International Journal of Recent Engineering Research and Development (IJRERD)
ISSN: 2455-8761
www.ijrerd.com || Volume 02 – Issue 07 || July 2017 || PP. 33-40

Feature Reduction and Robust clustering on Colon cancer data using principal component analysis

Rosy Mishra
Department of computer science and engineering,
Vikash Institute of Technology , Bargrah,,Odisha, India

Plaban Dash
Department of Electrical and Electronics engineering,
Vikash Institute of Technology , Bargrah,,Odisha, India

Abstract: In computational intelligence data as an example is the prominent tool to permute computation for grouping, prediction, recognition. Variation between data and centroid in the form of Euclidean distance, mahanabolis distance has created significant attention for computational researchers. There are so many statistical and signal processing techniques is being used to extract informative features from the data of various areas such as Biometrics, Biotechnology , Bioinformatics, image processing, speech processing , natural language processing in last decade . Unsupervised classification such as clustering has better focus than supervised classification. In this paper we have used a renown data dimension reduction technique as principal component analysis and an efficient K-mean clustering technique.

Keywords: clusters, Principal component analysis, *k*-means, modified *k*-means, Artificial neural networks

I. Introduction

PCA perform two functions like dimension reduction and extracting features. Using PCA we have to extract the features from a high dimensional data matrix by creating feature vector and computing covariance matrix.After that we have to calculate reduced matrix and use that matrix for clustering by K-mean clustering. A categories of Clustering methods are partitioning method, hierarchical method, density based method, Grid based method ,model based method. Examples of partitioning techniques are include k-means, adaptive k-means, k-medoids and fuzzy clustering. The k-mean and k-medoid algorithms used in partitioning cluster that is also called as nonhierarchical cluster. These iterative relocation techniques are use for improvement of partitioning by moving the object from one group to another group. The k-means clustering algorithm faces two major problems. First problem of obtaining non-optimal solutions. As the algorithm is greedy in nature, it is expected to converge to a locally optimal solution only and not to the global optimal solution, in general. This problem is partially solved by applying the k-means in a stochastic framework like simulated annealing (SA) and Genetic algorithm(GA)

K-Means Method

The k-means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting n intra-cluster similarity is low. Cluster similarity is low. Cluster similarity is measured in regard to the mean value of objects in a cluster, which can be viewed as the clusters centroid or centre of gravity.

"How does the k-means algorithm work?" First, it randomly selected k number of objects , each of which initially represents as a cluster mean or centre. For each of the remaining objects, an object is assigned to the cluster to which is most similar, based on the distance between the object and the cluster mean. It then computes the new mean for each cluster. This process iterates until the criterion function converges. Typically, the square-error

$$E\left(\Gamma,V\right)=\sum_{i=1}^{k}\sum_{j=1}^{n}\gamma_{ij}\left\|\overline{x}_j-\overline{v}_i\right\|^2$$

criterion is used, defined

Where E is the sum of square error for all objects in the data set; p is the point in space representing a given object; and mi is the mean of cluster ci.

Dimension Reduction Using Principal Component Analysis (PCA)

During cluster analysis of a given dataset, the selection of feature occurs to represent the data point. Features can be divided into two types, quantitative and qualitative. Quantitativefeatures include continuous, discrete and interval values, whereas qualitative features include nominal and ordinal values. The feature extraction methods

include, for example, principal component analysis, linear discriminant analysis, multidimensional scaling, self-organizing map and other projection, decomposition or transform methods. Normalization of the features may still be needed before the actual clustering in order to avoid the dominance of some features over others. In addition, if some features are strongly correlated with each other, the results can

