Market-Basket Analysis using Agglomerative Hierarchical approach for clustering a retail items

Rujata Saraf and Sonal Patil

†student of Computer Science and Engineering, North Maharashtra University, Jalgaon, India
††HOD of Information Technology Deparatment in G.H.R.I.E.M, North Maharashtra University, Jalgaon,India

Summary
With the advent of data mining technology, cluster analysis of items is frequently done in supermarkets and in other large-scale retail sectors. Clustering of items has been a popular tool for identification of different groups of items where appropriate programs and techniques in data mining like Market-Basket analysis have been defined for each group separately with maximum effectiveness and return. For example, items frequently purchased together are placed in one place in the shelf of a retail store. There are various algorithms used for clustering. Hierarchical algorithms find successive clusters using previously established clusters. These algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). The paper presents the Market-Basket Analysis using Agglomerative ("Bottom-up") hierarchical approach for clustering a retail items. Agglomerative hierarchical clustering creates a hierarchy of clusters which are represented in a tree structure called a Dendorogram. In agglomerative hierarchical clustering, dendrograms are developed based on the concept of 'distance' between the entities or, groups of entities. . The clustering will done in such a way that the Purpose of Market-Basket Analysis will achieve.

Key words:
Market-Basket analysis, Hierarchical Clustering, Agglomerative Hierarchical Clustering, Dendrogram etc…

1. Introduction

In recent years, customer relationship management (CRM) is the most important application in business world. The primary task of any type of business is to integrate 'relationship technology' with 'loyalty schemes'. By providing a different loyalty schemes, CRM is expected to enhance value to customers through raising satisfaction levels on transactions. If customers appreciate the value provided by a scheme in CRM, they are expected to continuously enhance the relationship with the firm through loyalty to the products/brands, purchasing more, advocating the firm to others, etc. In today's world, concept of Data Mining i.e Market-Basket analysis is widely used in retail stores to make the task of Planning such a schemes much easier and efficient.

Retail stores hence consider to be a Best place for achieving market-basket analysis as it's a place where an ample number of different quality products, in different quantities with different rates are made available to customers. The items in retail stores are organized in proper and systematic manner so that an Individual can select the product from many options available and buy according to individual needs. And hence choosing the correct location for a particular product in retail store is a challenging task.

One such a illustration have been presented in this paper with the help of clustering of retail items for the purpose of Market-basket analysis. In today's world the places like Super-market, Malls are at the center point for any type of shopping. Super-market is a large form of the traditional grocery store and a self-service shop offering a wide variety of food and household products, organized into large shelfs. These places are consider to be most crowdie places where a huge mob of customers are find out in order to purchase different items as per their needs. Customers are giving preference to such a supermarket as the items in supermarket are arranged in proper and systematic manner on shelf. People can easily find whatever they want to purchase because of such a systematic arrangement of product on shelf. The shelf contain different varieties of single product on single shelf or the shelf have been partitioned into different section where each section can contain a particular product along with its varieties [6].

Generally in super-market such arrangements of product is done manually, i.e. lots of human resources are require to make such arrangement with which the customers can easily get whatever they want more efficiently. The products are arranged in such a way that the items which are purchased together are placed in one shelf beside to each other and by providing the different loyalty schemes on such a items, the total sales of the product have been increased [5]. But while doing so it's very difficult to predict which products should be kept beside to each other and in which product the customer will shown their interest. For this purpose it is necessary to find out which products are frequently purchased by customers from the total sale and by using this we can easily achieve the market-basket analysis by placing the most frequently purchased items besides to those items which are

Manuscript received March 5, 2016
Manuscript revised March 20, 2016

necessary along with the purchased item but not compulsory to purchase.

By considering this scenario, it is necessary to group all such a items which can altemately increase the total sale of the product. For this purpose the Data Mining Techniques like Mining algorithms and Clutering techniques are useful. With the help of mining algorithms, we can easily find out in which items the users are interested and which items are frequently purchased by customers. Similarly after finding such a items, Clustering techniques are usefull to achieve the purpose of Market-Basket analysis.

It had been discovered that Market basket analysis is an important component of analytical system in retail organizations to determine the placement of goods, designing sales promotions for different segments of customers to improve customer satisfaction and hence the profit of the supermarket. In practical we can find the concept of Market-basket analysis in retail stores, super markets, malls, mega marts ect. These are the places where customer are mostly attracted for any type of purchase. CRM in such places expected to enhance value to customers through raising satisfaction levels on transactions by providing different loyalty schemes. If customers appreciate the value provided by a scheme in CRM, they are expected to continuously enhance the relationship with the firm involved through loyalty to the products/brands, purchasing more, advocating the firm to others, etc. and hence the retailers must know the needs of customers and it gives retailer a good information about related sales. Hence it is necessary to adopt the computerize techniques with which the total sales can be increase in supermarket.

