社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133（11）
2003／9／14

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

Warodom GEAMSAKUL†, Tetsuya YOSHIDA†, Kouzou OHARA†, Hiroshi MOTODA†, and Takashi WASHIO†

† Institute of Scientific and Industrial Research, Osaka University
8–1, Mihogaoka, Ibaraki, Osaka 567–0047, Japan
E-mail: †{warodom,yoshida,ohara,motoda,washio}@ar.sanken.osaka-u.ac.jp

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

Key words Data mining, graph-structured data, Decision Tree Graph-Based Induction (DT-GBI)

Decision Tree Graph-Based Induction による肝炎データからの診断知識の抽出

ワロドム・ジアムサクン†　吉田哲也†　大原剛三†　元田　浩†　鷲尾　隆†

† 大阪大学産業科学研究所　〒567–0047 大阪府茨木市美穂ヶ丘 8–1
E-mail: †{warodom,yoshida,ohara,motoda,washio}@ar.sanken.osaka-u.ac.jp

あらまし Decision Tree Graph-Based Induction (DT-GBI 法) はグラフ構造データから決定木を構築する手法である。DT-GBI 法では、部分グラフ（分類に効果的なパターン）を逐次ペア拡張（チャンキング）により抽出し、決定木の各ノードでの分岐パターンとして使用する。DT-GBI 法を千葉大学付属病院より提供された肝炎データセットに適用した。実験 1 では、線維化の段階（程度）をクラスとし、血液検査の時系列のみで第 4 段階（肝硬変）の患者とそれ以外の段階の患者を分類する決定木を構築し、第 2 実験では、肝炎の型（B または C）をクラスとし、第 1 実験と同様に肝炎の型を分類する決定木を構築する。初期実験の結果で得られた決定木，分類に効く検査パターンおよび予測精度を報告する。

キーワード データマイニング、グラフ構造データ、Decision Tree Graph-Based Induction（DT-GBI 法）

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 they are noticed. Although periodical inspection and proper treatment are important in order to prevent this situation, there are problems of expensove cost and physical burden on a patient. Although there is an alternative much cheaper method of inspection (blood test), the amount of data becomes enormous since the progress speed of condition is slow.

The hepatitis data set we are attempting to analyse is a real-world data provided by Chiba University Hospital.

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

Figure 1 Algorithm of GBI

There are some other analyses already conducted and reported on this dataset. [8] analysed the data by constructing decision trees from time-series data without discretizingg numeric values. [2] proposed a method of temporal abstraction to handle time series data, converted time phenomema to symbols and used a standard classifer. [9] used multiscale matching to compare time series data and clustered them using rough set theory. [5] also clustered the time series data of a certain time interval into several cateogories and used a standard classifier.

We have proposed a method called Decision Tree Graph-Based Induction (DT-GBI), which constructs a classifier (decision tree) for graph-structured data while simultaneously constructing attributes themselves for classification using GBI [10]. We conducted experiments to test our DT-GBI using this hepatitis data. The stages of fibrosis are used as classes in the first two experiments, and the types of hepatitis (B and C) are used as classes in the third experiment. The decision trees are constructed to discriminate between two groups of patients using no biopsy results but only the time sequence data of blood inspection.

2. Decision Tree Graph-Based Induction (DT-GBI)

GBI employs the idea of extracting typical patterns by stepwise pair expansion (we call this process "chunking"). In GBI, an assumption is made that 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 this, some typical patterns that exist in the input graph may not be extracted. However, GBI's objective is not to find all typical patterns nor all frequent patterns, but to extract only meaningful typical patterns of certain sizes.The stepwise pair expansion algorithm is summarized in Figure 1.

To increase the search space and extract more discriminative patterns while still keeping the computational complexity within a tolerant level, a beam search is incorporated to GBI, still, within the framework of greedy search [4]. 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 toal 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.

