A USER-CENTERED VISUAL APPROACH TO DATA MINING

THE SYSTEM D2MS

Tu Bao Ho, Trong Dung Nguyen, Duc Dung Nguyen
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa 923-1292, Japan

Abstract: We present a human-centered approach to model selection in machine learning and data mining that emphasizes and facilitates the active participation of the user in the knowledge discovery process with quantitative and qualitative evaluation of patterns/models. The key idea of such a model selection is it would result from a combination of a quantitative evaluation of model characteristics and performance metrics with a qualitative evaluation of patterns/model by the user. We develop data mining methods integrated with visualization tools in the user-centered visual system D2MS (Data Mining with Model Selection). We finally present a case-study of D2MS in mining stomach cancer data.

Key words:data mining, user-centered, visualization, model selection.

1. INTRODUCTION

The problem of *model selection* in machine learning and data mining—choosing appropriate learned models or algorithms and their settings for obtaining such models in a given application—is non-trivial and difficult because it requires empirical comparative evaluation of discovered models and meta-knowledge on models/algorithms.

Finding interesting patterns/models hidden in data is the objective of data mining. The interestingness of discovered patterns/models can be seen as a function of four criteria: validity, simplicity, usefulness, and novelty. Among these components, data mining systems can judge only the validity and simplicity of discovered patterns/models (e.g., the accuracy and support of

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35602-0_35

M. A. Musen et al. (eds.), *Intelligent Information Processing*

rules), but only the user can evaluate their usefulness and novelty. However, it can be observed that the support for the user in evaluation has received inadequate attention in the data mining community [1].

Though visualization techniques have progressed dramatically in the last decade, it can be seen that visual knowledge discovery still remains in its infancy [2], [3], [6]. Most data mining systems do not yet take advantage of visualization techniques, or visualization techniques still usually reside at the beginning and at the end of the knowledge discovery process [4].

In this paper we present a user-centered approach to data mining with visualization support for model selection, and the visual data mining system D2MS (Data Mining with Model Selection) that provides visualization integrated into the KDD process. We illustrate the approach and D2MS by a case-study in medical application.

2. VISUALIZATION SUPPORT FOR USER-CENTERED MODEL SELECTION

Recognizing the importance of the user in the knowledge discovery process [1], we have focused on a user-centered approach to knowledge discovery. Fig. 1 illustrates the basic idea of this approach in which the user plays a central role in the knowledge discovery process with the support of visualization tools.

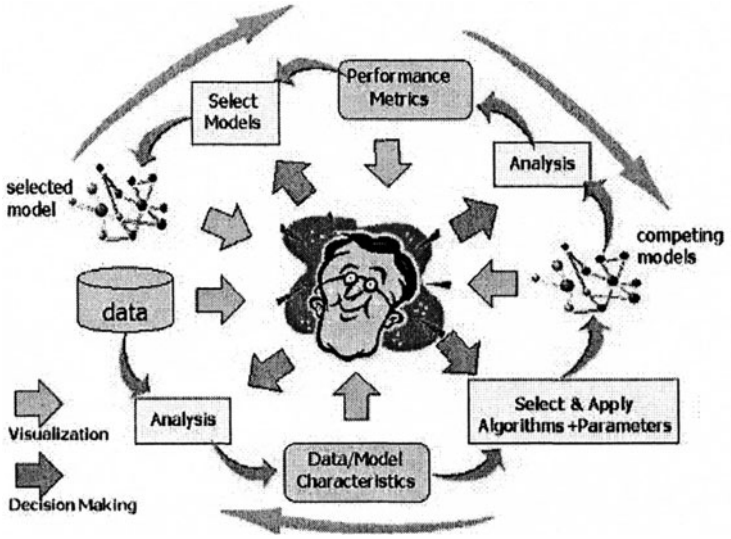

Fig. 1: The key idea of user-centered data mining with support for model selection and visualization

