Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

An Efficient Approach for Detection of Failure in Data Analysis by Using Proposed Modified K-Means Clustering

Ms. Sonia Yadav[1], Dr. Sachin Sharma[2]

[1]Associate Professor, Deshbandhu College, Delhi University.

[2]Associate Professor, FCA, MRIIRS.

Article Info
Page Number: 593-599
Publication Issue:
Vol. 72 No. 1 (2023)

Article History
Article Received: 15 October 2022
Revised: 25 November 2022
Accepted: 14 December 2022
Publication: 03 January 2023

Abstract:
The advent of modern scientific data collection techniques has led to the accumulation of large amounts of information in various fields. Traditional database reference methods are not sufficient to extract useful information from large amounts of data. Cluster analysis is one of the main methods of data analysis, and the k-mean cluster algorithm is widely used in many practical applications. However, the first k-mean algorithm is expensive to calculate, and the quality of the resulting clusters depends largely on the choice of the original centroid. Clustering is an uncontrolled data acquisition (machine learning) technique used to insert data elements into relevant groups without prior knowledge of the group definition. One of the most common and widely studied grouping methods that reduces the error in grouping points in Euclidean space is the K-mean grouping. However, it is known that the k-mean method approaches one of the many local minimums and that the final result depends on the starting points (tools). In this study, we introduced an algorithm for starting a k-tool using the appropriate starting points (tools). Sufficient starting points allow the k-tool to be brought closer to the local minimum; The number of iterations in all data sets is reduced.

Keywords: Data Analysis, Clustering, k-means Algorithm, Modified k-means Algorithm.

1. Introduction

Cluster analysis is based on different types of differences between objects and uses adaptations of the distance function to classify the model. Whether classification really matters depends on how the vector symbols of the model are distributed. If the contribution of vector points is grouped and test points from the same group are concentrated and test points from different groups are distant, it will be easy to use point ranking functions that, as far as possible, statistics in the same group. group will be similar and the statistics in another group will be different. The point distance function can act as a measure of model similarity. Depending on the proximity of the distance to the points, the measurement can be used to classify ideas. In this article, we combine the longer minimum distance algorithm and the traditional K-Means algorithm to offer an updated K-Means grouping algorithm. This updated algorithm can compensate for the shortcomings of the traditional K-Means algorithm for determining the starting focal point. The updated K-Means algorithm effectively solves the disadvantage that the traditional K-Means algorithm relies too much on the choice of starting focal points.

2. K-Means algorithm

"The division-based K-Means algorithm is a type of grouping algorithm and is proposed by J.B. MacQueen". This uncontrolled algorithm is often used for data analysis and pattern recognition. The square error and error criteria are the basis of this algorithm to reduce the cluster performance index. In order to find the optimization result, this algorithm tries to find the K division to meet certain criteria. First, select a number of points to represent the group's starting focus (usually we choose the first input in Example K to represent the group's starting point); second, we collect the remaining sampling points at the focal point according to the minimum distance criteria, take the first classification, and if the classification is unreasonable, we change it (recalculate the focus of each group) and repeat several times. until a reasonable estimate emerges. "The section-based K-Means algorithm is a type of cluster algorithm that has advantages in terms of briefness, efficiency, and speed". However, this algorithm is highly dependent on the differences in the selection of starting points and the initial sample, which always leads to different results. Moreover, this algorithm, based on objective functions, always uses the gradient method to achieve extremism. The search direction of the gradient method is always in the direction of decreasing energy, so that when the start focus of the cluster is not suitable, the whole algorithm will easily sink to the local lowest point.

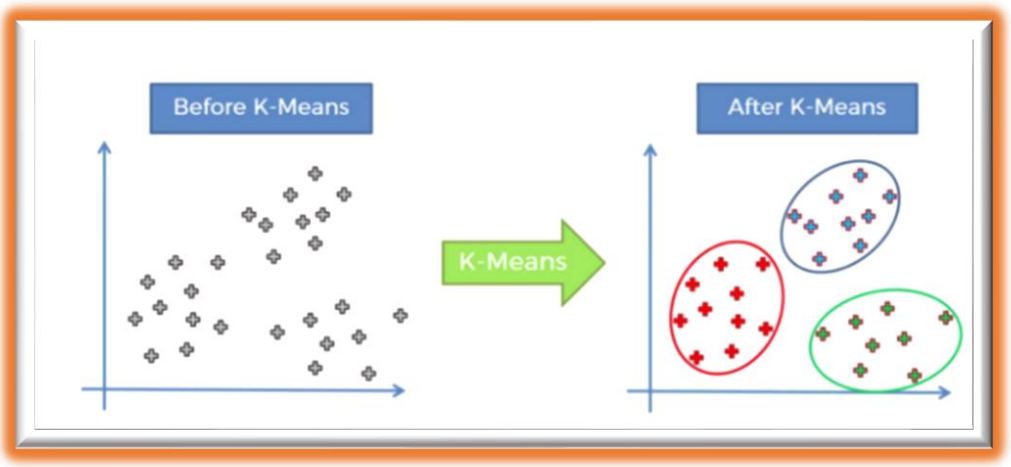

Figure 1: "K-means Clustering"

3. Working of K-Means Algorithm:

To process the training data, the K-tool in data extraction starts with a first group of randomly selected centroids, which are used as starting points for each group, and then performs iterative (repetitive) calculations to optimize the positions of the centroids.

