Performance Analysis of A Rule Based Mining Framework In Predicting Retail Sales

Roshni.M[1], Monika.R[2], Sowmiya.M [3], S. Sathya Bama [4]
Department of Computer science and Engineering
[1] Research Scholar, Sri Krishna college of Technology, Coimbatore, India- 641042.
[2] Assistant Professor, Sri Krishna college of Technology, Coimbatore, India- 641042.

Abstract- *Data Mining is the process of extracting interesting and previously unknown patterns and correlations from data stored in huge data repositories. association Rule Mining, a descriptive mining technique of Data Mining is the process of discovering items, which tend to occur together in transactions. Association rules are interesting correlations among attributes in a database. Fuzzy clustering based Association Rule Mining system is proposed in targeting customers to improve sales which improvises the G-MAR model by predicting sales based on customer needs and functional features. For a potential customer arriving the store, which customer group one should belong to according to customer needs, what are the preferred functional features or products that the customer focuses 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.*

Keywords- sales forecasting, neural networks, exponential smoothing, combined forecasts, retail sales.

I. INTRODUCTION

In recent years, Data-Mining (DM) has become one of the most valuable tools for extracting and manipulating data and for establishing patterns in order to produce useful information for decision-making. Nearly all areas of life activities demonstrate a similar pattern. Whether the activity is finance, banking, marketing, retail sales, production, population study, employment, human migration, health sector, monitoring of human or machines, science or education, all have ways to record known information but are handicapped by not having the right tools to use this known information to tackle the uncertainties of the future. Because it is an emerging discipline, many challenges remain in data mining.

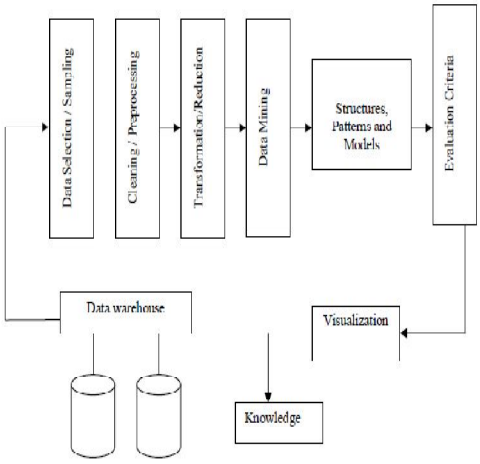

Figure 1 the KDD Process

Due to the enormous volume of data acquired on an everyday basis, it becomes imperative to find a methodology that determines which technique to select and what type of mining to do. Data sets are often inaccurate, incomplete, and/or have redundant or insufficient information.

The first stage is data preprocessing, which entails data collection, data smoothing, data cleaning, data transformation and data reduction. The second step, normally called Data Mining (DM), involves data modeling and prediction. DM can involve either data classification or prediction. The classification methods include deviation detection, database segmentation, clustering (and so on); the predictive methods include (a) mathematical operation solutions such as linear scoring, nonlinear scoring (neural nets), and advanced statistical methods like the multiple adaptive regression by splines (b) distance solutions, which involve the nearest-neighbor approach (c) logic solutions, which involve decision trees and decision rules. The third step is data post-processing, which is the interpretation, conclusion, or inferences drawn from the analysis in Step Two. The steps are shown diagrammatically in Figure 1.1.

The remainder of this dissertation is organized in the following fashion: In Chapter 2, a review of the published work on Association Rule Mining is presented. Chapter 3, describes the Group based Mining of Association Rule (G-MAR) model in detail. Chapter 4 The results were compared and the performance of the model and the technique were discussed. In Chapter 5, summarizes the main contributions of the research.

II. RELATED WORKS

Association rule mining is one of the most significant techniques applied in the field of data mining. It was first introduced in (Agrawal et al 1993), which aims in extracting interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases and other data repositories. The mined rules are used in various domains such as telecommunication, networks, retail store settings and layout, market analysis, business forecasting and risk management, etc. Hence, the different rule mining techniques and algorithms are briefly introduced and compared later. There are several association rule-mining algorithms proposed in the research literature. This chapter provides a review of a representative set of the major algorithms proposed.

