Discovery Challenge from Temporary Data of Thrombosis

[a]Katsuhiko Takabayashi, [a]Hideto Yokoi, [b]Shoji Hirano, [b]Shusaku Tsumoto
[a]Division for Medical Informatics, Chiba University, School of Medicine, Chiba, Japan
[b]Division for Medical Informatics, Shimane Medical College, Izumo-shi, Shimane, Japan

Abstract

We have provided medical database in discovery challenge for three years and evaluated the results obtained from several researchers from the medical viewpoint. From these experiences we summarized the characteristics of the medical database to analyze data mining, particularly in preprocessing and the different way of thinking in pursuing results between mathematicians and medical experts.

Keywords:
Data Mining, Temporal Data, Pretreatment

Introduction

Although data mining promises a new paradigm to discover medical knowledge from a database, there are many specific problems to be solved before real application is feasible in each field. We had the opportunity to provide a data set to be analyzed as a discovery challenge by using various data mining techniques at the PKDD conference for three years. From this experience as data providers, we learned many aspects of using medical data and analyze them here.

Approach and methods

In the challenge for discovery, we provided the database of a rheumatological disease called anti-phospholipid syndrome (APS) and a list of episodes of thrombosis in them. The declared goals of this discovery 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 from the data set.

The data set comprises three databases. They contain temporal laboratory data of 1241 collagen disease patients including those with thrombosis with 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 PKDD web site. No medical knowledge was provided except the normal value of laboratory data.

Data mining techniques and results

There were eight applicants for this discovery challenge of thrombosis data and each submitted an article before presentation.

Boulicaut and Cremilleux [1] used delta-strong classification rules and discovered many rules with 100% confidence, but most of them did not make sense from the medical point of view. 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 sufficient for determining the discriminate function to identify thrombosis. However, this discriminate function was too complicated to have physicians understand.

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. There are 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 the data using the cross-industry standard process (Clementine) for data mining and offered interesting suggestions. However, it was difficult to predict the time of thrombosis from the temporal data.

Discussion

There are two major aspects from which to consider the application of data mining tools in medicine.

First, we have to consider the specific pretreatment and preprocessing of data from the hospital database. Hospital data includes many characteristic problems to be analyzed, as shown below. :

1) The clinical data are not obtained routinely but occasionally according to the condition of the patients, which may produce some bias. Some patients have much data in the same period and we must create a balance among them.
2) During the clinical course, modification of the treatment and other changes may influence the results.
3) The standard values of the data sometimes changed over a long period and it is impossible to adjust all the data by applying regression formula more than twice.
4) For clusterizing the laboratory data, the significant changing range of the data value should be elucidated in order to avoid producing many meaningless rules.

One of the problems in cleaning a data set is the standardization of the data. Sometimes we have to merge the data from different databases, and we must confront differences in representation of the same item between different databases. Laboratory data have their own measurement and normal value ranges in different hospitals, and even in the same hospital over a long period. Not a few data mining researchers deduced the wrong rules because they could not clusterize the data properly as they cannot distinguish the significant change of those form noise. Interpretation of temporal changes is now a major topic in treating medical data. i

Second, and the more important thing is that data mining researchers have to think about what kinds of results the

users expect. :
1) The data that are expected by medical experts are not the rules with a good certainty factor that cover a small number of the cases, but results that have a statistically significant difference.
2) Unlike mathematics or physics, in medicine there are many uncertain facts in different stages mingled with clear evidences.
3) It is sufficient if doctors obtain simple results, while too complicated ones beyond their understanding or those that are not statistically significant are not practical.

From a medical perspective, 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 a negative control. An important medical discovery may be lurking somewhere between common sense and nonsense results, but to find it is problematic. If most of the results from a data mining technique are nonsense results, domain researchers do not trust other rules. To avoid this we should introduce essential medical knowledge before using such a data mining technique. On the other hand, even if most of the data mining results show a good correlation with current knowledge, the experts cannot say for certain that the remaining unclear results are also true. They can, however, investigate, these results with their own conventional prospective method in the next step. Another important point is the different way of thinking between mathematicians and medical doctors. Certainty factor which many computer researchers always pursue has no meaning if it covers a small number of cases, even though it is very high, while in medical science whether a statistically significant rule or fact is much more important and a rule needs a bigger population for that.

It is difficult for humans to interpret a complicated discrimination function even if it may have very high confidence. As is evident from the results of expert systems in the last two decades, experts may ignore findings generated by a black box simply because they cannot comprehend them. Therefore, close communication is very important among domain researchers, computer scientists, and developers of interactive tools. Interactive tools such as Infozoom and Clementine permit users to try data mining in various ways at each step until they obtain good results or discover a promising trend, and it seems very useful if medical experts could use them freely. Until the specific methods are developed to apply in medicine and the medical scientists can have the tools to use freely for their data mining, close cooperation between medical experts and data mining researchers is necessary to apply data mining in this field. A good example of this is that just one discussion after this conference made it possible to quickly obtain more interesting results. Medical study currently requires a prospective way or Cohort study as a good study design, which means a carefully planed experiment. If we think of a long-term experiment lasting over 10 years, however, it is impracticable to use a prospective study from the beginning. Retrospective studies are expected for these experiments and data mining techniques will have a major role to play in this filed to discover many hypotheses.