Stops cluster creation and optimization when:

- Centroids have stabilized: there are no changes in their values, as the grouping was successful.
- The defined number of iterations has been reached.

The simplest algorithm for the traditional K-means is as follows;

"**Input: D = {d1, d2, d3, ...dn} // set of n numbers of data points"**

"K // The number of desire Clusters Output"

A set of k clusters

1. Select k points as initial centroids.

2. Repeat

3. From K clusters by assigning each data point to its nearest centroid.

4. Recomputed the centroid for each cluster until centroid does not change.

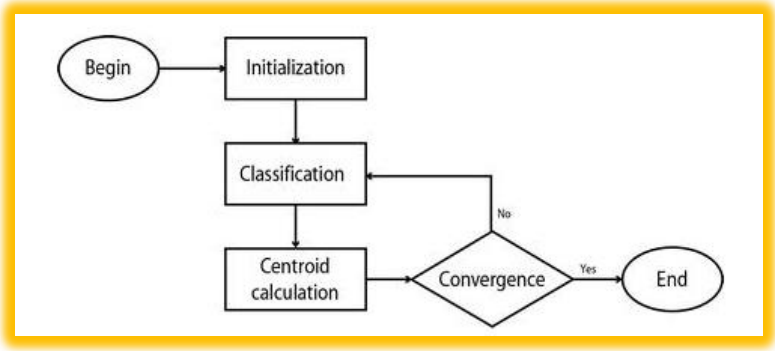

Figure 2: "Standard K-means Algorithm"

However, the algorithm has its advantages and disadvantages, which is as follows;

Advantages:

1. This is a relatively faster clustering technique.

2. Works quickly with a large set of data, since the complexity of time is O (nkl), where n is the number of patterns, k is the number of clusters and l is the number of iterations.

3. It relies on the Euclidean distance, which makes it good with numerical values of motivating geometric and statistical significance.

Disadvantages:

1. The initial assumption of a value for K is very important, but there is no adequate description for accepting the value of K and, therefore, for different values of K it will generate a different number of groups.

2. The initial centroids of the cluster are very important, but if the centroid is far from the centre of the data cluster, they lead to endless iterations, which sometimes lead to incorrect grouping.

3. K-means clustering is not good enough with the grouping of the noise data.

4. Modified K-means Clustering Algorithm (Proposed Algorithm)

Based on the research methodology based on some validated K- means algorithms, there were components that could be improved to achieve more accuracy and efficiency by changing the usual K-methods. These are the sections discussed in this section and refer to the validation books and methods. According to the study, it is clear that we should think of a better idea or find a method to determine the first centroids by providing data to groups of closed centroids after a successful upgrade to improve results of normal K -means.

Important Equations

Measuring the distance will determine how the similarity of two elements is calculated and will affect the shape of the clusters.

Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

Where n is the number of dimensions (attributes) and pk and qk are, respectively, the kth attributes (components) or data objects p and q.

The steps to initialize the centroids using Modified K-Means Clustering

The first cluster is selected equally at any given time from the data we want to collect. This is similar to what we do with K-Means, but instead of selecting all the centroids, we select only one centroid here:

- Find the average value for the given data points.
- Next, we calculate the distance (D (x)) of each data (x) from the average value by using equation.
- Then, select the new location from the database with the probability of x being proportional to (D(x))2
- Then we repeat steps 2 and 3 unless groups (k clusters) are selected.

Modified Technique

Part 1: Determine initial focal point

"Determine first focal point from the given number of same pattern (Dataset)".

Step 1.1: "Input Dataset Step"

Step 1.2: "Check the Each attribute of the Records Step".

"Check the same dataset from the given pattern".

Step 1.3: "Find the average value for the given pattern".

Step 1.4: "Find the distance for each data point from mean value using Equation".

If

"The distance between the average value is minimal, so it will be stored in. Then Divide data pattern into k cluster points don't needs to move to other clusters".

Else

"Recalculate distance for each data point from average value using Equation until divide data pattern into k cluster ".

Part 2: Allocating data points to nearest focal point

Step 2.1: "Calculate Distance from each data point to focal".

"Assign data points to its nearest focal to form clusters and stored values for each data".

Step 2.2: "Calculate new focal point for these clusters".

Step 2.3: "Calculate distance from all focal to each data point for all data points".

