A Visual Data Mining Framework for Convenient Identification of Useful Knowledge[1, 2]

Kaidi Zhao, Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan Street
Chicago, IL 60607. US
{kzhao, liub}@cs.uic.edu

Thomas M. Tirpak, Weimin Xiao
Motorola Labs
1301 E. Algonquin Rd.
Schaumburg, IL 60196. USA
{T.Tirpak, awx003}@motorola.com

Abstract

Data mining algorithms usually generate a large number of rules, which may not always be useful to human users. In this project, we propose a novel visual data-mining framework, called Opportunity Map, to identify useful and actionable knowledge quickly and easily from the discovered rules. The framework is inspired by the House of Quality from Quality Function Deployment (QFD) in Quality Engineering. It associates discovered rules, related summarized data and data distributions with the application objective using an interactive matrix. Combined with drill down visualization, integrated visualization of data distribution bars and rules, visualization of trend behaviors, and comparative analysis, the Opportunity Map allows users to analyze rules and data at different levels of detail and quickly identify the actionable knowledge and opportunities. The proposed framework represents a systematic and flexible approach to rule analysis. Applications of the system to large-scale data sets from our industrial partner have yielded promising results.

1. Introduction

Data mining algorithms usually generate a large number of patterns or rules [2] [20] that are hard to comprehend. Most of the discovered rules actually are not useful. A number of techniques have been proposed to help the user find interesting rules [1] [12] [18] [19] [21], either using objective measures, or subjective measures such as unexpectedness and actionability [1][15][18].

In our work, we propose a visual data mining framework called Opportunity Map. It integrates a set of visual data mining techniques, to quickly identify interesting and actionable knowledge. The visualization layout is inspired by the House of Quality in Quality Function Deployment [6] [23], specifically the Interrelationships Matrix in the House of Quality (HOQ) from Management Sciences. In the Opportunity Map, Customer Requirements in the HOQ are mapped to application requirements expressed as classes in data mining. Technical Requirements in HOQ are mapped to attributes and values. In this way, the framework is able to make use of well-established methodologies and business practices in product design and manufacturing from Management Sciences, such as fast identification of important activities and prioritizing them.

An initial prototype of the proposed Opportunity Map system is reported in [28]. In this paper, we enhance previous methods and also extend the above framework with a number of novel visual mining methods which significantly improve the usability of the system. We introduce them briefly below.

In the proposed Opportunity Map system, an integrated data mining rules visualization and distribution map visualization method is developed. Distribution maps (Distribution Bars/Correlation Charts) [5][11] are used in traditional statistical data analysis. This technique plots the data distribution of

[1] Parts of the work are under patent applications. For most recent advances please contact the authors.

[2] We would like to thank Tom Babin, Paul DeClerck, Dan DeClerck, Jeffrey Benkler and Michael Kramer for many useful discussions and suggestions. Thank you also to Mike Kotzin of Motorola's Mobile Devices Business for supporting this project through the Illinois Manufacturing Research Center.

(usually two) variables using bar charts, thereby giving users some ideas of how the two plotted variables are related. However, the distribution map usually is only able to tell the correlation of the currently plotted attributes. It does not explicitly show the relationships of the two plotted attributes and in relation to other attributes. Data mining rules, on the other hand, represent detailed relationships. However, due to the large number of rules that are typically discovered, the rules may not be easily comprehensible.

In the Opportunity Map, we propose an integrated visualization of the distribution map and data mining rules. The visualization process not only allows users to easily identify interesting and unusual spots in the map, but also to conveniently gain insights into the underlying reasons. The procedure is explained in more details in Section 3.3.

Trend is another important feature that Opportunity Map visualizes. This is especially true when a large number of attributes in our data are numeric or ordinal. The user is interested in knowing how the target classes change over those numeric/ordinal attributes. Opportunity Map provides a special visualization mode, which allows the user to see the changing behavior of an attribute with respect to the classes. Section 3.4 will discuss this in detail.

Opportunity Map also allows comparative study of rules with visualization to reveal important hidden knowledge. An extension to our previous work on this is presented in Section 3.5.

