International Journal of Artificial Intelligence
p-ISSN: 2407-7275, e-ISSN: 2686-3251

Original Research Paper

Optimizing K-Means Initial Number of Cluster Based Heuristic Approach: Literature Review Analysis Perspective

Harunur Rosyid[1,2*], Ramlah Mailok[2], Muhammad Modi Lakulu[2]

[1] *Universitas Muhammadiyah Gresik, Indonesia*
[2] *Universiti Pendidikan Sultan Idris, Malaysia*

Article History
Received:
26.09.2019

Revised:
01.11.2019

Accepted:
15.11.2019

***Corresponding Author:**
Harunur Rosyid
Email:
harun@umg.ac.id
mramlah@fskik.upsi.edu.my
modi@fskik.upsi.edu.my

Optimizing K-Means Initial Number of Cluster Based Heuristic Approach: Literature Review Analysis Perspective

Abstract: One popular clustering technique - the K-means widely use in educational scope to clustering and mapping document, data, and user performance in skill. K-means clustering is one of the classical and most widely used clustering algorithms shows its efficiency in many traditional applications its defect appears obviously when the data set to become much more complicated. Based on some research on K-means algorithm shows that Number of a cluster of K-means cannot easily be specified in much real-world application, several algorithms requiring the number of cluster as a parameter cannot be effectively employed. The aim of this paper describes the perspective K-means problems underlying research. Literature analysis of previous studies suggesting that selection of the number of clusters randomly cause problems such as suitable producing globular cluster, less efficient if as the number of cluster grow K-means clustering becomes untenable. From those literature reviews, the heuristic optimization will be approached to solve an initial number of cluster randomly.

Keyword: Clustering, Heuristic, K-Means, Number of Cluster.

DOI: 10.36079/lamintang.ijai-0602.40

1. Background

Clustering is an important tool for a variety of applications in data mining, statistical data analysis, data compression, and vector quantization. Clustering is an unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The goal of clustering is to group data into clusters such that the similarities among data members within the same cluster are maximal while similarities among data members from different clusters are minimal.

The clustering can be used in all research field include in education scope. The application of data mining methods in the educational sector is an interesting phenomenon. It sets to uncover the previously hidden data to meaningful information that could be used for both strategic as well as learning gains [1]. The detailed study of the access and revisit patterns of learners across groups and found that the difference in behavior is statistically significant [2]. It is able to classified the levels of respondent in a educational research. Common European Framework of Reference for Languages (CEFR) is one of the example of clustering levels of respondent proficiency. The framework includes six ascending levels of proficiency namely; Breakthrough (A1), Way stage (A2), Threshold (B1), Vantage (B2), Effective Operational Proficiency (C1) and Mastery (C2) Little (2007) in ÿaŁatay & Gürocak [3].

Clustering algorithms are basically divided into two categories: Partitioning Algorithms and Hierarchical algorithms. A partitioning clustering algorithm separates data set into a defined number of set in a single iteration while a hierarchical clustering divides data into smaller subsets in hierarchical manner [4]. K-means clustering is one of the classical and most widely used clustering algorithms developed by Macqueen [5].

K-means clustering is widely used for large number of datasets. K-Means clustering algorithm is very popular because of its ability to cluster a kind of huge data, and also outliers, quickly and efficiently. K-means clustering algorithm first randomly generates k initial cluster centroids. After several iterations of the algorithm, data can be classified into certain clusters by the criterion function, which makes the data close to each other in the same cluster and widely separated among clusters.

K-means is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. But it is very suitable for producing globular clusters [6]. Although K-means clustering shows its efficiency in many traditional applications its defect appears obviously when the data set become much more complicated. And adopting K-means algorithm to do cluster analysis on these data sets directly is less efficient. This problem is mainly caused by empirical selection of clusters number and random initial K-means centre [7]. In addition, as the number of clusters grow, for example to thousands of clusters, K-means clustering becomes untenable [8].

K-means requires a priori knowledge about the data or, in the worst case, guessing the number of clusters. The issue of determining “the right number of clusters” in K-Means has attracted considerable interest, especially in the recent years. Cluster intermix appears to be a factor most affecting the clustering results [9]. Different datasets have different number of clusters, which is difficult to known beforehand, and the initial cluster centroids are selected randomly, which will make the algorithm converge to the different local optimal. The number of clusters has to be known in advance for the conventional K-means clustering algorithm and moreover the clustering result is sensitive to the selection of the initial cluster centroids. This sensitivity may make the algorithm converge to the local optimal [9].

Literature analysis of previous studies suggesting that selection of number of clusters randomly cause problems such as suitable producing globular cluster, less efficient if as the number of cluster grow K-means clustering becomes untenable. Furthermore, different dataset has the different number of clusters and the previous known of number cluster in advance for K-means initial number of cluster is sensitive to the selection initial cluster centroids. This sensitivity may make algorithm converge to the local optimal. From those literature reviews the heuristic optimization will be approached to solve initial number of cluster randomly. The aim of this research to analyze, investigate and minimalized local minimal problem based clustering problem in K-means Clustering. In the next section will be describe several literature studies about K-means Problem.