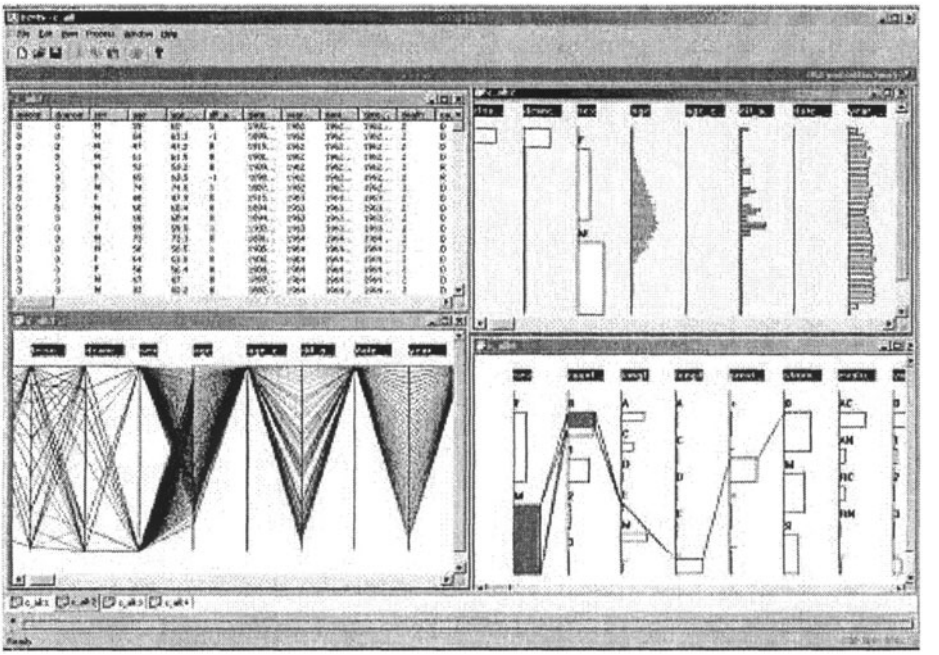

Figure 2. Data visualization in D2MS: The top-left window shows the dataset, the bottom-left window shows the original data view, the top-right window shows summarizing data view, and the bottom-right window shows the querying data view.

This user-centered approach requires an active participation of the user in the learning/mining process with both quantitative and qualitative evaluation of patterns/models. We have developed the user-centered visual data mining system D2MS to support such an active participation of the user in which the visualization module is linked to all other modules relating to model selection [5]. This visualization module consists of a data visualizer, a rule visualizer, and a tree visualizer. These visualizers are integrated to most methods available in D2MS for preprocessing, data mining, and postprocessing. We describe each visualizer with focus on its techniques and how it is linked to the steps in the KDD process.

2.1 Data Visualization

D2MS improves parallel coordinates [2] in several ways to adapt them to the knowledge discovery context. It is because when viewing a large dataset with many attributes, particularly categorical attributes, the advantage of parallel coordinates may lose as many polylines or their parts are partially overlapped, and certain kinds of summarization might be needed. Also, the user often needs to see subsets of the dataset in terms of cases and/or attributes.

Viewing original data. The basic idea of viewing a p-dimensional dataset by parallel coordinates is to use p equally spaced axes—which are parallel to one of the screen axes and correspond to attributes and the ends of the axes correspond to minimum and maximum values for each dimension—to represent each data instance as a polyline that crosses each axis at a position proportional to its value for that dimension. This view gives the user a rough idea about the distribution of data on values of each attribute; in particular colors of classes can show clearly how classes are distributed. The original stomach cancer data shown in the top-left window in Fig. 2 is visualized in the bottom-left window.

Summarizing data. This view is significant as the dataset may be very large. The key idea is not to view original data points but to view their summaries on parallel attributes. D2MS uses bar charts in the place of attribute values on each axis. The bar charts in each axis have the same height (depending on the number of possible attribute values) and different widths that signify the frequencies of attribute values. D2MS also provides interactively common statistics on each attribute as mean or mode, median, variance, boxplot, etc. The top-right window in Fig. 2 shows the summaries of the stomach cancer data.

Querying data. This view serves the hypothesis generation and hypothesis testing manually. It allows the user to view subsets of the dataset determined by queries. There are three types of queries: (i) based on a value of the class attribute where the query determines the subset of all instances belonging to the indicated class; (ii) based on a value of a descriptive attribute where the query determines the subset of all instances having this value, (iii) based on a conjunction of attribute-values pairs where the query determines the subset of all instances satisfied this conjunction. The subset of instances matched a query is visualized in viewing data mode and in summarizing data mode. The gray regions on each axis show the proportions of specified instances on values of this attribute as visualized in the bottom-right window in Fig. 2.

