AssocExplorer: An Association Rule Visualization System for Exploratory Data Analysis

Guimei Liu
School of Computing
National University of Singapore
liugm@comp.nus.edu.sg

Andre Suchitra
School of Computing
National University of Singapore
suchitra@comp.nus.edu.sg

Haojun Zhang
School of Computing
National University of Singapore
zhanghao@comp.nus.edu.sg

Mengling Feng
Data Mining Department
Institute for Infocomm Research, Singapore
mfeng@i2r.a-star.edu.sg

See-Kiong Ng
Data Mining Department
Institute for Infocomm Research, Singapore
skng@i2r.a-star.edu.sg

Limsoon Wong
School of Computing
National University of Singapore
wongls@comp.nus.edu.sg

ABSTRACT

We present a system called AssocExplorer to support exploratory data analysis via association rule visualization and exploration. AssocExplorer is designed by following the visual information-seeking mantra: overview first, zoom and filter, then details on demand. It effectively uses coloring to deliver information so that users can easily detect things that are interesting to them. If users find a rule interesting, they can explore related rules for further analysis, which allows users to find interesting phenomenon that are difficult to detect when rules are examined separately. Our system also allows users to compare rules and inspect rules with similar item composition but different statistics so that the key factors that contribute to the difference can be isolated.

1. INTRODUCTION

The main objective of exploratory data analysis is to summarize the main characteristics of datasets, often with visualization, to help users understand the data and to suggest hypotheses to test. Association rule mining is an important problem in the data mining area [1]. An association rule can be viewed as a summarization of a subset of the underlying dataset. Understanding association rules can help users understand the underlying dataset. Furthermore, the relationship and interaction among rules often reveal valuable information of the underlying data. Hence it is desirable to use association rules for exploratory data analysis.

Association rule mining often produces a large number of rules. Rules need to be organized and presented properly for easy comprehension. Here we present a system called AssocExplorer to support exploratory data analysis via association rule visualization and exploration. The main features of our system can be summarized as follows:

- We follow the visual information-seeking mantra [3] to design our system: we use a scatter plot to provide a global view of all rules; we allow users to filter rules based on various criteria; details of rules are displayed on demand.
- We use coloring to deliver information effectively. In particular, we color length-1 rules, and the whole collection of rules can be colored based on a selected attribute. This helps users find important attributes easily. Users can then start their analysis from there.
- If users find a rule interesting, they can explore related rules to have a deeper understanding of the rule. Bringing related rules together allows users to compare similar rules and find interesting phenomenon that are difficult to detect when rules are examined separately.
- Rules with similar item composition but very different statistics may represent inexpensive actions that we can take to make a big change. Our system allow users to inspect such rules under various contexts to isolate the key factors that contribute to the difference.
- We use proper structures to store and index rules and the original dataset, which ensures that our system can respond users' requests promptly.

In the rest of this paper, we first briefly describe the association rule mining problem and then present the AssocExplorer system.

2. ASSOCIATION RULE MINING

Association rule mining has been applied on both transactional datasets and attribute-valued datasets. In our system, we focus on attribute-valued datasets since they are prevalent in real-world applications.

Let $D = \{t_1, t_2, \cdots, t_n\}$ be a dataset containing a set of records. Each record is described by a set of attributes $\mathcal{A} = \{A_1, A_2, \cdots, A_m\}$. Attributes in $\mathcal{A}$ are either categorical or

continuous. We use $t[A]$ to denote the value of attribute A in record t. Let A be a categorical attribute and v be a value taken by A. We call attribute-value pair $A = v$ an *item*. Let t be a record. If $t[A] = v$, then we say t *contains* item $A = v$. We use letter i to denote items.

DEFINITION 1 (PATTERN). *A pattern is a set of items* $\{i_1, i_2, \cdots, i_k\}$*, and* k *is called the length of the pattern.*

We use letter X to denote patterns. If a record t contains all the items in X, then we say t *contains* X, denoted as $X \subseteq t$. We use $T(X)$ to denote the set of records containing X, that is, $T(X) = \{t | t \in D \wedge X \subseteq t\}$. The *support* of a pattern X in a dataset D is defined as the number of records in D containing X, that is, $supp(X) = |T(X)|$.

