Retail Market analysis in targeting sales based on Consumer Behaviour using Fuzzy Clustering - A Rule Based Model

D. Bhanu#, Dr. S. Pavai Madeshwari*

#Department of Computer Applications, Sri Krishna College of Engineering and Technology, Coimbatore
*Department of Computer Science and Engineering, RMK Engineering College, Chennai

Abstract — Product Bundling and offering products to customers is of critical importance in retail marketing. In general, product bundling and offering products to customers involves two main issues, namely identification of product taste according to demographic and product evaluation and selection to increase sales. The former helps to identify, analyze and understand customer needs according to the demographical characteristics and correspondingly transform them into a set of specifications and offerings for people. The latter, concerns with how to determine the best product strategy and offerings for the customer in helping the retail market to improve their sales. Existing research has focused only on identifying patterns for a particular dataset and for a particular setting. This work aims to develop an explicit decision support for the retailers to improve their product segmentation for different settings based on the people characteristics and thereby promoting sales by efficient knowledge discovery from the existing sales and product records. 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 a new location. The methodology is based on the integration of popular data mining approaches such as clustering and association rule mining. It focusses on the discovery of rules that vary according to the economic and demographic characteristics and concentrates on marketing of products based on the population.

Keywords — Decision Support, Product Segmentation, Knowledge Discovery, Data Mining, Association Rule Mining, Clustering

I. INTRODUCTION

Product Bundling and offering products to customers is of critical importance in retail marketing. In this work, 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 focusses on the discovery of rules that vary according to the economic and demographic characteristics and concentrates on marketing the products based on the population.

Today's manufacturers compete on price, adaptability and the variety of products which are all driven by customer satisfaction [1]. From a customer point of view, most customers purchase a product according to their individual preferences. Association rule mining is a significant data mining technique in generating rules based on the correlation between the sets of items that the customers purchase. To some extent, the rule generated gives us the dependencies among the set of customers and the products that they purchase. However, if the retail store plans to open a chain of stores at different location, the retailer has no idea about the customer attraction towards products at different locations. From the retailer's viewpoint, parameters / variables used to understand customer requirements are always fuzzy, uncertain and abstract. Researchers have analyzed the different aspects of the technique, but the shift in focus of increasing sales and customer targeting at different locations with the data available from a single point of sale is not much dealt with. It is without doubt that predicting sales at different locations with the data currently at hand need a solid methodology, which involves a large amount of data and knowledge inferred from the data. Therefore, the key element is to use the existing knowledge and patterns to generate new patterns or rules for different locations.

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. Fortunately, the existing relationship or patterns turn out to be more specific, which depicts the relationship between the customers and the products they go for.

The major challenge of our work is the identification of rules for a different demographic population based on the existing records. Past research has not much focused upon the knowledge discovery of customer needs for a new point with the current available data. This paper proposes a novel approach in integrating the rule mining technique with the clustering approach to discover rules driven by customer needs for a typical new situation. As a result, the approach has the opportunity to take advantage of the wealth of the customer needs information and the alternative rules for the new situation, which better aids the retailer in designing his store layout.

II. RELATED WORK

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 [2][3][4]. The method also called Apriori is the core of all known algorithms except the...


III. PROBLEM FORMULATION

For a potential customer arriving at the store, which customer group one should belong to according to customer needs, what are the preferred functional features or products that the customer focusses on and what kind of offers will satisfy the customer etc., finds to be the key factor in targeting customers to improve sales. Generally, a transactional database is created to record all the products purchased by the customer. To focus on the market segment that each customer falls into, the transaction database can be grouped into different clusters based on the customer needs.