Given a new task, Opportunity Map involves these steps in its visual data mining process:

1. Mine rules from the data. We use class association rules, i.e., association rules with only a class value on the right-hand-side [13].
2. Visualize the rules with data using the Opportunity Map. The visualization allows the user to create priority areas in the visualization that contain most important rules and knowledge.
3. Identify interesting spots in the priority areas, i.e., attributes and cells in the map that are interesting.
4. Drill down to a particular attribute with all the classes related to the attribute to find more specific rules.
5. Compare data and rule distributions, trends, etc., to discover actionable knowledge.

All these steps are performed in an iterative manner. Interesting and actionable rules are often identified in Steps 4 and 5.

Our work makes the following contributions:

1. To our knowledge, this is the first attempt to study rule actionability using a visual data mining approach.
2. Adapt ideas from Quality Function Deployment in management science to rule analysis. The framework, to some extent, bridges the Management Science and data mining, which effectively makes data mining results more attractive and acceptable to management teams.
3. Integrate distributions and data mining rules through a combined visualization, which are complementary to each other. Rule comparison and trend behavior visualization are also supported in the visualization. These new capabilities significantly enhance the Opportunity Map system. Our users confirm that the combined visualization enables them to quickly focus on interesting spots and discover actionable knowledge.

The Opportunity Map framework with the proposed techniques has been used for real-life, large-scale datasets from our industrial partner Motorola. The user feedbacks confirm the proposed methods are useful and easy to use.

2. Related work

Our research is related to three main areas of data mining: interestingness, rule query and visualization.

In rule interestingness analysis, our work is related to unexpectedness and actionability analysis. As was discussed in the introduction to this paper, little related work has been done using visualization for this purpose. In [8][24], query is used to retrieve certain rules. However, one cannot issue a query to find interesting and actionable rules if he does not already know what the rules are.

Regarding data mining results visualization, our work is related to rule visualization [11]. [9] proposes interactive mosaic plots to visualize the contingency tables of the association rules. In [7], classification rules are visualized using rule polygons. [25] introduces a method to visualize association rules in text domain. [26] visualizes the behavior of rules, i.e., changes of supports and confidences over time. In [4] parallel coordinates are used to visualize rules. [3] uses 3D graph to visualize rules by emphasizing their supports and confidences. In [10], a post-processing environment is proposed to browse and visualize association rules so that the user can divide a large rule set into smaller ones. In [17], important rules in terms of support and confidence values are highlighted with a grid view. In [16], ordering of categorical data is studied to improve the visualization, with the goal of less visualization clutter. It is mainly useful for parallel coordinates such as [26][27] and other general spreadsheet types of visualization.

All of the above approaches do not actively help the user find interesting and actionable knowledge. They differ from our proposed techniques in Opportunity Map in terms of both the goal and the visualization. Opportunity Map is a rule visualization system and a process for fast identification of interesting and actionable knowledge. It also can be used to analyze data distribution, trends of attributes, and rules comparative analysis.

Works presented in this paper extend those reported in [28] with a number of novel visual data mining methods as discussed in the Introduction section. We will not repeat them here.

3. Opportunity Map

3.1 Problem statement

The techniques used in Opportunity Map are designed for the type of applications in which data set has class labels. Usually, the user is interested in a subset of the classes, which represents important states. Such types of data are commonly used in classification or prediction. However, it is important to note that classification or prediction is not the primary objective here. Our goal is to find rules that help solve real-world problems. For example, a New Product Introduction Engineer may want to find the relationships between product attributes and abnormal classes in product performance data, thereby providing insights regarding how to modify the product design and avoid the abnormal cases.

3.2 Opportunity Map – basic layout

For adaptation of the concepts from Management of Science and the House of Quality, and details of the visualization layout, please refer to our previous report in [28]. Briefly, in the main Matrix visualization, the X-axis lists all the attributes, and the Y-axis lists all the classes. Figure 1 shows an illustration.

Important Classes: On Y-axis, the important classes are placed at the upper part of the visualization, while the less important classes are located at the lower part of the matrix.

Actionable Attributes: An attribute is actionable if the user is able to do something with that attribute to achieve some desired effects. Actionable attributes are placed on the left side of the matrix, while non-actionable attributes are placed on the right.

Please note that the important classes and actionable attributes are domain and application dependent. For each application/domain, the above placements create four priority sectors. Two prominent wide divider lines are used to divide the visualization into four sectors as illustrated in Figure 1.

Sector 1 (upper left) is the **Primary Area**, which contains important classes and actionable attributes. This is the most important sector and represents the best opportunities. It is thus the area on which the user should focus.