In real applications, users are often interested in finding the association between patterns and a target attribute. For example, in business analysis, companies are often interested in finding patterns that may be associated with customer spending; in medical domain, knowing which group of people are more likely to contract a disease and their associated patterns can help the prevention and the diagnosis of the disease. Therefore, we focus on studying the association between patterns and a pre-specified target attribute in our system. The target attribute is supplied by users based on their interest. It can be either categorical or continuous. Below we mainly focus on the case where the target attribute is categorical.

To simplify the problem, we assume that users specify one target attribute for one analysis task. Let A_{tgt} be the target attribute specified by users and it is categorical. We ask users to choose a value of A_{tgt} that is the most interesting to them, and we call this value *target value*, denoted as v_{tgt}. Given a pattern X, we formulate the corresponding association rule as $R : X \Rightarrow v_{tgt}$. As in traditional association rule mining, we define the support of R as $supp(X \bigcup \{A_{tgt} = v_{tgt}\})$, denoted as $supp(R)$, and the confidence of R as $supp(R)/supp(X)$, denoted as $conf(R)$. The support of X is called the *coverage* of R.

In our system, we use p-value as an additional interestingness measure. The p-value of a rule R, denoted as $p\text{-}value(R)$, measures the statistically significance of R. It is the probability of observing R or a rule more extreme than R given the two sides of R are independent. The lower the p-value, the less likely the rule occurs by random chance, thus the more statistically significant the rule is.

We use the algorithm described in [2] to mine patterns and store them in a structure called CFP-tree. The CFP-tree structure is specially designed for storing and querying patterns [2]. Sub-pattern search, super-pattern search, immediate sub-pattern/super-pattern search and exact match can be efficiently processed on a CFP-tree. These operations are frequently used in rule filtering, exploration and comparative analysis. The efficiency of CFP-tree in supporting various queries ensures that our system has a reasonable response time.

3. THE ASSOCEXPLORER SYSTEM

In this section, we use dataset *adult* downloaded from UCI machine learning repository [1] to illustrate our system. Dataset **adult** contains the demographical information of 32,561 adults. There is a class attribute on this dataset, and it indicates whether the annual income of a person exceeds 50K. The class attribute takes two values: "*>50K*" and "*<=50K*". We pick the class attribute as the target attribute, and "*>50K*" as the target value.

[1] `http://archive.ics.uci.edu/ml/datasets`

3.1 Global view

We use a scatter plot to provide a global view of all the rules. The X-axis is coverage of rules and the Y-axis is confidence. Rules are mapped to points in the two-dimensional space. From the scatter plot, users can also easily locate rules with high confidence and/or support. Figure 1 shows a snapshot of the AssocExplorer system. The lower-left panel displays the set of rules discovered from the *adult* dataset with minimum support threshold of 5%, minimum confidence threshold of 10% and maximum p-value of 0.05. The horizontal red line at 24.1% corresponds to the relative frequency of the target value "*>50K*" in the whole dataset.

Coloring length-1 rules. Length-1 rules are often of particular interest since they are simple and easy to comprehend. We highlight length-1 rules in the scatter plot using different colors as shown in Figure 1. Rules containing more than one item are colored in gray. Users can easily identify attributes and items that have strong association with the target value. Figure 1 shows that attributes *age*, *education*, *marital-status*, *occupation*, and *sex* are important for a person's income, while attribute *race* and *workclass* are not that important. The *native-country* attribute has no impact on income as most of the people in the dataset are Americans.