In general, customer needs can be described as a set of features or attributes, $A = \{a_1, a_2, a_3, ..., a_M\}$. Each attribute, $a_i \forall i \in \{1, 2, 3, ..., M\}$, may take one out of a finite set of options, $A_i = \{a_{i1}, a_{i2}, ..., a_{im}\}$. Suppose all customers comprise a set, $C = (c_1, c_2, ..., c_S)$, where ‘S’ denotes the total number of customers. The requirement information of a particular customer, $C_s \in C | \exists s \in \{1, S\}$, for the existing dataset can be depicted by a vector of certain options of the features. Hence, all customers can be grouped into a set, $A = \{\tilde{a}_1, \tilde{a}_2, ..., \tilde{a}_S\}$ which characterizes the customer domain. The functionality of each product is characterized by the set of functional requirements, $V = \{v_1, v_2, ..., v_n\}$. Each functional requirement, $v_r = v_{r1} v_{r2} ... v_{rn} \forall r \in \{1, 2, ..., n\}$, denotes the rth possible value of $v_r$. Suppose, let us assume that all existing products comprise a set $P = \{p_1, p_2, ..., p_T\}$, where T refers to the total number of products. The requirement specification for a particular product, $p_t \in P | \exists t \in \{1, 2, ..., T\}$ can be represented as a vector of certain requirement values.

Similarly, all the instances of requirements in the product domain constitute a set, $V = \{\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_T\}$. The problem domain and the specification of the problem between the customer needs and the functional requirements are shown in Figure 1.

IV. ARCHITECTURE OF THE PROPOSED SYSTEM

In order to obtain the association rules for a new store based on the analysis of customer transactions from the existing knowledge base, a 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 data pre-processing and data partitioning, data transformation, association rule mining and presentation modules. Before proceeding to the rule mining of datasets, raw data must be pre-processed in order to be useful for knowledge discovery. Due to the uncertainty of customer requirements and their behaviour, it is necessary to pre-process the knowledge base. Based on the raw data stored in the knowledge base, target datasets should be identified, involving data cleaning and filtering tasks as integration of multiple databases, removal of noise, handling of missing data files etc.

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.

Data Pre-processing

Many data types are available in fuzzy clustering, such as binary variables, ordinal variables, ratio-scaled variables and even a mix of these variables. Because of the influences in the demography at different locations, 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 \( F_i = (f_{i1}, f_{i2}, \ldots, f_{iM}) \) be the transaction record, where \( M \) denotes the product features. A standard deviation equation is applied on the raw data to standardize the data.

\[
\text{Suppose } f'_{ik} = f_{ik} - \bar{f}_k / S_k \quad (\forall \ i \in [1, 2, \ldots, T], \, \forall \ k = 1, 2, \ldots, M),
\]

where \( f_{ik} \) corresponds to the option value of the \( i \)th transaction record for the specification \( f_k \), where \( \bar{f}_k = 1/T \sum_{i=1}^{T} f_{ik} \).

It denotes the average value of all the transactions as to the \( k \)th specification, where \( S_k = \sqrt{1/T \sum_{i=1}^{T} (f_{ik} - \bar{f}_k)^2} \). Unlike the max-min standardization, it cannot transform the raw data to values between 0 and 1, but eliminate the influence of different dimensions. Figure 4 shows the method for pre-processing for customer need identification and Figure 5 shows the pre-processing for product requirement need identification.

Prioritization of variables by assigning weights

The specification of Functional requirements involves multiple variables, i.e., \( V = \{v_q \mid \forall q = 1 \ldots N\} \). These variables constitute to the overall functionality of a product purchase, which is unique in its own form. Hence, Functional requirements variables should be prioritized to differentiate their effects. The relative importance of requirements variables is usually quantified by assigning different weights. That is, each \( v_q \) is associated with a
weight, \( w_q \), subjective to \( \sum_{q=1}^{N} w_q = 1 \). For the proposed architecture, AHP \([20]\) is adopted for prioritization of functional requirements variables, owing to its advantages in maintaining consistence among a large number of variables through pair-wise comparisons.

**Computation of Distance to find the similarity measures**

The Functional Requirement clustering includes two steps as distance measure and fuzzy clustering. As the first stage, the distance measure module measures the dissimilarity between FR instances in order to define the fuzzy compatible relations among data objects.

**Distance measure**

In general, each FR instance \( v'_i = [v_{i1}, v_{i2}, \ldots, v_{in}] \in V' \), where \( \forall v_{iq} = v_{q}, \exists v_{iq} \in V_q, \forall r = 1, \ldots, n_q \), may involve three types of variables: numerical, binary and nominal variables. For example, \( v_{1q} \) may be a numerical value while \( v_{2q} \) may be a binary or nominal value. The distance between any two instances means the dissimilarity between them and thus is measured as a composite distance of the distance components corresponding to these three types of variables.

