IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

COMPARATIVE STUDY OF CLUSTERING TECHNIQUES IN MULTIVARIATE DATA ANALYSIS

Sabba Ruhi, Md. Shamim Reza
Department of Mathematics, Pabna University of Science & Technology, Pabna, Bangladesh

ABSTRACT

In present, Clustering techniques is a standard tool in several exploratory pattern-analysis, grouping, decision making, and machine-learning situations; including data mining, document retrieval, image segmentation, pattern recognition and in the field of artificial intelligence. In this study we have compared five different types of clustering techniques such as Fuzzy clustering, K-Means clustering, Hierarchical Clustering, Principal Components Analysis (PCA) and Independent Component Analysis (ICA) based clustering to identify multivariate data clustering.

KEYWORDS: Fuzzy clustering, K-Means, PCA, ICA, Hierarchical clustering.

INTRODUCTION

Clustering is well-recognized area in the research field of data mining. Data clustering plays the major research at pattern recognition, Signal processing, bioinformatics and Artificial Intelligence [8]. Clustering process is an unsupervised learning techniques where it generates a group of object based on their similarity in such a way that the objects belonging to other groups are similar and those belonging to other are dissimilar [5]. For last two decades, There are various clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density-based methods, and grid-based methods [2,7,8]. In some applications, for example in document retrieval, it may be useful to have a clustering that is not a partition. This means clusters are overlapping. Fuzzy clustering and functional clustering are ideally suited for this purpose. In partitioning method, a division data objects into non-overlapping clusters such that each data object is in exactly one subset. The hierarchical schemes are more versatile, and the partitioned schemes are less expensive [8]. Most common partitioning method is K-means and mixture models. Independent component analysis (ICA) recently developments technique successfully applied in multivariate data analysis such as outlier detection, data clustering, data visualization etc [19, 22].

ICA is the most natural tool for BSS [6] in instantaneous linear mixtures when the source signals are assumed to be independent. The plausibility of the statistical independence assumption in a wide variety of fields, including telecommunications, finance and biomedical engineering, helps explain the arousing interest in this research area witnessed over the last two decades. Many methods for data clustering try to identify cluster. Data clustering is carried out through the use of Principal Components Analysis (PCA) [12]. PCA is a dimension reduction procedure where some of the variables are highly correlated with each other. If this is to be used in a contaminated data, the nature of the estimated principal components may behave differently, implemented the principal components as a multivariate data clustering method.

Data analysis methods are essential for analyzing the ever-growing enormous quantity of high dimensional data. Some literatures of cluster analysis [2, 12,18] attempts to pass through data quickly to gain first order knowledge by partitioning data points into disjoint groups such that data points belonging to same cluster are similar while data points belonging to different clusters are dissimilar. One of the most popular and efficient clustering methods is the K-means method [2, 16] which uses prototypes to represent clusters by optimizing the squared error function.

In this article, we will begin a general description of different clustering techniques in the section-2, briefly describing the most popular methods of multivariate data clustering such as \Hierarchical, Fuzzy, K-Means, PCA, and ICA. In the section-3 we describe data set. In final section, we discuss results.

DIFFERENT CLUSTERING TECHNIQUES

Hierarchical Clustering

A hierarchical method creates a hierarchical decomposition of the given set of data objects. Here tree of clusters called as dendrograms is built. Every cluster node contains child clusters, sibling clusters partition the points covered by their common parent. A hierarchical algorithm yields a dendrogram representing the nested grouping of patterns and similarity levels [7, 8]. Most hierarchical clustering algorithms are variants of the single-link [23], complete-link [14]. Of these, the single-link and complete link algorithms are most popular. These two algorithms differ in the way they characterize the similarity between a pair of clusters. In the single-link method, the distance between two clusters is the minimum of the distances between all pairs of patterns drawn from the two clusters. In the complete-link algorithm, the distance between two clusters is the maximum of all pair wise distances between patterns in the two clusters. In either case, two clusters are merged to form a larger cluster based on minimum distance criteria. In hierarchical clustering we assign each item to a cluster such that if we have N items then we have N clusters. Find closest pair of clusters and merge them into single cluster. Compute distance between new cluster and each of old clusters. We have to repeat these steps until all items are clustered into K no. of clusters.

