in the fragment dictionary.

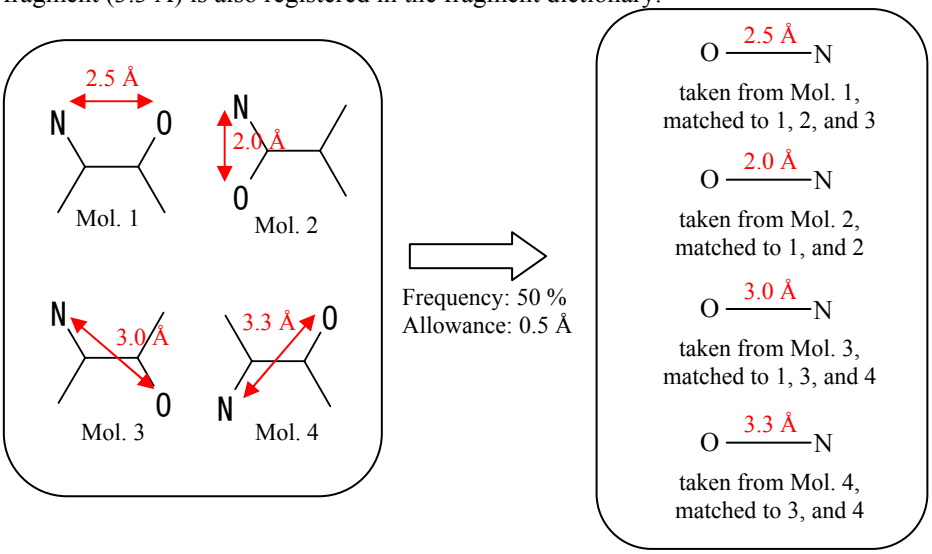


Fig. 2. Illustrative examples of an extraction and refinement for the similar fragments

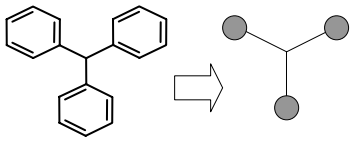
The algorithm mentioned above has been implemented into the program, FragSearch, by Java language, and all of computational work in the current paper were carried out on a PC (CPU: Pentium4 2.4GHz, main memory: 512MB) with Sun Java2 SDK, SE v1.4.0_01 on Windows XP.

2.3 Reduced Representation of Molecular Structure

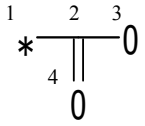
It is expected that a very large computational time will be required for finding the frequent fragmental patterns among large molecules and a large number of molecules. To decrease the computational time for the case, a reduced representation of molecular structure is also investigated here. In this representation method, a particular substructure such as a functional group or a ring is regarded as a pseudo-atom (superatom). Substructures described with their reduced representations can be defined by

the user arbitrarily. Furthermore, the user can specify a particular atom (or atomic group) to suppress trivial atom(s) in molecular structures. Using this utility, for example, the user can treat a molecular structure as only benzene-rings and hetero atoms excepting carbon hydrogen atoms.

The program for the super-atom transformation for a molecule was implemented on the basis of a substructure search technique in the ordinary chemical graph representation. The user can specify a set of atoms to reduce and the coordinate information for the super-atom. For example, in the case of benzene-ring (Fig. 3 (a)), the reduced atoms are six carbon atoms and a representative coordinate of these atoms is set at the center of the ring. On the other hand, in the case of carboxyl group, the reduced atoms are No. 2, 3, and 4 in Fig.3 (b), and the representative coordinate is usually approximated to that of the 2nd carbon atom. It is necessary to notice that these preprocesses cause some problem in the following 3D mining because of the user's bias to the data set. We believe that these representations allow us not only to reduce the searching space, but also to obtain more significant search results in chemical sense.



