IOSR Journal of Engineering (IOSR JEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 41-45

Survey on Selection of K and Initial Starting Points in K-Means Clustering Algorithm

Prof. Kedar Sawant[#1], Prof. Sagar Naik[*2]

[#]*Computer Engineering Department, Agnel Institute of Technology and Design, Goa University, Goa, India*

Abstract: *Data mining is a most popular and broad domain of computer science. It helps in extracting useful features and patterns from huge amount of available data. It involves clustering as one of its pre-processing steps which deals with grouping the similar items together in unsupervised manner. K-means is a basic and simple clustering algorithm used in various data mining applications. The initial selection of centroids and determining the number of clusters for a given dataset has a great impact on the final result also affecting its time complexity. This paper provides a brief survey on various initial centroid selection methods and finding the most desirable value of K in K-means clustering algorithm.*

Keywords*: Clustering, Unsupervised, Centroids*

I. I.Introduction

Identifying and analyzing the hidden patterns in the data, forms the prime objective of data mining domain of computer science. Patterns that are identified using various data mining concepts can be used in knowledge discovery which can aid the user or a business firm in decision making. In todays world, the amount of data collected is growing exponentially which needs to be analyzed properly to extract useful information out of it [1].

Clustering is a concept used to find similar groups, within the data set.The K-means algorithm is a popular data clustering algorithm. To use this algorithm requires the number of clusters in the data to be pre-specified. Appropriate number of clusters for a given data set is generally determined using a trial-and-error process. There are many clustering approached exists, including hierarchical clustering, minimum spanning tree, and k-means clustering [1]. In this survey paper the k-means clustering algorithm is selected for analyzing. It is one of the simplest and highly popular algorithms among all clustering techniques. The k-means algorithm determines k distinct clusters, when data set is submitted to it. The algorithm calculates the centroid, or midpoints, of each cluster and allocate each point to its nearest centroid. The algorithm is started by providing: 1) the number of desired clusters, k, and 2) initial starting k centroids. There is no standard best way to find above two parameters i.e. the number of clusters initial centroids. The result of the clustering technique depends on the specified choice of initial starting point values [2]. Commonly used method of choosing the initial centroids is to arbitrarily choose k of the actual sample data points. This method, along with 6 other techniques, are proposed and analyzed in this study. The decision for selecting the value of K also greatly impacts the k-means algorithm result. This survey paper also describes various methods for selecting optimal value of K. The remainder of the paper consists of five sections. Section 2 presents the well-known techniques for selecting K, number of clusters. Section 3 reviews the main known methods for selecting initial centroid points. Section 4 describes the proposed evaluation measure.

II. K-Means Algorithm - Overview

A data set of sample points will be given along with a desired number of clusters, k, and a set of k initial starting points. From this, the k-means clustering algorithm determines the desired number of clusters. A centroid is the average ie mean value of each of the coordinates of all the points of that cluster [2]. Generally, the k-means clustering algorithm follows the following steps.

1.Input the value of k i.e. the number of clusters.
2. Arbitrarily select k centroids.
3. Observe each point in the given data set and allocate it to the cluster whose centroid is near by to it.
4. Recompute the new k centroids after every iteration once all points have been assigned to some clusters.
5. Reiterate steps 3 and 4 until no point changes its assigned cluster, or until a maximum number of iterations pre-defined is performed or there is no change in centroids

III. Finding K

This section presents existing methods for selecting K required for the K-means algorithm as one of its input.

A. Variational Gaussian mixture model by Bishop

This method describes how the Bayesian formulation of a Gaussian mixture model can be used to automatically determine the number of components needed to adequately describe the data. A GMM is a soft clustering approach where each data point belongs to multiple clusters. Using GMM, set of weights are estimated which will determine amount by which a data point belongs to a cluster. Using Bayesian formulation, prior distribution is put on the parameters of the model which impose a tradeoff between the model complexity and the model fit. In this paper, he explains how the formulation can be used with "automatic relevance determination" to find out which combination components to keep at each training iteration [2].

B. User specified K

To implement K-means algorithm in many data-mining application, requires the number of clusters to be specified by the user. In order to get a best quality clustering output, usually, a number of iterations are needed where the user executes the algorithm with different values of K. The output of clustering is evaluated only visually without applying any specific performance measures. Using this approach, it is hard for naked eyes of the users to evaluate the clustering output for data sets with multiple dimensions [6].

C. The silhouette method

Using the average silhouette of the data, we can approximately calculate the natural number of clusters in a given dataset. In order to calculate the silhouette of a data point it is required to get the estimate of how closely it is similar to data within its cluster and how loosely it is similar to data of the neighbouring cluster. A silhouette value close to 1 indicates that the datum is in an appropriate cluster, while a silhouette close to −1 indicates that the datum is in the wrong cluster. Genetic algorithms can be used for determining the number of clusters that gives rise to the highest silhouette. With this approach it is possible to re-adjust the data points in such a way that the silhouette is most likely to be highest at the correct number of clusters in the given dataset [3].