Sector 2 (upper right) is the **Informative Area**, which is related to important classes, but the attributes in this sector are not actionable. Rules in this area may, however, help the user to better understand the application domain.

Sector 3 (lower left) is the **Secondary Area**, which contains less important classes with actionable attributes. The knowledge discovered from this area can be acted upon, but is generally of less interest to the user for the current application.

Sector 4 (lower right) is the **Uninteresting Area**, in which the rules are not important and not actionable.

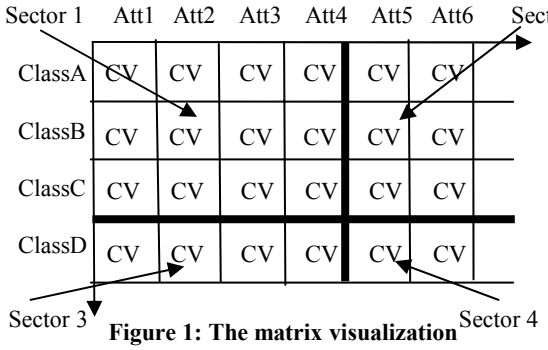


Figure 1: The matrix visualization

When the user finds an interesting attribute, he/she can study the details by using drill down visualization on that attribute. The visualization is of similar nature to Figure 1. Classes are still associated with Y axis, but the X axis now shows the values of the drill-down attribute (values are discretized/binned if the attribute takes continuous values).

Cell Visualization (CV) is used for each cell (grid) in Opportunity Map. It helps the user to decide where to focus and show important information. By default, a CV renders the rules in that grid, as in Figure 2 (on the later figure page).

The top part of a CV visualizes the set of rules in that cell in descending order of rules' confidence values. Each rule is visualized as a small bar. The height indicates the rule's confidence, and the width indicates its support. Orange and purple are used in a round robin fashion to color the bars so that the user is able to distinguish different rules. Center part of cell is colored in blue and the saturation value is used as a rough indication of number of data points in that grid.

By default, the two (optional) bars (A and B as in Figure 2) at the bottom of the CV proportionally indicate the number of data points in that cell (bar B) and the number of data covered by the rules in that cell (bar A). For the advantages of this CV layout and details, please refer to [28]. Also, CV is extensible and we will see its other rendering methods shortly.

3.3 Distribution map and data mining rules

Distribution map and data mining rules are two of the most frequently used techniques in traditional visual data analysis and data mining. However, when used separately, they are not sufficient for the following reasons:

A distribution map is visually structured and easily perceptible. It shows the data distribution of an attribute with respect to another attribute (usually class attribute), by a simple visualization. In this way, the user is able to recognize the common distributions, and unusual spots. However, due to the way that the distribution map is constructed, the visualization only allows the user to explore the relationships of the two attributes in question. It does not provide clues how they may be related to other attributes or values. In the real world applications, it is very likely that three or more attributes are correlated to express a piece of knowledge. Also, the distribution map can only indicate possible locations that have relationships, without explicitly showing the exact relationships.

Data mining rules, on the other hand, are capable of capturing multi-attribute relationships. The rules are able to tell the exact relationships and the quality of the relationships in terms of support and confidence values. However, rules are usually generated in large numbers, and they are hard to understand by human users. Also, although the rules can be ordered based on some statistical measures, they are not organized.

Opportunity Map integrates these two techniques into an intuitive and powerful visualization. In the drill down visualization of one attribute, the Opportunity Map shows two types of distributions:

1. **Distribution of values**: for each value of the drill-down attribute, a bar with height proportional to the data size for that value is drawn above the X-axis. Figure 4 illustrates this idea.
2. **Distribution of classes over each value**: for each column (a possible value of the drill down attribute), distribution bars on classes for data and data covered by rules are drawn in the corresponding cells in that column. This results in two bars (of different colors) for each cell. These two bars can either replace the whole Cell Visualization (such as those in Figure 4), or just act as the bar A and B (Figure 2) in CV (such as those in Figure 5). The user can choose either display mode. The system supports fast switching between these two modes.

