Proceedings of the 7th International Conference on Applied Informatics
Eger, Hungary, January 28–31, 2007. Vol. 1. pp. 261–267.

Data mining based on medical diagnosis

József Demeter, Barnabás Szász

University of Debrecen
e-mail: demeter@dote.hu, bszasz@gmail.com

Abstract

Our project is intended to investigate relations between laboratory examination results (mainly hematology and chemistry) and determined diagnoses at hospital cases using different methods of symbolic data mining. The work is carried out in the frame of DIP-CRC project owned by the University of Debrecen with several involved commercial partners. As source database we use the database provided by the HIS system (MedSolution) used in the Medical and Health Science Center of University of Debrecen, which is a real-life medical database, containing a huge number of patient cases in multiple medical areas. Laboratory results and diagnoses are extracted from the database for approximately 100,000 hospital cases, together with some auxiliary patient data like sex or age. Several data mining algorithms are tested on the input dataset, which will be prepared in multiple ways. Algorithm implementations provided by CORON toolkit and developed at Loria Laboratory, France, were used for the KDD process. The project is aiming to extract frequent as well as rare association rules in form of {interpreted laboratory values + auxiliary patient data} $\rightarrow$ {diagnose codes/groups}.

Keywords: CORON, Data mining, KDD, medical diagnosis, rules

1. Introduction

Extensive amounts of knowledge and data stored in medical databases require the development of specialized tools for storing and accessing of data, data analysis, and effective use of stored knowledge and data. The increase in data volume results difficulties in extracting useful and new information for decision support. The gap between data generation and data comprehension is widening in all fields of human activity. In medicine, overcoming this gap is particularly crucial since medical decision making needs to be supported by arguments based on basic medical knowledge as well as knowledge, regularities and trends extracted from data. This paper focuses on narrowing the increasing gap between data gathering and data comprehension. [5, 6, 3]

Knowledge Discovery in Databases (KDD) is frequently defined as a process consisting of the following steps: understanding the domain, forming the dataset and cleaning the data, extracting of regularities hidden in the data thus formulating knowledge in the form of patterns, rules, etc. Important role has the input data selection, which is collecting the relevant data from the source, usually real-world database. As the result of this step we have the rough input data, which has to be prepared to be suitable for data mining. In general different data conversion tasks are needed too. The input database of data mining – we call it also the context – in fact is a binary data table. In the rows are objects which correspond to real-world entities of the examined area. In the columns there are properties of the objects. Each property (which can be called an item too) could be present or absent for a certain object. This way we have true and false values in the input data table. The following steps in the overall KDD process are usually referred to as data mining (DM), post processing of discovered knowledge, and exploiting results. [7]

2. Database understanding

KDD is typically concerned with knowledge extraction from very large datasets. The database used for our research includes inpatient and outpatient cases as well as diagnostic examination data of a wide range of medical specialties conducted at the Medical and Health Science Center of The University of Debrecen. During 16 years a huge amount of raw data has been accumulated (ca. 900,000 patients with 9 million cases and 10 million diagnostic examinations), what was a rich base of the 63,184 selected inpatient hospital cases from the past 3 years. In the first phase of the research only the first cases were extracted from inpatient case chains, which are bundles of more than one subsequent cases.

The over sixty-thousand selected case has the following values:

- hematology and chemistry laboratory examination results (coded as low / normal / high values)
- diagnose codes associated with the cases (using the ICD-10 international standard of diagnose codes)
- auxiliary patient data: sex and age of the patient.

3. Data preparation

Data preparation is an important step because data itself may have been collected in an ad hoc manner, unfilled fields in records may be found, or mistakes in data entry may have been made. As a result, the KDD process cannot succeed without a serious effort of preparing the data. Without the data discovery phase, the analyst will have no idea if the data quality can support the task at all. Once the quality and details are assessed, serious work is usually needed to get the

data in shape for analysis. Preparation for mining involves looking at the variables individually as well as looking at the data set as a whole.

3.1. Laboratory data

The selected laboratory examination data contains values for 29 different hematology and 32 chemistry laboratory examinations, we call these keywords. We have numerical examination values which can interpreted as normal, low or high values according to the normal reference range set for the respective keyword. Having three possible values for a keyword we need binary properties. Two conversion methods could be used. We can assign a property for each value resulting normal, low and high properties or only a single property to the keyword-false if the value is normal and true if it is abnormal, in other words low or high values.

