Analysis of the data with multiple features in retail domain

Mrs.Bhavini A.Shah, Vaishali Suryawanshi
Lecturer,G.S.Moze College of Engineering,Pune
Assistant Professor,MIT College of Engineering,Pune

Abstract— In enterprise data mining applications it is crucial to develop effective techniques for mining rules combining necessary information from multiple relevant business areas, catering for real business settings and decision-making actions rather than just providing a single line of patterns. Here we adopt combined mining as a general approach to mining for informative rules combining components of multiple features. Association rules are one of the most impressive techniques for the analysis of attribute associations in a given dataset related to applications related to retail, bioinformatics, and sociology, but often produces large collections of association rules that are difficult to understand and put into action. Through classical association mining many redundant rules are generated which may not be useful for business analysis. The proposed framework helps in generating the combined rules which gives informative knowledge for business by combining static and transactional data mining combined patterns to extract useful and actionable knowledge from a large amount of learned rules. Experimental results on retail data set demonstrate the effectiveness and potential of the proposed approach in extracting actionable knowledge from complex data.

Key words- Data mining, Decision making, actionable knowledge

I. INTRODUCTION

With the accumulation of ubiquitous enterprise data, there is an increasing need to mine for such informative knowledge in complex data. It is challenging to mine for comprehensive and informative knowledge in such complex data suited to real-life decision needs by using the existing methods. The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple featured data sets. It is often very costly and sometimes impossible to join multiple features data set into a single data set for pattern mining. The idea of association rules [1] was proposed 15 years ago and is widely used today. However sometimes be very difficult for users to not only understand such rules, but also find them a useful source of knowledge to apply to their business processes. Therefore, to present associations in an interesting and effective way, and in order to find actionable knowledge from resultant association rules, an ideal approach of combined mining is proposed. Combined mining is a technique for analysing object relations and pattern relations, and for extracting and constructing actionable complex knowledge in complex situations. Combined patterns comprise combined association rules, combined rule pairs and combined rule clusters. The proposed combined patterns provide more interesting knowledge and more actionable results than traditional association rules. The paper is organized as follows. Section 2 briefly introduces some related work on domain-driven data mining and multi-relational data mining. The problem to be addressed and its business background are introduced in Section 3. The proposed idea and framework to mine the combined rules is described in details in Section 4. Section 5 gives experimental results and conclusions are given in Section 6.

II. Related Work

There are often too many association rules discovered from a dataset and it is necessary to conduct post-processing before a user is able to study the rules and identify interesting ones from them. There are many techniques proposed to summarize and/or post-analyse the learned association rules [6]. Hilderman et al. proposed to characterize item sets with information from external databases, e.g., customer or lifestyle data [2]. Their technique works by firstly mining frequent item sets from transactional data and then partitioning each frequent item set according to the corresponding characteristic tuple. This method likely results in a large number of rules when many characteristics are involved, with every characteristic having multiple values. Liu and Hsu proposed to rank learned rules by matching against expected patterns provided by user [4]. Rule Similarity and Rule Difference are defined to compare the difference between two rules based on their conditions and consequents, and Set Similarity and Set Difference are defined to measure the similarity between two sets of
rules. In another work, Liu et al. proposed to mine for class association rules and build a classifier based on the rules [5]. With their rule generator, the rule with the highest confidence is chosen from all the rules having the same conditions but different consequents. Liu et al. also proposed direction setting rules to prune and summarize association rules [6]. Domingos argues that multirelational data mining plays a key role in KDD. Cristosfor and Simovici designed a couple of algorithms to address the problem for mining association rules in databases consisting of multiple tables and designed using the entity-relationship model [9]. Chattratichat et al designed a Kensington software architecture for distributed enterprise data mining, which addresses the problem of data mining on logical and physical distribution of data and heterogeneous computational resources [8].

III. The Problem

The example that follows illustrates the target problem. Suppose that there are two datasets, transactional dataset and customer demographic dataset (see Tables 1 and 2), where “Income” class for dividing the customers depending on the salary of a customer i.e. High, Low and Moderate.

Table 1. Customer data

<table>
<thead>
<tr>
<th>Customer ID</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
</tr>
<tr>
<td>3</td>
<td>M</td>
</tr>
<tr>
<td>4</td>
<td>M</td>
</tr>
</tbody>
</table>

Table 2. Customer Transaction Data

L->Low, H->High, M->Moderate (c1, c2, c3, c4, c5)->Products
This table consist of the details of the transactions made by the customer for ex Customer id 1 purchased products c1, c2 with the income category low.

