知能と複雑系　132－14
(2003. 3. 14)

情報検索を用いたアクティブマイニングの螺旋サイクル

チャン・ナム・トアン† 市瀬龍太郎†† 沼尾 正行†

† 東京工業大学情報理工学研究科 〒152-8552 東京都目黒区大岡山 2-12-1
†† 国立情報学研究所知能システム研究系 〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: †tt-nam@nm.cs.titech.ac.jp, ††ichise@nii.ac.jp, †††numao@cs.titech.ac.jp

あらまし 本論文では，アクティブマイニングを実現するために，アクティブ情報収集とユーザー指向アクティブマイニングを組み合わせたマイニングシステムを提案する．文献データベースから，マイニングを行う属性とそれに関係がある文献数を調べ，属性の重要度を計算することで，属性の重要度に合わせた分類規則の生成を行う．さらに，文献から得られる属性とクラスの関係などの背景知識を活かすために，重要度を調整する機構も提案する．2種類の医療データを用いて評価を行った結果，提案した手法は面白い規則を生成できるのみならず，アクティブマイニングを実現するための一つの方法を提供できるということが示された．

キーワード アクティブマイニング、分類規則発見、MEDLINE、背景知識

Spiral Cycle of Active Mining by Using an Information Retrieval Approach

TuanNam TRAN†, Ryutaro ICHISE††, and Masayuki NUMAO†

† Department of Computer Science, Tokyo Institute of Technology
2-12-1, Oookayama, Meguro-ku, Tokyo, 152-8552 Japan
†† Intelligent Systems Research Division, National Institute of Informatics
2- 1-2, Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan
E-mail: †tt-nam@nm.cs.titech.ac.jp, ††ichise@nii.ac.jp, †††numao@cs.titech.ac.jp

Abstract This paper presents a method for realizing active mining. We have constructed a mining system which combines active information gathering and user-centered mining together. Our method modifies C4.5rules by taking into consideration the external weight of each attribute, which can be calculated by means of the number of corresponding documents found in the literature. A method for adjusting the weights introduced in order to reflect the background knowledge concerning the relationship between an attribute and an class is also proposed in this paper. The experiments on two kinds of medical data show that our proposed system is useful in terms of generating interesting rules as well as providing a solution for realizing active mining.

Key words active mining, classification rule mining, MEDLINE, background knowledge

1. Introduction

Knowledge discovery in databases (KDD) is defined as the non-trivial process of identifying *valid*, *novel*, *useful*, and ultimately *understandable* patterns in data [8]. Mining information and knowledge from large databases has been recognized by many researchers as a key research topic in database systems and machine learning, and by many industrial companies as an important area with an opportunity of major revenues.

However, currently, there is a lack of KDD researches for identifying and gathering relevant information related to the given data, which can be accessed easily through the Internet. We have constructed a data mining system called *RMAIG* (**R**ule **M**ining using **A**ctive **I**nformation **G**athering) based on the well-known propositional classification system C4.5 by using various heuristic functions taking into account the weight of each attribute of the given data in the literature. The rules obtained are filtered and evaluated using *support* and *confidence* as similar as in *association rule*

mining. The idea of weighting an attribute by looking for relevant documents related to that attribute is highly evaluated by domain experts as a solution for tracing domain's knowledge. *RMAIG* has been evaluated on two real-world medical data, showing many appropriate and interesting rules from the viewpoint of domain experts.

Our proposed weighted scheme is also useful for feeding back the results obtained to the mining process. Given the relation between a class and an attribute obtained by background knowledge, previous work or occurred in the literature, we can reweight that attribute so that *RMAIG* can produce rules concerning the focused attribute, even though these rules could not be discovered using the normal weighted scheme described above.

The remainder of this paper is organized as follows. Section 2 describes our proposed system. Section 3 describes the experiments on the hepatitis data set. Section 4 presents the results for adjusting weights in order to reflect the background knowledge in the literature. Some related work will be described in Section 5, and finally Section 6 presents our conclusion.

