Journal of Information Engineering and Applications
ISSN 2224-5782 (print) ISSN 2225-0506 (online)
Vol.3, No.11, 2013

The Improved K-Means with Particle Swarm Optimization

Mrs. Nidhi Singh, Dr.Divakar Singh
Department of Computer Science & Engg, BUIT,BU,Bhopal.India.(M.P)
nnidhi_sng14@yahoo.com divakar_singh@rediffmail.com

Abstract
In today's world data mining has become a large field of research. As the time increases a large amount of data is accumulated. Clustering is an important data mining task and has been used extensively by a number of researchers for different application areas such as finding similarities in images, text data and bio-informatics data. Cluster analysis is one of the primary data analysis methods. Clustering defines as the process of organizing data objects into a set of disjoint classes called clusters. Clustering is an example of unsupervised classification. In clustering, K-Means (Macqueen) is one of the most well known popular clustering algorithm. K-Means is a partitioning algorithm follows some drawbacks: number of clusters k must be known in advanced, it is sensitive to random selection of initial cluster centre, and it is sensitive to outliers. In this paper, we tried to improve some drawbacks of K-Means algorithm and an efficient algorithm is proposed to enhance the K-Means clustering with Particle Swarm Optimization. In recent years, Particle Swarm Optimization (PSO) has been successfully applied to a number of real world clustering problems with the fast convergence and the effectively for high-dimensional data.
Keywords: Clustering, K-Means clustering, PSO (Particle Swarm Optimization), Hierarchical clustering**.**

I. Introduction

A. Clustering

Data mining is the process of extracting patterns from large amount of data repositories or data bases. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. Cluster analysis or Clustering is the assignment of a set of objects into subsets (called clusters) so that objects in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique used in variety of fields including machine learning, data mining, pattern recognition, image analysis and bioinformatics [1]. A good clustering method will produce high quality of clusters with high intra-cluster similarity and low inter-cluster similarity.

B. Types of Clustering

Data Clustering algorithms mainly divided into two categories: Hierarchical and Partition algorithms. Hierarchical algorithm usually consists of either Agglomerative ("bottom-up") or Divisive ("top-down"). Agglomerative algorithms as we know is a bottom-up approach that is it begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms is a top-down approach and begin with the whole set and proceed to divide it into successively smaller clusters[1].Partition clustering algorithm partitions the data set into desired number of sets in a single step. Many methods have been proposed to solve clustering problem. One of the most popular and simple clustering method is K-Means clustering algorithm developed by Mac Queen in 1967.K-Means Clustering algorithm is an iterative approach; it partitions the dataset into k groups. Efficiency of K-Means algorithm heavily depends on selection of initial centroids because randomly choosing initial centroid also has an influence on number of iteration and also produces different cluster results for different cluster centroid [2].

C. Particle Swarm Intelligence

Swarm Intelligence (SI), was inspired by the biological behaviour of animals, and is an innovative distributed intelligent paradigm for solving optimization problems [3]. The two main Swarm Intelligence algorithms are (1) Ant Colony Optimization (ACO) and (2) Particle Swarm Optimization (PSO). This paper mainly deals with PSO. It is an optimization technique originally proposed by Kennedy and Eberhart [4] and was inspired by the swarm behaviour of birds, fish and bees when they search for food or communicate with each other.

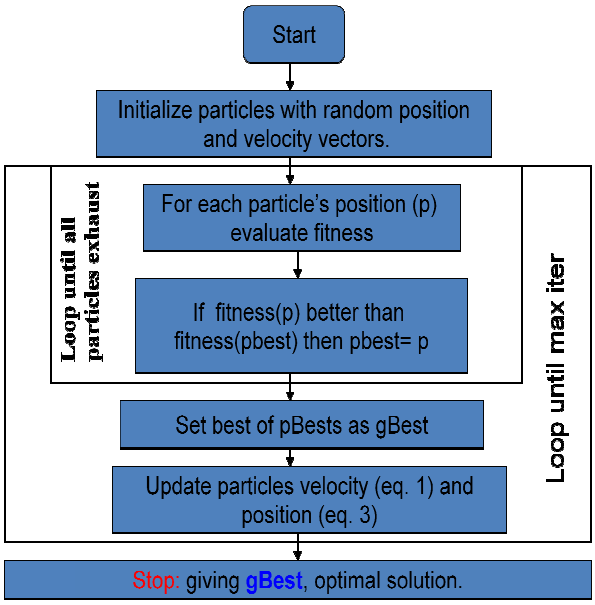

Figure 1.1. Basic flow diagram of PSO.

