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

2003－ＩＣＳ－133 (33)
2003／9／16

Stepwise Entropy Analysis

KenichiYoshida†

† Graduate School of Business Science, University of Tsukuba,
Otsuka 3-29-1, Bunkyo, Tokyo 112-0012, Japan
E-mail: †yoshida@gssm.otsuka.tsukuba.ac.jp

Abstract The importance of analyzing semi-structured data, such as hyper linked WWW text, XML data and chemical formulae, promotes research on a group of data-mining methods which can handle trees, graphs, and other non-table format data. Among these, GBI: Graph Based Induction [1,2] is one of the incipient methods to extract hidden rules from graph-format data. This paper describes an algorithm, SEA: Stepwise Entropy Analysis, using the GBI method. Though the conventional GBI algorithm contracts with the input graph during the analysis, this new algorithm does not use a contraction operation. Although this new algorithm is still greedy, its greediness is slightly weakened by omitting the contraction operation. Another characteristic of SEA, the simultaneous discovery of the association rule and the classification rule, is also described.

Key words Graph based induction, entropy

Stepwise Entropy Analysis

吉田 健一†

† 筑波大学 大学院ビジネス科学研究科
〒112-0012 東京都文京区大塚3-29-1
E-mail: †yoshida@gssm.otsuka.tsukuba.ac.jp

あらまし ここ数年WWW上のHTML/XMLデータや、化学物質の構造など、従来の統計・機械学習手法が扱ってきたテーブル形式では表現できないデータを解析したいというニーズが顕在下してきている。GBI(Graph Based Induction) 法 [1,2] は、そのようなニーズを受け比較的初期から研究が進められてきた色付有向グラフをデータ表現に用いた規則学習手法である。本研究ではGBIのための新しいアルゴリズムとして従来用いられてきたグラフの縮約操作を用いないSEA(Stepwise Entropy Analysis) というアルゴリズムを提案する。SEAは縮約操作がもたらすGBIのgreedinessを柔らげる事で、学習ルールの精度向上を狙ったアルゴリズムである。さらにSEAのもう1つの特徴である相関ルールと識別ルールの同時学習能力についても報告する。

キーワード Graph based induction, 情報量

1. Introduction

The importance of analyzing semi-structured data, such as hyper linked WWW text, XML data and chemical formulae, promotes research on a group of data-mining methods which can handle trees, graphs, and other non-table format data [1–8]. Among these, GBI: Graph Based Induction [1,2] is one of the incipient methods which extract hidden rules from graph format data.

From a performance point of view the original GBI algorithm has room for improvement. Recent research has also revealed its weaknesses and proposes improvements. In particular DT-GBI [9] uses a beam search to amend the greediness of the original GBI, and improves its prediction accuracy.

This paper proposes another improvement. In particular, it describes an algorithm, SEA: Stepwise Entropy Analysis, as a graph based induction method. Though the conventional GBI method contracts with an input graph during the analysis, this new algorithm does not use a contraction operation. Although this new algorithm is still greedy, its greediness is slightly weakened by omitting the contraction operation.