(a) An example of reduced representation of benzene rings



(b) A query for carboxyl group

Fig. 3. Illustrative examples of the reduced representation for molecular structures

3 Results and Discussion

3.1 3D Fragment Searching on Dopamine D1 Receptor Agonists

We have prepared a target data set that contains 66 molecular structures taken from MDDR-3D (MDL Drug Report) Ver.2001.1 [9]. These molecules have the same biological activity, the dopamine D1 agonist activity. The hydrogen atoms were omitted, and the benzene-ring reduced representation above mentioned was employed. We assumed that every 3D structure of molecules used in the present analysis is rigid. Search trials were carried out under the conditions that the distance allowance is 0.5Å, the different types of elements are distinguished, and the value of frequency is 100, 50, and 30%. The results are summarized in Table 1. Some graphical views of the extracted fragments are shown in Fig. 4 with their parent molecular structures. Obviously, lower value of frequency results in a larger and more characteristic fragments. The benzene-ring reduced representation allows us the efficient mining, and it gave us various structural features for this case.

Table 1. The result of the 3D fragment search for 66 drug molecules

Frequency [%] (Comput. Time)	Extracted Fragments	
	Size	Number
100 (13 sec.)	3	1
	2	6
	--	--
50 (8 min.)	7	1
	6	39
	5	219
30 (24 min.)	10	1
	9	2
	8	20

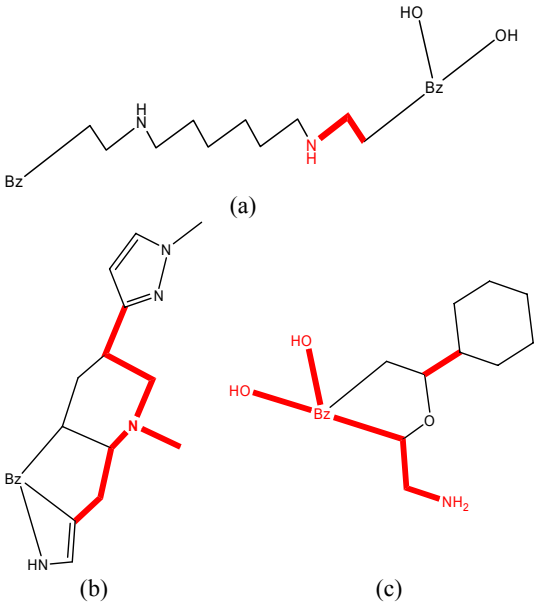


Fig. 4. Graphical views of three different 3D fragments and their parent molecular structures. The label "Bz" means the benzene-ring in the reduced representation. The fragment (a) was extracted with the value of frequency in 100%, (b) and (c) were in 30%, respectively

3.2 3D Structural Feature Mining Using the Fragment Dictionary

Alternatively, in the previous work, the authors reported a computer program for 3D substructure searching, which allows us to identify all occurrences of a user-defined 3D query pattern [3]. More extensive analyses of 3D structural features of drug molecules can be also executed by using our program with the fragment patterns registered in the 3D fragment dictionary (Fig. 5). In the present work, we have pre-

pared other six data sets that consist of dopamine agonists (for D2 receptor or autoreceptor), or antagonists (D1, D2, D3, or D4 receptors). For example, 3D pattern searching was carried out for the representative fragment shown in Fig.4 using these data sets. The number of molecules for each data set and the percentage of hit molecules are summarized in Table 2. Here, the fragment (a) is very small and very common. On the other hand, it seems that (b) is specific to the dopamine agonists, and (c) is closely related to D1 agonists, respectively. The total computational time was about 20 seconds. These results suggest that the present approach is quite useful for 3D structural data mining for drug molecules.

