International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 12 | Dec -2016 www.irjet.net p-ISSN: 2395-0072

AN EFFICIENT HYBRID COMPARATIVE STUDY BASED ON ACO, PSO, K-MEANS WITH K-MEDOIDS FOR CLUSTER ANALYSIS

S.Keerthana[1], Mrs. S. Akila[2]
[1]*Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India*
[2]*Assistant Professor, Dept. of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India*

Abstract - *Clustering is a popular data analysis and mining technique. A popular technique for clustering is based on k-means such that the data is partitioned into K clusters. However, the k-means algorithm highly depends on the initial state and converges to local optimum. The existing work presents a hybrid evolutionary algorithm to solve nonlinear partitional clustering problem. The evolutionary algorithm is the combination of FAPSO (fuzzy adaptive particle swarm optimization), ACO (ant colony optimization) and k-means algorithms, called FAPSO-ACO–K, which can find better cluster partition. Then k-means clustering is applied to get cluster results. K-means clustering is sensitive to the outliers and a set of objects closest to a centroid may be empty, in which case centroids cannot be updated. In k-means difficult to predict K-Value and different initial partitions can result in different final clusters. The objective of the proposed work is to overcome these problems, the K-medoids clustering algorithm where representative objects called medoids are considered instead of centroids. Because it uses the most centre located object in a cluster. The algorithm has excellent feature which requires the distance between every pairs of objects only once and uses this distance at every iterative step. It is less sensitive to outliers compared with the K-means clustering. It gives better performance than K-means clustering. Minimize the sensitivity of k-means to outliers. Pick the actual objects to represent clusters instead of mean values. Each remain object is clustered with the representative object (Medoid) to which is the most similar. The performance of the proposed work is evaluated through several benchmark data sets. The simulation result shows that the performance of the proposed work is better than the existing algorithm in terms of accuracy, recall, precision and F-measure.*

Key Words*: *Ant Colony Optimization, K-means clustering, K-medoids, Fuzzy Adaptive Particle Swarm Optimization

1. INTRODUCTION

Data cluster analysis is the task of grouping a set of objects in such a way that objects in the same group are more similar (in some sense or another) to each other than to those in other groups. The data clustering include in the some areas: mining, statistical data analysis, machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics. The traditional clustering algorithms can be divided into two categories: Hierarchical and Partitional Clustering. This paper concentrates on the partitional clustering. This clustering is used in wide variety of domains. In the K-means algorithms each cluster is represented by the center of gravity of the cluster. The k-means algorithm gave better results only when the initial partitions were close to the end solution. In other words, the results of k-means highly depend on the initial state and reach to local optimal solution. In order to overcome this problem, K-medoids clustering algorithm is used where representative objects called medoids are considered instead of centroids. Because it uses the most centrally located object in a cluster, the algorithm has excellent feature where it requires the distance between every pairs of objects is found only once and this distance is used at every iterative step. It is less sensitive to outliers compared with the K-means clustering.

2. LITERATURE REVIEW

TaherNiknam, BabakAmiri [10], proposed a hybrid evolutionary algorithm to solve nonlinear partitional clustering problem. The hybrid evolutionary algorithm is the combination of Fuzzy Adaptive Particle Swarm Optimization (FAPSO), Ant Colony Optimization and k-means algorithms, called FAPSO-ACO–K, which can find better cluster partition. The simulation results shows the better performance than other algorithms such as PSO, ACO, simulated annealing (SA), and k-means for partitional clustering problem.

Li-Yeh Chuang, Yu-Da Lin, and Cheng-Hong Yang, Member, IAENG [8], has proposed an improved particle swarm optimization based on Gauss chaotic map for clustering. It adopts a random sequence with a starting point as a parameter and relies on this parameter to update the positions of the particles. It provides the significant chaos distribution to balance the exploration and exploitation capability for search process. This is easy and fast function generates a random seed processes, and improve the performance of PSO due to their unpredictability. The method Gauss PSO is introduced to solve the data clustering problems.

TaherNiknam, BahmanBahmaniFirouzi and Majid Nayeripour [9], proposed an efficient hybrid evolutionary optimization algorithm based on combining Ant Colony Optimization (ACO) and Simulated Annealing (SA), called ACO-SA, for cluster analysis. The k-means algorithm is one of the most widely used clustering techniques. The algorithm is based on combination of the ant colony optimization and the simulated annealing. To evaluate the performance of the hybrid algorithm, it is compared with other stochastic algorithms viz. the original ACO, SA and k-means algorithms on several well known real life data sets.

