Research Journal of Applied Sciences, Engineering and Technology 8(11): 1355-1362, 2014
DOI:10.19026/rjaset.8.1107
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 14, 2014 Accepted: July 09, 2014 Published: September 20, 2014

Research Article

Distance Based Hybrid Approach for Cluster Analysis Using Variants of K-means and Evolutionary Algorithm

O.A. Mohamed Jafar and R. Sivakumar
Department of Computer Science, A.V.V.M. Sri Pushpam College (Autonomous), Poondi, Thanjavur, Tamil Nadu, India

Abstract: Clustering is a process of grouping same objects into a specified number of clusters. K-means and K-medoids algorithms are the most popular partitional clustering techniques for large data sets. However, they are sensitive to random selection of initial centroids and are fall into local optimal solution. K-means++ algorithm has good convergence rate than other algorithms. Distance metric is used to find the dissimilarity between objects. Euclidean distance metric is commonly used by number of researchers in most algorithms. In recent years, Evolutionary algorithms are the global optimization techniques for solving clustering problems. In this study, we present hybrid K-means++ with PSO technique (K++_PSO) clustering algorithm based on different distance metrics like City Block and Chebyshev. The algorithms are tested on four popular benchmark data sets from UCI machine learning repository and an artificial data set. The clustering results are evaluated through the fitness function values. We have made a comparative study of proposed algorithm with other algorithms. It has been found that K++_PSO algorithm using Chebyshev distance metric produces good clustering results as compared to other approaches.

Keywords: Cluster analysis, distance metrics, evolutionary algorithms, K-means, K-means++, K-medoids, particle swarm optimization

INTRODUCTION

With the fast development of information technology, huge amount of data collected from various fields has been stored electronically. The most challenging task of business analyst is to transform large volume of data stored in data warehouses into meaningful information called knowledge. Knowledge Discovery in Databases (KDD) is used to achieve this task. A part of KDD process is data mining. Data mining involves the use of data analysis techniques to discover previously unknown, valid patterns and relationship in large data sets. Clustering is one of the important data mining activities (Han and Kamber, 2001).

Cluster analysis is the process of grouping a set of data points in such a way that data points in the same group are more similar and data points from different groups are dissimilar. Clustering is called the unsupervised learning because there is no prior knowledge of patterns. The aim of clustering is to identify both dense and sparse regions in a data set. Clustering is used in many areas including pattern recognition, pattern analysis, artificial intelligence, image segmentation, image processing, bioinformatics, information retrieval and data mining and knowledge discovery. Therefore, it is an important research topic of diverse areas.

Data clustering can be broadly categorized into hierarchical methods, partitional methods, fuzzy clustering methods, hard clustering methods and model-based methods (Han and Kamber, 2001; Kaufman and Rousseeuw, 1990). *Hierarchical methods* create a hierarchical decomposition of the data points. They can be either top-down or bottom-up. Top-down algorithms start with one data point in a single cluster and then split into small groups until each data point is in one cluster. Bottom-up algorithms begin with each data point forming a separate cluster. They successively merge the data points that are close to one another, until all clusters are merged into one. *Partitional methods* partition the data set into predefined number of clusters. Given a data set of 'N' data points, they attempt to find 'k' groups, which satisfy the following requirements: each data point must belong to exactly one group and each group must contain at least one data point. In *fuzzy clustering methods*, each data point can belong to more than one cluster. The membership values are associated with each of the data points. The values lie between 0 and 1. In *hard clustering methods*, each data point can belong to only one cluster. *Model-based methods hypothesize* a model for each of the clusters and find the best fit of the data to the given model. They can be

Corresponding Author: O.A. Mohamed Jafar, Department of Computer Science, A.V.V.M. Sri Pushpam College (Autonomous), Poondi, Thanjavur, Tamil Nadu, India

either hierarchical or partitional depending on the structure.