Strengths:

- Do not have to assume any particular number of clusters. Each horizontal cut of the tree yields a clustering.
- Need only a similarity or distance matrix for implementation.

Disruptive:

- Start with one, all-inclusive cluster.
- At each step, split a cluster until each cluster contains a point (or there are k clusters).

Fuzzy Clustering

Generally clustering approaches generate partitions; in a partition, each pattern belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjoint. Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function [24]. The output of such algorithms is a clustering, but not a partition. In fuzzy clustering, each cluster is a fuzzy set of all the patterns. Fuzzy set theory was initially applied to clustering refer to the literature [21]. The most popular fuzzy clustering algorithm is the fuzzy *c*-means (FCM) algorithm. Fuzzy C-Means clustering (FCM), is the data clustering algorithm in which each data set belongs to a cluster to a degree assigned by a membership. This techniques works iteratively until no further clustering is possible. Even though it is better than the hard *k*-means algorithm at avoiding local minima, FCM can still converge to local minima of the squared error criterion. Also, fuzzy clustering algorithms can handle mixed data types. However, a major problem with fuzzy clustering is that it is difficult to obtain the membership values [8].

K-Means Clustering

In present K-Means is one of the key tools among all other partitioning based data clustering methods [4, 18] and popularly use for its simplicity. It is a partition method technique which finds mutual exclusive clusters of spherical shape. It generates a specific number of disjoint, flat clusters. Statically method can be used to cluster to assign rank values to the cluster categorical data. Here categorical data have been converted into numeric by assigning rank value . K-Means algorithm organizes objects into k-partitions where each partition represents a cluster. We start out with initial set of means and classify cases based on their distances to their centers. Next, we compute the cluster means again, using the cases that are assigned to the clusters; then, we reclassify all cases based on the new set of means. We keep repeating this step until cluster means don’t change between successive steps. Finally, we calculate the means of cluster once again and assign the cases to their permanent clusters.

Algorithm k-means

a) Decide on a value for *K*, the number of clusters.
b) Initialize the *K* cluster centers.
c) Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.
d) Re-estimate the *K* cluster centers, by assuming the memberships found above are correct.
e) Repeat 3 and 4 until none of the *N* objects changed membership in the last iteration.

The strength of K-Means clustering is relatively efficient scalable process for huge sum of data sets and easy to understand and implement. It has some drawbacks that process begins only after the mean of a cluster is initialized, user defined clusters constant and hard to handle data with noise and outliers [16].

Principal Component Analysis

Principal Component Analysis (PCA) is a well-recognized tool in multivariate statistical analysis and is often used to reduce the dimension of data for easy exploration. As a multivariate analysis technique for dimension reduction, it aims to compress the data without losing much information the original data contains. The process of how PCA is done here is based on Johnson, R. [2, 11]. It is concerned with explaining the variance-covariance structure of a set of variables through a few new variables. All principal components are particular linear combinations of the p random variables with three important properties which are:

i. The principal components are uncorrelated.
ii. The first principal component has the highest variance; the second principal component has the second highest variance, and so on.
iii. The total variation in all the principal components combined is equal to the total variation in the original variables.

Mathematically,
Let X and Y are $m \times n$ matrices related by a linear transformation P. X is the original recorded data set and Y is a representation of that data set.

$$PX = Y \qquad (1)$$

Equation 1 represents a change of basis and thus can have many interpretations.

1. P is a matrix that transforms X into Y.
2. Geometrically, P is a rotation and a stretch which again transforms X into Y.
3. The rows of P, $\{p_1, \ldots, p_m\}$, are a set of new basis vectors for expressing the columns of X. Where

$$PX = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix} (x_1 \quad x_2 \ldots \quad x_n)$$

$$Y = \begin{pmatrix} p_1x_1 & \cdots & p_1x_n \\ \vdots & \ddots & \vdots \\ p_mx_1 & \cdots & p_mx_n \end{pmatrix}$$

We can note the form of each column of Y. The new variable Y is linear combination of original variables X.

$$Y_i = \begin{bmatrix} p_1x_i \\ \vdots \\ p_mx_i \end{bmatrix}$$

The observations that are cluster with respect to the first few principal components or the major principal components usually correspond to cluster on one or more of the original variables. It is well known that PC's are uncorrelated but doesn't grantee the independence among PC's. To visualize multivariate data recently developed techniques ICA can be a more powerful tool than PCA, where IC's are independent. We will discuss details on ICA in the sub-section-2.5.

