ISSN - 2250-1991 | IF : 5.761 | IC Value : 79.96

Original Research Paper Computer Science

Market Basket Data Analysis Using A Novel Genetic Frequent Itemset MiningAlgorithm by Partitional Clustering Approach

Dr. D. Ashok Kumar | Associate Professor and Head, Department of Computer Science and Application, Government Arts College, Trichirappalli - 620 022, Tamilnadu, India.

Ms. T. A. Usha | Assistant Professor, Government Arts College, Trichirappalli - 620 022, Tamilnadu, India.

ABSTRACT

Frequent Itemsets Mining is very complex due to its high dimensionality, data sparsity, versatility and scalability. To improve the profit and quality of service of retail organizations or supermarkets or departmental stores, daily transactions need to be analysed to understand the customer behaviours. The number of itemsets grows exponentially with the number of items available so that memory consumption and processing efficiency is a major issue. Association rules will be derived using these frequent itemsets which can be implemented in the market to improve the sales. With these issues, analysing large datasets via traditional methods has moved from being tedious, to being infeasible. As a consequence, fast and scalable data mining techniques are increasingly more important. Thus, for the efficiency, it is motivated that the large market database need to be pre-processed using clustering and soft computing approaches viz., Genetic Algorithm etc., then frequent itemsets can be generated and association rules are formed to improve the business. It details the Genetic Algorithm and its significance with the Frequent Item Set Mining. It presents a novel approach based on Genetic Algorithm and Empirical Study Results, reveals that it out performs the existing methodology. Also, it presents a novel approach based on Partitional clustering for efficient Frequent Itemsets Mining and their Association Rules generation. The novel frequent itemset mining algorithm which is designed to work on high dimensional and very large databases and its effectiveness is measured using various measures. The proposed algorithm is applied for Market Basket data set. The performance of the proposed algorithm is evaluated and the results are tabulated. The results disclose the efficiency and efficacy of the algorithms with respect to various measures.

KEYWORDS | Frequent Itemset Mining, Genetic Algorithm, Partitional Clustering.

1. Introduction

The development of hardware, software and scientific advancements made the computerization of business easier. Scientific advancements have made easy to collect the data and digitizing the information which can be stored in database. This makes the collection and storing the data to grow at a phenomenal rate [1]. Such a raw data is rarely of direct use. A retailer must know the needs of customers and adapt to them. Market basket analysis has been intensively used in many companies as a means to discover product associations and base a retailer's promotion strategy on them. Market basket analysis gives retailer good information about related sales based on group of goods. Several aspects of market basket analysis have been studied in academic literature, such as using customer interest profile and interests on particular products for one-to-one marketing [19], purchasing patterns in a Supermarket [6] to improve the sales. Decisions can be made easily about product placement, pricing, promotion, profitability and also finds out, if there are any successful products that have no significant related elements. Similar products can be found so those can be placed near each other or it can be cross-sold. Market basket analysis is one possible way to find out which items can be put together in a retail centre so that customers can access them quickly. Such related groups of goods also must be located side-by-side in order to remind customers of related items and to lead them through the centre in a logical manner.

Market basket analysis is one of the data mining methods [3] focusing on discovering purchasing patterns by extracting associations or co-occurrences from a store's transactional data. Market basket analysis determines the products which are bought together, to reorganize the supermarket layout and also to design promotional campaigns such that products' purchase can be improved. Hence, the Market consumer behaviour needs to be analysed, which can be done through different data mining techniques.

Data mining is the system of extracting knowledge from huge amounts of data accumulated either in databases, data warehouses, or other information repositories. Data mining is a synonym for another popularly used term, Knowledge Discovery in Databases or KDD. Data mining operations are applied to mention the type of sequences to be identified in data mining tasks [11]. Broadly, data mining tasks are classified into two categories: descriptive and predictive. The common features of the data in the database are designated by the descriptive mining tasks. Predictions are inferred from the current data by the Predictive mining tasks.

In recent trends, data mining plays a vital role for framing association rules among the massive collection of itemsets. Frequent itemsets are produced from large data sets by utilizing association rule mining, takes ample of time to figure out all the frequent itemsets. By utilizing the Genetic Algorithm (GA) we improved the results of association rule mining. Genetic Algorithms are powerful and widely applicable stochastic search and optimization methods based on the concepts of natural selection and natural evaluation. On the whole, the objective of this study is to observe all the frequent itemsets and also generate the association rules in a short span of time from the large datasets by using the proposed genetic algorithm, GFISM.

In this study, Apriori algorithm and genetic algorithms have been used to identify the techniques for association rule mining. The performance measurement and complexities of algorithms have also been presented. The combination of Soft computing techniques with existing association rule mining yields fast results. In this study ARM algorithm with genetic approach has been analyzed.

