Knowledge Visualization in Hepatitis Study

DucDung Nguyen, TuBao Ho, and Saori Kawasaki

School of Knowledge Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi city, Ishikawa 923-1292, Japan
{ducdung, bao}@jaist.ac.jp

Abstract

Evidence-based medicine (EBM) is a shift in which medicine from being based on individual experience of doctors to evidence with clear background. In this shift, data mining can play a significant role as its ability of uncovering medical evidence from large volumes of medical data. Recognizing the crucial role of visualization in discovering such evidences, this work presents some developed tools integrated in our data mining system D2MS for appropriately visualizing knowledge, and their usage in hepatitis study. We emphasize on our two rule visualizers, one for individual rule and the other for rule in its relations with the others.

Keywords: Data and knowledge visualization, hepatitis study, model selection.

1 Introduction

Knowledge discovery in databases (KDD) or *data mining* in short— the rapidly growing interdisciplinary field of computing that evolves from its roots in database management, statistics, machine learning, and others— aims at finding useful knowledge (patterns/models) from large databases. Each of five steps in the KDD process (i.e., (1) understanding the application domain, (2) data preprocessing, (3) data mining, (4) evaluation of discovered knowledge, and (5) applying discovered knowledge) requires interaction and many decisions of the user. In other words, the KDD process can be alternatively viewed as a process of model selection, i.e., that of choosing by the user the most interesting discovered patterns/models or choosing algorithms and their settings to obtain interesting patterns/models in an application. In doing such a task, visualization has an indispensable role because it helps to understand complicated patterns/models in addition to using performance metrics [1], [5].

Medicine is a traditional application domain of artificial intelligence. In the new medicine trend of shifting from being based on individual experience of doctors to being based on evidence with clear background (evidence-based medicine, or EBM), KDD has been much expected to contribute to EBM [4]. However, so far KDD is just at a doorstep for medical applications and still too difficult to use by physicians in their data analysis [3]. Thus, there is a practical need of developing user-friendly KDD tools to support analyzing medical data. In the last four years we have been involved in the projects on mining hepatitis data [9] for which we have developed some new visualization tools, as well as used them in supporting the discovery process for medical knowledge [8].

The role of visualization in medical data analysis has been addressed in several works [4], however, only a few techniques were particularly developed for medical applications. The purpose of this paper is to introduce some of these visualization tools and demonstrate how they can support medical data mining. These tools include visualizers for multidimensional databases, discovered rules, hierarchical structures as well a synergistic visualization of data and knowledge. These tools are integrated in our knowledge discovery system D2MS (Data Mining with Model Selection) [7], [8].

2 The Visual Data Mining System D2MS

D2MS (Data Mining with Model Selection) is the visual data mining system that we have developed [8] in the framework of our active mining project. Figure 1 shows a conceptual architecture of D2MS where Data Mining component currently includes a decision tree learning (CABRO [12]), and a rule learning (LUPC [7]) subsystems.

2.1 D2MS: User-centered system

The *interestingnes*s of discovered patterns/models is commonly characterized by several criteria: *evidence* indicates the significance of a finding measured by a statistical criterion; redundancy amounts to the similarity of a finding with respect to other findings and measures to what degree a finding follows from another one; *usefulness* relates a finding to the goal of the users; *novelty* includes the deviation from prior knowledge of the user or system; *simplicity* refers to the syntactical complexity of the presentation of a finding, and *generality* is determined by the fraction of the population a finding refers to. The interestingness can be seen as a function of the above criteria, and strongly depends on the user as well his/her domain knowledge.

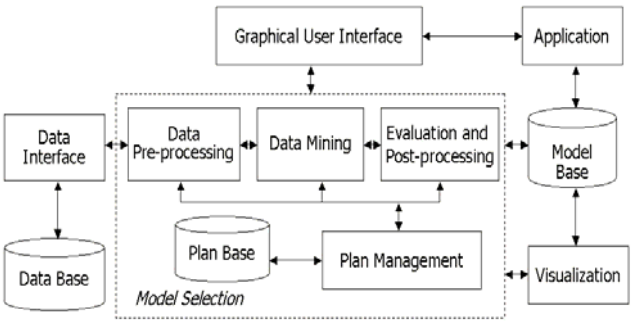

Figure 1: Conceptual architecture of the systemD2MS

The key idea of our solution to model selection in D2MS is to support an effective participation of the user in this process. Concretely, D2MS first supports the user in doing trials on combinations of algorithms and their parameter settings in order to produce competing models, and then it supports the user in evaluating them quantitatively and

qualitatively by providing both performance metrics values as well as visualization of these models (Figure 2).