PSO approach is based upon cooperation and communication among agents called particles. Particles are the agents that represent individual solutions while the swarm is the collection of particles which represent the solution space. The particles then start moving through the solution space by maintaining a velocity value *V* and keeping track of its best previous position achieved so far. This position value is known as its personal best position. Global best is another best solution which is the best fitness value which is achieved by any of the particles. The fitness of each particle or the entire swarm is evaluated by a fitness function [7]. The flow chart of basic PSO is shown in Figure 1.1:

D. K-Means Clustering Algorithm

This part briefly describes the standard K-Means algorithm. K-Means is a typical clustering algorithm in Data Mining which is widely used for clustering large set of data's. In 1967, Mac Queen firstly proposed the K-Means algorithm, it was one of the most simple, unsupervised learning algorithm, which was applied to solve the problem of the well-known cluster [9].The most widely used Partitioning Clustering algorithm is K-Means. K-Means algorithm clusters the input data into k clusters, where k is given as an input parameter.

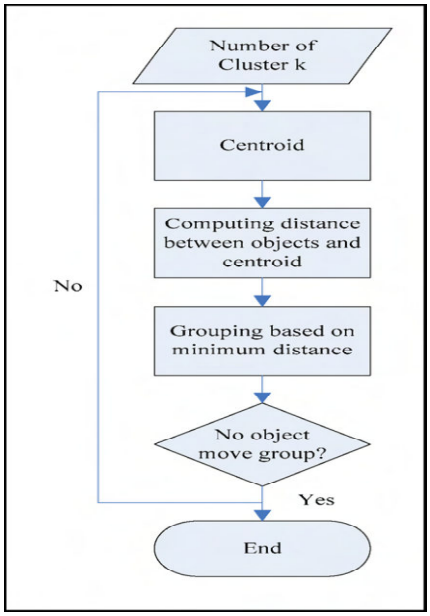

Figure 1.2 Process of K-Means algorithm

K-Means algorithm finds a partition that highly depends on the minimum squared error between the centroid of a cluster and its data points. The algorithm first takes k random data points as initial centroids and assigns each data point to the nearest cluster until convergence criteria is met. Although K-Means algorithm is simple and easy to implement, it suffers from three major drawbacks:
1) The number of clusters, k has to be specified as input.
2) The algorithm converges to local optima.
3) The final clusters depend on randomly chosen initial centroids.
4) High computational complexity
The flow diagram for simple K-Means algorithm is shown in Figure 1.2[11].

II. Related Work

Various researches have been carried out to improve the efficiency of K-Means algorithm with Particle Swarm Optimization. Particle Swarm Optimization gives the optimal initial seed and using this best seed K-Means algorithm produces better clusters and produces much accurate results than traditional K-Means algorithm.
A. M. Fahim et al. [5] proposed an enhanced method for assigning data points to the suitable clusters. In the original K-Means algorithm in each iteration the distance is calculated between each data element to all centroids and the required computational time of this algorithm is depends on the number of data elements, number of clusters and number of iterations, so it is computationally expensive.
K. A. Abdul Nazeer et al. [6] proposed an enhanced algorithm to improve the accuracy and efficiency of the K-Means clustering algorithm. In this algorithm two methods are used, one method for finding the better initial centroids. And another method for an efficient way for assigning data points to appropriate clusters with reduced time complexity. This algorithm produces good clusters in less amount of computational time.
Shafiq Alam [7] proposed a novel algorithm for clustering called Evolutionary Particle Swarm Optimization (EPSO)-clustering algorithm which is based on PSO. The proposed algorithm is based on the evolution of swarm generations where the particles are initially uniformly distributed in the input data space and after a specified number of iterations; a new generation of the swarm evolves. The swarm tries to dynamically adjust itself after each generation to optimal positions. The paper describes the new algorithm the initial implementation and presents tests performed on real clustering benchmark data. The proposed method is compared with K-Means clustering- a benchmark clustering technique and simple particle swarm clustering algorithm. The results show that the algorithm is efficient and produces compact clusters.
In this paper, Lekshmy P Chandran et al. [8] describes a recently developed Meta heuristic optimization algorithm named harmony search helps to find out near global optimal solutions by searching the entire solution space. K-Means performs a localized searching. Studies have shown that hybrid algorithm that combines the two ideas will produce a better solution.
In this paper, a new approach that combines the improved harmony search optimization technique and an enhanced K-Means algorithm is proposed.