A k-itemset that consists of k items from I, is frequent if it occurs in the Transaction(T) not less than s times, where s is a user-specified minimum support threshold and s <= n. In 1975, John Holland developed the Genetic Algorithm at the University of Michigan. Genetic Algorithm is an adaptive heuristic search algorithm based on the development of natural selection and genetics. This

directed search algorithm is based on the mechanics of biological evolution. Later in the year 1992, John Koza used Genetic Algorithm to evolve the programs to perform certain tasks and this termed as Genetic Programming. It is also a part of evolutionary computing. Genetic algorithms are inspired by Darwin's theory on the evolution, termed as "Survival of the Fittest". It generates a result by combining selection, recombination and mutation. It randomly searches the dataset to solve the optimization problems. It means that more desirable solutions evolve from the earlier generations until a close by ideal solution is obtained. It provides efficient, effective techniques for optimization and machine learning applications [7].

Wakabi-Waiswa, P.P., et al., proposed [18] "Generalized Association Rule Mining Using Genetic Algorithms". The Association rule mining is designed for combining the Genetic Algorithms and Apriori algorithm. It yields very fast results. It generalized a very large database of transactions, where each transaction contains a set of items, and a classification on the items, and then the associations between items at any level of the classification have been found. It improves the performance of minimum support number of items and number of transactions.

Ghosh S, Biswas S, Sarkar D and Sarkar P.P, "Mining Frequent Itemsets Using Genetic Algorithm", proposed [9] the algorithm to find frequent itemsets using genetic algorithm. The association rule mining algorithm like apriori, partition, fp-tree, etc., generates the frequent itemsets. However, it takes too much time to compute the frequent itemsets. The main aim to introduce genetic algorithm is to reduce the computing time. Genetic algorithm performs global search to generate the frequent itemsets. The time complexity and memory usage is less when compared to the association rule mining algorithm because the genetic algorithm is constructed by the greedy approximation.

This study compares and analyzes the Apriori with genetic algorithm for finding the frequent itemsets. The proposed Genetic algorithm GFISM uses Stratified Random Sampling technique for selecting the samples in each stratum. This algorithm identifies the frequent itemsets repeatedly using the following steps. First, the sample data is selected using Stratified Random Sampling method. Second, Fitness is calculated for each individual. Thirdly, Roulette Wheel selection method is used to select the individuals from the parents to be involved in recombination. Fourthly, new individuals can be created by using the genetic operators such as crossover and mutation. Finally, some of the new individuals are replaced with their parents.

Dou W, Hu J, Hirasawa K and Wu G, "Quick Response Data Mining Model Using Genetic Algorithm", [3] provided a base for this study to find the maximal frequent itemsets using Genetic algorithm.

2. AFISM over Large Data

The function of extracting association rules over market basket data is identified as essence of proficiency of locating the trends. Association rule mining furnishes a valuable procedure for detecting relationship in the group of items associated with the consumer dealings in a market basket database. In general, the association rule is indicated as X → Y, where X is the antecedent and Y is the consequent. Association rule counts the appearance of Y with respect to the existence of X, determined by the support and confidence value.

Frequent itemset: Assume A to be the set of items, T be the transaction database and σ be the user specified minimum support. An itemset X in A (i.e., X is a subset of A) is denoted as a frequent itemset in T with reference to σ, if support(X) T >= σ

Extracting association rules can be fragmented into two sub-problems as mentioned below:

1. Generating all itemsets that measures support higher than, or as same as the user specified minimal support. That is, generating all large itemsets.
2. Generating all the rules that have minimum confidence.

We can generate the association rule with more than one number of consequent items is generated by the following method:

1. Find the rule in which number of consequents =1.
2. For the given rules p(x → y) and p(x → z), the rule p(x → yz) is generated by the intersection of both the association rules and get a new rule p(x → yz) = p (xyz)/p(x).

3. Genetic Approach FISM for Large Data - GFISM

The GA tool from MATLAB R2006b had been used for GA implementation. The roulette wheel is used for selection process. The crossover and mutation points are randomly generated. The applying of GA factors is highly significant. For GA parameters, if the magnitude of the population is very less, it is hard to get the best solution and for a high magnitude, the convergence time will be prolonged.

Thus, the size is normally 40 - 60. If the crossover Pc is highly diminished, it is hard to hunt through and a Pc value is more, will abuse the individuals with modified values. Therefore, the Pc is regularly 0.3 - 0.9. If the mutation value, Pm is highly diminished, it is difficult to create new individuals and a Pm value is more, it is appreciable to hunt through GA at random. Thus, the Pm is generally 0.01 - 0.2. [15].

The itemset that satisfies the minimum support is selected for initial transaction. The sample data is selected by using stratified random sampling technique. The stratum size is measured for each Stratum. Fitness is identified for each item set. Roulette Wheel selection method is adopted for selecting the sample data set. Frequent Itemset is generated, based on the minimum support. Association rules are framed for Itemsets with min confidence. Rules are stored that satisfies the minimum support and confidence.

