Retail Marketing Segmentation and Customer Profiling
for Forecasting Sales

D.BHANU
Department of Computer Applications
Sri Krishna College of Engineering and Technology
Kuniamuthur, Coimbatore, TamilNadu
INDIA
bhanu.saran@gmail.com

S.PAVAI MADHESWARI
Department of Computer Science and Engineering
R.M.K.Engineering College
Chennai, TamilNadu
INDIA

Abstract: - Product Bundling and offering products to customers is of critical importance in retail marketing. In this paper, a predictive mining approach is presented that predicts sales for a new location based on the existing data. The major issue lies in the analysis of sales forecast based on the dependencies among the products and the different segment of customers, which helps to improve the market of the retail stores. The work presents a framework, which models an association relation mapping between the customers and the clusters of products they purchase in an existing location and helps in finding rules for an entirely new location. A novel methodology and model are proposed for accomplishing the task efficiently. The methodology is based on the integration of the popular data mining approaches such as clustering and association rule mining. It focuses on the discovery of rules that vary according to the economic and demographic characteristics and concentrates on marketing the products based on the population.

Key-Words: - Predictive Mining; Product Bundling; Data Mining; Association rules; Clustering

1 Introduction

Data mining is often defined as either finding hidden information in databases automatically or semi-
automatically. It involves many different techniques to accomplish different tasks. Our work is based on Association rule mining and clustering, which are among the frequently used approaches so far. Association rules generated for a location at a point of sale cannot be effective in another location since the complete and complex behaviour of customers and their approach in selecting products are different. The relationship between the customer requirements and the practical supplies that the stores can provide to the customers are often not readily available for a different store at an entirely different location. The major challenge of our work is the identification of rules for a different demographic population based on the existing records.

2 Literature Survey
Since the introduction of the problem of mining association rules [1], several generate-and-test type of algorithms have been proposed for the task of discovering frequent sets. An efficient breadth-first or level-wise method for generating candidate sets, i.e., potentially frequent sets has been proposed in [3] [4] [5]. The method also called Apriori is the core of all known algorithms except the original one [1] and its variation for SQL, which have been shown inferior to the level-wise method.

An alternative strategy for a database pass, using inverted structures and a general purpose DBMS, has been considered. Other work related to association rules includes the problem of mining rules with generalization [5], management of large amount of discovered rules [17] and a theoretical analysis of an algorithm for a class of KDD problems including the discovery of frequent sets [7]. A connectionist approach to mining rules is presented in [18]. All these approaches focus on generating rules, but a little attention is given to address the credibility of the rules outside of the dataset from which it was generated.

3 Architecture of the Clustering Based Association Rule Mining System
In order to obtain the association rules for a new store based on the analysis of customer transactions from the existing knowledge base, Clustering based Association Rule Mining System architecture [CARMS] to predict sales at a different location is proposed. The system involves different consecutive stages communicating with one another in generating rules as the data preprocessing and data partitioning, data transformation, association rule mining and presentation modules. Before proceeding to the rule mining of datasets, raw data must be preprocessed in order to be useful for knowledge discovery. Due to the uncertainty of customer requirements and their behavior, it is necessary to preprocess the knowledge base. Based on the raw data stored in the knowledge base, target datasets should be identified, involving such data cleaning and filtering tasks as integration of multiple databases, removal of noises, handling of missing data files etc.,

![Figure 1: Specification of Problem Domain](image1)

![Figure 2: Rule identification for a new setting based on Association Rule Mining](image2)
All target data should be organized into a usable transaction database. This involves the clear understanding of the variables, selection of attributes, which are more pertinent in generating rules for the new store. In the architecture proposed, the sales records and the product details are transformed into transaction data, which consists of a unique transaction identifier (TID). Transaction data consists of customer details and their affinity towards the products. Each customer is given a series of options on the selection of products based on the customer’s attributes such as income, age and gender, which are recoded as the key operational features. The options of products that the customer desires can be stated as related Functional Requirements, which can serve as mandatory information for the predicting of sales at a new location. The result of the Customer Needs and the Functional Requirement mappings are also recorded in the transaction database. Figure 3 shows the Block Diagram of CARMS, which maps the customer needs and the Functional requirements and helps in finding rules for a new setting based on the existing transaccional data.