Some of the algorithms discusses the generation of association rules using Apriori algorithm and discusses its variants with other constraints (Hegland 2003, Zheng et al 2001). In this chapter, we give an outline on the various currently used rule-mining techniques. We also discuss the notations and the basic terminologies used in the field of association rule mining. The chapter also deals with the efficiency of the various mining algorithms and the different types of databases upon which it can be applied.

Apriori is more efficient during the candidate generation process (Agrawal et al 1994). Apriori uses pruning techniques to eliminate the less frequent itemsets. However, there are two bottlenecks of the Apriori algorithm viz., the generation of candidate itemsets and multiple scans of the database. Hence, new algorithms were developed based on Apriori, which formed the basis for all algorithms.

To overcome these drawbacks, another algorithm called TreeProjection was proposed in (Agarwal et al 2000). The concept behind the algorithm is that it constructs a lexicographical tree and maps a large database into a set of reduced, item-based sub-databases based on the frequent patterns mined so far. The number of nodes in its lexicographic tree is exactly that of the frequent itemsets. The main advantage of this algorithm is that thetransaction projection limits the support counting in a small sample space and the lexicographical tree helps in managing and counting the candidates.

Wang and Tjortjis (2004) presented PRICES, an efficient algorithm for mining association rules. The algorithm reduces large itemset generation time, known to be the most time-consuming step, by scanning the database only once and using logical operations in the process. Another algorithm called Matrix (Yuan et al 2005) was proposed which enters 1 or 0 for the presence and absence of items and generates the large itemsets.

The next algorithm for the mining of association rules was proposed in (Toivonen 1996), using the sampling approach. The approach can be divided into two phases. During the first phase, a sample of the database is scanned for generating rules. These results are then validated against the entire database. To maximize the effectiveness of the overall approach, the author makes use of lowered minimum support on the sample. A probabilisticapproach is implemented during the first pass and hence not all rules may be found initially. In the second phase of the algorithm, the other part of the dataset that were not deemed to be frequent during the first pass were mined to generate the complete set of rules. Parthasarathy (2001) presented an efficient method to progressively sample for association rules.

Cheung et al (1996) presented an algorithm called FDM. FDM is a parallelization of Apriori to shared nothing machines, each with its own partition of the database. At every level and on each machine, the database scan is performed independently on the local partition. Then a distributed pruning technique is employed. Schuster and Wolff (2001) described another Apriori based D-ARM algorithm - DDM. As in FDM, candidates in DDM are generated level wise and are then counted by each node in its local database. The nodes then perform a distributed decision protocol in order to find out which of the candidates are frequent and which are not.

Wojciechowski and Zakrzewicz (2002) focus on improving the efficiency of constraint-based frequent pattern mining by using dataset filtering techniques. Dataset filtering conceptually transforms a given data mining task into an equivalent one operating on a smaller dataset. Tien Dung Do et al (2003) proposed a specific type of constraints called category-based as well as the associated algorithm for constrained rule mining based on Apriori. The Category-based Apriori algorithm reduces the computational complexity of the mining process by bypassing most of the subsets of the final itemsets. An experiment has been conducted to show the efficiency of the proposed technique.

A serious problem in association rule discovery is that the set of association rules can grow to be unwieldy as the number of transactions increases, especially if the support and confidence thresholds are small. As the number of frequent itemsets increases, the number of rules presented to the user typically increases proportionately (Zaki et al 1997a). Many of these rules may be redundant. A similar approach is described in (Yuan et al 2002). Wu et al (2004) derived a new algorithm for generating both positive and negative association rules. They add on top of the support-confidence framework another measure called minimum interest for a better pruning of the frequent itemsets generated. In (Teng et al 2002) the authors use only negative associations of the type X -Y to substitute items in market basket analysis.