Another characteristic of SEA, the simultaneous discov-

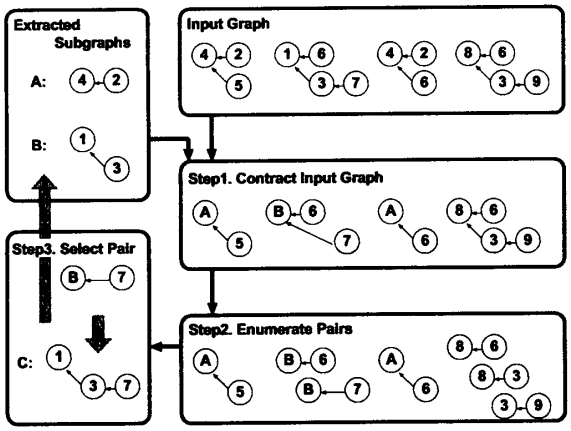

Figure 1 Stepwise Pair Expansion of GBI

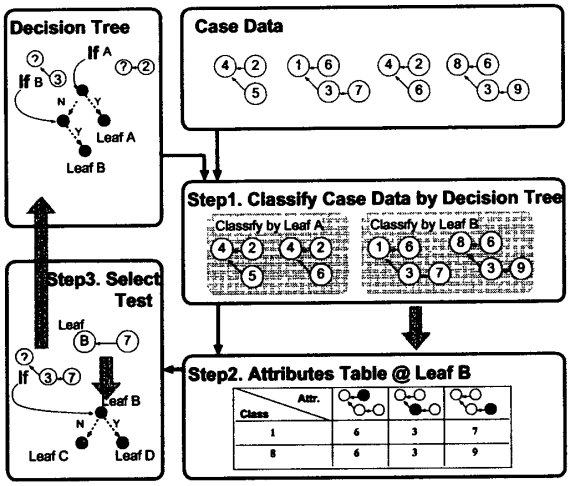

Figure 2 Interpretation of Stepwise Pair Expansion as Decision Tree Learning

ery of the association rule and the classification rule, is also described.

2. Stepwise Entropy Analysis

2.1 Original stepwise pair expansion [1]

GBI was originally designed as a frequent sub-graph finding algorithm [2]. The algorithm to extract frequent sub-graphs, named the stepwise pair expansion, is outlined bellow:

- First, the input graph is contracted according to the extracted Sub-graphs. Every occurrence of the extracted sub-graph in the input graph is replaced by a single node in Step 1. Figure 1 assumes that 1) the algorithm has already extracted two sub-graphs (sub-graph A: 4←2 and sub-graph B: 1←3), 2) each occurrence of the sub-graph 4←2 is replaced by a single node A, and 3) each occurrence of the sub-graph 1←3 is replaced by a single node B. In this step, the incoming edges are ordered, and the equivalence between the corresponding edges is examined.
- In Step 2, the contracted graph is analyzed and every possible Sub-graph, called a pair, that is made up of two linked nodes is extracted from the contracted graph. In Figure 1, seven pairs are extracted.
- Step 3 selects the best pair or pairs which satisfy certain criteria. In Figure 1, pair B←7 is selected as the best pair. The selected pair is then expanded to the original sub-graph, and added to the set of the extracted sub-graphs. In Figure 1, pair B←7 is expanded to sub-graph 1←3←7, since the node B is the replacement of the sub-graph 1←3.

Starting from the empty set of the extracted sub-graphs, this algorithm can extract various sub-graphs which appear frequently in the input graph. By repeating these steps, it extracts complex sub-graphs in a step-by-step manner. The expanding process in Step 3 contributes to finding complex sub-graphs.

This process can be interpreted as an decision tree learning process [1]. To use GBI for a decision tree (i.e., classification rule) learning, class information is represented as the color of root nodes (4, 1, 4 and 8 in Figure 1 and 2). The value of the attribute is coded as the color of the connecting nodes (2, 5, 6, 3, 7, 2, 6, 6, 3, and 9 in Figure 1 and 2). The information gain, which measures how a new test T affects entropy and contributes to the classification of data set D, is defined as:

$$\text{Information gain(D, T)} = Entropy(D) - \sum_{G_i \in G} \frac{|G_i|}{|D|} Entropy(G_i)$$

where

$$Entropy(D) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

and G_i is a subset of D classified by the test T. p_i is the probability of class i.

As shown in Figure 2, the above 3 steps in the stepwise pair expansion can be interpreted as the steps in the decision tree learning:

- Step 1 corresponds to the step of classifying training data set by the intermediate decision tree under construction.
- Step 2 corresponds to the step of analyzing the attribute table and calculates information gains.
- Step 3 corresponds to the step of selecting a new test condition and adds it to the intermediate decision tree under construction.

Note that the set of split training data set during the stepwise pair expansion is slightly different from that during the standard divide and conquer algorithm like ID3 [10]. At each step, GBI divides the training data set into two groups. A member of the first group has contracted nodes which is made up from a specific root node and a specific leaf node. A member of the other group might contain the specific leaf node only. On the other hand, the standard divide and con-

Main Routine *Stepwise Entropy Analysis*
begin
  while (Stopping Criteria are NOT Satisfied) **do**
    for (Each Classified Example Sets
        Divided by the Tree under Construction)
      Analyzing Entropy of each Pairs by *Entropy Analysis*
      Select Best Pair
      Add New Attribute Table based on Selected Pair
end

Procedure *Entropy Analysis*
begin
  for x **in** (Each Attribute: Class is treated as an Attribute)
    for y **in** (Each Attribute except x)
      for (Each Value: Color of y)
        Calc. $G \times (\frac{G_p}{G} \sum p_i log(p_i) + \frac{G-G_p}{G} \sum p_j log(p_j))$
          where G is the number of Nodes
            G_p is the number of Nodes
              with Target Pairs
            p_i,p_j are probabilities of
              each Color of Node$_x$
end

Figure 3 Stepwise Entropy Analysis

quer algorithm divides the training data set into two slightly different groups. Here, each member of the first group has a specific leaf node, and each member of the other group does not have this specific leaf node.