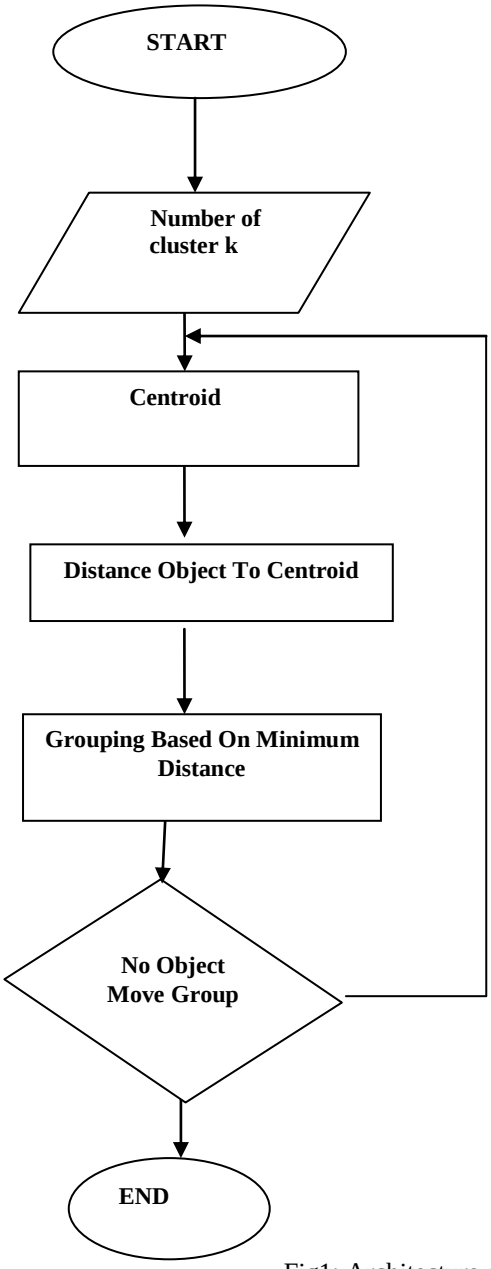

Fig1: Architecture of K-Means clustering

become skewed. Thus, not all the available features are necessarily needed. But our motive is gather the data points which are similar to each other. , the first thing to do is to define the similarity. If the chosen similarity measure is a metric, the results for search problem in metric spaces and for approximate searching can be used in clustering problems. A very popular clustering method, the k-means method, is also the best known squared error-based clustering algorithm. The method begins by initializing k cluster centers, and then proceeds by assigning data points to the center nearest to them, re-calculates the cluster centers, and assigns the points again. The process ends when there is no change in the cluster centers. The method is simple and easy to understand, but it has its drawbacks. The initial cluster centers and the number of clusters have to be given to the algorithm. The method is iterative and there is no guarantee that it converges to a global optimum, and the method is sensitive to outliers and noise. Furthermore, the k-means method can only detect hyper spherical clusters (if Euclidean distance is used) . There is still ongoing research aiming to improve the k-means method.

For example, an efficient exact algorithm for finding optimal k centers has been given, as well as a way to estimate the number of clusters present in the dataset with statistical methods during the k-means clustering procedure . Since clusters of different size, shape and density create problems in clustering, many solutions have been offered. An example of them is a method using the information about nearest neighbors of the data points in defining the similarity. The method uses core points to represent the clusters, and it has been shown to outperform k-means. It is also possible to use cluster skeletons instead of cluster centers. This approach can handle clusters with different shapes correctly. But It has been claimed that in higher-dimensional spaces (with 10–15 dimensions) the concept of "nearest neighbor" is not any more reasonable . This may have consequences in clustering procedures also. At that time Principal component analysis method is used to convert the high dimensional metrics to low dimensional metrics.