A broad review of the important clustering algorithms can be found in the literature (Jain and Dubes, 1998; Berkhin, 2002; Xu and Wunsch II, 2005). K-means algorithm was proposed by MacQueen (1967). It is a center-based clustering method. K-medoids algorithm (Han and Kamber, 2001; Kaufman and Rousseeuw, 1990) uses the most representative data points called medoids instead of centroids. K-means and K-medoids algorithms are the most popular and widely used partitional data clustering methods. However, they are easily struck at local optimal solution and are sensitive to random selection of initial centers. The number of clusters also must be known in advance. K-means++ (Arthur and Vassilvitskii, 2007) is one of the variants of K-means algorithm which uses a new technique of selecting initial centroids by random initial centers with specific probabilities. The new seeding method has better performance and convergence rate than other algorithms. In recent years, evolutionary algorithms (Yu and Gen, 2010) like Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) have been used to solve wide range of optimization problems including data mining tasks. They avoid the drawbacks of variants of K-means algorithms. The PSO algorithm was first proposed by Kennedy and Eberhart (1995). It has been successfully applied to solve clustering problems by the research community. It is a population-based global optimization technique (Chen and Fun, 2004).

Recently, hybrid techniques are more popular for solving variety of real-world optimization problems. Euclidean distance metric is traditionally applied for several clustering algorithms in the literature. In this study, we have made an attempt to study the performance of algorithms using other important distance metrics such as City Block and Chebyshev. Cluster analysis based on K-means++ and PSO algorithm (K++_PSO) is proposed in this research using different distance metrics. Through fitness function values, it is shown that K++_PSO algorithm reports good clustering result on four benchmark data sets such as teaching assistant evaluation, thyroid, seeds, breast cancer and an artificial data set for Chebyshev distance metric.

LITERATURE REVIEW

Omran *et al*. (2002) proposed a new image classification algorithm based on particle swarm optimization. Van der Merwe and Engelbrecht (2003) proposed two new methods for clustering data. Esmin *et al*. (2008) proposed new data clustering approaches using particle swarm optimization. Tsai and Kao (2010) developed a novel data clustering algorithm based on Particle Swarm Optimization with Selective Regeneration (SRPSO) which includes features, unbalanced parameter setting and particle regeneration operation. Mohamed Jafar and Sivakumar (2013) presented a study of particle swarm optimization algorithm to data clustering using different distance metrics.

Bandyopadhyay and Maulik (2002) presented an evolutionary technique based on K-means algorithm called KGA-clustering. This algorithm utilizes the searching capability of K-means and avoids the drawback of getting stuck at local optimization. Ye and Chen (2005) developed the hybrid PSO and K-means algorithm, called Alternative KPSO-clustering (AKPSO). They presented an evolutionary particle swarm optimization learning-based method to optimally cluster N data points into K clusters. Dong and Qi (2009) proposed a new hybrid clustering algorithm based on particle swarm optimization and K-means. The algorithm generates better solution than PSO and K-means algorithms. Yang *et al*. (2009) proposed a hybrid data clustering algorithm based on PSO and K-Harmonic Means (KHM). The performance of the proposed algorithm was compared with PSO and KHM clustering algorithms with different data sets. Kao and Lee (2009) presented a new dynamic data clustering algorithm based on K-means and particle swarm optimization, called KCPSO. Rana *et al*. (2010) presented a hybrid sequential approach for data clustering using K-means and particle swarm optimization. The proposed algorithm avoids the limitations of both algorithms. Niknam and Amiri (2010) proposed an efficient hybrid approach based on PSO, ACO and K-means algorithms, called PSO-ACO-K approach for cluster analysis. Danesh *et al*. (2011) proposed a data clustering algorithm based on an efficient hybrid of K-Harmonic Means, PSO and GA. The hybrid algorithm helps to solve the local optima problem and overcomes the limitation of slow convergence speed. Chuang *et al*. (2012) proposed an improved particle swarm optimization based on Gauss chaotic map for clustering. They used the intra-cluster distance as a measure to search data cluster centroids. Li *et al*. (2013) proposed the K-means clustering algorithm based on Chaos Particle Swarm (CPSOKM). The proposed algorithm solves the problem of K-means algorithm and optimizes the clustering result. Sethi and Mishra (2013) developed a linear Principle Component Analysis (PCA) based hybrid K-means clustering and Particle Swarm Optimization (PSO) algorithm (PCA-K-PSO). The algorithm uses the global searching ability of PSO and fast convergence of K-means algorithm. Aghdasi *et al*. (2014) proposed K-harmonic data clustering algorithm using combination of PSO and Tabu Search.

