IJCTA, 9(27), 2016, pp. 153-158
© International Science Press

ISSN: 0974-5572

Enhanced Clustering Algorithm to Improve the Cluster Accuracy

R. Govindaraj* and T.N. Ravi**

ABSTRACT

Data Mining is a process, where intelligent methods are applied to extract data patterns. Clustering technique plays an important role in the field of data mining. It is a process of partitioning a set of data in a meaningful sub classes called clusters. This technique is called unsupervised classification because the dataset has no predefined classes. K-medoids is the most popular and partition based clustering algorithm. In the past decades, various methods have been proposed to improve the performance of the k-medoids clustering algorithm. But still many research works are carried out to enhance the clustering accuracy by improving the K-Medoids algorithm. In this research a new clustering algorithm is proposed. The proposed algorithm is analysed with four different dataset and also a comparative analysis is done with the existing algorithm to prove the efficiency of the algorithm.

Keywords: Data Mining, Clustering, PCA and K-Medoids

1. INTRODUCTION

Data Mining is a process of extracting useful information from a huge database or data warehouse or any other data storage repository [1]. There are many other terms carrying a similar meaning to data mining, such as knowledge mining from databases, knowledge extraction, data or pattern analysis, data archeology, and data dredging. In real world, many data mining applications generate the datasets with a high volume of features and instances. The presence of high volume of data will degrades the machine learning algorithms. To infer knowledge from the high volume of data. The dataset should be preprocessed. In the data mining field Dimensionality Reduction (DR) is an important Technique to reduce the high volume of data.

Clustering is an important technique to classify the raw data reasonably and searches the hidden patterns that may exist in datasets [2, 3]. This technique is mainly used to group the data into disjoint clusters so that the data in the same cluster are similar, and data belonging to different cluster are differ. In the past decades many algorithms have been developed for clustering [4, 5]. The clustering algorithm typically considers all features. However, with high dimensional data, many features are redundant or irrelevant [6]. The redundant features are of no help for clustering; even worse, the irrelevant features may hurt the clustering results by hiding clusters in noises. In the field of data mining many approaches were developed to address this problem. But still the problem persist. To overcome the high dimensionality problem many approaches were developed among them one of important approach principal component analysis (PCA) and random projection.

The main purpose of clustering techniques is to partition a set of data into different groups, called clusters. In the literature the clustering process were performed using K-means clustering. But K-Medoids algorithm performs better than K-Means algorithm. The K-medoids clustering algorithm is very efficient in classifying cluster for many practical applications. Based on algorithm analysis. In this algorithm the initial centers are computed properly. To

* Research Scholar and Associate Professor, Department of Computer Science, Government Arts College, Ayyarmalai, Kulithalai–639 120, Karur-Dt, TamilNadu, India, *Email: pr.govindaraj63@gmail.com*

** Assistant Professor, Department of Computer Science, Periyar E.V.R. College (Autonomous), Tiruchirappalli–620 023, TamilNadu, India, *Email: proftnravi@gmail.com*

compute the initial center principal component analysis is used. In this method principal component analysis and the K-Medoids algorithm were combined to improve the clustering accuracy.

2. PRINCIPAL COMPONENT ANALYSIS

Component Analysis (PCA) [7] is a classical statistical technique which is widely used to reduce the dimensionality of a dataset consisting of enormous amount of interrelated variables. PCA reduces the dimensionality by transforming the original dataset into a new set of variables, called principal components, where the largest variance present in the original dataset is captured by the highest component in order to extract the most important information. The computational steps of the PCA algorithm is given below,

Step 1 Calculate the Mean:

$$\bar{y} = \frac{1}{M}\sum_{i=1}^{M} y_i$$

Step 2 Subtract the Mean from variables

$$\Phi_i = y_i - \bar{y}$$

Step 3 Form the Matrix A= [Φ1, Φ2,…. ,ΦM] (N × M matrix), then compute:

(sample covariance matrix, N × N, characterizes the scatter of the data)

Step 4 Calculate the Eigenvalues of $C : \lambda_1 > \lambda_2 < ... > \lambda_N$

Step 5 Calculate the Eigenvectors $C : u_1, u_2, ..., u_N$

Since C is symmetric, $u_1, u_2, ..., u_N$ form a basis, (i.e., any vector x or actually y – $\bar{y}$, can be written as a linear combination of the Eigenvectors):

Step 6 (dimensionality reduction step) keep only the terms corresponding to the K largest Eigenvalues:

$$\hat{y} - \bar{y} = \sum_{i=1}^{K} b_i u_i \quad whereK << N$$

The principal component of data obtained by considering the eigenvector which has the highest eigenvalue. Generally the eigenvectors are found form the covariance matrix. The next step is to order them by eigenvalue, highest to lowest. To reduce the dimensions, the first d (no. of principal components) eigenvectors are selected. The final data has only d dimensions. The main aim of applying PCA approach is to cluster the data in an accurate manner. This leads to the researcher to perform the cluster analysis in efficient way.