James Kennedy and Russell Eberhart [6], have proposed a concept for the optimization of nonlinear functions using particle swarm methodology. The relationships between particle swarm optimization and artificial life and genetic algorithms are described; Particle swarm optimization algorithm seems to be effective for optimizing a wide range of functions. The adjustment toward pbest and gbest by the particle swarm optimizer is theoretically similar to the crossover operation utilized by genetic algorithms. The particle swarm optimizer serves both of these fields equally well, social behavior is so ubiquitous in the animal kingdom, because it optimizes.

K. Krishna and M. NarasimhaMurty [7], proposed a novel hybrid genetic algorithm(GA) that find a globally optimal partition of a given data into a specified number of clusters. The hybridize GA with a classical gradient descent algorithm used in clustering, K-means algorithm. Hence, it is called genetic K-means algorithm. The performance of GKA has been compared with that of some representatives of evolutionary algorithms, which are used for clustering and are supposed to converge to a global optimum.

3. SYSTEM METHODOLOGY

Data Set

A data set is a collection of data. Input data is an integral part of data mining applications. The data used in experiment is either real-world data obtained from UCI data repository and commonly accepted during evaluation dataset is described by the data type being used, the types of attributes, the number of instances stored within the dataset. The implementation of K-means and k-medoid algorithm is done on Iris data in Mat lab.

The Iris flower dataset is a multivariate data set introduced by Ronald Fisher in his 1936. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). These four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.

Proposed System

The proposed K-medoids clustering, representative objects called medoids are considered instead of centroids. Because it uses the most centre located object in a cluster. This algorithm has excellent feature that it requires the distance between every pairs of objects only once and uses this distance at every iterative step. The mean value of the objects in a cluster, it is less sensitive to outliers compared with the K-means clustering. Therefore, the k-medoids method should be more suitable for spatial clustering purpose than the k-means method because of the better clustering quality it can achieve. Rather than using conventional mean/centroid, it uses medoids to represent the clusters.

System Architecture

The iris data as taken for input to produce a better performance evaluation. In this architecture, the hybrid evolutionary optimization algorithm is used for choosing cluster centers. The hybrid evolutionary optimization algorithm is a combination of FAPSO (Fuzzy Adaptive Particle Swarm Optimization), ACO (Ant Colony Optimization) and K-means algorithm, which can find the cluster partition. After choosing the cluster centers, K-means clustering is applied for the clustering process. As the same way k-medoids also begins with randomly selecting k data items as initial medoids to represent the K cluster. At the end both the clustering algorithm gives a result for performance evaluation.

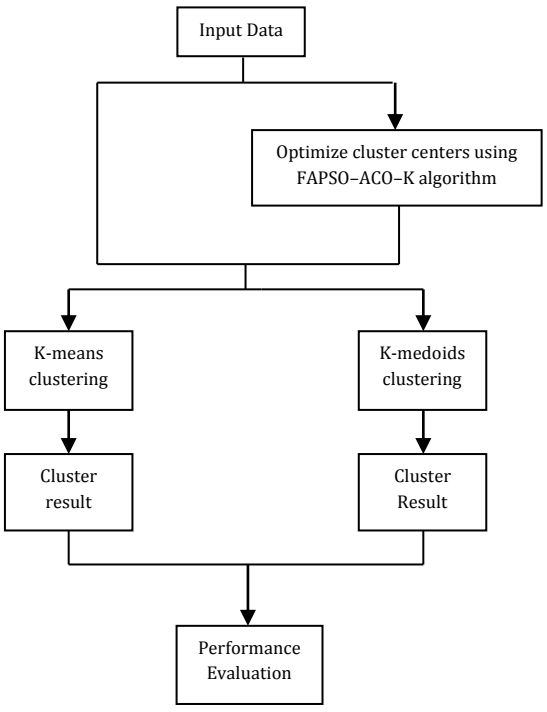

Fig-1: System Architecture

3.1 Module Description

The system is defined with the following modules

Optimize Cluster Centers FAPSO–ACO–K Algorithm

The PSO method should be taken as powerful techniques, which is efficient enough to handle various kinds of non linear optimization problems. It may be trapped into local optima if the global best and local best positions are equivalent to the particle's position over a number of iterations. A new method is proposed to incorporate intelligent decision-making formation of ACO algorithm into the original PSO where the global best position is particular for every particle. This algorithm uses randomly selection procedure of ACO algorithm to consign different global best positions to every distinct agent. For the clustering problem, the k-means algorithm tends to converge faster than the PSO and ACO algorithms as it requires less function evolutions. This algorithm uses the advantages of PSO and ACO algorithm to improve the final results of simulation.