Basic concepts: In this section, the concept of mathematical model of clustering problem, evolutionary algorithms and distance metrics are discussed.

Mathematical model of clustering problem: The mathematical model of clustering problem (Liu *et al*.,

2006) is described as follows: For a given data set of 'n' points, we have to allocate each data point to one of the 'k' clusters such that the sum of the Squared Euclidean Distance between data point and center of its belonging cluster should be minimum:

$$\text{Minimize } \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij} \| x_i - c_j \|^2 \quad (1)$$

where,

$$\sum_{j=1}^{k} w_{ij} = 1 \quad (2)$$

$$c_j = \frac{1}{N_j}\sum_{x_i \in C_j} x_i \quad (3)$$

where,
n = The number of data points
k = The number of clusters
w_{ij} = nxk 0-1 matrix
x_i = The location of the i-the data point
c_j = The center of the j-th cluster
N_j = The number of data points belonging to the cluster c_j

Evolutionary algorithms: Evolutionary algorithms (Eberhart and Shi, 2001) are stochastic optimization methods for solving real-life problems. Recently, many researchers have extensively used the evolutionary algorithms including Particle Swarm Optimization, Ant Colony Optimization, Tabu Search, Genetic Algorithm, Artificial Immune Systems, Differential Evolution and Simulated Annealing for solving wide range of real-world optimization problems. The important benefits of evolutionary algorithms are flexibility, communication, cooperation and self organization. The key application areas of evolutionary algorithms include classification, clustering, planning and decision making.

Particle swarm optimization: Particle swarm optimization was introduced by Kennedy and Eberhart (1995). It is based on the social behavior of a school of fish, a bacteria modeling, a flock of birds or a swarm of bees (Poli *et al*., 2007). In PSO system, the individuals are referred as particles. A population or swarm is a collection of particles. It is denoted by $P = (p_1, p_2, ..., p_n)$. Each particle flies through the search space, dynamically altering its position and velocity in the search space according to its own experience and that of neighboring particles. Therefore, particles tend to fly toward better and better searching areas. A predefined fitness function is used to measure the performance of a particle. Each particle maintains a memory of its previous best position, called *pbest* or local best (P_i). The best one among all the particles in the swarm is called *gbest* or global best (P_g). The position of the i-th particle and the velocity of the i-th particle are given by $X_i = (X_{i1}, X_{i2}, ...X_{id})$ and $V_i = (V_{i1}, V_{i2}, ...V_{id})$ respectively. The local best position and the global best position are represented as; $P_i = (p_{i1}, p_{i2}, ...p_{id})$ and $P_g = (p_{g1}, p_{g2}, ...p_{gd})$ respectively in a D-dimensional search space. The positions and velocities are adjusted and the fitness function is computed with new coordinates at each time step. The velocity and position of a particle are modified in each iteration, based upon its own *pbest* and *gbest*. The velocity update formula is calculated by the Eq. (4):

Table 1: Description of PSO parameters

Parameter	Description
d	Dimension, $d \in \{1, 2, ..., D\}$
n	Population size
i	Index, $i \in \{1, 2, ..., n\}$
ω	Inertia weight
c_1	Cognition component
c_2	Social component
r_1 and r_2	Uniformly generated random numbers from (0, 1)
V_{id}	Velocity of particle i on dimension d
X_{id}	Current position of particle i on dimension d
p_{id}	Personal best position of particle i on dimension d
p_{gd}	Global best position of particle i on dimension d

$$V_{id}^{'} = \omega V_{id} + c_1 r_1 (p_{id} - X_{id}) + c_2 r_2 (p_{gd} - X_{id}) \quad (4)$$

The position of the particle is updated using the Eq. (5):

$$X_{id}^{'} = X_{id} + V_{id}^{'} \quad (5)$$

The description of various parameters is shown in Table 1. Each particle X in the PSO system is constructed as follows:

$$X_i = (m_{i1}, m_{i2}, ...m_{ij}, ..., m_{iN_c}) \quad (6)$$

where,
m_{ij} = The j-th cluster center vector of the i-th particle in cluster C_{ij}
N_c = The total number of clusters

A swarm is a set of particles. Therefore, a swarm represents a number of candidate clustering solutions for a data set.

