Study of Hepatitis Data Using Visual Data Mining System D2MS

Saori Kawasaki, Duc Dung Nguyen, Trong Dung Nguyen, and Tu Bao Ho

Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa 923-1292 Japan

Abstract. The hepatitis data set has been decided as the common dataset for Active Mining Project with a careful privacy policy. It has been collected during 1982-2001 by the Chiba university hospital and consists of 6 sets of data. This paper reports our preliminary study of this hepatitis dataset using our developed visual data mining system D2MS (Data Mining with Model Selection) that supports users to play a more active role in the knowledge discovery process visualizing tools. We first describe our early solution of data preprocessing, and a derived data table for the purpose of studying the course difference between B and C hepatitis with D2MS. We then report some preliminary results of this processing. Finally, we address our limitations in this first trial, and what we think to do in coming time.

1. Introduction

Hepatocellular carcinoma (HCC) is the most common type of liver cancer, is the fifth most common cancer in the world. About three quarters of the cases of HCC are found in Southeast Asia (China, Hong Kong, Taiwan, Korea, and Japan). HCC is also very common in sub-Saharan Africa (Mozambique and South Africa). The exact cause of HCC is unknown. Viruses such as hepatitis B and hepatitis C have been shown to increase the risk of HCC [1].

Extensive amounts of data gathered in medical databases require specialized tools for data analysis and effective use of data. Medical informatics may use the methods developed in the new interdisciplinary field of knowledge discovery and data mining (KDD), encompassing statistics, databases, machine learning, and visualization tools to support the analysis of data and the discovery of useful patterns/models that are encoded within the data.

The hepatitis data set, collected during 1982-2001 by the Chiba university hospital and consists of 6 component data sets, has been selected as the common dataset for the Active Mining Project. This dataset presents many challenging features to the data mining research. For example, the data are not in the flat form suitable for most data mining systems, the data are temporal, mixture of numerical and categorical attributes, with many missing values, etc.

This paper reports our preliminary study of this hepatitis dataset using our developed visual data mining system D2MS (Data Mining with Model Selection) [2], [3] that supports users to play a more active role in the knowledge discovery process with visualization tools. In section 2, we briefly describe the dataset and system D2MS. Section 3 described our early data preprocessing, and a derived data table for the purpose of studying the course difference between B and C hepatitis with D2MS. Section 3 reports some preliminary results of this processing, and in section 5 we address our limitations in this first trial, and what we think to do in the next time.

2. The Hepatitis Dataset and System D2MS

2.1 The Hepatitis Dataset

The hepatitis dataset contains data of patients who have had the viral hepatitis type B or type C and had liver biopsy during 1982-2001. Among types of hepatitis, hepatitis B and C have possibility to result to liver cirrhosis or hepatoma. The study tasks specified by doctors include finding (a) relation between the pathological (fibrosis) and lab data, (b) the course of HCC

(validity of current staging), (c) period until HCC (normally it is more then 20 years), (d) factors related to HCC, (e) course difference between B & C hepatitis, (f) treatment of interferon (effectiveness and lab data). Doctors have shared some common knowledge about this disease like "interferon treatment is effective for hepatitis C" or "the typical path to hepatoma is observed mostly among the type C". Also, as abnormal values of GOT and GPT, which indicate the progress of hepatitis, may clearly appear at the late stage of hepatitis, it is expected to clarify the quantitive relation between progress and GOT/GPT values.

The dataset consists of six tables of basic information for (a) 771 patients, (b) laboratory test items, (c) the results for each patient of each test item, (d) additional out-sourced test results, (e) biopsy test results, (f) and interferon treatment history. Each patient has its management id (MID) that is transformed from its original patient id for privacy protection. Which kinds of test one patient had in one time can be identified by MID, test date and test time. According to the structure of tables, each patient can be described with a set of time series of laboratory test results with including interferon treatment. However, there are many test items with too many missing values because each test consists of a small number of items.

How to deal with the amount of attributes is one of most important problem for this dataset. In order to put this data into our mining system D2MS described in subsection 2.2, we had to solve these problems: how to select/generate attributes to be analyzed from original attributes, how to describe one record, how to cleanse data. For simplifying those we use processed results with join operation by Dr. Yoshida from Osaka University, and we select 42 test items according to the report by Prof. Yamaguchi's group [7]. We then made transformation of data into suitable form to our tools D2MS as described in the section 3.

2.2 **System D2MS**

D2MS is a visual data mining system with visualization support for model selection [2], [3]. The emphasis on model selection comes from the complexity of the knowledge discovery process: it is an iterative and interactive one requiring many steps each can be done by different methods. The problem of *model selection*—choosing appropriate discovered models or algorithms and their settings for obtaining such models in a given application—is difficult and non-trivial because it requires empirical comparative evaluation of discovered models and meta-knowledge on models/algorithms.