Function :GFISM Algorithm

Input: {Randomly ordered sample of feature vectors $X = \{X_1, X_2, \ldots, X_p\}$; and the number of clusters K, and $K < p$}
Output:{ K Set of clusters }

1. Select item with minimum support.
2. Select sample data using Stratified random Sampling.
3. Find the stratum size for stratum h using the Formula nh = (Nh / N) * n
4. Find fitness using the formula t = t + p(i,j) * support(j).
5. Select using Roulette Wheel selection.
6. Find frequent itemsets that satisfy the min support and min confidence.
7. Find Rules for the new population
8. Calculate the confidence of the rule by using the formula,
9. confidence = support / mean.
10. Store the rules that have min Support and minimum Confidence.
11. The fitness function for every rule $x \leftarrow y$ is acquired and the successive circumstances are probed.
12. If (fitness function > min confidence), Set $B = B \cup x \rightarrow y$
13. If the required propagation count is unachieved, then proceed to Step 3.
14. Stop.

4. Partitional Clustering

Partitional Clustering decomposes the dataset into a set of disjoint clusters. It determines an integer number of partitions that optimises a certain criterion function. It selects a criterion and evaluates it for all possible partitions containing K clusters. Symmetric similarity measure like Euclidean distance on binary points is acceptable, since a 0/0 match is as important as a 1/1 match on some dimension [5]. Optimized approach for partitional clustering is to start with an initial partition and move objects so that the value of the criterion function improves. The criterion

function may emphasize the local or global structure of the data and its optimisation is an iterative procedure. It is further subdivided into medoid based and centroid based clustering methods. In medoid based method, the cluster is represented by one of its points. Whereas, in centroid based the mean or average is used to represent the cluster. General procedure for partitioning methos is given below. curse of dimensionality and multi pass nature of these partitional techniques takes more computational time [17]. To improve clustering efficiency, the characteristics of each algorithm need to be understood. There are several new clustering techniques available which provides complete solutions [8].

Initial partition

A. Select k seed points at random or by taking the centroid as the first seed point and the rest at a certain minimum distance from this seed point.
B. Cluster the remaining points to the closest seed point.
a. Select an initial partition with K clusters. Repeat steps b through e until the cluster membership stabilizes.
b. Generate a new partition by assigning each pattern to its closest cluster center.
c. Compute new cluster centers as the centroids of the clusters.
d. Repeat steps b and c until an optimum value of the criterion is found.
e. Adjust the number of clusters by merging and splitting existing clusters or by removing small or outlier clusters.

To obtain a partition which, for a fixed number of clusters, minimizes the square-error where square-error is the sum of the Euclidean distances between each pattern and its cluster center. Data partitioning algorithms divides data into subsets. Since checking all possible subset systems is computationally not feasible, certain greedy heuristics are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the K clusters. Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, partitional algorithms gradually improve clusters. With appropriate input parameters which results in high quality clusters. One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found.

Another approach for partitioning starts with the definition of objective function depending on a partition. The pairwise distances or similarities can be used to compute measures of inter and intra cluster relations. Depending on how representatives are constructed, iterative optimization partitioning algorithms are subdivided into medoid and centroid based methods. Medoid based methods have the most appropriate data point within a cluster that represents it. In K-means algorithm, a cluster is represented by its centroid, which is a mean of points within a cluster. Centroids have the advantage of clear geometric and statistical meaning.

The K-means algorithm [12, 13] is the most popular clustering tool used in scientific and industrial applications. The name comes from representing each of K clusters by the mean or weighted average z of its points, the so called centroid. While this obviously does not work well with categorical attributes, it has the good geometric and statistical sense for numerical attributes. The sum of discrepancies between a point and its centroid expressed through appropriate distance is used as the objective function. Two versions of K-means iterative optimization are known. The first version is similar to Expectation Maximization (EM) [14] algorithm called Forgy's clustering algorithm and consists of two-step major iterations

1. Reassign all the points to their nearest centroids
2. Recompute centroids of newly assembled groups

Iterations continue until a stopping criterion is achieved with no reassignments. The second, classic in iterative optimization version of K-means iterative optimization reassigns points based on more detailed analysis of effects on the objective function caused by moving a point from its current cluster to a potentially new one. If a move has a positive effect, the point is relocated and the two centroids are recomputed. It is not clear that this version is computationally feasible, because the outlined analysis requires an inner loop over all member points of involved clusters affected by centroids shift. Besides these two versions, there have been other attempts to find minimum of K-means objective function.

Function : K-means Algorithm

Input: {Randomly ordered sample of feature vectors X = $\{X_1, X_2, \ldots, X_p\}$; and the number of clusters K, and $K < p$ }
Output:{ K Set of clusters }

Step 1: Choose initial cluster centers $Z_1, Z_2, \ldots, Z_K$ randomly from the N points; $X_1, X_2, \ldots, X_p$ where q is the number of features/attributes.

Step 2: Assign point X_i, $i = 1, 2, \ldots, ptocluster$ $C_j, j = 1, 2, \ldots, K$. if and only if $||X_i - Z_j|| < ||X_i - Z_u||$, u = 1, 2,. . . ,K. and $j \neq t$ .These are resolved arbitrarily.

Step 3: Compute the new cluster centers $Z_1^*, Z_2^*, \ldots, Z_K^*$ as $Z_i^* = \frac{1}{l_j}\Sigma_{X_j \in C_j} X_i$ where $i = 1, 2, \ldots, K$, l_j = Number of point s in C_j.