**Numerical Measures** – a number of methods of distance measure have been proposed for purpose of numerical clustering, including the Euclidean distance, Manhattan distance, Minkowski distance and weighted Euclidean distance measure \([13]\). The proposed architecture employs the weighted Euclidean distance. It is computed as following:

\[
d_{\text{numerical}} (v'_i, v'_j) = \sqrt{\sum_{q=1}^{Q} (w_q (N_{-v'_iq} - N_{-v'_jq}))^2},
\]

where \( d_{\text{numerical}} (v'_i, v'_j) \) indicates the numerical distance between two FR instances, \( v'_i \) and \( v'_j \), where \( \forall v'_i, v'_j \in V', w_q \) means the relative importance of the \( q \)-th numerical requirement variable, \( V_q \subseteq V, Q \) represents the total number of numerical FR variables among the total size – NFR variables \( (Q \leq N) \) and \( N_{-v'_iq} \) and \( N_{-v'_jq} \) denote the normalized values \( V_q \) and \( V_q \) respectively.

**Binary FRs** – A binary variable assumes only two states: 0 or 1, where 0 means the variable is absent and 1 means it is present. The architecture uses a well-accepted coefficient for assessing the distance between symmetric binary variables, called the simple matching coefficient\([10]\). It is calculated as the following,

\[
d_{\text{binary}} (v'_i, v'_j) = \alpha_2 + \alpha_3 + \alpha_4 + \alpha_5 + \alpha_6 + \alpha_7 \times \alpha_8,
\]

where \( d_{\text{binary}} (v'_i, v'_j) \) indicates the binary distance between two FR instances, \( v'_i \) and \( v'_j \), \( \forall v'_i, v'_j \in V', \alpha_1 \) is the total binary variables in \( V \) that equal to 1 for both \( v'_i \) and \( v'_j \), \( \forall v'_i, v'_j \in V', \alpha_2 \) is the total binary variables in \( V \) that equal to 1 for \( v'_i \) but 0 for \( v'_j \), \( \forall v'_i, v'_j \in V', \alpha_3 \) is the total binary variables in \( V \) that equal to 0 for \( v'_i \) but 1 for \( v'_j \), and \( \alpha_4 \) is the total number of binary FR variables that equal to 0 for both \( v'_i \) and \( v'_j \).

**Nominal FRs** – A nominal variable can be regarded as a generalization of binary variable in that it takes on more than two states. This type of variables cannot be expressed by numerical values but by qualitative expression with more than one option. Therefore the simple matching coefficient can be used here to measure the nominal distance between two FR instances containing nominal FR variables\([10]\).

\[
d_{\text{nominal}} (v'_i, v'_j) = \beta - \gamma / \beta, \text{where } d_{\text{nominal}} (v'_i, v'_j) \text{ indicates the nominal distance between two FR instances, } v'_i \text{ and } v'_j, \forall v'_i, v'_j \in V, \gamma \text{ means the total number of nominal FR variables in } V \text{ (i.e., } v_q \in V_{\text{nominal}} \subseteq V \text{) that assume the same states for } v'_i \text{ and } v'_j \text{ and } \beta \text{ is the total number of nominal variables among total size – NFR variables } (\beta \leq N). \text{ Given a set of functional requirement variables, } V = [v_{1}, v_{2}, \ldots, v_{N}], \text{ every functional requirement instance assumes a certain value for each variable, and thus consists of a combination of numerical, binary and /or nominal requirement values, that } V_{\text{numerical}} \cup V_{\text{binary}} \cup V_{\text{nominal}} = V. \text{ As a result, the overall distance between } v'_i \text{ and } v'_j \text{ comprises three components: the numerical, binary and the nominal distances. A composite distance can thus be obtained by the weighted sum: } d( v'_i, v'_j ) = w_{\text{numerical}} d_{\text{numerical}} (v'_i, v'_j) + w_{\text{binary}} d_{\text{binary}} (v'_i, v'_j) + w_{\text{nominal}} d_{\text{nominal}} (v'_i, v'_j), \sum (w_{\text{numerical}} + w_{\text{binary}} + w_{\text{nominal}}) = 1, \text{ where } w_{\text{numerical}}, w_{\text{binary}} \text{ and } w_{\text{nominal}} \text{ refer to the relative importance of numerical, binary and nominal distances respectively. These weights can be determined in the similar way as that of requirement variables applying the AHP.}