4. The Fuzzy Clustering Method for the formation of groups

The behavior of customers is often uncertain and vague and hence is frequently presented in linguistic forms. A concrete definition of customer behavior is not possible and is a rare case. For the proposed architecture, the Fuzzy clustering includes the following steps as distance measure and fuzzy clustering. As the preparatory stage for fuzzy clustering, the distance measure module measures the dissimilarity between the functional requirement instances in order to find the fuzzy compatible relations among such data objects.

4.1 Data Preprocessing

Many data types are available in fuzzy clustering, such as binary variables, ordinal variables and ratio-scaled variables and even a mix of these variables. Because of the influences in the demography at different locations, the data standardization should also be considered. The data standardization for numerical features alone is specified for the current discussion, as all the other data types can be converted to the numerical data type. Let $F_i = (f_{i1}, f_{i2}, \ldots, f_{iM})$ for all $i \in [1, 2, \ldots, T]$ be the transaction record, where $M$ denotes the product features. A standard deviation equation is applied on the raw data to standardize the data.

Suppose $f_{ik}' = f_{ik} - \bar{f}_k / \sigma_k$ for all $i \in [1, 2, \ldots, T]$, $k \in [1, 2, \ldots, M]$, where $f_{ik}$ corresponds to the option value of the $i$th transaction record for the specification $f_k$, where

$$f_{ik} = \frac{1}{T} \sum_{i=1}^{T} f_{ik}' \text{.}$$

It denotes the average value of all the transactions as to the $k$th specification, where

$$\sigma_k = \sqrt{\frac{1}{T} \sum_{i=1}^{T} | f_{ik}' - \bar{f}_k |^2} \text{.}$$

Unlike the max-min standardization, it cannot transform the raw data to values between 0 to 1, but eliminate the influence of different dimensions. Figure 4 shows the method for preprocessing for customer need identification and Figure 5 shows the preprocessing for product requirement need identification.
4.1.1 Fuzzy clustering

Three main methods are available to facilitate fuzzy clustering, including fuzzy netting graph, maximum generated graph and transitive closure method. The former two methods can be implemented through the fuzzy compatible matrix directly, while the latter methods need to convert fuzzy compatible matrix to fuzzy equivalent matrix. It is then transformed into a Boolean equivalent matrix. It means that besides the fuzzy compatible relation that is composed of both symmetric and reflexive relations, the fuzzy equivalent relation utilized to construct the fuzzy equivalent matrix should have the transitive relation.

Here, we adopt the Boolean equivalent matrix to obtain the clusters of customer needs and product features, respectively. For the customer transaction records \( C_i \), \( R \) is transitive iff

\[ R^k R \leq R \quad (\Leftrightarrow \bigvee_{k=1}^{T} (\rho (C_i, C_k) \land \rho (C_k, C_i))) \]  

(2)

Similarly, for the transaction record \( F_i \), \( G \) is transitive iff \( G^k G \leq G \). The process is performed by the max-min transitive closure for \( R \) and \( G \) respectively. This method is based on the following model:

\[ t (R) = \bigcup_{k=1}^{T} R^k, \]  

(3)

where \( t (R) \) is transitive closure and \( R^k \) denotes the max-min operation for \( k \) items. The above model implies that the transitive closure \( t (R) \) can be obtained by max-min operation for most \( T \) times. In this way, the transitive relation is satisfied and the fuzzy compatible matrix \( R_{T \times T} \) and \( G_{T \times T} \) are converted into the fuzzy equivalent matrix \( t (R) \) and \( t (G) \) respectively.

Figure 6 shows the Fuzzy clustering method for grouping customers by converting the customer needs and functional requirements onto a fuzzy equivalent matrix.

4.1.2 Rule Evaluation

The Apriori algorithm is the most well known association rule-mining algorithm. At first, we give the following definitions [10]:

Definition 1 : Given a set of items \( I = \{ I_1, I_2, \ldots, I_n \} \), and the database of transaction records \( D = \{ t_1, t_2, \ldots, t_m \} \), where \( t_i = \{ I_{i1}, I_{i2}, \ldots, I_{ik} \} \) and \( I_{ji} \in I \), an association rule is an implication of the form \( X \Rightarrow Y \) where \( X, Y \subset I \) and \( X \cap Y = \Phi \).
Definition 2: The support (s) for an association rule \( X \Rightarrow Y \) is the percentage of transactions in the database that contain \( X \cup Y \). That is, support \( (X \Rightarrow Y) = P(X \cup Y) \), \( P \) is the probability.

Definition 3: The confidence or strength (\( \Phi \)) for an association rule \( (X \Rightarrow Y) \) is the ratio of the number of transactions that contain \( X \cup Y \) to the number of transactions that contains \( X \). That is confidence \( (X \Rightarrow Y) = P(Y|X) \).