The system called D2MS (Data Mining with Model Selection) provides the user with the ability of trying various alternatives of algorithm combinations and their settings. The quantitative evaluation can be obtained by performance metrics provided by the system while the qualitative evaluation can be obtained by effective visualization of the discovered models. Two main features of D2MS are its data mining methods and its visualization tools. The data mining methods consists of CABRO to learned decision trees [6], CABROrules [5] and LUPC [4] to learn prediction rules. These methods have been carefully evaluated and shown to be comparable to well-known methods. The visualization module is linked to most other modules in D2MS in particular those directly concerned with model selection. It currently consists of a data visualizer, a rule visualizer, and a tree visualizer (for hierarchical structures). These visualizers are integrated to most methods mentioned above in preprocessing, data mining, and postprocessing [3].

3. Preprocessing the Hepatitis Dataset

The original hepatitis dataset contains all available data observed from patients. The basic idea of data preprocessing is to obtain a derived dataset in the flat form appropriate to each study task and that can be processed by our current available tools in D2MS. Our preprocessing of hepatitis data according to the general scheme includes: (i) data cleaning, (ii) data integration, (iii) data reduction and transformation, and (iv) data discretization.

3.1 Data cleaning

Generally, this step requires to fill up missing values and to eliminate noisy data. In this attempt we do no investigation on measuring the inconsistency and noise in data but consider to

do it in next period. In a dataset with fixed attributes, how to deal with missing values may have a lot influence on the mining results. To fill missing values with global constant or other kinds of simple statistical values like mean and mode when there are too many, it may bias the data and results in the incorrect way. One of the best is to predict most probable values for each attribute according to the class each tuple belongs to and replace by them. Unfortunately we could not evaluate suitable values so that we remain those missing values this time. Also we remove unexpected characters "H" or "L" or others following numeric values, because they are redundant to be derived from values and the normal range definition.

3.2 Data integration

This is one of the main tasks of preprocessing this dataset. It consists of integrating six relational data tables into one data table suitable to our tools. In order to have data integration, we employed the results of Dr. Yoshida that he kindly provided us. The dataset from four tables (the patient information, laboratory test results, interferon treatment history and biopsy test results) are merged by attributes of MID, the test date and test time.

3.3 Data reduction and transformation

We employ the results of analyzing the frequency of attributes presented in [7]. According to this analysis the following attributes have been selected. The most frequent attributes include GPT, GOT, LDH, ALP, TP, T-BIL, ALB, D-BIL, I-BIL, UA, UN, CRE, オウダン, ニュウビ, ヨウケツ, LAP, G-GTP, CHE, ZTT, TTT, T-CHO. The highly frequent attributes include NA, CL, K. The frequent attributes include F-ALB, F-A2.GL, G.GL, F-A/G, F-B.GL, F-A1.G. The less frequent but significant attribute include F-CHO, U-PH, U-GLU, U-RBC, U-PRO, U-BIL, U-SG, U-KET, TG, U-UBG, AMY, CRP. Note that all of these attributes are numerical.

In order to be able to utilize our D2MS system (as well as many other existing data mining systems) to find out the relations between causes of hepatitis type B and type C, it is necessary to transform the original data into a data set that describe the information about each patient (the transformation of data observed on each patient in a sequence of days into one patient record). The class attribute contains the type of hepatitis B or C of each patient. Others descriptive attributes are those of selected tests. Because one patient may take a test many times then the values that describe information of this test (on one patient) should be summarized to characterize the test. The summarization should retain the information of the original data as much as possible. In our first work on the hepatitis data, we utilize some statistical measurements to do this conversion. For each patient the values of a test are transformed into three statistical values: the mean, standard variation, and the value called *trend*. Suppose that patient P take n times on the test T_i, the value of each test are T_{ij}, $j = 1,...,n$. The synthesized data about this test will be described by three values:

$$T_{ij}Mean = \frac{1}{n}\sum_{j=1}^{n} T_{ij}$$

$$T_{ij}StdVar = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(T_{ij} - T_{ij}Mean)^2}$$

$$T_{ij}Trend = \sum_{j=2}^{n}(T_{ij} - T_{ij-1})$$

This process results a data set in which each record contains synthesized data about one patient with a type of hepatitis.

3.4 Data discretization

Two kinds of discretization are done, one for derived attributes that have upper and lower thresholds of normality (e.g., synthesized mean value of each test), and one for derived attributes without background knowledge (e.g., synthesized variance and trend value). In the first case, the

conversion we applied is to replace each value by one of four intervals decided by five special values: the *minimum* and *maximum* values of the attribute (given fron the original dataset), and three *sample quartile* values. In the second case we use the entropy-based discretization.