**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 \( R^{0} R \leq R \). Similarly, the transaction record \( F_p, G \) is transitive iff \( G^{0} G \leq G \). The process is performed by the max-min transitive closure for \( R \) and \( G \) respectively. This method is based on the following model: \( t(R) = \bigcup_{k=1}^{T} R^k \), 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.
The following step is to transform the fuzzy equivalent matrix to the Boolean equivalent one. Assuming \( t( R ) = \{ \rho'( C_i, C_j ) | \forall ( C_i, C_j ) \in T \times T \} \), the \( \alpha \)-cut of \( t( R ) \) is represented as \( R_\alpha = \{ \nu( C_i, C_j ) | T \times T \} \), where \( \nu( C_i, C_j ) = 1 \), if \( \rho'( C_i, C_j ) \geq \alpha \); otherwise \( \nu( c_i, c_j ) = 0 \). It is obvious that Boolean matrix \( R_\alpha \) is an equivalent matrix too. If all the elements of two rows in \( R_\alpha \) are identical, the transaction records corresponding to the rows are classified into the same cluster. The clusters are described as \( TU / R_\alpha = \{ g_1, g_2, ..., g_s \} \), where \( TU \) is the universe of all the transaction records and \( s \) is the number of clusters. Similarly, we can derive the clusters with respect to product specifications by setting different threshold values for \( \beta \). The clusters are depicted as \( TU / G_\beta = \{ p_1, p_2, ..., p_l \} \). Because the equivalent matrix \( R_\alpha \) and \( G_\beta \) are defined by a corresponding binary equivalent relation, they can be also equally viewed as the equivalent relation. Figure 6 shows the Fuzzy clustering method for grouping customers by converting the customer needs and functional requirements onto a fuzzy equivalent matrix.

**Fig. 6 – Fuzzy Clustering**

**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 \( I = \{ I_1, I_2, ..., I_s \} \), and the database of transaction records \( D = \{ t_1, t_2, ..., t_n \} \), where \( t_i = \{ I_{i1}, I_{i2}, ..., I_{ik} \} \) and \( I_{ij} \in I \), an association rule is an implication of the form \( X \Rightarrow Y \) where \( X, Y \subseteq I \) 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:

Choosing the suitable product clusters, which make the dependency maximum, calculating the mean values and the corresponding variation as to specification for each transaction in the cluster.

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. 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 show the overall architecture of CARMS.

**Fig. 7 – CARMS overall architecture**

**V. 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 customer’s choice of certain feature options. The AHP is implemented to prioritize the nine features and we get the relative vector

\[ \hat{W} = \{0.115+0.351+.067+.034+.021+.075+.146+.083+.106\} \]

Due to the different metrics used for the functional requirement variables, all the FR instances in Table 3 are 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 Customer Needs

| Feature | Description | Options |
|---------|-------------|---------|
| a1      | Safe keeping| a11 - High, a12 - Low |
| a2      | Resilience  | a21 - High, a22 - Medium, a23 - Low |
| a3      | User Satisfaction | a31 - High, a32 - Medium, a33 - Low |
| a4      | Strength    | a41 - High, a42 - Low |
| a5      | Cost        | a51 - High, a52 - Medium, a53 - Low |

### Table 2: List of Functional Requirements

| Feature | Description | Options           |
|---------|-------------|-------------------|
| v1      | Frequency of Purchase | v11 - Once in 2 weeks, v12 - Once in 4 weeks, v13 - Once in 8 weeks |
| v2      | Cost        | v21 - Expensive, v22 - Medium, v23 - Less expensive |
| v3      | Performance | v31 - Maximum, v32 - Medium, v33 - Low |
| v4      | Size        | v41 - Large, v42 - medium, v43 - Low |
| v5      | Age         | v51 - 0-20, v52 - 21-40, v53 - 41-60 |
| v6      | Gender      | v61 - Male, v62 - Female |
| v7      | Income      | v71 - < 15000, v72 - 16000 – 30000, v73 - 31000-50000 |
| v8      | Usage       | v81 - High, v82 - Medium, v83 - Low |
| v9      | Manifestation| v91 - yes, v92 - No |