With such visualization, Opportunity Map allows the user to perform various interesting studies:

1. He/she can inspect various distributions in the drill down mode, either by values (X-direction) or by classes of one value (Y-direction), and compare them visually. The visualization makes it easy for the user to recognize unusual distributions, which act as good starting points for further interactive knowledge discovery.
2. Once an interesting distribution is located in a cell, the user can either compare it with its neighbor values (cells), or refer to the rules visualized in the cell for possible reasons. This usually gives clues of closely related attributes or values.
3. Visually comparing two bars related to data and data covered by rules may give clues to the quality of the data and the rules, such as: if the bar for the rules is much smaller than the bar for the data, it means the data there is hard to be characterized by rules. If there are only a few big rules which cover most of the data (prominent by the large bar size), they indicate that the data for that cell has very good patterns that characterize the data very well. In general, for one cell, the more rules there are, and the less data the rules cover, the more random are the data in the cell, i.e., with few significant patterns. This type of comparison can give important insights to the domain expert when he/she is presented with the visualization.

Usage of the Opportunity Map has confirmed that visualizing distribution map and data mining rules together turns out to be one of the most effective tools in our proposed framework. Interesting knowledge is usually discovered in those cells with unexpected distributions, with rules revealing exact relationships and related attributes / values for further study.

3.4 Trends behavior visualization

When a certain attribute increases it value, a certain target class may become less likely to happen. This kind of knowledge is helpful to identify key attributes in the data, and also useful for the user to improve his products and applications. In Opportunity Map, a special Cell Visualization called Trends Behavior Visualization is provided to help the user find such useful information using visualization.

Trends behavior visualization plots the changing

trends of one attribute (ordinal type) with respect to all the classes in the drill down visualization.

1. First, values of the attribute (on X axis) are ordered. Bins will be used instead of raw values if applicable.
2. A normalized value is calculated for each cell, using: $Count_{Cell}$ / $Count_{Column}$, where $Count_{Cell}$ is the data count in that cell, and $Count_{Column}$ is the total data count of that column.
3. A bar is rendered in each cell, with its height proportional to the above normalized value.

Figure 6 shows an example. Please note that different values of the attribute may have different counts on X axis (either due to the nature of the data, or due to the way we discretize / bin the data). However, due to the way we normalize and render the bars, on horizontal direction for each class (row), the height of the bars will correctly reflect the data percentage distribution trends of the corresponding class for each value. For example, in Figure 6, for class8473, class3118, we can see the clear going up behavior for these two classes over all the values of that attribute. This explains to the user that these two classes are more likely to happen when the value of the attribute increases. If that attribute is actionable and the user wants to avoid these two classes, then the obvious action should be trying to decrease the value of that attribute in his products or applications.

3.5 Comparative study

Comparative study of rules may reveal interesting results that are not easily observable from individual rules alone. For example, in the product design domain, one may want to compare the rules for two products to find out why one performs better than the other (product model is an attribute).

A simplistic way of doing this is to list rules from two products side-by-side. This does not work well because rules in the two sets can be quite different due to minimum support/confidence constraints. Thus, we proposed a method to mine and compare the rules as reported in [28]. This method enables the user to see the difference of two values on a given rule set.

When there are many possible values for comparison, Opportunity Map provides another comparative study method in the drill-down visualization, called value-based rule comparison. This method is useful when the user finds one interesting value of an attribute, e.g., a specific product model, in a set of rules. He can test this set of rules on all other values of that attribute, e.g., other products, to see whether this set of rules represent any general knowledge. The testing is done by creating new rules by replacing the value in question with other possible values, and computing the support and confidence values for the new rules and visualizing them. Figure 7 shows one example (See Section 4 for description). For our applications, users use rule hypothesis testing as a convenient way to test and confirm the knowledge that they learned.

Although the above two methods are similar, they are proposed for different scenarios. The first method compares rules on two data subsets, and the visualization is designed for finding rules that behave very differently on the two data subsets. The value-based rule comparison method is more appropriate for locating unusual and abnormal cells, so that the user can focus on them to do detailed analysis. For the value-based method, the system allows the user to do the value-based rule comparison on either values of an attribute (as discussed above), or on different classes. The process of comparing different classes is similar. When applied to classes, this technique allows the user to see the strength and applicability of the rules on different classes. Due to space limitation, we do not discuss this further.

4. Case study

In general, it is difficult to have an objective measure of effectiveness for a visual data mining system. It can only be evaluated subjectively by the people who use it in real-life applications. This section presents a case study based on our real-life applications using the Opportunity Map for Motorola's Mobile Devices Business unit.