For this purpose ,Several aspects of market basket analysis have been studied, such as using customer interest profile and interests on particular products for one-to-one marketing purchasing patterns in a multi-store environment to improve the sales[2][4]. But the existing technique related to clustering of such a retail items which can directly affect or increase the total revenues of the market has some pitfalls while working. The proposed work in this paper will try to overcome the encountered drawbacks in existing system by using the Agglomerative Hierarchical clustering.

The complete paper is organized as per follows: Section 1 introduces the paper title in short concepts. Section 2 describes backgroung for the topic including Data mining process, Basic techniques in Data Mining with main task of Data Mining. Also gives the details about Clustering and Various clustering Techniques in Data Mining with their respective advantages and disadvantages. Section 3 gives details regarding Agglomerative Hierarchical Clustering and the Dendrogram concept used in Aglomerative hierarchical clustering with its importance have been decribed. Section 4 gives the details regarding complete working with cluster analysis, Section 5 is about some evaluation done while proposed work execution. Section 6 describes General result analysis for the proposed work. Section 7 finally conclude the complete topic with final result as clustering of retail items is necessary in retail stores as its directly proportional to the total sales and profit. And the Agglomerative Hierarchical Clustering is Best technique to achieve Market-basket analysis in retail stores.

2. Literature Survey

There are several major data mining techniques have been developed and used in data mining projects recently including association, classification, clustering, prediction and sequential patterns etc., are used for knowledge discovery from databases [10].

Association : Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction[9]. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together.

Classification : Provide an overview of the classification problem and introduce some of the basic algorithms. Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. For Example, Teachers classify students' grades as A, B, C, D, or F. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics [13].

Clustering: Clustering is "the process of organizing objects into groups whose members are similar in some way". A cluster is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.For example. In a library, books have a wide range of topics available. The challenge is how to keep those books in a way that readers can take several books in a specific topic without irritate. By using clustering technique, we can keep books that have some kind of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in a topic, he or she would only go to that shelf instead of looking the whole in the whole library.

Prediction : The prediction as it name implied is one of a data mining techniques that discovers relationship between independent variables and relationship between dependent and independent variables[8].

Sequential Patterns : Sequential patterns analysis in one of data mining technique that seeks to discover similar patterns in data transaction over a business period [7]. The

uncover patterns are used for further business analysis to recognize relationships among data.

2.1. Clustering in Data Mining

Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships. The aim is the objects in a group should be similar (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better the clustering. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters. There are various types of clusters.

1. Well-Separated Cluster Definition: A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster. Sometimes a threshold is used to specify that all the points in a cluster must be sufficiently close (or similar) to one another [11].

2. Center-based Cluster: A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster. The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the "most representative" point of a cluster.

3. Contiguous Cluster (Nearest Neighbor or Transitive Clustering): A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

4. Density-based: A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. This definition is more often used when the clusters are irregular or intertwined, and when noise and outliers are present.

5. Similarity-based Cluster: A cluster is a set of objects that are "similar", and objects in other clusters are not "similar." A variation on this is to define a cluster as a set of points that together create a region with a uniform local property [15], e.g., density or shape. The proximity measure (and the type of clustering used) depends on the attribute type and scale of the data. The three typical types of attributes are shown in Table 1, while the common data scales are shown in Table 2

Table 1. Different types of Attributes

Attribute Type	Description
Binary	Two values,e.g.True & False
Discrete	A finite number of values OR an integer, e.g. counts
Continuous	An effectively infinite number of real values, e.g. Weight

Table 2. Different Scale type

Scale Type		Description
Qualitative Scales	Nominal	The values are just different names, e.g., colors or zip codes.
	Ordinal	The values reflect an ordering, nothing more, e.g., good, better, best
Quantitative Scales	Ratio	The scale has an absolute zero so that ratios are meaningful. Examples are physical quantities such as electrical current, Pressure
	Interval	The difference between values is meaningful, i.e., a unit of Measurement exits. For example, temperature on the Celsius

2.2. Clustering Techniques:

In clustering, similarities and dissimilarities are assessed based on the attribute values describing the objects. It is an unsupervised learning and faces many challenges such as a high dimension of the dataset, arbitrary shapes of clusters, scalability, domain knowledge, ability to deal with noisy data and insensitivity to the order of input records. Large number of clustering methods [8] had been proposed till to address these challenges.

1) Partition Based Clustering:

Partitioning clustering divide data into several subsets. The reason of dividing the data into several subsets is that checking all possible subset systems is computationally not feasible; there are certain greedy heuristics schemes are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the k clusters. For this purpose Relocation algorithms are used which gradually improve clusters.