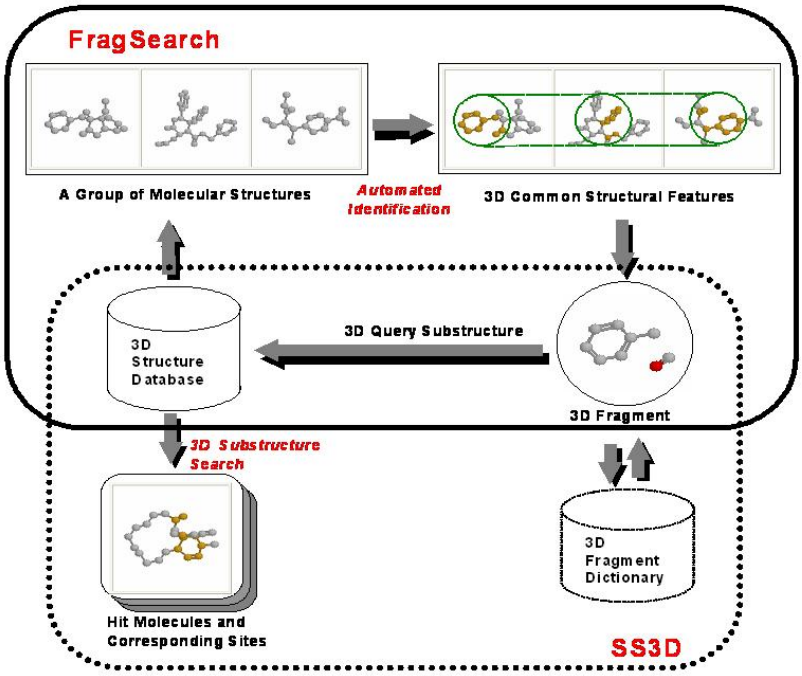


Fig. 5. 3D structural feature mining for drug molecules using the present approaches

Table 2. The results of 3D substructure searching for the 3D fragment patterns obtained in dopamine D1 receptor agonists. The hit molecules for each data set are indicated in percentage. The labels (a), (b), and (c) are corresponded to the fragments shown in Fig. 4. The number in parentheses shows the number of molecules in the data set.

Fragment	Dopamine Agonist			Dopamine Antagonist			
	D1	D2	Auto	D1	D2	D3	D4
	(66)	(143)	(191)	(173)	(430)	(254)	(574)
(a)	100	100	100	100	99	98	100
(b)	33	45	10	1	4	9	0
(c)	30	0	0	0	0	0	0

4 Conclusion

A computer program used for 3D structural fragments search, FragSearch, has been developed. This program can identify all of the fragment patterns that appear with the frequency of occurrences specified for the given set of molecules. The search experiment was carried out using a data set of the dopamine D1 agonist molecules. Our tool successfully found several fragments that are seemed to be characteristic for a particular activity class.

The comprehensive analysis of dopamine agonists and antagonists are now under progress using the tools described here. We believe that the 3D fragment information help us to understand structure-activity relationships of drug molecules. The graphical user interface for these tools will be also required in future works.

Acknowledgement

The authors wish to thank Prof. Takashi Okada and Dr. Masumi Yamakawa of Kwansei Gakuin University for their valuable comments. 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. Willett, P.: Chemical similarity searching, *J. Chem. Inf. Comput. Sci.*, **38** (1998) 983-996
2. Takahashi, Y. et al.: Automated recognition of common geometrical patterns among a variety of three-dimensional molecular structures, *Anal. Chim. Acta*, **200** (1987) 363-377
3. Kato, H., Takahashi, Y.: Development of a three-dimensional substructure search program for organic molecules, *Bull. Chem. Soc. Jpn.*, **70** (1997) 123-127
4. Kato, H. et al.,: Data mining based on 3D structural similarity of drug molecules, *IEICE Tech. Rep.*, **102** (2003) AI2002-88
5. De Raedt, L., Kramer, S.: The level-wise version space algorithm and its application to molecular fragment finding, *Proc. 17th International Joint Conference on Artificial Intelligence*, (2001) 853-862
6. Inokuchi, A. et al.: Applying the apriori-based graph mining method to mutagenesis data analysis, *J. Comput-Aided Chem.*, **2** (2001) 87-92
7. Nishimura, K. et al.: Fast apriori-based graph mining algorithm and application to 3-dimensional structure analysis, *Trans. Jpn. Soc. Artificial Intelligence*, **18** (2003) 257-268
8. Koshika, T. et al.: Development of a three-dimensional fragment search system for SAR, *26th Symp. Chem. Inf. Comput. Sci.*, (2003) 109-110
9. MDL Information Systems, Inc., MDL Drug Data Report -3D, Ver. 2001.1 (2001)