ISSN No. 0976-5697

Volume 3, No. 2, March-April 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Improved Clustering Technique with Informative Genes for High Dimensional Data

Clement Sherlin.C* and N. Tajunisha
*Research Scholar, Associate Professor
Sri Ramakrishna College of Arts & Science for Women,
Coimbatore, India
csherlin66@gmail.com*
tajkani@gmail.com

Abstract: Data mining is often defined as the process of finding hidden information in a database. Cluster analysis is a powerful tool in the study of gene expression data. Clustering is the process of grouping data objects into a set of disjoint classes, called clusters. K-Means clustering algorithm is one of the most frequently used clustering method in data mining, due to its performance in clustering massive data sets. In addition, the number of distance calculation increases exponentially with the increase of dimensionality of data. Microarray data is taken as high dimensional data and the dimension is being reduced with the use of dimension reduction techniques. In clustering gene expression data, the data may contain noise, irrelevant data and missing data. Preprocessing is an essential step to improve clustering. In this thesis, a new algorithm is proposed which handles high dimensional data. The dimension of the dataset is reduced using FastICA. Then, the dataset is partitioned into k equal sets. In each set, the MODE of each dimension is fixed as initial centroids for K-Means and the data points are clustered using K-Means clustering. The proposed algorithm provides better results in terms of accuracy compared to the existing algorithm.

Keywords: Clustering, High Dimensional Data, K-Means, FastICA.

I. INTRODUCTION

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. [1]Clustering is a descriptive task that seeks to identify homogeneous groups of objects based on the values of their attributes (dimensions)[2][3].Clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning ,data mining, pattern recognition, image analysis, and bioinformatics. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups [4]. Therefore, a cluster is a collection of objects that are similar among themselves and dissimilar to the objects belonging to other clusters One of the most popular clustering method is K-Means clustering algorithm developed by Mac Queen in 1967 [5].

The K-Means clustering algorithm is a partitioning clustering method that separates data into k groups [6]. The K-Means clustering algorithm is more prominent because it clusters massive data rapidly and efficiently. Because of the initial cluster centers produced arbitrarily, K-Means algorithm does not promise to produce the consistent clustering results. Efficiency of the original K-Means algorithm heavily relies on the initial centroids [7][8].Initial centroids also have an influence on the number of iterations required while running the original K-Means algorithm. The computational complexity of the original K-Means algorithm is very high; specifically for massive data sets [9].

The quality of the clustering algorithm greatly depends on the similarity measure used by the method and its implementation and also by its ability to discover some or all of the hidden patterns. The computational complexity of original K-Means algorithm is very high, especially for large data sets [10].In addition the number of distance calculations increases exponentially with the increase of the dimensionality of the data. When the dimensionality increases usually, only a small number of dimensions are relevant to certain clusters, but data in the irrelevant dimensions may produce much noise and mask the real clusters to be discovered.

Moreover when dimensionality increases, data usually become increasingly sparse, due to which data points located at different dimensions can be considered as all equally distanced and the distance measure, which, essentially for cluster analysis, becomes meaningless. Hence, attribute reduction or dimensionality reduction is an essential data-preprocessing task for cluster analysis of datasets having a large number of attributes. [10] Traditional K-Means algorithm for cluster analysis developed for low dimensional data; often do not work well for high dimensional data. Also the computational complexity increases rapidly as the dimension increases. Hence, to improve the efficiency, FastICA is applied on original data set; so that the correlated variables exist in the original dataset would be transformed to possibly uncorrelated variables, which are reduced in size. Before applying FastICA the dataset needs to be normalized, so that any attribute with larger domain will not dominate attributes with smaller domain. The resulting reduced data set is obtained from the application of FastICA and then it selects the efficient initial clusters center using the most frequent values (mode) of sorted reduced data set and the K-Means clustering algorithm is used to perform clusters.

In [11], Adam Schenker, Mark Last, Horst Bunke, Abraham kandel said, global k-means method [10] provide a way of determining good initial cluster centers for the k-means algorithm without having to use random initialization. The experiment result has shown clustering performance under global k-means to be as good as or better than using random initialization. But the execution times for the global k-means are much greater than random, due to the need to compute the initial cluster centers.