2.2 Rule Visualization

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 of 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:

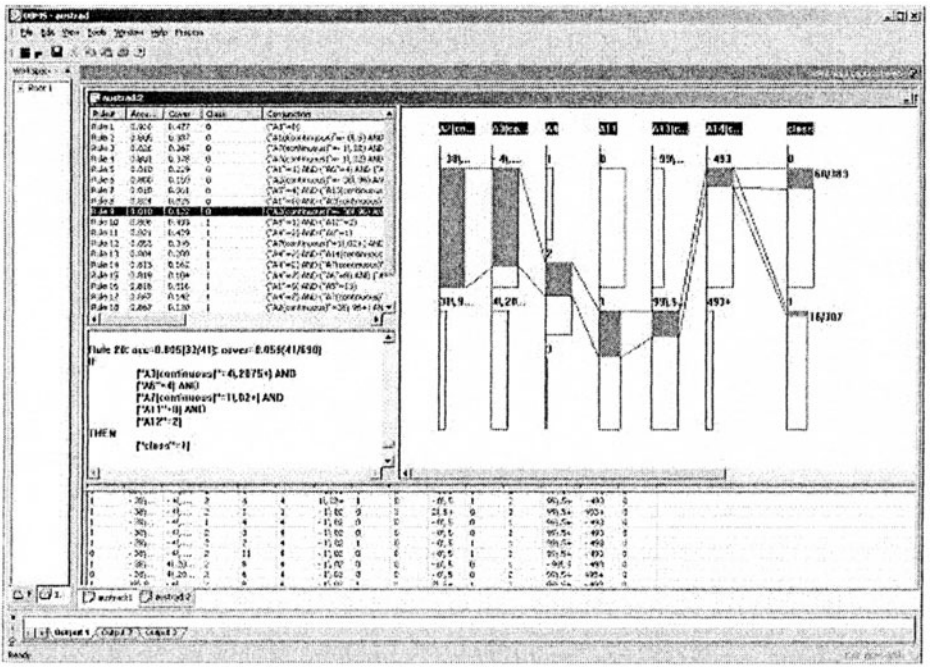

Figure 3. Rule Visualization in D2MS: The top-left window shows the list of discovered rules with a highlighted rule, the middle-left window shows this rule content, the top-right window visualizes the rule by parallel coordinates, the bottom window shows the set of instances covered by the rule.

Viewing rules. Each rule is displayed by polylines 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 color. 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.

Viewing rules and data together. 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. The rules for a class can be displayed together, and all instances covered by these rules are displayed.

Rule visualization in model selection. There are several ways that support the user in evaluating the quality of the rule together with other measure such as coverage and accuracy of the rule. For example, two rules predicting a target class having the same support and confidence but the one wrongly covered more instances belonging to classes different from the target class would be considered worse. Fig. 3 illustrates rule visualization in D2MS where the top-left and middle-left windows display a discovered rule, and the top-right visualizes the rule and the bottom window shows the instances covered by that rule.

2.3 Tree Visualization

D2MS provides several visualization techniques that allow the user to visualize effectively large hierarchical structures. The *tightly coupled views* display simultaneously a hierarchy in normal size and tiny size that allows the user to determine quickly the field-of-view and to pan to the region of interest. The *fish-eye view* distorts the magnified image so that the center of interest is displayed at high magnification, and the rest of the image is progressively compressed. Also, our new proposed technique *T2.5D* is implemented in D2MS for visualizing very large hierarchical structures.

Different modes of viewing hierarchical structures. The D2MS tree visualizer provides multiple-views of trees or hierarchical structures.

- *Tightly coupled views*: The global view (on the left) shows the tree structure with nodes in same small size without labels and therefore it can display a tree fully or a large part of it, depending on the tree size. The detailed view (on the right) shows the tree structure and nodes with their labels associated with operations to display node information. The global view is associated with a field-of-view or panner (a wire-frame box) that corresponds to the detailed view. These two views are tightly coupled as the field-of-view can be moved around in the global view in order to pan the detailed view. Also, when the detailed view is scrolled the position of the field-of-view will be updated accordingly. The user can resize windows for these views, and the field-of-view shape and size will be automatically changed. The top-left and top-right windows in Figure 3 show the tightly coupled views of D2MS for the stomach cancer data.

