A Linear PCA based hybrid K-Means PSO algorithm for clustering large dataset

Chetna Sethi, Garima Mishra

ABSTRACT - Clustering is a process of grouping together same data vectors into a specified number of clusters or groups. Clustering algorithms are important data mining task that have been applied on variety of fields. Partitional Clustering algorithms such as k-means algorithm are most popular for clustering large data-sets. The drawback of k-means clustering is that initially the number of clusters is not known and the clustering result depends on selection of initial centroids and may converge to local optimum. Also when the data set is of high dimension, k-means algorithm lost its accuracy.Linear Principal Component Analysis is used for dimension reduction. With the deficiency of global search strategy for k-means clustering algorithm and clutersing large data sets we develop a Linear PCA based hybrid k-means clustering and PSO algorithm (PCA-K-PSO) .In (PCA-K-PSO) algorithm the fast convergence of k-means algorithm and the global searching ability of Particle Swarm Optimization(PSO) are combined for clustersing large data sets using Linear PCA. .

Index terms- Data clustering, Linear PCA, k-means, PSO, Global Search strategy, large data set

1 INTRODUCTION

Due to its numerous applications Data Clustering is one of the widely studied research topics. Clustering means to divide a set of objects into specified number of clusters based on similarity. There are two major clustering techniques- "Partitioning and Hierarchical" [1]. K-means algorithm is an effective partitioning clustering used for solving data clustering problem. This algorithm try to minimize a certain criteria and therefore can be treated as an optimization problem [2]. The time complexity of k-means clustering is almost linear. The major drawback of k-means clustering method is that it is sensitive to the selection of initial cluster centres and it converges to local optimum. This drawback of k-means clustering is solved by an efficient global search strategy called as Particle Swarm Optimization (PSO).

Particle Swarm Optimization (PSO) Algorithm is global population based stochastic optimization technique modeled after the social behavior of bird flock [3, 4]. PSO can converge rapidly .Various literatures reveal that PSO can be successfully applied to data clustering technique. The initial selection of cluster centroids decides the processing of PSO and partition result on data set as well. The main advantage of PSO is that it has very few parameters to adjust and it finds the best value for interaction of particles

This paper shows the applicability of PSO to data clustering. The global search capability of PSO is utilized for clustering data. Here we present a hybrid K-means +PSO Clustering algorithm that performs fast clustering. This is used for clustering efficiently high- dimensional data sets .

The rest of paper is organized in following manner: Section 2 provides related work in data clustering using PSO. Section 3 provides an overview of k-means clustering. Section 4 provides Particle Swarm Optimization. A hybrid improved PSO based k-means clustering algorithm is proposed in Section 5. Section 6 gives Conclusion. In Section 7 Future Scope is discussed.

2 RELATED WORK

The k-means clustering algorithm and its various variants are well-known data clustering algorithm [5, 8]. The drawback of this k-means algorithm is that the clustering result is sensitive to selection of the initial cluster centres and it easily converges to local optima. Recently, many new approaches have been proposed which are inspired from biological

characteristics to solve the data clustering problem. These include Genetic Algorithm [6], PSO, Ant Clustering and Self Organizing Maps (SOM) [7]. Authors presented a hybrid document clustering algorithm capable of performing fast Clustering[5] .Reference [11] proposed a discrete PSO with crossover and Mutation operators of GA for Clustering. [10]presented a mountain clustering based on an improved swarm optimization (MCBIPSO) algorithm.[9] reported that when data set is large and has high dimension, the time single swarm is incapable of exploring the entre solution space. To solve such kind of problems they proposed a multiple swarm based clustering algorithm based on PSO. Reference [13] proposed a discrete-coded implementation of the PSO algorithm to identify and select variables with a significant clustering tendency. In particular, to cope with the problems that can arise when performing variable selection on data sets with a high number of variables and to account for possibility of using this method, an unsupervised approach was used. In[12] authors presented a hybridized fuzzy clustering technique based on fuzzy PSO (FPSO) and FCM which combines the merits of both these algorithms. In Reference[14], authors proposed a novel approach to particle swarm optimization (PSO) using digital pheromones to co-ordinate swarms within a n-dimensional design space to improve reliability and efficiencyof search. In [15], authors investigated the application of EPSO to cluster data vectors. Comparision was made between the EPSO and the PSO algorithm, it was showed that the EPSO converge slower while PSO converges faster than the former. In [16], k-means clustering was proposed for improving the perfomance .