Step 4: If $Z_i^* = Z_i, i = 1, 2, \ldots, K$ then terminate. Otherwise $Z_i \leftarrow Z_i^*$ and go to step 2.

Reassigning points and immediately re-computing centroids works much better. The wide popularity of K-means algorithm is well deserved. It is simple, straightforward, and is based on the firm foundation of analysis of variances.Most k-means type algorithms have been proved convergent [4, 19]. When the data are continuous and normally distributed K-means algorithm performs well.

Binary dataset clustering based upon Partitional clustering technique is considered here because Partitional clustering technique is simple, fast and convergent [16]. Partitional clustering approaches are in high need since it works for all data types like continuous (factor scores for example), binary indicators (0/1 flags) and ranks (1=low, 2=middle, 3=high).

Binary data streams are clustered by [5]. On-line K-means, Scalable K-means and Incremental K-means algorithms. Here clustering is based on mean-based initialization and incremental learning. These algorithms use extra data structures like summary tables and have some extra matrix operations which will increase the processing time and memory usage. From the above survey it is found that the standard K-means provides better solution for normal binary data clustering.

5. A Novel Approach K- GFISM Algorithm

Association rule mining finds interesting association or correlation relationships among a large set of data items. Association rules are derived from the frequent itemsets using support and confidence as threshold levels. The most influential algorithm for efficient association rule discovery from market databases is K-GFISM algorithm [2] which is proposed by us. This algorithm shows good performance with sparse datasets hence it is considered. The K-GFISM algorithm extracts a set of frequent itemsets from the data and then pulls out the rules with the highest information content for different groups of customers by dividing the customers in different clusters.

K-GFISM Algorithm [2] is based on the GFISM property [10] and the Association rule generation procedure of the GFISM algorithm. The data is partitioned using the multi-pass K-means algorithm. Then each partition is kept in the main memory and then the GFISM procedure is executed, in which the Frequent Itemsets are found iteratively based on minimum support. For the K clusters database scanning required is K times which improves the FISM process. Using these Frequent Itemsets based on confidence, Association Rules are derived. The items in the clusters are very

similar, so that multiple and high informative frequent itemsets are effectively generated in the K-GFISM algorithm.

A novel merged approach, K-GFISM Algorithm for mining Frequentitemsets and deriving Association rules from binary database has been proposed. K-GFISM [2] is based on the GFISM property and the Association rule generation procedure of the GFISM algorithm.

The data is clustered using the K-means algorithm so there will be K clusters or groups. As the database is partitioned, each partition can be placed in the main memory so that repeated database scanning can be avoided. The items in the clusters are very similar, so that multiple and highly informative frequent itemsets are generated in theK-GFISM algorithm.

The GFISM algorithm should be executed multiple (K) times using the K clusters, so K-GFISM is a multi-pass algorithm. Algortihm uses knowledge from previous iteration phase to produce frequent itemsets.

Algorithm : K-GFISM Algorithm

Input: { Binary data matrix X of size p X q, K }
Output:{ Frequent Itemsets and Association rules }

1. Calculate K clusters $(C_1, C_2, \ldots, C_K)$ for Y using K-means algorithm
2. Call function kmeans(Y, K)
3. For each cluster C_i
4. Call function GFISM Algorithm
5. For cluster C_i
6. Return frequent items of cluster C_i
7. End

In K-GFISM algorithm, the data is partitioned using the multi-pass K-means algorithm for the data so that candidate itemsets generated will be very less and Database scanning will also be done for adequate data so that large number of efficient frequent itemsets and Association rules can be generated.

6. Empirical Case Study

6.1 Supermarket

Supermarket is a retail organization which contains large number of items, includes enormous customers and has more competitors. Market basket analysis is used to learn more about customer behaviour of the Anantha Supermarket. The methodology of market basket analysis in Supermarket is to discover the invoices for the purchase transactions. This logic is valid for item-related market basket analysis. Market Basket Data is taken from Supermarket for the duration of 6 months from August 2016 till January 2017. Supermarket is organised in seven separate sections. a) Household items b) Fruits and vegetables c) Bakery d) Kitchenwares e) Toys f) Gifts g) Textiles and h) Pharmacy.

Supermarkets have three floors with eight departments. In the Ground floor, to attract customers special promotions are placed near to the entrances. Bakery items are sold near to it. It includes all baked products and also leading branded packed food items. The household items come next which includes more than 500 items with different brands and prices. This is the main section of the store which provides the major revenue. It provides approximately 75% of the profit for the supermarket.

Nearby is the vegetable and fruit section with more than 120 items with various varieties. Automatic weighing is made here, so that customers can get their required quantity.

In the first floor, comes the kitchenware which includes stainless steel cook wares, plastic household items and much more. Very near to this is the Gifts' section which includes variety of gifts for all sorts of age groups and occasions.

Customers include small retail shops, products' agents and normal individuals. The supermarket makes almost 45% of its sales revenues by selling goods in wholesale for small retail shops. Then, 18% revenue comes from hotels and remaining 37% from the normal retailers. Wholesale has business relations with more than 250 buyers, and Wholesale issues approximately 3000 invoices with total 2,200 items weekly. A Retail shop sells goods to about 600 end consumers daily.