Typical issue is not all the keywords have values for each object (inpatient case). These absent values can be treated two ways. The first method is to skip these values, which means to set each related property to false (note that this method can not be chosen if we use a single property for a keyword). The second method is to treat these values as normal values, assuming that these examinations were not ordered because their values are probably normal.

In addition we have determined a minimal threshold percentage for the keywords. The keywords available for too few objects were completely dropped. Originally there were 187 different keywords and the limit was set to 1000 object.

3.2. Diagnoses

In the queried data set a number of 4302 different diagnoses occur. This is a noticeable subset of the ICD-10 diagnose set which contains a total number of approximately 11,500 diagnose codes. It has to be mentioned that to a certain inpatient case there can be assigned several different diagnose. The ICD-10 diagnose coding system uses codes built from one letter and four digits. The code set has a hierarchical structure. Considering only the first letter and the first 2 or 3 digits we get diagnose groups.

Exploiting the hierarchical structure we can assign binary properties to diagnoses by two ways. We can assign one property to each diagnose code. Note that this will result a large number of properties. Or we can assign properties to diagnose groups getting this way a more manageable number of properties. Truncating to 3 digits does not have a big effect, it results in 4167 diagnose groups, but truncating to 2 digits is more effective, since this way we get only 1952 diagnose groups.

3.3. Auxiliary patient data

Beside of keywords and diagnoses we put into the input database properties for auxiliary patient data, namely the sex and the age of the patient. To the sex two properties are assigned, one for the male and another for the female. Handling the age of the patient needs more properties. We establish multiple age intervals then

we assign one property to each interval. For a certain object exactly one from these properties will be true.

3.4. File format

As result of the data preparation process the source data is available in form of 3 simple text files with the following structure:

1. The main data file contains medical information and has the following structure:

```
<case seq.no.>|<age>|<sex>|<laboratory value 1.>|...
|<laboratory value n.>|<diagnose code>{[|<diagnose seq.no.>]}
<sex> := "M" or "F"
<laboratory value> := "L", "N", "H" or empty
```

2. The keyword definition file gives the definition of the keywords, it is a table with the following columns:

- keyword seq.no. -Ű this defines the position of the keyword in the main data file
- keyword category (hematology or chemistry)
- keyword code
- keyword description

3. The diagnose definition file contains a list of the diagnose codes that can be found in the main data file. Available diagnose data:

- diagnose seq.no – this number is referred in the main data file
- diagnose code
- diagnose description

The code and description are from the official ICD-10 diagnose code system. The file in fact a subset of the ICD-10 code set.

4. KDD Algorithms

An item set consist of multiple items which are together present at multiple objects. In other words an item set is the common set of properties contained by a given set of subjects. The main property of an item set is its support. The support of the item set is equal to the number of objects which contain the respective items. The support can be expressed in two ways: the absolute support is the number of objects, while the relative support is the proportion of the number of objects to the

total number of subjects in the input database. We say an item set to be frequent if its support is greater or equal than a certain value called the minimal support.

An association rule is a relation between two item sets A and B. The rule means if item set A is present at an object than item set B is also present with a given plausibility.

The two main properties of the association rules are the support and the confidence. The support of an association rule is equal to the support of the union of the two participant item sets. The confidence of an item set we call the proportion of the support of the rule and the support of the item set on the left side, called the antecedent. An association rule is valid if it has at least the minimal support and the confidence is grater or equal to a given value called minimal confidence. [1]

- support: $\text{supp}(A \to B) = \text{supp}(A\ B)$
- confidence: $\text{conf}(A \to B) = \text{supp}(A\ B)/\text{supp}(A)$
- valid association rule: $\text{supp}(A \to B) \geqslant \text{min\_supp}$ and $\text{conf}(A \to B) \geqslant$ minimal confidence (min_conf) [2].