K-MEDOIDS Clustering

K-medoids clustering algorithm is well efficient in classifying cluster categories. K-medoids algorithm is a center point k algorithm is mainly intended to overcome the shortcomings of K-means algorithm, especially the sensibility of outliers. K -medoids algorithm, first select k clustering centre points randomly from n data objects before computing the distance of other objects to each clustering centre, then select the one which is closest to clustering centre to set up an initial partition, and then use the iteration methods to change the clustering centre continuously until the most suitable fixed partition is found. This cycle is repeated till no medoid changes its placement. These marks the end of the process and have the resultant for end clusters with their medoids defined.

Algorithm 1: FAPSO-ACO–K

FAPSO-ACO–K algorithm used randomly selection procedure of ACO algorithm to assign different global best positions to every distinct agent. For clustering problem, the k-means algorithm tends to converge quicker than the PSO and ACO algorithms as it requires fewer function evolutions, but it usually results in less accurate clustering. FAPSO-ACO–K algorithm uses the advantages of this algorithm to develop the final results of simulation. In other word, the results of the PSO–ACO algorithm are used as the initial condition of the k-means algorithm.

Step 1: Generate the initial population and initial velocity

$$Population = \begin{bmatrix} C_1 \\ C_2 \\ \dots. \\ C_{Nswarm} \end{bmatrix}$$

C_i= [Center$_1$, Center$_2$,. . . ,Center$_k$],i = 1, 2, 3.,. . . ,N_{Swarm}

Center j = [c_1,c_2, . . . ,c_d]

$C_i^{min}<C_i<C_i^{max}$

$$Velocity = \begin{bmatrix} V_1 \\ V \\ \dots. \\ V_{Nswarm} \end{bmatrix}$$

V_i = [Center_V_1, Center_V_2,.....,Center_V_k], i = 1, 2, 3.,. . . ,N_{Swarm}

Center_V_j= [v_1,v_2....v_d] $V_i^{min}<v_i<v_i^{max}$

Where, Center$_j$ is the jth cluster center for the ith individual.

Center_V_j is the velocity of the jth cluster center for the ith individual.

V_i and C_i are the velocity and position of the ith individual, respectively. d is the dimension of each cluster center.

v_i^{max} and v_i^{min} are the maximum and minimum value of the velocity of each point belonging to the jth cluster center, respectively.

c_i^{max} and c_i^{min} are the maximum and minimum value of each point belonging to the jth cluster center, respectively.

Step 2: Generate the initial trail intensity

$$Trail_{Intensity} = [\tau_{ij}]_{N_{Swarm} \times N_{Swarm}}$$

$$\tau_{ij} = \tau_0$$

Where,τ_{ij} and τ_0 are trial intensity between the ith and jth swarms and initial trial intensity, respectively.

Step 3: Calculate the objective function is evaluated for each individual.

Step 4: Then sort the initial population based on the objective function values.

The initial population is ascending order based on the the objective function values (Gbest).

Step 5: Select the best global position.

The individual that as the minimum objective function is selected as the best global position.

Step 6: Select the best local position

The best local position ($Pbest_i$) is selected for each individual.

Step 7: Select the i^{th} individual

$$S_i = \left\{ C_j \parallel C_i - C_j \parallel \leq 2D_0 \left(\frac{1}{1 - \exp(-\frac{at}{t_{max}})} \right), i \neq j \right\}$$

Where D_0 is the initial neighborhood radius

a is a parameter used to tune the neighborhood radius over the iteration, t,

$\| \cdots \|$is the Euclidean distanceoperator.

Step 8: Calculate the next position for the i^{th} individual

In this case, at first, the transition probabilities between the C_i and each individual in S_i are calculated as

$$[Probability]_i = \left[P_{i1}, P_{i2}, \ldots . P_{i,M}\right]_{1 \times M}$$

$$P_{ij} = \frac{(\tau_{ij})^{\gamma_2}\left((1/L_{ij})\right)^{\gamma_1}}{\Sigma_{j=1}^{M}(\tau_{ij})^{\gamma_2}\left((1/L_{ij})\right)^{\gamma_1}}$$

$$L_{ij} = \frac{1}{|J(C_i) - J(C_j)|}$$

Where P_{ij} is the state transition probability between Ci and the j^{th} individual in S_i

M is the number of members in S_i

Step 9: If all of the individuals are chosen, go to the next step, otherwise i = i + 1 and go back to step 7.

Step 10: Check the termination criterion.

If the current iteration number reach the predetermined maximum iteration number, go to the next step, otherwise the initial population is replace with the new population of swarms and then goes back to step 3.