6.2 Marketing and Sales Promotion Campaigns

When sales campaigns are prepared, promoted items must be chosen very carefully. The main goal of a campaign is for entire customers to visit Supermarket and to buy more than they usually do. Margins on promoted items are usually cut; therefore, additional non-promoted items with higher margins should be sold together with promoted items. Therefore, the related items must be chosen to make effective promotions such that promoted items must generate sales of non-promoted items.

Customers who buy a kitchen appliance often also buy several other kitchen appliances. It makes sense that these groups are placed side by side in a retail centre so that customers are attracted and can access them quickly. Such related groups of goods also must be located side-by-side in order to remind customers of related items and to lead them through the centre in a logical manner.

When different additional brands are sold together with the basic brands, the revenue from the basic brands is not decreasing, but increasing. Based on market basket analyses, sets of products are defined and sold together with discount presents.

6.3 Information Systems

Market basket analysis targets customer baskets in order to monitor buying patterns and improve customer satisfaction . Market basket analysis is an important component of analytical CRM in retail organizations. By analysing, recurring patterns in order to offer related goods together can be found and therefore the sales can be increased. Sales on different levels of goods classifications and on different customer segments can be tracked easily. Market basket analysis will be taken into consideration to improve the sales in Supermarket.

Different analyses and reports were performed in Supermarket' transactional information systems, much of the analytical data was held in Excel spreadsheets and Access databases. The inventory levels of each item in the supermarket on a monthly basis are stored in Access database and enables detailed inventory analyses and detection of critical items. All the time it tries to use adequate analytical and data mining methodologies in order to improve the whole system of business reporting. Key success factors such as net margin, net margin per item, net margin per customer, number of new customers are measured and reported on monthly basis.

6.4 Binary Data Pre-processing

In this dissertation, the transaction is observed from copy bills or invoice copies which contain the items purchased by different customers. Copy bill is the duplicate copy of the bills generated in the system which is used for future reference. Each copy bill is considered as a transaction. On an average 962 transactions are done per day. There are around 500 household items, 45 vegetables, 90 bakery products, 290 kitchenwares, 450 toys and gifts. Since the household section provides major profit of the store, household items are considered for this market basket analysis. Using the copy bills item names' are coded as I1 to I850 for the different transactions which is numbered as TR00001 to TR09620 for 10 days. The data are converted into a 9620 X 302 binary data. For easy and effective processing matrix format is considered with Transactions as rows and the item names' as columns for the binary data.

The goal is to find the frequent items which often occur together and so transactions with one or two items is rejected for effectiveness. Transactions with more number of items provide useful information about customers' behavior. For a specific

transaction i, if an item j is purchased then the Transaction Binary Matrix position (i,j) is made as 1. If the item j is not purchased in the transaction i then the Transaction Binary Matrix position (i,j) will be made as 0. If Some dummy transactions are there with no items, it should be rejected.

6.5 Customer Segmentation

The complexity and especially the diversity of phenomena have forced society to organize the customers based on their similarities on their purchase behaviour. Clustering partition a data set into several dis-joint groups such that points in the same group are similar to each other according to some similarity metric. Clustering is useful to build and identify the different clusters or segments of a market. In K-GFISM algorithm, initially the binary data is clustered such that the customers are categorized and then the clusters' frequent itemsets are generated. The binary data is clustered using the multi-pass K-means algorithm. K-GFISM algorithm addresses different customer groups' satisfaction using clustering property.

7. Experimental Results

From the household section of the Supermarket, sample market basket dataset is taken using the invoice copies or copy bills of the supermarket. 9620 X 300 sample Binary dataset is manipulated with GFISM and K-GFISM algorithm and the results are shown below. For K-GFISM algorithm, K is selected as 3 (3 clusters) for this comparison. The result analysis of GFISM and K-GFISM for super market dataset with confidence of 100% is represented in Table 1.

Table 1. GFISM & K-GFISM Result Analysis for Supermarket Dataset with Confidence of 100%

Support (%)	Max. FIs		FIs		ARs	
	GFISM	K-GFISM	GFISM	K-GFISM	GFISM	K-GFISM
4	3	5	45	2035	30	9684
5	2	5	29	2009	2	9656
6	2	5	27	2009	2	9656
10	1	5	1	1830	0	8916
20	0	5	0	1829	0	1562

Where

- Max FIs – Maximum possible K-itemset
- FIs – Total Number of Frequent Itemsets generated
- ARs – Total Number of Association Rules generated

From the Table 1, GFISM and K-GFISM algorithms are compared based on the frequent itemsets and association rules generated. GFISM algorithm provides output only for very low support values. Very low support values are meaningless because it shows nothing about the customers' behavior.

The frequent itemsets (FIs) generated for GFISM are given below, 1-itemset are {I3}, {I11}, {I12}, {I13}, {I15}, {I20}, {I22}, {I23}, {I24}, {I25}, {I26} {I29}, {I32}, {I34}, {I35}, {I39}, {I40}, {I44}, {I47}, {I50}, {I61}, {I179}, {I190}, {I291}, {I293}, {I297} 2-itemset are {I22, I23} and Association Rules (ARs) generated are I22 → I23, I23 → I22 are the 2 exact rules of K-GFISM algorithm for 50% support. It implies that "if I22 item is purchased, then I23 are purchased" and " if I23 item is purchased together, then I22 are purchased" with 100% confidence.