2. The *RMAIG* System

This section will describe *RMAIG*, a system which is based on a decision-tree-based approach by using MEDLINE information concerning the given data. *RMAIG*'s algorithm consists of two stages. The first stage modifies the classification system C4.5 by using various heuristic functions taking into account the weight of each attribute of the given data in the literature. The second stage filters the rules obtained in the previous stage that satisfy the user-specified minimum support (called *minsup*) and minimum confidence (called *minconf*) constraints.

The core of a decision tree algorithm is to repeat the process of selecting the attribute with highest information ratio. The characteristic of our method is to consider "weighting" by external information when calculating information ratio for each attribute. If the importance of an attribute (i.e. the number of literature documents from previous biomedical research) occurred in the given data can be calculated, the importance of that attribute can be calculated easily.

Suppose T is a set of training examples of a decision tree consisting of attributes $A_1, A_2, \ldots, A_m$, and $freq(C_i, S)$ is the number of cases in a set of examples S that belong to class C_i. The *entropy* of T is defined as:

$$info(T) = -\sum_{j=1}^{k} \frac{freq(C_j,T)}{|T|} \times log_2\left(\frac{freq(C_j,T)}{|T|}\right) \quad (1)$$

where $freq(C_j,T)$ stands for the number of cases in T that belong to class C_j.

Suppose T has been partitioned in accordance with the n outcomes $T_1, T_2, \ldots, T_n$ of a test X corresponding to the attribute A_j. Then, according to [9], *gain* and *gain ratio* for the attribute A_j at the given stage in the construction of the decision tree can be calculated as follows:

$$info_j(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} \times info(T_i) \quad (2)$$

$$gain_j = info(T) - info_j(T) \quad (3)$$

$$split\ info_j = -\sum_{i=1}^{n} \frac{|T_i|}{|T|} \times \log_2(\frac{|T_i|}{|T|}) \quad (4)$$

$$gain\ ratio_j = gain_j / split\ info_j \quad (5)$$

We have introduced *gain'* and *gain ratio'* based on *gain* and *gain ratio* as follows

$$gain_j^{'} = G(gain_j, \omega_j) \quad (6)$$

$$gain\ ratio_j^{'} = G(gain\ ratio_j, \omega_j) \quad (7)$$

Here, ω_j is the *external weight* of the attribute A_j, and is calculated by

$$\omega_j = \frac{F(|A_j|)}{\sum_{i=1}^{m} F(|A_i|)} \quad (8)$$

where $|A_j|$ stands for and the number of MEDLINE documents found related on the given data and the attribute A_j. The heuristic functions F and G will be described below.

Three kinds of functions G, namely G_1, G_2, and G_3 have been employed by the following

$$G_1(x,w) = x \times w \quad (9)$$

$$G_2(x,w) = x^2 \times w \quad (10)$$

$$G_3(x,w) = (2^x - 1) \times \frac{w}{1+w} \quad (11)$$

Note that these functions have been used in the previous work on *costs of tests*, which will be mentioned further in Section 5. We modified the heuristic functions in that the external weight of an attribute jth is inversely proportional to the cost of measuring this attribute.

As for the function F, we have currently defined two types of $F(x)$, namely F_1 and F_2 as follows:

$$F_1(x) = x \quad (12)$$

$$F_2(x) = \begin{cases} 0 & \text{if } x \leqq 0 \\ \lfloor (log_k(x) + 1) \rfloor & \text{if } x > 0 \end{cases} \quad (13)$$

where $k > 0$.

C4.5 selects the attribute that maximizes the information gain ratio (*gain ratio*), which is a function of the information gain, and we modified C4.5 so that it selects the attribute that maximizes *gain ratio'*.

Table 1 shows the list of heuristic functions used in *RMAIG*.