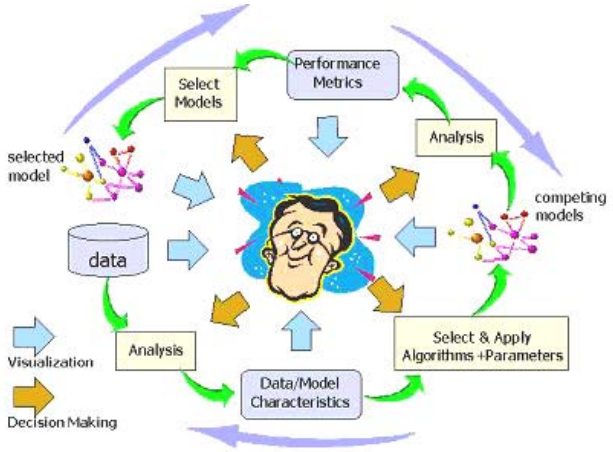

Figure 2: The idea of user-centered system in D2MS and the visualization support

We have chosen the parallel coordinates technique for visualizing 2D tabular datasets defined by n rows and p columns. D2MS improves parallel coordinates in several ways to adapt.

2.2 Knowledge Visualization

2.3.2 Rule visualizer-1

This visualizer aims to view individually a rule. A rule is a pattern related to several attribute-values and a subset of instances. The importance in visualizing a rule is how this local structure is viewed in its relation to the whole dataset, and how the view supports the user's evaluation on the rule interestingness. D2MS's rule visualizer allows the user to visualize rules in the form *antecedent* → *consequent* where *antecedent* is a conjunction of attribute-value pairs, *consequent* is a conjunction of attribute-value pairs in case of association rules, and is a value of the class attribute in case of prediction rules. A rule is simply displayed by a subset of parallel coordinates included in *antecedent* and *consequent*. The D2MS's rule visualizer has the following functions:

Each rule is displayed by a polyline that goes through the axes containing attribute-values occurred on the antecedent part of the rule leading to the consequent part of the rule that are displayed with different colour. In the case of prediction rules, the ratio associated with each class in the class attribute corresponds to the number of instances of the class covered by the rule over the total number of instances in the class, giving a view on the rule quality.

2.3.3. Rule visualizer-2

This aims to view a rule in its relations with others. We developed a graph-based rule visualization technique to support user in finding out interesting patterns.

There have been numerous rule visualization techniques using two dimension matrices (), grid view (), tree view (), and interactive mosaic plots. The common point of the technique is that they stress on visualizing individual rule, or the relation between the left-hand side (antecedent) and the right-hand side (consequent) of a rule. It is still very difficult to user to see which rule is potential new.

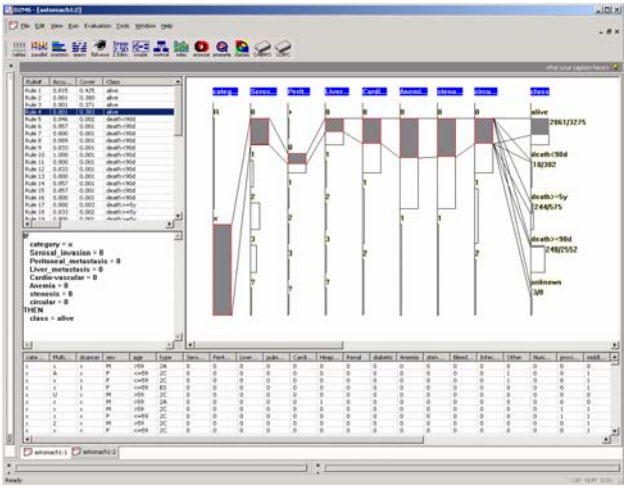

Figure 3: View an individual rule in D2MS: top-left window shows the list of discovered rules, the middle-left and the top-right windows show a rule under inspection, and bottom window displays the instances covered by that rule.

