International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 4 (2015) pp. 10087-10094

http://www.ripublication.com

Innovative Approach on Algorithmic Implementation For Effective Clusters

Jessie Monica.J [1],Mohana Prasad [2]
[1] Student, Department of Computer Science, Sathyabama University, Chennai, Tamil Nadu,India
Email: jjessiemonica@gmail.com
[2]Asst. Professor, Faculty of Computing, Sathyabama University Chennai, Tamil Nadu,India
Email: mohanaprasad1983@gmail.com

Abstract

Nowadays, the huge number of data set is generated by various real time applications such as e- marketing, astronomical, data center etc. Clustering consists of a group of data sets in turn every data set in the cluster contains centroid. It will be easy to have centroids for datasets so that the data sets are retrieved and use additional operations on the data sets. Different types of clustering methods are used in grouping the generated data sets such as k-means, k-medoid etc. These techniques have two restrictions such as k-value selection and centroid selection where the cluster size is assigned manually and the selection of centroid value is done randomly. These two parameters provide a massive impact on the clustering performance. This paper discusses about how to design a method to select k value automatically and also to design a method to select exact cluster center automatically. The system proposed here defines two methods for resolving these two issues using domain based k value selection and centroid selection.

Key Terms: Centroids, Data Set, K-means Clustering, Differential Evolution, Fuzzy Sets,

Introduction

A data set is nothing but a collection of data.Generally a data set corresponds to the contents of a single database table which is a single statistical data matrix, in which every column of the tableis represented as a particular variable, where as each row in the table corresponds to a given member of the data set.

Data clustering is done for data sets. It is an important mechanism used in many applications such as data mining, statistical data analysis, math programming etc.

Many approaches are proposed for data clustering and used in many applications. Many algorithms are used for clustering of data. One of the important algorithms in data clustering is the k means algorithm. This method will put number of data points into clusters with randomly initializing k means data points.

K-means is one of the simplest unsupervised learning algorithms which solves the well known problems that occurs during clustering. A simple and easy way is followed to classify a given data set through a certain number of clusters fixed a priori. The objective is to define k centroids, one for each of the cluster formed. The next step is to associate every points which belong to a given data set to the nearest centroid. When there is no point pending, the first step is completed and groupage is done. From the previous step, It is needy tocalculate k new centroids as barycentre of the clusters once more. A new binding is made between the nearest new centroid and the same data set points after obtaining these k new centroids. A loop has been formed. Thusit is noticed that these k centroids keep on changing their location step by step until no more changes are done.

Related Works

Fuzzy k-Modes Algorithm

A fuzzy k-modes algorithm for clustering categorical data is very effective for identifying cluster structures from categorical data sets [1]. The K modes type algorithms are effective in recovering the inherent clustering structures from categorical data if such structures exist. The fuzzy partition matrix provides more information to help the user to determine the final clustering and to identify the boundary objects. Such information is extremely useful in applications such as data mining in which the uncertain boundary objects are sometimes more interesting than objects which can be clustered with certainty [2].

Point Symmetry Based Clustering

By using the point symmetry based distance a variable string length genetic clustering technique is developed which can automatically evolve the number of clusters present in a data set [3] [9]. Here the proposed cluster validity index, *Sym*-index is used for detecting the proper number of clusters present in a data set and the proper partitioning . Finally, development of some multi objective clustering technique using symmetry, connectivity, compactness etc. as different objective functions so that it can work well for partitioning any type of data sets needs to be investigated [6].

Differential Evolution Algorithm

Differential evolution (DE) is considered as the fastest, efficient global search heuristics of current interest and it is meant for its robustness. The application of DE is it automatically clusters large unlabeled data sets..This algorithm requires no prior knowledge of the data that has to be classified, When compared to most of the existing clustering techniques. Rather, the optimal number of partitions of the data it determined on the run [4] [7]. DE-based strategy is presented for crisp clustering of

real-world data sets. An important feature of this technique is that it can automatically find the optimal number of clusters even for data sets with very high dimension, where tracking the number of clusters may be impossible(i.e., the number of clusters does not have to be known in advance). The proposed Adaptive Co-evolutionary Differential Evolution (ACDE) algorithm is able to outperform two other state-of-the-art clustering algorithms in a statistically meaningful way when compared to a majority of the benchmark data sets.

Artificial Bee Colony (ABC) algorithm

Several modern heuristic algorithms have been developed for solving combinatorial and numeric optimization problems. Classification of these algorithms can be done in different groups depending on the criteria, population, iterative, stochastic, deterministic, etc are taken in consideration. This algorithm proposes real time example compared with some parameters. The research branch, Swarm intelligence models the population of swarms or the interacting agents that are able to organize itself. The typical example of a swarm system are, a flock of birds,an ant colony, or an immune system [8]. Another good example of swarm intelligence is Bees' swarming around their hive. Artificial Bee Colony (ABC) Algorithm is an optimization algorithm which is based on the intelligent behaviour of honey bee swarm. ABC algorithm is used for optimizing multivariable multimodal function optimization [5].