Table 1 A list of heuristic functions used in *RMAIG*

No.	Heuristic method	Function F	Function G
1	none		
2	log10	F_2 $(k = 10)$	G_1
3	log	F_2 $(k = e)$	G_1
4	linear	F_1	G_1
5	log10 + CS-ID3	F_2 $(k = 10)$	G_2
6	log + CS-ID3	F_2 $(k = e)$	G_2
7	linear + CS-ID3	F_1	G_2
8	log10 + EG2	F_2 $(k = 10)$	G_3
9	log + EG2	F_2 $(k = e)$	G_3
10	linear + EG2	F_1	G_3

2.1 Definition the support and confidence values of rules

We assume that a rule consists of two conditions called the antecedent and consequent, and is denoted as $A \rightarrow C$ where A is the antecedent and C is the consequent. The support of a condition A is equal to the number of separate MIDs in the data set for which A evaluates to true, and this value is denoted as $sup(A)$. The support of of a rule $A \rightarrow C$, denoted similarly as $sup(A \rightarrow C)$, is equal to the number of of separate MIDs in the data set for which both A and C evaluate to true. The antecedent support of a rule is the support of its antecedent alone. The confidence for a rule is the probability with which the consequent evaluates to true given that the antecedent evaluates to true in the input data set, defined as follows:

$$conf(A \rightarrow C) = \frac{sup(A \rightarrow C)}{sup(A)}$$

It should be noted that the support and confidence of a rule defined above do not consider the number of classes. We have applied the Laplace function which is commonly used to rank rules for classification purposes [1], [2] as follows:

$$laplace(A \rightarrow C) = \frac{sup(A \rightarrow C) + 1}{sup(A) + k}$$

where k is the number of classes when building a classification model $(k > 1)$.

Figure 1 shows the proposed system *RMAIG*.

3. Experiments

First, we applied the *RMAIG* system to the *hepatitis data set*, provided by Chiba University Hospital, contains administrative information as well as long time-series data of laboratory examinations of 771 patients with chronic hepatitis B and C who took biopsy in the period 1982-2001. We have currently considered the following problems:

- Discover knowledge concerning the stage of liver fibrosis using laboratory examinations
- Discover knowledge which indicates whether the interferon therapy is effective or not.

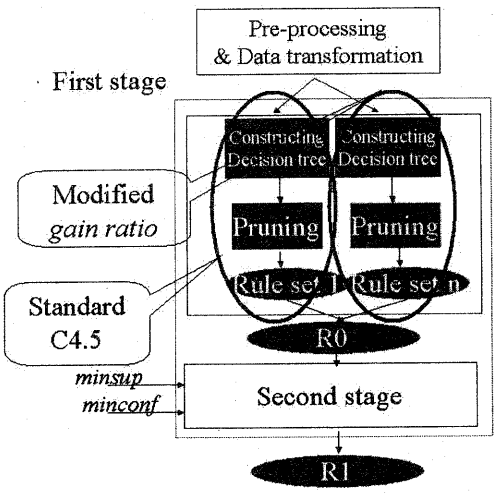

Figure 1 Proposed system *RMAIG*

- Discover knowledge which distinguishes hepatitis B and C.

In general, the purpose of the pre-processing stage is to generate a single table from the given six tables after conducting data cleaning, attribute selection, attribute generation using MID, a primary key given for each patient, and the examination date of the patients. The pre-processing of hepatitis data set in detail is described in [12].

3.1 Some rules obtained by the proposed method

Figure 2 shows some rules discovered only by our proposed method which were highly evaluated by the domain experts. Each rule is attached by the corresponding support, confidence and the weighted method. For example, the first rule

> IF $(CHE > 176)$
> THEN $Fibrosis = F1$ [support : 124, confidence : 51.9%, method : log + CS-ID3]

means that there are 124 patients out of total 771 patients satisfying this rule, while the number of patients satisfying the antecedent part is

$$124 * 100/51.9 = 239$$

3.2 Discussion

The way of choosing best attributes using the number of MEDLINE's corresponding documents makes the attributes with high external weighting values being taken priority over the low ones. That is, the occurring probability of those attributes paid attention to in the literature will become higher. The merit of using information gathered from biomedical literature is that, the users are able to know the "importance" of each attribute without any assistance of the domain experts. Moreover, the generated rules of our al-