The fitness function value of the cluster analysis is calculated by the Eq. (7):

$$\sum_{j=1}^{N_c}\sum_{\forall z_p C\, C_{ij}} \mathrm{d}\left(\mathrm{z_p}, \mathrm{m_j}\right) \quad (7)$$

Table 2: List of different distance metrics

Distance metric	Formula
Euclidean	$d(x_i, z_j) = \sqrt{\sum_{k=1}^{d} (x_{i,k} - z_{j,k})^2}$
City block	$d(x_i, z_j) = \sum_{k=1}^{d} \lvert x_{i,k} - z_{j,k} \rvert$
Chebyshev	$d(x_i, z_j) = \max_{i=1,2...n} \lvert x_{i,k} - z_{j,k} \rvert$ k = 1, 2, …d

The fitness function value should be minimized.

Distance metrics: Distance metrics are used to determine the similarity or dissimilarity between two objects. They play a vital role in clustering data objects. The distance between two objects x_i and x_j is denoted by d (x_i, x_j). The important properties of distance metrics are (Gan *et al*., 2007):

- $d(x,y) \geq 0$, ∀ x and y
- $d(x,y) = 0$ only if x = y
- $d(x,x) = 0$, ∀ x
- $d(x,y) = d(y,x)$, ∀ x and y
- $d(x,z) \leq d(x,y) + d(y,z)$, ∀ x, y and z

The various distance metrics and their formula are shown in Table 2.

METHODOLOGY

In this section, K-means clustering algorithm, K-medoids clustering algorithm and Hybrid algorithm are described.

K-means clustering algorithm: The aim of clustering is to classify the given data set $X = \{x_1, x_2 ... x_N\}$ into set of clusters satisfying the following conditions (Niknam and Amiri, 2010):

- $z_i \neq \emptyset$, i = 1, 2,…c
- $z_i \cap z_j = \emptyset$, i, j = 1, 2,…c, i ≠ j
- $\cup_{i=1}^{c} z_i = X$

Given a set of 'N' data points and the number of clusters 'c', the objective is to select 'c' cluster centers so as to minimize the mean squared distance. It generates the fast solution. The K-means clustering algorithm is described as follows.

Input: Data set $X = \{x_1, x_2 ... x_N\}$, a set of data points; select the number of cluster centers $1 < c < N$; Initialize the random cluster centers selected from the data set.

Output: Cluster centers $z = \{z_1, z_2 ... z_c\}$

Repeat: For p = 1, 2 …..

Step 1: Compute the selected distance of each object in the data set from each of cluster centroids.
Step 2: Select the points for a cluster with the minimal distances, they belong to that cluster.
Step 3: Calculate the cluster centers:

$$z_i^{(p)} = \frac{\sum_{k=1}^{N_i} x_k}{N_i}, \text{ i = 1, 2, …c} \quad (8)$$

where, N_i is the number of data points in the i-th cluster until $\prod_{j=1}^{n} \max \lvert z^{(p)} - z^{(p-1)} \rvert \neq 0$

K-medoids clustering algorithm: In this algorithm, the centers are located among the data points themselves. A medoid is defined as the data point of a cluster, whose mean dissimilarity to all the data points in the cluster is minimum.

Input: Data set $X = \{x_1, x_2 ... x_N\}$, a set of data points; select the number of cluster centers $1 < c < N$; Initialize the random cluster centers selected from the data set.

Output: Cluster centers $z = \{z_1, z_2 ... z_c\}$

Step 1: Choose c objects at random to be the initial cluster centroids.
Step 2: Assign each object to the cluster associated with the closest cluster centers.
Step 3: Recalculate the positions.