4. Preliminary Study of the Hepatitis Data with D2MS

For this preliminary study, we focus on the classification problem of type B or C based on other attributes of the preprocessed dataset. In this section, we describe the experimental results with two classification algorithms CABRO and LUPC. Table 1 summaries the experimental results of CABRO and LUPC on the preprocessed dataset by 10-fold stratified cross validation. The size is the number of leaf nodes of a decision tree (CABRO), or number of rules (LUPC).

4.2. Experiment with CABRO

CABRO is a decision tree learning subsystem in D2MS based on rough sets [6]. Table 1 shows both sizes and error rates of output decision trees before and after pruning. A part of the pruned tree obtained from the whole dataset is showed below:

```
GPT-trend = <=- 679: B (29.0/1.4)
GPT-trend = - 679<:
|  LAP-variance = <=50:
|  |  GPT-variance = <=171:
|  |  |  K-mean = [-0.500;-0.500): C (0.0)
|  |  |  K-mean = [-0.500;-0.500): C (0.0)
|  |  |  K-mean = [-1.000;-0.500):
```

That tree contains 20 attributes: GPT-trend, LAP-variance, GPT-variance, K-mean, CHE-trend, T-BIL-trend, F-A/G-variance, F-CHO-mean, CL-mean, inf, LDH-trend, ZTT-variance, F-ALB-variance, LDH-variance, T-BIL-variance, GPT-mean, G-GTP-trend, U-PRO-variance, F-A2.GL-mean, and LDH-mean. In other word CABRO uses only these variable to classify type B or C. Our observation on that tree and trees obtained at individual trials in the 10-fold cross validation shows that GPT-trend, LAP-variance, inf, and GPT-variance may be the most important among the above attributes. Figure 1 shows that tree visualized by the tightly-coupled views in D2MS [2].

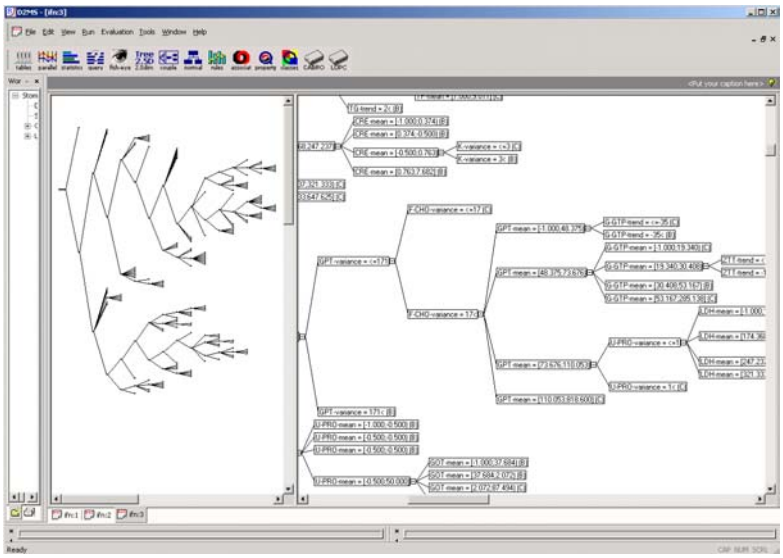

Figure 1: Discovered decision tree by CABRO that classifies hepatitis B and C

With the help of the tree visualizers, we can have the following observation from two most general levels of the decision tree:

- If GPT-trend ≤ -697, that means GPT decrease very quickly, type of hepatitis will be B with 4% support and 95% confidence.
- Otherwise, LAP-variance will be used for further classify:
  - If LAP-variance ≤ 50, that means LAP does not change much during the treating course, type largely will be C (51% support, and 85% confidence).
  - Otherwise, type largely will be B (44% support, and 85% confidence).

4.3. Experiment with LUPC

LUPC is developed to learn prediction rules from supervised data. Its performance depends on several parameter specified by the user: α for min accuracy of rules, β for min coverage of rules, γ for maximal number of candidate rules in the beam search, and η for maximal number of attribute-value pairs to be consider. By varying these parameters we can find different sets of rules [4].

If we use a trial with α = 98%, β = 3, γ = 200, and η = 50. From the whole dataset of 771 patients, LUPC discovered 79 rules characterizing the hepatitis B and 72 rules characterizing hepatitis C. Most of them cover from 5% to 15% of the whole patients (35 to 120 patients) and with accuracy (on training data) of at least 98%. Examples of these rules are

IF num-variance = (0, 1]
F-B.GL-variance = 2<
F-CHO-variance = 17<
LAP-mean = [118.370; 463.542]
NA-mean = [-1.000; 60.852)
TG-variance= 30<
U-PRO-variance = 1<
THEN Hepatitis B

IF CHE-trend = <=-10
CRE-variance = <=1
CRP-mean = [-1.000; 0.099)
F-A/G-variance = <=1
GPT-variance = <=171
LAP-trend = -115< <=114
T-BIL-variance = <=2
U-KET-mean = <=-1
THEN Hepatitis C

Figure 2 shows the above mentioned rule for class C that matches 120 patients of class C and only mismatches 2 patients belonging to class B. By using a stratified cross evaluation with LUPC, we obtained an estimation of average error rate for these rules when diagnosing unknown patients as 17.820% ± 4.933%.