1. IF ($CHE > 176$)
THEN $Fibrosis = F1$ [support : 124, confidence : 53.1%, method : log + CS-ID3]
Evaluation: This rule is in accordance with knowledge of domain experts, since the fibrosis is not progressive when CHE is high.
2. IF (($PT\_Sex = M$) AND ($Age = fifties$) AND ($ALP = +$) AND ($LAP = +$))
THEN $IFN\_effect = response$ [support : 11, confidence : 61.1%, method : log10, log10 + EG2]
Evaluation: This rule is interesting because both ALP and LAP occurred in the antecedent part are abnormal. In contrast, most of other mined rules containing normal values of ALP and LAP. Such a rule is unique and has never been discovered by other groups in the Active Mining Project.
3. IF ($CHE \leqq 12.58$)
THEN $hepatitis\_type$ = B [support : 117, confidence : 87.2%, method : log, log10]
Evaluation: In general, hepatitis B shows a lower value of CHE compared to hepatitis C, and it is possible to say that hepatitis B is more progressive than hepatitis C.

Figure 2 Some interesting rules obtained by the weighted proposed method. It should be noted that these rules were not obtained by the normal C4.5 method.

gorithm are able to reflect the state-of-the-art research in biomedical literature, since the number of documents related to an attribute, and as a result its external weight changes with time. One more merit of our system is that it is easy and flexible to update the weighting scheme. That is, we can increase the weight of the attributes that were highly evaluated by the domain experts, and by repeating the mining process we may obtain new interesting rules. We will give detailed discussion on this issue in Section 4.

In general, we have obtained some rules which were evaluated as "interesting" by domain experts. The domain experts also gave us valuable comments on the idea of weighting attributes using MEDLINE search as necessary and at the same time, it may generate knowledge which has already been known since our approach focuses on those attributes which are cited frequently in the literature, and it is impossible to conduct mining effectively without taking into account the past knowledge accumulated in the literature.

As same as other data mining methods, it is noteworthy that pre-processing is the key point to the success of the mining process. The set of rules obtained depends on the set of attributes used in the mining process, however it is not trivial to determine which examinations should be removed beforehand. This relates to the spiral effects of the *active mining* framework, since domain experts play an essential role before and after evaluating the rules obtained by *RMAIG*.

4. Adjusting Weights

Two previous sections have introduced the proposed system *RMAIG* and experimental evaluation on two kinds of medical data sets. In these two sections, the relationship between *active information gathering* and *user-centered mining* is mainly based on the idea of retrieving MEDLINE documents relating to each attribute occurred in the given data and using the number of documents obtained to calculate the weight of the corresponding attribute. That is, the weight of each attribute is unchanged during the mining process. In this section, we will investigate the correlation between *active information gathering* and *user-centered mining* by considering the *user reaction* part. First, using the rules obtained from the mining process may be useful for finding relevant documents in the *active information gathering*. For example, on the task of distinguishing chronic hepatitis B and C described in Section 3, *RMAIG* generated the following rule:

IF (CHE > 12.58) THEN hepatitis type = C
IF (CHE $\leqq$ 12.58) THEN hepatitis type = B

The medical doctors gave us a comment on this rule that, in general, hepatitis B shows a lower value of CHE compared to hepatitis C, and it is possible to say that hepatitis B is more progressive than hepatitis C.

On the contrary, there are many cases in which we have some background knowledge between an attribute and a class obtained by domain experts or by *active information gathering* from MEDLINE, but neither *RMAIG* nor C4.5 could easily discover such patterns from the given data. To solve the problem, a natural way is to increase the weight of the corresponding attribute so that its occurring probability in the mined rules will become higher. In this section, we conduct experiments on the *meningoencephalitis data set* donated by Prof. Tsumoto, Department of Medical Informatics, Shimane Medical University, since we have already known some background knowledge from previous work on this data set. This data set contains 140 patients of meningoencephalitis hospitalized sometimes between 1979 and 1992. Here, "meningoencephalitis" stands for those patients of meningitis who have some symptoms of brain damage. The given data set contains 38 attributes, and the goal of the mining task is to find factors important for diagnosis (DIAG and DIAG2), for detection of bacteria or virus (CULT_FIND and CULTURE) and for predicting prognosis (C_COURSE and COURSE). Although the hepatitis data set is not used in this section, we believe that the method described here is also applicable for the hepatitis data set.

