Comparative Performance Of Using PCA With K-Means And Fuzzy C Means Clustering For Customer Segmentation

Fahmida Afrin, Md. Al-Amin, Mehnaz Tabassum

Abstract: Data mining is the process of analyzing data and discovering useful information. Sometimes it is called knowledge Discovery. Clustering refers to groups whereas data are grouped in such a way that the data in one cluster are similar, data in different clusters are dissimilar. Many data mining technologies are developed for customer segmentation. PCA is working as a preprocessor of Fuzzy C means and K- means for reducing the high dimensional and noisy data. There are many clustering method apply on customer segmentation. In this paper the performance of Fuzzy C means and K-means after implementing Principal Component Analysis is analyzed. We analyze the performance on a standard dataset for these algorithms. The results indicate that PCA based fuzzy clustering produces better results than PCA based K-means, and is a more stable method for customer segmentation.

Index Terms: Data Mining, Clustering, K-means, Principal component analysis, Fuzzy C means, Customer segmentation, Crisp Set

1 INTRODUCTION

Data clustering is an unsupervised data analysis and data mining technique. Hundreds of clustering algorithms have been developed by researchers. The development of clustering methods is very interdisciplinary. Contributions have been made, for example, by psychologist, biologists, statisticians, social scientists, and engineers. There exist huge amount of clustering applications from many different fields, such as, biological sciences, life sciences, medical sciences, behavioral and social sciences, earth sciences, engineering and information, policy and decision sciences to mention just a few. Customer Segmentation is one of the important fields. Customer segmentation, also referred to as market segmentation, is the process of finding homogenous sub-groups within a heterogeneous aggregate market. Typically this approach is used in direct marketing to target and focus on increasingly well-defined and profitable market segments. The process of segmentation begins with observing customer actions and continues with learning about the demographic and psychographic characteristics of these customers [1]. Customers are the most important asset of an organization. There cannot be any business prospects without satisfied customers who remain loyal and develop their relationship with the organization. That is why an organization should plan and employ a clear strategy for treating customers [2]. Customer segmentation is a term used to describe the process of dividing customers into homogeneous groups on the basis of shared or common attributes. Marketers can decide which marketing strategies to take for each segment and then allocate scarce resources to segments. To cluster massive amount of data, there are lot of technique which reduce the dimension of data such as partition based clustering algorithms. For segmentation K-Means and Fuzzy C-Means are analyzed in this research work. FCM is soft clustering algorithms that retain more information from the original data than those of crisp or hard. PCA is employed as preprocess K-Means and FCM for relieving the curse of high dimensional. These algorithms are implemented by means of practical approach to analyze its performance, based on their computational time [7]. The wholesale customer's data is the source data for this analysis. The computational complexity (execution time) of each algorithm is analyzed and the results are compared with one another.

- *Fahmida Afrin has completed her B.Sc in CSE at Jagannath University, Dhaka, Bangladesh. E-mail: afrincsejnu@gmail.com*
- *Md. Al-Amin has completed his B.Sc in CSE at Jagannath University, Dhaka, Bangladesh.PH-+8801924836158. E-mail: alamin2293@yahoo.com*
- *Mehnaz Tabassum is currently doing job as an Assistant Professor in CSE at Jagannath University, Dhaka, Bangladesh, PH-+8801913497661. E-mail: mtabassum2013@gmail.com*

2 LITERATURE REVIEW

The paper Gayathri .A, Mohanavalli S[6], is intended to implement soft clustering and hard clustering to enhance CRM (Customer Relationship Management). This paper focuses on applying apt clustering algorithm to identify the soft partitions of the customers namely the fuzzy c means (FCM) clustering algorithm to determine the churn ratio accurately. Limei Zhang [3]. This paper wants to promote the importance of data mining within the customer relationship management. Tajunisha and Saravanan [5].In this paper, analyzed the performance of Principal Component Analysis (PCA) for dimension reduction and to find the initial centroid for k-means. Next used heuristics approach to reduce the number of distance calculation to assign the data point to cluster. D.Napoleon, S.Pavalakodi [4].K-means clustering algorithm often does not work well for high dimension, hence, to improve the efficiency, apply PCA on original data set and obtain a reduced dataset containing possibly uncorrelated variables. In this paper principal component analysis and linear transformation is used for dimensionality reduction and initial centroid is computed, then it is applied to K-Means clustering algorithm. Ananthi Sheshasayee and P. Sharmila [7], as the number of records increases the time execution both the technique gets increased but the fuzzy c means performance is found to be better than the k means algorithm. Tejwant & Manish [2] compare of Fuzzy C Means with Respect to other Clustering Algorithm and found that fuzzy requires take more computation time.