2) Hierarchical Clustering: A hierarchical method creates a hierarchical decomposition of the given set of data objects. This hierarchical structure is represented by a tree stricture where every cluster node contains child clusters, sibling

clusters partition the points covered by their common parent. In hierarchical clustering each item to a cluster is assigned such that if we have N items then we have N clusters [14]. It Finds the closest pair of clusters and merge them into single cluster. Compute distance between new cluster and each of old clusters. This method works on both bottom-up and top-down approaches.Based on the approach hierarchical clustering is further subdivided into agglomerative and divisive.

The agglomerative hierarchical technique follows bottom-up Approach and begin with each element as a separate cluster and merge them into successively larger clusters.and hence these clustering also called as "Bottom-Up clustering". Whereas Divisive hierarchical clustering follows the top-down approach. This method starts with a single cluster containing all objects, and then successively splits resulting clusters until only clusters of individual objects remain [12].

3) Density based Clustering:

The density based method has been developed based on the notion of density, which is the no of objects in the given cluster, in this context. To discover clusters with arbitrary shape, density based clustering methodology have been developedThe general idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold; that is for each data point within a given cluster; the neighbourhood of a given radius has to contain at least a minimum number of points[11].

4) Grid Based Clustering:

As the name suggest, grid based clustering methods uses a multidimensional grid data structure. It divides the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. Grid based clustering maps the infinite number of data records in data streams to finite numbers of grids [10]. The grid based methods use the single uniform grid mesh to partition the entire problem domain into cells and the data objects located within a cell are represented by the cell using a set of statistical attributes from the objects. One of the distinct features of this method is the fast processing time, as it depends not on the number of data objects but only on the number of cells. The grid-based clustering algorithms are STING, Wave Cluster, and CLIQUE [12].

As per all the topics discussed so far, Cluster Analysis in data mining used for grouping the objects, called as a cluster/s, which consists of the objects that are similar to each other in a given cluster and dissimilar to the objects in other cluster. Now a days the concept of Market-basket analysis Is widely encounter in retail stores, super markets, malls, mega marts ect. These are the places where customer are mostly attracted for any type of purchase. In such places it is expected to enhance value to customers through raising satisfaction levels on transactions by providing different loyalty schemes. If customers appreciate the value provided by a scheme, they are expected to continuously enhance the relationship with the firm involved through loyalty to the products/brands, purchasing more, advocating the firm to others, etc. and hence the retailers must know the needs of customers and it gives retailer a good information about related sales. Hence it is necessary to adopt the computerize techniques with which the total sales can be increase in supermarket.

But the existing technique related to clustering of such a retail items which can directly affect or increase the total revenues of the market, has some pitfalls while working. With the application of clustering in all most every field of science and technology, large number of clustering algorithms had been proposed which satisfy certain criteria such as arbitrary shapes, high dimensional database, and domain knowledge and so on. It had been also proved that it is not possible to design a single clustering algorithm which fulfils all the requirement of clustering. Therefore it is very difficult to select any algorithm for a specific application. So as a final outcome of the complete literature survey the hierarchical clustering technique have been chosen to cluster a retail items as hierarchical clustering have an ease of headlining any forms of similarity or distances also this technique can be applicable to any type of attribute. By considering this points proposed work will try to overcome the encountered drawbacks in existing system by using the Agglomerative Hierarchical clustering.

3. Proposed System:

3.1. Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called a dendrogram[17]. A dendrogram is a branching diagram that represents the relationships of similarity among a group of entities. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations. Agglomerative hierarchical clustering algorithms can be characterized as *greedy*, in the algorithmic sense. Such hierarchical algorithms may be conveniently broken down into two groups of methods. The first group is that of linkage methods – the single, complete, weighted and unweighted average linkage methods [.These are methods for which a graph representation can be used. The second group of hierarchical clustering methods are methods which allow the cluster centers to be specified. The proposed work is based on the second type of agglomerative hierarchical clustering methods where cluster centers are specified with the consideration of dissimilarities between two different

points. It can be said as stored Dissimilarity based approach as it s an alternative to the general dissimilarity based algorithm for clustering. The stored data approach for the agglomerative hierarchical algorithm is as follows :

