ISSN 2395-1621

Pairwise Deterministic Approach of Centroid Swapping for K – Means Clusters

[#1]SampadaM.Chaudhari, [#2]R.H.Kulkarni

[1]schowdhari007@gmail.com
[2] rkpv2002@gmail.com

[#1]Department of Computer Engineering
TSSM's BSCOER,Narhe,Pune

ABSTRACT

Cluster analysis entertains clustering algorithm and validity of clusters with equal significance ratio. Many applications include clustering process as a vital need to accomplish even the primary objectives and social benefits. k – means clustering algorithm has faced many challenges to accomplish precise clustering because of arbitrary initialization of centroids. Such arbitrary initial centroids have led the algorithm to converge in local optimal solutions (or) clusters, which are not possible clusters of the data. Hence, efforts have been made in the literature to improve k – means clustering algorithm by selecting initial centroids meaningfully. Taking forward in this paper we are Proposing Pairwise deterministic centroid swapping strategy for k – means clustering. The proposed deterministic swapping strategy will work based on feedback relevance model, which is a function of transition state of objective function between consecutive stages of centroid updates.DB index is use to improve the performance of systematic swapping.

Keywords— Clustering, K-means clustering, cluster validation

ARTICLE INFO

Article History

Received :18th June 2015

Received in revised form :

19th June 2015

Accepted : 23rd June 2015

Published online :

29th June 2015

I. INTRODUCTION

Clustering is a technique used in data mining to place data into groups without having any information of the group.As computer and database technologies advance rapidly, data accumulates in a speed unmatchable by human's capacity of data processing. Millions of databases have been used in business management, government administration, scientific and engineering data management, and many other applications. Therefore, developing approaches and tools to discover knowledge hidden in these databases is the need of the hour. Knowledge Discovery in Databases (KDD) is defined as "The non-trivial extraction of implicit, previously unknown, and potentially useful information from data". KDD involves several steps which are as follows: Data cleaning, Data Integration, Data selection, Data transformation, Data mining, Pattern evaluation and knowledge representation. Of these, Data mining is the pivotal step. Hierarchical clustering algorithms can usually find satisfiable clustering results. Although the hierarchical clustering technique is often portrayed as a better quality clustering approach, this technique does not contain any provision for the reallocation of entities, which may have been poorly classified at the early stage. Furthermore, most of the hierarchical algorithms are very computationally intensive and require much memory space Partitional clustering techniques create a one-level (unnested) partitioning of the data points. If K is the desired number of clusters, then partitional approaches typically find all K clusters at once. The partitional clustering technique is well suited for clustering a large dataset due to their relatively low computational requirements.

II. LITERATURE SURVEY

MadjidKhalilian*et al.* [1] have discussed that dimension reduction by means of vertical data reduction performed before employing clustering methods for exceedingly large and high dimensional data sets has the main disadvantage of reducing the quality of results. Still, extra carefulness has been recommended because dimensionality reduction methods unavoidably cause some loss of information or may impair the comprehensibility of the results, even disfiguring the real clusters. They have proposed a method for use in high dimensional datasets that improves the performance of the K-Means clustering method by employing divide and conquer technique with equivalency and compatible relation concepts. The proper precision and speed up of their proposed method have been proved by experimental results.

De Amorim, R.C [2] has presented a method for clustering by employing two pair-wise rules (must link and cannot link) and a single-wise rules (cannot cluster) single-wise rule that uses extremely restricted quantity of labeled data. They have demonstrated that the precision of results could be improved by including these rules in the intelligent k-means algorithm and verified the same by means of experiments where the actual number of clusters in the data has not been previously known to the method.

Dharmveer Singh Rajput [3] proposed a basic framework by integrating the hypothesis of rough set theory (reduct) and k-means algorithm for efficient clustering of high dimensional data. First, by discarding the superfluous attributes by means of the (reduct) concept of rough set theory, it has identified the low dimensional space in the high dimensional data set. Then, it has identified suitable clusters by employing the k-means algorithm on this low dimensional data (reduct). The fact that the framework increases the efficiency of the clustering process and the precision of the resultant clustering has been proved by their experiment on test dataset.