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

Figure 2 Algorithm of DT-GBI

If pairs are expanded in a step-wise fashion by B-GBI and discriminative ones are selected and further expanded while constructing a decision tree, discriminative patterns (subgraphs) can be constructed simultaneously while constructing a decision tree. The algorithm of DT-GBI is summarized in Figure 2. Since the values for an attribute are yes (contains pair) and no (does not contain pair), the construcuted decision tree is represented as a binary tree. Every time when an attribute (pair) is selected to split the data, the pair is chunked into a larger node in size. Thus, although initial pairs consist of only two nodes and one link between them, attributes useful for classification task are gradually grown up into larger pair (subgraphs) by applying chunking recursively.

Recursively partitioning data until each subset in the partition contains data of a single class often results in overfitting to the training data and thus degrades the predictive accuracy of decision trees. To improve the predictive accuracy, a pessimistic pruning used in C4.5 [7] is implemented by growing an overfitted tree first and then pruning it based on the confidence interval for binomial distribution.

3. Analysis of Hepatitis Data

The data set contains long time-series data (from 1982 to

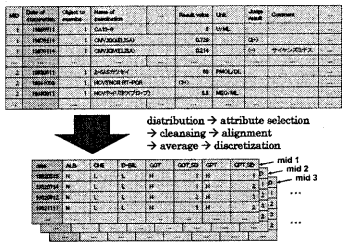

Figure 3 An example of graph conversion in phase 1-2

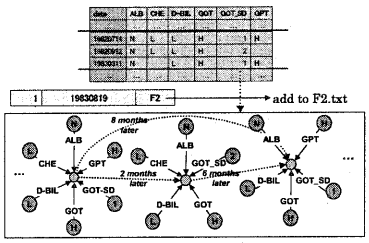

Figure 4 An example of graph conversion in phase 3

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

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 [6] for finding discriminative patterns after chunking.

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 (Remark1), and the numeric attributes are averaged over two-month interval and for some of them, standard deviations are calculated over six month interval and added as new attributes. Numerical 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 from 500 days before to 500 days after the first biopsy of each patient were converted into a graph. Here, the date of first biopsy and the result, to be treated as class (Remark2), of each patient are searched from the biopsy data file. In case that the result of the second biopsy or after differs from the result of the first one, the result from the first biopsy is defined as the class of that patient for the entire 1,000-day time-series.

In the last phase of data preparation, one patient record is mapped into one directed graph. Assumption is made that there is no direct correlation between two sets of pathological tests that are more than a predefined interval (here, two years) apart. Hence, time correlation is considered only within this interval. Figure 4 shows an example of conversion of data to graph. In this figure, a star-shaped subgraph represents values of a set of pathological examination in the two-month interval. The centre node of the subgraph is a hypothetical node for the two-month interval. An edge pointing to a hypothetical node represents an examination. The node connected to the edge represents the value (processed result) of the examination. And the edge linking two hypothetical nodes represents time difference.

3.2 Classifying Patients with Fibrosis Stages

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

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

A decision tree was constructed in either of the following

(Remark1) : Letters and symbols such as H, L, +, or - are deleted from numeric attributes.

(Remark2) : Activity, progress of fibrosis, hepatitis type, etc. can be taken as class.

Table 1 Size of graphs (classified by fibrosis stage)

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

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

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

two ways: 1) apply chunking N_r=20 times at the root node and only once at the other nodes of a decision tree, 2) apply chunking N_e=20 times at every node of a decision tree. Decision tree pruning is conducted by postpruning: conduct pessimistic pruning by setting the confidence level to 25%.