D. Elbow Method

By visually inspecting the data set, the number of clusters can be estimated, but we will soon understand that there is a lot of ambiguity in this procedure for almost all dataset excluding the simpler one. Here, having previous experience with that particular problem or something similar will help you choose the right value. If we want some hint about the number of clusters that you should use, we can apply the Elbow method. First of all, compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is calculated as the sum of the squared distance between each point of the cluster and its centroid. If we plot k against the SSE, we will see that the error decreases as k gets larger; this is because when the number of clusters increases, they should be smaller, so distortion is also minimal. The idea of the elbow method is to choose the k at which the SSE decreases abruptly [4].

E. K equal to the number of classes

With this method, it accepts the dataset, reviews it and computes the number of classes. This value is equated to the number of clusters in the data sets. The result of the observations is fed back to the clustering algorithm to improve its performance, thus making it supervised. Obtaining the value of k using this method, the assumption is made that the data clustering method could form clusters, each of which would consist of only points belonging to one class. Unfortunately, most real problems do not satisfy this assumption [6].

F. K calculated using a neighbourhood distance

A neighbourhood measure could be added to the cost function of the K-means algorithm to determine K. Although this technique is promising, it needs to prove its potential in practical applications that could be used in real life. Because the cost function has to be modified, this technique cannot be applied to the original K-means algorithm.

G. Values of K determined by statistical measures

These statistical measures for getting value of K, are often applied in combination with probabilistic clustering approaches. They are calculated with certain assumptions about the underlying distribution of the data. The set of Gaussian distribution are used to compute the Bayesian information criterion on the dataset. The technique applied here is based on the assumption that the given data set fits the Poisson distribution. The methods related to the null hypothesis, such as Monte Carlo, are used for assessing the clustering results and

also for determining the number of clusters [6].

IV. Selecting K Number Of Initial Centroids

Once the value of K is calculated using one of the above mentioned methods, next step is to implement actual K-means algorithm. Standard K- means algorithm selects K number of initial centroids randomly. Which points are selected as initial centroids may significantly affect the quality of final clusters obtained. It may also affect the time requirements of the algorithm. Various methods are proposed for selecting K number of initial centroids which are discussed in below section.

A. Random Selection

In this method of selection of centroids, 'K' numbers of points are selected randomly [1]. Standard K-means follows this strategy for selecting the K number of initial centroids. The quality of final clusters obtained depends on which points are selected by randomly.

The drawback of this method is, in some cases, different initial centroids might give you different final clusters. This can be clearly seen in below figures.

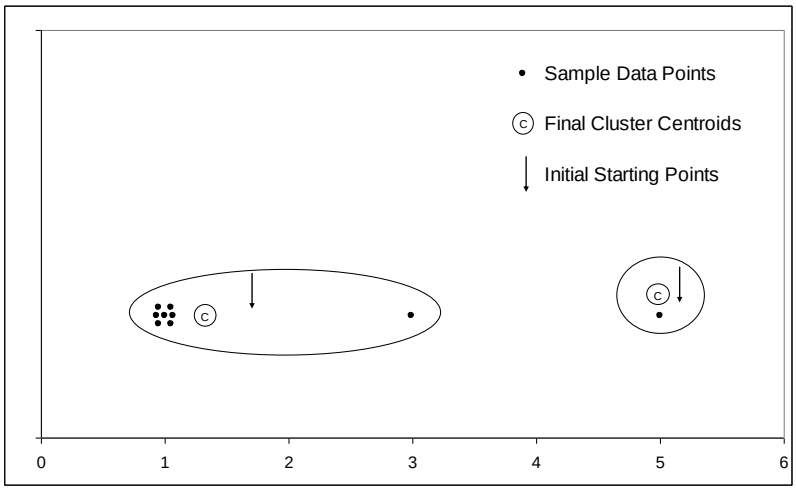

Figure 1: Clustering output

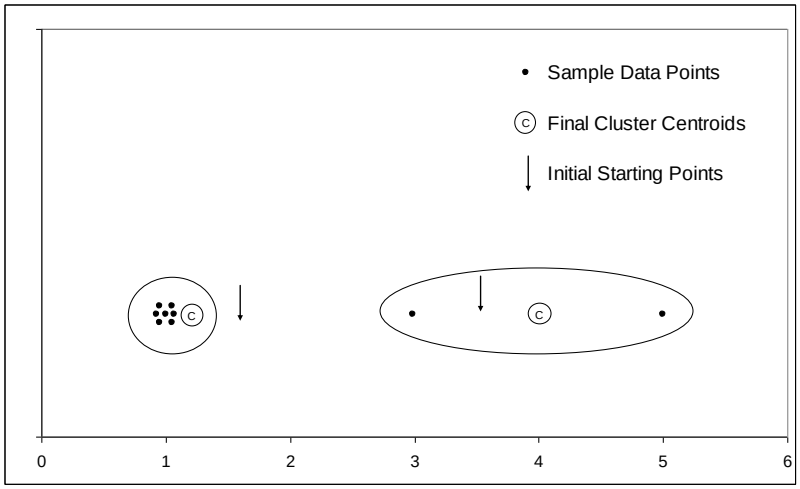

Figure 2: Clustering output

Also, if there are outliers present in the dataset, then there are chances that those outlier points getting selected as initial centroids. If this happens, then the K-means algorithm will take long time to terminate and also, the quality of clusters might not be that good.