Mohammad Al Hasan*et al.* [4] have proposed, ROBIN, a method for initial seed selection in k-means types of algorithms. It imposes constraints on the chosen seeds that lead to better clustering when k-means converges. The constraints make the seed selection method insensitive to outliers in the data and also assist it to handle variable density or multi-scale clusters. Furthermore, the constraints make the method deterministic, so only one run suffices to obtain good initial seeds, as opposed to traditional random seed selection approaches that need many runs to obtain good seeds that lead to satisfactory clustering. They did a comprehensive evaluation of ROBIN against state-of-the-art seeding methods on a wide range of synthetic and real datasets.

Domenico Daniele Bloisi and Luca Iocchi*et al.* [5] have presented a clustering method based on k-means that has been implemented on a video surveillance system. Rek-means does not require specifying in advance the number of clusters to search for and is more precise than k-means in clustering data coming from multiple Gaussian distributions with different co-variances, while maintaining real-time performance. Experiments on real and synthetic datasets were presented to measure the effectiveness and the performance of the proposed method.

In [6], a Centroid Ratio is firstly introduced to compare two clustering results. This centroid ratio is then used in prototype-based clustering by introducing a Pairwise Random Swap clustering algorithm to avoid the local optimum problem of k-means. The centroid ratio is shown to be highly correlated to the mean square error (MSE) and other external indices. Disadvantages are Random swapping lead more computation complexity and also, the algorithm missed the global information to avoid the local optimum problem.

III. PROBLEM DEFINITION

From the date of research, various works have been done on clustering algorithms to ensure precise data partitioning. Despite various clustering algorithms come in practice, the evolution has begun from the k – means clustering algorithm as the primary source. However, k – means clustering algorithm has faced many challenges to accomplish precise clustering because of arbitrary initialization of centroids. Such arbitrary initial centroids have led the algorithm to converge in local optimal solutions (or) clusters, which are not possible clusters of the data. Hence, efforts have been made in the literature to improve k – means clustering algorithm by selecting initial centroids meaningfully. These improved algorithms have required understanding the data characteristics to initiate the cluster centroids. The viability behind the works have become complex because of wide data distribution, time series characteristics and high dimension.

In contrast, an ideology on modifying the cluster centroids at run – time has been introduced. These enhancements have shown promising outcome, because they have intended to find the global centroids throughout the process, rather than focusing on initialization part. Pairwise random swap is one of such enhancement strategies that have made k – means algorithm to elude from sticking with local optima. However, the strategy remains uncertain because of lack of knowledge of the data characteristics. Further, the strategy has not made the converging procedure to be aware of the quality improvement of centroids.

IV. PROPOSED METHODOLOGY

The proposed method of data clustering using systematic approach is explained in this section. In the existing method, random swapping method was used using MSE value. In order to improve the performance, systematic swapping approach is used here using DB index.

A. Block diagram

The block diagram of the proposed approach is given in figure 3. At first, input data is read out with user input k value and R value. Then, k means algorithm is performed ‘R’ number of times to find R different cluster centroids. Here, random initialization, distance finding and grouping is performed. The output set of R centroids is then given for systematic process where, centroids are swapped systematically based on the distance function. The swapped cluster centroids are then used to find DB index. The centorid set which have the minimum DB value is taken as final set.

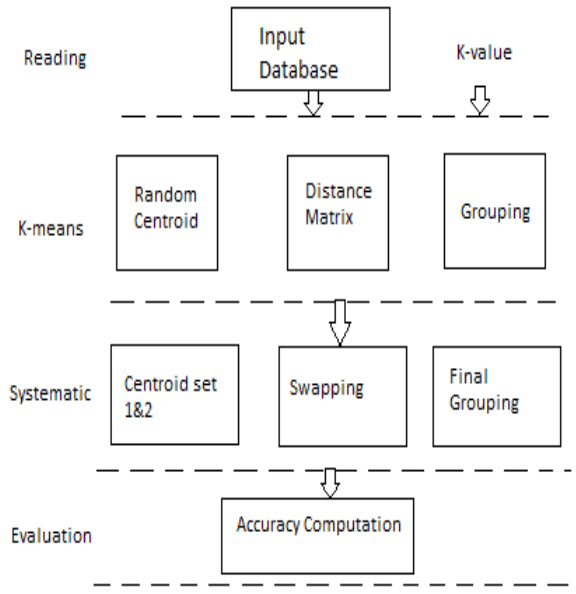

Figure 1. Block diagram of the proposed deterministic swapping clustering

B. *Description*

Step 1: K-means clustering

Data clustering is a common technique for statisticaldata analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The computational task of classifying the data set into *k* clusters is often referred to as ***k*-clustering**. The K-means algorithm is an iterative technique that is used to partition an image into *K*clusters. The *K*-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.

