An Efficient Discovery of Class-Restricted MARs

Hyontai Sug
Division of Computer & Information Engineering
Dongseo University
Busan, 617-716
REPUBLIC OF KOREA
hyontai@yahoo.com    http://kowon.dongseo.ac.kr/~sht

Abstract: - Because the target databases of conventional multidimensional association rule algorithms have no distinction in the role of attributes, a lot of rules can be found and the computing time can be enormous. So, some efficient way is needed to find multidimensional association rules for a specific class for target database table that has a decision attribute and many conditional attributes. In order to overcome the problem of intensive computing time and possibly generating a lot of uninteresting rules, a preprocessing technique that can narrow down the search space is suggested. The method can generate smaller table for multidimensional association rule so that computing time can be saved and smaller number of rules are generated. Experiments with a real world data set showed a very good result.

Key-Words: - Multidimensional association rules, efficient discovery, conditions

1 Introduction
Association rules are rules describing association patterns that indicate how likely a set of items occur together. These patterns reveal some information on regularities that exist in a database. For example, suppose we have collected a set of transactional data records for sales in a super market for some period of time. We might find an association rule from the data stating, “bread => jam (70%),” which means “70% of customer who buy bread also buy jam.” Such association rules can be helpful for decision making on sales, e.g., designing an appropriate layout for items in the store.

Multidimensional association rules are rules that can be found from a table-like database. The difference between association rules and multidimensional association rules is that each item in multidimensional association rules has distinct attributes, and there is no such attribute in the items of association rules. In that sense association rules are one dimensional multidimensional association rules. So, each item in multidimensional association rules is a pair {attribute, a value of the attribute}. If the attributes of the table-like database can be divided into two kinds of attributes, condition and decision attributes, we can find multidimensional association rules for classification.

Unlike conventional multidimensional association rule discovery algorithms in which there is no predefined distinction between the role of attributes, we want to give attention to the table that has a decision attribute and many conditional attributes, and each conditional attribute is independent on each other, and dependent only on the decision attribute. So, computing time to generate rules can be saved and smaller number of rules can be generated than the case when rules are generated by conventional multidimensional association rule algorithms. Moreover, if we want to find multidimensional association rules for a specific class having specific feature vectors, for example, we want to find rules to decide HIV patient is sick, and the disease is minor compared to other disease, so the instances that indicate this disease belongs to a minor class, so that it is highly possible that we may have many uninteresting rules, if we apply conventional multidimensional association rule finding method without doing any step. Therefore, we want to find multidimensional association rules more effectively for the case.

In section 2, we provide the related work to our research, and in sections 3 we present our method. Experiments were run to see the effect of the method in section 4. Finally section 5 provides some conclusions.

2 Related Work
Association rule discovery systems [1, 2, 3] were developed to find association rules that indicate how often a set of items occur together in transaction databases. These rules give information on association patterns that exist in the database. Many good algorithms were suggested to find association rules efficiently. For example, a standard association algorithm, Apriori, large main memory-based algorithm like AprioriTid [1], the hash-table based algorithm, DHP [4], random sample based algorithms [5], tree structure-based algorithm [6] or even a parallel version of the algorithm [7]. Other related work is multidimensional association rules.

Multidimensional association rules are basically an
application of general association rule algorithms to table-like databases. The table-like databases may consist of condition attributes and decision attributes. In papers like [8, 9], multidimensional association rules have better accuracy than decision trees for most of example data in small size. However, these algorithms generate, potentially, a large number of rules and manifest facts, so rule selection based on interest measure [10, 11] and generalization is important [12, 13]. Hybrid association rules are a generalization technique to discover more interesting association rules. Hybrid-dimension association rules were found by treating attributes as main and subordinate [14]. In order to mine the database having index structures hybrid dimensional association rules are suggested [15]. There was also efforts to reduce computing time. MPNAR algorithm [16] divides the datasets into infrequent itemsets and frequent itemsets and discovers multidimensional association rules for infrequent itemsets as well as frequent itemsets. The multidimensional association rules mined from infrequent itemsets are called negative association rules, and the multidimensional association rules mined from frequent itemsets are called positive association rules. So, the algorithm neglects itemsets between being frequent and infrequent, and negative association rules may not be useful due to the infrequency.