III. Proposed Work

The proposed algorithm works in two phases. Phase I is Algorithm 3.1 describes the Particle Swarm Optimization and Phase II is Algorithm 3.2 describes the original K-Means Algorithm. The Algorithm 3.1 which gives better seed selection by calculating forces on each particle due to another in each direction and the total force on an individual particle. The output of Algorithm 3.1 is given as input to Algorithm 3.2 which generates the final clusters. The cluster generated by this proposed algorithm is much accurate and of good quality in comparison to K-Means algorithm.

Algorithm 3.1 Particle Swarm Optimization

Step 1.initialization of parameters// number of particles, velocity.
Step 2.select randomly three particles goals
- 2.1 Initial goal
- 2.2 Average goal
- 2.3 Final goal

Step 3. Repeat step 2 until finds which particle goal is optimal.
Step 4 .Do
- 4.1 The total force on one particle due to another, in X direction due to distance factor will be $FDX_{a,i}$ such that: $$FDx_{a,i} = \min\left[Q.\left(1/d_{a,i}\right)\cos\Phi\right]$$

Where,

$d_{a,i}$, is distance between the two particles, Φ is the angle with X-axis and Q is a constant.

4.2 In the same manner force in Y direction is calculated.

4.3 Now total force acted on a particle can be calculated as $FD_{xa} = \sum_{\substack{i=1 \\ i \neq a}}^{n} FDx_{a,i}$

Where n is the number of particles in the system.

Step 5.Similarity measure is calculated. A low value of squared error of corresponding values of data vector is considered as a measure of similarity.

$$\varepsilon = \sum_{j=1}^{m} \left(x_{j,1} - x_{j,i} \right)^2$$

Step 6.distance matrix V_c is calculated as $\vec{v}_c = \begin{bmatrix} \sum_{i=1}^{p} v_{i,1} \\ \sum_{i=1}^{p} v_{i,2} \\ . \\ . \\ \sum_{i=1}^{p} v_{i,j} \end{bmatrix} .1/n$

Step 7. Clusters are formed when one clusters combines with another. If the mean value of cluster having higher cluster ID is within 1**sigma* limit of the other cluster, the two clusters are merged together. *Sigma* is the standard deviation of distances of the particles from the mean value.

Step 8. Stop

Algorithm 3.2 K-Means Clustering Algorithm [6]

Require: *D = {d1, d2, d3, ..., di, ..., dn }* // Set of n data points.

Initial cluster centroids calculated from Algorithm 3.1.

Ensure: A set of *k* clusters.

Steps:

Step 1. Arbitrarily choose *k* data points from *D* as initial centroids;

Step 2. Repeat

Assign each point *d*i to the cluster which has the closest centroid;

Calculate the new mean for each cluster;

Step 3.Until convergence criteria are met.

The squared error can be calculated using Eq (3.1):

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} \left| x - x_i \right|^2 \quad \text{.....(3.1)}$$

The accuracy can be calculated using Precision and Recall:

Precision is the fraction of retrieved documents that are relevant to the search. It can be calculated using Eq(3.2).

Recall is the fraction of documents that are relevant to the query that are successfully retrieved. It can be calculated using Eq(3.3).

$$\text{Precision} = \frac{t_p}{t_p + f_p} \quad \text{.....(3.2)}$$

$$\text{Recall} = \frac{t_p}{t_p + f_n} \quad \text{.....(3.3)}$$

Where

t_p= true positive (correct result)

t_n= true negative (correct absence of result)
f_p= false positive (unexpected result)
f_n= false negative (Missing result)
Overall accuracy can be calculated using Eq (3.4).

$$\text{Accuracy}=\frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad \text{.....(3.4)}$$

IV. Result Analysis

The technique proposed in this paper is tested on different data sets such as breast cancer [10], Thyroid [10], E-coli [10], against different criteria such as accuracy, time, error rate, number of iterations and number of clusters. The same data sets are used as an input for the K-Means algorithm. Both the algorithms do not need to know number of clusters in advance. For the K-Means algorithm the set of initial centroids also required. The proposed method finds initial centroids systematically. The proposed methods take additional inputs like threshold values. The description of datasets is shown in Table 4.1.