The k-means algorithm is an algorithm to cluster*n* objects based on attributes into *k*partitions, *k*<*n*. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance, or, the squared error function:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where there are *k* clusters S_i, i = 1, 2, ..., *k*, and μ_i is the centroid or mean point of all the points $x_j S_i$.

The most common form of the algorithm uses an iterative refinement heuristic known as Lloyd's algorithm. Lloyd's algorithm starts by partitioning the input points into k initial sets,either at random or using some heuristic data. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and algorithm repeated by alternate application of these two steps until convergence, which is obtained when the points no longer switch clusters (or alternatively centroids are no longer changed). A drawback of the k-means algorithm is that the number of clusters *k* is an input parameter. An inappropriate choice of *k* may yield poor results. The algorithm also assumes that the variance is an appropriate measure of cluster scatter.The initial centres are generated randomly to demonstrate the stages in more detail.

Step 2: Swapping

In swap-based clustering, the centroids are perturbed by a certain strategy in order to not get stuck in local minima. A swap is accepted if it improves the clustering quality. This trial-and-error approach is simple to implement and very effective in practice. The Random Swap algorithm (RS), originally called Randomized Local Search, is based on randomization: a randomly selected centroid is swapped to another randomly selected location. After that, a local repartition is performed and the clustering is fine-tuned by twok-means iterations. To ensure a good clustering quality, the number of iterations for random swap should be set large enough to find successful swaps..Deterministic swap aims at finding good swaps by a systematic analysis rather than by trial-and-error. In general, the clustering can be found in a few swaps only if the algorithm knows the centroid that should be swapped and the location where it should be relocated. Several heuristic criteria have been considered for the selection of the centroids to be swapped, but simple criteria such as selecting the clusters with the smallest size or variance do not work very well in practice. Other approaches remove one cluster, or merge two existing clusters as in agglomerative clustering. Deterministic removal takes Ndistance calculations for each of the Mclusters. Thus, the overall time complexity of the deterministic removal step becomes O(MN).

The replacement location of the swapped centroid can be chosen by considering the locations of all possible data points: this, however, would be very inefficient. In order to find the correct location, the task can be divided into two parts: select an existing cluster and select a location within this cluster. One heuristic selection is to choose the cluster that has the largest distortion (Eq. 1). The exact location within the cluster can be chosen considering the following heuristics: 1) current centroid of the cluster with small movement; 2) furthest data point; 3) middle point of the current centroid and furthest data point; 4) random.

The final objective is to do the systematic swapping among the cluster centroid. For example, C_1 and C_2 are taken and the distance among these two centroids are found out and swapping is done. After swapping, the Davies Bouldin Index is computed instead of MSE to obtain final cluster set. Let C_i be a cluster of vectors and D_j be d dimensional feature vector assigned to cluster C_i. A_i is the centroid of C_i and T_i is the size of the cluster i. S_i is a measure of scatter within the cluster. Usually the value of p is 2, which makes this a Euclidean distance function between the centroid of the cluster. $M_{i,j}$ is a measure of separation between cluster C_i and cluster C_j . $a_{k,i}$ is the k th element of A_i. The centroid set which having the minimum Davies Bouldin Index is considered as final centroid set.

V. MATHEMATICAL FORMULATION

Let us consider D be database having N points represented as, $D=\{d_1,d_2,...,d_N\}$. Here, every data point d_i have d-dimensional feature value, $d_i=\{f_1,f_2,...,f_d\}$. The objective is to find the k-centroids, $C=\{c_1,c_2,...,c_k\}$ by minimizing the most common cost function as like,

$$f=Min\left(\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{d}\left\|d_i-c_j\right\|^2;d_i\in c_j\right).$$

Based on this objective function, the final centroids is obtained. The same procedure is repeated for R number of times. Now, R number of centroid set is generated and it is denoted as, $C_R=\{C_1,C_2,...,c_{CR}\}$.