Independent Component Analysis

Independent component analysis (ICA) is a Statistical and computational technique in which the goal is to find a linear projection of the data that the source signals or components are statistically independent or as independent as possible. The ICA algorithms are able to separate the sources according to the distribution of the data. Independent component analysis (ICA) [6], and projection pursuit (PP) [13], are closely related techniques, which try to look for interesting directions (projections) in the data. To achieve separation of mixed data into independent components ICA exploits the independence between the sources in order to achieve their separation from mixed data. In order to formally define ICA model, consider $X = (x_1 \quad x_2 \cdots \quad x_n)$ as a random vector, representing n sensor signals that are observable, and $S = (s_1 \quad s_2 \cdots \quad s_p)$ as a random vector of latent mutually independent sources, where $p \leq n$ .The ICA model is then given by

$$X = AS$$

Where A is a $n \times p$ matrix with full column rank, called the mixing matrix. A is assumed to be fixed, but unknown. ICA consists of estimating both the matrices A and S, when only X is known, i.e., finding a matrix W such that $S = WX$. Here, S is obtained by ICA based on the following two main assumptions on each source signals S_i in S: i) S_i is statistically independent of S_j in S $(i \neq j)$, ii) S_i is non-Gaussian random variable.

Although numerous application of ICA in different fields but its main drawback to determine order of IC's [9]. In principal component analysis, PC's are ordered by Eigen value where first Eigen value is first pc, second Eigen value second pc and so on. But in independent component analysis, these components have no order [6, 17, 22].For practical reasons to define a criterion for sorting these components to our interest. One measurement which can match our interest very well, is kurtosis. Kurtosis is a classical measure of non-Gaussianity, and is computationally and theoretically relatively simple. However, for outlier identification super Gaussian distributions (positive kurtosis) are more interesting. Negative kurtosis can indicate a cluster structure or at least a uniformly distributed factor [19, 20]. Thus the components with the most negative kurtosis can give us the most relevant information.

Data (*Country Attributes Data*)

To classify countries in groups into developed, emerging and underdeveloped based on some of their attributes such as per capita income, literacy, infant mortality rate and life expectancy [18]. To analyze their similarity and assign them to the groups, the following attributes should be taken into account:

- ❖ Per capita income (PI)
- ❖ Literacy (LI)
- ❖ Infant mortality (IM)
- ❖ Life expectancy (LE)

As the presenting problem consists of dividing countries into similar groups, it is plausible that Hierarchical, K-means, Fuzzy, PCA and ICA can be applied to this task. We want to compare which techniques capture the scenario where countries need to be classified into the three already mentioned groups: developed, emerging and underdeveloped.

Table-1: Data on country developing index

Country	PI	LI	IM	LE
Bangladesh	1314	71	45.67	70.65
Brazil	10326	90	23.6	75.4
Germany	39650	99	4.08	79.4
Mozambique	830	38.7	95.9	42.1
Australia	43163	99	4.57	81.2
China	5300	90.9	23	73
Argentina	13308	97.2	13.4	75.3
UK	34105	99	5.01	79.4
SA	10600	82.4	44.8	49.3
Zambia	1000	68	92.7	42.4
Namibia	5249	85	42.3	52.9
Georgia	4200	100	17.36	71
Pakistan	3320	49.9	67.5	65.5
India	2972	61	55	64.7
Turkey	12888	88.7	27.5	71.8
Sweden	34735	99	3.2	80.9
Lithuania	19730	99.6	8.5	73
Greece	36983	96	5.34	79.5
Italy	26760	98.5	5.94	80
Japan	34099	99	3.2	82.6

Result

By applying hierarchical clustering we found three clusters from cluster dendrogram in Fig.-1, where Japan, Australia, Sweden, Greece, Germany, U.K, Italy and Lithuania consists a cluster represent developed country. In the second group we see that Bangladesh, China, Mozambique, Georgia, Pakistan, India, Zambia and Namibia have the lowest values for all attributes and, therefore, represent the undeveloped countries. The cluster formed by the other countries, Brazil, South Africa, Turkish and Argentina represents the group of emerging countries.