Table 3. Support, Confidence and Lift calculated using traditional algorithms

<table>
<thead>
<tr>
<th>Rules</th>
<th>Supp</th>
<th>Conf</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-&gt;L</td>
<td>3/10</td>
<td>3/5</td>
<td>1.2</td>
</tr>
<tr>
<td>F-&gt;H</td>
<td>3/10</td>
<td>3/5</td>
<td>1.2</td>
</tr>
<tr>
<td>M-&gt;L</td>
<td>2/10</td>
<td>2/5</td>
<td>0.8</td>
</tr>
<tr>
<td>M-&gt;H</td>
<td>3/10</td>
<td>3/5</td>
<td>1.2</td>
</tr>
</tbody>
</table>

In this table we are applying traditional method to find association rules. From the above table the rules generated are not completely informative

Table 4. Combined Association Rules

From this table we get a clear picture that a female customer is more likely to buy product c1 who belong to low income category and product c2 who belong to high income category.

The traditional association rules discovered are shown in Table 3, and the four rules with lift greater than one are F
L, M
H, and c1
L, and c3
H. If partitioning the whole population into two groups, male and female, based on the demographic data in Table 2, and then mining the two groups separately, some rules are shown in Table 4, where Lift 1 and Lift 2 denote respectively the lift of the first/second part of the left side, and Irule is the interestingness of the combined rule. We can see from Table 4 that more rules with high confidence and lift can be found by combining the rules from two separate datasets. Although all the rules in Table 4 are of the same confidence and lift, their interestingness is not the same, which is shown by the last column Irule. For example, for the first rule in Table 4, F
L, its interestingness Irule is 0.8, which indicates that the rule is not interesting at all. The explanation is that its lift is the same as the lift of c1
L (see Table 3), which means that F contributes nothing in the rule. Therefore, the new measures are more useful than the traditional confidence and lift. It is more interesting to organize the rules into contrasting pairs shown in Table 5, where Irule is the interestingness of the rule pair. P1 is a rule pair for male group, and it shows that c1 is associated with income but c2 with stay. P1 is actionable in that it suggests c2 is a preferred product.
purchased by the male customers. Moreover, male customers should be excluded when initiating purchase of c1. P2 is a rule pair with the same campaign but different demographics. With the same action c2, male customers tend to purchase, but female do not. It suggests that c2 is a preferable product for male customers but an undesirable product for female customers. From the previous example, we can see that rule pairs like P1 and P2 provide more information and are more useful and actionable than traditional simple rules shown in Table 3 and in this paper, they are referred to as combined patterns. A straightforward way to find the rules in Table 4 is to join Tables 1 and 2 in a pre-processing stage and then apply traditional association rule mining to the derived table. Unfortunately, it is often infeasible to do so in many applications where a dataset contains hundreds of thousands of records or more.

<table>
<thead>
<tr>
<th>Table 5. Combined Association Rule Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pairs</td>
</tr>
<tr>
<td>P1</td>
</tr>
<tr>
<td>P2</td>
</tr>
</tbody>
</table>

Here we are generating pairs depending on different features. It is more interesting to organize the rules into contrast pairs as shown in Table 5, where IPair is the interestingness of a rule pair. For instance, P1 is a rule pair for the male group, and it shows that c3 is associated with low income group customers but c2 is not. P1 is actionaline in that it suggests that c2 is a prefered product for purchase by male customers. P2 is a rule pair with the same policy but different demographics. With action c2, male customers buy belonging to high income category while the females tend to c2 belonging to low income category. It suggests that c2 is a preferable product for males but undesirable for females.

IV. Generation of combined data

For generating combined data we apply the following algorithm

Domain Knowledge Method:
1. List both the tables i.e. transactional and demographic tables at run time
2. Select the tables in which the user is interested
3. List the characteristics/features of selected table
4. List the distinct values of each selected characteristics of selected table
5. Select only the interested values of listed features

Algorithm for pruning: // to get the pruned rules Input:
Confidence, Lift, and Patterns Generated on combined data above the support
1. For each pattern p
   a. If Confidence(p) >=minimum confidence and Lift(p) >=1
   b. End if
   c. Return true
   d. End if
Output: set of pruned rules // Pattern string checking function Check_PATTERN_STRING(p)
1. For each pattern p (X → T)
   a. Check for left hand side of the rule
   b. If X contains {A1 ^ B1}
   i. Consider it for combined rule
   ii. Return true
   iii. Return false
   iv. End for

