Medical (Thrombosis) Data Description

Jan M. Żytkow[1], Shusaku Tsumoto[2] and Katsuhiko Takabayashi[3]

[1] Computer Science Dept. Univ. of North Carolina, Charlotte, N.C. 28223. USA
[2] Dept. of Medical Informatics, Shimane Medical University
Izumo-city, Shimane 693-8501 Japan
[3] Dept. of Internal Medicine, Higashi Matsudo Hospital
Matudo-city, Chiba 270-2222 Japan
e-mail: zytkow@uncc.edu ; tsumoto@shimane-med.ac.jp ; takaba@ho.chiba-u.ac.jp

Abstract. Collagen diseases are ofter dangerous and can be lethal. A severe complication common to those diseases of auto-immune system is called thrombosis. It occurs when coagulation of blood clogs blood vessels. Data relevant to the analysis of patients with collagen diseases have been donated to the PKDD Discovery Challenge in the hope that the discovered knowledge will illuminate the mechanisms responsible for collagen diseases and will help to diagnose and predict attacks of thrombosis. Discovery Challenge at PKDD-99 in Prague brought preliminary results, but it seems that the data offer a potential for much more knowledge. We describe a number of improvements made to the data, and their role in pursuing knowledge useful to doctors. We also describe a number of challenges caused by unconventional values recorded by physicians. The enhanced data are available for PKDD-2000 Discovery Challenge.

1 Medical problems for knowledge miners' attention

Collagen diseases are disorders of auto-immune system. Patients generate anti-bodies which attack their own bodies. That may result in a loss of life, when anti-bodies paralyze the organ where they develop. For example, if a patient generates anti-bodies in lungs, (s)he will chronically lose the respiratory function and finally will lose life. Little is known about the mechanisms responsible for those diseases and their classification is still fuzzy. Some patients may generate many kinds of anti-bodies and their manifestations may include all the characteristics of collagen diseases.

In collagen diseases, thrombosis is one of the most important and severe complications and one of the major causes of death. Thrombosis is an increased coagulation of blood which clogs blood vessels. Usually it lasts several hours and can repeat. It has been found that this complication is closely related to anti-cardiolipin antibodies. This was discovered by physicians, one of whom donated the dataset to discovery challenge.

Thrombosis must be treated as emergency. It is important to predict the possibility of its occurrence. It is also important to detect that it occurred and to capture temporal patterns specific and sensitive to attacks of thrombosis.

Doctors are moreover interested in classifying collagen diseases and in temporal patterns specific and sensitive to each collagen disease.

2 The raw data: PKDD Challenge 1999

The Challenge data may be a source of answers to such questions. The data were collected at Chiba University Hospital. For the 1999 Discovery Challenge they were organized into three tables, that we named TSUM_A.CSV, TSUM_B.CSV, TSUM_C.CSV (for simplicity we will skip the extensions .CSV). The tables can be connected by the ID number unique for each patient.

Each patient first came to the Hospital's Outpatient Clinic on collagen diseases, as recommended by a home doctor or a general physician in a local hospital. The primary data on the patient were recorded at that time. TSUM_A table consisted of approximately 1240 records and contained that information. The table was defined in detail by Tsumoto (1999). Besides ID the attributes included sex birthday, the first date when patient's data were recorded, the date when the patient came to the hospital, whether the patient was admitted to the hospital or followed in the outpatient clinic. The last attribute was DIAGNOSIS. This was a multi-valued attribute and upon closed examination the values turned out to belong to several categories, only some directly related to collagen diseases. In section on multi-valued attributes we will discuss the treatment of DIAGNOSIS.

The table TSUM_B included special results obtained in the Laboratory on Collagen Diseases. The data were input by doctors. They only include the patients who underwent those special tests. The data include patient ID, examination date, concentrations of three anti-cardiolipin antibodies (IGG, IGM, IGA), anti-nucleus antibody concentration (ANA), ANA patterns (a multi-valued attribute), three measures of degree of coagulation (KTC, RVVT, LAC). One attributes described degree of thrombosis, while two other multi-valued attributes described diagnosis and symptoms. The problems with multi-valued attributes are similar to diagnosis in TSUM_A.

The examination date was frequently close to the date of thrombosis, but upon closer inquiry it turned out that some of the patients suffered multiple attacks of thrombosis, and many of relevant data were not included in TSUM_B. We will present the new data in section on data enhancements.

The third table, TSUM_C, included ordinary laboratory examinations, one record per one date of the tests. Distinct attributes permit storage of values of 42 specific tests recorded. ID is a foreign key to TSUM_A and TSUM_B. Many records with dates that stretch over a long time are available on some patients, raising a possibility of time-series analysis.