The example shown here is based on a dataset of 50,000 data points, with 160 attributes and over 5,000 rules. Here, we emphasize the findings from the system, rather than the process (which was addressed in the previous sections of the paper). Due to confidentiality reasons, all attributes, classes and values are replaced by generic names and values. To save space, images are cropped and resized.

We started our analyses by speaking with the domain experts to identify the important classes and actionable attributes. Figure 3 shows the initial visualization screen after data were loaded. The main window (on the left) displays the Matrix visualization of the Opportunity Map. All the user interaction with the system is performed in this window. The two wide green lines are used to divide the Matrix into four sectors as described in Section 3.2. In the subsequent figures, these two lines are not visible due to the limited page size (In these examples, there are many

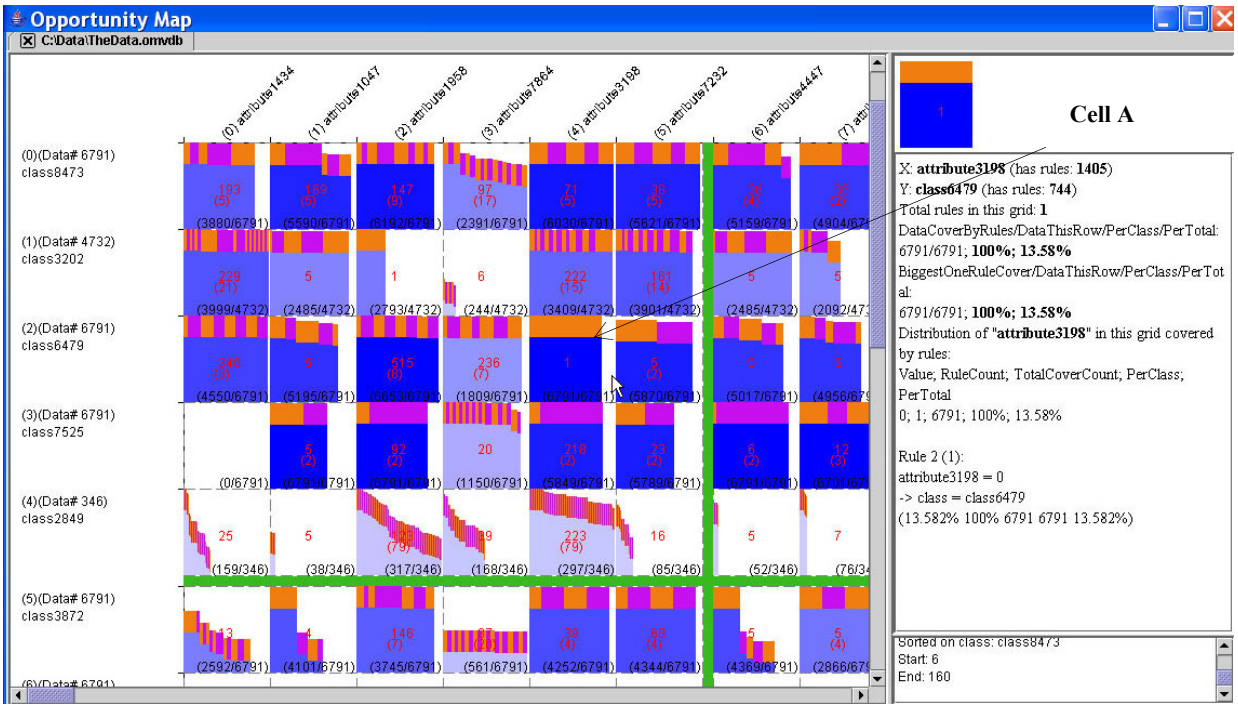


Figure 3. Initial visualization after loading data

actionable attributes and important classes). The information window on the right displays the detailed information as the user moves the mouse cursor over the main window. A small log window on the lower-right corner logs the operations performed, and provides hints when applicable.

The analysis starts with the prominent bars, which are indications of possibly strong relationships or rules. In Figure 3, cell A is prominent for having a large orange bar. The information window confirms that it has one strong rule. We can drill down on this attribute to see its distribution map as in Figure 4.

The distribution map clearly shows that class 6479 only happens when the value of attribute3198 is 0 (cell B). In this case, both the rule and the distribution map can reveal this piece of relationship. If the attribute in that rule is actionable, then the user could use this discovered knowledge to improve the product design.

