Implementation of Sales Forecasting Prediction System using Hybrid Classification Algorithm

Mr. Anil Hingmire[1], Bhavik Gore[2], Swapnil Badgujar[3]

[1, 2, 3] Vidyavardhini's College of Engineering and Technology, Vasai

Abstract- *Data is very important for every organization and business. Data that was measured in gigabytes until recently, is now being measured in terabytes, and will soon approach the peta byte range. In order to achieve our goals, we need to fully exploit this data by extracting all the useful information from it. Sale data classification has different market trends. Some clusters or segments of sale may be growing, while others are declining. The information produced is very useful for business decision making. Decision can take place on the basis classification of Dead-Stock (DS), Slow- Moving(SM) and Fast-Moving (FM) of the sale. Segment by- segment sales forecasting can produce very useful information.*

I. INTRODUCTION

Data mining techniques like clustering and associations can be used to find meaningful patterns for future predictions. Clustering is used to generate groups of related patterns, while association provides a way to get generalized rules of dependent variables. Patterns from a huge stock data on the basis of these rules can be obtained. This is a useful approach to distinguish the selling frequency of items on the basis of the known attributes, e.g. we can examine that a "black coat of imperial company in winter season has high ratio of sale", here we have basic property related to this example, i.e. colour, type, company, season, and location. Similarly we can predict that certain products of certain properties have what type of sale trends in different locations. Thus on the basis of this scenario we can predict the reason of dead-stock, slow moving and fast moving items. Data mining techniques are best suited for the analysis of such type of classification, useful patterns extraction and predictions.

II. RELATED WORK

In recent years, it has been recognized that the partitioned clustering technique is well suited for clustering a large dataset due to their relatively low computational requirements. The time complexity of the partitioning technique is almost linear, which makes it widely used. The best known partitioning clustering algorithm is the K-means algorithm and its variants. This algorithm is simple, straightforward and is based on the firm foundation of analysis of variances. In addition to the K-means algorithm, several algorithms, such as Particle Swarm Optimization (PSO) is another computational intelligence method that has already been applied to image clustering and other low dimensional datasets.

Data mining techniques and clustering techniques have a greater scope in creating software which will not only enable companies and organizations in decision making, but also help in following areas:

- Sales forecasting.
- Identifying new and emerging trends.
- Maintaining accounts and inventory.
- Prospecting the potential customers by reports generated.

III. CLUSTER ALGORITHMS

1) Hierarchical clustering methods

In Hierarchical clustering, we group data items into a tree consisting of a definite number of clusters. Hierarchical clustering methods can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up (merging) or top-down (splitting) fashion. Hierarchical methods examines all the clusters present so far at each stage of merging, the clustering methods we examined work incrementally, instance by instance. At any stage the clustering forms a tree with instances at the leaves and a root node that represents the entire dataset. At the beginning of clustering, the tree consists of only the root node. Instances are added one by one, and the tree is updated appropriately at each stage. The key to deciding how and where to update the tree is a quantity called category utility that measures the overall quality of a partition of instances into clusters. The important algorithms in hierarchical clustering are -

2) Agglomerative Clustering (Bottom-up):

This method builds the hierarchy from the individual elements by progressively merging clusters. In our example, we have six elements {a} {b} {c} {d} {e} and {f}. 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.

Optionally, one can also construct a distance matrix at this stage, where the number in the i-th row j-th column is the distance between the i-th and j-th elements. Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. This is a common way to implement this type of clustering, and has the benefit of caching distances between clusters. A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage

3) Partition Clustering Methods

In Partition clustering clusters are created for optimizing a predetermined criterion. The criteria adopted for our study is dissimilarity function which is essentially based on object distances within and across clusters. The grouping is done as follows – objects within a same cluster are grouped as ―similar‖ and objects belonging to different clusters are grouped as ―dissimilar‖. The object attributes play a crucial role in this grouping of clusters. Given D, a data set of n objects, and k, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions (k ≤ n), where each partition represents a cluster

1. K-MEANS:

The inputs of this algorithm are the number of clusters to be formed i.e., k and the data to be clustered. The algorithm starts with an initial set of cluster centers, chosen at random or according to some heuristic procedure. In each iteration, each instance is assigned to its nearest cluster centre according to the Euclidean distance between the two.

Then the cluster centres are re-calculated. The centre of each cluster is calculated as the mean of all the instances belonging to that cluster. This algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster c j centre, is an indicator of the distance of the n data points from their respective cluster centres.

2. K-Medoid

The k-medoids algorithm is related to the k-means algorithm and the medoid shift algorithm. Both the k-means and k-medoids algorithms are partitional (breaking the dataset up into groups) and both attempt to minimize the distance between points labeled to be in a cluster and a point designated as the center of that cluster. In contrast to the k-means algorithm, k-medoids chooses datapoints as centers.