If

"The Distance stored previously is equal to or less then Distance stored in Step 2.1"

Then

"Those Data points don't need to move to other clusters".

Else

"From the distance calculated assign data point to its nearest focal point by comparing distance from different focal points".

Step 2.4: "Calculate focal points for these new clusters again".

"Until The convergence condition met".

"Output A Set of K clusters".

5. Conclusion:

A modified K-means be based on the two phases:

- Determine initial Focal Point.
- Distribution of data points to the closest focal point.

Conventional K-media clustering is a widely used method, but it relies on the prospect of starting centroids and sharing data in close quarters. There are advantages rather than disadvantages in classifying k-averages but some improvements are still needed. This study describes methods that improve methods for pre-determining centroids and provide visual acuity in their proximity to more complex and time-sensitive O (n), which is much faster than normal k characteristics. The initial value of K (number of clusters) is still a concern because

it can improve the accuracy of the cluster, which will be improved by improving the normal appearance in the future.

References

1. Zhai, D.; Yu, J.; Gao, F.; Lei, Y.; Feng, D. K-means text clustering algorithm based on centers selection according to maximum distance. *Appl. Res. Comput.* **2014**, *31*, 713–719.
2. Sun, J.; Liu, J.; Zhao, L. Clustering algorithm research. *J. Softw.* **2008**, *19*, 48–61.
3. Li, X.; Yu, L.; Hang, L.; Tang, X. The parallel implementation and application of an improved k-means algorithm. *J. Univ. Electron. Sci. Technol. China* **2017**, *46*, 61–68.
4. Kanungo, T.; Mount, D.M.; Netanyahu, N.S.; Piatko, C.D.; Silverman, R.; Wu, A.Y. An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 0–892.
5. Wagstaff, K.; Cardie, C.; Rogers, S.; Schrödl, S. Constrained k-means clustering with background knowledge. In Proceedings of the Eighteenth International Conference on Machine Learning, Williamstown, MA, USA, 28 June–1 July 2001; pp. 577–584.
6. Huang, Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Min. Knowl. Discov.* **1998**, *24*, 283–304.
7. Narayanan, B.N.; Djaneye-Boundjou, O.; Kebede, T.M. Performance analysis of machine learning and pattern recognition algorithms for Malware classification. In Proceedings of the 2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation
8. Summit (OIS), Dayton, OH, USA, 25–29 July 2016; pp. 338–342.
9. Narayanan, B.N.; Hardie, R.C.; Kebede, T.M.; Sprague, M.J. Optimized feature selection-based clustering approach for computer-aided detection of lung nodules in different modalities. *Pattern Anal. Appl.* **2019**, *22*, 559–571.
10. Narayanan, B.N.; Hardie, R.C.; Kebede, T.M. Performance analysis of a computer-aided detection system for lung nodules in CT at different slice thicknesses. *J. Med. Imag.* **2018**, *5*, 014504.
11. Wang, Q.; Wang, C.; Feng, Z.; Ye, J. Review of K-means clustering algorithm. *Electron. Des. Eng.* **2012**, *20*, 21–24.
12. Ravindra, R.; Rathod, R.D.G. Design of electricity tariff plans using gap statistic for K-means
13. clustering based on consumers monthly electricity consumption data. *Int. J. Energ. Sect. Manag.* **2017**, *2*, 295–310.
14. Han, L.; Wang, Q.; Jiang, Z.; Hao, Z. Improved K-means initial clustering center selection algorithm. *Comput. Eng. Appl.* **2010**, *46*, 150–152.
15. UCI. UCI Machine learning repository. Available online: **http://archive.ics.uci.edu/ml/** (accessed on 30 March 2019).
16. Tibshirani, R.; Walther, G.; Hastie, T. Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. Ser. B (Statist. Methodol.)* **2001**, *63*, 411–423.
17. Xiao, Y.; Yu, J. Gap statistic and K-means algorithm. *J. Comput. Res. Dev.* **2007**, *44*, 176–180.

18. Kaufmn, I.; Rousseeuw, P.J. *Finding Groups in Data an Introduction to Cluster Analysis*; New York John Wiley&Sons: Hoboken, NY, USA, 1990.
19. Esteves, K.M.; Rong, C. Using Mahout for clustering Wikipedia's latest articles: A comparison between K-means and fuzzy c-means in the cloud. In Proceedings of the 2011 Third IEEE International Conference on Science, Cloud Computing technology and IEEE Computer Society, Washington, DC, USA, 29 November–1 December 2011; pp. 565–569.
20. Yu, C.; Zhang, R. Research of FCM algorithm based on canopy clustering algorithm under cloud environment. *Comput. Sci.* **2014**, *41*, 316–319.
21. Mccallum, A.; Nigam, K.; Ungar, I.H. Efficient clustering of high-dimensional data sets with application to reference matching. In Proceedings of the Sixth ACM SIUKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, USA, 20–23 August 2000; pp. 169–178.