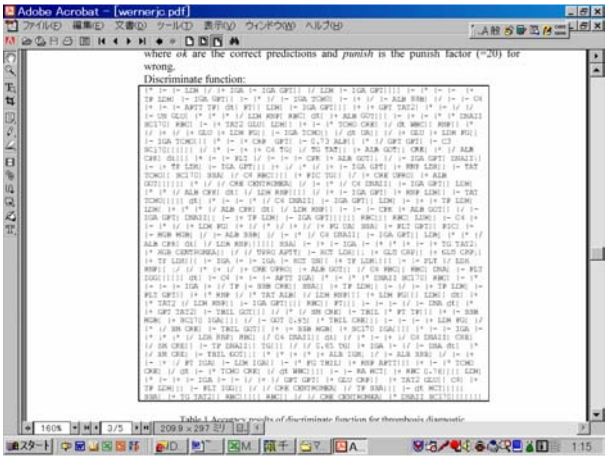
Conclusion

We mentioned the problems and solutions peculiar to medicine in applying data mining techniques. From the experience of discovery challenge, we hope to find hints to develop general tools and approaches applicable in medicine.

References

1 Boulicaut JF and Cremilleux B. -strong classification rules for predicting collagen diseases.
Discovery challenge on thrombotic data, PKDD, 2001.
2 Coursac I, Duteil N and Lucas N. PKDD 2001 Discovery Challenge- Medical Domain. Discovery challenge on thrombotic data, PKDD, 2001.
3 Zytkow J and Gupta S. Mining Medical Data using SQL Queries and Contingency Tables. Discovery challenge on thrombotic data, PKDD, 2001.
4 Werner JC and Fogarty TC. Genetic programming applied to collagen diseases and thrombosis.
Discovery challenge on thrombotic data, PKDD, 2001.
5 Jensen S. Mining medical data for predictive and sequential patterns. Discovery challenge on thrombotic data, PKDD, 2001.
6 Spenke M. Visualization and interactive analysis of blood parameters with InfoZoom. Artif Intell Med 22:159-72, 2001

Figure 1 The complicated discrimination function
Though this formula has high confidence to identify thrombosis patients, it is too complicated to be explained. (Courtesy of Dr. Werner J)


Adobe Acrobat - [wernerjc.pdf]
ファイル(F) 編集(E) 文書(D) ツール(T) 表示(V) ウィンドウ(W) ヘルプ(H)
160%
3/5
209.9 × 297 ミリ
スタート
1:15

where *ok* are the correct predictions and *punish* is the punish factor (=20) for wrong.
Discriminate function:
[illegible]

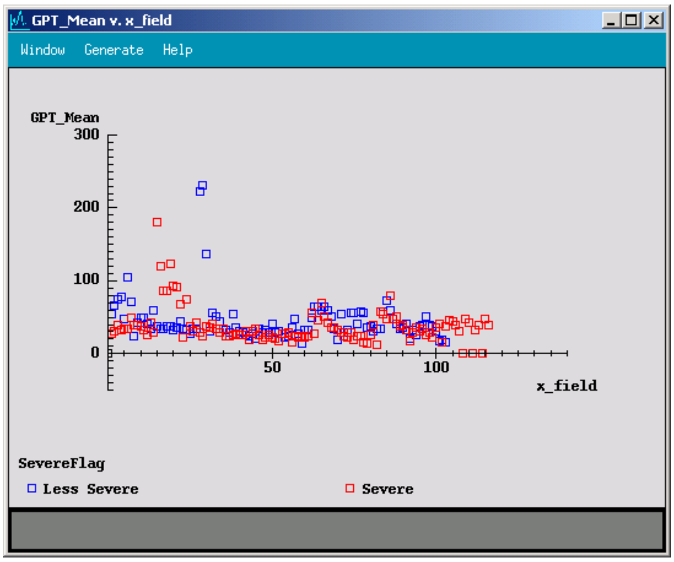
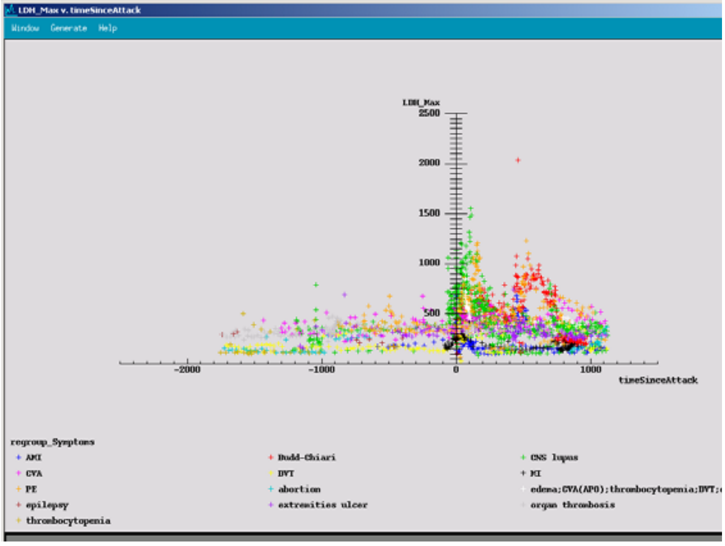


Fig 2 Relationship between GOT, GPT, and thrombosis dates

Peaks of GOT and GPT are observed in the time course.

(Courtesy of Ms. Jensen S)



(Courtesy of Ms. Jensen S)

Figure 3 Relationship between LDH change and thrombosis

Thromboses are classified by diagnosed diseases and indicated in different color. Each thrombosis disease has characteristic change in the time course. This unique discovery was one of the results after discussion between a data mining specialist and a medical expert.