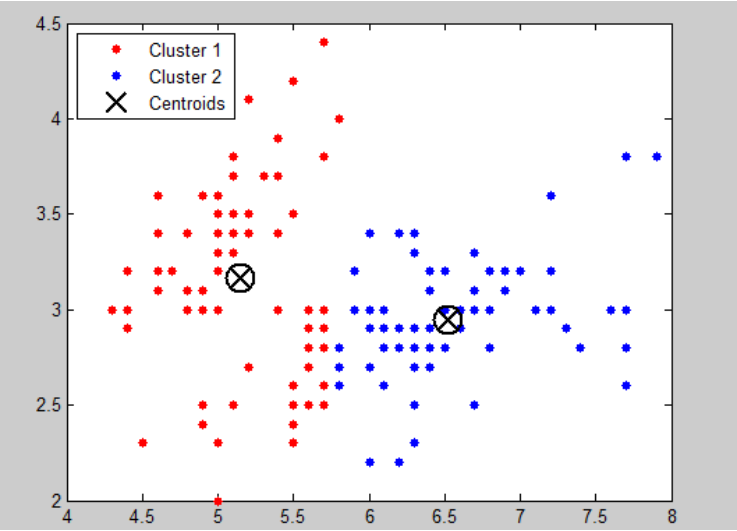

Fig2: clustering of two patterns using k-means algorithm plotted using Matlab

Supervised Learning is a machine learning paradigm for acquiring the input-output relationship information of a system based on a given set of paired input-output training samples. As the output is regarded as the label of the input data or the supervision, an input-output training sample is also called labeled training data, or supervised data. Supervised learning. The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.

Unsupervised Learning, the training data for Supervised Learning need supervised or labeled information, while the training data for unsupervised learning are unsupervised as they are not labeled (i.e., merely the inputs). In clustering, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.

II. Proposed Model

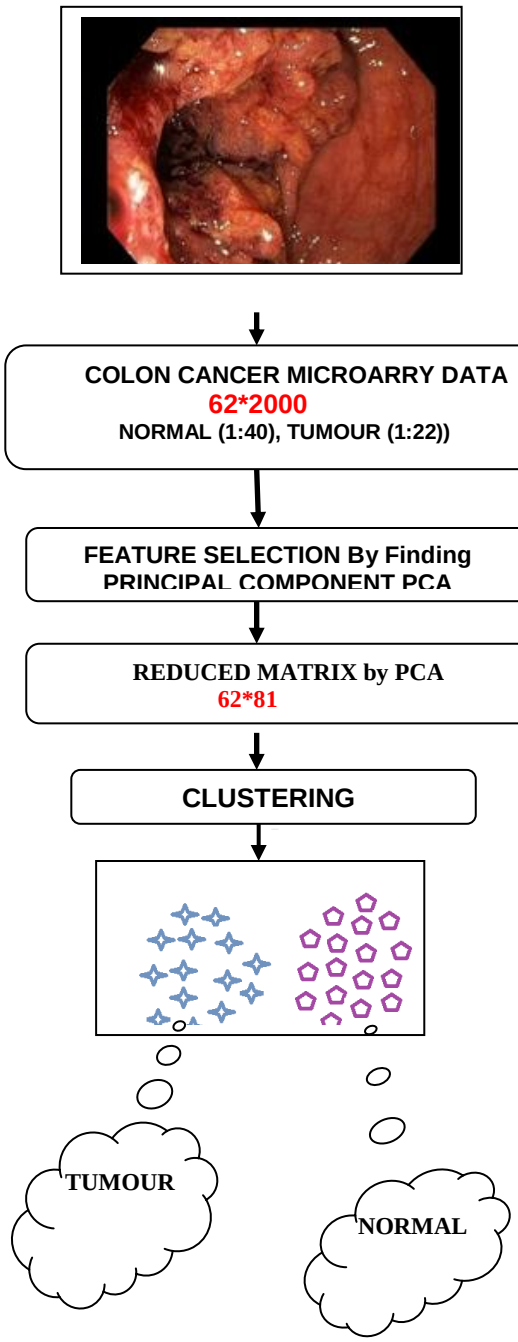

Fig3:Steps to cluster on colon cancer data

Dimension Reduction Using Principal Component Analysis (PCA):

Colon cancer

Below is the figure 9 which shows 2-D plotting of colon cancer data having data dimension 62*2000 using matlab simulation software.