V. EXPERIMENTAL RESULTS

The proposed technique is tested on the subset of retail demo dataset used for presenting Microsoft Business Intelligence products which is available on microsoft.com. The sample subset of data contained 290 transactional records of 104 customers for 5 different products. Customer data was classified as High, Moderate and Low customers based on their features. The aim of the experiment was to find the association of demographic features of customer, product buying pattern and the class of customers which could help to give a promotional campaign on different products based on customer class and their demographic feature. We used SQL server 2008 and .net technology for implementation purpose. For experiment purpose 5 products from transactional dataset and from the customer table two features Gender (Male, Female) and Marital Status(S-Single, M-Married) were selected through domain knowledge concept. First traditional association mining is applied on transactional and demographic data separately. Subset of traditional association Rules for Transactional dataset and Customer dataset are shown respectively in Table 3 and Table 4. Minimum support was set to 10% and minimum confidence was set to 20%. We got 56 rules for transactional dataset 19 rules for demographic dataset which are above minimum support, confidence and Lift>=1.

Table 3. Traditional association rules for transactional data set
The concept of feature selection is given numbers of rules.

1. **Domain driven concept** gives the idea of reduced and also readable and understandable to human to take business decision and can reduce the cost of promotion. For specific products, the rules can be generated with domain knowledge technique. On the framework designed, the concept of feature selection is given which reduces the computational cost and space complexity also.

### Table 4. Traditional Association rules for demographic data set

<table>
<thead>
<tr>
<th>Rule (B -&gt; T)</th>
<th>Support (%)</th>
<th>Confidence (%)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wireless Headphones, PlayStation-&gt;Low</td>
<td>13/104</td>
<td>43</td>
<td>1.96</td>
</tr>
<tr>
<td>Wireless Headphones, PlayStation-&gt;Moderate</td>
<td>17/104</td>
<td>57</td>
<td>1.03</td>
</tr>
<tr>
<td>LCD, Portable DVD Player-&gt;Moderate</td>
<td>13/104</td>
<td>22</td>
<td>0.98</td>
</tr>
<tr>
<td>LCD, Portable DVD Player-&gt;Low</td>
<td>37/104</td>
<td>62</td>
<td>1.13</td>
</tr>
</tbody>
</table>

### Table 5. Combined Association Rules

<table>
<thead>
<tr>
<th>Rule (A -&gt; T)</th>
<th>Support (%)</th>
<th>Confidence (%)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>M, Female-&gt;High</td>
<td>12/104</td>
<td>38</td>
<td>1.62</td>
</tr>
<tr>
<td>M-&gt;High</td>
<td>19/104</td>
<td>31</td>
<td>1.35</td>
</tr>
<tr>
<td>Female-&gt;High</td>
<td>16/104</td>
<td>30</td>
<td>1.31</td>
</tr>
<tr>
<td>S, Male-&gt;Low</td>
<td>15/104</td>
<td>68</td>
<td>1.24</td>
</tr>
<tr>
<td>S, Male-&gt;Moderate</td>
<td>6/104</td>
<td>27</td>
<td>1.23</td>
</tr>
<tr>
<td>S-&gt;Low</td>
<td>28/104</td>
<td>65</td>
<td>1.19</td>
</tr>
</tbody>
</table>

After pruning and combined mining, the rules generated with both transactional and demographic features are shown in Table 5. Total 329 combined rules have been generated. Such types of knowledge help to give the promotional campaign on products with respect to the class of customer and their demographic features. The rules are readable and understandable to human to take business decision and can reduce the cost of promotion. For specific products, the rules can be generated with domain knowledge technique. On the framework designed, the concept of feature selection is given which reduces the computational cost and space complexity also.

### VI. Conclusion

This paper presents the new idea about pruning the association rules before making the rule pairs or rule clusters. The domain driven concept gives the idea of selection of data features before generating the frequent patterns. So the user have freedom to select the product on which the company wants to give promotional campaign before generating the rules. Through domain driven user can select the customer characteristics from static customer data. The rule pairs can also be generated from pruned frequent patterns, or closed frequent patterns or maximal frequent patterns. So the numbers of rules generated get reduced and also readable and understandable to the user. This new technique is giving the cluster of rules for similar type of customers with 10 their change of class as the transactional characteristic changes. This proposed technique gives more actionable rules than traditional association technique which help to improve business process.

### References


[5]. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD 1998).

[6]. Liu, B., Hsu, W., Ma, Y.: Pruning and summarizing the discovered associations. In: Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 1999).