We pay attention to this difference as an potential weakness of GBI which might result in an inaccurate prediction, and therefore we develop an algorithm discussed in the next sub-section.

2.2 Stepwise Entropy Analysis

Figure 3 shows SEA, the Stepwise Entropy Analysis (SEA), algorithm. It removes the contraction operation from the stepwise pair expansion algorithm described in the previous sub-section. The characteristics of the SEA algorithm are:

- Though the conventional GBI algorithm using the stepwise pair expansion contracts an input graph during the analysis, SEA does not use the contraction operation. Step 2 of the stepwise pair expansion analyzes the attribute tables which are made from the contracted graph. SEA just adds corresponding attribute tables into the analysis in a stepwise manner. Though the contraction operation in the stepwise pair expansion sometimes removes attribute tables from the analysis, SEA does not remove such tables. Although this new algorithm is still greedy, its greediness is slightly weakened by omitting contraction operation.
- Another characteristic of SEA is its simultaneous discovery of an association rule and classification rule. Though step 1 of the stepwise pair expansion contracts a combination of nodes where each of the nodes corresponds to a class of information and of attribute information (See Figure 4), SEA also considers the combination of attribute nodes (Figure 5).

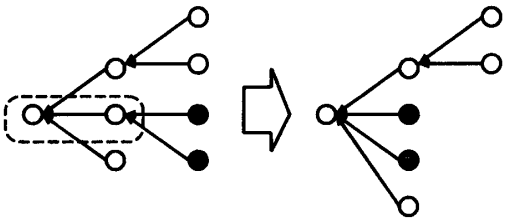

Figure 4 Contraction for Classification Rule Learning

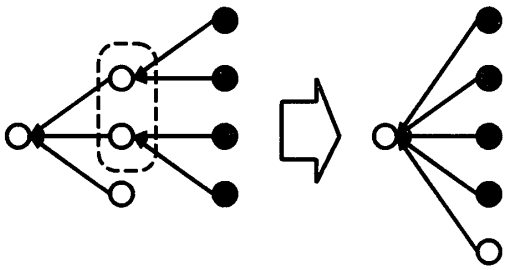

Figure 5 Contraction for Association Rule Mining

PlayTennis	Outlook	Temperature	Humidity	Wind
No	Sunny	Hot	High	Weak
No	Sunny	Hot	High	Strong
Yes	Overcast	Hot	High	Weak
Yes	Rain	Mild	High	Weak
Yes	Rain	Cool	Normal	Weak
No	Rain	Cool	Normal	Strong
Yes	Overcast	Cool	Normal	Strong
No	Sunny	Mild	High	Weak
Yes	Sunny	Cool	Normal	Weak
Yes	Rain	Mild	Normal	Weak
Yes	Sunny	Mild	Normal	Strong
Yes	Overcast	Mild	High	Strong
Yes	Overcast	Hot	Normal	Weak
No	Rain	Mild	High	Strong

Table 1 Training Examples

To handle both types of combinations in a fair method, the gain index is slightly modified so that it reflects the number of occurrences of such combinations. (Figure 3).

3. Learning Example

Suppose we have data on weather conditions and a decisions is to be made if it is suitable for playing tennis depending on the weather (Table 1). Figure 6 shows the decision tree made by the standard decision tree learning algorithm.

