Refining Diagnostic Knowledge extracted for Interferon Therapy by Graph-Based Induction

Tetsuya Yoshida
Meme Media Laboratory, Hokkaido University
N-13, W-8, Sapporo, Japan
yoshida@meme.hokudai.ac.jp

Akira Mogi
ISIR, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan
mogi@ar.sanken.osaka-u.ac.jp

Kouzou Ohara
ISIR, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan
ohara@ar.sanken.osaka-u.ac.jp

Hiroshi Motoda
ISIR, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan
motoda@ar.sanken.osaka-u.ac.jp

Takashi Washio
ISIR, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan
washio@ar.sanken.osaka-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) within the framework of greedy search. GBI has been extended to 1) Beam-wise GBI (B-GBI) by incorporating a beam search to improve its search capability, and 2) Decision Tree Graph-Based Induction (DT-GBI) to construct a decision tree for graph-structured data. We applied 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 was used as class label and measurement patterns that were strongly correlated with the response were extracted. Descriptive patterns were extracted by B-GBI and discriminative ones by DT-GBI using only the time sequence data of blood inspection and urinalysis in order to extract knowledge (typical patterns) from the dataset. The discriminative patterns extracted by DT-GBI tend to be included in only relatively small number of patients and thus too specific. Thus, we tried to extract patterns which are both discriminative and descriptive by B-GBI. Furthermore, since there are the exceptional situations (patients) with the extracted patterns, these patterns are further utilized to extract refined knowledge from the dataset. The preliminary results are reported in this paper with some of extracted patterns.

1. Introduction

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

We have applied Graph-Based Induction (GBI) for the analysis of hepatitis dataset provided by Chiba University Hospital [2, 9]. 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 has been extended to Beam-wise GBI (B-GBI) to improve its search capability without imposing much computational complexity by incorporating a 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 at-

tribute 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 are gradually grown up into larger pairs (subgraphs) by applying chunking recursively.

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 inspection and urinalysis. 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. Since the discriminative patterns extracted by DT-GBI (those used at the nodes of the constructed decision tree) tend to be included in only relatively small number of patients and thus too specific, we tried to extract patterns with both relatively high discriminative and descriptive power by B-GBI. Furthermore, since there are the exceptional situations (patients) with the extracted patterns, these patterns are further utilized to extract refined knowledge from the dataset.

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.

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 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 greedy search, some typical patterns that exist in the given graph may not be extracted. However, GBI's objective is not to find all typical patterns nor all frequent patterns, but to extract only meaningful typical patterns of certain sizes. The stepwise pair expansion algorithm is summarized in Figure 1. As described above, currently frequency is used as chunking criterion at line 3 in Figure 1.

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

Figure 1. Algorithm of GBI

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 due to greedy search. 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. at line 3 in Figure 1. To prevent each branch growing exponentially, the 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

DT-GBI(D)

1: Create a node DT for D
2: **if** termination condition reached
3: return DT
4: **else**
5: P := GBI(D) (with the number of chunking specified)
6: Select a pair p from P
7: Divide D into D_y (with p) and D_n (without p)
8: Chunk the pair p into one node c
9: D_{yc} := contracted data of D_y
10: **for** D_i := D_{yc}, D_n
11: DT_i := DT-GBI(D_i)
12: Augment DT by attaching DT_i as its child along yes(no) branch
13: **return** DT

Figure 2. Algorithm of DT-GBI

class 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

Table 1. class label for interferon therapy

between them, attributes useful for classification task are gradually grown up into larger pairs (subgraphs) by applying chunking recursively. The above process is summarized in Figure2.

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

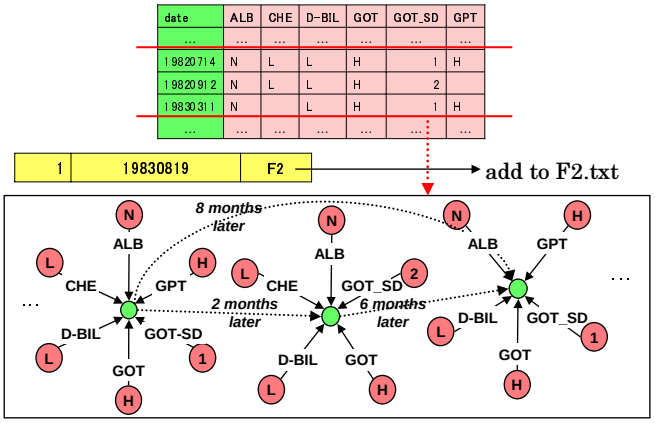

Figure 3. An example of converted graph structured data

for the patients with label R and N (38 and 56 patients, respectively).