3 METHODOLOGY

3.1 K-Means Clustering

K-means clustering is the crisp clustering technique which attempts to cluster data by grouping related attributes in uniquely defined clusters. Each data point in the dataset is assigned to only one cluster. In partitioning the data, only the centers of the clusters are moved and condition of all the data points are fixed. Clustering is an iterative process of finding better and better cluster centers. Distance metric measures calculate how far away a point is from a cluster center [6]. The K-means algorithm may be described as follows:

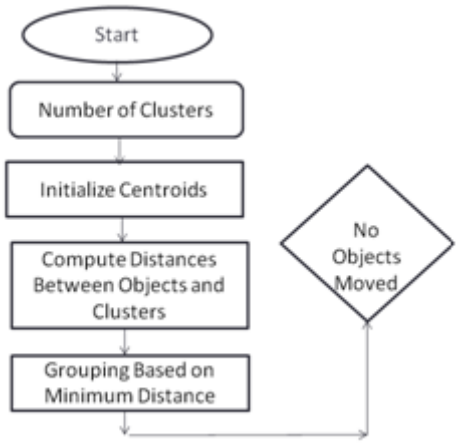

Figure 3.1: *Flow Chart of K-means Algorithm*

3.2 PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) includes a mathematical procedure that maps a number of correlated variables into a smaller set of uncorrelated variables, called the principal components. The first principal component represents as much of the variability in the data as possible. The succeeding components describe the remaining variability. The steps involved in PCA are:

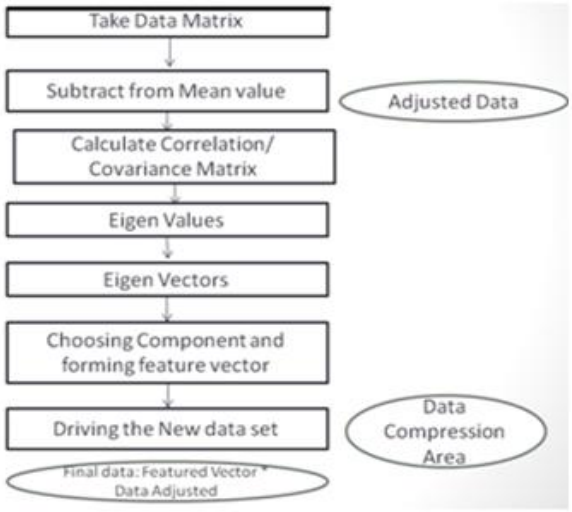

Figure 3.2: *Flow Chart of PCA*

3.3 FUZZY C-MEANS

Fuzzy C-Mean (FCM) is an unsupervised clustering algorithm based on fuzzy set theory that allows an element to belong to more than one cluster. In FCM the number of cluster are randomly selected. FCM is the advanced version of K-means clustering algorithm and doing more work than K-means. K-Means just needs to do a distance calculation, whereas fuzzy c means needs to do a full inverse-distance weighting. This, plus the overhead needed for computing and managing, explains why FCM is quite slower than K-Means.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2$$

,

$$1 \le m < \infty$$

where *m* is any real number greater than 1, *uij* is the degree of membership of *xi* in the cluster *j*, *xi* is the *i*th of d-dimensional measured data, *cj* is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. Fuzzy partitioning is carried out through an iterative optimization of the objective function shown above, with the update of membership *uij* and the cluster centers *cj* by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\left\|x_i - c_j\right\|}{\left\|x_i - c_k\right\|}\right)^{\frac{2}{m-1}}} \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

,

This iteration will stop when $\max_{ij}\left\{\left|u_{ij}^{(k+1)} - u_{ij}^{(k)}\right|\right\} < \varepsilon$, where ε is a termination criterion between 0 and 1, whereas *k* is the iteration steps. This procedure converges to a local minimum or a saddle point of *Jm*.

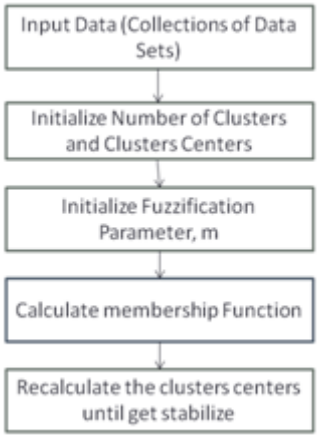

Figure 3.3: *Flow chart for Fuzzy C means algorithm*

4 DATA RESOURCES & SOFTWARE USES

This dataset is taken from UCI repository web site. The data set refers to clients of a wholesale distributor. The entire work is carried out using "RStudio-0.98.1103" to simulate the algorithm.