Figure 3 indicates some important patterns of extracted

- **If CSF_CELL is high then CULT_FIND=T.** This knowledge indicates that the number of cells in Cerebulospinal Fluid is important for detecting bacteria or virus,
- **LOC is important for C_COURSE.** This knowledge shows the importance of early treatment for meningoencephalitis (or meningitis in general)
- **The relationships between SEX, AGE and DIAG (DIAG2) are interesting.** This knowledge means that personal information plays an important role for diagnosis, and this knowledge is interesting because it is not mentioned in the medical literature so far.

Figure 3 Some interesting knowledge on meningoencephalitis data set previously evaluated by the domain experts

rules indicated in [13] which is considered to be interesting by the domain experts.

For each class attribute, we vary the weight of the corresponding attribute in Figure 3, and compare with the original result in terms of number of rules concerning that attribute. That is, the weight CSF_CELL should be increased in the case of class $CULT\_FIND$ or class $CULTURE$. Similarly, the weight of LOC should be increased in the case of class C_COURSE or class $COURSE(Grouped)$, and the weight of SEX, AGE should be increased in the case of class $DIAG$ and $DIAG2$.

We modified the equation concerning the external weight ω_j to the following:

$$\omega_j = \frac{F(\alpha_j \times |A_j|)}{\sum_{i=1}^{m} F(\alpha_i \times |A_i|)} \tag{14}$$

$$\alpha_j = \begin{cases} \alpha & \text{if } A_j \text{ is "focused"} \\ 1 & \text{otherwise} \end{cases} \tag{15}$$

where α is a parameter.

Table 2 shows the number of rules obtained concerning the re-weighted attribute and the number of total rules for each class attribute. The number enclosed by parentheses means the number of total rules obtained. Here, $\alpha = 1$ stands for using the original $RMAIG$. The effectiveness of re-weighted scheme can be seen for every class attribute. For example, for the class $CULTURE$, we could not obtain any rules concerning CSF_CELL by standard C4.5 or $RMAIG$. However, at $\alpha = 100$, we found one out of five rules, and at $\alpha = 1000$ two out of six rules concerning CSF_CELL. We see from this example that adjusting the weight of focused attribute is a simple but effective way for discovering rules related to that attribute. Therefore, we conclude that the proposed method is very useful if we have already known the attribute to be focused on. However, the number of rules relating to the focused attribute and the total rules obtained may be saturated or even be decreasing at a very large value of α as in the case of the class C_COURSE at $\alpha = 1000$.

Table 2 Comparison of C4.5rules and our proposed method on the interestingness of rules found based on domain experts' background knowledge when *minsup*is 10 and *minconf* is 0.5

Class attribute	C4.5 rules	Proposed method			
		$\alpha = 1$	$\alpha = 10$	$\alpha = 100$	$\alpha = 1000$
CULTURE	0(0)	0(0)	0(4)	1(5)	2(6)
CULT_FIND	0(0)	0(0)	0(8)	0(7)	2(9)
CULTURE+ CULT_FIND	0(0)	0(0)	0(12)	1(12)	4(15)
C_COURSE	1(2)	3(9)	3(9)	5(7)	3(5)
COURSE(Gr.)	1(1)	1(3)	1(3)	1(4)	1(2)
C_COURSE+ COURSE(Gr.)	2(3)	3(9)	4(12)	6(11)	4(7)
DIAG (SEX)	0(0)	4(14)	3(12)	5(14)	5(14)
DIAG2 (SEX)	0(0)	1(8)	1(8)	1(8)	1(8)
DIAG+DIAG2 (SEX)	0(0)	5(22)	4(20)	6(22)	6(22)
DIAG (AGE)	0(0)	1(14)	2(16)	2(15)	2(15)
DIAG2 (AGE)	0(0)	1(8)	3(11)	3(11)	3(11)
DIAG+DIAG2 (AGE)	0(0)	2(22)	5(27)	5(26)	5(26)

5. Related Work

Regarding decision tree induction, some work has been conducted on *cost-sensitive classification* which consider either the *costs of tests* (features, measurements) or the *costs of classification errors*. For instance, there are several machine learning algorithms that consider the costs of tests such as IDX [6], CS-ID3 [10], and EG2 [7]. Our algorithm is similar to these algorithms in terms of modifying the information gain for selection of attributes, however it should be noted that they aim to minimize the costs of tests, while our purpose is to find interesting patterns for the domain experts. Some other studies consider the weighting scheme for instances such as the boosting algorithm [3] or attempt to adapt the boosting algorithm for cost-sensitive classification [11]. Turney [14] introduced a method which uses a genetic algorithm with the fitness function is the average cost of classification when using the decision tree, including both the costs of tests and the costs of classification errors.