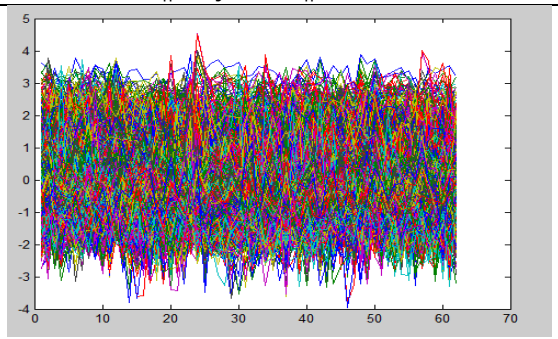

Fig4: [Colon dataset] 62*2000 plotted using matlab

In figure application of Principal Component Analysis step to find covariance matrix of the colon cancer dataset is plotted using matlab which reflects the mean value of the product of the deviations of two variants from their respective mean.

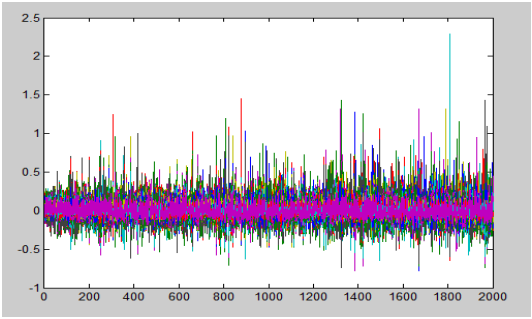

Fig5:[Colon covariance dataset] 2000*2000 plotted using matlab

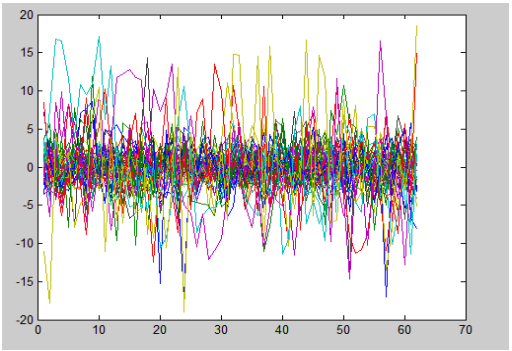

Fig6: [colon reduced data] 62*81 by PCA plotted using matlab

Dataset Used For Analysis

Without the proper data any research area is incomplete. To train a machine by a machine learning algorithm right dataset is more essence. Microarray technology has given a platform to generate genome data by using intensity color depth measurement from DNA chip.

Tumor	**Size**	**Category**	
Colon Cancer	62*2000	NORMAL (1:22)	TUMOR (23:62)

Table1:Cancerous genome data with its categories

Tumor	**Size**	**PCA**
Colon Cancer	62*2000	62*81

Table2:Cancerous genome data decomposition using Principal Component Analysis

III. Result Analysis

Experimental work was designed to compare the performance of proposed K-mean algorithm. Number of data elements selected was 155 in case of lung cancer and 81 for colon cancer after PCA based reduction. And for the sake of experiment, 5 numbers of clusters (k) were entered at run time for lung cancer and 2 numbers of clusters (k) for colon cancer. The process was repeated 10 times for different data sets generated by MATLAB. The proposed K-mean algorithm is efficient because of less number of iterations and improved cluster quality, as well as reduced elapsed time. In Figure 4, Basic and proposed K-mean clustering algorithms are compared in terms of different data sets. For each run different data sets are generated by MATLAB and entered, to observe the number of iterations. In Figure, basic and proposed K-mean clustering algorithms are compared in terms of same data set. For each run same data set is entered, to observe that at each time numbers of iterations are different in basic K-mean clustering algorithm. The numbers of iterations are fixed in proposed K-mean clustering algorithm because initial centroids are not selected randomly. Basic K-mean clustering algorithm gives different clusters, as well as clusters size differs in different runs. Table 1 shows different results for same data set as wells elapsed time. Proposed K-mean clustering algorithm gives same clusters, as well as clusters size is same in different runs. Comparison with Other K-Mean Clustering Algorithms: Comparison of proposed K-mean clustering algorithm with basic and other enhanced algorithm is given. Comparison of Basic and Proposed K-Mean Clustering Algorithm: Proposed K-mean algorithm is efficient from basic K-mean algorithm in terms of iterations, cluster quality as well as elapsed time. As in basic K-mean algorithm, initial centroids are selected randomly from the input data, so clusters vary from one another, because of which the number of iterations and total elapsed time also changes in each run of the same data.