3. K-MEDOIDS ALGORITHM

The k-Medoids algorithm is a clustering algorithm. In this algorithm a medoid is obtained, which is the most centrally located object in a cluster [8,9]. The basic idea of kMedoids clustering algorithms is to find k clusters in n objects by first arbitrarily finding a representative object (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 partitioned among a set of m objects. Let us consider *m* objects having *n* variables and these variables will be divided *k*(*k* < *m*) into clusters the following steps shows the K-Medoids Algorithm:

Step 1 Select the Initial Medoids

- * The distance between the each pair of *m* objects are calculated by the following Euclidean distance

$$d_{xy} = \sqrt{\sum_{s=1}^{n} \left(Z_{xs} - Z_{ys} \right.^{2}}$$

where x = 1 ……n; y = 1 ….... .n

- * Calculate c_{ij} and make the initial guess of the center

$$c_{ij} = \frac{d_{xy}}{\sum_{l=1}^{m} d_{il}}$$

where x = 1……n; y = 1 …..... n

- * Calculate the center medoids of all data and arrange them in ascending order and select the k objects
- * Assign each objects to the nearest Medoids

Step 2 Find the new Medoids

- * Selected medoids in each cluster are changed by the new object which has the minimum distance in its cluster

Step 3 * Assign each object to the nearest new medoid.

- * Based on the distance measure of new medoids new optimal value has computed
- * The algorithm works till the new optimal value and the existing one is not similar.

4. UCI MACHINE LEARNING REPOSITORY

The UCI (University of California, Irvine) Machine Learning Repository was created by David Aha and fellow graduate students at University of California, Irvine in 1987. The repository consists of different kinds of databases which were used by the machine learning community to validate the machine learning algorithms. But nowadays, this repository is accessed by students, educators, and researchers for data mining. This repository contains 256 datasets for classification analysis, 61 datasets for regression analysis, 52 datasets for clustering analysis and 51 datasets for other analysis. The UCI repository is widely used in the field of computer science especially in data mining [UCI, 2012].

4.1. Dataset

The proposed algorithms are evaluated with four different datasets which are obtained from the UCI Machine Learning Repository. The motivation of choosing these datasets is due to their popularity. Many researchers use them for their evaluation. They are regarded as benchmark among Data Mining researching community.

5. PROPOSED WORK

Clustering is an important technique commonly used in data mining to find some hidden features embedded in the dataset. Traditionally this techniques are broadly classified into hierarchical and partitioning [10,]. A partitioning-based algorithm such as K-medoids has been widely reported in the literature for the clustering of data. For the

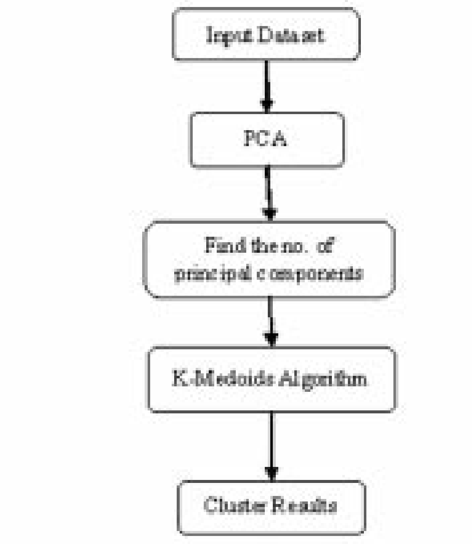

Figure1: Work Flow of PK Algorithm

past few years many improvement were done in the K-medoids algorithm [11, 12]. Even now many research work were carried out to improve the clustering accuracy. For the betterment of clustering accuracy a new algorithm is proposed namely, PK algorithm (first letter of PCA & K-Medoids) by combining PCA and K-medoids algorithm.

In the proposed algorithm PCA and the K-medoids are used. PCA is used to reduce the dimension of the data. The K-medoids algorithm used to cluster the data. In this model the dataset is given as an input. The principal component analysis is applied to given input to find the number of principal components. The obtained principal components are assigned as initial centroids for the k-medoids algorithm. Using the initial centroids it starts to form the initial clusters based on the relative distance of each data-point from the initial centroids.

PK Algorithm

Step 1 Reduce the dimension of the data using PCA and Find the number of principal components(pc)

Step 2 Arrange the pc's in the ascending order to make it as initial centroids

Step 3 Compute the distance of each data-point to all the centroids using Euclidean distance formula.

Step 4 For each data object find the closest centroid and assign to the cluster with nearest centroid and store them in array.

Step 5 Repeat the Step 3 and 4 till the end of the datasets

Step 6 Finally, the data are clustered accurately.

6. EXPERIMENTS

In this section, the experiment is carried out to prove the efficiency of the proposed algorithm. The experiment is carried out on a well-known publicly available dataset for UCI Machine Learning Repository. The efficiency of the proposed algorithm is analyzed in terms of clustering accuracy. The proposed algorithm is compared with two existing clustering selection algorithms; they are K-Medoids and Modified K-Medoids (M k-medoids).

