研究論文

Learning First-Order Rules to Handle Medical Data

Ryutaro ICHISE
National Institute of Informatics
Masayuki NUMAO
Tokyo Institute of Technology

ABSTRACT

Since information systems are used in large hospitals, a large amount of medical data is provided to physicians which often constitutes an information overload. Consequently, computers must extract useful information from such data. The most difficult issue in handling medical data is that included time-series data have irregularities. Here, we describe handling this type of data using a first order logic learning system named DAMS.

[Keywords]
Knowledge Discovery, Machine Learning, Inductive Logic Programming

1 Introduction

Hospital information systems that store medical data are very popular especially in large hospitals. Such systems hold medical records of patients, laboratory data and other types of information. Knowledge can be extracted from the medical data that can be useful for physicians when deciding treatment strategies. However, when physicians attempt to manually extract such information, the volume of data is too large to do so efficiently. Therefore, physicians require the support of computers to extract relevant information.

Medical data has three features.[8] The amount of records increases each time patients visit a hospital, values are often missing, usually because patients do not always undergo all examinations and time-series attributes with irregular time intervals are included. To handle medical data, a mining system must have a function supporting these features. Methods for mining data include K-NN, decision trees, neural nets, association rules and genetic algorithms.[1] However, these methods are unsuitable for medical data due to the inclusion of multiple relationships and time relationships with irregular intervals.

One way of handling multiple relationships is Inductive Logic Programming (ILP), [4][5] because it uses horn clauses that constitute a subset of first order logic. We propose a new system[3] with which to induce horn clauses from data in the manner of ILP systems. We applied this system to a medical data mining task with association rule criteria.

This paper is organized as follows. Section 2 characterizes the medical data with some examples. Section 3 presents a data mining system named DAMS along with algorithms and mechanisms. Section 4 applies DAMS to medical databases and Section 5 discusses the results and related studies.

2 Medical Data

As described above, the sample medical data shown here have three features. Table 1 shows an example of laboratory examination data including eight attributes. The first attribute, ID, means personal identification. The second is Examination Date, which is date when the patient is consulted by a physician. The remaining attributes designate results of laboratory tests.

The first feature shows that the data contain a large amount of records. The amount of data in this table increases quickly because new records with many attributes are added every time a patient undergoes an examination.

The second feature is that many values are missing from the data. Table 1 shows that many values are absent from the attributes that indicating the results of laboratory

Table 1: Medical Data Example

ID	Examination Date	GOT	GPT	WBC	RBC	RNP	SM
14872	19831212	30	18				
14872	19840123	30	16				
14872	19840319	27	17	4.9	4.33		
14872	19840417	29	19	18.1	4.18		
14872	…						
5482128	19960516	18	11	9.1	4.01	-	-
5482128	19960703	25	23	9.6	3.9		
5482779	19980526	52	59	3.6	4.3	4	-
5482779	19980811			4	4.16		
5482779	…						

examinations. Since this table is an extract from medical data, the number of missing values is quite low. However, this number is far higher in the actual data. That is, most of the data is missing values because each patient underwent only some tests at one examination.

The other feature of the medical data is that it contains time-series attributes. When a table does not have these attributes, then the data only contain a relationship between ID and examination results. Under these circumstances, decision tree learning or any other propositional learning method can be applied to the data. However, relationships between examination test dates are also included, that is, multiple relationships.

In contrast to other time-series information such as stock market data, medical data is not collected at regular intervals. Consequently, traditional time-series analytical methods are not suitable for direct application to medical data.

3 Data Mining with Medical Data

We developed a new mining system for medical data that can handle multiple relationships like ILP systems. We refer to it as DAMS (DAta Miner with Syngip) and it is an extension of SYNGIP for data mining. We outline SYNGIP then present DAMS.

3.1 SYNGIP

SYNGIP is a relational learning system that uses horn clauses found using an evolutional search. To do this, it considers an individual as a set of horn clauses. At the same time, it considers a gene as a literal when the individual contains one horn clause, and considers a gene as a horn clause when the individual contains two or more horn clauses. The SYNGIP algorithm is shown in Figure 1. SYNGIP creates an initial population by randomly using refinement operators,[7] then it evaluates all individuals by fitness function. SYNGIP then chooses suitable individuals and places them in the next generation population. In other words, an elite strategy is used. The remainder of the new generation is formed by applying a crossover operator for the selected candidates. A mutation operator is then randomly applied to the population. When SYNGIP applies the genetic operators, it uses type and mode information to generate valid individuals. In addition, this information restricts hypothesis space. SYNGIP then re-evaluates all individuals by returning to step 2 and repeats this procedure until the maximal generation number defined by the user is reached. It then outputs horn clauses with the maximal fitness value.