Problem Statement

The task of the system is to

- To compute k value to automatically using feature selection model in the specific domain.
- To select exact centroid value to automatically using domain based dataset.

Proposed System

This system assumes the k value with fuzzy logic system.The fuzzy logic will automatically generate the centroids of the data sets based on some assumptions and preplanning system. The k value with fuzzy logic is implemented using feature selection based k value selection and domain data set are based on centroid selection.The advantages of this research work is that it Converges the data set within short time and it reduces the number of clustering iteration as well as the Clustering accuracy is increased when compared to other approaches.

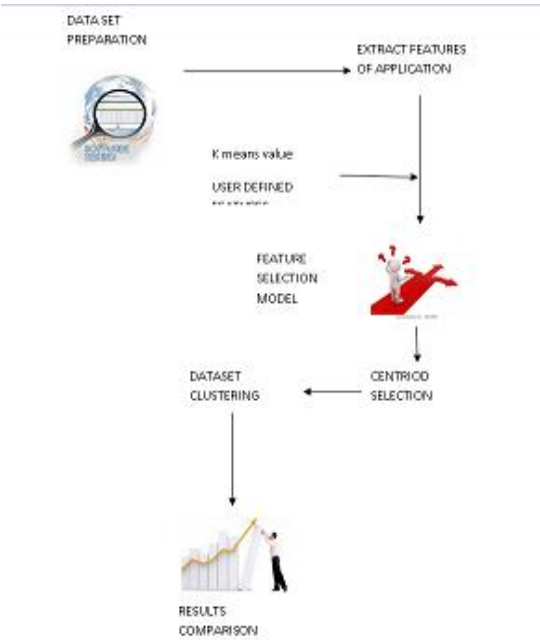


Figure 1: System Architecture

The working of system architecture, consider a real-time dataset initially which is taken from any real time applications. The dataset includes some works, operations, process of that application. Extract some features of that application. These features are necessary to find the K-value and Centroid value according to their computed distance.

Assign K value that will select the centroid value based on the features assumed in the datasets. The feature selection model will implement the features to the k means value based on user specification. Next, a single iteration is performed on the data set then compute the number links between every dataset. To this end, an average is computed to select the K means value. An average value is calculated and computes a k means value. Centroid value is selected for clustering the data set into single group. The centroid is selected based on the domain data set at initial iteration. Finally the data sets are converging into K groups. Initially, the k-means value is selected automatically by using abovementioned k-means value selection method. Next, the centroid is selected by using abovementioned centroid value selection method then computes the distance between every data points with number of iterations. Finally, a threshold value is applied to group the data sets.

The phases involved in the selection process of the centroid using k-means method are stated below

- Dataset Preparation
- Feature Selection Model using K-means
- Domain Based Data Set using Centroid Selection
- Dataset Clustering

Dataset preparation

In this module, the dataset will be taken from any real time applications. The dataset is including with some works, operations, process of that application. Extract some features of that application. These features are necessary to find the K-value and Centroid value according to their computed distance.

An algorithm for clustering (partitioning) N data sets into K disjoint subsets Sj containing data sets so it can minimize the sum-of-squares criterion .

$$J=^{K}\sum_{j=1}\sum_{n\epsilon SJ}\left|\, X_n- \mu_j \,\right|^2$$

Where xn is a vector representing the nth data point and uj is the geometric centroid of the data points in Sj.

How Clustering algorithm works, Start with a decision on the value of k = number of clusters.Now make an initial partition randomly that classifies the data into k clusters. The training samples can be taken randomly,or systematically it begins with k training sample as single-element clusters and assigning each of the remaining (N-k) training samples to the cluster with the nearest centroid. Then finally after assigning training samples to the cluster, re-compute the gaining cluster's centroid. In the next phase each and every sample of sequence is taken and distance from the centroid for each cluster is computed. In case if a sample is not in the cluster with closest centroid, thenthe sample is added to that cluster for updating the centroid of the cluster thus it gains the new sample and the cluster losing the sample. The processis repeated until convergence is achieved.

K means value selection: Feature Selection Model

In this module , the K means value will select the centroid value based on the features assumed in the datasets. The feature selection model will implement the features to the k means value based on user specification. After feature extraction an initial iteration is performed on the data set then compute the number links between every dataset. From the results of iteration process similar datasets with identified features are grouped. From the identified similar datasets an average is computed to select the K means value.

Centroid selection: Domain Based Data Set

In this module, an average value is calculated and computes a k means value. Centroid value is selected for clustering the data set into single group. The centroid is selected based on the domain data set at initial iteration. The centroid value is determined to relate the similar group of datasets and also for efficient clustering. A random centroid value is Choosen and based on the centroid value datasets will be clustered.

Dataset Clustering

In this module, the data sets are converging into K groups. Initially, the k-means value is selected automatically by using abovementioned k-means value selection method. Next, the centroid is selected by using abovementioned centroid value selection

method then computes the distance between every data points with number of iterations. Finally, a threshold value is applied to group the data sets.

Let *q* be the centroid value and *t* be the target data point. The Euclidean distance between the centroid and target data point then,

$D_{EUC}(q_i,t_i)=^{N}\sum_{i=1}\sum(\sum q_i - t_i)^2$

Results and Disscussion

A discussion about the result and performance measures of the proposed system is discussed.