Figure 7 shows the decision tree made with SEA. In Figure 7, nodes 4, 9, 19 and 16 are test conditions. Other nodes , i.e., 2, 3, 5, 6, 7, 8, 11, 12, 13, 14 and 15, are frequent item-

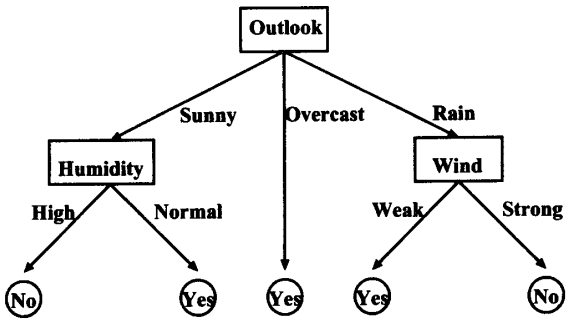

Figure 6 Standard Decision Tree

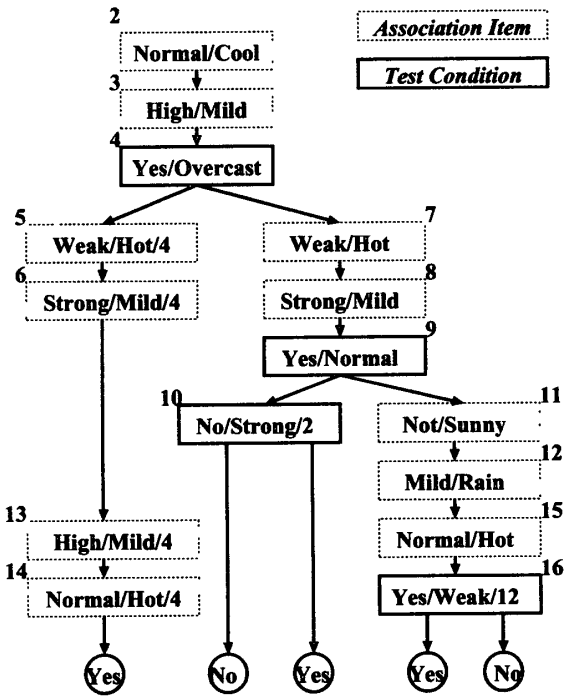

Figure 7 Decision Tree by SEA

sets. For example, the leftmost leaf of this tree represents the classification rule: *if the outlook is overcast, then play tennis.* The intermediate nodes, i.e., nodes 2, 3, 5, 6, 13 and 14, represents frequent itemsets. More precisely, nodes 2 and 3 represents the association of two attributes normal/cool and high/mild, and nodes 5, 6, 13 and 14, represents the association of three attributes, e.g., weak/hot/overcast.

Note the node 10 represents a test condition of the learned decision tree. It uses the found frequent itemset, Normal/cool, as a part of its test condition.

4. Related Works

The importance of analyzing semi-structured data, such as hyper linked WWW text, XML data and chemical formulae, promotes researches on a group of data-mining methods which can handle trees, graphs, and other non-table format data [1–8]. GBI [1,2] is one of the incipient methods and its weaknesses and improvements are studied in [9]. Our study is derived from these researches.

Another characteristics of our study is the use of an entropy based index for rule mining. The basic algorithm of APRIORI, which mines association rules from given data, was introduced by Agrawal et al. in [11]. Although it is the most widely used and well studied algorithm in the data mining field (See [12] for the general survey and comparison), it always requires careful tuning of the MinSup parameter. Suppose we have a data base and there exists; 1) rule X which has 100 supporting data items in the data base with 200 contradicting data items, and 2) rule Y which has 99 supporting data items with no contracting data items. If we set MinSup as 100, APRIORI only finds rule X. To find rule Y, we have to use a smaller MinSup which tends to produce more noisy results. The use of an entropy based index alleviates this defect.

The combination of the association rule mining system and the classification rule learning system has also been thoroughly studied in the data- mining field. Liu et al. [13] undertook a pioneering study on the combination of the association rule mining system and the classification rule learning system. Li et al. [14] reported on a method to improve both the efficiency and accuracy of a combined system. Liu et al. [15] and Wang et al. [16] also reported on such combined systems. Our study is also derived from such researches. One of the characteristics of our approach is the *simultaneous* discovery of the association rule and the classification rule. Most of the previous studies use the classification rule learning system to post-process the results of association rule mining systems.

Another important trend is the use of other indexes to select rules. Liu et al. [15] uses the ratio of the class in the data base. Li et al. [14] and Liu et al. [17] use χ^2 test. Smyth et al. [18] and Meretkis et al. [19] use entropy related indexes. Bayardo et al. [20] compares various indexes for rule selection.