Step 1: Examine all interpoint dissimilarities &form cluster from two closest points.
Step 2: Replace two points clustered by representative point (center of gravity) or by cluster fragment
Step 3: Return to step 1, treating clusters as well as remaining objects, until all objects are in one cluster.
In steps 1 and 2, "point" refers either to objects or clusters, both of which are defined as vectors in the case of cluster center methods. This algorithm is justified by storage considerations, since we have O(n) storage required for n initial objects and O(n) storage for the n−1 (at most) clusters. In the case of linkage methods, the term "fragment" in step 2 refers (in the terminology of graph theory) to a connected component in the case of the single link method and to a clique or complete subgraph in the case of the complete link method. Without consideration of any special algorithmic "speed-ups", the overall complexity of the above algorithm is O(n3) due to the repeated calculation of dissimilarities in step 1, coupled with O(n) iterations through steps 1, 2 and 3.While agglomerating the two closest point in one cluster, the inversion criteria of arbitrary points[5] must be taken into consideration. The inversion situation while hierarchy construction is explain as follows with an example.
For example: Consider the five arbitrary points as shown in figure 1

Fig 1. Five arbitrary Points with their respective nearest niebours.

A nearest Niebour chain consists of an arbitrary point a in Fig 1 followed by its nearest neibour b which is followed by the nearest neibour from among the remaining points c, d, and e in Fig. 1 of this second point; and so on until we necessarily have some pair of points which can be termed reciprocal or mutual nearest neibours. Such a pair of reciprocal nearest neibours may be the first two points in the chain; and with assumption that no two dissimilarities are equal. While constructing a Nearest Neibour chain, irrespective of the starting point, we may agglomerate a pair of Reciprocal nearest neibours as soon as they are found as shown in figure 3(B). Because there is no guarantees that whether we can arrive at the same hierarchy as if we used traditional "stored dissimilarities" or "stored data" algorithms.

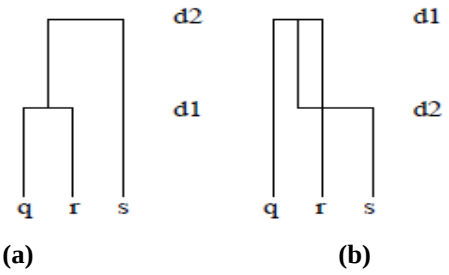

Fig 2. Hierarchy representation (a) without inversion and (b) with an inversion.

Essentially this is the same condition as that under which no inversions (figure 2(a)) or reversals are produced by the clustering method. Fig.2 gives an example of this, where s is agglomerated at a lower criterion value (i.e. dissimilarity) than was the case at the previous agglomeration between q and r. Our ambient space has thus contracted because of the agglomeration. This is due to the algorithm used – in particular the agglomeration criterion – and it is something we would normally wish to avoid.

3.2. Dendrogram in Agglomerative Hierarchical Clustering :

The agglomerative hierarchical clustering build a cluster hierarchy that is commonly displayed as a tree diagram called a Dendrogram. They begin with each object in a separate cluster. At each step, the two clusters that are most similar are joined into a single new cluster. A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to Individual observations. Any valid metric may be used as a measure of similarity between pairs of observations. The choice of which clusters to merge or split is determined by a linkage criterion [17] which is a function of the pair-wise distances between observations. Different clusters are obtained at different levels of the tree diagram of dendrogram. This gives an opportunity to understand how multiple Levels of particular Dendrogram have been read to make sure that every cluster of that tree different from another one.
The process of Agglomerative Hierarchical Clustering (AHC) starts with the single observation clusters and progressively combines pairs of clusters, forming smaller numbers of clusters that contain more observations [17] . Then clusters successively merged until the desired cluster structure is obtain.
For Example: in fig 3., six elements {a}, {b}, {c}, {d}, {e} and {f} are shown in the Euclidean field. Clustering is to be done on the basis of Euclidean distance of similarity distance The most important distinction one should be

concider while clustering is whether the clustering uses symmetric or asymmetric.The distance function used in the work have the property that distances are symmetric i.e. the distance from object A to B is the same as the distance from B to A. The first step is to determine which elements to merge in a cluster. Usually, we want to take the two closest elements, according to the chosen distance metric.

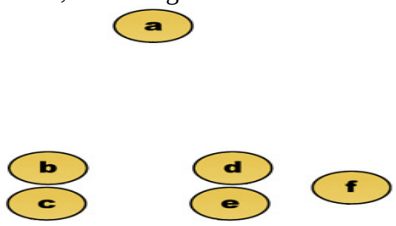

Fig 3 : Objects for clustering in A Field

Based on Euclidean distance metric, the Dendrogram will be constructed. Dendrogram is a hierarchy of clusters represented in a tree like structure as shown in following figure. Dendrograms are made up of subtrees, and those subtrees, in turn, have subtrees nested within them. Each cluster of execution profiles in a dendogram comprises a subtree of the dendrogram, and each subtree or cluster has several attributes that can be examined and used in the refinement technique [2]. The root of the tree consists of a single cluster containing all observations, and the leaves correspond to individual observations. Any valid metric may be used as a measure of similarity between pairs of observations. The choice of which clusters to merge or split is determined by a linkage criterion, which is a function of the pair-wise distances between observations. Different clusters are obtained at different levels of the tree diagram of Dendrogram. This gives an opportunity to compare the performance of various clustering in different levels with respect to a selected performance criterion. By examining the way in which executions are arranged in clusters and subtrees, their similarity to each other and to other clusters may be evaluated. The height of any subtree in a dendrogram indicates its similarity to other subtrees - the more similar two Executions or clusters are to each other, the further from the root their first common ancestor is.