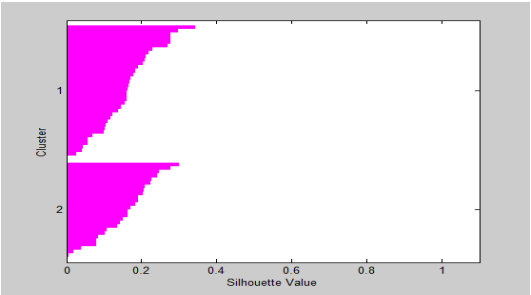

Fig7:Silhouette plot for clustering of colon cancer original data having 2 classes

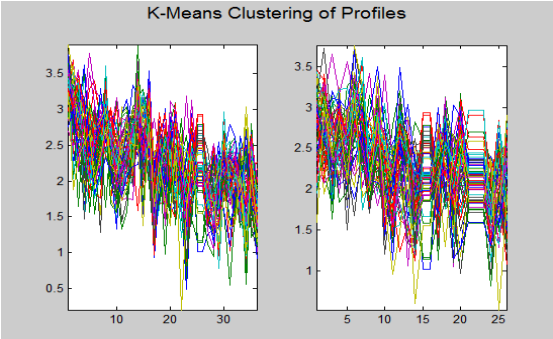

Fig8:k-means plot for clustering of colon cancer original data having 2 classes

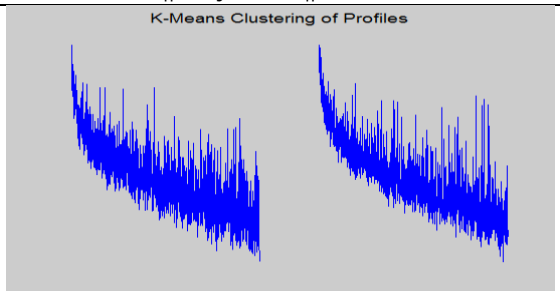

Fig9:Plot for clustering of colon cancer having 2 classes 62*2000

Silhouette categorizes similar data sample group based on Euclidean distance('Euclidean') , Squared Euclidean distance ('sqEuclidean') , Sum of absolute differences ('cityblock'), One minus the cosine of the included angle between points (treated as vectors- 'cosine') , One minus the sample correlation between points (treated as sequences of values-'correlation'), Percentage of coordinates that differ('Hamming') and plots it. The default parameter is Squared Euclidean distance. Silhouette value ranges from -1 to +1. A high silhouette value indicates that i is well-matched to its own cluster, and poorly-matched to neighboring clusters. If most points have a high silhouette value, then the clustering solution is appropriate. Silhouette(X, clust) plots cluster silhouettes for the *n-by-p* data matrix X, with clusters defined by clust. Rows of X correspond to points, columns correspond to coordinates. Cluster can be a categorical variable, numeric vector, character matrix, or cell array of strings containing a cluster name for each point. Silhouette treats NaNs or empty strings in cluster as missing values, and ignores the corresponding rows of X. By default, silhouette uses the squared Euclidean distance between points in X. Some of the matlab silhouette function explained below

s = silhouette(X, clust) returns the silhouette values in the *n*-by-1 vector s, but does not plot the cluster silhouettes.

[s,h] = silhouette(X, clust) plots the silhouettes, and returns the silhouette values in the *n*-by-1 vector s, and the figure handle in h.

[...] = silhouette(X, clust, metric) plots the silhouettes using the inter-point distance function specified in metric.