Table4.1. Description of Datasets

Dataset	Number of attributes	Number of instances
Breast Cancer	9	286
e-coli	8	336
Thyroid	21	7200

The comparative analysis for different attributes like time, accuracy, error rate and number of iterations are tabulated in Table 4.2. Based on these attributes the performance of the K-Means and proposed algorithm are calculated for a particular threshold. In this proposed algorithm there is no need to explicitly define number of clusters K.

Table4.2. Performance Comparison of the Algorithms for Different Datasets for Threshold 0.45

Dataset	Time(in sec)		Error Rate		Iteration		Accuracy(%)	
	K-Means	Proposed Algorithm	K-Means	proposed Algorithm	K-Means	Proposed Algorithm	K-Means	proposed Algorithm
Breast Cancer	3.01	2.79	4.26	1.50	4	5	84.80	89.05
e-coli	4.83	4.01	4.99	2.14	5	6	86	90.62
Thyroid	4.80	4.32	3.97	1.27	4	5	84.63	88.84

The graphical result based on comparison shown in table 4.2 is shown in Figure 4.1, Figure 4.2, Figure 4.3, and Figure 4.4, with threshold value 0.45.
Figure 4.1 shows the comparison of different datasets with respect to Time. Results have shown that the proposed algorithm takes comparative less time as compared to K-Means.

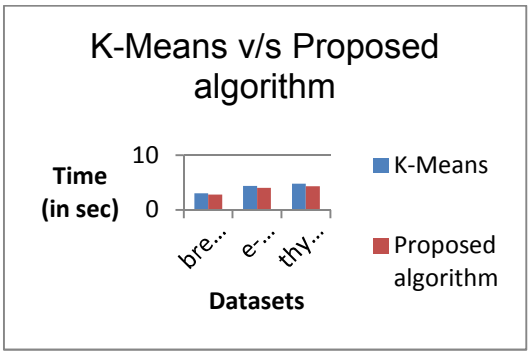

Figure 4.1. Comparison of different dataset with respect to Time at threshold 0.45

Figure 4.2 shows the error rate for different datasets, and the error rate for K-Means much larger than the proposed algorithm.

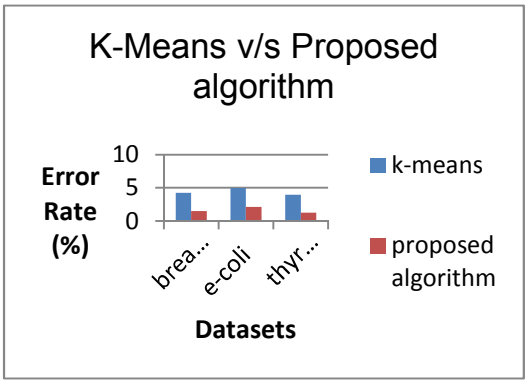

Figure 4.2 Comparison of different dataset with respect to Error rate at threshold 0.45

Figure 4.3 shows number of iterations for K-Means and proposed algorithm, the proposed algorithm take much number of iterations.

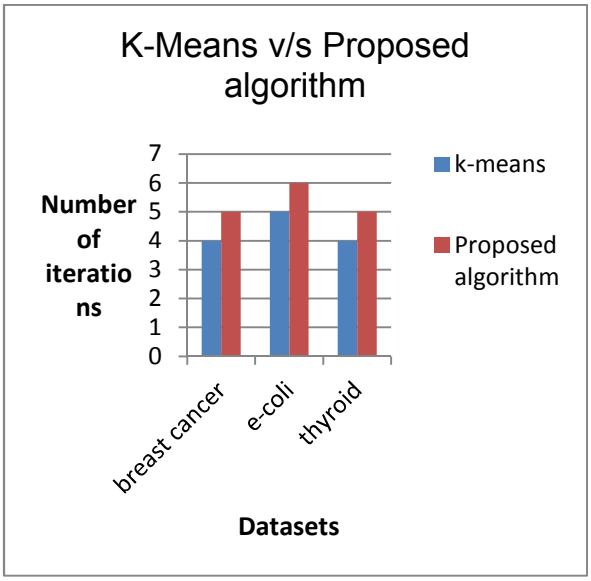

Figure 4.3 Comparison of different dataset with respect to number of iterations at threshold 0.45

Figure 4.4 shows the accuracy of K-Means and proposed algorithm, result shows that the proposed algorithm is much accurate than K-Means.

