ISSN No. 0976-5697

Volume 3, No. 3, May-June 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

PSO Optimized Hybridized K-Means Clustering Algorithm for High Dimensional Datasets

H.S Behera* and Madhusmita Mishra
Department of Computer Science & Engineering
Veer SurendraSai University of Technology, Burla
Sambalpur, Odisha, India
hsbehera_india@yahoo.com* madhusmita.cs@gmail.com

Abstracts-Clustering is a widely used concept in data mining which finds interesting pattern hidden in the dataset that are previously unknown. K-means is the most efficient partitioning based clustering algorithm because it is easy to implement. However, due to rapid growth of datasets in practical life, the computational time, accuracy and efficiency decreases while performing data mining task. Hence an efficient dimensionality reduction technique should be used. Due to sensitiveness to initial partition k-means clustering can generate a local optimal solution. Particle Swarm Optimization (PSO) is a globalized search methodology but suffers from slow convergence near optimal solution. In this paper, a PSO optimized Hybridized K-Means is proposed to cluster high dimensional dataset. The proposed algorithm generates more accurate, robust and better clustering with reduced computational time.

Keywords*:* clustering, k-means algorithm, Dimensionality Reduction, Principal Component Analysis, Particle Swarm Optimization

I. INTRODUCTION

Data Mining is a convenient way of extracting useful information from large data sets in terms of knowledge, pattern or rules. Data mining can be performed on various types of databases and information repositories, but new patterns can be found by various data mining techniques like association, correlation analysis, classification and clustering techniques. Clustering is an unsupervised classification technique [1] because it does not have any prior knowledge of data before classification. Clustering is the process of finding groups of objects such that objects in a group are similar to one another and different from the objects in other groups. A good clustering method will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity. To achieve different application purposes, a large number of clustering algorithms have been developed. Clustering techniques can be broadly divided into five categories [2]: hierarchical methods, partitioning methods, density-based methods, grid based method, model based method. Partitioning clustering algorithms, such as K-means, K-medoid, which assign objects into k (predefined cluster number) clusters, and iteratively reallocate objects to improve the quality of clustering results . Hierarchical clustering algorithms assign objects in tree-structured clusters, i.e., a cluster can have data points or representatives of low level clusters.

Hierarchical clustering algorithms can be classified into categories according their clustering process: agglomerative and divisive [2].Density-based clustering is that for each instances of a cluster the neighborhood of a given radius has to contain at least a minimum number of instances. Grid-based clustering first quantize the clustering space into a finite number of cells. Cells that contain more than certain number of points are treated as dense and the dense cells are connected to form the clusters. In model based method in addition to the observed or predictive attributes, there is a hidden variable which reflects the cluster membership for every case in the data set.

K-means is a commonly used partitioning based clustering technique that tries to find a user specified number of clusters(k), which are represented by their centroids [2,3,4,5], by minimizing the sum of square error. Although K-means is simple and can be used for a wide variety of data types, it is very sensitive to initial position of centroids. So, a suitable technique should be used to find out the initial centroids. Moreover computational complexity of original K-means algorithm is very high for high dimensional data sets because the number of distance calculation increases exponentially with the increase in the dimensionality of the dataset. When the dimensionality increases, only a small number of dimensions are useful to certain clusters, but data in the irrelevant dimensions may produce much noise and produce erroneous clusters. Hence dimensionality reduction is an essential data-preprocessing task for cluster analysis of datasets having a large no. of features or attributes.

To improve the efficiency of K-means clustering algorithm an efficient dimensionality reduction technique is used i.e., Principal Component Analysis (PCA). After applying PCA the correlated variables exist in the original dataset would be transformed to possibly uncorrelated variables [5], which are reduced in size. Before applying PCA the dataset needs to be normalized, so that any attribute with larger domain will not suppress the attribute with smaller domain. The resulting dataset from PCA is then used for clustering.

The Particle Swarm Optimization (PSO) is a nature inspired population based optimization algorithm [9]. It is a swarm intelligence technique based on the observation of the collective behavior in decentralized and self-organized systems. Its example are bee colonies, ant colonies, bird flocking, animal herding, fish schooling. The particles search locally but the interaction with each other leads to the emergence of global behavior. The PSO algorithm can be used to generate good initial cluster centroids for K-means by avoiding being trapped in a local optimal solution. In this paper we have combined both PCA and PSO for better clustering.