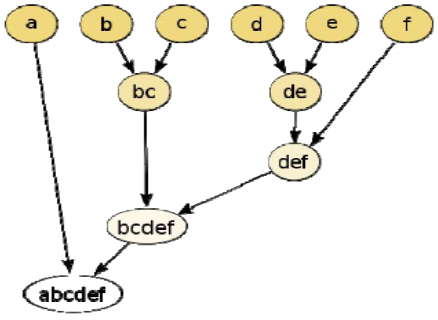

Fig 4. Hierarchical Clustering – Dendrogram

4. Implementation:

4.1. Algorithm for Clustering Retail Items :

For n samples, agglomerative algorithms begin with n clusters and each cluster contains a single sample or a point. Then two clusters will merge so that the similarity between them is the closest until the number of clusters becomes 1 or as specified by the user. [4] [13]. The algorithm works as per following steps :

1. Start with n clusters, and a single sample indicates one cluster.
2. Find the most similar clusters Ci and Cj then merge them into one cluster.
3. Repeat step 2 until the number of cluster becomes one or as specified by the user.

The distances between each pair of clusters are computed to choose two clusters that have more opportunity to merge. There are several ways to calculate the distances between the clusters Ci and Cj, the Linkage method is one of such a method which is basically used to calculate the distance between two clusters which we have to merge together. All such a Linkage methods are describe in table 3.

Table 3. Linkage Methods to calculate the association between two clusters

Single Linkage	$d_{12} = \min_{ij} d(X_i, Y_i)$	This is the distance between the closest members of the two clusters.
Complete Linkage	$d_{12} = \max_{ij} d(X_i, Y_i)$	This is the distance between the farthest apart members.
Average Linkage	$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j)$	This method involves looking at the distances between all pairs and averages all of these distances

Notation:

X1, X2, ... , Xk = Observations from cluster 1

Y1, Y2, ... , Yl = Observations from cluster 2

d (x,y) = Distance between a subject with observation vector x and a subject with observation vector y.

4.2. Cluster Merging and Splitting :

Hierarchical clustering constructs a hierarchy of clusters by either repeatedly merging two smaller clusters into a larger one or splitting a larger cluster into smaller ones. The crucial step is how to best select the next cluster(s) to split or merge. Several methods have been proposed for this purpose in and it had been find out that the average similarity is the best method in divisive clustering and the MinMax linkage is the best in agglomerative clustering. Cluster balance is a key factor to achieve good performance. We also introduce the concept of objective function saturation and clustering target distance to effectively assess the quality of clustering. The

hierarchical cluster structure provides a comprehensive description of the data, which is quite useful for a number of applications. Given a dataset, however, the hierarchical cluster structure is not unique. It depends crucially on the criterion of choosing the clusters to merge or split.

Besides hierarchical methods, there are many other clustering methods[15] A popular method is the K- means method, which is essentially a function minimization, where the objective function is the squared error. In most applications, one initializes K mean-vectors and directly optimize the objective function to obtain the optimal clusters. One can also follow a hierarchical divisive approach, and split a current cluster into two using the K-means method [18]. The gaussian mixture model using EM algorithm directly improves over the K-means method by using a probabilistic model of cluster membership of each object. Both K-means method and Gaussian mixtures uti- lize directly the coordinates (attributes or variables) of the data points. From a general data clustering perspective, given all distances between data points, the cluster structure of the dataset is uniquely determined. Using the concept of similarity, we can equivalently say that given all pairwise similarities, the clustering is uniquely decided. Recently, aMinMaxCut algorithm[19] is developed using similarity concepts. It is based on a min-max clustering principle: data should be grouped into clusters such that similarity between different clusters is minimized while the similarities within each clusters are maximized individually. MinMaxCut is extensively analyzed and experimented on two-cluster problems in [2], and is shown to be more effective than other current competitive methods such as the normalized cut [19] and PDDP [18].