Coloring rules based on an attribute. Another coloring scheme is to color rules based on a selected attribute. Let A be an attribute selected by users. Rules that do not contain items of A are colored in gray. Rules that contain items of A are colored based on the items of A. This coloring scheme help users understand the impact of the items of A on the whole set of rules. For exmaple, users can have a rough idea on the frequency of the items of A in the whole set of rules. It is also easy to know whether attribute A or its items dominate other attributes/items. Figure 2 shows the rules colored based on attribute *education*. Most rules containing "*education=Bachelors*" have higher confidence than rules containing "*education= Some-college*" or "*education=HS-grad*". It means that people with a bachelor degree are more likely to earn more than 50K per year than people with a lower degree. We observe that one rule containing "*education=Bachelors*" has very low confidence. It lies at the lower-left corner of the plot. By taking a closer look, we find the rule contains "*marital-status=never-married*". It implies item "*marital-status=never-married*" is more dominant than item "*education=Bachelors*".

3.2 Zooming and filtering

The number of rules generated from a dataset can be very large. Points in the scatter plot may get over-crowded. Users can utilize zooming and filtering to focus on rules that are interesting to them. Using the zooming function, users can focus on a particular region of the scatter plot. It is equivalent to filtering by support and confidence.

The filtering functions are located at the upper-left panel of Figure 1. Users can do filtering based on p-value by specifying a range for the p-value. Rules which fall out of the specified range are removed from the scatter plot.