4.1 DATA SET (ATTRIBUTES)

	Description	Type
CHANNEL	customers channel	Nominal
REGION	customers region	Nominal
FRESH	annual spending on fresh products	Continuous
MILK	annual spending on milk products	Continuous
GROCERY	annual spending on grocery products	Continuous
FROZEN	annual spending on frozen products	Continuous
DETERGENTS PAPER	annual spending on detergents and paper products	Continuous
DELICATE-SSEN	annual spending on delicatessen products	Continuous

The data set refers to 440 customers of a wholesale: 298 from the Horeca (Hotel/Restaurant/Cafe) channel and142 from the Retail channel. They are distributed into two large Portuguese city regions (Lisbon and Oporto) and a complementary region.

Table 4.1: *Region Frequency*

Region	Frequency	Percentage
Lisbon	**77**	**17.5**
Oporto	**47**	**10.5**
Other	**316**	**31.8**

5 EXPERIMENTAL RESULTS

The intermediary quantitative variables related to the Wholesale customer's data, variables area bout in the middle of the data frame, so we can visualize all of the mat unceasing scatter plot matrix, which is the default for R's output if plot () is called on a data frame. Exploring the data reveals that channels contain more variability than the regions. Strong correlation found in the Grocery and Detergent-0.92, Milk and Detergent-0.66 & Milk & Grocery-0.73. There is somewhat high correlation between Channel and Detergents paper / Grocery.

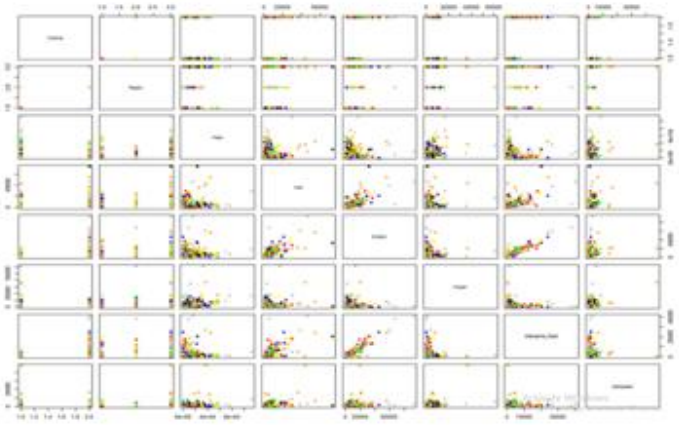

Figure5.1: *Data set being used*

5.1 PCA IMPLEMENTATION

As this is a data set of many variables (440), this would be better to apply dimensionality reduction method to make the information easier to visualize and analyze.

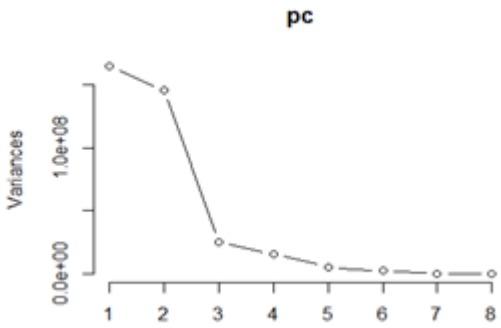

Figure 5.2: *Scree plot for PCA*

There are various rules for selecting the number of principal components:

- Use the 'elbow' method of the scree plot (on right).
- Pick the number of components which explain 85% or greater of the variation.

Here retain the first three principal components.

5.2 PCA BASED FUZZY C MEANS IMPLEMENTATION

Table 5.2 shows the result after applying Fuzzy C means on reduced dimensions by PCA.

Table 5.2: *Result of Fuzzy C-Means*

No of Clusters	Elapsed Time	Iterations	Centers
3	47.09	36 (num)	24
5	44.92	61 (num)	40
7	39.0	88 (num)	56

5.3 PCA BASED K-MEANS IMPLEMENTATION

Table 5.3 shows the result of after applying K-means on reduced dimensions by PCA.

Table 5.3: *Result of K-Means via PCA*

No of Clusters	Elapsed Time	Iterations	Centers
3	57.47	4 (int)	9
5	54.55	4 (int)	15
7	46.58	3 (int)	21

drawn by this experiment it may be safely stated that PCA+ Fuzzy C means clustering algorithm less time consuming than PCA+K- Means algorithm and hence superior.

