International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 6 (2015) pp. 15493-15501

http://www.ripublication.com

Comparative Analysis of Clustering In Super Market Data Set

A.Bamini[1], Dr. S. Franklin John[2] and Dr. Ranjit JebaThangaiah P.[3]
[1]Research Scholar, Department of Computer Applications,
Karunya University, Coimbatore, Tamil Nadu-India
[2]Professor and Principal,
Nehru College of Management, Coimbatore, Tamil Nadu-India
[3]Assistant Professor (SG), Department of Computer Applications,
Karunya University, Coimbatore, Tamil Nadu-India

Abstract

Now a days the growth of human population has been a source for increase in food production. Foods produced of high quality are often obtained in super markets. As an initiative super markets helps the retailers of food production community to trade or import bulk purchases through the frequent sales of an item in a supermarket. Market basket analysis is a technique used to understand product purchase patterns at the customer level in the super market. This paper focuses on implementing the comparative analysis of market basket through clustering techniques with combined logics of neural networks, to identify the frequent purchases.

Introduction:

Data Mining is the basic process employed to analyze patterns in data and extract information [3]. Data mining is actually the core of a bigger process, known as knowledgeable discovery in databases (KDD). KDD is the process of taking low-level data and turning it into another form that is more useful, such as a summarization or a model. [2]

According to Insightful Miner in his User Guide, "Data mining is the application of statistics in the form of exploratory data analysis and predictive models to reveal patterns and trends in very large data sets. According to the great Gartner Group, “Data mining is the process of discovering meaningful new correlation, pattern and trend by shifting through large amounts of data stored in repository, using pattern recognition technology as well as statistical and mathematical tools.”

According to Hand (1998) “data mining is the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or

value to the database owners." Data Mining is being referred to as a statistical process of analyzing data stored in a warehouse (Decker, 1998).

As we know the importance of data mining in marketing area of banking industry, similar applications of data mining can also be seen in marketing area of retail industry. In this "market basket analysis" is a marketing method used by many retailers to determine the optimal locations for the establishment of marts or shopping malls and promote their goods. The basic information behind this technique is to identify what type of goods customers most often purchase together so that the promotion campaigns and advertisements are planned in such a way that maximum sales can be generated, thus increase the revenue of the supermarket.

Data mining tools proved to be the best way of discovering patterns in the data for the successful customer relationship, since, nowadays business is relations based the data is extremely important because retailers reach customers best through the data. Retailers can study about customers past purchasing histories and know with what kinds of promotions and incentives to target customers. By collecting customer and transactional data, data mining tools helps the retailers to identify best customers and offer exclusive extras to them. Data mining applications like Target achievement, which identifies the prospective customers by observing the past purchase behavior, can reduce the expenditure in campaigning and promotional activities; Attrition prediction and churn analysis used to prevent loss of customers and avoids adding churn-prone customers, this data mining application uses the neural nets and time series analysis, the ideal benefits includes retention of customers and more effective promotions.

What is cluster analysis?

Cluster analysis (CA) is an exploratory data analysis tool for organizing observed data (e.g. people, things, events, brands, companies) into meaningful taxonomies, groups, or clusters, based on combinations of IV's, which maximizes the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown.

This paper focus on the analysis of hierarchical and k-means Techniques using data sets collected from the nearby Super market. Two months transactions have been collected and the Groceries transactions were segregated and the following experimental study has been done.

K-means clustering with SOM method

K-means clustering-method is integrated with SOM to find the cluster count (ie. k). Once the analysis have been made using kohonen map the clusters are formed using k-means.

SOM:

The Kohonen map is considered to be the original or traditional SOM and is composed of a two-dimensional grid of units. The grid can be composed of square or hexagonal units that are initialized using random numbers or Eigen values from the data.

During a training process, the SOM is exposed to vectors from the training data one at a time and the SOM is adjusted to have more similar values. The training process begins by activating the single unit in the SOM that has the smallest Euclidean distance from an input vector. The active unit then creates a neighborhood by selecting all the adjacent units up to a certain distance. All the units in the neighborhood are then adjusted so that all their Euclidean distances from the input vector are smaller [1].

The creation of a Kohonen map requires the user to specify the topology, neighborhood type, x-dimension, y-dimension, and the dimensionality of the SOM. The topology is the connections between the grid units visible at the adjacent edge and is usually rectangular or hexagonal. The neighborhood type is the way in which connections are made between units and is usually Gaussian. The x-dimension and the y dimension are the number of units in the x-direction and y-direction. The dimensionality of the SOM is the number of variables the SOM is designed for, which depends exclusively on the number of variables in the data.