Given the dataset and similarity measure (Euclidean distance in K-means and similarity graph weight in Min-MaxCut), the global optimal value of the objective function is a function of K. An important property of these clustering objective functions is the monotonicity. Wecan prove that as K increases K = 2, 3, ・ ・ ・, the Min-MaxCut objective increases monotonically, while the K-means objective decreases monotonically. Based on this the theorem has been stated as:

Theorem 1. : Given the dataset and the similarity metric, as K increases,

1) the optimal value of the K-means objective function decreases monotonically:

$$J^{\text{opt}}_{\text{Kmeans}}(K) > J^{\text{opt}}_{\text{Kmeans}}(K+1) \quad (1)$$

2) the maximun log-likelihood of the Gaussian mixture increases monotonically:

$$\ell^{\text{opt}}(K) < \ell^{\text{opt}}(K+1) \quad (2)$$

3) the optimal value of the MinMax Cut objective function increases monotonically:

$$J^{\text{opt}}_{\text{MMC}}(K) < J^{\text{opt}}_{\text{MMC}}(K+1) \quad (3)$$

4.3. Working:

The proposed work complements the existing literature on the technique of association clustering. The proposed technique in this work considers only the positive associations for clustering. The detail working is explained as follows:

There are two steps in this technique. First step is mining association rules with threshold value of '1.00' for lift. Suitable threshold values for support and confidence are also chosen. In the second step, these association rules along with the support, confidence and lift are taken as input for clustering. Hence, this step gives an algorithm for developing the dendrogram. The proposed algorithm for clustering makes use of 'all strength measures of association rule/s' instead of only 'support of itemset'.

As discussed, there are two phases involved in the proposed clustering technique as mentioned below.

Phase I : Mine association rules from the transaction data with some threshold values of rule support and confidence with lift more than 1.00. The threshold values of rule support and confidence are chosen with low values so that all the items under consideration find place at least in one rule. Amongst these rules, only those rules will mined which has lift value greater than 1 as main target is to mine positive association rule. The input and output in this step are as given below.

Input:

I = {I1, I2,……,Im} //set o items

D = {t1, t2,……tn} //a transaction database with ti as one transaction

Threshold rule support = s

Threshold rule confidence = c & Threshold lift = 1.00

Output:

A = { {r1, s1, c1, l1}, {r2, s2, c2, l2},……,{ rn, sn, cn, ln}}

// set of positive association rules (ri) with rule support (si) , confidence (ci) and lift (li)

Phase-II: Make use of the output in phase-1 for developing the dendrogram. Key steps involved in this phase are described below.

Step 1: Input to this phase is the output of phase-I, i.e., the set of association rules With the values of rule support, confidence and lift, given by set A.

Step 2: Obtain the set of individual items present in at least one of the association rules in A by

I = {I1, I2,……,Im}.

Step 3: Start with tree level s initiated as 1, the itemset similarity is defined as very high value (tending to infinity), and number of clusters (itemsets) is m (the total number of itemsin I). Hence, the set of clusters at level 1, L{1}, contains all 1-item clusters in I.
That is, L{1} = {{I1}, {I2},…….,{Im}}.
Step 4: To generate a set of candidate itemsets for next level (C(s+1)) each pair of itemsets in the previous level are joined.
Step 5: To evaluate itemset similarity, i.e., similarity amongst the items in a cluster, each of the association rules is checked if all the items in the candidate set exist in the rule (either in the antecedent or in the consequent). If all the items exist in a rule and no other item is present in the rule, then sum up rule support, confidence and lift for the rule. Similarly, sums are obtained for all other rules where all the items are present in a rule. Sum of all such sums is taken as the measure of similarity.
Step 6: To generate L(s+1) (i.e., the set of itemsets in level (s+1)), the two itemsets are merged if their similarity is the highest value among all itemsets in c(s+1).
Hence, L{s+1} = { L{s} - La{s} - Lb{s}} U {La {s} U Lb {s} }.
Step 7: The steps 4-6 are iterated with updating the dendrogram (DE) by adding the tuple < s, sim, k, L{s} > into DE.
where s =s+1, sim = sim{La {s} , Lb{s}, k = k-1, L{s} = L{s+1}.
Itereation stops when there is no association rule with all items of any pair of combined clusters/itemsets in a level and this level is the last level of clustering. Hence, all items may not be merged in one cluster as per the proposed algorithm in most of the cases.

