International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Theoretical Approach to Optimizing A k-Means Clustering Algorithm in Data Science/Big Data: with a view to Artificial Intelligence

Dasaka VSS Subrahmanyam[1], K. Venkatesh Sharma[2], V. Padmakar[3], M. Mohan Veer[4]

[1]Professor, Dept of CSE, Keshav Memorial Engineering College, Hyderabad
[2]Professor, Dept of CSE, CVR Engineering College, Hyderabad
[3]Assoc. Professor, Dept of CSE, Neil Gogte Institute of Technology, Hyderabad
[4]Asst. Professor, Dept of CSE, Keshav Memorial Engineering College, Hyderabad

Abstract
Applications of Artificial Intelligence have been penetrating deeply into various kinds of domains, at faster rates, such as data science. The general k-means clustering algorithm may not properly deal with larger data sets. So, optimization techniques such as an optimized clustering algorithm for efficient decision makings are necessary to improve the performance efficiency of k-means clustering algorithm further by considering standard deviation and variance of the given data set, to deal with large data sets in data science with a view to Artificial Intelligence.

Keywords: k-means clustering algorithm, Mean, Median, Mode, Optimization technique, Partitions, Standard Deviation, Variance.

Introduction

k-means clustering algorithm is aimed at partitioning n elements (observations) into a finite number of k clusters in which every single observation belongs to any one of the clusters, with the nearest mean, without leaving even a single observation[1]. Here, the letter k indicates the number of divided partitions (clusters), of any given data. If the entire data set is divided into 20 partitions (clusters), then is called 20-means problem (algorithm). Then this algorithm will have 20 centroids. If the data set was divided into 45 partitions, then it is called 45-means algorithm. Then this algorithm will have 45 centroids. The main aim of this clustering algorithm is to find a centroid (also called as cluster centroid or cluster center) of every individually divided partitions. Finding an accurate centroid for every individual partition is crucial in k-means clustering algorithm [2].

The nature and behavior of data are to be observed carefully before going for partitions. Once the entire chosen data set is divided into possible finite number of partitions, then density of observations is to be considered. In clustering algorithms, in general, a few observations are selected randomly selected in every divided partition set as per the density or concentration of distribution of observations. The distances of all other observations, from randomly selected few points, are calculated by Euclidian distance method. Here, let the coordinates of one of the randomly selected points be P1(x1, y1) and let the coordinates of

the point, to which distance is to be measured, be P2 (x2, y2). All these points are to be considering in 2-dimensional x y-plane [2].

The distance is measured by using the following formula

$$D(P1, P2) = \sqrt{\{(x2 - x1)^2 + (y2 - y1)^2\}} \quad \text{---------- (1)}$$

where D (P1, P2) indicates the Euclidian distance between the points P1 and P2 in x y- plane.

In this way, all Euclidian distances in a particular partition are computed and observations with closer distances with respect to the randomly selected points for centroids are detected, for every randomly selected point. This process is repeated continuously for all randomly selected points. To which randomly selected point for centroid, all surrounding observations are closer, that point will be selected as a centroid of that particularly divided partition [1]. This method is repeated iteratively till a suitable centroid is found, for all other divided partitions. These partitions, in general, are done on the basis of thickly populated observations, thinly populated observations and moderately populated observations. The division of partitions is efficient up to a particular size of the data set.

The performance analysis of the clustering algorithm may be degraded, once the quantity of the data set increases proportionately [2]. In this paper, some measures are incorporated to improve the performance analysis of k-means clustering algorithm with some optimization techniques, which is called optimized k-means clustering algorithm. This step is inevitable with a view to Artificial Intelligence, as there may not be limitations on the quantum of input data to be given to a data set. It is necessitated as Artificial Intelligence has been penetrating into many fields at faster rates with vast applications.

Clustering algorithms are applied in finding solutions to cybercrime detections, delivery route optimizations, fraud detection cases, finding consumer behavior, predicting attrition rate, moment of consumer durables, and so on. Thus, an optimized k-means clustering algorithm is needed to strengthen its wings of providing advantages to many other fields as well. Some important measures are to be taken into consideration for strengthening this algorithm on the basis of the nature and behavior of observations.

The following measures are to be incorporated to optimize k-means clustering algorithm –

- To begin with, observe the nature of the data, whether homogeneous data or non-homogeneous data, to be considered. Since, all data are unlabeled data (of unsupervised learning), every observation is to be considered as a prominent one (as it is not the case of supervised learning). Proper distinction is needed to detect between homogeneous data and non-homogeneous data. Otherwise, data should be segregated.
- Detect for any extreme observations (values) involved. In most of the cases, it is fair not to consider those values into account for determining centroids. The determining position of centroids won't be affected by excluding some extreme observations from the given data. Otherwise, these extreme values might affect the determining the positions of centroids of the concerned divided partitions of the given data set.
- Inspect for the way of scattering of observations on x y – plane. Verify for density of observations, whether they are thickly scattered, thinly scattered or uniformly scattered. If observations are thickly scattered on x y - plane, then partitions are to be made with more practical observations. Finding centroids are easy in case of thickly scattered observations as almost all observations are very closed to the centroid.
- If more observations are thickly populated and they have some thinly populated observations surrounding it, then some sub-partitions are advised within a partition to include all thinly populated

observations in sub-partitions. Then corresponding centroids are to be evaluated. Thus, dividing a partition into sub-partitions will provide for more meaningful centroids with respect to their positions which covered all observations. Thus, find a major centroid of all sub-centroids of sub-partitions. This provides an easy way to know the distribution analysis of all scattered observations of all sub-partitions within a partition.