Algorithm

-Select initial medoids
-Iterate while the cost decreases:

- In each cluster, make the point that minimizes the sum of distances within the cluster the medoid
- Reassign each point to the cluster defined by the closest medoid determined in the previous step.

-Repeat
-End

Most Frequent Pattern [MFP] ALGORITHM:

Association rule mining is one of the most important and well defines technique for extract correlations, frequent patterns, associations or causal structures among sets of items in the transaction databases or other repositories. Association rules are widely used in various areas such as risk management, telecomm, market analysis, inventory control, and stock data. Apriori algorithm for strong association among the patterns is highly recommended. In this work we proposed a new algorithm MFP that is more efficiently generates frequent patterns and strong association between them. For this purpose a property matrix containing counted values of corresponding properties of each product has been used as shown below.

Let we have set X of N items in a Dataset having set Y of attributes. This algorithm counts maximum of each attribute values for each item in the dataset.

```
Most Frequent Pattern(MFP)
Input: Datasets (DS)
Output: Matrix
Frequent Property Pattern (FPP):
FPP (DS)
Begin
for each item Xi in DS
```

```
        a. for each attribute
                i. count occurrences for Xi
                    C=Count (Xi)
             ii. Find attribute name of C
                    Mi=Attribute (Ci)
                   next [End of inner loop]
   b. Find Most Frequent Pattern
                 MFP=Combine(Mi)
                      next [End of outer loop]
End
```

System Architecture:

This is a two phased model. First we generate clusters using K-Mean algorithm, and then MFP is designed for counting frequencies of items under their specified attributes. The block diagram of the whole process is given in figure. In phase-1 the first step is to collect sample data from real store inventory data. We have process the data to remove the noise first, so the incomplete, missing and irrelevant data are removed and formatted according to the required format.

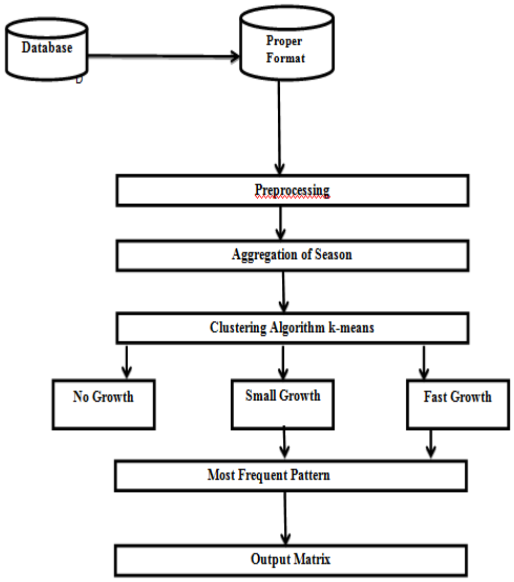


Figure 1. Architecture.

This Architecture is divided into 2 distinct phases described as follows:

- The data is collected from the database and pre-processing is done on data. After the pre-processing is done k-mean clustering algorithm is applied to the data. Clusters are formed on bases of dead stock, slow moving stock and fast moving stock.
- Later in 2nd phase we apply the most frequent pattern (MFP) for analyzing the patterns.

IV. EXPERIMENTATION AND RESULTS

First we have transformed inventory data in required format by removing noise and any other inconsistencies which is then used for clustering. Clusters are formed from data on basis of the quantity sold. As Experimenting all the algorithm we found k-means with MFP algorithm is best suited.

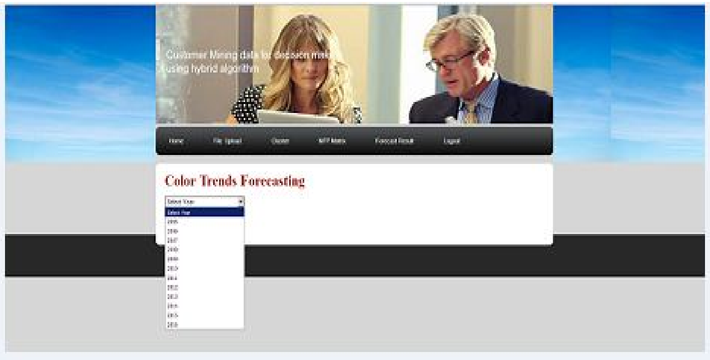
Clustering done by means of K-Means Algorithm. The results obtained from both phases are as follows:

Phase One:

In first phase, data is divided into three clusters by K-Means Clustering Algorithm. The Clusters obtained are Dead Stock, Slow Moving and Fast Moving. The Dead Stock cluster contains records of those products with small selling quantity. Slow Moving stock cluster contains records of products with medium sales. Fast Moving Stock Cluster contains records of products with large selling quantity.