- *Customizing views*: Initially, according to the user's choice, the tree is either displayed fully or with only the root node and its direct sub-nodes. The tree then can be collapsed or expanded partially or fully from the root or from any intermediate node.

- *Tiny mode with fish-eye view*: Note that no current visualization technique allows us to display efficiently the entire tree when it has, says, ten thousands nodes. The tightly coupled views are extended with three viewing modes according to the user's choice: normal size, small size and tiny size. The tiny mode uses much more efficiently the space to visualize the tree structure, on which the user can determine quickly the field-of-view and pan to the region of interest. It allows the user to be able to see the tree structure while focusing on any particular part so that the relationship of parts to the whole can be seen and the focus can be moved to other parts in a smooth and continuous way.

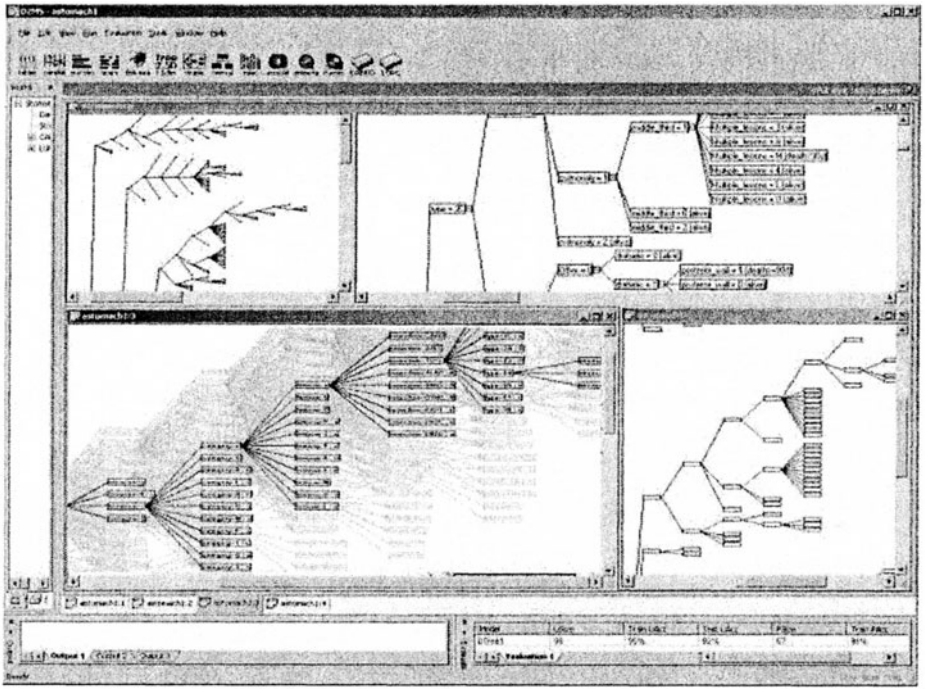

Figure 4. Multiple views of trees: Tightly coupled, fish-eye, overall, and T2.5D views.

- *Trees 2.5 Dimensions.* The user might find it difficult to navigate a very large hierarchy, even with tightly coupled and fish-eye views. To overcome this difficulty, we have been developing a new technique called T2.5D (stands for Trees 2.5 Dimensions) [5]. Different from tightly-coupled and fish-eye views that can be seen as location-based views, T2.5D can be seen as a relation-based view in the sense that highlighted parts of a tree are relations determined by queries. The starting point of T2.5D is the observation that a large tree consists many subtrees that are not usually and necessarily viewed simultaneously. The key idea of T2.5D is to represent a large tree in a virtual 3D space (subtrees are overlapped to reduce occupied space) while each subtree of interest is displayed in a 2D space.

To this end, T2.5D determines the fixed position of each subtree (its root node) in two axes X and Y, and in addition, it computes dynamically a Z-order for this subtree in an imaginary axis Z.