The algorithm uses the following property: If an itemset satisfies the minimum support threshold, so do all its subsets. The key of Apriori algorithm is to generate the large itemsets and then to generate association rules. The details of the algorithm are more specifically given in [10].

The unions of transaction records in the clusters that make the dependency maximum are often more representative than other transaction ones. Therefore, we can partition the target transaction table with them to decrease the scale of data mining without loss of the information content. In general, the focus must be more on the cluster groups than the individual customers, since the groups can reflect the characteristics of individual customers.

Based on the previous considerations, the rule mining process can be divided into three steps:

1. Choosing the suitable product clusters, which make the dependency maximum;
2. Calculating the mean values and the corresponding variation as to specification for each transaction in the cluster.
3. Choosing the unions of transaction records of product specifications in the clusters that make the dependency maximum replacing items in the chosen transactions with the new items represented with mean value and variation range.
4. Supposing \( X \) represents the items of product features and \( Y \) represents the items of the requirement alternatives in the same transaction record, by implementing the Apriori algorithm with minimum support (s) and confidence (\( \phi \)), the association rule \( X \Rightarrow Y \) depicts the relation between the product specifications and the requirements alternative. Figure 7 shows the overall architecture of CARMS.

5. Case Study

The proposed architecture has been tested for a retail store, which has its branches at more than 10 locations. The store has a large variety of different products to be offered for the customer. The practical problem is that the retailer is confused with the store layout and the items to be put on sale at a different location when he opens a new store at an entirely new location. Since customers at different age groups, gender and with different income levels have distinct needs in procuring things from the store, a new strategy has to be developed for the new store with the study of the existing demographic study at the new location. In the existing methodology, some approaches are used to discover the potential relationships between the customer groups and the products they purchase, the extension of predicting sales at an entirely new location is not predicted. Therefore, the methodologies seem to be impractical from the new store perspective. In order to predict sales for a new location, the customers with similar preferences are grouped into the same category and the properties associated with an individual customer are shifted towards the total group.

The dependency between the group behaviour and their purchase of items plays a key factor in
generating association rules. According to the marketing survey, irrespective of the customer category, the main concern of people towards the product can be summarized as in Table 1. The features expected on the product requirements and their options are shown in Table 2. Among the features, some are interval-based variables, which are specified in ranges. Based on the existing records, target data are identified and organized into a transaction database as shown in Table 3. The target transaction database is obtained from the existing database. For illustrative simplicity, only 15 records were considered from the transaction database. As shown in Table 3, each customer order indicates the customers choice of certain feature options. The AHP is implemented to prioritize the nine features and we get the relative vector

\[ \vec{W} = (0.115+0.351+0.067+0.034+0.021+0.075+0.146+0.083+0.106) \]  \hspace{1cm} (4)

Due to the different metrics used for the functional requirement variables, all the FR instances in Table 3 is standardized based on the max-min standardization. The results of the distance measures for the binary and numerical requirement instances are shown in Figures 10 and 11 respectively. Based on the max-min standardization and relative weights, the dissimilarity matrix is obtained as shown in Figure 8. Subsequently, the weighted Euclidean distance is used to get the dissimilarity matrix and the fuzzy clustering module is adopted to obtain the fuzzy equivalent matrix as shown in Figure 9. By setting different similarity thresholds for the fuzzy equivalent matrices, Boolean equivalent matrices can be obtained.

Table 1: List of Functional Needs

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>a1</td>
<td>Safe keeping</td>
<td>a_{11} - High a_{12} - Low</td>
</tr>
<tr>
<td>a2</td>
<td>Resilience</td>
<td>a_{21} - High a_{22} - Medium</td>
</tr>
<tr>
<td>a3</td>
<td>User Satisfaction</td>
<td>a_{31} - High a_{32} - Medium a_{33} - Low</td>
</tr>
<tr>
<td>a4</td>
<td>Strength</td>
<td>a_{41} - High a_{42} - Low</td>
</tr>
<tr>
<td>a5</td>
<td>Cost</td>
<td>a_{51} - High a_{52} - Medium a_{53} - Low</td>
</tr>
<tr>
<td>a6</td>
<td>Ease of Use</td>
<td>a_{61} - High a_{62} - Medium a_{63} - Low</td>
</tr>
</tbody>
</table>