GFISM algorithm implies that I22 and I23 items are frequently purchased together with 100% confidence for 5% of the population. Only one 2-itemset is generated with twenty two 1-itemsets. GFISM algorithm provides 3-itemsets for 4% support and 2-itemsets for 5% support.

K-GFISM provides 5-itemsets up to 20% support values. GFISM derives only 2 ARs for 5% support which provides no useful customer information and nothing for higher support values. But, K-GFISM generates 1830 FIs for 10% and 20% support respectively. To get the consumer behaviour of the store at least 40% support is needed but with GFISM it is impossible. It is possible for K-GFISM in higher support values. K-GFISM generates FIs and their ARs which are described in the following tabulations. K-GFISM provides large number of FIs and ARs for lower support values. In market basket analysis, to analyse the frequency of item purchase some higher values of support is required, hence K-GFISM is better compared to GFISM. Figure \ref{ch5_fig1} analyses the frequent itemset generation for GFISM and K-GFISM algorithm with confidence as 100% for various support values.

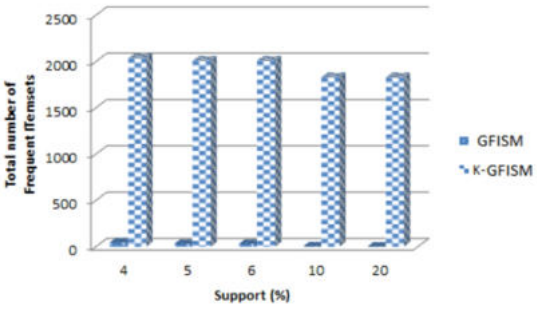

Figure 1: GFISM and K-GFISM Algorithm Analysis Based on Fis

Figure 1 depicts the performance analysis of GFISM and K-GFISM algorithms based on Association rule generation for various support values with 100 % confidence. Exact rules are ARs with 100% confidence.

Figure 2 gives the number of exact rules generated for the GFISM and K-GFISM algorithms. 213 FIs are generated by K-GFISM algorithm with 2 clusters for support = 35% and the ARs generated for various confidence levels are illustrated in Table 2. As the confidence increases the number of ARs generated decreases.

For a supermarket support value 36% is nominal hence K-GFISM is analysed with that value for different confidence levels in Table 2. K-GFISM generates 576 ARs for 40% confidence. As the confidence increases the number of ARs generated decreases. For 100% confidence, K-GFISM generates 220 ARs which proves the trustworthiness rules. Figure 3 depicts the K-GFISM algorithm AR generation for various confidence levels.

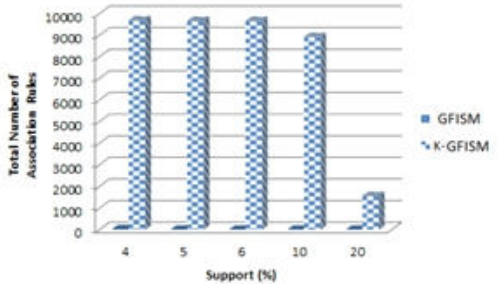

Figure 2: GFISM and K-GFISM Algorithm Analysis Based on Ars

Table 2. K-GFISM Algorithm Analysis Based on ARs Generated for Various Confidence Thersolds with Support of 35%

Confidence (%)	Total Number of Association Rules
40	576
60	561
80	405
90	307
100	220

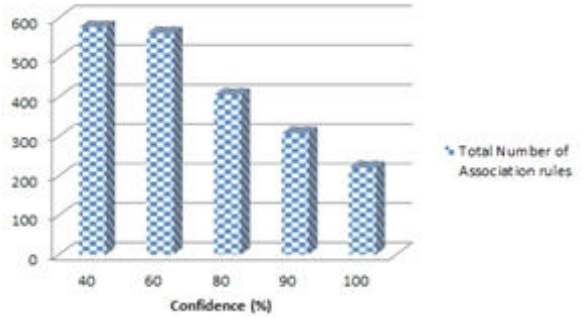

Figure 3: K-GFISM Algorithm Analysis for Various Confidence Thersold Levels

K-GFISM algorithm divides the customers into different segments /groups / clusters initially. Then it finds the frequent itemsets and association rules for those categories separately. K-GFISM algorithm attempts to find consumer behaviours as groups, so that those specific groups of people can be satisfied effectively. Consider for example seasonal promotions can be provided for particular groups like Deepavali festive season which can improve the purchase and the profit.

Table 3 result shows that the total number of frequent itemsets increases as the number of clusters increases for the same support and confidence levels. If the number of groups increases means different ARs need to satisfy the specific groups with more number of ARs which is depicted in Figure 4. We can neglect some groups with low information due to its negligible information.