A subtree with a given Z-order is displayed "above" its siblings those have higher Z-orders. When visualizing and navigating a tree, at each moment the Z-order of all nodes on the path from the root to a *node in focus* in the tree is set to zero by T2.5D. The *active wide path* to a node in focus, which contains all nodes on the path from the root to this node in focus and their siblings, is displayed in the front of the screen with highlighted colors to give the user a clear view. Other parts of the tree remain in the background to provide an image of the overall structure. With Z-order, T2.5D can give the user an impression that trees are drawn in a 3D space. The user can easily change the active wide path by choosing another node in focus. We have experimented T2.5D with various real and artificial datasets. In an experiment, T2.5D can handle well trees with more than 20,000 nodes, and more than 1,000 nodes can be displayed together on the screen. Fig. 4 illustrates a pruned tree of 1795 nodes learned from stomach cancer data and drawn by T2.5D (the bottom-left window).

Tree Visualization in Model Selection. In D2MS, visualization is integrated with the steps of the KDD process and closely associated with the plan management module in support for model selection. The user can have either views in executing a plan or comparative views of discovered models. If the user is interested in following the execution of one plan, he/she can view, for example, the input data, the derived data after preprocessing, the generated models with chosen settings, the exported results. Thus, the user can follow and verify the process of discovery by each plan, and change settings to reach alternative results.

With the three modes of viewing of data, D2MS integrates data visualization into different KDD steps by displaying and interactively changing these views of data at any time. Data visualization supports doing data preprocessing and examining the relation between data and discovered knowledge. In the first step of collecting data and formulating the problem, the user can and often need to view the original dataset and its summarization. The visual analysis of collected data may help the user to identify important or redundant attributes or new attributes to be added. The data visualization has shown to be significant in the data preprocessing step that consists of functions on data cleaning, integration, transformation and reduction. Many discretization algorithms provide alternative solution of dividing a numerical attribute into intervals, where the visual data query on the discretized attribute and the class attribute can give insights for decision.

The data visualization is also very significant in data mining step with data query mode, and particularly in the evaluation step in its synergistic combination with rule and tree visualization. If the user is interested in

comparative evaluation of competing models generated by different plans, he/she can have multiple views on these models. The user can compare performance metrics of all activated plans that are always available in the summary table. Whenever the user highlights a row in the table, the associated model will be automatically displayed. Several windows can be opened simultaneously to display competing models in forms of trees, concept hierarchies, or rule sets. For example, two rules predicting a target class having the same support and confidence but the one wrongly covered more instances belonging to classes different from the target class would be considered worse (Figure 6).

3.ACTIVE PARTICIPATION OF THE USER IN MINING STOMACH CANCER DATA

3.1 The dataset

The stomach cancer dataset collected at the National Cancer Center (NCC) in Tokyo during the period 1962-1991 is a very precious source for the research. It contains data of 7,520 patients described originally by 83 numeric and categorical attributes. These include information on patients, symptoms, type of cancer, longitudinal and circular location, metastasis, pre-operative complication, post-operative complication, etc.

One problem is to use of attributes containing patient information before operation to predict the patient status after the operation. The domain experts are particularly interested in finding predictive and descriptive rules for the class of patients who "death within 90 days" after operation among totally 5 classes ("alive", "death after 5 years", "death after 90 days", "death within 90 days", "unknown"). This class has only 302 available cases from the total 6,712 cases in the dataset (4%). We face here with the problem of mining *minority classes in imbalanced datasets* that is very popular in medicine.

3.2 Mining stomach cancer data with D2MS

The task of extracting patterns/models from unbalanced datasets in general, and from the class "death within 90 days" in particular, is very difficult because of such a class distribution and there is no sharp boundary between this target class and the others. The NCC experts evaluate discovered patterns/models by three criteria of "acceptability", "novelty", and "utility" with points from 1 (lowest) to 5 (highest).

We first applied well-known and popular data mining methods such as CBA [7], C4.5 [8] and its commercial version See5 to this task. However, the obtained results were far from expectation. We could not find a trade off between support and confidence to discover rules. Moreover, discovered

rules usually cover less than a half of patients of the target class. Also, it happens that the found rules are often acceptable (5 points) in terms of validity but not new (1 point) to the domain experts and thus are not interesting to them.