To find frequent association rules we use the Zart algorithm developed to find a lossless representation subset of all valid association rules, the set of MNR rules. Rare association rules identification is time and process demanding task, mainly in case of the huge number of records. However the BtB algorithm is designed to find the MRG rules, a subset of all rare association rules. Both algorithms were developed recently implemented in the CORON toolkit, a domain independent, multi-purpose data mining platform including a large number of symbolic data mining algorithms and auxiliary tools. The toolkit is product of the Loria Laboratory, France. [8, 9]

5. Finding association rules

Properties set up for keywords, diagnoses and auxiliary patient data are put together into a binary table, the input database. Using different property determination methods described for keywords and diagnoses we can prepare multiple different input databases which can result in different association rules. Association rules are extracted from the input database using appropriate data mining algorithms. Principally we are interested in extracting valid association rules, generated from frequent item sets. Besides of these also rare but confident association rules can be interesting regarding to rare but serious diseases.

The resulted set of association rules are filtered, and only those rules are retained where the antecedent item set contains only keyword and/or auxiliary patient data related properties and the consequent item set contains only diagnose properties.

Since the diagnoses follow a hierarchical structure, it allows extracting semantic relationships from the rule-set that might be not explicit rules. We except rules

with a diagnose group on the left side and value intervals on the right side. A rule {diagnose} → {N, H, VH, L, VL} can be transformed into {diagnose} → {N, H, L} or {diagnose} → {N, A} A rule {diagnose group} → {value} can be transformed into {diagnose (subgroup)} → {value} Using this background information we can save some round of algorithms run. Diagnose groups are determined by ICD-10 code system. Using ontology as source of relations later it will be possible to define new type of relations additionally to the ICD-10 hierarchy. [4]

6. Future perspectives

For the present project we use an input database containing only MHSC data of only the last 3 years. The data mining process can be extended to the whole dataset of more than 16 years, however processing this amount of data needs much more computation time.

Future plans include developing an expert system based on these rules which will be capable to propose diagnoses with different probability for individual patient cases based on laboratory examination results. The base knowledge of the expert system consists of the association rules extracted during data mining. The planned medical diagnostic expert system is intended to act as a tool for doctors, which helps determining diagnoses by suggesting multiple possible diagnoses. The suggestion is made based on hematology and general chemistry laboratory examination results.

The background database of the system will contain the definition of the supported laboratory examinations (the keywords) as well as the definition of diagnoses and diagnose groups. In addition for laboratory examination value interpretation the laboratory reference ranges and conversion tables for the different measurement units used are also needed.

And last but not least the extracted association rules may contain so far undiscovered relations between laboratory examinations and diagnoses so it can be a good idea to examine in depth these rules from scientific point of view involving medical experts.

References

[1] d'Aquin, M., Badra, F., Lafrogne, S., Lieber, J., Napoli, A., Szathmary, L., Case Base Mining for Adaptation Knowledge Acquisition, *In Proceedings of the 20th International Joint Conference on Artificial Intelligence – IJCAI '07*, Hyderabad, India, (2007), 750–755.

[2] Agrawal, R., Imielinshki, T., Swami, A., Mining Association Rules between Sets of Items in Large Databases, *In Proceeding of ACM-SIGMOD International Conference on Management of Data*, (1993), 207–216.

[3] Cios, K., Moore, G., Uniqueness of medical data mining, In *Artificial Intelligence in Medicine*, (2002).

[4] Doddi, S., Marathe, A., Ravi, S. S., Torney, D. C., Discovery of Association Rules in Medical Data, In *Med Inform Internet Med.*, (2001), 26, 25–33.

[5] Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., Hammond, W. E., Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse, (1997).

[6] Rao, R. B., Rosales, R., Niculescu, S., Sriram, K., Bogoni, L., Zhou, X. S., Krishnapuram, B., Mining Medical Records for Computer Aided Diagnosis, *Siemens Medical Solutions*, Malvern, PA, USA, (2006).

[7] Stilou, S., Bamidis, P. D., Maglaveras, N., Pappas, C., Mining Association Rules from Clinical Databases: An Intelligent Diagnostic Process in Healthcare, In *Medinfo*, (2001), 10, 1399–1403.

[8] Szathmary, L., Mithodes symboliques de fouille de donnies avec la plate-forme Coron, *PhD Thesis, Universiti Henri Poincari – Nancy 1*, France, (2006).

[9] Szathmary, L., Napoli, A., CORON: A Framework for Levelwise Itemset Mining Algorithms, In *Supplementary Proceedings of the Third International Conference on Formal Concept Analysis – ICFCA '05*, Lens, France, (2005), 110–113.