The larger the set of frequent itemsets the more the number of rules presented to the user, many of which are redundant. This is true even for sparse datasets, but for dense datasets it is simply not feasible to mine all possible frequent itemsets, let alone to generate rules, since they typically produce an exponential number of frequent itemsets, finding long itemsets of length 20 or 30 is not uncommon. Although several different strategies have been proposed to tackle efficiency issues, they are not always successful.

III. SYSTEM DESIGN

The conventional hierarchical clustering algorithms such as singlelink and complete-link suffer higher time complexity. As a result, a recent trend is to develop hybrid-clustering algorithms that exploit the advantages of both hierarchical and partitioned algorithms. Hence, a clustering algorithm Birch (Tian Zhang et al 1996) has been utilized for finding factions which are treated as a group of clusters. The idea is to use a standard clustering algorithm to identify the intervals of interest followed by the construction of Coalescent Dataset in order to check the applicability of the rules outside of the dataset. The clustering algorithm uses a single partitioning of the attributes into disjoint sets (Xi) over which there is a meaningful metric. Most often, each Xi an individual attribute or a small set of closely related attributes oversimilar domains. The clusters are created incrementally and represented by a compact summary. The summaries produced in the first phase are then used for the construction of the Coalescent Dataset approach.

This section, describes the faction generation procedure using the clustering algorithm. The clustering phase is divided into two phases, where the first phase consists of the identification of clusters and the next phase on combining clusters to generate factions. Later, the Coalescent Dataset was constructed using the factions generated.

The experiments cover a range of databases and mining workloads and the only difference with the existing algorithms is that the database sizes were considered significantly larger than the available main memory.

The performance metric in all the experiments is the total execution time taken by the mining operation. The databases used in the experiments were synthetically generated using the technique described in (Agrawal et al 1994) and attempt to mimic the customer purchase behaviour seen in retailing environments.

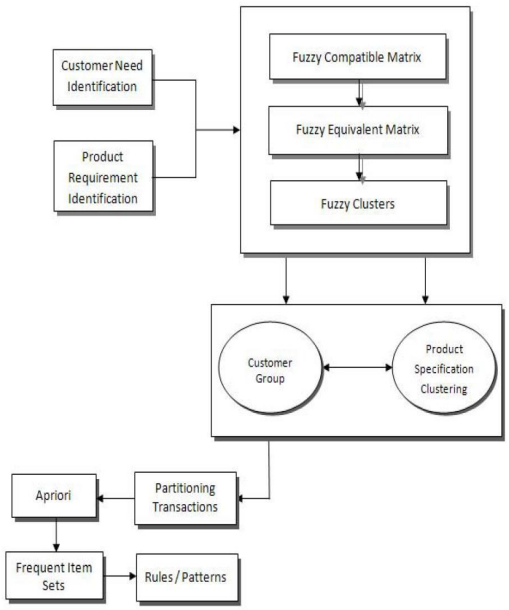

Figure 2 the Model Architecture

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.

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; calculating the mean values and the corresponding variation as to specification for each transaction in the cluster.
2. 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.

3. 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, the association rule X Y depicts the relation between the product specifications and the requirements alternative.

The specification of functional requirements involves multiple variables . 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 vq is associated with a weight, wq, subjective to for the proposed architecture, AHP (Saaty 1980), 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.

IV. RESULT AND DISCUSSIONS

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. Due to the different metrics used for the functional requirement variables, all the FR instances in Table 3 are standardized based on the maxmin standardization. The results of the distance measures for the binary and numerical requirement instances are shown in Figures 3 and 4 respectively. Based on the max-min standardization and relative weights, the dissimilarity matrix is obtained as shown in Figure 3. 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 4. By setting different similarity thresholds for the fuzzy equivalent matrices, Boolean equivalent matrices can be obtained. 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.