Our approach firstly starts from the question that which could be a good indication of a new rule. According to [13], a rule $A \rightarrow B$ is potentially new with respect to a given rule $X \rightarrow Y$ if

(a) B AND Y |= FALSE. (B and Y logically contradict each other)
(b) A AND X holds on a statistically large subset of tuples in dataset D.
(c) The rule A AND $X \rightarrow B$ holds (so the rule A AND $X \rightarrow \neg Y$ holds)

We also view that a rule $A \rightarrow B$ is an unexpected conclusion rule if A and X are similar but B and Y are very different. It is clear that visualizing individual rule could not help to reveal the about relations, but several rules must be viewed together. Our method is to construct a graph based on all rules discovered by data mining algorithms, then visualize the graph and focus on nodes that reflect above relation between precedent and consequence parts of a rule.

In our implementation, each condition (or an attribute-value pair) is represented as a node, and there will be an edge between a conclusion attribute-value pair with any attribute-value pair in the condition part. When focusing on one node, a conclusion or a condition, the magnetic-spring algorithm is applied to visualize a local part of the whole graph. Because we concern a rule and its related rules and conditions, the local part of the graph includes all nodes with the distance less than or equal 2 from focusing node. Figure 4 shows an example of viewing a rule.

2.3.4 Viewing rules and data

The subset of instances covered by a rule is visualized together with the rule by parallel coordinates or by summaries on parallel coordinates. From this subset of instances, the user can see the set of rules each of them cover some of these instances, or the user can smoothly change the values of an attribute in the rule to see other related

possible rules. These possible operations facilitate the user in evaluating the quality of this rule: a rule is good if instances covered by it are not recognized by other rules, and vice-versa (see later in 3.3). The rules for a class can be displayed together, and instances of the class as well of other classes covered by these rule are displayed.

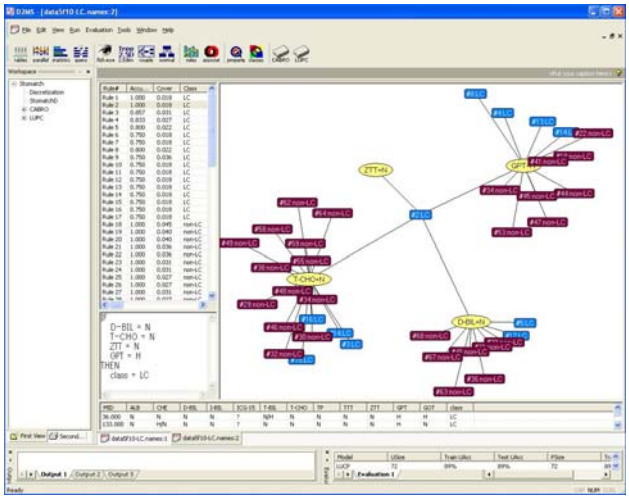

Figure 4: A rule viewed in its relations with others.