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 [1], 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 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 patho-

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

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

Table 2. Size of graph structured data

logical tests that are more than a predefined interval (here, 8 weeks) apart. Figure 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 the converted graph structured data in this paper.

3.2. Analysis by DT-GBI

Two criteria were used to apply DT-GBI for selecting pairs: frequency for selecting pairs to chunk, and 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]. Prediction accuracy of the constructed decision trees constructed by DT-GBI was evaluated as 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 overall error rate was 22.60%.

Patterns used at the upper nodes in two constructed decision trees [2] out of 100 are shown in Figures 4, 5. Since these patterns are used at the upper nodes in the decision trees, and the left pattern in Figure 4 was used at the root node of many decision trees, it is considered that these patterns are sufficiently discriminative for classifying patients for whom

2 One with the best prediction accuracy and the other with time-correlated patterns

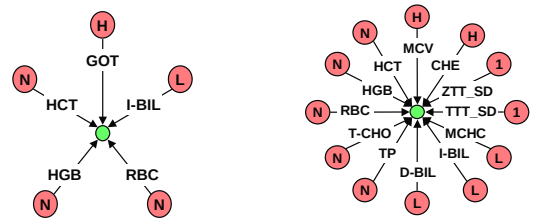

Figure 4. Examples of patterns extracted by DT-GBI (if exist, then R)

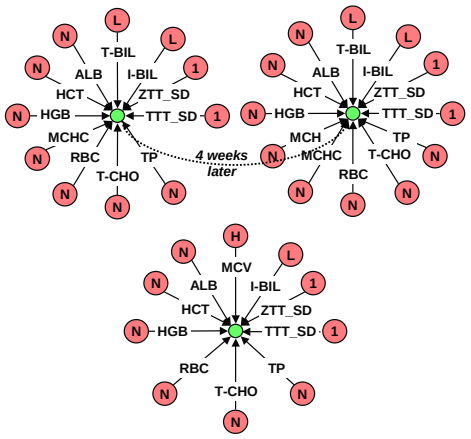

Figure 5. Examples of patterns extracted by DT-GBI (if exist, then R)

interferon therapy was effective (with class label R). However, although these patterns are discriminative in terms of information gain, these are included only in the very limited number of patients (left one in Figure 4 is included in only 5 patients out of 94 and right one in 6) and thus too specific. Furthermore, the constructed 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.

3.3. Analysis by B-GBI

Besides the prediction accuracy, the decision trees constructed by DT-GBI were rather unbalanced. This is because the patterns with large discriminative power (information gain) have 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 from the patients analyzed in subsection 3.2 in terms of not only the discrimina-

	94 Graphs in Table 2				41 Graphs with patterns in Figure 6			
	only R	only N	common	Total	only R	only N	common	Total
No. of patterns	2604	3468	5467	11539	1392	4587	3678	9657
Maximum No. of nodes per pattern	35	38	38	38	39	42	39	42
Average No. of nodes per pattern	15.4	15.0	10.9	13.1	15.9	16.3	10.8	14.1

Table 3. Summary of extracted patterns by B-GBI

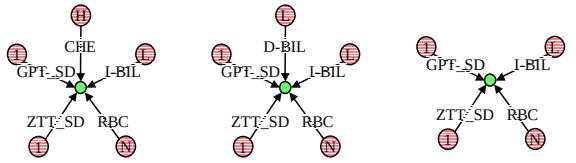

Figure 6. patterns with large support (41 patients (R:N = 10:31) out of 94)

tive power but also the support. B-GBI was terminated wh the support of all the extracted patterns [3] became less th 0.1. Beam width b was set to 3 as in subsection 3.2. The e tracted patterns were divided into 3 groups: 1) patterns i cluded only in the patients with class label R, 2) those wi class label N, and 3) those with both label R N (these grou are called only R, only N, common, respectively). The nur ber and size of the extracted patterns from the graphs in T ble 2 are summarized in the left-hand side of Table 3.

To seek for patterns with both discriminative and descriptive, first, 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 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 Figure 6. These patterns are included in 10 patients with label R and 31 patients with label N.