Step 11: Consider the previous Gbest value as the primary solution for the k-means algorithm

In this step to use the k-means clustering algorithm, the Gbest is considered as a primary solution of the k-means clustering problem. If the results of k-means algorithm are better than the Gbest value, the k-means results are considered as the final results, otherwise the last Gbest is considered as the concluding results.

Algorithm 2: K-MEANS clustering Algorithm

K-means is the one of the simplest unsupervised learning algorithms that solves the well known clustering problem. The procedure follows a simple to classify a given dataset through a certain number of clusters assume the k clusters fixed a priori. The major idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes dissimilar result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and connect it to the nearby centroid. When no point is pending, the first step is completed and an early group age is done. At this point need to recalculate the k new centroids as bar centers of the clusters resulting from the before step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest centroid. A loop has been generated. As an end of this result, loop may notice that the k centroids change their location step by step until no other changes are done. In other words centroids do not move any other.

Finally, this algorithm aim at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j-1}^{k} \sum_{i-1}^{x} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\| x_i^{(j)} - c_j \|^2$ is a choose the distance measure between a data points $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their respective cluster centres.
The algorithm is composed of the following steps:

- Place the K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
- Consign each object to the group that has the closest centroid.
- When all objects have been assigned, and recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a division of the objects into groups from which the metric to be minimized can be calculated.

K-MEDOIDS Clustering Algorithm

K-medoids is also a partitioning method of clustering that clusters the data set of *n* objects into *k* clusters with *k* known a priori. The k-Means algorithm is sensitive to outliers since an object with a very large value may substantially distort the distribution of data. Instead of, taking the mean value of the objects in a cluster as a reference point. The k-medoid can be used, which is the mainly center located object in a cluster. Thus, the partitioning method can still be performed based on the principle of minimizing the sum of the dissimilarities among each object and its corresponding reference point. This forms the basis of the k-Medoids method. The basic strategy of k-Medoids clustering algorithms is to find k clusters in n objects by first arbitrarily finding a object for the medoids for each cluster. Each remaining object is clustered with the medoid to which it is the most similar. The algorithm takes the input parameter k, the number of clusters to be partition among a set of n objects.

A typical k-medoids algorithm for partitioning based on medoid or central objects is as follows:

Input: K: Is the number of clusters
D: the data set containing n items

Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearby medoids

$$Z = \sum_{i-1}^{k} \sum |X - m_i|$$

Where, Z: Sum of absolute error for all items in the data set
x: the data point in the space representing a data item
m_i: is the medoid of cluster C_i

The algorithm is composed of the following steps:

- Arbitrarily choose k data items as the initial medoids.
- Assign each remaining data item to a cluster with the nearest medoid.
- Randomly select a non-medoid data item and compute the total cost of swapping old medoid data item with the current selected non-medoid data item.
- If the total cost of swapping is less than zero, then perform the swap operation to generate the new set of k-medoids.
- Repeat steps 2, 3 and 4 till the medoids stabilize their locations.

The working of K-Medoids clustering algorithm is related to K-Means clustering. It also begins with randomly selecting k data items as initial medoids to represent the k clusters. All the other remains items are included in a cluster which has its medoid nearby them. Thereafter a new medoid is determined which can represent the better cluster. All the remaining data items are yet again assigned to the clusters having contiguous medoid. In the each iteration, the medoids alter their location. The methods minimize the sum of the dissimilarities between each data item and its corresponding medoid. This process is repeated till no medoid changes its placement. At the end of the process and have the resultant final clusters with their medoids defined. K clusters are formed which are centered around on the medoids and all the data members are placed in the proper cluster based on nearest medoid.

4. RESULT AND DISCUSSION

The experimental analysis is indented to be of use to researchers from all fields to study algorithms experimentally, that analysis was conducted using MATLAB environment. The existing and proposed methodologies are compared with each other in terms of varying parameter values. The performance that are considered for proving the improvement of the proposed methodology are accuracy, precision, recall and F-measure.

Metrics	HYBRID ACO-PSO-K-means	HYBRID ACO-PSO-K-medoids
Accuracy	83.2215	95.9732
Precision	0.8438	0.9600
Recall	0.8327	0.9597
F-Measure	0.8382	0.9599

Table-1 Comparison Table

Accuracy

The accuracy is the proportion of true results, both true positive and true negative among the total number of case examined.

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \frac{\text{True positive + True negative}}{\text{True positive + True negative+False positive+ False negative}}$$

Precision

The precision is calculated as follows:

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive+False positive}}$$

In the task, precision for a class is the number of true positives divided by the total number of elements label as belonging to the positive class.

Recall

The calculation of the recall value as follows:

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive+False negative}}$$

Recall in this context is defined as the number of true positives divided by the total number of elements that truly belong to the positive class.