Also in Figure 4, with prominent bars in cell C and cell D, the distribution map suggests that class7525 is very related to values of (on X axis) "389025024" and "1778412288". Although the distribution map tells us neither the exact relationships nor other detail, we can easily get some clues of the reasons by checking the rules, either moving the mouse over them or switching the visualization to rules as in Figure 5 (cells E and F correspond to cell C and D in Figure 4). The switch between visualization modes is just a click away.

Given the above specific goal, the user can study the rules in cell E to find possible reasons, which may suggest other closely related attributes and values. After that, he/she can go back to Figure 3 and examine other attributes/values suggested in the above steps, with enriched knowledge and clearer view of the nature of the data. We do not discuss it further here due to limited space.

For Cell F, though the distribution map (Figure 4) suggests that it acts as an important value for class7525, the rule generation system failed to generate any rules for that grid. This is a very common phenomenon either because of the minimum support/confidence constraints, or because of the rule pruning. Opportunity Map allows the user to compare these two values in order to find their differences in terms of the rules. The result (after normalization) is shown in Figure 7. The visualization surprisingly informs the user that these two values behave very similarly in terms of the rules on that class. Given the underlying meaning of the values, the user is able to enrich or adjust his/her knowledge based on this fact.

Figure 6 demonstrates the trends behavior visualization of attribute6376, which is a continuous attribute and binned into 3 bins. Clearly, we can see that increasing the value of this attribute will increase the likelihood for class8473, class3872, etc., and will decrease the likelihood of class6479, class7525, etc. Some other classes remain stable, or have no clear pattern. If this attribute is actionable, then the trends behavior information would be very useful in helping the user improving his products and applications.