II. RELATED WORKS

K. A. Abdul Nazeer [3] proposed an algorithm where he has calculated the initial centroid by taking most closest data points in to a group up to a limit .75*(n/k) ,and then calculated their mean to take them as initial centroid for K-means algorithm. Madhu Yedla [4] proposed an enhanced method where if the dataset contain both positive and negative value then each attribute is subtracted from the minimum attribute, then the distance from origin is calculated, data points are sorted according to distances from which k subsets are taken and their mean value is calculated to take initial centroid for clustering using K-means. Rajashree Dash [5] proposed an algorithm where she has used a dimensionality reduction technique i.e., PCA to reduce the high dimensional dataset and calculated the initial centroid by an enhanced method to use them in k-means clustering. D.Napoleon [6] has applied PCA and taken the median of 'k' subsets of maximum variances of the dataset as initial centroid then applied k-means. Sandeep Rana [7] uses a global search method i.e., Particle Swarm Optimization to find the initial centroid then these centroid are applied to k-means clustering algorithm for better solution. All the above methods have not used a good method to calculate the initial centroid, some of them applied dimensionality reduction technique but have not taken any good concept to choose the number of principal components and also they have not used any optimization method to get a better result.

III. MATERIALS AND METHODS

A. K-Means Clustering Algorithm:

K-means is a commonly used unsupervised partitioning based clustering algorithm [1]. This algorithm aims to find user specified number of cluster (k) of a given dataset which are represented by their centroid by minimizing the sum of square error. The K-means algorithm starts by initializing the k cluster centers. The input data points are then allocated to the closest centroid by calculating the square of Euclidean distance to the centroids [3,5]. The mean of each cluster is then computed to update the cluster center. Then again the data points are assigned, distance is calculated ,mean is updated following the same procedure until no changes in the centroid of each cluster occurs. The steps of the K-means clustering are as follows:

a. Chose randomly K data points to initialize the clusters.
b. For each data point, find the closest cluster center, and assign that input data point to the corresponding cluster.
c. Update the cluster centers in each cluster using the mean of the data points assigned to that cluster.
d. Repeat steps 2 and 3 until no more change in the value of the centroids.

B. Principal Component Analysis:

Principal component analysis is a variable reduction procedure. It is useful when you have a large number of data on a number of variables having some redundancy. In this case, redundancy means that some of the variables are correlated with one another, because they are measuring the same construct. Because of this redundancy, it should be possible to reduce the observed variables into a smaller number of principal components that will account for most of the variance in the observed variables. Principal Component Analysis is an unsupervised Feature Reduction method for projecting high dimensional data into a new lower dimensional representation of the data that describes as much of the variance in the data as possible with minimum reconstruction error. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on[5]. It transforms a number of possibly correlated variables into a smaller number of uncorrelated variables called principal components.

The first component extracted in a principal component analysis accounts for a maximal amount of total variance in the observed variables. The second component extracted will account for a maximal amount of variance in the data set that was not accounted for by the first component and it will be uncorrelated with the first component. When the principal component analysis will complete, the resulting components will display varying degrees of correlation with the observed variables, but are completely uncorrelated with one another. PCs are calculated using the Eigen value decomposition of a data covariance matrix/ correlation matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute [8]. Covariance matrix is preferred when the variances of variables are very high compared to correlation.

For many datasets, the 1st several PCs explain the most of the variances, so that rest can be eliminated with minimal loss of information. In our proposed algorithm we have taken the variances hiving value greater than mean Eigen value as the 1st Principal Components.

C. Particle Swarm Optimization:

Particle Swarm Optimization is a computational method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality. PSO optimizes a problem by having a population of candidate solutions, here particles are initialized, and moved around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position and is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles [9]. This is expected to move the swarm toward the best solutions.