Way to find K-means cluster:

a) Place K objects points into the space that are to be clustered object points always represent initial group centroids.
b) Assign each object point to the group that has the closest centroid.
c) When all object points have been assigned, re-calculate the positions of the K centroids.
d) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the object points into groups from which the metric to be minimized can be calculated.
e) It is an indicator of the distance of the n data points from their respective cluster centers.

K-Means and SOM:

- Step 1: Transaction data from a super market has been collected for nearly 2 months.
- Step 2: After Data cleaning has been performed manually using Office package, around 475 transactions has been identified. The essential dataset S_i has been segregated into Groceries, Cosmetics, Stationaries, Snacks and Household.
- Step 3: Normalization has been carried out by converting categorical data to numeric data.
- Step 4: Kohonen's SOM algorithm has been used to find the number of clusters.
- Step 5: The outcome of SOM is passed as input to k in K-Means algorithm along with the transaction dataset.
- Step 6: Similar transactions are grouped together using k-means. The working of K-Means has been evaluated using Euclidean distance measures.

$$p_{ij=(} \Sigma \, |x_{ik} - x_{jk}|^2)^{1/2}$$

- Step 7: The k-Means technique minimizes the intra cluster variance, or squared error:
  $E = \Sigma \Sigma \|p - m_i\|^2$
  Where there are k clusters Ci, i = 1,2, ...,k and mi is the centroid of all the points p ε Si.
- Step 8: Negligible results are considered to be Outliers and are removed.

Hierarchical Clustering

This is a statistical method for finding relatively homogeneous clusters of cases based on measured characteristics. It starts with each case as a separate cluster, and then combines the clusters sequentially, reducing the number of clusters at each step until only one cluster is left. The clustering method uses the dissimilarities or distances between objects when forming the clusters. A hierarchical tree diagram, can be produced to show the linkage points. The clusters are linked at increasing levels of dissimilarity. Clustering can be summarized as follows:

- The distance is calculated between all initial clusters. In most analyses, initial clusters will be made up of individual cases.
- Then the two most similar clusters are fused and distances recalculated.
- Step 2 is repeated until all cases are eventually in one cluster

Experimental Results

The analysis has been carried using SPSS tool. First analysis has been carried using artificial data set Then the grocery real data set is given as input to the below techniques and the analysis table and graphical display has been produced.

K-Means and SOM (Euclidean distance):

Table 1: Distance between final clusters

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1		1.203	1.312	1.155	2.344	1.302	.999	1.774	1.440	1.033	.860	1.360	1.399	.955	1.037
2	1.203		1.152	1.117	1.976	1.279	.920	1.231	1.090	.841	1.070	1.158	1.212	.920	1.118
3	1.312	1.152		1.233	2.341	1.309	.998	1.294	1.097	1.028	1.062	1.451	.988	1.026	1.067
4	1.155	1.117	1.233		2.166	1.298	1.100	1.333	.993	1.321	1.074	1.271	1.128	1.050	1.255
5	2.344	1.976	2.341	2.166		1.886	2.107	1.978	2.114	2.198	2.352	2.073	2.276	2.276	2.218
6	1.302	1.279	1.309	1.298	1.886		1.204	1.481	1.369	1.326	1.340	1.471	1.326	1.298	1.221
7	.999	.920	.998	1.100	2.107	1.204		1.310	1.119	.911	.947	1.196	1.138	.956	.993
8	1.774	1.231	1.294	1.333	1.978	1.481	1.310		1.048	1.376	1.496	1.268	1.327	1.352	1.244
9	1.440	1.090	1.097	.993	2.114	1.369	1.119	1.048		1.114	1.194	1.019	1.199	1.054	1.140
10	1.033	.841	1.028	1.321	2.198	1.326	.911	1.376	1.114		.882	1.180	1.237	.730	.870
11	.860	1.070	1.062	1.074	2.352	1.340	.947	1.496	1.194	.882		1.154	1.063	.737	.889
12	1.360	1.158	1.451	1.271	2.073	1.471	1.196	1.268	1.019	1.180	1.154		1.518	1.100	1.121
13	1.399	1.212	.988	1.128	2.276	1.326	1.138	1.327	1.199	1.237	1.063	1.518		1.095	1.190
14	.955	.920	1.026	1.050	2.276	1.298	.956	1.352	1.054	.730	.737	1.100	1.095		.922
15	1.037	1.118	1.067	1.255	2.218	1.221	.993	1.244	1.140	.870	.889	1.121	1.190	.922	