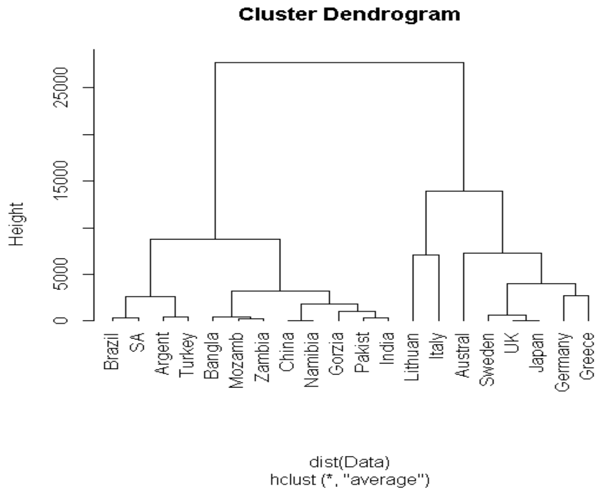

***Fig.1**.By applying hierarchical clustering in distance method to the total data, the graph shows three cluster identified based on countries attributes.*

From the dendrogram countries attributes data has three cluster based on attributes. From Fig.2, by applying Fuzzy, K-Means and PCA to the total data, the graph shows firs two pc's, K-Means and Fuzzy wrongly identified three cluster e.g- in first two PC's plot some countries has treated in the underdeveloped countries which contradicts the dendrogram, but in IC's, and ICA on PC's gives strongly identified three cluster among the countries.

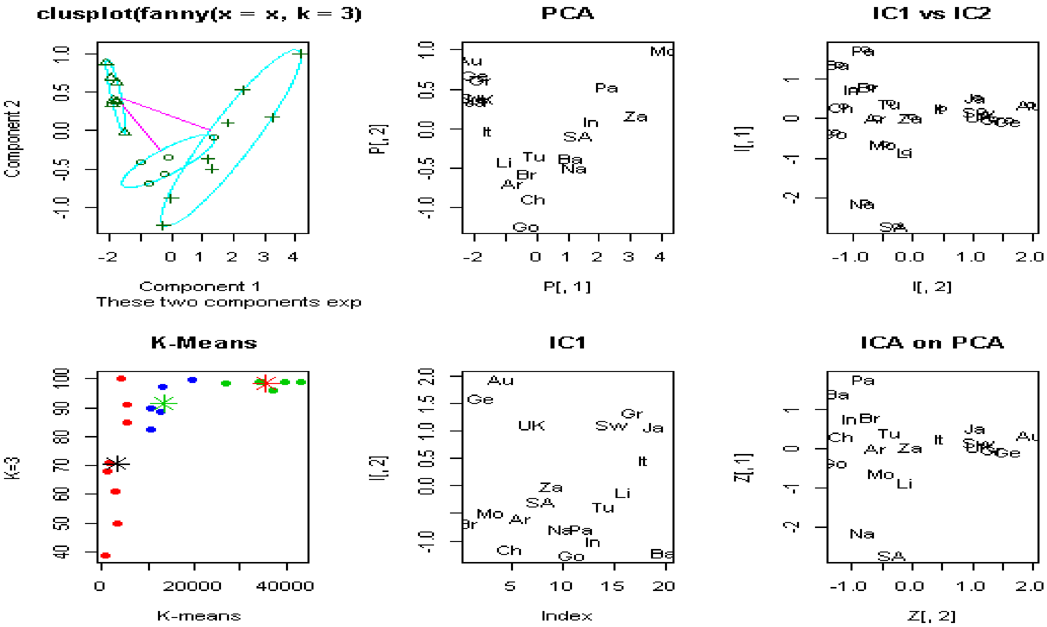

Fig. 2. On the left, by applying Fuzzy, K-Means and PCA to the data, the result is worse than the result of ICA. However, by using PCA for preprocessing before applying ICA, a more strongly cluster can be identified.