3.2 Data Miner with SYNGIP

The data mining system named DAMS follows the SYNGIP approach and it is an adaptation of the SYNGIP algorithm. DAMS adopts the criteria defined in WARMR[2] for rule evaluation in the context of first order logic.

In association rule learning, two criteria evaluate a rule, namely support and confidence. According to the framework of WARMR, support and confidence in the context of first order logic are summarized as follows: When a rule is applied to the examples, Table 2 is obtained.

Support and confidence are defined by the following formulae.

Table 2: True and False Table

	positive example	negative example	sum
cover	a	b	a+b
not cover	c	d	c+d
sum	a+c	b+d	a+b+c+d

1. Create an initial population $M(\phi)$ using refinement operators.
2. Calculate the fitness value for the current population $M(t)$.
3. Move the top n% of the elite population that has a high fitness value into the next population $M(t+1)$.
4. Generate the remainder of the next population $M(t+1)$ by selecting two individuals by the tournament method and performing a crossover.
5. Randomly perform a mutation for next population $M(t+1)$.
6. $t = t+1$. Return to step 2.

Figure 1: SYNGIP algorithm

$$Sup = \frac{a}{a+b+c+d} \quad (1)$$

$$Conf = \frac{a}{a+b} \quad (2)$$

DAMS uses a fitness function based on the above definitions of support and confidence. Since the Equation 1 implies that the variable range of support is not from 0 to 1, the support was normalized in the following manner:

$$Sup_{normal} = \frac{Sup}{Sup_{max}} = \frac{\frac{a}{a+b+c+d}}{\frac{a+c}{a+b+c+d}} = \frac{a}{a+c}$$

The fitness function can then be obtained as follows:

$$Fitness_{normal}(I) = Sup_{normal} \times Conf = \frac{a}{a+c} \times \frac{a}{a+b} \quad (3)$$

This fitness function implies that the more support and confidence that are present, the higher the fitness function will become.

4 Experiment

4.1 Settings

A computational experiment using DAMS was performed to obtain useful knowledge from medical databases. The databases were those designated as a discovery challenge at PKDD-99[6] and were collected from patients with collagen related diseases at a Japanese university hospital. The experimental data were composed of three databases in tables. Table A, contains basic information about patients. Table A consists of seven attributes including those with multi-values. Table B, contains laboratory results of tests for thrombosis, which is an important complication. This table has thirteen attributes that also include those with multi-values. Table C, contains laboratory examination data. Since laboratory tests were repeated, the table has multiple records for individual patients obtained at various times between 1980 to 1999 and therefore has forty-four attributes.

The data tables were preprocessed through four steps. At first, patient records were selected from Tables A and B when the same patient appeared in both tables. In this manner, 379 patients were selected. Examination records relevant to the selected patients were selected from Table C. The tables were modified by deleting some attributes that had no meaning or by fixing values that were not readable by DAMS. Finally background knowledge was constructed using those tables. The background knowledge also contained horn clauses that can handle time intervals.

The target predicate was whether or not the patient had thrombosis. The number of positives, that is the number of patients with thrombosis, was 63 and the number of negatives was 316.

The major DAMS parameters were set as follows. The maximal number of clauses for an individual was 10, the maximal number of literals a clause can contain was 8, the

population size was 200, the mutation rate was 20 percent and the generation gap was 0.9. The termination criterion was set to achieve 1000 generations.

4.2 Results

Since all results cannot be explained in the allotted space, we introduce only three of the rules obtained by DAMS.

- *If a diagnosis is APS or suspected APS and the patient has taken a LAC examination, then the patient suffers thrombosis.*

  Among the 379 patients, 34 satisfied the antecedent of this rule, and 23 of those were actually affected by thrombosis. The rule held for 36.5 percent of the patients with thrombosis. These statistics imply that the rule has high support but not so high confidence. In a data mining context, rules with high support are useful. However, a rule like that obtained by DAMS cannot be induced by normal ILP systems because the rule lacks perfect confidence.