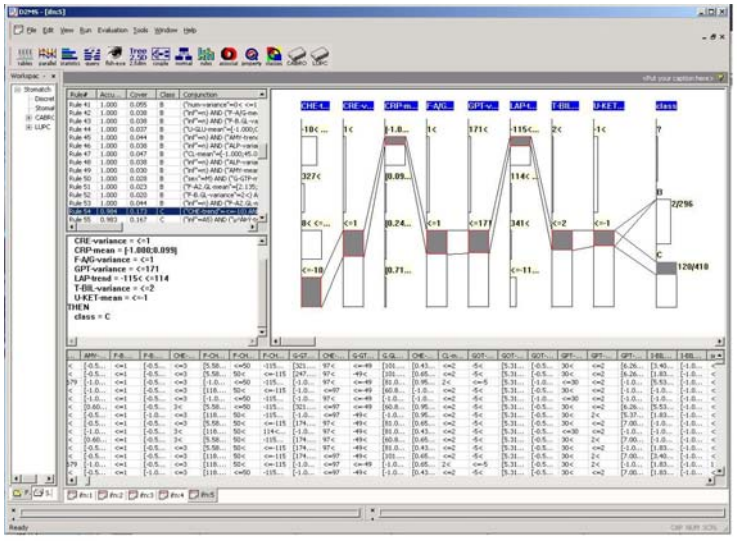

Figure 2: Discovered rules by LUPC that distinguish hepatitis B and C

Table 1. Experimental results

	Before pruning		After pruning	
	Size	Error rate	Size	Error rate
CABRO	194.4	18.1%	88.9	16.7%
LUPC (α = 75%, β = 3)			170	21.6%
LUPC (α = 98%, β = 3)			140	17.8%

5. Discussion and Conclusion

We have presented basic ideas and results of our first attempt in studying the hepatitis data. In our understanding, this early study of the hepatitis dataset has a number of limitations.

(a) In terms of preprocessing, the main research left to do in the near future is the treatment of temporal data. We plan to investigate the *qualitative analysis of trend* and apply the technique to detect the trend in the patient data. This processing certainly is important for the task of (b) to find the course of HCC and the task (c) to find the period until HCC.

(b) More general, we need to *model* the patient data when transforming temporal data of patients into several synthesized values that characterize the patient information.

(c) We need to find other derived data sets for other research purposes. The need of a tight *collaboration* between the doctors and data miners when learning trees or rule sets is indispensable so that background knowledge can be used to achieve and to improve the accuracy as well the utility of discovered knowledge.

(d) *Interactive and visual tools* are valuable in each step of the discovery process, especially they are very useful for us to understand and analyze the data and discovered knowledge. However, we also learn that many features of these tools need to be improved in order to meet several new requirement arose from this study.

References

[1] MedicineNet.com http://www.focusoncancer.com/script/main/art.asp?articlekey=1917&rd=1)

[2] Ho T.B., Nguyen T.D., Nguyen D.D., Kawasaki S., "Visualization of Data and Knowledge in the Knowledge Discovery Process", *Active Mining: New Directions of Data Mining*, H. Motoda (Ed.) IOS Press (in press).

[3] Ho, T.B., Nguyen, T.D., Nguyen, D.D., Kawasaki, S., "Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining", *International Journal of Artificial Intelligence Tools*, Vol. 10 (2001), No. 4, 691-713.

[4] Ho, T.B., Nguyen, D.D., Kawasaki, S., "Mining Prediction Rules from Minority Classes", *International Workshop Rule-Based Data Mining RBDM 2001*, Tokyo, October 20-22, 2001, 254-264.

[5] Nguyen, D.T., Ho, T.B., Shimodaira, H., "A Scalable Algorithm for Rule Post-Pruning of Large Decision Trees", *5th Pacific-Asia Conference on Knowledge Discovery and Data Mining PAKDD00*, April 2001, Hongkong. Lecture Notes in Artificial Intelligence 2035, Springer, 2001, 467-476.

[6] Nguyen, D.T., Ho, T.B., "An Interactive-Graphic System for Decision Tree Induction", *Journal of Japanese Society for Artificial Intelligence*, Vol. 14, N. 1, 1999, 131-138.

[7] 山口高平，,畑澤寛光，佐藤芳紀「慢性肝炎データセットのクレンジングとマイニングの試み」『情報洪水時代におけるアクティブマイニングの実現　平成１３年度科学研究費補助金　特定領域研究（B）研究成果報告書）』, 2002, 205-221.