Table 3. K-GFISM Algorithm Performance Analysis for Different Number of Clusters with 100% Confidence

Dataset 9620 X 300	Support(%)	Number of Clusters		
		2	3	5
FIs	62	9	74	69
	50	33	109	570
ARs	62	4	99	292
	50	37	195	1791

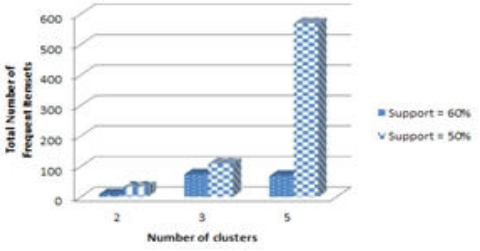

Figure 4: K-GFISM Algorithm Analysis Based on FIs for Different Clusters

Based on the number of clusters, K-GFISM algorithm provides different number of ARs. If the number of clusters increases then the number of FIs and ARs generated also increases. It means that if the supermarket has more number of customer groups then inorder to satisfy them different numbers of ARs are generated which is depicted in Figure 5.

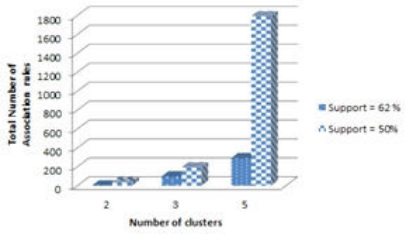

Figure 5:K-GFISM Algorithm Analysis Based on ARs for Different Clusters

As the number of clusters plays an important role in K-GFISM algorithm, clustering must be done effectively. In K-GFISM algorithm, clustering is done using K-means algorithm. The clustering efficiency is measured using the popular metrics like Inter-cluster and Intra-cluster distances.

Inter-cluster distance means the distances between different clusters, and it should be maximized i.e distance between their centroids. Intra-cluster distance is the sum of distances between objects in the same cluster, and it should be minimized i.e., distance between the centroid and all objects in the cluster

Table 4 Perfomance Analysis of K-GFISM Algorithm Based on Customer Segmentation

Number of Clusters	Intra- Cluster Distance	Inter- Cluster Distance
5	4.37	15.8
3	5.23	18.71
2	8.23	33.33

From Table 4, it shows that the efficient clusters are generated with high inter-distance between clusters. Compact clusters with low intra-distance between elements. The efficiency in clustering implies the effective customer segmentation of the Supermarket. Clustering performance analysis depicted in Figure 6.

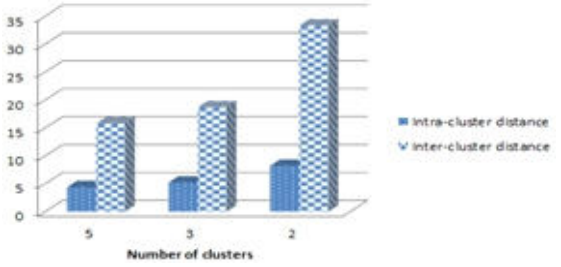

Figure 6: Clustering Performance Analysis

Table 5. Dimensionality Variation on 2 clusters of K-GFISM with Support of 62% and Confidence as 100%

Dataset Size	FIs	ARs
9620 X 500	31	27
9620 X 400	9	4
9620 X 300	7	2

K-GFISM algorithm FI generation are described below.
4-itemsets for 500 attributes are [I26,I34,I303,I311].
Ars derived from the 4-itemsets are
[I26,34→303,311]
[I26,I303→34,311] [I26,I311→I34,I303]
[I34,I303→I26,I311] [I34,I311→I26,I303] [I303,I311→[I26,34]
[I26,I34,I303→I311] [I26,I34,I311→I303] [I34,I303,I311→I26]

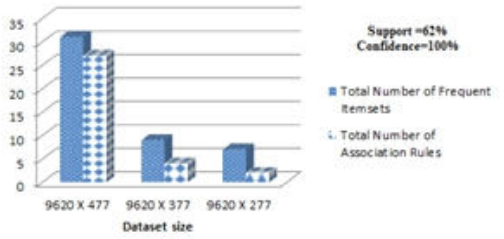

Figure 7: K-GFISM Algorithm Performance Analysis on Various Dimensions

Figure 7 depicts the K-GFISM Result analysis for Supermarket dataset with support of 45%.

In Supermarket, there are approximately 500 household items, the number of items normally increase and decrease. From the copy bills it is found that some items are purchased rarely hence these are neglected from analysis. To find the frequency of items common items are considered.

K-GFISM algorithm is executed with more number of attributes for the supermarket dataset and it is tabulated in Table 5.4. It shows that the total Number of FIs and ARs generated for K-GFISM algorithm is directly proportional to the number of attributes. Market basket analysis of Supermarket based on various number of transactions are done in Table 6. Two weeks has 13468 transactions, 10 days has 9620 and 1 week has 6438 transactions which are analysed based on the K-GFISM in this table. This result shows that the total number of FIs and ARs generated increases as the number of records increases.