2. Related Work

This section describe several studies related to K-means and optimization of K-means are examined to show the approach that has been taken by several researchers in order to find the problem of K-means

DOI: 10.36079/lamintang.ijai-0602.40

and problems that are common in the process of clustering using the K-means algorithm. Clustering is a fundamental problem that frequently arises in a great variety of application fields such as pattern recognition, machine learning, statistics, etc. It is a formal study of algorithms and methods for grouping or classifying objects without category labels. The resulting partition should possess two properties: (1) homogeneity within the clusters, i.e. objects belonging to the same cluster should be as similar as possible, and (2) heterogeneity between the clusters, i.e. objects belonging to different clusters should be as different as possible. Clustering algorithms are generally classified as hierarchical clustering and partitioned clustering. Hierarchical clustering groups data objects with a sequence of partitions, either from singleton clusters to a cluster including all individuals or vice versa. Hierarchical procedures can be either agglomerative or divisive: agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters; divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Partitioned Clustering attempt to divide the data set into a set of disjoint clusters without the hierarchical structure. The most popular partitioned clustering algorithms are the prototype-based clustering algorithms where each cluster is represented by the centre of the cluster and the used objective function (a square-error function) is the sum of the distance from the pattern to the centre. The most popular class of clustering algorithms is K-means algorithm which is a centre based, simple and fast algorithm. It partitions the input dataset into clusters (k). Each cluster is represented by an adaptively-changing centroid (also called cluster centre), starting from some initial values named seed-points.

2.1. K-Means Problem

K-means is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. But it is very suitable for producing globular clusters [6]. Although K-means clustering shows its efficiency in many traditional applications its defect appears obviously when the data set become much more complicated. And adopting K-means algorithm to do cluster analysis on these data sets directly is less efficient. This problem is mainly caused by empirical selection of clusters number and random initial K-means [7]. In addition, as the number of clusters grow, for example to thousands of clusters, k-means clustering becomes untenable [8].

K-means requires a priori knowledge about the data or, in the worst case, guessing the number of clusters. The issue of determining "the right number of clusters" in K-Means has attracted considerable interest, especially in the recent years. Cluster intermix appears to be a factor most affecting the clustering results [9]. Different datasets have different number of clusters, which is difficult to known beforehand, and the initial cluster centroids are selected randomly, which will make the algorithm converge to the different local optimal. The number of clusters has to be known in advance for the conventional K-means clustering algorithm and moreover the clustering result is sensitive to the selection of the initial cluster centroids. This sensitivity may make the algorithm converge to the local optimal [9].

2.2. Initial Number of Cluster

K-means starts with K number of predefined clusters and then assigns each data member to its closest cluster. It is clear that the number of clusters cannot be easily specified in many real world applications and datasets; therefore, the above mentioned algorithms requiring number of clusters as a parameter cannot be effectively employed. On behalf of these understanding, finding the "optimum" number of clusters in a data set has become an important research area.

2.3. Initial Centroid

K-means starts with randomly initial cluster centroids and keeps reassigning the data objects in the dataset to cluster centroids based on the similarity between the data objects and the cluster centroids.

2.4. Grouping Object into Cluster

K-means computes the squared distances between the inputs (also called input data points) and centroids, and assigns inputs to the nearest centroid. An algorithm for clustering N input data points $x_1,x_2,...,x_N$ into k disjoint subsets C_i , $i=1,...,k,$ each containing n_i data points,$0<n_i<N$, minimizes the following mean-square-error (MSE) cost-function:

DOI: 10.36079/lamintang.ijai-0602.40

$$Jmse = \left(\sum_{i=1}^{K} \sum_{x_i \in C_i} \left(\| X_i - C_i \| \right) \right)^2 \quad (1)$$

x_t is a vector representing the *t-th* data point in the cluster C_i and c_i is the geometric centroid of the cluster C_i. Finally, this algorithm aims at minimizing an objective function, in this case a squared error-function, where $\| X_i - C_i \|$ is a chosen distance measurement between data point *xt* and the cluster centre *ci* The K-means algorithm assigns an input data point x_t into the *ith* cluster if the cluster membership function $I(x_t,i)$ is 1.

$$I(x_t, i) = \begin{cases} 1 & if \quad i = \arg\min \left(\| x_t - c_j \| \right)^2 \quad j = 1,...k \\ 0 & otherwise \end{cases} \quad (2)$$

Here $c_1, c_2, c_j, ..., c_k$ are called cluster centres which are learned by the following step

Step 1 : Initialize *k* cluster centres $c_1, c_2, ..., c_k$ by some initial values called seed-points, using random sampling. For each input data point x_t and all *k* clusters, repeat steps 2 and 3 until all centres converge.

Step 2 : Calculate cluster membership *function* $I(x_t,i)$ *by Eq.(2) and decide the* membership of each input data point in one of the *k* clusters whose cluster centre is closest to that point.

Step 3 : For all *k* cluster centres, set c_i to be the centre of mass of all points in cluster C_i.