Rough set theory considers dependency between attribute values solely based on data. Researchers tried to investigate attribute dependency in algebraic aspects [17], or in statistical aspects [18]. ROSETTA [19] and RSES [20] are some examples of data mining tools to find decision rules based on rough set theory. Because the target database for data mining system based on rough set theory and multidimensional association rules is similar, there are some research results that consider both techniques. Yang et al. [21] pre-processed a transaction database based on constraints of profit and frequency to make a decision table which can be supplied to rough set theory-based data mining method.

3 Problem Solution

The following is a formal definition of association rules we are interested. Let $I = \{i_1, i_2, ..., i_m\}$ be a set of items that are in a table, where $i_j$ is an attribute-value pair. Let $T$ be a collection of records, where each record has a set of itemset $X \subseteq I$.

A multidimensional association rule is an implication of the form $Y \rightarrow Z$ where $Y \subseteq I$, $Z \subseteq I$, and $Y \cap Z \neq \phi$. The confidence $C\%$ of the association rule $Y \rightarrow Z$ implies that among all the records which contain itemset $Y$, $C\%$ of them also contains itemset $Z$. For an itemset $X \subseteq I$, $\text{support}(X) = s$ is the fraction of the records in $T$ containing $X$. So, the confidence of a rule $Y \rightarrow Z$ is computed as the ratio, $\text{support}(Y \cap Z)/\text{support}(Y)$. Note that $Y \cap Z$ means itemsets in records. If we consider actual count of itemsets, $\text{support number}$ can used to represent how many times an itemset occurs in $T$. The itemsets that occur more frequently than some given support level are called frequent itemsets.

We are interested in more efficient way in discovering multidimensional association rules for specific classes like minor classes. Conventional multidimensional association rule finding algorithms also can find those rules, but all the other multidimensional association rules are found also so that much computing time is needed as well as possibly a lot of rules are generated. Because we are interested in multidimensional association rules for a specific class, we can limit our search to the records having itemset $Y$ only, if we want to find multidimensional association rule like $Y \rightarrow Z$. So, we can reduce the size of data record a lot. The following is a brief description of the procedure of the method.

```
INPUT:
T: a decisional table for data mining,
CF: minimum confidence of rules,
MS: minimum support number,
S: the table of records of items in the interesting class.

OUTPUT: a set of multidimensional association rules.

Begin
D := Select all records from T that have any one item in each record of S;
Apply multidimensional association rule finding algorithm to D with MS;
Generate multidimensional association rules having greater confidence than or equal to CF;
Calculate the confidence of rules based on T;
Eliminate such rules that have confidence < CF;
If a rule’s condition part is shorter and the confidence of the rule is greater than or equal to some other rule Then
select the rule of shorter condition part;
End;
```

In the algorithm each record of $S$ consists of items in each record that belong to the interesting class. We select records that have any same item in the records of interesting class from the target database $T$. By doing so, the target database size for data mining can be reduced. As a result, we may have smaller number of multidimensional association rules and save computing time. By Eliminating any such rules that have smaller confidence than given CF after calculating the confidence of rules based on $T$, we can have correct confidence of
each rule. In the following experiment the given CF value is 65% and minimum support number is two, because the interesting class consists of relatively small number of instances.

4 Experimentation

Experiments were run using a database in UCI machine learning repository [22] called 'Statlog(shuttle)' [23] to see the effect of the method. The number of instances in the data set for training is 43,500. The data sets were selected, because it is relatively large and consists of many classes. The total number of attributes is ten, named A to J, and there are seven classes. All attributes are continuous attributes for stalog data set. Class distribution is in table 1.

<table>
<thead>
<tr>
<th>Class</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78.41%</td>
</tr>
<tr>
<td>2</td>
<td>0.09%</td>
</tr>
<tr>
<td>3</td>
<td>0.3%</td>
</tr>
<tr>
<td>4</td>
<td>15.51%</td>
</tr>
<tr>
<td>5</td>
<td>5.65%</td>
</tr>
<tr>
<td>6</td>
<td>0.01%</td>
</tr>
<tr>
<td>7</td>
<td>0.03%</td>
</tr>
</tbody>
</table>

We are interested in multidimensional association rules for class 6. Before applying association rule finding algorithm, decritization based on entropy minimization heuristic [24] is performed. So, the records that have any same item with the records of class 6 are selected. The total number of the selected records is 4,902. So the size is reduced to 11.27%.