3. 2. 1 Experiment 1: F4 stage vs {F0+F1} stages

All four instances in F0 and 61 instances in F1 stage for non-cirrhosis class were used in this experiment. We performed 10 cycles of 10 fold cross-validation. In the first cycle,

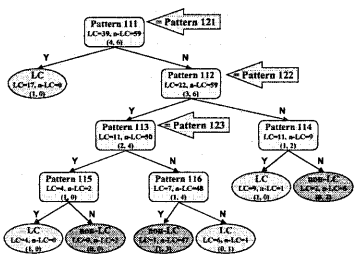

Figure 5 One of trees from the best cycle in exp.1 (N_e=20)

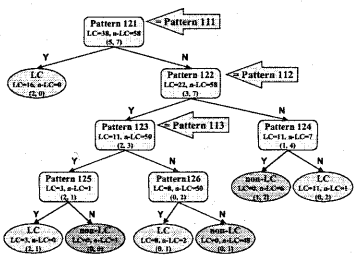

Figure 6 One of trees from the worst cycle in exp.1 (N_e=20)

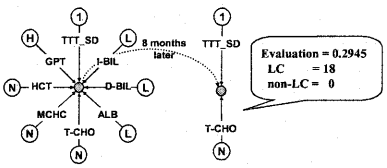

Figure 7 Pattern 111 = Pattern 121, if exist then LC

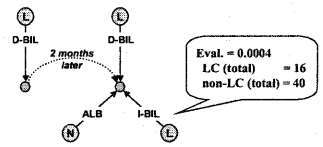

Figure 8 Pattern 112 = Pattern 122

the beam width was changed from 1 to 15. The lowest prediction error rates were obtained when the width was 15 for both methods (N_r=20 and N_e=20). Thus, for the rest nine cycles, we set the beam width to 15 when running DT-GBI.

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

3. 2. 2 Experiment 2: F4 stage vs {F3+F2} stages

In this experiment, we used all instances in F3 and 28 instances in F2 stage for non-cirrhosis class. As in experiment 1, we performed 10 cycles of 10 fold cross-validation.

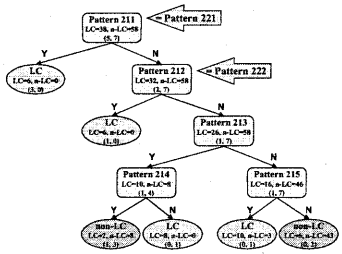

Figure 9 One of trees from the best cycle in exp.2 (N_e=20)

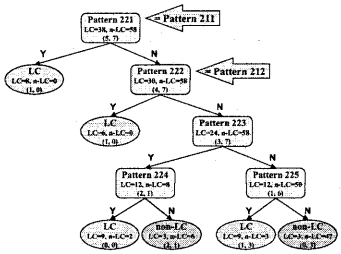

Figure 10 One of trees from the worst cycle in exp.2 (N_e=20)

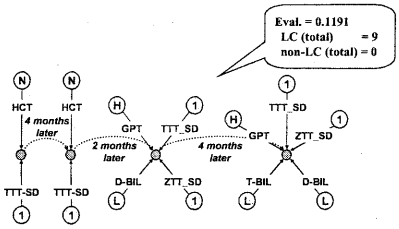

Figure 11 Pattern 211 = Pattern 221, if exist then LC

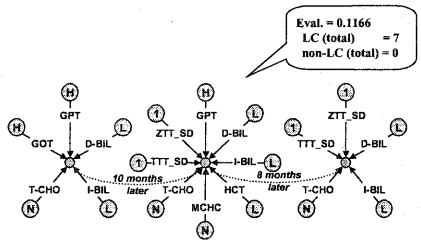

Figure 12 Pattern 212 = Pattern 222

The lowest prediction error rate was obtained in the first cycle when beam width was set to 14 for both 1) and 2). Thus, we set beam width to 14 when operating DT-GBI in the remaining nine cycles.

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

3.2.3 Discussion

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

Patterns shown in Figure 7, 8, 11, and 12 are sufficiently discriminative since all of them are used at the nodes in the upper level of all decision trees. The certainty of these patterns is ensured as, for almost patients, they appear after the biopsy.