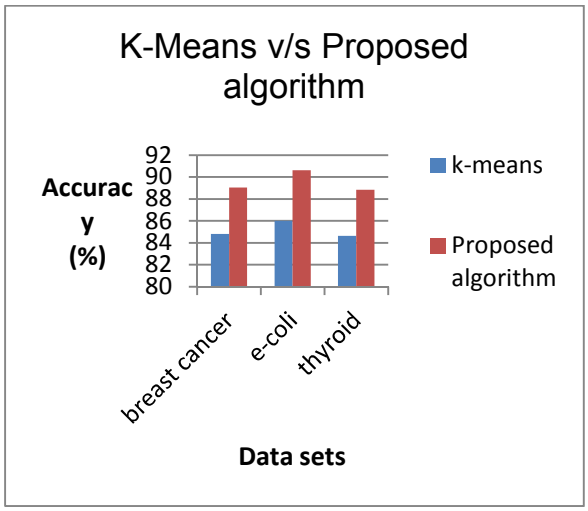

Figure 4.4 Comparison of different dataset with respect to accuracy at threshold 0.45

Matlab 7.8.0 is used for programming in experiment. Computer configuration is Intel Pentium 2.80GHz CPU, 256MB RAM.

V. Discussion

In this paper we have discussed the improved K-Means Clustering with Particle Swarm Optimization. One of the major drawback of K-Means Clustering is the random selection of seed, the random selection of initial seed results in different cluster which are not good in quality. The K-Means Clustering algorithm needs the steps (1) declaration of k clusters (2) initial seed selection (3) similarity matrix (4) cluster generation. The PSO algorithm is applied in step (2) the PSO gives the optimal solution for seed selection. In this paper, the standard K-Means is applied with different PSO to produce results which are more accurate and efficient than K-Means algorithm. In this paper there is no need to given k number of clusters in advance only a threshold value is required.

References

[1]. M.V.B.T.Santhi, V.R.N.S.S.V. Sai Leela, P.U. Anitha, D. Nagamalleswari, “Enhancing K-Means Clustering Algorithm”, IJCST Vol 2, Issue 4, Oct- Dec 2011.

[2]. Madhu Yedla, Srinivasa Rao Pathokota, T M Srinivasa, “Enhancing K-Means Clustering Algorithm with Improved Initial Center”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 1 (2), pp:121-125, 2010.

[3]. Ajith Abraham, He Guo, and Liu Hongbo, “Swarm Intelligence: Foundations, Perspectives and Applications”, Swarm Intelligent Systems, Nedjah N, Mourelle L (eds.),Nova Publishers, USA, 2006.

[4]. J. Kennedy, and R. C. Eberhart, “Particle Swarm Optimization,” Proc. Of IEEE International Conference on Neural Networks (ICNN), Vol. IV, Perth, Australia, 1942- 1948, 1995

[5]. A. M. Fahim, A. M. Salem, F. A. Torkey and M. A. Ramadan, “An Efficient enhanced K-Means clustering algorithm,” journal of Zhejiang University, 10(7): 16261633, 2006.

[6]. 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, July, London, UK, 2009.

[7]. Shafiq Alam, Gillian Dobbie, Patricia Riddle, "An Evolutionary Particle Swarm Optimization Algorithm for Data Clustering", Swarm Intelligence Symposium St. Louis MO USA, September 21-23, IEEE 2008.

[8]. Lekshmy P Chandran,K A Abdul Nazeer, “An Improved Clustering Algorithm based on K-Means and Harmony Search Optimization”, IEEE 2011.

[9]. Shi Na, Xumin Liu, Guan Yong, “Research on K-Means Clustering Algorithm -An Improved K-Means Clustering Algorithm”, Third International Symposium on Intelligent Information Technology and Security Informatics, pp: 63-67, IEEE, 2010.

[10]. The UCI Repository website. [Online].Available: http://archive.ics.uci.edu/,2010.

[11]. Juntao Wang,Xiaolong Su,"An improved K-Means clustering algorithm",IEEE,2011.