5. Evaluation in Proposed Work:

The strategy for using dendrograms to form a hierarchy of cluster has three phases, that are as follows :
1) Select the number of clusters into which the dendrogram will be divided, using a method such as the Calinski-Harabasz metric.
2) Examine the individual clusters for homogeneity by choosing the two executions in each cluster with maximally dissimilar profiles according to the chosen similarity metric, and determining whether these two executions have the same cause. If the selected executions have the same or related causes, it is probable that all of the other failures in the cluster do as well.
3) Choose appropriate candidates for splitting and merging of clusters, according to the properties outlined in the previous section.
(a) Split a cluster if it is found to be non-homogenous, and one or both of the resulting clusters would be a non-singleton or contain a largest homogeneous subtree.
(b) Merge two clusters if they are homogeneous, are siblings, and if their failures have the same cause.
The clusters with the largest amount of internal dissimilarity should be examined first. Internal dissimilarity may be measured using the two maximally dissimilar profiles, or by calculating the average similarity between all individual execution profiles in the cluster. If the clusters with high internal dissimilarity are homogenous, it is reasonable to assume that the others are as well, though it is still useful to examine clusters with more internal similarity. If a cluster resulting from a splitting operation is still an appropriate candidate for splitting, it may be advantageous to split the new cluster as well. Doing so has the effect of splitting the original cluster twice.

5.1. Evaluating Desired Characteristics of clustering :

While conducting many experiments during the implementation work, the desired characteristics of a clustering algorithm have been examines at different places. These characteristics are depends on a particular problem under consideration. The following is a list of characteristics:
1) Scalability : Clustering techniques for large sets of data must be scalable, both in terms of speed and space. It is not unusual for a database to contain millions of records, and thus, any clustering algorithm used should have linear or near linear time complexity to handle such large data sets. Furthermore, clustering techniques for databases cannot assume that all the data will fit in main memory or that data elements can be randomly accessed. These algorithms are, likewise, infeasible for large data sets. Accessing data points sequentially and not being dependent on having all the data in main memory at once are important characteristics for scalability. This property have been used with the data set used for the proposed work.
2) Effective means of evaluating the validity of clusters that are produced : It is common for clustering algorithms to produce clusters that are not “good” clusters when evaluated later. To check this whether the generated clusters are good or not, Validation for those clusters have been tested regularly.
3) The ability to find clusters in subspaces of the original space : Clusters often occupy a subspace of the full data space. Hence, the popularity of dimensionality reduction techniques is there. Many algorithms have difficulty in finding items to be keep in particular cluster in particular space, for example, a 5 dimensional cluster in a 10 dimensional space. The proposed work supports to the fact that all the cluster must be find in subspaces of the original space.
4) Ability to function in an incremental manner : In certain cases, e.g., data warehouses, the underlying data used for

the original clustering which can change over time. If the clustering algorithm can incrementally handle the addition of new data or the deletion of old data, then this is usually much more efficient than re-running the algorithm on the new data set. The proposed work have an ability to handle the new data

6. General Result Analysis:

The work regarding the topic have been proposed by using the online shopping scenario. And the expected results according to the proposed work are as follows:

Fig 5. Searching a product (Laptop here)

Figure 6. Outcomes for the searching product

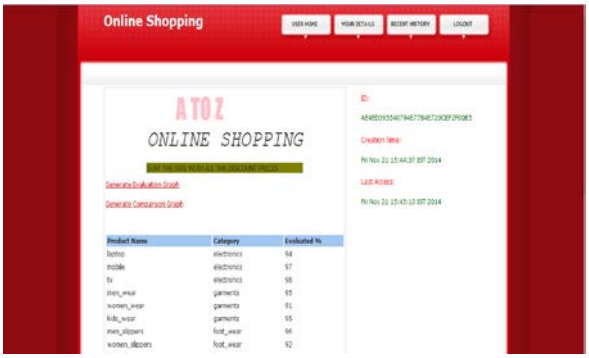

Fig 7 : Performance Evaluation

Fig 8: General Result Analysis:Evaluation Graph

Figure 7 shows the graphical representation of general evaluation of the proposed work. In this diagram, the X-axis of a graph belongs to the no of items on which clustering have to be perform while Y-axis belongs to the efficiency of the proposed work in terms of some threshold values. The red line in graph represents the complete efficiency of a proposed work with respect to the items present in the data set. As shown in graph, for the 30 items in a data set the complete efficiency of a product is almost more than 90%. I.e. we can get at least 90% exact cluster by using the proposed work to achieve the Purpose of Market-Basket Analysis.

7. Conclusion:

Market-basket analysis is an integral part of today's business world. Customer satisfaction is at the center point in Market-Basket Analysis and to achieve this it is necessary to find the main interest of customer in a particular product. The agglomerative hierarchical clustering used in the proposed work creates the clusters by considering each item or product as a individual cluster from its starting with which the retailars can easily identify which products are frequently purchased by the customer from a huge dataset. The clustering of retail items with this technique gives more efficient and reliable result than other techniques as here clustering of item start from a individual element The placement of product in retail with the help of such clustering will not only effective and impressive but also helpful to achieve the goal of market-Basket Analysis. The technique presented is useful in the area of failure classification in retail stores or in supermarket,since the current failure classification methods do not have a definitive way to determine the number of clusters into which a set of program executions should be divided.