The final objective is to do the systematic swapping among the cluster centroid. For example, C_1 and C_2 are taken and the distance among these two centroids are found out and swapping is done. After swapping, the Davies Bouldin Index is computed instead of MSE to obtain final cluster set. Let C_i be a cluster of vectors and D_j be d dimensional feature vector assigned to cluster C_i. A_i is the centroid of C_i and T_i is the size of the cluster i. S_i is a measure of scatter within the cluster. Usually. the value of p is 2, which makes this a Euclidean distance function between the centroid of the cluster. $M_{i,j}$ is a measure of separation between cluster C_i and cluster C_j . $a_{k,i}$ is the k th element of A_i .

$$DB\equiv\frac{1}{N}\sum_{i=1}^{N}D_i$$

$$D_i\equiv\max_{j:i\neq j}R_{i,j}$$

$$R_{i,j}=\frac{S_i+S_j}{M_{i,j}}.$$

$$S_i=\frac{1}{T_i}\sum_{j=1}^{T_I}\left\|X_j-A_i\right\|_p$$

$$M_{i,j}=\left\|A_i-A_j\right\|_p=\left(\sum_{k=1}^{n}\left|a_{k,i}-a_{k,j}\right|p\right)^{\frac{1}{p}}$$

The centroid set which having the minimum Davies Bouldin Index is considered as final centroid set.

$$C_F=\underset{DB\in C_j}{MIn}\{Cj\}$$

VI. ALGORITHM

Input: D, k, R

Output: C_F

Procedure

1. **For** i=1 to R do

2.**Initialize Centroid** $C=\{c_1,c_2,...,c_k\}$

3. **Find** distance among c_i and d_j

4.**Assign**dj to relevant cluster c_i based on minimum value

5. **Find** new centroid C

6. **Go** to step 2 until convergence

7. **Add** to centroid list C_R

8. **End for**

9. **while** $S\neq 1 do$

10. $(C_1^{'},C_2^{'},DB_1^{'},DB_2^{'})=system\_Swap(D,M,C_1,C_2,DB_1,DB);$

11.**Find** C_F related to minimum DB

12. **end**

13. **return** C_F

VII. RESULTS AND DISCUSSION

This section presents the experimental results of the Pairwise Deterministic Centroid swapping Strategy for K – Means Clustering and the detailed discussion of the results obtained. Here, three different datasets are used for experimentation and the performance of the proposed algorithm is analysed in terms of clustering accuracy.

A. Description of Datasets

Iris: The data set contains three categories of 50 objects each, where each category refers to a type of iris plant. One category is linearly separable from the other two; the latter are not linearly separable from each other. There are 150 instances with four numeric features in iris data set.

B. Performance Metrics

A number of metrics for comparing clustering algorithm were recently proposed in the literature. The performance metrics used here is clustering accuracy. **Clustering Accuracy*:*** Accuracy refers to the degree of closeness of measurement of

a quantity to its actual value. Clustering Accuracy is the measure of closeness of the cluster formed as a result of the proposed algorithm to the required value which means how much accurate the members of a cluster are. The clustering accuracy is computed using the following formula.

$$CA = \frac{1}{N}\sum_{i=1}^{T} X_i$$

Where, N is the number of data point and T is the number of class.

C. *Experimental results*

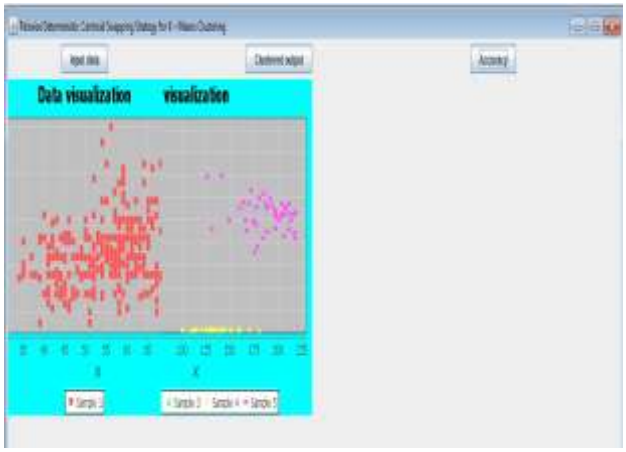

Figure 2. GUI of the proposed algorithm

D. *Performance Analysis*

The performance of the proposed Pairwise Deterministic Centroid swapping Strategy for K – Means Clustering is analysed in this section.