F-Measure

F-measure is calculated as:

F-measure = $2 \times (precision \times recall / precision + recall)$

F-measure is calculated from the precision and recall value. A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score.

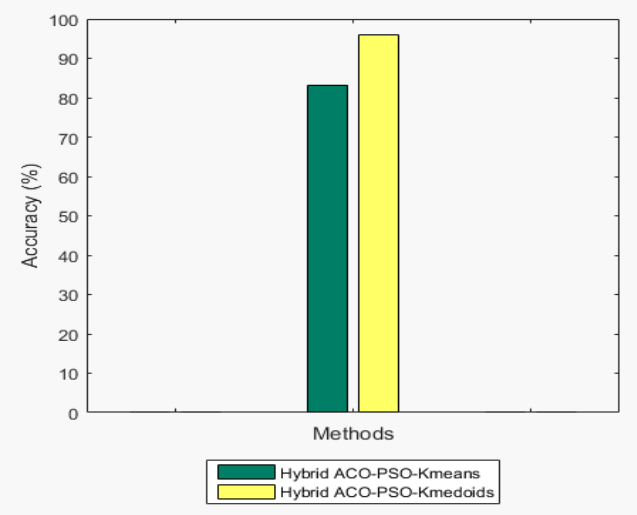

Chart-1: Accuracy Table

5. CONCLUSION AND FUTURE WORK

Clustering is a popular analysis and data mining technique. A popular technique for clustering is based on k-means such that the data is partitioned into K clusters. However, the k-means algorithm extremely depends on the initial state and converges to local optimum solution. The base work presents a hybrid evolutionary algorithm to solve nonlinear partitional clustering problem. It is called FAPSO-ACO–K, which can find better cluster partition. Then k-means clustering is applied to get cluster result. This K-means clustering have some issues like sensitive to the outliers and a set of objects nearest to a centroid may be empty, in which case centroids cannot be updated.

To overcome these problems we proposed K-medoids clustering, where representative objects called medoids are considered instead of centroids as it uses the most centrally located object in a cluster. This algorithm has excellent feature as it requires the distance between every pairs of objects only once and uses this distance at every iterative step. It is less sensitive to outliers compared with the K-means clustering. The simulation results show that the performance of the proposed algorithm is better than the existing algorithm in terms of accuracy, recall, precision and F-measure.
K-mediods improves the noise sensibility of k-means. Actual objects to represent clusters instead of mean values. Each object is clustered with the representative medoid to which is the most similar. Less sensitive to outliers compare with k-means. K-Medoids method is very time- consuming.

In the future wok, if we use SOM (Self Organization Map) or Hierarchical clustering algorithm with the hybrid ACO-PSO algorithm a better clustering result than the current approach may be obtained.

REFERENCE

[1] Agustın-Blas .L.E, Salcedo- Sanz.S, Jimenez-Fernandez .S, Carro-Calvo .L, Del Ser.J,Portilla-Figuer .J.A, “A new grouping genetic algorithm for clustering problems”, Elsevier ad hoc [2012].
[2] Cao.D.N, Krzysztof.J.C, GAKREM: a novel hybrid clustering algorithm, Information Sciences, 2008
[3] Gnanapriya.S and Shivaranjani.P, ”Initialization K-Means using ant colony optimization” ISSN 2319-5991 www.ijerst.com,vol. 2, no. 2, may [2013].
[4] Kao .Y.T, Zahara .E, Kao.I.W, A hybridized approach to data clustering, Expert Systems with Applications , 2008 .
[5] Kapil Agrawal, RenuBagoria, ”Ant Colony Optimization: efficient way to find shortest path International Journal of Advanced Technology & Engineering Research (IJATER)”, [2014].

[6] Kennedy.J, Eberhart.R, Particle swarm optimisation, vol. 4, in: Proceedings of the IEEE International Conference on Neural Networks, Piscataway, NJ, 1999.
[7] Krishna.K, Murty, Genetic k-means algorithm, IEEE Transactions of System Man Cybernetics Part B-Cybernetics, 1999 .
[8] Li-Yeh Chuang, Yu-Da Lin, and Cheng-Hong Yang, Member, IAENG, An Improved Particle Swarm Optimization for Data Clustering 2012.
[9] Niknam. T, BahmaniFirouzi. B, Nayeripour. M, An efficient hybrid evolutionary……..algorithm for cluster analysis, World Applied Sciences Journal ,2008.
[10] TaherNiknam ,BabakAmiri, An efficient hybrid approach based on PSO, ACO and *k*-means for cluster analysis,2010.