Resource Utilization:

Table 2: Period of Execution

Processor Time	0:00:00.093
Elapsed Time	0:00:00.131
Workspace Required	29768 bytes

Table 3: Number of Cases in each Cluster

Cluster	1	104.000
	2	33.000
	3	33.000
	4	25.000
	5	2.000
	6	16.000
	7	35.000
	8	27.000
	9	19.000
	10	40.000
	11	32.000
	12	12.000
	13	30.000
	14	39.000
	15	27.000
	Valid	474.000
	Missing	.000

Proximities:
Rice UradDal ToorDal Salt Tamarind Sugar KalaChanna Ginger SambarPowder RasamPowder MasalaPowder DryChilli ChilliPowder Turmeric Powder FriedGram Anise Ghee CuminSeeds Peanut Pepper WheatFlour MaidaFlour GramFlour Pappad RajmaDal YellowGram GreenGramDal MustardSeeds PoppySeeds Cardamom Cinnamon Cloves Mace Oil Ragi Millet Fenugreek Raisin Cashew Semolina Jaggery Sago Asafoetidia Vermicelli MealMaker Cowpea InstantMix Garlic Coriander

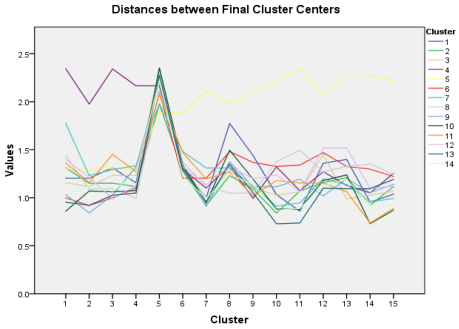


Figure 1: cluster distance

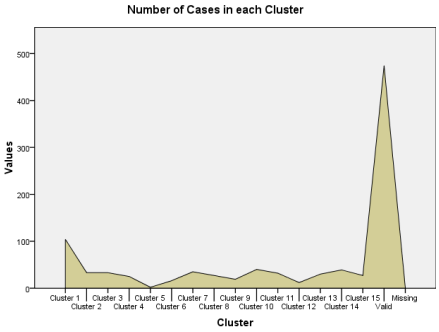


Figure 2: Cluster Cases

Hierarchical cluster analysis (Wards Method)