Users can also narrow down the set of rules to be examined

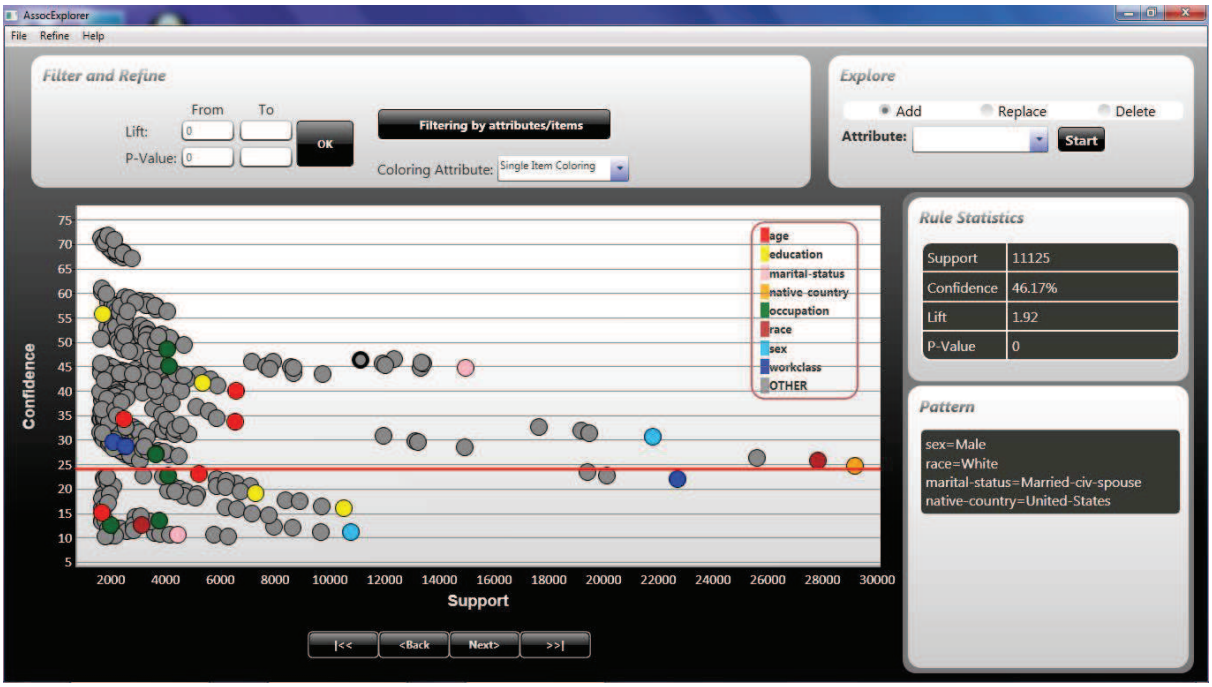

Figure 1: Overview of the AssocExplorer system

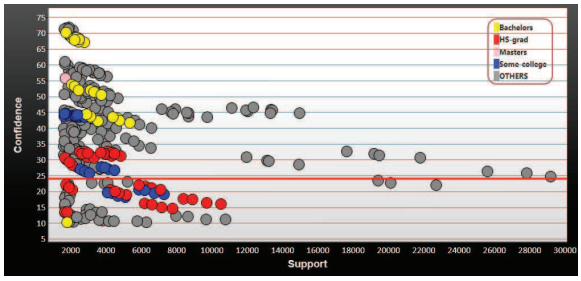

Figure 2: Coloring rules based on *education*

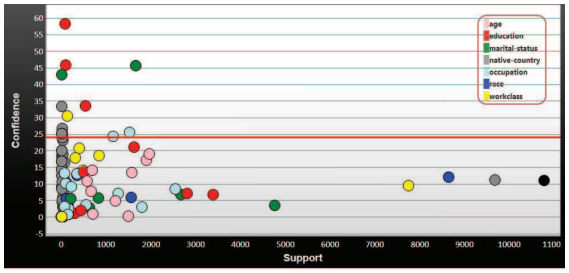

Figure 3: Adding one attribute to {sex=female}

by specifying constraints on attributes and items. Attributes are put into three categories: attributes that must appear in rules (the "include" category), attributes that cannot appear in rules (the "exclude" category) and attributes that may or may not appear in rules (the "optional" category). Similarly, items are put into the three categories in accordance with the status of attributes. Initially, all the attributes and items are set to be optional. When users change the status of attributes and items, the consistency between attributes and items is automatically enforced by our system.

3.3 Details on demand

The detailed information of a rule is displayed at the lower-right panel when the rule is clicked. In Figure 1, rule {*sex=Male*, *race=white*, *marital-status= Married-civ-spouse*, *native-country=United-State*} is clicked and it is highlighted in a bold black circle.

3.4 Exploring rules

When one rule is deemed to be interesting, it is often helpful to look at rules that are similar to it. We provide several operations to explore rules starting from a selected rule R: adding one more item to R, removing one item from R and replacing one item in R. Users can specify an attribute to be added, removed or replaced. If no attribute is specified, then any item whose attribute is outside R can be added, and any item inside R can be removed or be replaced by items from the same attribute. The resultant rules are still displayed in a scatter plot.

For the adding operation, the resultant rules are colored based on the added attributes. Rules that contain the same added attribute have the same color. Figure 3 shows the set of rules generated by adding one more attribute to rule {*sex=female*}. The original rule is colored in black, and it lies below the red line, which means that females are less likely to earn more than 50K than males. If we divide females into different age groups, we find that all the groups still lie below the red line. The situation is almost the same for attribute *occupation*. A few rules that are generated by adding items of attribute *education* or *marital-status* have high confidence. It implies that females can have higher income by getting a higher degree or changing marital status.

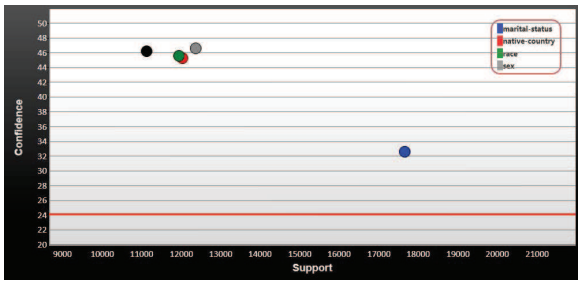

Figure 4: Remove one attribute from {*sex=Male, race=white, marital-status=Married-civ-spouse, native-country=United-State*}

Figure 4 shows the rules generated by removing one item from {*sex=Male*, *race=white*, *marital-status=Married-civ-spouse*, *native-country=United-State*}. The original rule is colored in black. If item "*marital-status=Married-civ-spouse*" is removed, the confidence of the rule (the blue point) decreases a lot. It suggests that this item is key to the confidence of the original rule. On the contrary, removing any of the other three items does not change the confidence of the rule much. It implies that the other three items do not have much impact on the association between the original pattern and the target value.