In PSO swarm is composed of a set of particles. The position of particle corresponds to a candidate solution of the optimization problem. Each particle is associated with position X_i and velocity V_i. The best position that a particle has ever visited is known as personal best and represented by $pbest_i$. The best position of all the particles is known as global best and represented by $gbest_i$. The PSO consists of 3 steps: Evaluate the fitness of each particle, update individual and global best fitness and positions, update velocity and position of each particle[10]. First the position and velocity of each particle is initialized by taking the upper and lower bounds variables PUB(Particle Upper Bound), PLB(Particle Lower Bound) and VUB(Velocity Upper Bound), VLB(Velocity Lower Bound) on the search space, as expressed in equations (1) and (2), rand is a variable that can take any value between 0 and 1.

x_i^0= (rand-1/2)*(PUB-PLB)+1/2*(PUB+PLB) (1)

v_i^0= (rand-1/2)*(VUB-VLB)+1/2*(VUB+VLB) (2)

Fitness evaluation is conducted by supplying the candidate solution to the objective function. Individual and global best fit nesses and positions are updated by comparing the newly evaluated fit nesses against the previous individual and global best fit nesses, and replacing the best fit nesses and positions as necessary. The velocity of each particle in the swarm is updated using the following equation:

$v_i(t + 1) = wv_i(t) + c1rand\ [pbest_i(t) − x\ _i(t)] + c2rand[gbest_i(t) − x_i(t)]$ (3)

$v_i(t)$ and $x_i(t)$ are the velocity and position of the particle at time t. Parameters w,c1&c2 (0<=w<=1.2,0<=c1<=2,0<=c2<=2) are user supplied coefficients. $pbest_i(t)$ is the individual best candidate solution at time t and $gbest_i(t)$ is the swarm's best candidate solution at time t. The 1st part in equation(3) is the inertia component, responsible for keeping the particle moving in the same direction it was originally heading. The 2nd part in equation(3) is the cognitive component which influence the particle to return to the region of the search space in which it has experienced high individual fitness. The 3rd part in the equation is called as social component which causes the particle to move to the best region the swarm has found so far[10].Once the velocity of each particle is updated, each particle's position is updated by applying the new velocity to particle's previous position with the help of following equation:

$x_i(t + 1)= x_i(t)+v_i(t + 1)$ (4)

PSO algorithm is very fast, simple, easy to understand, implement and requires little memory for computation. But it has a major drawback that due to its fast convergence nature it may converge in mid optimum points instead of global optimum point.

IV. PROPOSED OPTIMIZED CLUSTERING METHOD

As original K-means clustering does not perform well for high dimensional dataset, hence we used PCA to obtain a reduced dataset containing possibly uncorrelated variables. Here the number of clusters to be formed is equal to the number of PCs found from PCA. Another problem in k-means algorithm is that it finds a local minimum instead of global minimum which is very important in clustering, basically when the dataset is very large and important like medical, security, finance etc. The bio-inspired algorithm i.e., Particle Swarm Optimization which follows the process of random searching and information sharing is best for finding global solutions Here we have used PSO to optimize the centroids due to its fast convergence speed and then the result is tuned by the help of K-means clustering algorithm. Here we have chosen the data points as initial centroid whose squared Euclidean distance is maximum among all the data points.

Proposed Optimized Algorithm:

Phase-1: Aplication of PCA to reduce the dataset

a. Organize the dataset in a matrix A.

b. Normalize the dataset using the formula: $\frac{(V-mean\,(A))}{std\,(A)}$

c. Calculate the covariance of the normalize matrix B.

d. Calculate the Eigen vector and Eigen value of the matrix B.

e. Sort the Eigen vectors in increasing order of Eigen value.

f. Choose the 'p' principal components/ variances from matrix B which have Eigen value greater than mean Eigen value.

g. Form a transformation matrix P consisting of these PCs.

h. Find the reduced projected dataset Y in a new coordinate axis by applying P to B.

Phase-2: Find the initial centroid using PSO

i. Particles' initial position are taken as in matrix Y(obtained by applying PCA).

j. The velocity initialization is done by equation(2).
The fitness value of each particle is computed by the following fitness function. $f=\Sigma\|x_i-z_j\|$, i=1,…,n,j=1,…,c (5)
Where, x_i is data point and z_j is cluster center.

k. Velocity and position are updated according to equation (3) and (4).
As particles are initialized within a boundary value, so if they move out of boundary they are reset to the boundary value.

l. Steps 3-4 are repeated until maximum iteration is reached.

m. The final gbest position are taken as initial centroid.

Phase-3: Apply K-means clustering with the initial centroid

n. Assign each data point to the group that has closest centroid.

o. Update the cluster centers in each cluster using the mean of the data point, assigned to that cluster.

p. Repeat the steps 14-15 until there is no more change in the value of the centroids.

The PSO algorithm is used at the initial stage to find the optimal solution by global search and the result is used in K-means for refining and generating the final result.

