From The Discovery Challenge on Thrombosis Data

Katsuhiko Takabayashi[a], and Shusaku Tsumoto[b]

[a] *Chiba University, School of Medicine, Chiba*
1-8-1 Inohana, Chuou-ku, Chiba-shi, Chiba, 260-0867, Japan
Tel: +81-43-226-2346, Fax: +81-226-2373, E-mail: takaba@ho.chiba-u.ac.jp
[b] *Shimane Medical College*
89-1 Enya, Izumo-shi, Shimane, 693-8501, Japan
Tel: +81-853-23-2111, Fax: 81-853-20-2170, E-mail: tsumoto@computer.org

Abstract

Although data mining promises a new paradigm to discover medical knowledge from a database, there are many problems to be solved before real application is feasible. We had the chance to provide a data set to be analyzed as a discovery challenge by using various data mining techniques at the PKDD conference. As data providers, we evaluated and discussed results and clarified problems.

Keywords:

Data mining, preprocessing, database

Introduction

For the 5th European Conference on Principles and Practice of Knowledge Discovery (PKDD) which was held in Freiburg, Germany in 2001, we provided medical data of patients with collagen diseases and received five analyses by using different data mining techniques. From this trial we learned much about being a data provider..

Approach and Methods

Data set Data mining techniques

The data set is composed of three databases, Tsum-A,B and C. In Tsum-A there are demographic data of patients with or suspected of having anti-phospholipid syndrome. Tsum-B is produced from a patient database containing 7 specific laboratory and other data of thrombosis for 806 cases. Seventy-six patients had some thrombotic events in their clinical courses. In Tsum-C there are temporal laboratory data of 1241 collagen disease patients including patients with thrombosis. It contains 41 different items for a total of 57,543 laboratory data collected from the main database of the hospital information system over a period of 17 years.

The data set was preprocessed by us and opened on the web site of PKDD. No medical knowledge was provided except the normal value of lab data. The declared goals were to find rules for detecting patients with thrombosis and to find specific relations between the date of thrombosis and changes in laboratory test values.

Data mining techniques and results

There were five applicants to this challenge of data discovery and they each submitted an article before presentation in Freiburg.

Boulicaut and Cremilleux [1] used delta-strong classification rules, and many rules with 100% confidence, but most of them were not useful. We learned that introduction to medical knowledge is necessary before using this data mining technique, otherwise too many rules will be produced and analyzed by domain experts.

Coursac et al [2] applied genetic programming and mentioned that they could predict the health state from the data in 99.28% of the cases. Werner and Fogarty [3] utilized similar genetic programming and found the lab data as sufficient for determining the discriminate function to identify thrombosis. However, this discriminate function is too complicated to be applied by physicians.

Zytkow and Gupta [4] identified the patterns in a data set by SQL queries and contingency tables. They mentioned that they obtained the same results as Infozoom as well as additional medically reasonable results. However, some of these were simply definitions which we had defined. We should have made our definitions and categorizations clearer to avoid such misunderstandings. Still, these interactive methods between users and computers are useful to help to avoid the feeling of using a black box and seem to produce reasonable results.

Jensen et al [5] analyzed by using the cross-industry standard process (Clementine) for data mining and offered interesting suggestions. However, it was difficult to predict time of thrombosis from the temporal data.

Discussion

As a data provider and a domain expert, we could judge the

results only on the basis of whether they could be explained by current medical knowledge. From a medical point of view, results are classified as common sense results that can be used as a positive control, probable results, possible results, unclear results that are difficult to evaluate, and nonsense results that serve as negative control. An important medical discovery may be lurking somewhere between common sense and nonsense results, but finding it is problematic. For instance if most of the results from a data mining technique are nonsense results, domain researchers are apt to consider the rest of the unclear results as nonsense as well. On the other hand they cannot say for certain that unclear results are true even if another data mining results show many good accordance with current knowledge. Then they can investigate these results with their own conventional prospective method in medicine.