3.5 Comparative analysis

When we replace one item in R and the confidence of the resultant rule R' is quite different from that of R, we get a situation where a small change in items causes a big change in confidence. It is often worthwhile to further examine such situations as they may suggest inexpensive actions that we can take to change things toward a desired direction. Such knowledge is called actionable knowledge [4].

Rules R and R' differ on only one attribute. Let X_c be the set of common items of R and R', A_{diff} be the attribute on which R and R' differ, and v and v' be two values of A_{diff} that appear in R and R' respectively. Pattern X_c constitutes the context where the impact of A_{diff} on the target value v_{tgt} is studied. We call A_{diff} a *comparing attribute*, and its two values, v and v', *comparing items*.

To have a comprehensive understanding of the impact of A_{diff} on A_{tgt}, we provide operations to examine the impact of A_{diff} on A_{tgt} in other contexts: (1) on the whole dataset, that is, X_c is empty; (2) add one item to X_c; (3) remove one item from X_c; and (4) X_c contains one single item, and this item can be any item except for items of A_{diff}.

Figure 5 shows the comparison between *occupation= Admin-clerical* and *occupation = Craft-repair* in the context of {*race =White*}. The bar plot on the left shows that when *race =White*, craft repairers are more likely to earn more than 50K per year than administrative clerks. The two comparing items exhibit similar difference in confidence on the whole dataset. The right panel of Figure 5 displays how the difference changes when we add one more attribute to the context. The top two bar plots show that when items of attributes *age* and *education* are added to the context respectively, the difference between the two comparing items does not change much. The bottom bar plot shows that for both males and females, craft-repairer are slightly less likely to earn more than 50K than administrative clerks. This contradicts the

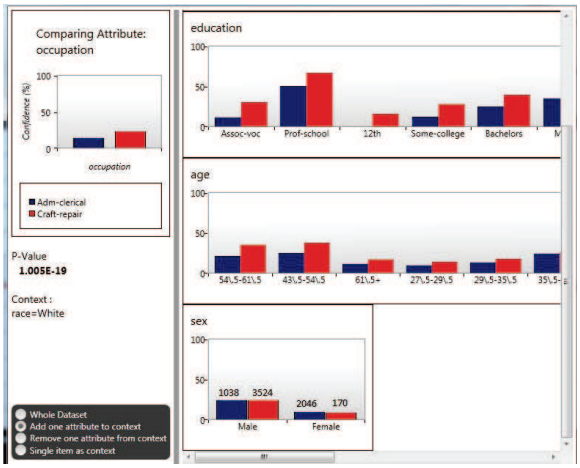

Figure 5: Comparing {*race=White, occupation= Admin-clerical*} with {*race=White, occupation= Craft-repair*}. The numbers on top of the bars are the support of the corresponding patterns.

original finding shown in the left panel. We have a Simpson's paradox. By taking a closer look, we find that attribute *sex* is a confounding factor here. It has association with both attribute *occupation* and the target value: men generally earn more than women and many more men work as craft repairers than women. We use pie charts to visualize the interactions between comparing items and the added attribute, which are accessible from the right-click menu.

3.6 Managing history

In a data mining task, users often do not have a clear idea on what they are looking for at the beginning. They may need to iteratively examine rules from different points of view to find things that are interesting to them. To help users go back to results produced by previous operations, we store previous results in a sequence and we call it the history sequence. Users can move backward and forward along the history sequence. In Figure 1, the history bar is located under the scatter plot.

4. ACKNOWLEDGMENT

This work is supported in part by Singapore Agency for Science, Technology and Research grant SERC 102 101 0030.

5. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In *Proc. of the 1993 ACM SIGMOD Conference*, pages 207–216, 1993.

[2] G. Liu, H. Lu, and J. X. Yu. Cfp-tree: A compact disk-based structure for storing and querying frequent itemsets. *Information Systems*, 32(2):295–319, 2007.

[3] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proc. of the IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[4] Q. Yang, J. Yin, C. X. Ling, and R. Pan. Extracting actionable knowledge from decision trees. *IEEE Trans. Knowl. Data Eng.*, 19(1):43–56, 2007.