EXECUTION TIME COMPARISON GRAPH

Generally the run time depends on input data points and number of clusters. Different type of approach yield different types of results. Usually time complexity is varies on one processor to another and also depends on speeds and systems and also varies on different simulation software.

Figure 5.3: *Execution Time Comparison*

Wholesale customer segmentation in various clusters:

Table 5.4: Size of Cluster for K=3

Algorithm	Cluster 1	Cluster 2	Cluster 3
Fuzzy C Means	270	84	86
K-Means via PCA	330	60	50

Table 5.5: *Size of Cluster for K=5*

Algorithm	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Fuzzy C Means	110	31	31	93	175
K Means via PCA	10	104	80	223	23

On the basis of the result if we select the number of clusters is =3. Then we found the following results:

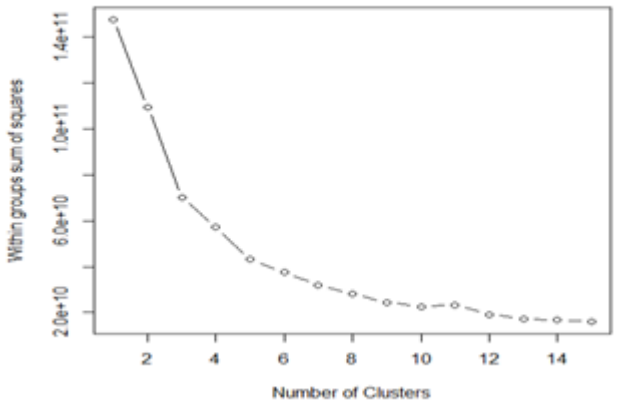

Figure 5.4: *Determine Number of Clusters*

We can summarize the following segmentation. Using Fuzzy C means algorithm:

Cluster 1: Tends to spend on Grocery, Milk and Det_Paper categories. So this clusters customers are high sender in Retail channel.

Cluster 2: Low Sender in Both Channels.

Cluster 3: Tends to spend on Fresh, Frozen and Deli categories. So this clusters customers are high sender in Horeca (Hotel/Restaurant/Café).

6 CONCLUSION

A variety of research has been done to compare the k means and fuzzy c means. In this paper Principal component Analysis has added before K – means and Fuzzy C – means algorithm. The standard K-Means algorithm is used in many fields. The efficiency of the algorithms for the wholesale customer's data is analyzed by various executions of the programs in this work. Finally, this research work concludes that the computational time of PCA+Fuzzy C means is less than PCA+K- means algorithm for the chosen application. Hence, the performance of FCM algorithm is comparatively better than the k-Means algorithm.

7 FUTURE WORK

In future research, we want to employ other kinds of datasets to do the experiment, such as financial industry or education industry and want to propose a new algorithm.

REFERENCES

[1] Customer Segmentation, http://www.statsoft.com/Textbook/Customer-Segmentation, [Access Date : 23th May, 2015].

[2] Tejwant Singh, M. M. (2014). Performance Comparison of Fuzzy C Means with Respect to Other Clustering Algorithm. International Journal of Advanced Research in Computer Science and Software Engineering, 89-93.

[3] Zhang, L. (2010). Data mining application in customer relationship management, International Conference on Computer Application and System Modeling (ICCASM) (pp. V14-171 - V14-174). Taiyuan: IEEE.

[4] D.Napoleon, S.Pavalakodi. A New Method for Dimensionality Reduction using K Means Clustering Algorithm for High Dimensional Data Set, International Journal of Computer Applications, Volume 13, No.7 (2011), pp. 41-46.

[5] Tajunisha, S. (2010). Performance analysis of k-means with different initialization methods for high dimensional data. International Journal of Artificial Intelligence & Applications, 44-52.

[6] Gayathri . A, M. (2011). Enhanced Customer Relationship Management . International Journal of Computer Science & Engineering Technology, 163-167.

[7] Ananthi Sheshasayee1 ,P. Sharmila, "Comparative Study of Fuzzy C Means and K Means Algorithm for Requirements Clustering", Indian Journal of Science and Technology,2014, Vol .7,No. 6, pp. 853–857.

[8] Soumi Ghosh, S. K. (2013). Comparative Analysis of K-Means and Fuzzy CMeans Algorithms. International Journal of Advanced Computer Science and Applications, 35-39.