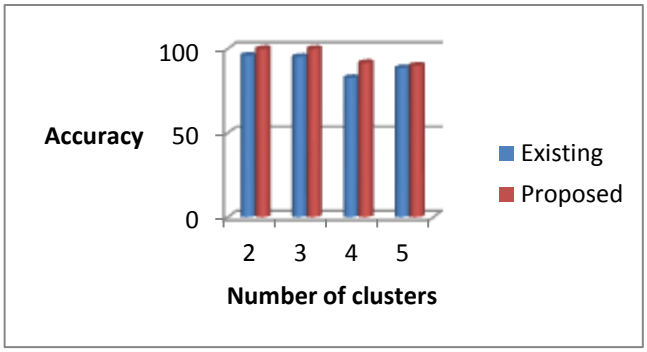

Figure 3. Clustering accuracy graph of iris data

VIII. CONCLUSION

Clustering algorithm and cluster validity are two highly correlated parts in cluster analysis. Here, idea for cluster validity and a clustering algorithm based on the validity index are introduced. A Centroid Ratiois firstly introduced to compare two clustering results. This centroid ratio is then used in prototype-based clustering by introducing a Pairwise systematic Swap clustering algorithm to avoid the local optimum problem of k-means. The swap strategy in the algorithm alternates between simple perturbation to the solution and convergence toward the nearest optimum by k-means. The centroid ratio is shown to be highly correlated to the DB indices. Moreover, it is fast and simple to calculate.

REFERENCES

[1] Madjid Khalilian, Norwati Mustapha, MD Nasir Suliman and MD Ali Mamat "A Novel K-Means Based Clustering Algorithm for High Dimensional Data Sets", In Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong, 17-19 March, Vol.1,pp.503-507, 2010.

[2] de Amorim, R.C., "Constrained Intelligent K-Means: Improving Results with Limited Previous Knowledge", In Proceedings of the Second International Conference on Advanced Engineering Computing and Applications in Sciences, ADVCOMP '08, Valencia, Sept.29-Oct. 4, pp.176, 2008.

[3] Dharmveer Singh Rajput , P. K. Singh, Mahua Bhattacharya, "An Efficient and Generic Hybrid Framework for High Dimensional Data Clustering", In proceedings of International Conference on Data Mining and Knowledge Engineering (ICDMKE 2010), World Academy of Science, Engineering and Technology, Rome, April, pp.174-179, No.64, 2010.

[4] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed J. Zaki, "Robust Partitional Clustering by Outlier and Density Insensitive Seeding", Pattern Recognition Letters, Vol: 30, No: 11, pp: 994-1002, 2009.

[5] Domenico Daniele Bloisi and Luca Iocchi, "Rek-Means: A k-Means Based Clustering Algorithm", Lecture Notes in Computer Science, Springer Berlin, Vol: 5008, pp: 109-118, 2008.

[6] Qinpei Zhao and PasiFränti, "Centroid Ratio for a Pairwise Random Swap Clustering Algorithm", IEEE transactions on knowledge and data engineering, Vol. 26, No. 5, May 2014.

[7] Li, M.J. Ng, M.K. ; Yiu-ming Cheung ; Huang, J.Z., "Agglomerative Fuzzy K-Means Clustering Algorithm with Selection of Number of Clusters", IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 11, 2008.

[8] HuiXiong, Junjie Wu, Jian Chen, "K-Means Clustering Versus Validation Measures: A Data-Distribution Perspective", IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, Vol. 39, no. 2, 2009.

[9] Tzortzis, Likas, C.L., "The Global Kernel k -Means Algorithm for Clustering in Feature Space", IEEE Transactions on Neural Networks. vol. 20, n0.7, 2009.

[10]PradiptaMaji,"Fuzzy–Rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data", IEEE Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 41, No. 1, pp. 222-233, Feb. 2011.