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

3.3 Classifying Patients with Types (B or C)

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

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

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

Figure 13 Data of No.203 patient

Table 3 Size of graphs (classified by type)

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

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

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

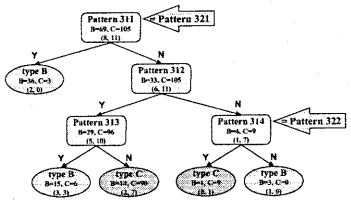

Figure 14 One of trees from the best cycle in exp.3 (N_e=20)

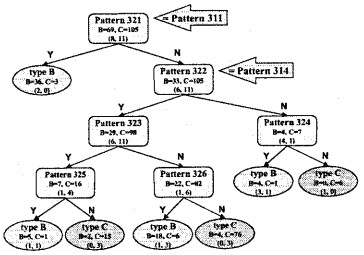

Figure 15 One of trees from the worst cycle in exp.3 (N_e=20)

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

4. Conclusion

This paper reports the preliminary results of analysing the hepatitis data set from Chiba University Hospital by using DT-GBI. Decision trees were constructed to distinguish patients at the most severe stage of fibrosis and those at the other stages in the first two experiments, and decision trees distinguishing patients of type B and those of type C were constructed in the third experiment. We believe that

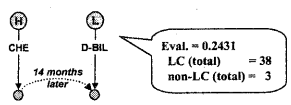

Figure 16 Pattern 311 = Pattern 321, if exist then type B

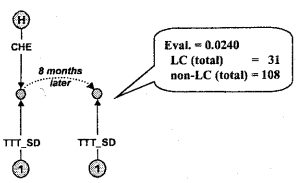

Figure 17 Pattern 322 = Pattern 314

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

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

Acknowledgement

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

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth & Brook/Cole Advanced Books & Software, 1984.

[2] T. B. Ho, T. D. Ngyuen, S. Kawasaki, and D. D. Ngyuen. Abstracting Temporal Data for Mining Tasks in the Hepatitis Domain. In *Active Mining Report (Fiscal year 2002)*, pages 133–138, 2003.

[3] T. Matsuda, T. Horiuchi, H. Motoda, and T. Washio. Extension of graph-based induction for general graph structured data. In *Knowledge Discovery and Data Mining: Current Issues and New Applications, Springer Verlag, LNAI 1805*, pages 420–431, 2000.

[4] T. Matsuda, H. Motoda, T. Yoshida, and T. Washio. Knowledge discovery from structured data by beam-wise graph-based induction. In *Proc. of the 7th Pacific Rim International Conference on Artificial Intelligence, Springer Verlag, LNAI 2417*, pages 255–264, 2002.

[5] M. Ohsaki, Y. Sato, S. Kitaguchi, and T. Yamaguchi. A Rule Discovery Support System for Sequential Clinical Data. In *Active Mining Report (Fiscal year 2002)*, pages 147–152, 2003.

[6] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[7] J. R. Quinlan. *C4.5: Programs For Machine Learning*. Morgan Kaufmann Publishers, 1993.

[8] E. Suzuki, T. Watanabe, Y. Yamada, F. Takeuchi, Y. Choki, K. Nakamoto, S. Inatani, N. Yamaguchi, M. Nagahama, H. Yokoi, and K. Takabayashi. Toward Spiral Exception Discovery. In *Active Mining Report (Fiscal year 2002)*, pages 153–160, 2003.

[9] S. Tsumoto, K. Takabayashi, M. Nagira, and S. Hirano. Trend-evaluation Multiscale Analysis of the Hepatitis Dataset. In *Active Mining Report (Fiscal year 2002)*, pages 191–197, 2003.

[10] Warodom G., T. Matsuda, T. Yoshida, H. Motoda, and T. Washio. Classifier construction by graph-based induction for graph-structured data. In *Advances in Knowledge Discovery and Data Mining, Springer Verlag, LNAI 2637*, pages 52–62, 2003.

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