Backround knowledge available on attributes in TSUM_C included the range of normal values of each test and the meaning of each test described in one or a few words.

3 Brief history of the challenge on thrombosis data

Three contributions were made to the September 1999 Prague challenge chaired by Petr Berka (Beilken & Spenke, 1999; Levin et al. 1999; Taylor, 1999). Also in September 1999, four contributions were made to a workshop in Japan chaired by Shusaku Tsumoto (Ichise & Numao, 2000; Nakamoto, Yoshida & Suzuki, 2000; Negishi, Suyama & Yamaguchi, 2000; Tsukada, Inokuchi, Washio & Motoda, 2000). The contributions are interesting, but results are preliminary. The most interesting results were obtained by Beilken and Spenke's Infozoom, which captures not only reasonable rules from TSUM_B, but also very interesting temporal patterns of laboratory tests before the thrombosis episode. Compared with their results, although other rule induction methods obtain reasonable results from TSUM_B, they are not capable to induce interesting temporal patterns.

4 Data enhancements: PKDD Challenge 2000

The past challenges demonstrated that multi-relational and multi-valued data are difficult for knowledge miners. Tools are not available and problems go beyond traditional tasks of KDD. Many problems are presented by string-valued attributes. Upon closer inspection, additional data can provide new information essential for doctors' business problems.

4.1 Odd values; string values

Consider values such as "> 107" for the predominantly numerical attribute PLT in TSUM_C. Many attributes in TSUM_C include such values. They are allowed since data types are strings rather than numbers. We can understand the convenience of the value "> 107" when the test is not exact. But this value is hard to compare with numerical values. String values allow neither the use of number ordering, nor other numerical relations.

Unfortunately, there is no quick solution. The normal values of PLT are between 100 and 400, so we can include "> 107" into the normal range, but any detailed number assignment may cause significant error. On the other hand a combination of numerical and non-numerical values impedes the use of many knowledge discovery tools.

Reluctantly, the number 108 was entered as replacement for "> 107", and the same convention was used in other cases. The original data are also available, so that any Challenge participant can revise the homogenization policy.

String format caused other problems, too. It is vulnerable to misspelled values, different spacing in disease names, and other non-essential changes. While some values could be easily identified by commas ("SLE, PM, PSS"), many cases required help from database provider, for instance `"ANA¤$BM[@-$N$_¤(B"` and `"Spleen infarction+R[-784]C, PH,thrombophlebitis"`

The same diagnosis occurred under different names, such as

```
CHRONIC EB
CHRONIC EB VIRUS INFECTION
CHRONIC EBV
CHR EB
```

Value identification is a case-by-case effort that is especially helpful when the number of records with a particular string value is small, so that by recognizing the same values records can be grouped and significance of findings can improve.

4.2 Multi-valued attributes

Together, three multi-valued attributes in TSUM_A and TSUM_B were replaced by relational tables two diagnoses and one ANA PATTERN attribute. Actually, the values were single strings, but many strings included multiple values. In the process of separation of individual values, many were determined identical, and were represented by the same name.

4.3 Meaning of diagnosis

Upon inspection, different values of diagnosis turned out to belong to different categories. The values indicated not only collagen diseases but also other diseases and various observations and symptoms. It was important to create a table that identifies the category of each value of DIAGNOSIS, since the main focus of the data was collagen diseases.

4.4 Temporal information missing on thrombosis attacks

The guide to 1999 TSUM_B says that the tests in the Laboratory on Collagen Diseases were related to thrombosis attacks, so that the date of attack and date of the exam were similar. But this can be true only to a degree. For instance, the thrombosis attribute indicates that many patients did not suffer from thrombosis, so the exam recorded in TSUM_B may not follow a thrombosis attack. Second, doctors know that some patients suffered more than one attack. Upon closer investigation it turned out that the dates of multiple attacks can be retrieved from hospital database. Symptoms observed during each attack are also available and may be useful. The new TSUM_B includes up to four attacks per patient, which is the maximum number registered for any single patient. Each attack is described by date and symptoms observed during the attack.

The data on up to four attacks, each on a specified date enable a better use of tests in TSUM_C. Now we can distinguish data relevant to prediction of thrombosis: those are test results prior to the onset. We can also distinguish the data that can lead to detection of a past attack of thrombosis: they include test that follow the attack.

Now, TSUM_C can be JOINed with the new TSUM_B on records selected by their relevance to prediction or to diagnosis of an attack. Tests before an attack can be compared to tests after the attack. It is always important to compare such tests with a control group of patients who did not suffer thrombosis.

4.5 Summary of the PKDD-2000 Medical Challenge tables

The enhanced thrombosis data were placed into seven tables.