A modified Ariori algorithm is used to generate multidimensional association rules. Note that we can find frequent itemsets more efficiently than the original algorithm, because an item consists of attribute-values pair and there is no duplicate attribute in an itemset.

The following 15 rules are the result. If a rule’s condition part is shorter and the confidence of the rule is greater than or equal to some other rules, then the rule with shorter condition part has been selected. In attribute value ‘{’ means open bracket, and ‘[’ means closed bracket, and ‘freq.x/y’ means the frequency of condition itemsets is x, and the frequency of condition and decision itemsets is y.

If A=(54.5~55.5] and B=(586~∞) Then CLASS=6 (CF: 1, freq.=2/2);
If A=(56.5~68.5] and B=(586~∞) Then CLASS=6 (CF: 1, freq.=2/2);
If B=(586~∞) and G=(15.5~20.5] Then CLASS=6 (CF: 1, freq.=2/2);
If B=(586~∞) and I=(21~27] Then CLASS=6 (CF: 1, freq.=2/2);
If B=(586~∞) and I=(31~37] Then CLASS=6 (CF: 1, freq.=2/2);
If B=(586~∞) and E=(17~25] and F=(-0.5~0.5] Then CLASS=6 (CF: 1, freq.=2/2);
If B=(586~∞) and F=[-0.5~0.5] and H=(51.5~57.5] Then CLASS=6 (CF: 1, freq.=2/2);
If B=(586~∞) and D=(-0.5~0.5] and E=(17~25] and F=(-0.5~0.5] and H=(51.5~57.5] Then CLASS=6 (CF: 1, freq.=2/2);
If B=(586~∞) and D=(-0.5~0.5] and E=(17~25] and F=(-0.5~0.5] and H=(51.5~57.5] Then CLASS=6 (CF: 0.67, freq.=2/3);
If B=(586~∞) and D=(-0.5~0.5] and E=(17~25] and F=(-0.5~0.5] and H=(51.5~57.5] Then CLASS=6 (CF: 0.67, freq.=2/3);
If B=(586~∞) and D=(-0.5~0.5] and E=(17~25] and F=(-0.5~0.5] and H=(51.5~57.5] Then CLASS=6 (CF: 0.67, freq.=2/3);
If B=(586~∞) and D=(-0.5~0.5] and E=(17~25] and F=(-0.5~0.5] and H=(51.5~57.5] Then CLASS=6 (CF: 0.67, freq.=2/3);

5 Conclusion

Multidimensional association rules are association rules that can be found from a table-like database. The difference between association rules and multidimensional association rules is that each item in multidimensional association rules has distinct attributes, while there is no such attribute in the items of association rules. Moreover, if the attributes of the table-like database can be divided into two different kinds of attributes, condition and decision attributes, we can find multidimensional association rules for classification task.

Because the target databases of conventional multidimensional association rule algorithms have no distinction in the role of attributes, a lot of rules can be found and the computing time can be enormous. So, we are interested in some efficient way to find multidimensional association rules for a specific class.
from the table that has a decision attribute and many conditional attributes. Moreover, if we find multidimensional association rules for a specific class, it is highly possible that we may have more interesting rules for the class, otherwise we may have many uninteresting rules by the conventional multidimensional association rule finding method.

In order to overcome the problem of intensive computing time and possibly generating a lot of uninteresting rules, a preprocessing technique that can narrow down the search space is suggested, and it can generate smaller table for the data mining of multidimensional association rules so that computing time can be saved and smaller number of rules are generated. Experiments with a real world data set showed a very good result.

References:
[3] J. Han, J. Pei, Y. Yin, R. Mao, Mining frequent patterns without candidate generation, Data Mining and Knowledge Discovery, Vol. 8, 2004, pp. 53-87.
[5] H. Toivonen, Discovery of Frequent Patterns in Large Data Collections, PhD thesis, Department of Computer Science, University of Helsinki, Finland, 1996.
[6] J. Han, J. Pei, Y. Yin, R. Mao, Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach, Data Mining and Knowledge Discovery, Vol. 8, 2004, pp. 53-87.
[8] W. Li, J. Han, J. Pei, CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules, Proceedings 2001 Int. Conf. on Data Mining (ICDM'01), 2001, pp. 369-376.