Finding the object i within the cluster that minimizes:

$$\sum_{j \in C_i} d(i,j) \quad (9)$$

where, C_i is the cluster containing the object i and d (i, j) is the distance between object i and j.

Step 4: Repeat step 2 and 3 until converges.

Hybrid algorithm: Hybrid algorithms are the integration of two or more optimization techniques. Nowadays, hybrid algorithms are popular due to capability in handing various real-world applications that involve uncertainty and complexity. They make use of qualities of individual algorithms. In this study, we have combined the K-means++ and particle swarm optimization algorithm, called (K++_PSO) for cluster analysis. Euclidean distance is the commonly used metric in most of the clustering algorithms. We have

also made an attempt to study the performance of various algorithms with different distance metrics such as City Block and Chebyshev.

Description of K++_PSO algorithm:

Input: Data set $X = \{x_1, x_2 ... x_N\}$, a set of data points; select the number of cluster centers $1 < c < N$

Output: Cluster centers $z = \{z_1, z_2 ... z_c\}$

Step 1a): Select an initial center z_1 uniformly at random from the data set X
Step 1b): While |z| <c do

Choose the next center z_i randomly from X, where every $x \in X$ has a probability of:

$$\frac{d^2(x,z)}{\sum_{x \in X} \min_{i=1,2,...c} \| x - z_i \|^2} \quad (10)$$

of being selected.
end While

Step 2a): For iter = 1 to max_it do
Step 2b): Compute the selected distance of each object in the data set from each of cluster centroids of Step 1
Step 2c): Select the points for a cluster with the minimal distances, they belong to that cluster
Step 2d): Calculate the new cluster centers using:

$$z_i^{(p)} = \frac{\sum_{k=1}^{N_i} x_k}{N_i}$$

where N_i represents the number of data points in the i-th cluster.

Step 2e): Interchange the new cluster centers to old cluster centers
Step 3: The final cluster centers of step 2 to be taken as the initial cluster centers for particle 1 and N_c randomly selected cluster centroids for remaining particles
Step 4: For t = 1 to max_it do
Step 5: For each particle i do
Step 6: For each data vector z_p:

- Calculate selected distance d (x_p, m_{ij}) to all cluster centroids C_{ij}
- Assign z_p to cluster C_{ij} such that distance d (z_p, m_{ij}) = $\min_{\forall c = 1, ... ,N_c} \{d(z_p, m_{ic})\}$
- Calculate the fitness value (intra-cluster distance) using the Eq. (7)

Step 7: Update the global best and local best positiosns
Step 8: Update the cluster centroids using the Eq. (4) and (5)

EXPERIMENTAL RESULTS AND DISCUSSION

We compare the performance of the proposed hybrid algorithm with other clustering algorithms on four benchmark UCI machine learning repository data sets (http://archive.ics.uci.edu/ml/) which include data sets of teaching assistant evaluation, thyroid, seeds, breast cancer and an artificial data set.

The *teaching assistant evaluation* data set consists of 151 objects and 3 different types of classes characterized by 5 features. The data consist of evaluations of teaching performance over three regular semesters and two summer semesters of 151 teaching assistant assignment at the Statistics Department of the University of Wisconsin-Madison. The scores were divided into 3 roughly equal-sized categories ("low", "medium" and "high") to form the class variable.

The *thyroid dataset* consists of 215 instances. Each instance has 5 features including T3-resin uptake test, total serum thyroxin, total serum triiodothyronine, basal Thyroid-Stimulating Hormone (TSH) and maximal absolute difference of TSH value after injection of 200 micro grams of thyrotropin-releasing hormone as compared to the basal value. Each of the samples has to be categorized into one of the three classes: Class 1: normal (150 instances), Class 2: hyper (35 instances), Class 3: hypo functioning (30 instances).

The seeds data set contains 210 patterns belonging to 3 different varieties of wheat: Kama, Rosa and Canadian. Each pattern has 7 geometric parameters of wheat kernels such as area, perimeter, compactness, length of kernel, width of kernel, asymmetry coefficient and length of kernel groove.

The breast cancer data set consists of 683 records characterized by 9 features such as clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. The two categories are benign cases (239 records) and malignant cases (444 records).