S. Deelers and S. Auwatanamongkol [4] proposed an algorithm to compute initial cluster centers for K-Means clustering. Data in a cell is partitioned using a cutting plane

that divides cell in two smaller cells. The plane is perpendicular to the data axis with the highest variance and is designed to reduce the sum squared errors of the two cells as much as possible, while at the same time keep the two cells far apart as possible. Cells are partitioned one at a time until the number of cells equals to the predefined number of clusters, k. The centers of the K cells become the initial cluster centers for K-Means. The experimental results suggest that the proposed algorithm is effective, converge to better clustering results than those of the random initialization method. The research also indicated the proposed algorithm would greatly improve the likelihood of every cluster containing some data in it.

Kohei Arai and Ali Ridho Barakbah [6] proposed a new approach to optimize the initial centroids for K-Means. It utilizes all the clustering results of K-Means in certain times, even though some of them reach the local optima. Then, the result is transformed by combining with Hierarchical algorithm in order to determine the initial centroids for K-Means. The experimental results provide accurate results and improved clustering results as compared to some clustering methods.

Gouchol Pok, Jyh-Charn Steve Liu, and Keun Ho Ryu [12] proposed a feature subset selection framework that is effective in selecting a subset of influencing genes from microarray data. It provides an explicit representation of class-specific features. It will be useful to identify biologically meaningful genes associated with a certain diagnosis. In this method a distinct typical dimension reduction methods are used that do not consider preserving the unit property of individual features in the reduced representation. Typical dimension reduction of microarray data is carried out either by reducing only the number of rows (gene expression levels) or by creating new reduced dimensions without considering the unity of original features, such as employed by PCA. In this paper, a row-wise dimension reduction by using features selection technique, while applying the clustering technique for column-wise reduction. In the existing algorithm First it selects the most relevant dimensions from the high dimensional data set using the minimum value of median absolute deviation (MAD) and then selects the efficient initial clusters centers using the most frequent values (mode) [9] of sorted and portioned (in *K* groups) reduced data set. Finally, it uses the K–means algorithm to perform the clustering

To improve the accuracy, a new method is proposed to improve the clustering by fixing the initial centroid and reduce the dimension using FastICA.

II. METHODOLOGY

A. Proposed Algorithm:

In the proposed work, FastICA have been used to find the most relevant dimensions, and then it uses the most frequent value (MODE) of selected dimensions to determine the initial clusters centers. Finally these initial clusters centers are used in the K-Means algorithm for optimum clustering. The efficient Fast Independent Component analysis selects the relevant feature from the large data set. The purpose of this method is to remove noise and redundant features from the original feature subspace. The Euclidean distance is calculated for all the data points. The data point that has the maximum distance is chosen as initial centroid.

Algorithm: The proposed algorithm is as follows:

a. Feature selection using Fast Independent Component Analysis
b. Cluster Initialization For K-Means
c. Performance Measures

Feature selection using Fast Independent Component Analysis:

a. N x D data set, where N represents number of objects in data set and D represents the dimensions of an object.
b. Z-ICs s1. . . sz with zero mean and unit variance are extracted from the gene expression dataset using ICA.
c. For gene l (l = 1 . . . p), the absolute score on each component |slj | is computed. These z scores are synthesized by retaining the maximum one, denoted by gl = max j |slj |.
d. The p genes are sorted in increasing order according to the maximum absolute scores {g1. . . gp}, and for each gene, the rank r (l) is computed. In our experiments, we found that ICA is not always reproducible when used to analyze gene expression data.

Cluster Initialization for K-Means:

i. Select the l dimensions which have minimum values of FastICA
ii. Sorted the selected dimensions in ascending order based on the dimensions having minimum FastICA value and k partitions of the reduced dataset.
iii. Calculate the most frequent value (MODE) of dimensions for each part. These values form tuples which serve as initial cluster centers.
MODE= mode (x_i)
iv. Assign each object to its closest centroid based on minimum Euclidean distance.
v. Update the centroid by calculating mean value of objects in the cluster.
vi. Repeat steps 4 to 5 until no change occur or no object move to other clusters

B. Performance Measures:

The **Rand index** or **Rand measure** is a measure of the similarity between two data clustering. This Rand index simply measures the number of pair wise agreements between clustering K and a set of class labels C, normalized so that the value lies between 0 and 1:

$$J(C,K) = \frac{a+d}{a+b+c+d}$$

where a denotes the number of pairs of points with the same label in C and assigned to the same cluster in K, b denotes the number of pairs with the same label, but in different clusters, c denotes the number of pairs in the same cluster, but with different class labels and d denotes the number of pairs with a different label in C that were assigned to a different cluster in K. The index produces a result in the range [0, 1], where a value of 1.0 indicates that C and K are identical.