Phase Two:

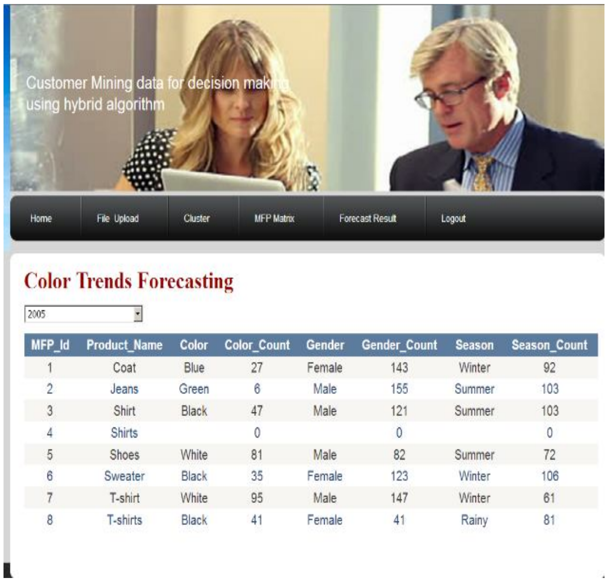
In this phase MFP algorithm has been used to generate a property matrix, containing counted values of corresponding properties of each product. This procedure receives data sets from clusters. The first loop scans all the records of the data set. The inner loop counts occurrences of the attribute for a given item and placed in the MFP matrix. Finally maximum occurrences of attributes values within a row give a single pattern. On the basis of these patterns, we can say that why a certain product falls in particular cluster.



Color Trends Forecasting

Select Year

Figure 2.



Color Trends Forecasting

2005

MFP_Id	Product_Name	Color	Color_Count	Gender	Gender_Count	Season	Season_Count
1	Coat	Blue	27	Female	143	Winter	92
2	Jeans	Green	6	Male	155	Summer	103
3	Shirt	Black	47	Male	121	Summer	103
4	Shirts		0		0		0
5	Shoes	White	81	Male	82	Summer	72
6	Sweater	Black	35	Female	123	Winter	106
7	T-shirt	White	95	Male	147	Winter	61
8	T-shirts	Black	41	Female	41	Rainy	81

Figure 3. Cluster Result

V. CONCLUSION

After analyzing and predicting we conclude that hybrid classification algorithm, which is k-means and MFP is good for sales forecasting. In this system, the clustering association mining approach is used to classify stock data and find associated patterns of sales. From the experimental results it is clear that the approach is very efficient for mining patterns of huge stock data and predicting the factors affecting the sale of products.

REFERENCES

[1] Abubakar, Felix "Customer satisfaction with supermarket retail shopping", 2002.

[2] http://www.roselladb.com/sales-trendforecast.htm, visited January, 2010.

[3] M. Braglia, A. Grassi, R. Montanari "Multiattribute classification method for spare parts inventory management" 2004.

[4] Association Analysis of Customer Services from the Enterprise Customer Management System- ICDM-2006.

[5] Terry Harris,"Optimization creates lean green supply chains", 2008.

[6] Matt Hartely "Using Data Mining to predict inventory levels" IEEE, 2005

[7] Jiawan Han, Micheline Kamber "Data Mining Concepts and Techniques" 2nd edition 2004

[8] L. Frans, Wei, Paul, "Towards an agent based framework for online after sales services" 2006.

[9] Gebouw D, Diepenbeek, Belgium "Building an Association Rules Framework to Improve Product Assortment Decisions" B-3590, 2004.

[10] Brijs, Bart, Gilbert, Koen, Geert "A Data Mining Framework for Optimal Product Selection in Retail Supermarket Data: The Generalized PROFSET Model" 2000.

[11] R. C. Wong, A. W. Fu, K. Wang"Data Mining for Inventory Item Selection with Cross-Selling Considerations" 2005.

[12] L. Cao, C. Luo, J. Ni, DanLuo, C. Zhang "Stock Data Mining through Fuzzy Genetic Algorithm" Proceeding of JCIS s , 2008.

[13] P.Thomas, Macredie "Knowledge Discovery and Data Mining" 1999 .

[14] Artigan, J. A. Clustering Algorithms. Ohn Wiley and Sons, Inc., New York, NY. 1975.

[15] Kennedy J., Eberhart R. C. and Shi Y., 2001. Swarm Intelligence, Morgan Kaufmann, NewYork.

[16] Merwe V. D. and Engelbrecht, A. P., 2003. Data clustering using particle swarm optimization. Proceedings of IEEE Congress on Evolutionary Computation 2003(CEC 2003), Canbella, Australia.

[17] Omran, M., Salman, A. and Engelbrecht, A. P., 2002. Image classification using particle swarm optimization. 2002.

[18] http://en.wikipedia.org/wiki/cluster_analys is, visited 2009.

[19] Jo Ting, "Mining of stock data: inter- and inter-stock patternassociative classification" proceddings of 2006 international conference on data mining Las Vegas,USA, June 2006.