This proposed hybrid approach is efficient and can give encouraging results.

3 K-MEANS CLUSTERING

K-means clustering groups data objects into a number of clusters.These clusters are predefined The similarity measure used is Euclidean distance. Data vectors or objects belonging to the same cluster have small Euclidean distances from one object to another. These are associated with one central centroid vector, which is actually the "midpoint" of that cluster. The "mean value" of all the data objects is the centroid vector that belongs to that given cluster.

The standardized form of k-means algorithm is given below:

1. Initialize the ***"Nc"*** cluster centroid vectors. This *Nc* is initialized randomly.

2. Repeat

(a)Assign the data vector to the class with the closest centroid vector, for each data vector where the distance to the centroid vector is calculated using the Euclidean distance formula given as:

$$d(\mathbf{z}_p, \mathbf{m}_j) = \sqrt{\sum_{k=1}^{N_d} (z_{pk} - m_{jk})^2} \qquad (1)$$

(b)Recalculate the new cluster centroid vectors, using the formula given below

$$\mathbf{m}_j = \frac{1}{n_j} \sum_{\forall \mathbf{z}_p \in C_j} \mathbf{z}_p \qquad (2)$$

3. Continue the process until a stopping criterion is met

Where the symbols which are used above are defined as follows =>

- N_c denotes the number of cluster centroids (given by the user). This is actually the number of clusters or groups which are to be formed
- z_p denotes the p^{th} data vector.

□ m_j denotes the centroid vector for the cluster j.

□ n_j denotes the number of data vectors in cluster j

□C_j denotes the subset of data vectors that form cluster j.

k-means clustering algorithm is stopped when any one of the criterion is satisfied. i.e. when the maximum number of iterations have been exceeded or when there is a little changeor no change in the centroid vectors or when there are no changes in cluster membership.

4 LINEAR PRINCIPAL COMPONENT ANALYSIS

The k-means clustering algorithm loses its accuracy and effectiveness if the dataset or the dataobjects for clustering is large.Hence a technique should be employed for reducing the dimension of the data set.

Linear Principal Component analysis is a technique for reducing the dimension of large datasets. It is a second-order method and is based on covariance matrix of the variables.No assumptions have to be made in PCA .The observed variables are transferred to principal components of the

data set that consists of a large number of variables. It is a statistical method to determine key variables in a high dimension data set and then explain the difference in the observations that can be used for simplifying the dataset.

4.1 Steps involved in PCA

Step1: Make a input matrix Table

Step2: Compute the covariance matrix

Covariance matrix is calculated as

$$S = (xi - \mu_x)(xi - \mu_x)^T = YY^T$$

Where $\mu_x = \sum x/n$

Step3: Now Calculate eigen vectors and eigen values of this covariance matrix.The principal eigen vectors "u_K" of this covariance matrix are the principal directions of the data Y . The eigen vectors with the highest value are the principal components of the data set. In general, once when the eigen vectors are found from this covariance matrix, In the next step they are ordered by Eigen value, from highest to lowest. The first d (no. of principal components) eigen vectors are selected to rduce the dimension. Now the final data set has only d dimensions.

Step4: Choose the components and form a feature vector.

Step5: Derive the new dataset

5 PARTICLE SWARM OPTMIZATION

Particle swarm optimization is proposed by American social psychology James Kennedy and Russell Eberhart in 1995[4].It follows the simple basic idea that biotic population share information. The algorithm is easy to implement and converge rapidly. It can be applied when there is large number of samples. Each particle is a point of N-dimensional solution space and has a speed which is also a N-dimensional vector. Each particle has a fitness function(value) associated with it. Each particle adjusts its position and move closer to optimal point.

The position of i-th particle is denoted as: X_i □(x_{i1} , x_{i2} , , x_{iN}).The current velocity of particle is denoted as: V_i □(v_{i1} , v_{i2} , , v_{iN}) .Initially,the particles are initialized

with a position ,

$$x(0), x_{min} < x(0) < x_{max},$$

And the initial velocity is assigned to V(0) zero.

Next there is update of position and velocity of each particles which are classified as local and global update.