References

[1] Ashok Kumar D and Loraine Charlet Annie M.C , "Market Basket Analysis for a Supermarket Based on Frequent

Itemset Minng" , IJCSI, Vol. 9 . Issue 5, No.3,September 2012.
[2] Aastha Joshi, Rajneet Kaur , "Comparative study of Clustering Techniques in Data mining" , IJARCSSE, 2012.
[3] Berry, M.J.A., Linoff, G.S.: Data Mining Techniques: for Marketing, Sales and Customer Relationship Management (second edition), Hungry Minds Inc., 2004.
[4] "Cluster analysis" in http://en.wikipedia.org/ wiki/Cluster_Analysis
[5] Chen, Y.-L., Tang, K., Shen, R.-J., Hu, Y.-H.: "Market basket analysis in a multiple store environment, Decision Support Systems", 2004.
[6] Erik Buchmann, Leonardo Weiss Ferreira Chaves and Klemens Bohm, "Finding misplaced items in retail by clustering RFID data" , EDBT 2010,March 22-26,2010, Lausanne, Switzerland.
[7] Er. Arpit Gupta 1 ,Er.Ankit Gupta 2,Er. Amit Mishra 3, " Research Paper On Cluster Techniques Of Data Variations", International Journal of Advance Technology & Engineering Research (IJATER).
[8] Fionn Murtagh and Pedro Contreras, "Methods of Hierarchical Clustering",arXiv:1105.021v1 [CS:IR], 30th April 2011.
[9] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.
[10] Kalyani M Raval, " Data Mining Techniques", IJARCSE, Volume 2, Issue 10, October 2012
[11] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining",International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012
[12] MR ILANGO, Dr V MOHAN, "A Survey of Grid Based Clustering Algorithms", International Journal of Engineering Science and Technology, pp.3441-3446, 2010.
[13] "Measuring Association d12 Between Clusters 1 and 2" in http://www.stat.psu.edu/online/courses/stat505/18_cluster/0 5_cluster_between.html
[14] Neha Soni, Amit Ganatra, ― "Categorization of Several Clustering Algorithms from Different Perspective: A Review‖", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2,no.8,pp.63-68,Aug. 2012.
[15] P. Berkhin. (2001) ― "Survey of Clustering Data Mining Techniques" [Online]. Available: http://www.accure.com/products/rp_cluster_review.pdf
[16] Rui Xu, Donald C. Wunsch II, ― "Survey of Clustering Algorithms‖", IEEE Transactions on neural Networks, vol. 16, pp. 645-678, May 2005.
[17] Rahmat Widia Sembiring, JAsni Mohamad Zain, Abdullah Embong, "A Comparative Agglomeartive Hierarchical Clustering Method to Cluster Implemented Course" , journal of Computing, volume 2,Issue 12 , December 2010.
[18] S.M. Savaresi and D. Boley , "On the performance of bisecting K-means and PDDP" , . Proc. SIAM Data Mining Conf, 2001.
[19] W. Bishop, " Documenting the value of merchandising. Technical report, National Association for Retail Merchandising Service", 2000.

Rujata Saraf pursuing M.E. Degree in Computer Science and Engineering from North Maharashtra University,Jalgaon and received the B.E., degree from Mumbai Univ. in 2011. After degree worked as an assistant professor (from January 2012) in the Dept. of Information Technology, the G.H.Raisoni Institute of Egineering and Management, Jalgaon. , and now working as a Lecturer (from January 2015) at ST. Francis Institute of Technology, Borivali.in Mumbai University

Sonal P. Patil received the Diploma and B.E. . degrees, from MSBTE and NMU Univ. in 2005 and 2008, respectively. She received the M.Tech degree from Bhopal Univ. in 2013. She have been Appreciated as Best Outgoing Student during Diploma & Degree College. Also Appreciated by North Maharashtra University, Jalgaon for active involvement and management of YUVARANG 2009. She worked as an assistant professor (from 2009 to 2014) in the Dept. of Information Technology, in G.H.Raisoni Institute of Engineering and Management,Jalgaon and now (from July 2014) working as a HOD of Information Technology department in Same college. She has successfully published book having a title "Computer Organization" for Second Year CSE & IT Students of Engineering in January 2014. Her research interest includes Data Mining and its Application. She is a member of CSI, CAP process of North Maharashtra University from Dec 2010 Exam, Syllabus setting committee member for the Computer & I.T. department subjects and Certified & lifetime member of ISTE.