Table 2: List of Functional Requirements

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>v1</td>
<td>Frequency of Purchase</td>
<td>v_{11} – Once in 2 weeks v_{12} – Once in 4 weeks v_{13} – Once in 8 weeks</td>
</tr>
<tr>
<td>v2</td>
<td>Cost</td>
<td>v_{21} – Expensive v_{22} – Medium v_{23} – Less expensive</td>
</tr>
<tr>
<td>v3</td>
<td>Performance</td>
<td>v_{31} – Maximum v_{32} – Medium v_{33} - Low</td>
</tr>
<tr>
<td>v4</td>
<td>Size</td>
<td>v_{41} – Large v_{42} – medium v_{43} - Low</td>
</tr>
<tr>
<td>v5</td>
<td>Age</td>
<td>v_{51} – 0-20 v_{52} – 21-40 v_{53} – 41-60</td>
</tr>
<tr>
<td>v6</td>
<td>Gender</td>
<td>v_{61} – Male v_{62} - Female</td>
</tr>
<tr>
<td>v7</td>
<td>Income</td>
<td>v_{71} - &lt; 15000 v_{72} – 16000 – 30000</td>
</tr>
</tbody>
</table>
### Table 3 Transaction database

<table>
<thead>
<tr>
<th>Transaction Records (TID)</th>
<th>Customer Needs</th>
<th>Functional Requirements</th>
</tr>
</thead>
<tbody>
<tr>
<td>T001 a11,a21,a3, 1,a44, a51,a62</td>
<td>v11,v21,v31,v42, v52,v62,v72, v81,v92</td>
<td></td>
</tr>
<tr>
<td>T002 a11,a21,a4, 3,a51</td>
<td>v11,v21,v31,v41, v51,v62,v72,v81</td>
<td></td>
</tr>
<tr>
<td>T003 a12,a22,a3, 3,a62</td>
<td>v12,v21,v33,v43, v52,v61,v72,v83,v91</td>
<td></td>
</tr>
<tr>
<td>... ... ...</td>
<td>... ... ...</td>
<td>... ... ...</td>
</tr>
<tr>
<td>T014 a13,a21,a3, 2, a42,a63</td>
<td>v12,v22,v43,v52,v62, v72,v81,v91</td>
<td></td>
</tr>
<tr>
<td>T015 a12,a21,a3, 2,a41, a61</td>
<td>v11,v22,v31,v42,v53, v71,v82</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8: Dissimilarity matrix based on distance measures

Figure 9: Fuzzy Equivalent Matrix

Figure 10 – Result of distance measures for binary equivalent
The characteristics of each cluster involves the specification of a set of base values together with the related variation ranges, and therefore can be used to suggest standard settings for a new location. The items are then added to the transaction database. The link of each customer preferences is then linked to the corresponding cluster that the customer belongs to. All data that are recorded in the transaction database is fed as input for the Apriori algorithm, which generates rules based on the support and confidence measures.

The output is guaranteed such that only those rules with the highest values for the specified measures are found according to user-defined threshold settings. Figure 13 shows the screenshot of the implemented association rule-mining algorithm. At the end of mining, the system generates 35 rules, based on the specified support and confidence.

The chosen association rules

1. If Age = 35 and Gender = M ⇒ beer and cigarettes is common (Support = 0.4, Confidence = 0.6)
2. If Age = 15 and Gender = M ⇒ Coke and Chips is common (Support = 0.4, Confidence = 0.6)
3. If Age = 35 and Gender = F ⇒ milk and sugar is common (Support = 0.4, Confidence = 0.6)

4. If Age = 15 and Gender = F ⇒ Ice-cream and chocolates is common (Support = 0.4, Confidence = 0.6)

......

In terms of the results of the customer clustering, we can see that the transactions with similar preferred customer needs are clustered into the same class. If for a new potential customer, things are to be planned, then, we can place the new customer into the corresponding customer group based on the mean value of the clusters obtained.

6. Conclusion
In this paper, an efficient architecture is proposed to discover customer group-based rules if a retailer want to open his outlet at an entirely new location. In order to obtain the rules, both the customer and the product domains are bridged based on fuzzy clustering. Association rule mining and Fuzzy clustering were incorporated to analyze the similarity between customer groups and their preferences for products. The complete set of rules generated can be stored in a separate knowledge base. Then, for the stated or required customer needs, we can categorize the corresponding customer groups and can find the clusters to which the customer belongs. Finally, with the different options that the customer would prefer upon, we can predict the layouts and the items for the new store.

References