It is difficult for humans to interpret a complicated discrimination function even though it may have very high confidence. Though most physicians probably use many variants of similar complicated algorithms in their brains unconsciously, they are reluctant to accept complicated rules from a computer. As with findings for expert systems in the last two decades, experts may ignore findings generated by a black box simply because they cannot comprehend them. Therefore intimate communication is very important between domain researchers, computer scientists, and developers of interactive tools. Interactive tools such as Infozoom and Clementine permit the users to try data mining in various ways at each step until they find good results or a promising trend and seem very useful if medical experts can use them freely. Since this was a challenge of computer scientists, there was no contact between the data provider and applicants before the conference. This trial was different from the normal usage of data mining techniques. Currently, contact between end-users and computer scientists is essential even if they use an interactive tool. As an example of this, the day after the conference we met with one of the applicants and tried using their interactive tool again with the same database, and we were easily able to obtain more interesting results in a short time.

Another important key is in cleaning or preprocessing. One of the problems in cleaning a data set is the standardization of the data. The target of data mining in medicine is usually huge and continues to grow with the prevalence of electronic patient records. When we collect data from different databases however, we must confront differences in representation of the same item between different databases. Lab data have different values in different hospitals, and even in the same hospital they are not always measured using the same method. Also when data are defined or categorized by more than two physicians, there are often differences in interpretation, classification and granularity. This process of carefully comparing and assessing data between databases or even in the same database and then altering them is a difficult but inevitable step in obtaining reasonable and confidential results. This data cleaning is not only necessary in the beginning but also throughout the course of data mining repeatedly. In addition, it is important that computer scientists have some medical knowledge until physicians can use the data mining technique freely.

Conclusion

In the near future, data mining techniques will be important tools to analyze and discover new medical knowledge. However, at the present time close cooperation between medical experts and computer scientists is still necessary to develop data mining techniques for use in clinical medicine.

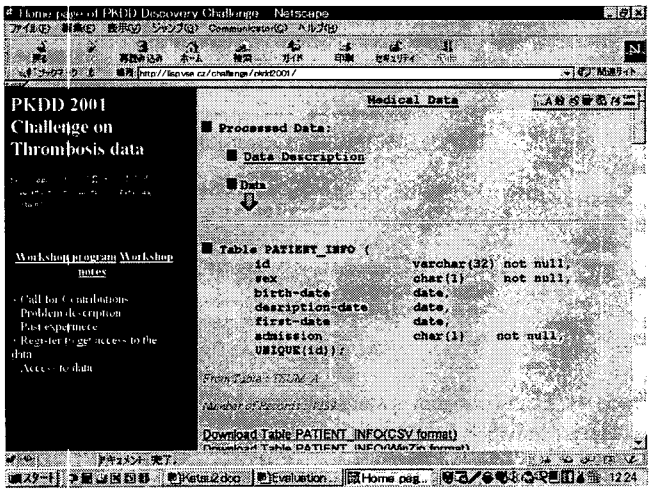

Figure 1 The explanation of this challenge on web site before the meeting.

Three databases, tsum-A, B, C were provided after preprocessed. (by courtesy of Prof Berka P)

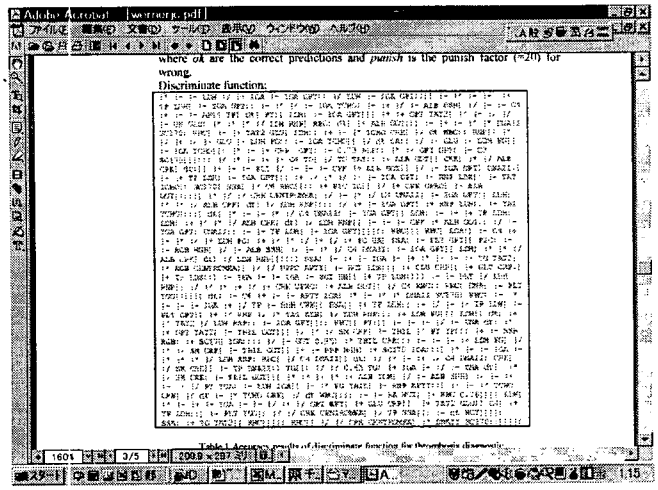

Figure 2 – the complicated discrimination function

Though it has high confidence to identify thrombosis patients, it is too complicated to be explained. We should have int-roduced medical domain to this applicant.