5. Conclusion

A new algorithm, SEA: Stepwise Entropy Analysis, using the GBI method is proposed. The characteristics of SEA are:

- Though the conventional GBI method contracts with the input graph during the analysis, this new algorithm does not use the contraction operation.
  Although this new algorithm is still greedy, its greediness is slightly weakened by omitting the contraction operation.
- SEA simultaneously discovers the association rule and the classification rule.
  Thought most of the previous studies use the classification rule learning system to post-process the results of

association rule mining systems, SEA discovers them simultaneously by using entropy to guide its rule finding process.

We are now developing an intrusion detection system using SEA as it's main data-mining engine [21]. Although the conventional approaches, which use data-mining techniques for intrusion detection, share common defects, the use of SEA alleviates such defects.

References

[1] H. Motoda and K. Yoshida, "Machine learning techniques to make computers easier to use," *Artificial Intelligence*, vol. 103, no. 1, pp. 295–321, 1998.

[2] K. Yoshida and H. Motoda, "Clip: Concept learning from inference patterns," *Artificial Intelligence*, vol. 75, pp. 63–92, 1995.

[3] T. Asai, H. Arimura, K. Abe, S. Kawasoe, and S. Arikawa:, "Online algorithms for mining semi-structured data stream," in *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, 2002, pp. 27–34.

[4] M. Kuramochi and G. Karypis, "Discovering geometric frequent subgraphs," in *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, 2002, p. 258.

[5] X. Yan and J. Han, "gspan: Graph-based substructure pattern mining," in *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, 2002, p. 721.

[6] N. Vanetik, E. Gudes, and E. Shimony, "Computing frequent graph patterns from semi-structured data," in *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, 2002, p. 458.

[7] S. Parthasarathy and M. Coatney, "Efficient discovery of common substructures in macromolecules," in *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, 2002, p. 362.

[8] A. Termier, M.-C. Rousset, and M. Sebag, "Treefinder: a first step towards xml data mining," in *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 9-12 December 2002, Maebashi City, Japan*, 2002, p. 450.

[9] W. Geamsakul, T. Matsuda, T. Yoshida, H. Motoda, and T. Washio, "Improvement of search capability of decision tree - graph based induction (in japanese)," in *Proceedings of the 17th Annual Conference of JSAI*, 2003, pp. 2F2–01.

[10] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[11] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, P. Buneman and S. Jajodia, Eds., Washington, D.C., 26–28 1993, pp. 207–216. [Online]. Available: http://citeseer.nj.nec.com/agrawal93mining.html

[12] J. Hipp, U. Güntzer, and G. Nakhaeizadeh, "Algorithms for association rule mining – a general survey and comparison," *SIGKDD Explorations*, vol. 2, no. 1, pp. 58–64, July 2000. [Online]. Available: http://citeseer.nj.nec.com/hipp00algorithms.html

[13] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Knowledge Discovery and Data Mining*, 1998, pp. 80–86. [Online]. Available: http://citeseer.nj.nec.com/liu98integrating.html

[14] W. Li, J. Han, and J. Pei, "CMAR: Accurate and efficient classification based on multiple class-association rules," in *ICDM*, 2001, pp. 369–376. [Online]. Available: http://citeseer.nj.nec.com/li01cmar.html

[15] B. Liu, Y. Ma, C.-K. Wong, and P. S. Yu, "Scoring the data using association rules," *Applied Intelligence*, vol. 18, no. 2, pp. 119–135, 2003.

[16] K. Wang, S. Zhou, and Y. He, "Growing decision trees on support-less association rules," in *Knowledge Discovery and Data Mining*, 2000, pp. 265–269. [Online]. Available: http://citeseer.nj.nec.com/wang00growing.html

[17] B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations," in *Knowledge Discovery and Data Mining*, 1999, pp. 125–134. [Online]. Available: http://citeseer.nj.nec.com/liu99pruning.html

[18] P. Smyth and R. M. Goodman., "An information theoretic approach to rule induction from databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 4, pp. 301–316, Aug 1992.

[19] D. Meretkis and B. Wuthrich, "Extending naive bayes classifiers using long itemsets," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, 1999, pp. 165–174.

[20] R. Bayardo and R. Agrawal, "Mining the most interesting rules," in *Knowledge Discovery and Data Mining*, 1999, pp. 145–154.

[21] K. Yoshida, "Entropy based intrusion detection," in *Proc. of 2003 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 2003 (In Press).