- Try for various ways of partitions to make all observations are closer to the corresponding centroid. Euclidian distances are iteratively evaluated to make adjustments in the positions of centroids. Always try for sub-partitions within a partition for locating an exact sub-centroid for every sub-partition. It provides a proper mathematical analysis over the entire data set.
- In some cases, thinly populated observations can be separately grouped together. It is, keenly, to be verified that whether thinly scattered observations are to be grouped together with thickly populated observations are not. A sincere decision is to be taken upon calculating Euclidian distances for every single observation.
- To make accurate locations of centroids, it is advisable to follow the standard deviation (SD) (σ) and variance (ν) of all observations of all sub-partitions of the partition by using the following formulae –

SD (σ) = $\sqrt{\{(x_i - \mu)^2 / n\}}$, where x_i indicates every individual data for i = 1, 2, 3, ……..n and n is the number of elements or values or observations, of the given data set.

And variance $\nu = \sigma^2$.

Thus, find standard deviation (σ) and variance (ν) of every sub-partition of the partition, then find the same for the corresponding partition.

Compare these two values. If the difference between these two evaluations are of very small value, then the chosen centroid position is accurate.

Accurate location of centroid = {(σ of a partition) – (μ of σ of all sub-partitions) and (ν of the partition) – (μ of ν of all sub-partitions)} ≈ very small negligible value, where μ is mean.

If any one of the above two observations differs with high values, then modifications are to be done in highly varying domains to reduce their high deviations.

- At every stage, mean (μ), median and mode are to be calculated for the concerned data set of every sub-partition and partitions. The same are to be computed for the entire data set as a whole. All these are to be tabulated in a table for further observations. Find the discrepancy among means of sub-partitions, partitions and the entire date set. Similarly, find the same with mean and mode operations too. Verify for it whether there are any fluctuating differences or a uniform difference. In case of a uniform difference, it can be assumed that centroids are properly located. Otherwise, modifications are to be altered accordingly.
- Once all centroids are evaluated, then an ultimate final centroid of the entire data set is to be evaluated. This final centroid provides final observations of the standard deviation and variance of the entire data set. Thus, for every data set, different kinds of centroids are to be found out for making different kinds of analysis for various purposes.
- All unsupervised algorithm techniques can be converted in supervised learning algorithms by means of using statistical tools such as standard deviation and variance. Thus, an optimized k-means clustering algorithm can be under the category of supervised learning techniques.
- A finite set of centroids is thus formed to make analytical decisions from a given data set. Finding an overall centroid (only one centroid for the entire data set), from an evaluated set of centroids, provides further deep analysis by applying the concepts of circumcenter and incenter (of 2-Dimensional

Geometry). It is an efficient method of binding all observation of a given data set to evaluate centroids. This algorithm is very much useful in problems like cybersecurity, consumer behavior, financial applications, fraud detection, computer networks, network security and so on.

- Sometimes, the physical shape of data sets may be Geometrical like circles, semi-circles, triangles (isosceles, equilateral, right-angled triangles), rectangle, square, polygon etc., Computing centroids will be easy, in all these cases, as they directly follow Geometric properties. So, it will be efficient to reduce some possible partitions data sets into possible Geometric shapes. If needed, pruning techniques can be applied to curtail some observations. Pruned observations are not to be excluded from the concerned partitions, instead they will be added to other partitions having the same characteristics.
- In rare cases, the physical shape of some data sets may have regression characteristics. These are special cases and the process of partitions will be an easy task, as these kinds of algorithms are directly used for predictions. Thus, all possible ways of making partitions are to be very keenly noticed, which has been the decision maker of deciding centroids. Every partition is to be observed with a mathematical view. So, it is important to know that the factor of deciding centroids of various partitions depends not only upon the density factor, but also depends upon their Geometrical shapes and their nature and behavior.
- Partitions can also be done on the basis of Median and Mode as well. But, in majority of case, mean-based partitions are followed. All observations are to arranged in an ascending order or descending order before going for partitions. This method works out, if majority of observations of any given data set are unique, otherwise, this method is not an efficient one (not for many repeated observations). Mode-based partitions are allowed in some cases as this method is based upon values with high frequency. This method is not suggestable one, in general.
- Thus, optimized algorithms are possible by using the available Mathematical and Statistical tools in an efficient way in deciding accurate centroids of partitions. This kind of approach is capable of handling huge data sets, irrespective of their sizes. It is very much needed with respect to the advancement of Artificial Intelligence. Sizes of data sets are dynamically changing from time to time.

Conclusion

By using the practical applications of statistics, many optimizing solutions can be obtained for various existing problems. Many optimization techniques can be used according to the existing initial and boundary conditions of problems. Mathematical solutions with Statistical approaches will provide more suitable solutions to complicated problems in data science and its allied branches with a view to Artificial Intelligence. Mathematical and Statistical tools and their applications provide more exact and accurate solutions to many problems in data science (Artificial Intelligence).

References

1. Peter Bruce, Andreco Bruce, Bruce Gedeck, “Practical Statistics for Data Scientists”, OTREILY publishers.
2. James D Miller, “Statistics for Data Science”, PACKT publishers. www.packt.com