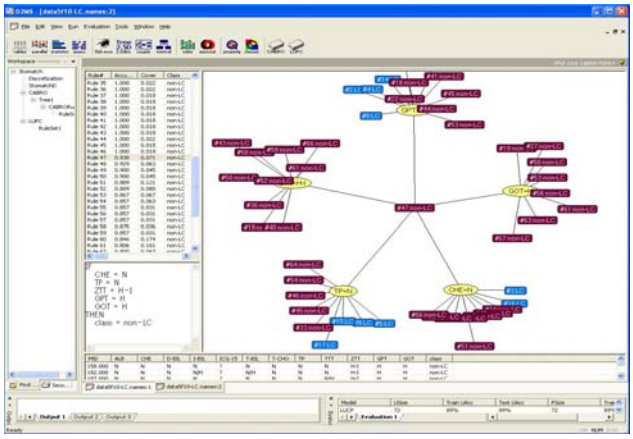

Figure 5: Visualization of rule No.2 for class LC

3 Visualization in mining hepatitis data

3.1 The data mining projects in hepatitis study

The hepatitis database is composed of 6 relational tables of time-series data on 983 laboratory examinations of 771 patients of hepatitis B and C. The data are broadly split into two categories. The first includes administrative information such as patient's information (age and date of birth), pathological classification of the disease, date of biopsy, result of biopsy, and duration of interferon therapy. The second includes temporal records of blood examination and urinalysis that can be further split into two subcategories, in-hospital and out-hospital examination data. In-hospital examination data contain the results of 230 examinations that were performed using the hospital's equipment. Out-hospital examination data contain the results of 753 examinations, including comments of staffs, performed using special equipment on the other facilities. Consequently, the temporal data contain the results of 983 types of examinations. The database is given recently to challenge the data mining research community. Among six problems posed by physicians, we focus on the following problems:

P1. Discover the differences in temporal patterns between hepatitis B and C.

P2. Evaluate whether laboratory examinations can be used to predict the stage of liver fibrosis. A variant of this problem is prediction of LC (liver cirrhosis, related fibrosis stages F3 and F4) and non-LC (related to fibrosis stages F0, F1, and F2).

P3. Evaluate whether the interferon therapy is effective or not.

Our mining methods of temporal abstractions to this database were described in detail in [9].

3.2 Visualization support in finding interesting rules in hepatitis study

We illustrate the use of rule visualization in support of interpreting and understanding discovered rules in hepatitis study, in particular the LC vs. non-LC problem.

Figure 6 shows a typical screenshot of rule visualizer2 in investigating this problem. A total of 22 rules were found for LC and 59 rules for non-LC with program LUPC (75% minimum confidence and 4 as minimum support count. Assume that we want to assess its interestingness, says, rule #47 of non-LC whose precedent is the conjunction of "CHE = N and TP = N and ZTT = H−I, GPT = H and GOT = H" as shown in Figure 6.

Rules of each class are displayed with a color, says, red for LC class and dark green for non-LC class, while attribute-value pairs are displayed with yellow color. The rule #47 non-LC is shown at the center in connecting with its four attribute-value pairs in the precedent part (Figure 6). Each attribute-value pair then linked to all rules where it appears in the precedent parts. We can easily observe that conditions "GPT=H", "TP=N", and "CHE=N" occurred

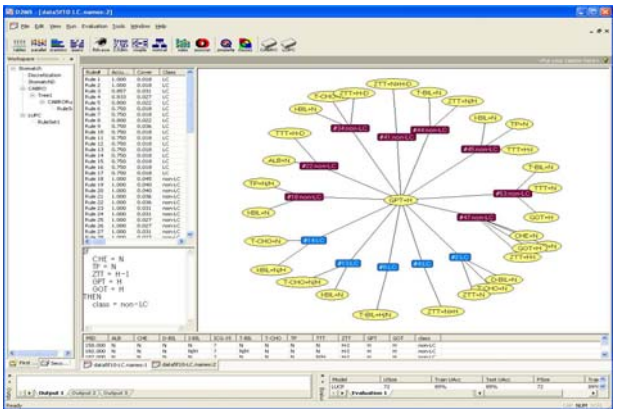

Figure 6: Rules that contain"GPT=H" in the precedent part, obtained by a double click on node "GPT=H" in Figure 10.

in both LC and non-LC classes while the other two are only in non-LC. By a double click on a attribute-value node in Figure 10, says, "GPT=H" we can see in another

screen (Figure 11) the links from this node to all the rules that contains "GPT=H" in the precedent part. In fact, we can switch between these two modes in rule visualizer-2 to observe each rule or attribute-value pair and its related conditions or rules, i.e., its neighborhood information. Many views of rule interestingness can be assessed with the visualizers. For example, conditions (a), (b) and (c) of view in [13] addressed in 2.3.3 can be assessed by rule visualizer-1 and rule visualizer-2. In fact, condition (a) can be applied to any two rules each describes one of two classes LC or non-LC; condition (b) can be accessed by the checking the sets of instances covered by the two rules shown in the window under the graph in Figure 10 (or by clicking the mode that shows the interaction of "*X* AND *A*"; (c) can be easily checked by using rule visualizer-1 for the rule "$A \rightarrow B$" as seen in Figure 6, then adding conditions X to A to check "X AND A $\rightarrow$B".