(Courtesy of Dr. Werner J)

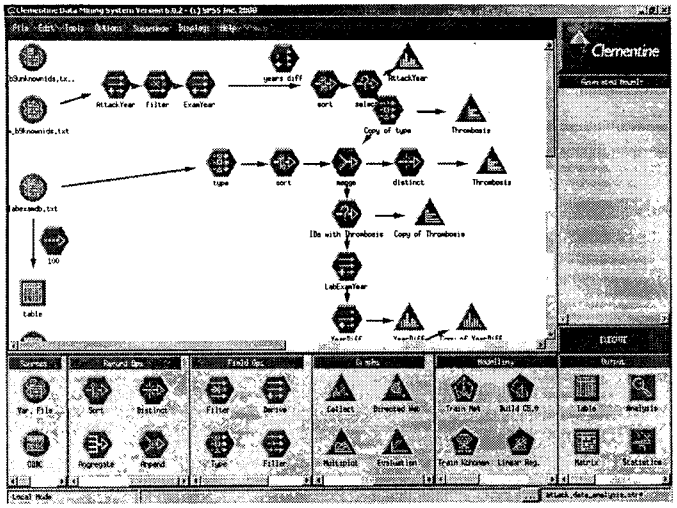

Figure3 Clementine

With this interactive tool we tried to find the relation between thrombosis attack and changes in lab data values.

(Courtesy of Dr. Jensen S)

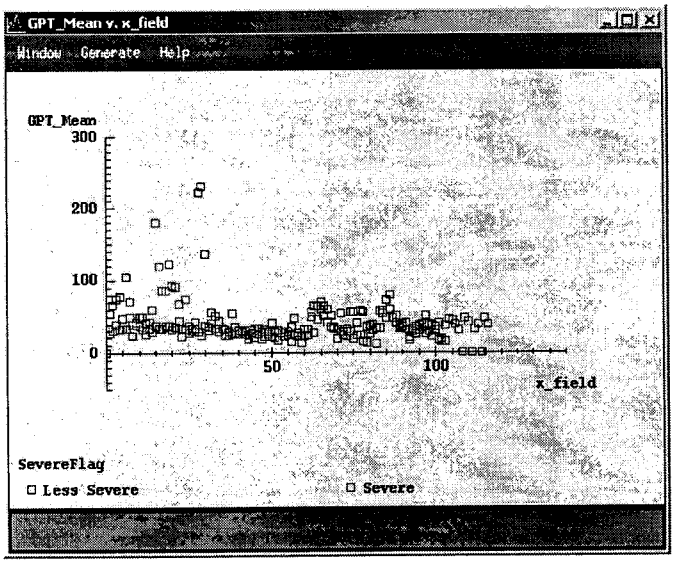

Fig 4 Relationship between GOT GPT and thrombosis dates

Peaks of GOT and GPT are observed in the time course. (Courtesy of Dr Jensen S)

Acknowledgments

We would like to show our gratitude for Professor Pert Berka and all the applicants for their cooperation to this discovery challenge of thrombosis data of PKDD. We would really like to express our regret over the loss of Prof Zytokow who started to hold this fruitful challenge.

References

[1] Jean-Francis Boulicaut and Bruno Cremilleux. Delta-strong classification rules for predicting collagen diseases. Lecture Notes in Artificial Intelligence 2168,2001.

[2] Ivan Coursac, Nicolas Duteil and Noel Lucas. PKDD 2001 Discovery Challenge-Medical Domain. Lecture Notes in Artificial Intelligence 2168,2001.

[3] James Cunha Werner and Terence C. Fogarty. Genetic programming applied to Collagen diseases and thrombosis. Lecture Notes in Artificial Intelligence 2168,2001.

[4] Jan Zytokow and Shishir Guputa Mining Medical Data using SQL Queries and Contingency Tables. Lecture Notes in Artificial Intelligence 2168,2001.

[5] Susan Jensen Mining Medical Data for predictive and sequential patterns. Lecture Notes in Artificial Intelligence 2168,2001.