Equation-1 shows the local update of the position

$$y(t+1) = \begin{cases} y_i(t) \text{ if } f(x_i(t+1)) \geq f(y_i(t)) \\ x_i(t+1) \text{ otherwise,} \end{cases} \quad (1)$$

Where the local optimum of particles is denoted by ***y(t)*** and the fitness value(function) of the local optimization is denoted by ***f(x(t))***

Equation 2 – shows the global optimization of particle's position.

$$Y(t+1) = \begin{cases} y_i(t) \text{ iff } f(y_i(t+1)) \geq f(Y_i(t) \; \forall y_i(t)) \\ y_i(t) \text{ if } \exists y_i(t) | \; f(y_i(t)) < f(Y_i(t)) \end{cases} \quad (2)$$

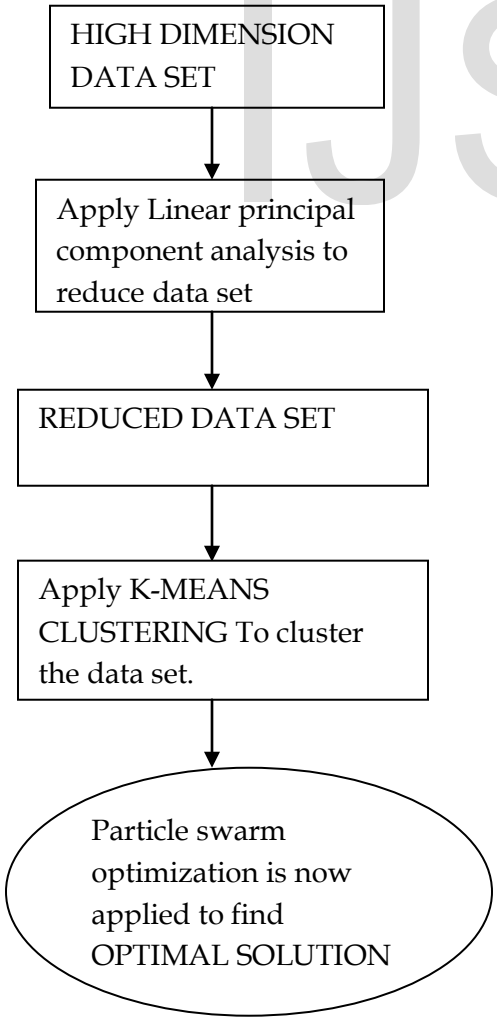
In each step, according to PSO algorithm given by Kennedy, particles update their velocity and position according to the given formula :

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (3)$$

$$v_i(t+1) = v_i(t) + c_1 r_1 \left(y_i(t) - x_i(t)\right) + c_2 r_2 (Yi(t) - x_i(t)) \quad (4)$$

Where c_1 and c_2 denote accelerating factor According to the experience of PSO algorithm, they are actually set c_1-c_2=2, r_1 and r_2 are two random numbers between zero and one. .A constant 'V max" is used to limit the speed of particles and improve search results.
This updating of particle's position and velocity continues until a stopping criterion is satisfied.