IV. Conclusion and Future Work:

Clustering large data sets is a difficult task. Researchers work to classify objects into similar sets based on their images. To group together similar documents based on the words they contain or based on different citations. Marketers clusters of similar shoppers based on their purchasing quality and history. Shop bots is a website where cluster takes place by identifying similar products and price of the products. In all of these cases, the objects are characterized by large numbers of features like text, image, and different dataset sequence. Given large data sets with hundreds of thousands or millions of entries, computing all Pair wise also possible now a days. This is generally takes place by computing the large matrix into a small matrix by using PCA. In this paper we have focused on reference matching, a particular is the merge-purge problem. When information is extracted from the web, the reference matching problem is even more severe.

V. Reference

[1]. Aleix M. Martinez, Member, IEEE, and Avinash C. Kak ," PCA versus LDA" IEEE Transcation On Pattern Aanlysis And Machine Intelligence, Vol. 23, NO. 2, Feburary 2001

[2]. Haitao Zhao, Pong Chi Yuen ; Kwok, J.T.," A novel incremental principal component analysis and its application for face recognition", Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on (Volume:36 , Issue: 4

[3]. Haibo Yao , Urbana, IL, Lei Tian ," A genetic-algorithm-based selective principal component analysis (GA-SPCA) method for high-dimensional data feature extraction.
Chein-I Chang, Qian Du, "Interference and Noise-Adjusted Principal Components Analysis", IEEE Transcations On Geoscience And Remote Sensing, VOL. 37, NO. 5, SEPTEMBER 1999

[4]. Chris Ding ,Xiaofeng He," K-means Clustering via Principal Component Analysis", Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 2004.

[5]. Zhong, W. ; Altun, G. ; Harrison, R. ; Tai, P.C. ." Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property .

[6]. Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed , "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity",959-963 ,2012ISSN 1990-9233©IDOSIPublications,2012DOI:10.5829/idosi.mejsr.2012.12.7.1845

[7]. Kuo-LungWu,Miin-ShenYang ,"Alternative c-means clustering algorithms", 0031-3203/02/$22.00 ? 2002 Pattern Recognition Society. Published by Elsevier Science
[8]. Dr. M.P.S Bhatia and Deepika Khurana, , : "Experimental study of Data clustering using k- Means and modified algorithms", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, May 2013DOI
[9]. Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE,Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" , IEEE Transcation ON Pattern Analysis AND Machine Intelligence, VOL. 24, NO. 7, JULY 2002
[10]. Malay K. Pakhira Kalyani,, "A Modified k-means Algorithm to Avoid Empty Clusters" West Bengal, INDIA International Journal of Recent Trends in Engineering, Vol 1, No. 1, May 2009
[11]. Kiri Wagsta ,Claire Cardie, , ." Constrained K-means Clustering with Background Knowledge", CA 94304 USA Proceedings of the Eighteenth International Conference on Machine Learning, 2001
[12]. Oyelade, O. J , Oladipupo, O. O, Obagbuwa, I. C , "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", 2002 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved. PII: S0031-3203(01)00197-2
[13]. Youping Deng , Dheeraj Kayarat, Mohamed O. Elasri, Susan J. Brown," Microarray Data Clustering Using Particle Swarm Optimization K-means Algorithm",
[14]. Zhong, W. ; Altun, G. ; Harrison, R. ; Tai, P.C. "Improved K-means clustering algorithm for exploring local protein sequence motifs representing common structural property, NanoBioscience, IEEE Transactions on (Volume:4 , Issue: 3)
[15]. Chris Ding and Hanchuan Peng,"Minimum redundancy feature selection from microarray gene expression data" , Journal of Bioinformatics and Computational Biology Vol. 3, No. 2 (2005) 185–205

Rosy Mishra: Btech in Information Technology From Trident Academy Of Technology Bhubaneswar. Mtech in computer science engineering From Gandhi Institute Technology Bhubaneswar. Currently She is working as Astt. Professor in the department of Computer science & Engineering, in Vikash Institute of Technology, Bragarh.

Plaban Dash: B.Tech Degree in Electrical and Electronics Engineering from National Institute of Science and Technology(NIST),Berhampur in 2012 .He is continuing his M.Tech in Electrical Engineering from NIT,Rourkela.Currently working as a Lecturer in Electrical and Electronics Engineering at Vikash Institute of Technology,Bargarh.