2.5. Optimization Number of Cluster

Solving the selection of a correct cluster number has been tried in two ways. The first one invokes some heuristic approaches. The clustering algorithm is run many times with the number of clusters gradually increasing from a certain initial value to some threshold value that is difficult to set. The second is to formulate cluster number selection by choosing a component number in a finite mixture model [10]. One of these techniques is heuristic optimization methods based on the bees behavior that try to optimize a pre-defined function that can be very useful in data clustering. Karaboga & Ozturk [11] successfully applied Artificial Bee Colony algorithm to clustering for the purpose of classification. And also to improve perform of ABC, several studies of hybrid ABC for K-means clustering problem has been proposed, for instance determined number cluster and initial cluster automatically has improved version of discreter artificial bee(disABC) which specified in many real world application and datasets. Forsati et al. [8] has presented an improved bee colony optimization algorithm, dubbed IBCO, by introducing cloning and fairness concepts into the BCO algorithm and make it more efficient for data clustering. And the latest is Kishor, Singh, & Prakash [12] develop SABC: Non-dominated sorting based multi-objective artificial bee colony algorithm and its application in data clustering. In other side researchers solved the clustering problem by stochastic optimization methods using genetic algorithms. presented a genetic algorithm for evolving the cluster centres in the K-means algorithm [13]. Liu, Wu, & Shen [14] develop a genetic algorithm based clustering method called automatic genetic clustering for unknown K(AGCUK) to automatically find the number of clusters and provide the proper clustering partition. ABC and GA are algorithm that used researchers to optimize K-means clustering and other optimization problem, there is some way to use or combine both these algorithm to solve clustering and other optimization problem.

3. Conclusion

Research in K-Means clustering algorithms as applied to educational context and will also be working towards generating a unified clustering approach such that it could easily be applied to any educational institutional dataset without any much overhead. Although K-means has been widely used in data analyses, pattern recognition and image processing, it has three major limitations: The number of clusters must be previously known and fixed, number of cluster (K) influence centre of cluster (centroid). The results of K-means algorithm depend on initial cluster centres (initial seed-points). The algorithm contains the dead-unit problem.

Based on literature to optimize initial number of cluster without known in advance this research will be develop, especially for the case of data clustering unknown number of cluster especially segmentation and big data heuristic approach can be adopted in future research.

REFERENCES

[1] Dutt, A., Aghabozrgi, S., Akmal, M., Ismail, B., & Mahroeian, H. (2015). Clustering Algorithms Applied in Educational Data Mining, 5(2), 112–116. https://doi.org/10.7763/IJIEE.2015.V5.513

[2] Roy, D. (2017). Synthesis of clustering techniques in educational data mining.

[3] ÿaŁatay, S., & Gürocak, F. ÿnveren. (2016). Is CEFR Really over There? Procedia - Social and Behavioral Sciences, 232(April), 705–712. https://doi.org/10.1016/j.sbspro.2016.10.096

[4] Dunham, M. H. (2003). Data Mining Introductory and Advanced Topics. Prentice Hall/Pearson Education.

[5] Macqueen, J. (n.d.). Some Methods For Classification And Analysis Of Multivariate Observations, 233(233), 281–297.

[6] Sharmilarani, D., & Kousika, N. (2014). Modified K-Means Algorithm for Automatic Stimation of Number of Clusters Using Advanced Visual Assessment of Cluster Tendency, 236–239.

[7] Wang, X., Jiao, Y., & Fei, S. (2015). Estimation of Clusters Number and Initial Centers of K-means Algorithm Using Watershed Method, (0). https://doi.org/10.1109/DCABES.2015.132

[8] Forsati, R., Keikha, A., & Shmasfard, M. (2015). Author ' s Accepted Manuscript An Improved Bee Colony Optimization Algorithm with an Application to Document Clustering. Neurocomputing. https://doi.org/10.1016/j.neucom.2015.02.048

[9] Xiao, J., Yan, Y., Zhang, J., & Tang, Y. (2010). Expert Systems with Applications A quantum-inspired genetic algorithm for k -means clustering. Expert Systems With Applications, 37(7), 4966–4973. https://doi.org/10.1016/j.eswa.2009.12.017

[10] Z, K. R. (2008). An efficient k-means clustering algorithm, 29, 1385–1391. https:// doi.org/10.1016/j.patrec.2008.02.014

[11] Karaboga, D., & Ozturk, C. (2011). A novel clustering approach : Artificial Bee Colony (ABC) algorithm, 11, 652–657. https://doi.org/10.1016/j.asoc.2009.12.025

[12] Kishor, A., Singh, P. K., & Prakash, J. (2016). NSABC: Non-dominated sorting based multi-objective artificial bee colony algorithm and its application in data clustering. Neurocomputing, 216, 514–533. https://doi.org/10.1016/j.neucom.2016.08.003

[13] Michael Laszlo, S. M. (2006). A genetic algorithm using hyper-quadtrees for low-dimensional k-means clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(4), 533–543.

[14] Liu, Y., Wu, X., & Shen, Y. (2011). Automatic clustering using genetic algorithms. Applied Mathematics and Computation, 218(4), 1267–1279. https://doi.org/10.1016/j.amc.2011.06.007