Conclusion

In this paper we have considered five techniques to identify multivariate data cluster of human developments index based on some attributers by applying Fuzzy, Hierarchical, K-Means, PCA and ICA. We also applied ICA on PCA for multivariate data clustering through normal dot plot. In our data set, our proposed method ICA on PCA a new visualization technique correctly diagnosis clusters than only PCA or others techniques. In further study, we have to use neural networks methods and comparing their performance ICA based techniques.

REFERENCES

[1] Cluster R package.(http://cran.r-project.org/web/packages/ cluster/index.html).

[2] Ding, C., & He, X. K-Means clustering via principal component analysis. Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

[3] Duda, R. O., Hart, P. E., & Stork, D. G. (2000). Pattern classification, 2nd ed. Wiley.

[4] Hartigan, J., & Wang, M. (1979). A K-means clustering algorithm. Applied Statistics, 28, 100–108.

[5] Hastie, T., Tibshirani, R., & Friedman, J. (2001). Elements of statistical learning. Springer Verlag.

[6] Hyv¨arinen, A. and Oja, E.: Independent component analysis: Algorithms and applications. Neural Networks. 4-5(13):411-430. 2000.

[7] Improved Outcome Software, Agglomerative Hierarchical Clustering Overview. Retrieved from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Agglomerative_Hierarchical_Clustering_Overview.htm

[8] Jain A.K, Murthy M.N and Flynn P.J., Data Clustering: A Review, ACM Computing Reviews, Nov 1999.

[9] Jain, A., & Dubes, R. (1988). Algorithms for clustering data. Prentice Hall.

[10] J.C. Salagubang and Erniel B. Barrios, Outlier detection in high dimensional data in the context of clustering, 12th National Convention on Statistics (NCS) EDSA Shangri-La Hotel, Mandaluyong City October 1-2, 2013

[11] Johnson, R. and Wischern, D. (2002). Applied Multivariate statistical analysis, 5th ed. Prentice-Hall, Inc.

[12] Jolliffe, I. (2002). Principal component analysis. Springer. 2nd edition.

[13] Jones,M. and Sibson, R. What is projection pursuit? J. of the Royal Statistical Society, Ser. A, 150:1-36. 1987.

[14] King, B. 1967. Step-wise clustering procedures. *J. Am. Stat. Assoc. 69*, 86 101.

[15] Leela, V. K. Sakthi priya and R. Manikandan, 2013. “Comparative Study of Clustering Techniques in Iris Data Sets” World Applied Sciences Journal 29 (Data Mining and Soft Computing Techniques): 24-29, 2014 ISSN 1818-4952.

[16] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proc. 5th Berkeley Symposium, 281–297.

[17] Matthias Scholz, Yves Gibon, Mark Stitt and Joachim Selbig, Independent component analysis of starch deficient pgm mutants. Proceedings of the German conference on Bioinformatics. Gesellschaft fur info mark, Bonn, pp.95-104, 2004.

[18] MeiraJr.,W.;Zaki,M.FundamentalsofDataMining Algorithms.(http://www.dcc.ufmg.br/miningalgorithms/DokuWiki/doku.php).

[19] Reza, M.S., Nasser, M. and Shahjaman, M. (2011) An Improved Version of Kurtosis Measure and Their Application in ICA, International Journal of Wireless Communication and Information Systems (IJWCIS) Vol 1 No 1.

[20] Reza M.S., Ruhi S., Multivariate Outlier Detection Using Independent Component Analysis, Science Journal of Applied Mathematics and Statistics, Science Publishing Group, USA, Vol. 3, No. 4, 2015, pp. 171-176. doi: 10.11648/j.sjams.20150304.11.

[21] Ruspini, E. H. 1969. A new approach to clustering. *Inf. Control 15*, 22–32.

[22] Scholz, M., Gatzek, S., Sterling, A., Fiehn, O., and Selbig, J. Metabolite fingerprinting: detecting biological features by independent component analysis. Bioinformatics 20, 2447-2454, 2004.

[23] Sneath, P. H. A. and Sokal, R. R. 1973.*Numerical Taxonomy*. Freeman, London, UK.

[24] Zadeh, L. A. 1965. Fuzzy sets. *Inf. Control 8*, 338–353.

[25] Zha, H., Ding, C., Gu, M., He, X., & Simon, H. (2002). Spectral relaxation for K-means clustering. Advances in Neural Information Processing Systems 14 (NIPS’01), 1057–1064.