The experimental results of the proposed algorithm are depicted in table 1. Fig 2 shows the comparative analysis of the feature selection algorithms. From the results, it is proved that the proposed algorithm shows an improved clustering accuracy than the existing algorithms.

Table 1
Clustering Accuracy of PK with Existing Algorithms

Dataset	*Accuracy Percentage (%)*		
	K-Medoids	*MK-Medoids*	*PK*
Iris	92.00	93.11	95.28
Wine	94.32	95.15	96.11
Glass	88.13	90.22	92.86
Leukemia	89.29	91.86	93.47

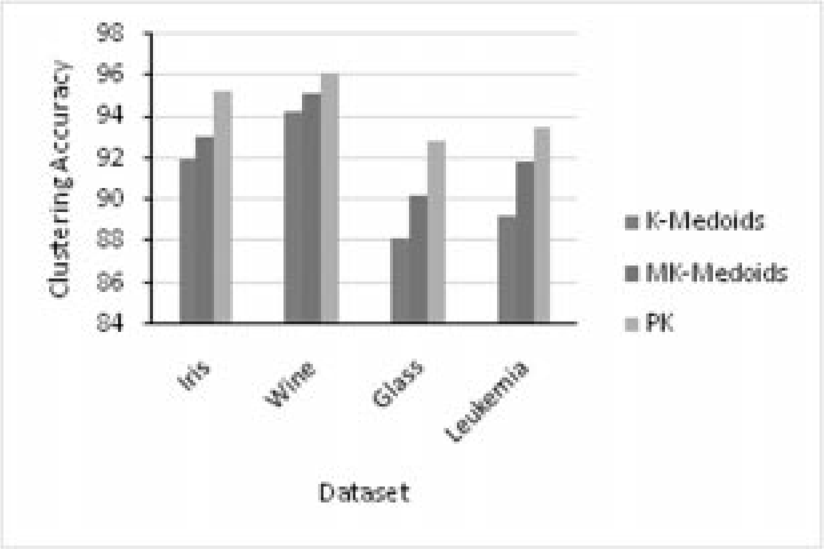

Figure2: Comparative Analysis of Clustering Algorithms

7. CONCLUSION

Data mining is a convenient way of extracting patterns, which represents knowledge implicitly which are stored in large data sets. In this field clustering plays an important role to group the data and this is called unsupervised learning. Clustering the data is a challenging task. Nowadays many research works were carried out to improve the clustering accuracy. But still the problem is under research. In this research article a new algorithm is proposed by enhancing the K-Medoids algorithm to improve the accuracy of the clusters. To check the efficiency of the proposed algorithm, an experiment is conducted on a publicly available datasets. The experiment results on four datasets exhibit that the proposed algorithm shows the promising improvement in clustering accuracy.

REFERENCES

[1] Data Mining: “A Prediction for Performance Improvement of Engineering Students using Classification”, *World of Computer Science and Information Technology Journal*, **2**, 51-56, 2012.

[2] Kavita Nagar, “Data Mining Clustering Methods: A Review”, *International Journal of Advanced Research in Computer Science and Software Engineering,* 575-579, 2014.

[3] R. G. Mehta, N. J. Mistry, M. Raghuwanshi, “Towards Unsupervised and Consistent High Dimensional Data Clustering”, *International Journal of Computer Applications,* 40-44, 2014.

[4] R. Amutha, Renuka. K, “Different Data Mining Techniques And Clustering Algorithms”, *International Journal of Technology Enhancements and Emerging Engineering Research*, 15-17, 2015.

[5] Sharaf Ansari, Sailendra Chetlur, “An Overview of Clustering Analysis Techniques used in Data Mining ”, *International Journal of Emerging Technology and Advanced Engineering*, 284-286, 2013.

[6] Nebu Varghese, Vinay Verghese, Gayathri. P, N. Jaisankar, “A Survey of Dimensionality Reduction and Classification Methods”, *International Journal of Computer Science & Engineering Survey,* **3**, 45-54, 2012.

[7] R. Shenbakapriya, M. Kalimuthu, P. Sengottuvelan, “Improving Clustering Performance on High Dimensional Data using Kernel Hubness”, *International Journal of Computer Applications*, 27-30, 2014.

[8] Abhishek Patel, “New Approach for K-mean and Kmedoids algorithm”, *International Journal of Computer Applications Technology and Research*, 71-82, 2013.

[9] T. Velmurugan,and T. Santhanam, “A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach” *An experimental approach Journal of Information. Technology*, 478-484, 2011.

[10] Fang Yuag, Zeng Hui Meng, “A New Algorithm to get initial centroid”, *Third International Conference on Machine Learning and cybernetics, 1119-1124*, 2014.

[11] Maria Camila N. Barioni, Humberto L. Razente, Agma J. M. Traina, “An efficient approach to scale up k-medoid based algorithms in large databases”, *International journal Recent Technologies,* 265-279, 2015.

[12] Gopi Gandhi, Rohit Srivastava, “Analysis and Implementation of Modified K-Medoids Algorithm to Increase Scalability and Efficiency for Large Dataset”, *International Journal of Research in Engineering and Technology, 150-153,* 2014.