PATIENT_INFO is the rest of TSUM_A, after DIAGNOSIS was put into a separate DIAGNOSIS table. PATIENT_INFO contains 1239 records.

```
PATIENT_INFO
  id
  sex
  birth-date
  description-date
  first-date
  admission
```

DIAGNOSIS includes values of diagnosis from TSUM_A and TSUM_B, and separates them to a single value per record. It includes 1942 records.

```
DIAGNOSIS
  id
  disease           % single value of diagnosis
  diag              % suspected vs. confirmed
  fn_test           % from what table
```

Information from TSUM_B was distributed in four tables, including DIAGNOSIS (see above, for a multivalued attribute), ANA_PATTERN (multivalued attribute, 656 records), OTHER_SYMPTOM (all thrombosis attacks, 195 records) and the remainder was left in SPC_EXAM (801 records).

```
OTHER_SYMPTOM
  id
  attack-date
  osymptom
  attack-number

ANA_PATTERN
  id
  pattern           % P: peripheral, S:speckled, N: nucleolar,
                    % D: discrete speckled
SPC_EXAM
  id
  exam-date
  aCL_IgG           % anti-Cardiolipin antibody IgG concentration
  ...               % etc. 7 other attributes as in TSUM_B
```

TSUM_C was included without changes as LAB_EXAM (57542 records). Additional table DISEASE categorizes the values of diagnosis as collagen diseases, non-collagen diseases, observations and no diagnosis (65 records)

```
DISEASE
  disease-name
  disease-type
  comments          % varchar(64) - space for comments
```

5 Conclusion

We augmented the 1999 data with information relevant to doctors' business problems. Better data on the thrombosis attacks, especially on their dates allows us better utilization of laboratory tests in TSUM_B. Categorization of values of the diagnosis attributes is necessary to focus on collagen diseases. Conversion from multi-valued attributes to additional relational tables facilitates the use of popular relational DBMS such as Access or Oracle.

Many individual values were improved in the 2000 data. Japanese strings of characters were converted to English. For the discovery systems which disqualify entire records when one value is missing, that is a big data enhancement.

Many attributes contained values in the category of > 20. They must reflect imprecise results of some tests. Incomplete dates belong to the same category. They are meaningful for diagnosis of individual patients, but they form serious problems for knowledge discovery as it is difficult to use them jointly with regular numerical values. Such values were transformed to numbers, at the cost of distorting the data.

Since data transformation may cause data distortion, we provided access to both old (raw) data and new data, so that the each participant in the challenge can decide on data preparation.

References

1. Tsumoto, S. 1999. Guide to the Medical Data Set. In: Berka P. ed. *Workshop Notes on Discovery Challenge, PKDD-1999, Prague, Sep.15-18*, Univ.of Economics, Prague, p.45-47.
2. Beilken, C. & Spenke, M. 1999. Visual, Interactive Data Mining with InfoZoom –the Medical Data Set. In: Berka P. ed. *Workshop Notes on Discovery Challenge, PKDD-1999, Prague, Sep.15-18*, Univ.of Economics, Prague, p.49-54.
3. Ichise Ryutaro & Numao Masayuki. 2000. Knowledge Discovery from Medical Database with Multistrategy Approach. *Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence*, p.1-4.
4. Levin, B., Meidan, A., Cheskis, A., Gefen, O. & Vorobyov, I. 1999. PKDD99 Discovery Challenge – Medical Domain. In: Berka P. ed. *Workshop Notes on Discovery Challenge, PKDD-1999, Prague, Sep.15-18*, Univ.of Economics, Prague, p.55-57.
5. Nakamoto Kazuki, Yoshida Mieko & Suzuki Einoshin, 2000. Analysis of Collagen-Disease Data Set Based on KDD Process Model, *Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence*, p.9-153.
6. Negishi Naoya, Suyama Akihiro & Yamaguchi Takahira. 2000. Automatic Composition of Inductive Applications to Collagen Diseases Database Using Inductive Learning Method Ontologies. *Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence*, p.5-8.
7. Taylor, C. 1999. PKDD'99 Discovery Challenge: Medical Data Set. In: Berka P. ed. *Workshop Notes on Discovery Challenge, PKDD-1999, Prague, Sep.15-18*, Univ.of Economics, Prague, p.59-64.
8. Tsukada Makoto, Inokuchi Akihiro, Washio Takashi & Motoda Hiroshi. 2000. Discretization of Numerical Attributes on Structured Data for Basket Analysis, *Proceedings of SIG-FAI/KBS-9902, Japan Assoc. of Artificial Intelligence*, p.17-24.

This article was processed using the LaTeX macro package with LLNCS style