Since 31 patients out of 41 with the patterns in Figure 6 have label N, these patterns can be considered as indicating "patients with these patterns *tend to* have label N". However, there are exceptions (namely, 10 patients who have class label R with these patterns) and thus it is necessary to further refine this extracted knowledge. Thus, the patients with these patterns were further analyzed by B-GBI and the extracted patterns were also divided into 3 groups as before. This process is illustrated in Figure 7. In Figure 7, $\boldsymbol{G}$ stands for the set of 94 patients, $\boldsymbol{G_R}$ and $\boldsymbol{G_N}$ for the set of patients with label *R* and *N*. $\boldsymbol{G'_R}$ and $\boldsymbol{G'_N}$ are the set of patients with the patterns in Figure 6 and the patterns are extracted from these patients by B-GBI. $\boldsymbol{P'_R}$ is the set of patterns extracted from $\boldsymbol{G'_R}$ and $\boldsymbol{P'_N}$ from $\boldsymbol{G'_R}$. The number and size of the extracted patterns from $\boldsymbol{G'_R}$ and $\boldsymbol{G'_N}$ are summarized in the right-hand side of Table 3.

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

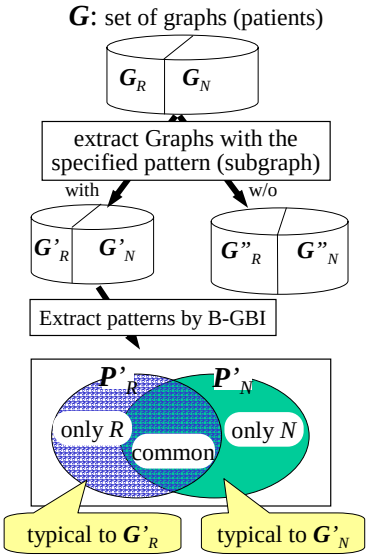

Figure 7. Refinement of extracted knowledge

Examples of extracted patterns with a large information gain are shown in Figure 8. Upper pattern is included *only* in the patients with label R and lower one *only* in the patients with label N. Thus, these patterns can be considered as refining the extracted knowledge as: "patients with upper pattern in Figure 8 in conjunction with the patterns in Figure 6 are actually with R" and "patients with lower pattern in Figure 8 in conjunction with the patterns are definitely with N". Furthermore, by analyzing the extracted patterns, it was revealed that extracted patterns in $\boldsymbol{P'_R}$ and in $\boldsymbol{P'_N}$ share many nodes and links with the patterns in Figure 6 and differ only with respect to some nodes and links. Thus, the different portions of nodes and links can be considered as effective for refining the extracted knowledge. In addition, 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 Figure 6.

The same domain expert who evaluated the results by DT-GBI also evaluated the results by B-GBI. Unfortunately,

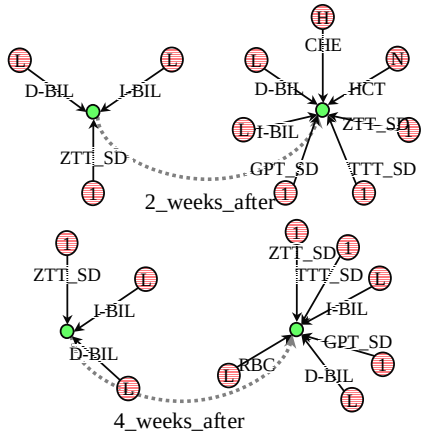

Figure 8. Example of extracted patterns (upper: in "only R", lower: in "only N")

many patterns appear in both class label R and N, as shown in the column "common" in Table 3 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. Since the extracted patterns are discriminative but tend to be too specific, B-GBI was applied to the dataset to seek for both discriminative and descriptive patterns. Furthermore, the extracted knowledge (patterns) by B-GBI was refined. 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 appropriate way, and 3) extract more time-correlated patterns using some bias for time interval edges.

Acknowledgment

This work was partially supported by the grant-in-aid for scientific research 1) on priority area "Active Mining" (No. 13131101, No. 13131206) and 2) No. 14780280 funded by the Japanese Ministry of Education, Culture, Sport, Science and Technology. Special thanks are due to Chiba University for providing us with the hepatitis dataset.

References

[1] T. B. Ho, T. D. Nguyen, S. Kawasaki, S. 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, 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. In *Working note of International Workshop on Active Mining (AM2003)*, pages 106–117, 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, 2003.

[11] K. Yoshida and H. Motoda. Clip : Concept learning from inference pattern. *Journal of Artificial Intelligence*, 75(1):63–92, 1995.