- *If a patient has an examination for A2PI at intervals of over 3 months, then the patient has thrombosis*.

  Among the 379 patients, 8 patients satisfied the antecedent of this rule, and 7 of those 8 were actually affected by thrombosis. This rule held for 12.7 percent of the patients with thrombosis and had 87.5 percent confidence.

- *If the ANA pattern is D and the patient has a PLT examination at intervals of over 3 months and the results show a high value in former examination, then the patients suffers thrombosis.*

  Two patients satisfied the antecedent of this rule. This rule had 100 percent confidence. However, this rule is hard to handle with a propositional learning system because it includes multiple relationships.

Of course, DAMS can also induce rules with high confidence like that learned by normal ILP systems. These results were shown to a physician who evaluated some of them as valid.

5 Discussion

The main difference between normal ILP systems and DAMS is that the latter is controlled by support and confidence criteria, whereas normal ILP systems are controlled only by confidence criteria. The purpose of normal ILP systems is to find clauses that entail only positive examples. Hence, normal ILP systems are designed only with confidence criteria. As a result, they cannot find clauses having high support like those that DAMS can handle.

Although normal ILP systems cannot handle support criteria, some non-monotonic ILP systems can. One of such system is WARMR[2]. This system utilizes support and confidence criteria. However, the method used in WARMR must determine in advance the level of minimal values for support and confidence. In addition to requiring these values, it also asks for the specification of candidate literals in a fixed form. In contrast, our system does not have such restrictions. DAMS can induce association rules in the first order without specifying minimal values for support and confidence. Furthermore, DAMS does not require predefined candidates and clauses with variable length literals can be handled.

Our system adopts missing value as non positive data which is treated as negative data. This treatment is obviously incorrect in the view of logic. Some missing values could be positive also. Hence, when an attribute contains many missing values, the system possibly does not find out a rule related to the attribute because of low confidence.

6 Conclusion

In the present paper, a new method for treating the data mining task and a new data mining system called DAMS were proposed. The performance of DAMS was tested experimentally.

The experimental results show that DAMS can induce knowledge from medical databases. It can be used for mining types of knowledge that cannot be handled by existing methods. Since the method used in DAMS is not specialized for application solely to medical databases, the system can also be applied to other types of data mining tasks to provide good solutions.

Despite our encouraging results, several research possibilities remain open in the fitness function. Our fitness function is based on support and confidence criteria. It

should be mathematically analyzed to see if the function is the best. In addition, there could exist another criteria for interesting rules such as exception rules. Extensions to the fitness function need to be investigated. We plan to investigate these extensions in our future work.

Reference

[1] Adriaans, P.; Zantinge, D., "Data Mining", London, Addison Wesley, 1996.

[2] Dehaspe, L.; Raedt, L. D., "Mining Association Rules in Multiple Relations", *Proceedings of the 7th International Workshop on Inductive Logic Programming*, 1297 of LNAI, Lavrač, N.; Džeroski, S., ed., Berlin, Springer, 1997, pp. 125-132.

[3] Ichise, R., "Synthesizing Inductive Logic Programming and Genetic Programming", *Proceedings of the 13th European Conference on Artificial Intelligence*, ECAI-98, Prade, Henri, ed., Chichester, John Wiley & Sons, pp. 467-468, 1998.

[4] Lavrač, N.; Džeroski, S., "Inductive Logic Programming Techniques and Applications", New York, ELLIS HORWOOD, 1994.

[5] Lavrac, N.; Džeroski, S.; Numao, M., "Inductive Logic Programming for Relational Knowledge Discovery", *New Generation Computing*, 17(1), pp. 3-23, 1999.

[6] "Discovery Challenge Guide to the Medical Data Set", *PKDD-99*, 1999, (http://lisp.vse.cz/pkdd99/tsumoto.htm).

[7] Shapiro, E. Y., "Algorithmic Program Debugging", Massachusetts, MIT Press, 1983.

[8] Tsumoto, S., "Rule Discovery in Large Time-Series Medical Databases", *Proceedings of Principles of Data Mining and Knowledge Discovery: Third European Conference (PKDD-99)*, 1704 of LNAI, Żytkow, J. M.; Rauch, J., ed., Berlin, Springer-Verlag, pp. 23-31, 1999.