Various rules found by D2MS with its visualization tools from stomach cancer data and hepatitis data have been encouragingly evaluated by physicians and they really interested in the tools [7], [8], [9]. A number of rules found recently were judged to be potentially new and useful, for example:

R#10: NonLC: "GPT in very high state with peaks" AFTER "TTT in high state with peaks" AND "GOT in very high state with peaks" ENDS "GPT in very high with peaks" AND "GOT in very high state with peaks" AFTER "TTT in high state with peaks" (support count = 10, conf. = .80).

R#8: LC: "GPT in very high state with peaks" AFTER "TTT in very high state with peaks" AND "GPT in very high state with peaks" BEFORE "TTT in high state with peaks" AND "GOT in very high state with peaks" AFTER "TTT in high state with peaks", (support count = 8, conf. = .80)

R#42: NonLC: "IDH changed from normal to low state" BEFORE "TTT in normal state with peaks" AND "GOT in very high state with peaks" BEFORE "ZTT in high state with peaks" AND "GOT in very high state with peaks" AFTER "GPT in very high state with peaks" (support count = 6, conf. = .86).

4. Conclusion

We have shortly presented the visual data mining system D2MS. We emphasized the central role of the user's participation in the knowledge discovery process and have developed data and knowledge visualizers in D2MS, in particular rule visualizer-1 (for individual rules) and rule visualizer-2 (for rules and their neighbourhood information) to support such participation. The visualizers of D2MS have been used in mining stomach cancer and hepatitis databases and shown their usefulness.

References

1. Ankerst, M., Grinstein, G., Keim, D. (2002): Visual Data Mining: Background, Techniques, and Drug Discovery Applications, *Tutorial Notes of ACM SIGKDD'02*, 277-245.
2. Card, S. K., Mackinlay, J. D., Shneiderman, B. (1999): *Readings in Information Visualization*, Morgan Kaufmann.
3. Chittaro, L. (2001): Information Visualization and Its Application to Medicine, *Artificial Intelligence in Medicine* **22**, 81-88.
4. Cios, K.J. (2001): *Medical data mining and knowledge discover* (Ed.) Physica-Verlag.
5. Fayyad, U.M., Grinstein. G.G., and Wierse, A. (2002): *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann.
6. Han, J. and Cercone, N. (2000): RuleViz: A Model for Visualizing Knowledge Discovery Process, *Sixth Inter. Conf. on Knowledge Discovery and Data Mining ACM SIGKDD'00*, 244-253.
7. Ho, T.B., Nguyen, T.D., Nguyen, D.D., Kawasaki, S. (2001): Visualization Support for User-Centered Model Selection in Knowledge Discovery and Data Mining, *International Journal of Artificial Intelligence Tools*, World Scientific, **10**(4), 691-713.
8. Ho, T.B., Nguyen, T.D., Nguyen, D.D. (2002): Visualization Support for a User-Centered KDD Process, *ACM International Conference on Knowledge Discovery and Data Mining ACM SIGKDD'02*, 519-524.
9. Ho, T.B., Nguyen, T.D., Kawasaki, S., Le, S.Q., Nguyen, D.D., Yokoi, H., Takabayashi, K. (2003): Mining Hepatitis Data with Temporal Abstraction, *ACM International Conference on Knowledge Discovery and Data Mining ACM SIGKDD'03*, 24-27.
10. Kumar, H. P., Plaisant, C., Shneiderman, B. (1997): Browsing Hierarchical Data with Multi-Level Dynamic Queries and Pruning", *Inter. Journal of Human-Computer Studies*, **46**(1), 103-124.
11. Lee, H.Y., Ong, H.L., Quek, L.H. (1995): Exploiting Visualization in Knowledge Discovery, *ACM International Conference on Knowledge Discovery and Data Mining ACM SIGKDD'95,* 198–203.
12. Nguyen, T.D. and Ho, T.B. (1999): An Interactive Graphic System for Decision Tree Induction, *Journal of Japanese Society for Artificial Intelligence*, **14**(1), 131-138.
13. Padmanabhan, B. and Tuzhilin, A. (1999): Unexpectedness as a Measure of Interestingness in Knowledge Discovery, *Decision Support* Systems, 27(3), 303-318.