V. EXPERIMENTAL ANALYSIS AND RESULT DISCUSSION

In this section we have discussed the overall result of the proposed algorithm and compared its result with original K-means and Hybridized K-means proposed by Rajashree Dash(5). We have taken a sample of pima Indian diabetes data set which has 65 data point and 8 attribute. After applying PCA the dataset is reduced to three PCs or three attributes, so three clusters will be formed. The complete experiment is done by using MATLAB and the result is shown below:

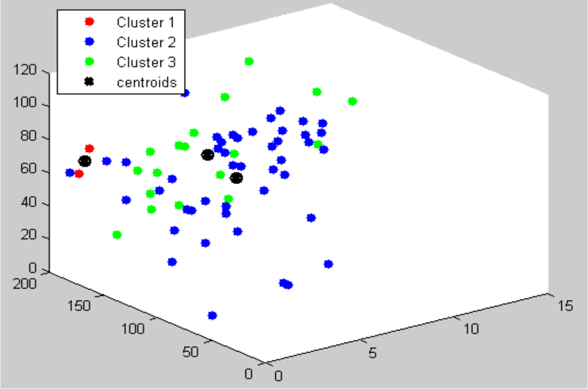

Figure-1:Clustering using original K-means

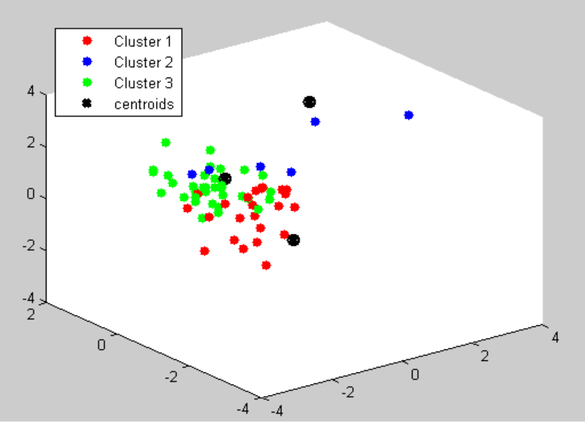

Figure-2:Clustering using hybridized clustering using PCA

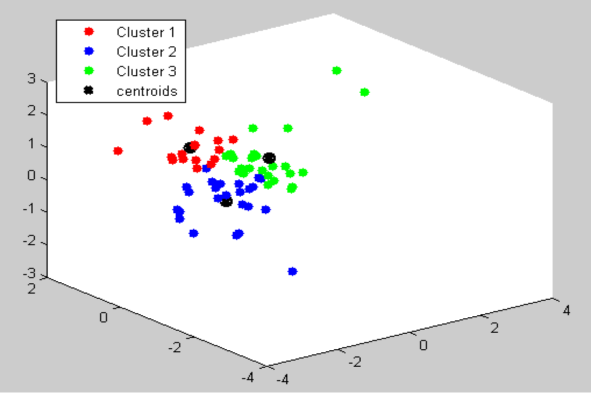

Figure-3:Clustering using proposed optimized method using PSO

In the table below we have taken Pima Diabetic dataset and Breast cancer dataset(having 50 data points and 10 attributes), applied them in hybridized K-means algorithm, Optimized K-means algorithm and compared the computational time and sum of squared error in all the cases.

Table-1:Time taken in secs and SSE value obtained with Hybridized K-means and proposed Algorithm

Data Sets	Hybridized K-Means Algorithm using PCA		Optimized K-Means Algorithm using PSO	
	Execution time	Sum of Squared Error	Execution time	Sum of Squared Error
Pima Indian Diabetes Dataset	2.000034	195.396	1.246423	93.8078
Breast cancer Dataset	1.192823	99.4578	1.119686	42.1449