Graph 1: Time Complexity

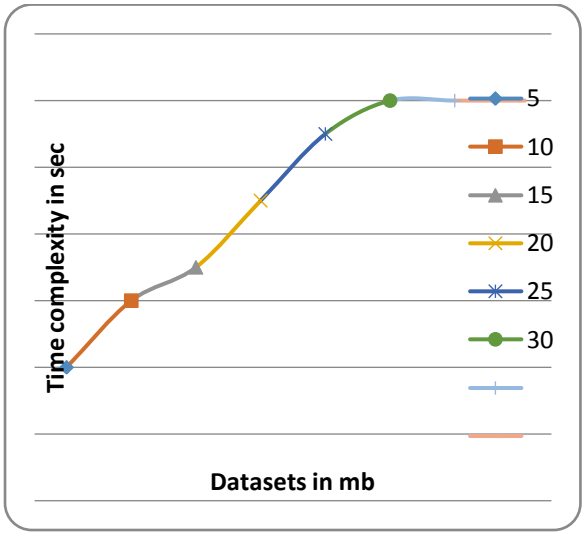
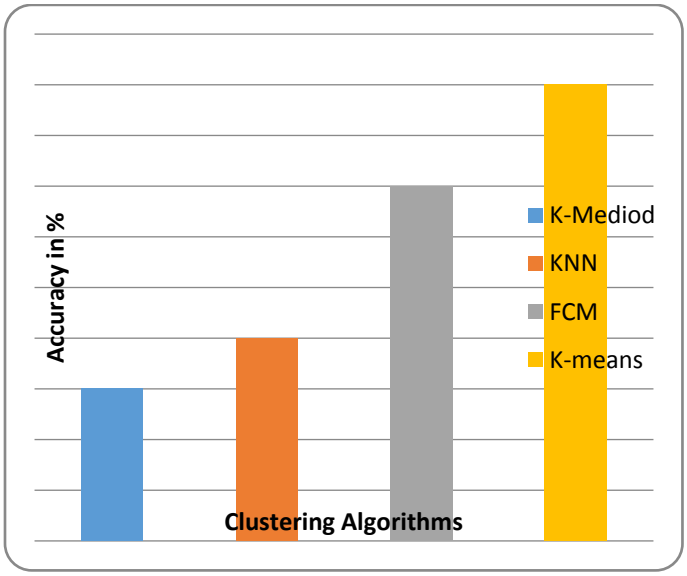


Figure 2: Demonstration of Time Complexity using K-means

The above graph states the time complexity in clustering datasets using k-means. This indicates that the time complexity gradually increases with increase in number of datasets.

Graph:2Accuracy

The second graph shown below depicts the comparison of various clustering algorithms. It shows that K-means clustering produces high accuracy when compared to other clustering methods.



Figue 3: Depicting the accuracy of efficient clustering of datasets

Conclusions and Future Work

Selection of k value and exact cluster centre is done automatically using domain based k value selection and centroid selection.The dataset is converged within short time and it reduces the number of clustering iteration as well as the Clustering accuracy is increased.

The k means value will automatically find the centroids based on some features assigned by the user. The future enhancement includes some approaches without selecting the features automatically generates the exact centroids value.

References

[1] Zhexue Huang and Michael K. Ng "A Fuzzy k-Modes Algorithm for Clustering Categorical Data" IEEE transactions on fuzzy systems, vol. 7, no. 4, August 1999.

[2] James C. Bezdek , Robert Ehrlich, William Full "FCM Fuzzy c-Means Clusteringalgorithm" Computers & Geosciences Vol. 10, No. 2-3, pp. 191-203, 1984.

[3] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 12, pp. 1650–1654, 2002.

[4] Swagatam Das, Ajith Abraham "Automatic Clustering Using an Improved Differential Evolution Algorithm" IEEE Transactions on Systems, Man, and Cybernetics—part a: systems and humans, vol. 38, no. 1, January 2008.

[5] DervisKaraboga, BahriyeBasturk"A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm" J Glob Optim (2007) 39, pp. 459–471.

[6] Sara C. Madeira and ArlindoL.Oliveira" Biclustering Algorithms for Biological Data Analysis: A Survey" IEEE Transactions on computational biology and bioinformatics, vol. 1, no. 1, January-march 2004.

[7] S. Paterlini and T. Minerva, "Evolutionary approaches for cluster analysis," in Soft Computing Applications, A. Bonarini, F. Masulli, and G. Pasi, Eds. Berlin, Germany: Springer-Verlag, 2003, pp. 167–178.

[8] M. Omran, A. Salman, and A. Engelbrecht, "Dynamic clustering using particle swarm optimization with application in unsupervised image classification," in Proc. 5th World Enformatika Conf. (ICCI), Prague, Czech Republic, 2005.

[9] R. H. Eduardo and F. F. E. Nelson, "A genetic algorithm for cluster analysis," Intelligent Data Analysis, vol. 7, pp. 15–25, 2003.