B. The centroid selection method proposed in the paper [6] provides a systematic way to calculate initial centroid points which may provide better accuracy in less number of iterations as compared to traditional algorithm. Steps can be summarized as follows.

Step 1: Accept the dataset and the value of K
Step 2a: calculate the distance of first point to all other points in the dataset using:

$$d(Pi) = \sum_{i=1}^{n} (\text{distance}(\text{Pi}, \text{Xi}))$$

Step 2b: Arrange all the points in the dataset according to the above sorted distance using Sort {d(Pi)}
Step 3a: Divide the entire dataset into K equal partitions.
Step 3b: Choose the first point of every partition initial centroid.
The drawback of this method is it can be computationally expensive if used with large data sets since it makes use of sorting. Also, calculating the distance from first point to all other points is also an overhead.
C.[7] proposed a method of centroid selection in which first centroid was the mean of all the points. Subsequent centroids are selected based on following algorithm.
Step1. Calculate the mean of all points in dataset and consider this mean as first centroid.
Step2. Select next centroid in such a way that its distance from selected centroids is maximum.
Step3. Repeat step 2 till K number of initial centroids are obtained.
The method proposed here worked well for small value of K. For larger value of K, this method was found to be inefficient.

D. *Unscrambled midpoints* [8]
Step1. Divide each feature into K partitions each of equal size.
Step2. For each partition of each feature, calculate the midpoint as the initial starting value for each partition.
Step3. Midpoint values computed for partition 1 of each feature forms first centroid. Similarly, midpoint values computed for partition 2 of each feature forms the second centroid. This continues till K initial centroids are obtained.

***E. scrambled midpoints* [8]**
Step1. Each feature values are divided into K equal sized partitions.
Step2. For each partition of each feature, calculate the midpoint as the initial starting value for each partition.
Step3. For each feature, randomly select one of the partition's midpoints as the starting point. Do this k times to construct the needed k starting points.

Unscrambled and scrambled midpoint methods gave a systematic approach for finding the initial centroids by dividing the each feature into K equal sized partitions and then by selecting the initial starting value from each partition.

***F. Feature value sums* [8]**
[ ] proposed a simple method for selecting the initial centroids which involves adding the feature values of data points and then sorting the points based on their feature sums. Steps can be summarized as follows.
Step1. For each point, add all its feature values to give its feature sum.
Step2. Sort entire dataset based on this feature sum values for all dataset points.
Step3. Divide the sorted points into K equal sized partitions.
Step4. From each partition, select the median and take that corresponding point as the initial centroid.
Again, this method involves sorting procedure which might lead to increase in time requirements of the algorithm.

V. Conclusion

The paper presents various methods for determining the number of clusters i.e. value of K and initial centroids for given data set for K-means clustering algorithm. For determining the number of clusters, nine methods are discussed out of which Elbow method looks more efficient. For selecting K number of initial centroids, six methods are discussed from which unscrambled midpoints method might give better results. To validate the correctness of the above mentioned methods, it is recommended to implement and test them on s standard dataset.

References

[1]. Han, J. and Kamber, M. Data Mining: Concepts and Techniques, 2000 (Morgan Kaufmann, San Francisco, California).
[2]. Al-Daoud, M. B., Venkateswarlu, N. B., and Roberts, S. A. Fast K-means clustering algorithms. Report 95.18, School of Computer Studies, University of Leeds, June 1995.
[3]. Alsabti, K., Ranka, S., and Singh, V. An efficient K-means clustering algorithm. In Proceedings of the First Workshop on High-Performance Data Mining, Orlando, Florida, 1998;

[4]. Bradley, S. and Fayyad, U. M. Refining initial points for K-means clustering. In Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98) (Ed. J. Shavlik), Madison, Wisconsin, 1998, pp. 91–99.
[5]. Bradley, S. and Fayyad, U. M. "Refining initial points for K-means clustering.," In Proceedings of the Fifteenth International Conference on Machine Learning, ICML'98,1998.
[6]. Kedar B. Sawant , *"Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance"* International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015, PP 22-27 ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)
[7]. Anand M. Baswade, Prakash S. Nalwade, *"Selection of Initial Centroids for k-Means Algorithm"*, International Journal of Computer Science and Mobile Computing, Vol. 2, Issue. 7, July 2013, pg.161 – 16, ISSN 2320–088X
[8]. Frank Robinson, Amy Apon, Denny Brewer, Larry Dowdy, Doug Hoffman, Baochuan Lu "*Initial Starting Point Analysis for K-Means Clustering: A Case Study"*