Table 6. K-GFISM Algorithm for Different Number of Transactions for 2 Clusters with Support as 6% and Confidence of 100%

Dataset Size	13468 X 300		9620 X 300		6438 X 300	
	GFISM	K-GFISM	GFISM	K-GFISM	GFISM	K-GFISM
FIs	32	19523	29	2009	27	11731
ARs	2	69048	2	9656 2	2	88072

Figure 8 depicts the performance analysis of different number of transactions with GFISM and K-GFISM.

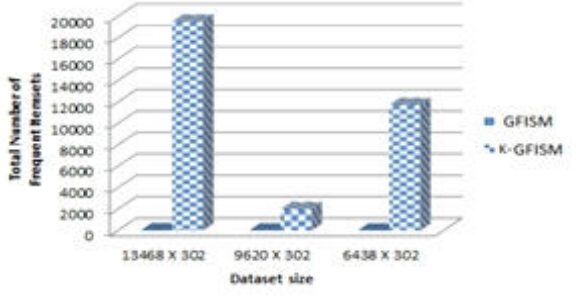

Figure 8: GFISM and K-GFISM Algorithms Performance Analysis for Different Number of Transactions

Computational complexity is directly proportional to dimensionality and number of records. For sparse dataset like market databases, K-GFISM algorithm is the best algorithm for market basket analysis. The market basket analysis results based on K-GFISM algorithm are suggested to the Supermarket. The management accepted to implement these frequent itemsets and association rules in the near future.

8. Conclusion

K-GFISM algorithm effectively generates highly informative frequent itemsets and association rules for the Supermarket. Supermarket widely used the market basket analyses to manage the placement of goods in their store layout. Related products are placed together in such a manner that customers can logically find items he or she might buy which increases the customer satisfaction and hence the profit. Customers are segmented and association rules are separately generated to satisfy their specific needs in a cost effective manner using some special promotions for the common groups. From the results it is shown that the market basket analysis using K-GFISM algorithm for Supermarket improves its revenue.

REFERENCES

[1] Ananthanarayana.V.S.,NarasimhaMurthy.M. and Subramanian.D.K., Efficient Clustering of Large Data Sets, Pattern Recognition, vol. 34, pp. 2561-2563, 2001.
[2] Ashok Kumar D and Loraine Charlet Annie M.C.,"Frequent Item set mining for Market Basket Data using K-Apriori algorithm" , International Journal of Computational Intelligence and Informatics, Volume 1, No. 1, pp.14-18, 2011.
[3] Berry, M.J.A., Linoff, G.S.: Data MiningTechniques: for Marketing, Sales and Customer Relationship Management (second edition), Hungry Minds Inc., 2004.
[4] Bezdek J.C., A Convergence Theorem for the Fuzzy ISODATA Clustering Algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 2, No. 8, pp.153-155, 1980.
[5] Carlos Ordonez, Clustering Binary Data Streams with Kmeans, DMKD03: 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2003.
[6] Chen Y.-L., Tang K., Shen, R.-J., Hu, Y.-H.: "Market basket analysis in a multiple store environment", Decision Support Systems, 2004.
[7] Das S and Saha B, "Data Quality Mining using Genetic Algorithm", International Journal of Computer Science and Security, Vol. 3, No. 2, pp. 105-112, 2009.
[8] Georg Peters., Some Refinements of Rough K-means Clustering, Pattern Recognition, 39, pp. 1481 - 1491, 2006.
[9] Ghosh S., Biswas S., Sarkar D and Sarkar P.P., "Mining Frequent Itemsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications, Vol. 1, No. 4, pp.133-143, 2010.
[10] Han .J., and Kamber .M., "Data Mining: Concepts and Techniuqes", Morgan Kaufmann Publishers, San Francisco, CA,2001.
[11] Han .J., and Kamber .M., Data Mining Concepts and Techniques. Morgan Kanufmann, 2000.
[12] Hartigan.J. andWong.M., A K-Means Clustering Algorithm, Applied Statistics, 28, pp. 100-108, 1979.
[13] Highleyman.W.H., The Design and Analysis of Pattern Recognition Experiments, Bell Syst. Tech,J., vol. 41, pp. 723-727, 1962.
[14] Mclachlan.G. andKrishnan.T., The EM Algorithm and Extensions. John Wiley and Sons, New York, 1997.
[15] S.N. Sivanandam, and S.N. Deepa, " Introduction to Genetic Algorithms, New York: Springer-Verlag Berlin Heidelberg.
[16] Selim.S.Z. andIsmail.M.A., K-Means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality, IEEE Transactions On Pattern Analysis and Machine Intelligence, vol. 6, No.1, pp. 81-87, 1984.
[17] Ting Su, Dy J., A Deterministic Method for Initializing K-means Clustering, 16th IEEE International Conference on Tools with Artificial Intelligence, pp. 784 - 786, 2004.
[18] Wakabi-Waiswa P.P., Baryamureeba V and Sarukesi K, "Generalized Association Rule Mining Using Genetic Algorithms", International Journal of Computing and ICT Research, Vol. 2 No. 1, pp. 59-69, 2008.
[19] Weng S.-S., Liu J.-L.: "Feature-based recommendations for one-to-one marketing", Expert Systems with Applications, Vol. 26, pp. 493-508, 2004.