III. RESULTS AND DISCUSSION

The proposed algorithm is implemented in MATLAB 7.10 with Intel(R) Core 2 duo CPU with a RAM capacity of 2 GB. The algorithm is tested with six datasets WDBC (Wisconsin Diagnostic Breast Cancer), SRBCT (Small Round Blue Cell Tumor), and Lymphoma, Colon, Leukemia, and Iris data sets. The datasets are taken from the UCI machine learning repository. The result is found to be more accurate. The cluster accuracy has been calculated for the Mode + K-Means, FastICA +K-Means.

The Figure 1 shows the Sample output screen for proposed algorithm for WDBC dataset.

Cluster formation for WDBC dataset:

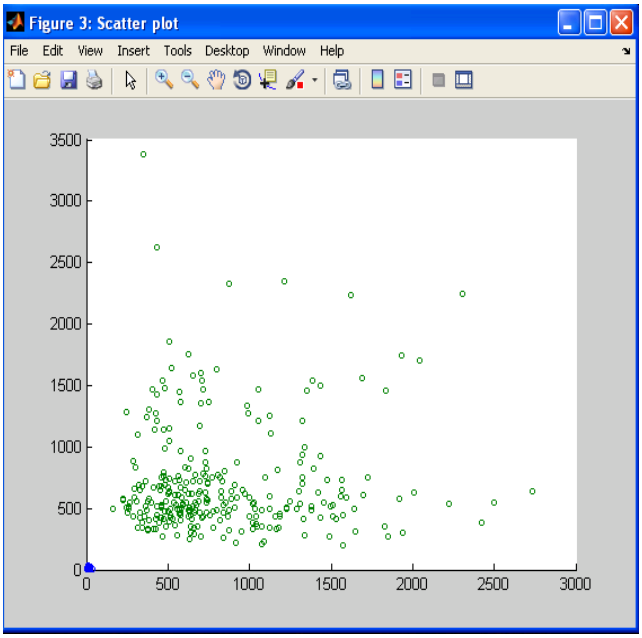

Figure 1 Sample output screen for cluster formation for WDBC Dataset.

The accuracy comparison result of WDBC dataset:

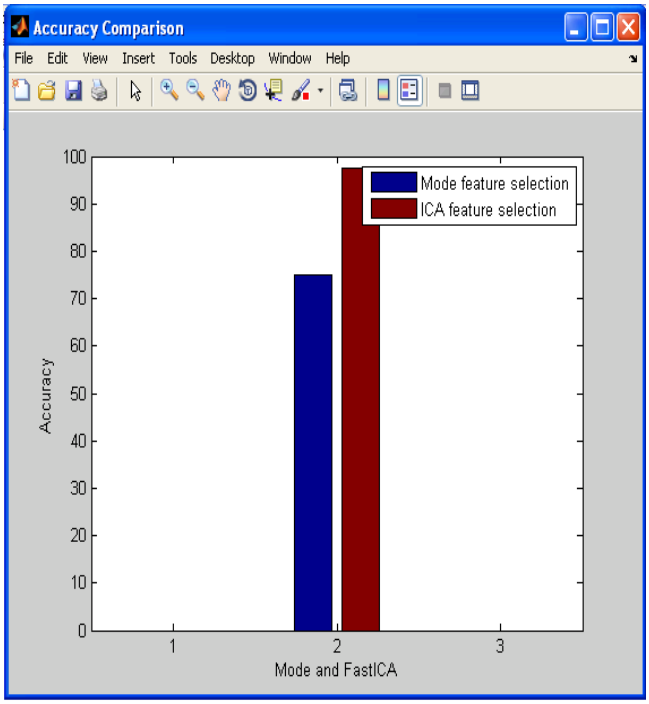

Figure 2 The graph shows the accuracy for Mode and FastICA with feature selection of WDBC dataset.

The accuracy comparison result of Iris Dataset:

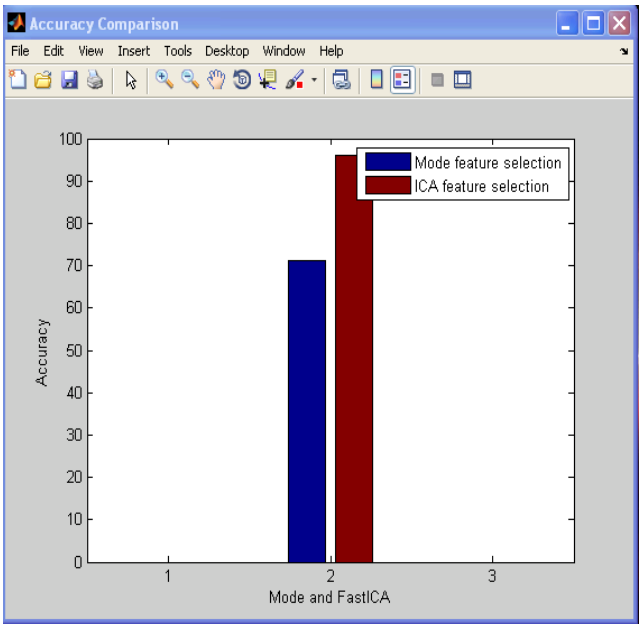

Figure 3 The graph shows the accuracy for Mode and FastICA with feature selection of Iris Dataset.

Table 1 Accuracy Comparisons of Various Datasets.

DATASETS	Mode (Accuracy in %)	FastICA (Accuracy in %)
WDBC	75.03	97.56
SRBCT	70.08	95.85
LYMPHOMA	56.52	97.72
COLON	49.39	97.67
LEUKEMIA	60.56	95.81
IRIS	71.15	97.05

The above table points out the accuracy results, which have been obtained from five different cancer data sets used by Mode +K-Means and FastICA+ K-Means. It has been clear that FastICA + K-Means are greater when compared with the previous algorithm.

IV. CONCLUSION

In this paper a new algorithm is proposed, Fast Independent Component analysis method is presented to select the best features for the recognition of the cancer and assigning the initial centroids for the minimum frequent value. The K-Means clustering algorithm is applied to form the optimal clusters. The performance evaluation of the proposed system was evaluated and compared with existing approaches. The K-Means algorithm is applied for assigning cluster centers and Fast Independent Component Analysis for sample clustering which has given more accuracy with less computing time compared to the existing work. For clustering validation, Rand Index measurement is used to compute the accuracy between the class label and clustered samples.

V. REFERENCES

[1] Pavel Berkhin "Survey of Clustering Data Mining Techniques Grouping Multidimensional Data Volume: Cl,

Issue: c, Publisher: Springer, Pages: 25-71. DOI: 10.1007/3-540-28349-8_2, 2006.

[2] Siyoung Park, Daewoo Choi, Chi-Hyuck Jun," A Clustering Method for Discovering Patterns Using Gene Regulatory Processes, Genome Informatics Series, 12: 249–251, 2001.

[3] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan,"Automatic Subspace Clustering of High Dimensional Data", Data Mining and Knowledge Discovery, 11, 5–33. Springer Science 2005.

[4] S. Deelers and S. Auwatanamongkol, "Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance," International Journal of Computer Science, Vol. 2, Number 4.

[5] Mac Queen J, "Some Methods for Classification and Analysis of Multivariate Observations," Proc. 5th Berkeley Symp. Math. Statist. Prob. (1): 281–297, 1967.

[6] Koheri Arai and Ali Ridho Barakbah, "Hierarchical K-means: An Algorithm for Centroids Initialization For K-means," Department of Information Science and Electrical Engineering Politechnique in Surabaya, Faculty of Science and Engineering, Saga University, Vol. 36, No.1, 2007.

[7] K. A. Abdul Nazeer and M. P. Sebastian, "Improving the Accuracy and Efficiency of the K-Means Clustering Algorithm," in International Conference on Data Mining and Knowledge Engineering (ICDMKE), Proceedings of the World Congress on Engineering (WCE-2009), Vol 1, London, UK July 2009.

[8] Madhu Yedla, Srinivasa Rao Pathakota and T M Srinivasa "Enhanced K-Means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies, Vol.1 (2), 121-125, 2010.

[9] Shehroz S. Khan, Shri Kant "Computation of Initial Modes for K-modes Clustering Algorithm using Evidence Accumulation"IJCAI-2007.Proceedings of the 20th International Joint Conference on Artificial Intelligence.

[10] Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath and Milu Acharaya "A Hybridized K-Means Clustering Approach for High Dimensional Dataset", International Journal of Engineering, Science and Technology, Volume 2, No.2, pp. 59- -66. 2010.

[11] Adam Schenker, Mark Last, Horst Bunke, Abraham Kandel "Comparison of Two Noval Algorithms for Clustering Using Web Documents, WDA." 2003.

[12] Gouchol Pok, Jyh-Charn Steve Liu, and Keun Ho Ryu, Bioinformation,"Effective Feature Selection Framework for Cluster Analysis of Microarray Data."PMCID: PMC29516662010; 4(8): 385–389. Published online 2010 February 28.