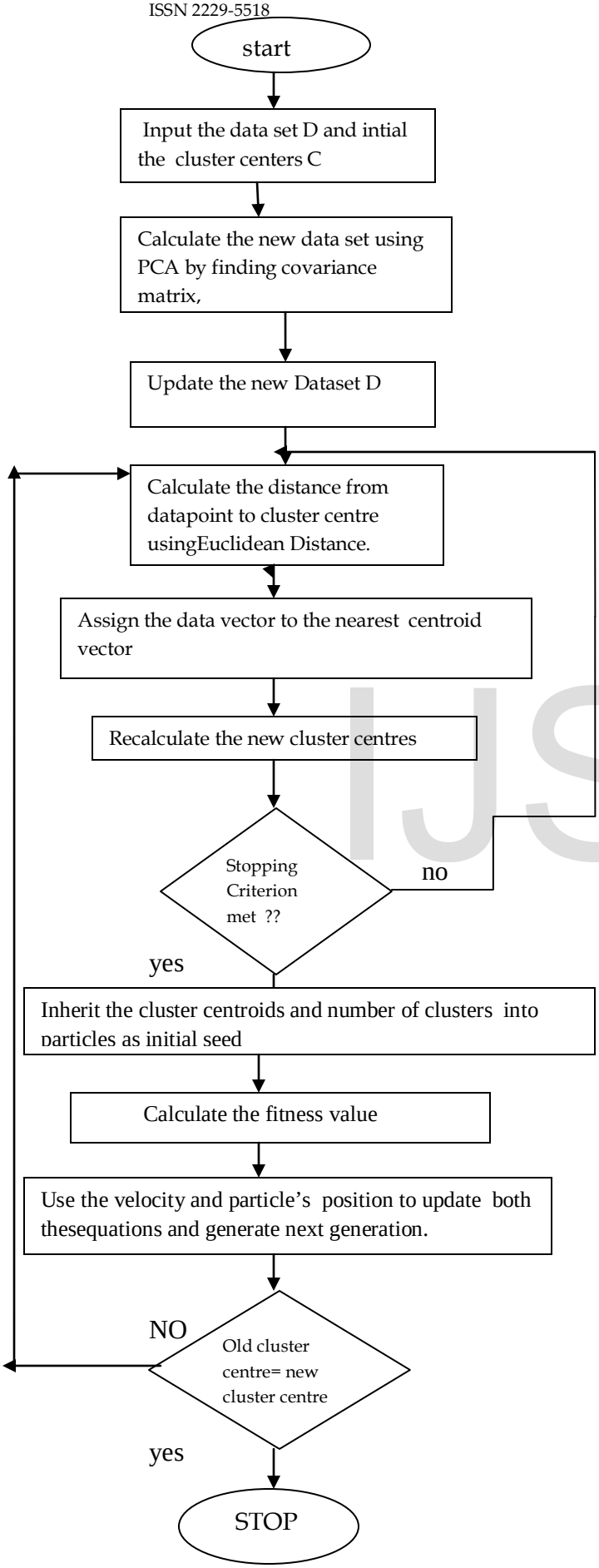
6 PROPOSED METHOD



The previous figure1 shows the proposed methodology for clustering high-dimensional data set and to find optimal solution using PSO.

7 PCA BASED HYBRIDIZED K-MEANS AND PSO

In this paper, a clustering algorithm based on popular swarm intelligent-based PSO technique and K-Means method has been proposed to cluster large data sets using PCA. The algorithm utilizes both global search capability of PSO and fast convergence capability of K-means in order to achieve global optima and faster convergence. The problem space is defined by the multidimensional vector space . In the problem space each vector is represented as a dot. The whole dataset is represented as a multiple dimension space with the large number of dots in the space. First the linear PCA is applied to the large data set.The reduced data set is the output from PCA.The output From PCA is the input to the k-means module. Then the hybrid K-Means + (PSO) clustering algorithm includes two modules, the k-means clustering module and PSO module. First the k-means module is executed. The k-means clustering process is stopped when maximum number of iterations have been exceeded. The result of k-means algorithm is used as particle for PSO algorithm and this information is being passed to the PSO Module for refininement and generation of the optimal clustering solution. The flow chart of the Proposed PCA based hybrid k-means + (PSO) clustering is depicted graphically in Figure 2.



8 CONCLUSION

Data clustering is one of the comprehensive KDD techniques, which is gaining more and more importance in optimization research. In the recent past various optimization techniques were used to improve one or another aspect of clustering analysis to achieve optimal results. Swarm intelligence is one of such optimization techniques. This paper investigated the application of the K-MEANS + PSO algorithm, which is a hybrid of PSO and K-means algorithms to cluster High dimension data vectors. First the PCA module is executed to convert high dimensional data to low dimension using covariance matrix. Then the K-means clustering algorithm is made to execute for searching for the cluster's centroid locations using the Euclidean distance similarity metric. This information is passed to the PSO module for the generation of the final optimal clustering solution as the result. In general, PSO conducts a global search for the optimal clustering, but more iteration numbers is required. The PSO is helped by k-means clustering to start with good initial cluster centroid that converge faster thereby giving a more compact result. The algorithm overall includes three modules, the PCA section, the k-means module and the PSO optimization module. The k-means module is executed first . The result from the k-means module is treated as the initial seed for the PSO module for discovering the optimal solution by a globalized search and also it avoids to consume high computation. The PSO algorithm will be applied for refinement and generation for the final result. Better clustering results from PCA based HYBRID (K-PSO) algorithm as compared to using ordinary PSO.