Artificial data set: In this data set, there are 5 classes and each class has 50 samples consisting of 3 features. Each feature of the class is distributed according to Class 1~Uniform (80, 100); Class 2~Uniform (60, 80); Class 3~Uniform (40, 60); Class 4~Uniform (20, 40); and Class 5~Uniform (1, 20). The Characteristics of above mentioned data sets are shown in Table 3.

Table 3: Characteristics of selected data sets

Data set	Sample size	No. of classes	No. of features	Size of classes
Teaching assistant evaluation	151	3	5	(49, 50, 52)
Thyroid	215	3	5	(150, 35, 30)
Seeds	210	3	7	(70, 70, 70)
Breast cancer	683	2	9	(239, 444)
Artificial	250	5	3	(50, 50, 50, 50, 50)

Table 4: Comparison of fitness function value for the seven clustering algorithms on teaching assistant evaluation data set

Distance	K-means	K-med	K++	PSO	K_PSO	K-med_PSO	K++_PSO
Euclidean	1505.562	1532.526	1505.562	1505.121	1499.192	1504.400	1494.048
City block	2366.832	2460.000	2366.626	2338.159	2209.711	2216.166	2184.582
Chebyshev	1253.934	1285.000	1230.359	1228.708	1216.683	1218.455	1211.850

Table 5: Comparison of fitness function value for the seven clustering algorithms on thyroid data set

Distance	K-means	K-med	K++	PSO	K_PSO	K-med_PSO	K++_PSO
Euclidean	2001.636	2027.247	2001.636	2250.458	1962.502	1964.722	1930.333
City block	2985.348	3062.700	2985.348	3463.440	2929.856	2944.442	2925.505
Chebyshev	1678.176	1691.800	1678.176	1752.753	1632.316	1639.195	1622.335

Table 6: Comparison of fitness function value for the seven clustering algorithms on seeds data set

Distance	K-means	K-med	K++	PSO	K_PSO	K-med_PSO	K++_PSO
Euclidean	313.217	315.989	313.217	338.982	312.161	312.949	312.159
City block	545.621	552.858	544.590	672.034	543.683	546.218	543.589
Chebyshev	261.505	264.672	261.501	286.535	258.016	258.414	257.987

Table 7: Comparison of fitness function value for the seven clustering algorithms on breast cancer data set

Distance	K-means	K-med	K++	PSO	K_PSO	K-med_PSO	K++_PSO
Euclidean	2988.428	3089.114	2988.428	3741.141	2967.178	3031.565	2966.431
City block	7326.375	6555.000	7326.375	8243.116	6512.534	6518.526	6454.468
Chebyshev	1933.127	2105.000	1933.127	2179.059	1886.595	1980.579	1880.628

Table 8: Comparison of fitness function value for the seven clustering algorithms on artificial data set

Distance	K-means	K-med	K++	PSO	K_PSO	K-med_PSO	K++_PSO
Euclidean	2293.511	2342.943	2293.511	3779.582	2291.415	2299.133	2290.905
City block	3547.000	3620.000	3547.000	5107.456	3538.582	3562.208	3535.108
Chebyshev	1828.260	1875.000	1828.260	2693.166	1817.595	1831.180	1813.231

The algorithms perform best under the following selected parameter values: The number of particles (p) is set to 10. The cognitive component (c_1) and social component (c_2) are set to 2.0. The inertia weight (ω) is 0.9→0.4. ω decreases linearly from 0.9 to 0.4 throughout the search process. ω is calculated by the following Eq. (11):

$$\omega = \omega_{max} - \frac{\omega_{max} - \omega_{min}}{I_{max}} * I \quad (11)$$

where, ω_{max} and ω_{min} are the initial and final value of weighting coefficient, respectively; ω_{max} = 0.9 and ω_{min} = 0.4; I_{max} is the maximum number of iterations; I is the current iteration number. The maximum number of iterations is 100. The experiments are conducted through 10 independent runs for all the algorithms. The iteration error (ε) is 0.00001.