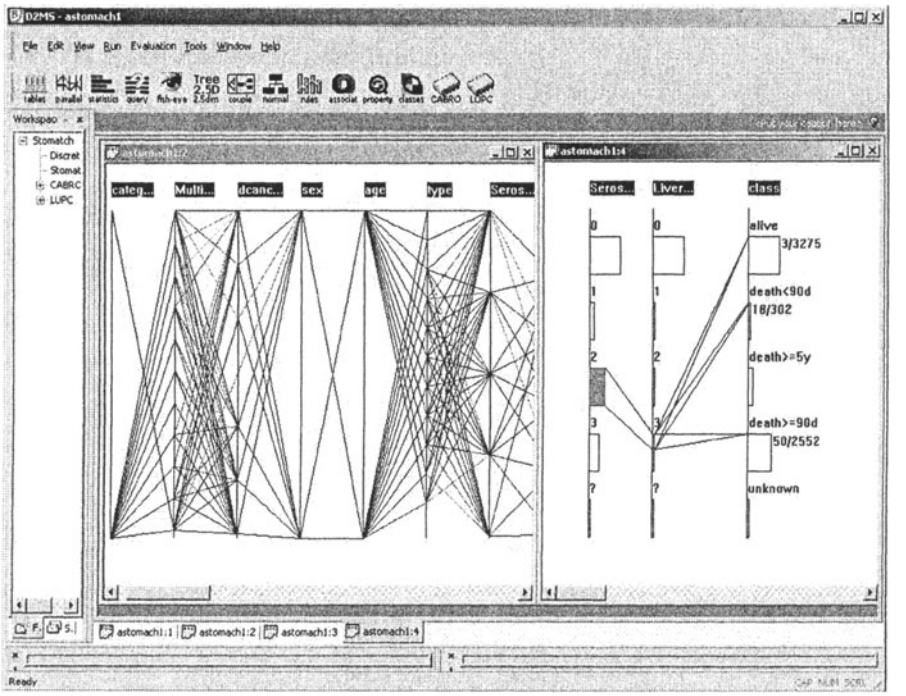

Figure 5. Data visualization supports finding unusual phenomena.

LUPC (Learning Unbalanced Positive Class) is our data mining method implemented in D2MS [5] to deal with the problem of mining in unbalanced datasets. It can find efficiently prediction and description rules by allowing the user (1) to vary its parameters and search strategies to detect possible rules with different values of support and accuracy, (2) to impose constraints on discovered patterns/models.

The visualization tools in D2MS allow the us to examine interactively the data and to gain better insight into complex data. The summarizing and querying modes can suggest hypotheses to be investigated. An illustration of identifying unusual phenomena is shown in Fig. 5. It is known that patients who have symptoms of "liver metastasis" of all levels 1, 2, or 3 will not be able to survive. Also, "serosal invasion = 3" is a typical symptom of the class "death within 90 days". With the visualization tools, we found several unusual events such as 5 of 2329 patients in class "alive" have heavy metastasis of level 3; as well 1 and 8 of them are of metastasis level 2 and 1, respectively. Moreover, the querying data allow us to verify some significant

combination of symptoms such as "liver metastasis = 3" and "serosal invasion = 3" as shown in Fig. 5.