9 FUTURE SCOPE

The PSO still requires much investigation to improve performance and other key features that would make such algorithms suitable techniques for clustering high-dimensional data. Some future works and research trends to address the clustering of high-dimensional data are as follows: developing of multiple swarm cluster algorithms; dynamic PSO clustering algorithms; multi-objective PSO clustering algorithms; developing new fitness and measure functions; integration with feature selection and other techniques for

dimensionality reduction; developing new PSO variants algorithms for clustering, the use of new similarity measures; developing and use of a multi similarity measure with multiple swarm algorithms; sensitivity analysis of PSO parameters; new strategies to find the optimal number of the clusters without any prior knowledge; and the applications to different real-world problems.

REFERENCES

[1] Michael R. Anderberg , "Cluster Analysis for Applications". Academic Press Inc., New York, 1973.

[2] Mariam El -Tarabily, Rehab Abdel-Kader,Mahmoud Marie, Gamal Abdel-Azeem, "A PSO- Based Subtractive Data Clustering Algorithm " International Journal of Research in Computer Science, 3 (2): pp. 1-9, March 2013 doi:10.7815/ijorcs. 32.2013.060.

[3] J Kennedy, RC Eberhart, "Particle Swarm Optimization", Proceedings of The IEEE International Joint Conference on NeuralNetworks, Vol. 4, pp1942-1948, 1995.

[4] X.Cui, P.Palathingal, T.E.Potok, "Document Clustering using Particle Swarm Optimization ". IEEE Swarm Intelligence Symposium 2005, Pasadena, California, pp. 185 - 191. doi: 10.1109/SIS.2005.1501621

[5] J Kennedy, RC Eberhart, Y ***Shi***, "Swarm Intelligence", Morgan Kaufmann, 2002.

[6] Everitt, B. "Cluster Analysis". 2nd Edition, Halsted Press, New York, 1980.
Snásel, "Fuzzy Clustering using Hybrid Fuzzy c-means and Fuzzy Particle Swarm Optimization", World Congress on Nature & Biologically Inspired Computing, NaBIC 2009. In Proc. NaBIC, pp.1690-1694, 2009. doi: 10.1109/NABIC.2009.5393618

[7] A. K. Jain , M. N. Murty , P. J. Flynn, "Data Clustering :A Review". ACM Computing Survey,Vol. 31, No.3, pp: 264-3231999. doi:10.1145/331499.331504

[8] Selim, Shokri Z.,"K-means type algorithms: A generalized convergence theorem and characterization of local optimality". Pattern Analysis and MachineIntelligence,IEEE Transactions Vol. 6, No.1, pp:81–87,1984 doi: 10.1109/TPAMI.1984.4767478

[9] Abbas Ahmadi, Fakhri Karray, Mohamed Kamel,Multiple Cooperating Swarms for Data Clustering, In the Proceedings of IEEE Swarm Intelligence Symposium, pp. 206-212, 2007 .

[10] Mahamed G. Omran, Ayed Salman, Andries P. Engelbrecht, "Image classification using particle swarm optimization". Proceedings of the 4th Asia-Pacific Conference on Simulated Evolution and Learning 2002, Singapore, pp: 370-374. doi: 10.1142/9789812561794_0019

[11] K. Premalatha, A.M. Natarajan, "Discrete PSO with GA Operators for Document Clustering". International Journal of Recent Trends in Engineering, Vol. 1, No. 1, 2009 .

[12] H. Izakian, A. Abraham, and V. relevant clustering directions in high-dimensional data using particle swarm optimization.J Chemom25(7):366-374

[13] Marini F, Walczak B (2011) Finding relevant clustering directions in high-dimensional data using particle swarm optimization.J Chemom25(7):366-374

[

14] Vijay Kalivarapu, Jung-Leng Foo, Eliot Winer,"Improving solution characteristics of particle swarm optimization using digital pheromones" Structural and Multidisciplinary Optimization - STRUCT MULTIDISCIP OPTIM, Vol. 37, No. 4, pp: 415-427, 2009. doi: 10.1007/s00158-008-0240-9

[15] Neveen I. Ghali, Nahed El-dessouki, Mervat A. N, Lamiaa Bakraw, "Exponential Particle Swarm Optimization Approach for Improving Data Clustering" International Journal of Electrical & Electronics Engineering,Vol. 3, Issue 4, May2009.

[16] P.Prabhuet et al. "Improvising the performance of K-means clustering for high dimensional data set" International journal on computer science and engineering vol 3, Jun 2011.

.

.

.