The inter-relationship of grocery products are shown below in the clusters

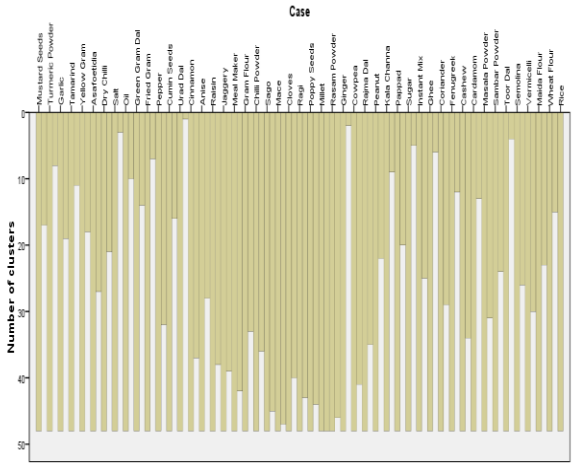
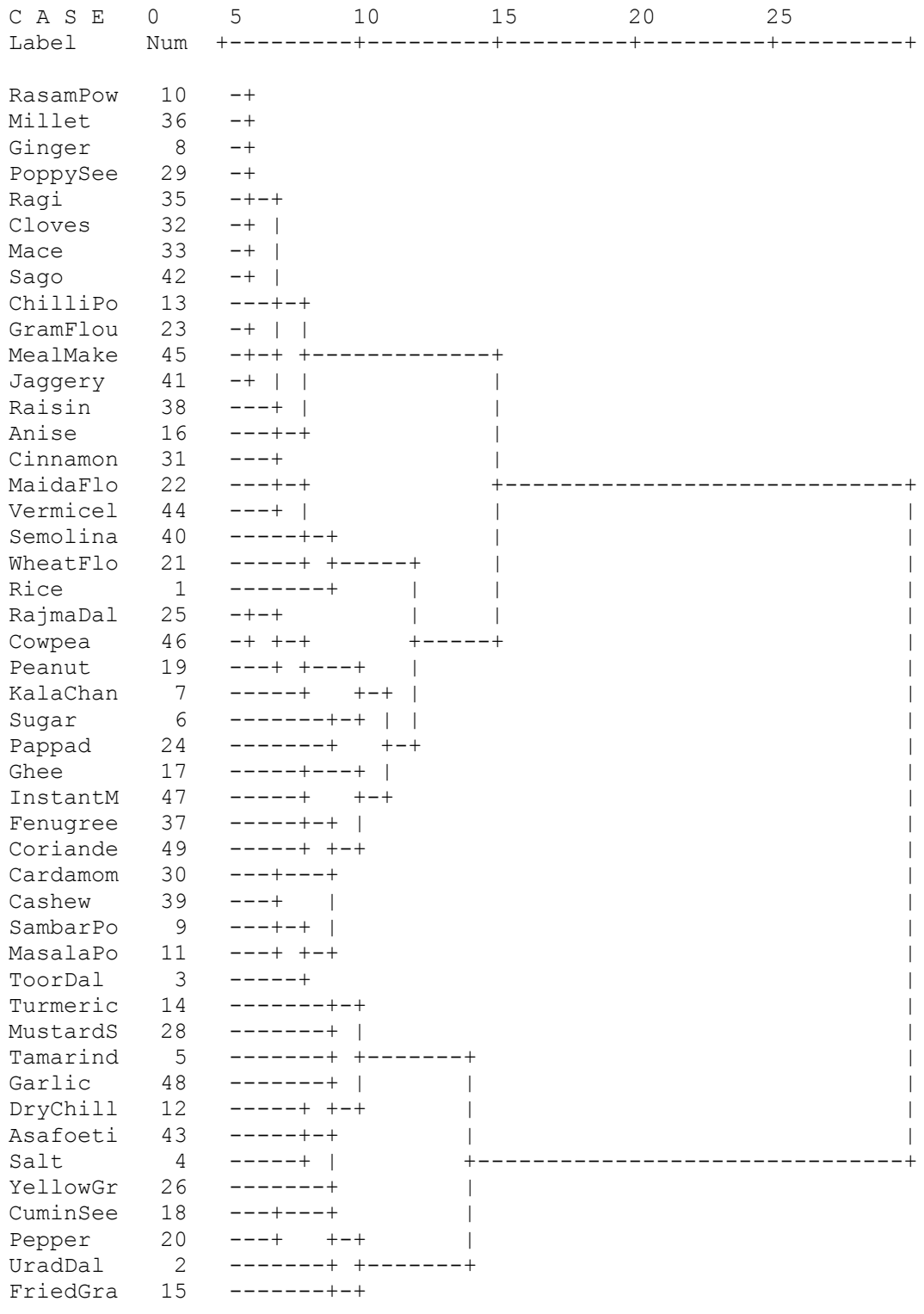


Figure 3: Hierarchical Cluster

Dendrogram – Fusion of clusters using Ward Method (analysis of variance approach to evaluate the distances between clusters)



```
GreenGra    27   -------+ |
Oil         34   ---------+
```

Conclusion

This experimental evaluation scheme was created to provide the comparison of SOM – KMeans and hierarchical clustering. Cluster analysis determines how many 'natural' groups there are in the sample. It also allows you to determine who in your sample belongs to which group. The main aim is to minimize variability within clusters and maximize variability between clusters. From these experiments on the datasets, it is observed that the Self Organizing Map based K-means algorithm has provided correct results in terms of finding frequent purchasing patterns than hierarchical clustering, because hierarchical clustering leads to the fusion of dissimilar transactions. This approach can increase the marketing and improve sales in the super market.

References

[1] Ackoff, R. F. (1989). From Data to Wisdom. Journal of Applied Systems Analysis, 16, 3-9. Alahakoon, D., Halgamuge, S. K., & Srinivasan, B. (2000). Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery. IEEE Transactions on Neural Networks, 11(3), 601-613.

[2] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37-54.

[3] Trybula, W. J. (1997). Data mining and knowledge discovery. In M. E. Williams (Ed.) Annual Review of Information Science and Technology, 32, 196-229. Medford, NJ: Information Today.