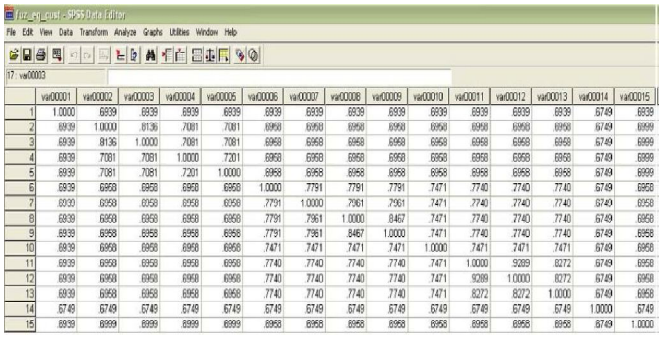

Figure 3 Fuzzy Equivalent Matrixes

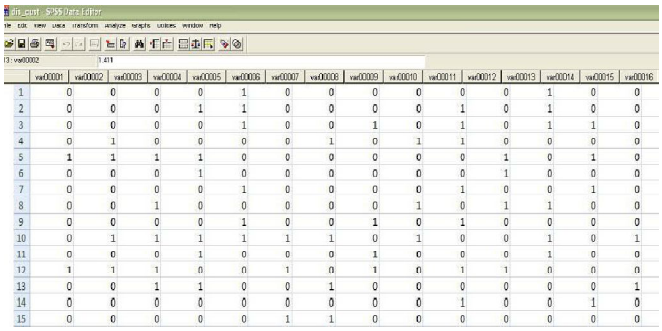

Figure 4 Results of Distance Measures for Binary Requirement Instances

The chosen association rules

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

V. CONCLUSIONS

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] Agard B. and Kusiak A. (2004), 'Data Mining Based Methodology for design of Product Families', International Journal of Production Research, Vol. 42(15), pp. 2955-2969.

[2] Agrawal D. and Aggarwal C.C. (2001), 'On the Design and Quantification of Privacy preserving Data Mining Algorithms', Proceedings of the ACM symposium on Principles of Database Systems, Santa Barbara, California, United States, pp. 247-255.

[3] Agarwal R.C., Aggarwal C.C. and Prasad V.V.V. (2000), 'Depth first generation of Long Patterns', Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, United States, pp. 108-118.

[4] Agarwal R.C., Aggarwal C.C. and Prasad V.V.V. (2001), 'A Tree Projection algorithm for generation of frequent itemsets', Journal of Parallel and Distributed Computing, Vol. 61, Issue 3, pp. 350-371.

[5] Agrawal R., Imielinski T. and Swami A. (1993), 'Mining Association rules between sets of items in Large Databases', Proceedings of ACM SIGMOD International Conference on Management of Data, Washington, United States, pp. 207-216.

[6] Agrawal R. and Srikant R. (1994), 'Fast Algorithms for mining Association rules', Proceedings of International Conference on Very Large Databases (VLDB), San Francisco, United States, pp. 487-499.

[7] Agrawal R. and Srikant R. (1995), 'Mining Sequential Patterns', Proceedings of International Conference on Data Engineering (ICDE), pp. 3-14.

[8] Agrawal R. and Shafer J.C. (1996), 'Parallel mining of Association rules', IEEE Transactions on Knowledge and Data Engineering, Vol. 8, Issue 6, pp. 962-969.

[9] Agrawal R. and Srikant R. (2000), 'Privacy-preserving Data Mining', Proceedings of ACM SIGMOD International Conference on Management of Data, pp. 439-450.

[10] Aggarwal C.C. and Yu P.S. (1998), 'Online generation of Association rules', Proceedings of International Conference on Data Engineering (ICDE), Washington, United States, pp. 402-411.

[11] Anumann Y., Feldman R., Lipshtat O. and Manilla H. (1999), 'Borders: An Efficient Algorithm for Association generation in Dynamic Databases', Journal of Intelligent Information Systems, Vol. 12, Issue 1, pp. 61-73.