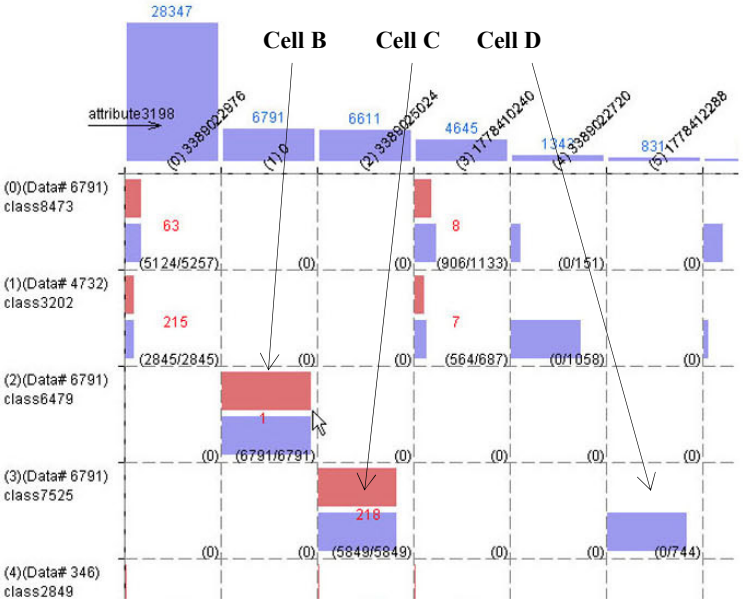


Figure 4. Distribution in drill-down visualization

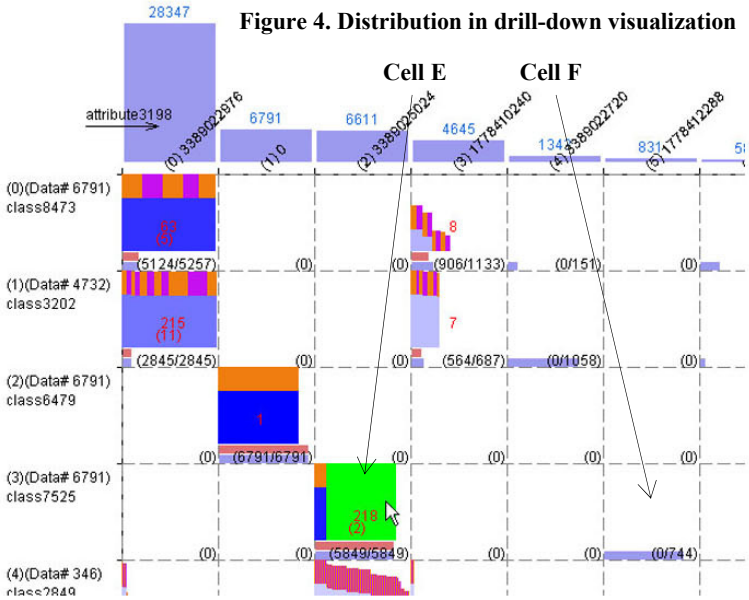
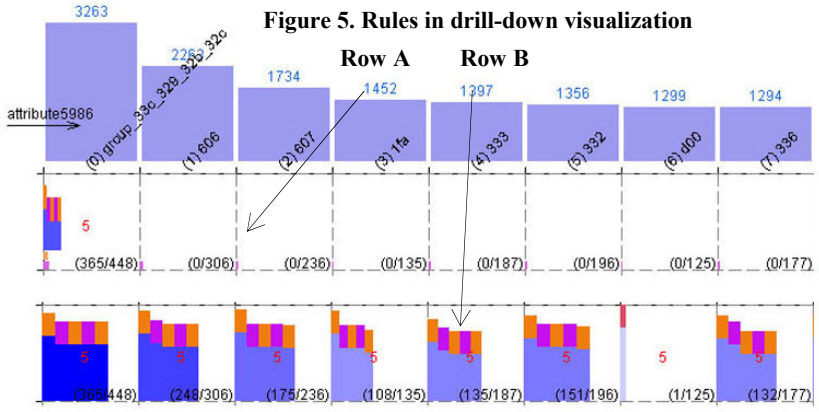


Figure 5. Rules in drill-down visualization



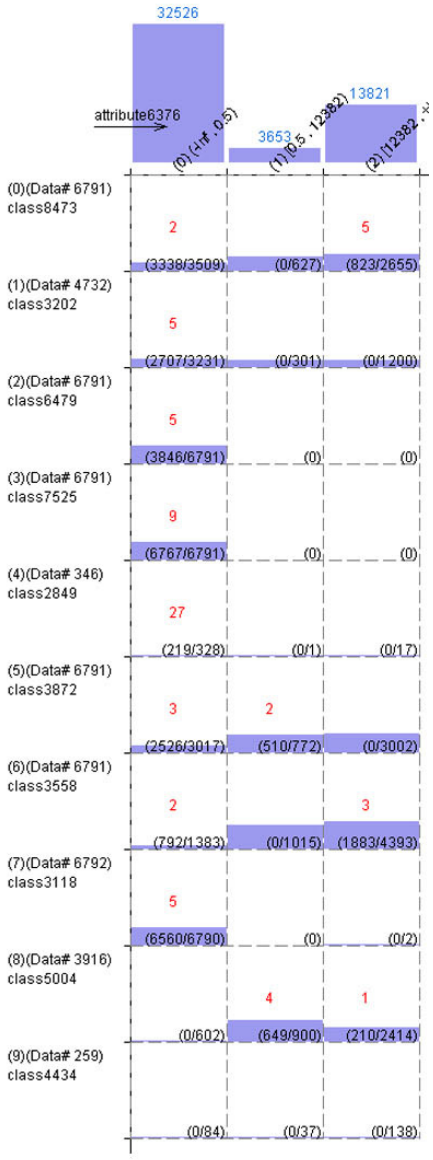


Figure 6. Trends behavior visualization

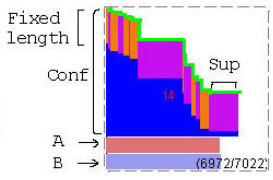


Figure 2.
Example of cell visualization

Figure 7 (left).
Test rules over other values

We now demonstrate the value-based rule comparison using another attribute. Row B in Figure 7 demonstrates the results of the 5 rules in the first cell of Row A. The visualization clearly shows that most values of that attribute behave very similarly on the rules, while the "d00" is an exception. This leads the user to further analyze the reasons using his/her domain knowledge.

In summary, the users confirm that Opportunity Map helps them to find many pieces of useful knowledge.