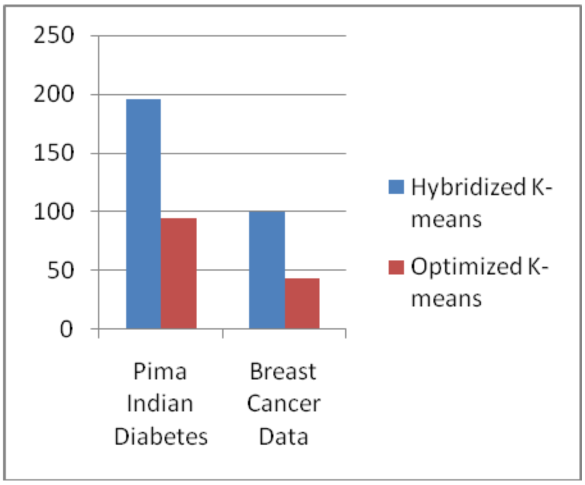

Figure-4: SSE results on Pima Indian Diabetes and Breast Cancer Dataset

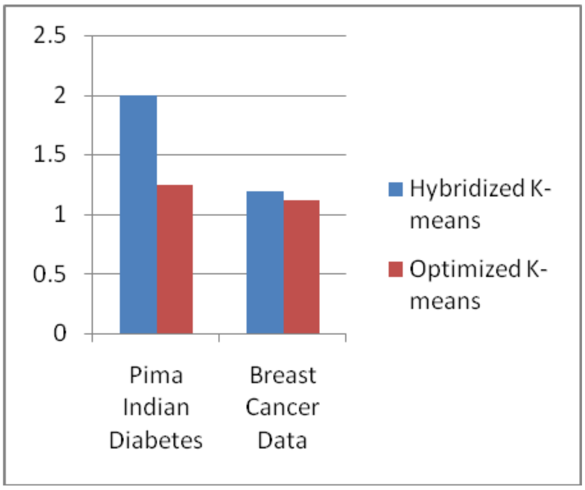

Figure-5: Comparision of Execution time of Pima Indian Diabetes and Breast Cancer Dataset

VI. CONCLUSION

In this paper we have proposed an optimized K-means clustering algorithm which first follows the dimensionality reduction technique i.e., PCA, determine the initial centroid according to maximum distance, centroids are optimized using bio-inspired algorithm and used as initial centroid for K-means algorithm to fine tune the result. Using the proposed method a given dataset is partitioned in such a way that the sum of total clustering error is reduced to a large extent and inter cluster distance is increased. By comparing our result through experiment with the hybridized K-means algorithm using PCA we found better performance and accuracy of our algorithm. In our algorithm the sum of squared error and execution time is reduced to a large extent when it is applied to various practical datasets. Again the method for finding the initial centroid may not work well for every dataset , so it is suggested to do future research on how to calculate the initial centroids, find better technique to handle high dimensional dataset and to determine proper value of ‘k’.

VII. REFERENCES

[1] N. Grira, M. Crucianu, N. Boujemaa, “Unsupervised and Semi-supervised Clustering: a Brief Survey”, August 15, 2005.

[2] S.B. Kotsiantis, P. E. Pintelas, ”Recent Advances in Clustering: A Brief Survey”,pp.1-9.

[3] K. A. Abdul Nazeer, M. P. Sebastian,”Improving the Accuracy and Efficiency of the k-means Clustering Algorithm” WCE , July 1 - 3, 2009, pp.978-988.

[4] M.Yedla, S. R. Pathakota, T .M. Srinivasa,”Enhancing K-means Clustering Algorithm with Improved Initial Center”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2) , 2010, pp.121-125.

[5] R. Dash, D. Mishra, A. K. Rath, M. Acharya,”A hybridized K-means clustering approach for high dimensional dataset”, International Journal of Engineering, Science and Technology,Vol. 2, No. 2, 2010, pp. 59-66.

[6] D.Napoleon, S.Pavalakodi, “A New Method for Dimensionality Reduction using KMeans Clustering Algorithm for High Dimensional Data Set”, International Journal of Computer Applications, Volume 13– No.7, January 2011,pp.41-46.

[7] S. Rana, S. Jasola, R. Kumar, “A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm”, International Journal of Engineering, Science and Technology,Vol. 2, No. 6, 2010, pp. 167-176.

[8] Tajunisha and Saravanan,”An efficient method to improve the clustering performance for high dimensional data by Principal Component Analysis and modified K-means”, International Journal of Database Management Systems (IJDMS), Vol.3, No.1, February 2011,pp.196-205.

[9] J. Kennedy, R. Eberhart, “Particle Swarm Optimization” IEEE Int'l. Conf. on Neural Networks, IEEE Service Center, Piscataway, NJ, vol IV:1942-1948.

[10] J. Blondin,” Particle Swarm Optimization: A Tutorial”, September 4, 2009,pp.1-5.

[11] P. Valarmathie, DR M. V Srinath, K. dinakaran,” An Increased Performance of Clustering High Dimensional Data Through Dimensionality Reduction Technique”, Journal of Theoretical and Applied Information Technology, pp. 731-733.