The aim of this paper is to study the effect of hybrid algorithm for data clustering using different distance metrics. Clustering algorithms are implemented using Java. For conducting various experiments, we used a PC Pentium IV (CPU 3.06 GHZ and 1.97 GB RAM) with the selected parameter values. Each algorithm is tested through 100 iterations and 10 independent runs. In this study, the quality of clustering of data clustering algorithms is measured by fitness function values. Table 4 to 8 present a comparison among the results of different clustering algorithms on selected data sets in terms of fitness function values.

Fitness function values: The distance between each data point and within a cluster and the cluster center of that cluster is computed and added up. It is calculated by using the Eq. (12):

$$\sum_{j=1}^{K} \sum_{x_i \in C_j} d(x_i, c_j) \quad (12)$$

where, $d(x_i, c_j)$ is the distance between the data point x_i and the cluster center c_j. The minimum function value indicates the higher quality of clustering. Table 4 to 8 show that the proposed algorithm has the minimum function values 1494.048, 2184.582 and 1211.850 on teaching assistant evaluation data set; 1930.333, 2925.505 and 1622.335 on thyroid data set; 312.159, 543.589 and 257.987 on seeds data set; 2966.431, 6454.468 and 1880.628 on breast cancer data set; 2290.905, 3535.108 and 1813.231 on artificially

generated data set for Euclidean, City block and Chebyshev distance metrics, respectively.

Hence, the K++_PSO hybrid algorithm has better performance than other clustering algorithms in terms of fitness function values. It is also observed that the proposed algorithm using Chebyshev distance produces better result than those of other distance metrics.

CONCLUSION

Clustering is a NP complete problem which group the data points that are more similar to one another than to members of other clusters. The K-means and K-medoids algorithms are easily trapped in local minimum and are sensitive to initial values and noisy environment. K-means++ algorithm produces better performance than other algorithms. PSO is a population-based stochastic optimization algorithm. The hybrid algorithm improves the performance of clustering results. Euclidean distance is commonly applied in many data clustering algorithms. In this study, K++_PSO algorithm is proposed using different distance metrics including City Block and Chebyshev. The performance of different algorithms is evaluated through fitness function values. The proposed algorithm is compared with other clustering algorithms on four benchmark data sets such as teaching assistant evaluation, thyroid, seeds, breast cancer and an artificial data set using different distance metrics. Experimental results show that the K++_PSO algorithm has better clustering result in terms of fitness function value as compared to other algorithms: K-means, K-medoids, K-means++, PSO, K-PSO, K-med_PSO. It is also recorded that the proposed algorithm produces good performance for the Chebyshev distance than other distance metrics.

REFERENCES

Aghdasi, T., J. Vahidi and H. Motameni, 2014. K-harmonic means data clustering using combination of particle swarm optimization and tabu search. Int. J. Mech. Electr. Comput. Technol., 4(11): 485-501.

Arthur, D. and S. Vassilvitskii, 2007. K-means++: The advantages of careful seeding. Proceeding of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, pp: 1027-1035.

Bandyopadhyay, S. and U. Maulik, 2002. An evolutionary technique based on K-means algorithm for optimal clustering in R^n. Inform. Sci., 146: 221-237.

Berkhin, P., 2002. Survey of clustering data mining techniques. Technical Report, Accrue Software, San Jose, California.

Chen, C.Y. and Y. Fun, 2004. Particle swarm optimization algorithm and its application to clustering analysis. Proceeding of IEEE International Conference on Networking Sensing and Control, 2: 789-794.

Chuang, L.Y., Y.D. Lin and C.H. Yang, 2012. An improved particle swarm optimization for data clustering. Proceeding of International MultiConference of Engineers and Computer Scientists (IMECS, 2012). Hong Kong, Vol. 1, March 14-16.

Danesh, M., M. Naghibzadeh, M.R.A. Totonchi, M. Danesh, B. Minaei and H. Shirgahi, 2011. Data clustering based on an efficient hybrid of K-harmonic means, PSO and GA. In: Nguyen, N.T. (Ed.), Transactions on CCI IV. LNCS 6660, Springer-Verlag, Berlin, Heidlberg, pp: 125-140.