Another approach for integrating classification and association rule mining is described in [4]. The difference between classification rule mining and association rule mining is that classification rule mining aims to discover a small set of rules in the database that forms an accurate classifier, while association rule mining finds all the rules existing in the database that satisfy some minimum support and minimum confidence

constraints. The key point of this study in the first stage is to focus on mining a special subset of association rules, called *class association rules*. After that, the second stage will build a classifier based on the set of subset discovered. The same group also proposed a technique [5] for post-precessing of association rule mining. This technique prunes the discovered associations to remove those insignificant associations, and then finds a special subset of the unpruned associations to form a summary of the discovered associations.

6. Conclusions

We have developed a new approach for realizing active mining. We have constructed a system which modifies the standard decision tree system C4.5 by using various heuristic functions taking into consideration the weight of each attribute of the given data in the literature. Each attribute of the given data is labelled a weight based on the frequency of those documents relating to that attribute. The merit of using information gathered from biomedical literature is that the users are able to know the "importance" of each attribute without any assistance of the domain experts.

One more merit of our system is that it is easy and flexible to update the weighting scheme. That is, we can increase the weight of the attributes that are highly evaluated by the domain experts or by previous work, and we show that our system with the updated weights is able to find rules concerning the focused attribute.

Acknowledgements

This research is supported by the grant-in-aid for scientific research on priority area "Active Mining" from the Japanese Ministry of Education, Culture, Sports, Science and Technology. The authors would like to thank Prof. Katsuhiko Takabayashi and Dr. Hideto Yokoi of Chiba University Hospital for their useful comments.

References

[1] R. J. Bayardo and R. Agrawal. Mining the most interesting rules. In *Proc. of the 5th International Conference Knowledge Discovery and Data Mining (KDD)*, pages 145–154, 1999.

[2] P. Clark and R. Boswell. Rule induction with CN2: Some recent improvements. In *Proc. of the Fifth European Working Session on Learning*, 1991.

[3] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proc. of the 13th International Conference on Machine Learning (ICML)*, pages 148–156, 1996.

[4] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Knowledge Discovery and Data Mining (KDD)*, pages 80–86, 1998.

[5] B. Liu, W. Hsu, and Y. Ma. Pruning and summarizing the discovered associations. In *Knowledge Discovery and Data Mining (KDD)*, pages 125–134, 1999.

[6] S. W. Norton. Generating better decision trees. In *Proc. of the 11th International Join Conference on Artificial Intelligence (IJCAI)*, pages 800–805, 1989.

[7] M. Núñez. The use of background knowledge in decision tree induction. *Machine Learning*, 6(3):231–250, 1991.

[8] G. Piatetsky-Shapiro and W. J. Frawley, editors. *Knowledge Discovery in Databases*. AAAI Press, 1991.

[9] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[10] M. Tan. Cost-sensitive learning of classification knowledge and its application in robotics. *Machine Learning*, 13(1):7–33, 1993.

[11] K. M. Ting. A comparative study of cost-sensitive boosting algorithms. In *Proc. of the 17th International Conference on Machine Learning (ICML)*, 2000.

[12] T. N. Tran, R. Ichise, and M. Numao. Mining hepatitis data set using information gathered from biomedical literature. In *Proc. of International Workshop on Active Mining (AM-2002), the IEEE International Conference on Data Mining (ICDM)*, pages 136–141, 2002.

[13] S. Tsumoto and K. Takabayashi. Comparison and evaluation of knowledge obtained by KDD methods. *Journal of the Japanese Society for Artificial Intelligence (JSAI)*, 15(5):790–797, 2000. In Japanese.

[14] P. D. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.