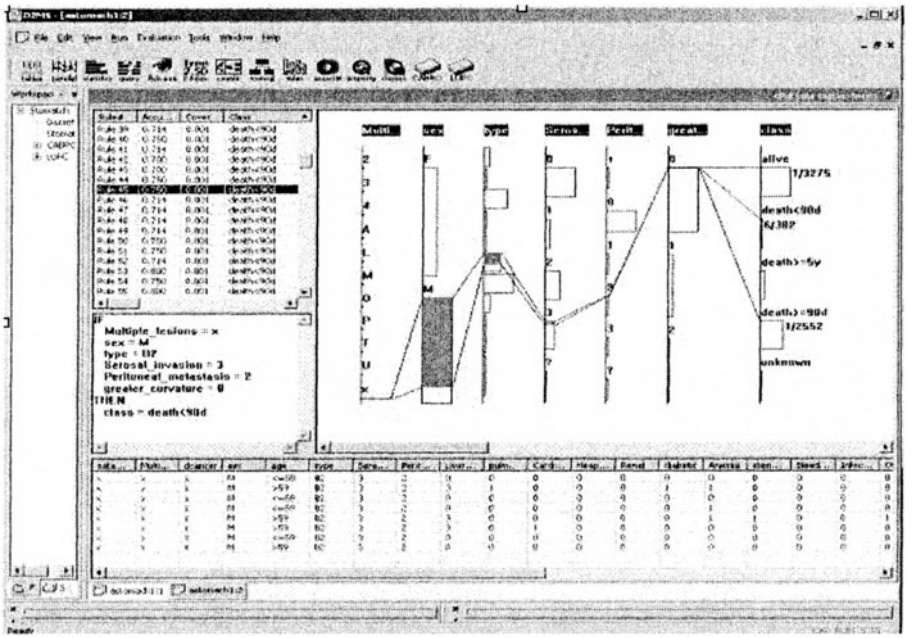

Figure 6. Rule visualization supports the user in evaluating discovered knowledge.

Using visual interactive D2MS we have tried different trials with varying parameters and constraints to find only rules that do not contain the characterized attribute "liver metastasis" and/or its combination with other two typical attributes "Peritoneal metastasis", "Serosal invasion" (Fig. 6). Below is a rule with accuracy 100% discovered by D2MS that brings new and *irregular* patterns in the class "death with 90 days".

Rule 8: accuracy = 1.0 (4/4), cover = 0.001 (4/6712)
IF category = R AND sex = F AND proximal_third = 3 AND middle_third = 1
THEN class = death within 90 days

Thank to the supported interaction with D2MS we can also discover *rare* patterns. D2MS allow us examine effectively the hypothesis space and identify rare rules with any small given support or confidence. An example is to find rules in class "alive" that contain the symptom "liver metastasis". Such patterns are certainly rare and influence the human decision making. We found events in the class "alive" such as male patients getting "liver metastasis" at serious level 3 who can survive with the accuracy of 50%.

Rule 1: accuracy = 0.500 (2/4); cover = 0.001(4/6712)
IF sex = M AND type = B1 AND liver_metastasis = 3 AND middle_third = 1

THEN class = alive

4. CONCLUSION

We have presented the visualization techniques in the knowledge discovery system D2MS for supporting model selection. We emphasize the central role of the user's participation and visualization in the model selection process of knowledge discovery. Our basic idea is to provide the user with the ability of trying various alternatives of algorithm combinations and their settings, and to provide the user with performance metrics as well effective visualization so that the user can get insight into the discovered models before making his/her final selection. D2MS with its visualization support in model selection has been used and shown advantages in extracting knowledge from a real-world application on stomach cancer data.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. N. Yamaguchi and his colleagues at the National Cancer Center in Tokyo for the permission of using the stomach cancer dataset. This research is supported by the programme Grand-in-Aid for Scientific Research on Priority Areas (B).

REFERENCES

[1] Brachman, R. and Anand, T., "The Process of Knowledge Discovery in Databases", in *Advanced in Knowledge Discovery and Data Mining*, Fayyad, U. et al. (Eds.), Morgan Kaufmann, 1996, 38-57.

[2] Card, S. K., Mackinlay, J. D., and Shneiderman, B., *Readings in Information Visualization*, Morgan Kaufmann, 1999.

[3] Fayyad, U.M., Grinstein. G.G., and Wierse, A. *Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann*, 2002.

[4] Han, J. and Cercone, N., *Visualizing the Process of Knowledge Discovery, J. of Electronic Imaging*, No. 4, 404-420, 2000.

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

[6] Keim D. A. and Kriegel H.P., "Visualization Techniques for Mining Large Databases: A Comparison", *J. IEEE Transactions on Knowledge and Data Engineering*, 923-938, 1996.

[7] Liu, B., Hsu, W., and Ma, Y., "Integrating Classification and Association Rule Mining", *Fourth Int. Conf. on Knowledge Discovery and Data Mining KDD'98*, 80-86, 1998.

[8] Quinlan, R., C4.5: *Programs for Machine Learning*, Morgan Kaufmann, 1993.