5. Conclusions

In this paper, we have proposed a number of visual data mining techniques, which are implemented in our Opportunity Map framework, for fast identification of interesting and actionable knowledge. The framework is inspired by the House of Quality from Industrial Engineering. It visualizes the user's needs (e.g., problem classes), attributes and rules as a matrix. The visualization clearly shows how attributes and their values are linked to the important classes. With cell visualization, drill down visualization, integrated distribution map and rule visualization, and comparative study, Opportunity Map represents a systematic way of post-analysis of discovered rules with convenient visualization support. In our applications with real-life, large-scale data sets from our industrial partner, it has been possible to find the interesting and useful rules and patterns from a large number of rules generated by data mining. Thus, the technology transfer team of our industrial partner is working to make the Opportunity Map a part of product reviews and ongoing decision-making by new product designers and managers.

6. References

[1] Adomavicius G. and Tuzhilin, A. "Discovery of actionable patterns in databases: the action hierarchy approach". KDD-97, 1997.

[2] Agrawal R. and Srikant R. "Fast algorithms for mining association rules". VLDB-1994.

[3] Blanchard J, Guillet F, Briand H. "Exploratory Visualization for Association Rule Rummaging". KDD-03 Workshop on Multimedia Data Mining, 2003.

[4] Bruzzese D., Davino C. "Visual Post-Analysis of Association Rules" ECML/PKDD VDM Workshop 2002.

[5] Chambers J.M, Cleveland W. S, Kleiner B, and Tukey P.A. Graphical Methods for Data Analysis. Chapman & Hall, 1983.

[6] Cohen L. Quality Function Deployment. Prentice Hall, 1995.

[7] Han J, Cercone N. "RuleViz: A Model for Visualizing Knowledge Discovery Process". KDD-00, 2000.

[8] Han J., Fu, Y., Wang W., Koperski, K. and Zaiane, O. "DMQL: a data mining query language for relational databases". SIGMOD Workshop on DMKD, 1996.

[9] Hofmann H, Siebes A., Wilhelm, A. "Visualizing association rules with interactive mosaic plots". KDD-00, 2000.

[10] Jorge A., Pocas J., Azevedo P. "Post-processing environment for browsing large sets of association rules". PKDD-02 VDM Workshop, 2002.

[11] Keim D. "Information Visualization and Visual Data Mining". IEEE Trans. Vis. Comput. Graph, 2002.

[12] Liu B., Hsu W., and Chen S. "Using general impressions to analyze discovered classification rules". KDD-97, 1997.

[13] Liu B., Hsu W, Ma Y. "Integrating Classification and Association Rule Mining". KDD-98, 1998.

[14] Liu B., Hsu W., and Ma Y. "Mining Association Rules with Multiple Minimum Support". KDD-99. 1999.

[15] Liu, B., Hsu, H., Ma, Y. "Identifying Non-Actionable Association Rules". KDD-01, 2001.

[16] Ma S., Hellerstein J. "Ordering Categorical Data to Improve Visualization ". INFOVIS-99, 1999.

[17] Ong K-H, Ong K-L, Ng W-K, Lim E-P. "CrystalClear: Active Visualization of Association Rules". ICDM-02 Workshop on Active Mining (AM-02), 2002.

[18] Padmanabhan B. and Tuzhilin A. "A belief-driven method for discovering unexpected patterns". KDD-98.

[19] Piatesky-Shapiro G., and Matheus C. "The interestingness of deviations". KDD-94, 1994.

[20] Quinlan J. R. 1992. C4.5: program for machine learning. Morgan Kaufmann.

[21] Silberschatz A., and Tuzhilin, A. "What makes patterns interesting in knowledge discovery systems." IEEE Trans. on Know. and Data Eng. 8(6), 1996.

[22] Suzuki E., "Autonomous discovery of reliable exception rules". KDD-97, 1997.

[23] Terninko J. "Step-by-Step QFD: Customer-Driven Product Design". Saint Lucie Press. 1997.

[24] Tuzhilin A., and Liu, B. "Querying multiple sets of discovered rules". KDD, 2002

[25] Wong P-C, Whitney P, Thomas J. "Visualizing Association Rules for Text Mining". INFOVIS-99.

[26] Zhao K. Liu B. "Visual Analysis of the Behavior of Discovered Rules". KDD-2001 Workshop on Visual Data Mining.

[27] Zhao K. Liu, B., Tirpak, T., and Schaller, A. "V-Miner: Using Enhanced Parallel Coordinates to Mine Product Design and Test Data". KDD-2004.

[28] Zhao K., Liu B., Tirpak T., and Xiao W. "Opportunity Map: A Visualization Framework for Fast Identification of Actionable Knowledge". CIKM 2005.