### Table 3: Transaction Database

| Transaction Records (TID) | Customer Needs | Functional Requirements |
|---------------------------|----------------|-------------------------|
| T001                      | a11,a21,a31, a44,a51,a62 | v11,v21,v31,v42, v52,v62,v72,v81,v92 |
| T002                      | a11,a21,a43, a51       | v11,v21,v31,v41,v51, v62,v72,v81 |
| T003                      | a12,a22,             | v12,v21,v33,v43,v52,  |
Figure 8: Dissimilarity matrix based on distance measures

Figure 9: Fuzzy Equivalent Matrix

Figure 10: Result of distance measures for binary requirement instances

Figure 11: Result of distance measures for numerical requirement instances

Figure 12 shows the result of mapping between the customer groups and product clusters in order to get the customer groups that have the maximum dependency to clusters of product features. 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.

![Customer Needs and Clusters](image1)

**Fig. 12: Mapping between customer needs and clusters**

![Results after Mining](image2)

**Fig. 13 – The results after Mining**

VI. CONCLUSION

In this work, 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

1. Agrawal R., Imilienksi T. and Swamy, A. 1993. Mining association rules between sets of items in large databases. Proceedings of the ACM SIGMOSD International Conference of Management of Data, Washington DC., pp:207-216.

2. Agrawal R. and Srikant R., 1994. Fast algorithm for mining association rules in large databases. Proceedings of 1994 International Conference on VLDB, pp: 487-499.

3. Barbara D. and Chen P. 2003. Using Self-Similarity to Cluster Large Datasets. Data Mining and Knowledge Discovery, 7: 123-152.

4. Crespoa F and Weberb R. 2005. A methodology for dynamic data mining based on fuzzy clustering. Fuzzy Sets Systems. 150: 267 – 284.

5. De-Chang P.Q., Xiao-lin G., Wang-Feng and Ran C. 2005. STBAR. A more efficient algorithm for association rule mining. The 4th International Conference on Machine Learning and Cybernetics, 18-21 August, Guangzhou, China, pp: 1529 – 1533.

6. Grabmeier J. and Rudolph R. 2002. Techniques of cluster algorithms in data mining. Data Mining and Knowledge Discovery, 6: 303 – 360.

7. Groenen P. and Jajuga J.F.K. 2001. Fuzzy clustering with squared Minkowski distances. Fuzzy Sets Syst., 120: 227 – 237.

8. Hirano S. and Tsumoto S. 2003. Dealing with Relative Similarity in Clustering: An Indiscernibility Based Approach. PAKDD 2003, Springer – Verlag Berlin Heidelberg, LNAI., 2637 : 513 -518.

9. Oh, S.H. and Lee W.S.2003. Optimized Clustering for Anomaly Intrusion Detection. PAKDD 2003, Springer – Verlag Berlin Heidelberg, LNAI, 2637 – 576-581.

10. Padmanabhan B. and Tuzhilin, 1999. Unexpected as a measure of interestingness in knowledge discovery. Decision Support Syst., 27: 303 – 318.

11. Ravi T.V. and Gowda C.K. 1999. An ISODATA clustering procedure for symbolic objects using a distributed genetic algorithm. Pattern Recognition Lett., 20 : 659-666.

12. Timm H. Doring C and Kruse R. 2004. Different approaches to fuzzy clustering of incomplete datasets. Int. Journal of Approx. reason., 35: 239-249.

13. Xiaoyong D., Suzuki S. and Ishii N. 1999. A Distance Based Clustering and Selection of association Rules on Numeric Attributes. RSFDGrC’99, Springer Verlag, Berlin Heidelberg, LNAL, 1711: 423-433.

14. Yih-Jen H.C., Shyi-Ming C., Yu-Chuan and Chia-Hoang L. 2005. A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques. IEEE Trans. Fuzzy Syst., 13: 216-227.