Dong, J. and M. Qi, 2009. A new algorithm for clustering based on particle swarm optimization and K-means. Proceeding of International Conference on Artificial Intelligence and Computational Intelligence (AICI'09), pp: 264-268.

Eberhart, R.C. and Y. Shi, 2001. Particle swarm optimization: Developments, applications and resources. Proceeding of the 2001 Congress on Evolutionary Computation, 1: 81-86.

Esmin, A.A.A., D.L. Pereira and F. de Araujo, 2008. Study of different approach to clustering data by using the particle swarm optimization algorithm. Proceeding of the IEEE World Congress on Evolutionary Computation, pp: 1817-1822.

Gan, G., C. Ma and J. Wu, 2007. Data Clustering: Theory, Algorithms and Applications. SIAM, Philadelphia, PA.

Han, J. and M. Kamber, 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco. Retrieved form: http:// archive.ics.uci.edu/ml/.

Jain, A. and R. Dubes, 1998. Algorithms for Clustering Data. Prentice Hall, New Jersey.

Kao, Y. and S.Y. Lee, 2009. Combining K-means and particle swarm optimization for dynamic data clustering problems. Proceeding of the IEEE International Conference on Intelligent Computing and Intelligent System, pp: 757-761.

Kaufman, L. and P.J. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons Inc., New York.

Kennedy, J. and R. Eberhart, 1995. Particle swarm optimization. Proceeding of IEEE International Conference on Neural Networks. Piscataway, NJ, 4: 1942-1948.

Li, Y.R., Z.Y. Yong and Z.C. Na, 2013. The K-means clustering algorithm based on chaos particle swarm. J. Theor. Appl. Inform. Technol., 48(2): 762-767.

Liu, Y., J. Peng, K. Chen and Y. Zhang, 2006. An improved hybrid genetic clustering algorithm. In: Antoniou, G. *et al*. (Eds.), SETN 2006. LNAI 3955, Springer-Verlag, Berlin, Heidlberg, pp: 192-202.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceeding of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp: 281-297.

Mohamed Jafar, O.A. and R. Sivakumar, 2013. A study of bio-inspired algorithm to data clustering using different distance measures. Int. J. Comput. Appl. (IJCA), 66(12): 33-44.

Niknam, T. and B. Amiri, 2010. An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis. Appl. Soft Comput., 10(1): 183-197.

Omran, M., A. Salman and A.P. Engelbrecht, 2002. Image classification using particle swarm optimization. Proceeding of the 4th Asia-Pacific Conference on Simulated Evolution and Learning. Singapore, pp: 370-374.

Poli, R., J. Kennedy and T. Blackwell, 2007. Particle swarm optimization-an overview. Swarm Intell., 1(1): 33-57.

Rana, S., S. Jasola and R. Kumar, 2010. A hybrid sequential approach for data clustering using K-means and particle swarm optimization algorithm. Int. J. Eng. Sci. Technol., 2(6): 167-176.

Sethi, C. and G. Mishra, 2013. A linear PCA based hybrid K-means PSO algorithm for clustering large dataset. Int. J. Sci. Eng. Res., 4(6): 1559-1566.

Tsai, C.Y. and I.W. Kao, 2010. Particle swarm optimization with selective particle regeneration for data clustering. Expert Syst. Appl., 38: 6565-6576.

Van Der Merwe, D.W. and A.P. Engelbrecht, 2003. Data clustering using particle swarm optimization. Proceeding of the IEEE Congress on Evolutionary Computation. Canberra, Australia, pp: 215-220.

Xu, R. and D. Wunsch II, 2005. Survey of clustering algorithms. IEEE T. Neural Networ., 16(3): 645-678.

Yang, F., T. Sun and C. Zhang, 2009. An efficient hybrid data clustering method based on K-harmonic means and particle swarm optimization. Expert Syst. Appl., 36: 9847-9852.

Ye, F. and C.Y. Chen, 2005. Alternative KPSO-clustering algorithm. Tamkang J. Sci. Eng., 8(2): 165-174.

Yu, X. and M. Gen, 2010. Introduction to Evolutionary Algorithms. Springer, London.