ISSN No: - 2456 - 2165

Efficient Seed and K Value Selection in K-Means Clustering Using Relative Weight and New Distance Metric

Premsagar Dandge[1]
Department Computer Engineering,
JayawantraoSawant College of Engineering,
Hadapsar Pune-28, SavitribaiPhule Pune University,
Pune, India.

Prof. A.K. Gupta[2]
Professor, Computer Engineering,
JayawantraoSawant College of Engineering,
Hadapsar Pune-28,
SavitribaiPhule Pune University, Pune, India.

Abstract:- K-mean clustering algorithm is used for clustering the data points which are similar to each other. K-means algorithm is popular due to its simplicity and convergence tendency. The general distance metrics used in this algorithm are Euclidean distance, Manhattan distance etc. which are best suited for numeric data like geometric coordinates. These distance metrics does not give full proof result for categorical data. We will be using a new distance metric for calculating the similarity between the categorical data points. The new distance metric uses dynamic attribute weight and frequency probability to differentiate the data points. This ensures the use of categorical properties of the attributes while clustering. The k-mean algorithm needs the information about number of clusters present in the data set in advance before proceeding for cluster analysis. We will be using a different technique for finding out the number of clusters which is based on the data density distribution. In k-means algorithm, the initial cluster seeds are selected in a random fashion which may lead to more iteration required for convergent solution. In proposed method, seeds are selected by considering the density distribution which ensures the even distribution of initial seeds. This will reduce the overall iterations required for convergent solution.

Keywords*: k-means clustering, categorical data, dynamic attribute weight, frequency probability, data density.*

I. INTRODUCTION

The k-means method is highly used clustering technique that which tries to minimize the average squared distance between points in the same data set within a cluster. This technique is not fool proof but its simplicity made it popular. There is a possibility in which this algorithm generates arbitrarily bad clustering even when number of data points i.e. n and number of clusters i.e. k are fixed.

There are multiple clustering algorithms proposed in which the data set getting divided into partition and hierarchy broadly. Also there are algorithms which are derived from mixed techniques. In the partition-based clustering type of algorithms K-means algorithm is the most famous. K-means algorithm includes K-means, k-modes and K-Prototypes basically, of which, K-means algorithm is used for numerical data. K-means has easy operations and generic clustering. It can beapplied on the cluster analysis of several data types such as text and images. But this algorithm goes into processing of multiple iterations which are not dependent on the quantity or nature of the data. There is a fluctuation in the results due to random initial selection of the seeds. Seeds are the initial data points which are assumed as local centroids of the cluster. Due to the fact that the clustering is often applied to the data of the cluster quality which end-users can't predefine and this fluctuation is difficult to be accepted in the application. Hence there is a huge scope to improve the quality and stability of clustering results.

The K-means algorithm defines a mathematical function to compute Euclidean distance or any other distance metric to find out the difference between two numeric objects. But for categorical data it is not possible to find the difference between data points using the Euclidean or any other numeric measure. An example for the similar situation can be finding Euclidean distance between numeric points A to B is 25 and A to C is 10 which means A is closure to C and not B. Categorical values are not the numbers or the distance. They are the enumerations such as 'grapes', 'pineapple' and 'mangos'. Euclidean distance cannot be used to compute distances between the above values of fruits. We cannot say that pineapple is closer to grapes or mangos since Euclidean distance is not useful to handle such information. Hence it is required to modify the cost function other than the Euclidean distance. The possible options could be the use of distance metrics like hamming distance or other metric which can be used for categorical data. In hamming distance if two categorical values are same then make the distance 0 or else 1. Instead of mean the frequency of the attribute is considered.

In k-means algorithm the seeds are selected at random. This randomness introduces the problems like selecting the seeds from the same cluster or not selecting the seed from a cluster.

ISSN No: - 2456 - 2165

The improper selection of the initial seed tends to increase the number of iterations required for calculating the clusters from the given database. If algorithm selects the seeds in proper manner, ideally one seed per cluster then the cluster becomes faster and more accurate. The number of seeds must be equal to number of clusters available in the data set.Considering more or less number of seeds from the actual clusters generates wrong cluster calculations. Hence selecting proper cluster seeds is important

The algorithms which are used to determine the number of clusters are iterative in nature i.e. they use trial and error methodology for calculation. The iteration which gives more optimal output for the cost function (mostly squared sum in case of numerical data) is taken as a standard value for number of clusters. For example if there are 100 data points in a data set and 5 likely clusters then these algorithm keeps calculating the cost with cluster value starting from 1,2,3, and so on. It generally gives optimum cost value for cluster value near to 5. In each iteration 100 points are used for calculations. This method is monotonous and gives degraded results. Hence an efficient method with less time complexity is required to determine the number of clusters.

In case of categorical value, each possible value has some property or context associated. If we use normal cost function for clustering like Euclidean distance then it discards the contextual value of that data point. For example A+,A-,B+,B-,O+,O- etc. are just signed characters for normal reader but these are blood groups for a medical personal. Out of these blood groups few are universal donor or universal acceptor. If general distance metrics are used then the background information may get discarded. Hence it is important to use a separate distance metric for categorical data.

The k-means method is highly used clustering technique that which tries to minimize the average squared distance between points in the same data set within a cluster. This technique is not fool proof but its simplicity made it popular. There is a possibility in which the algorithm generates arbitrarily bad clustering even when number of data points i.e. n and number of clusters i.e. k are fixed.

II. RELATED WORK

Hong Jia, Yiu-ming Cheung and andJiming Liu 2016, has published a paper for a new distance metric which is suitable for categorical data. The distance metrics like Euclidean distance are not suitable for categorical data. They can be used with numerical data. Use of these metrics for categorical data does not refer the properties of the categorical data. This paper introduces frequency probability and dynamic weight as a new distance metric for categorical data. The frequency probability of the data is calculated on the basis of similarity between objects. In most of the cases more weight is given to few attributes which are more important for evaluation. If the attribute values are different then it contributes for the overall distance between two objects else the attribute is ignored while calculating the distance. Also the relationship between the data objects is discarded. The matching and mismatching is highlighted in this method. The weight of attribute is directly proportional to the similarity of objects. The objects with dissimilar values has high values and vice verse. [1]

Greg Hamerly and Charles Elkan has published a paper which describes a method for finding the input required for k-means clustering i.e. k value which is number of clusterspresent in the data set. When the value is smaller than the actual number of clusters then the data gets divided into wrong clusters. Same situation happens when the value is bigger than the actual value. It is difficult to find k value programmatically. This paper has given a technique for determining the k-value. It uses Gaussian distribution. For every k value the squared sum is calculated and the optimum one is selected. [2]

David Arthur and Sergei Vassilvitskiihas published a paper to provide details about importance of proper seed selection. The method given in the paper gives importance result for kmeans clustering algorithm. This technique tries to separate out the seeds across the data set. The distance from center is calculated which is already selected. The next center is selected at random. This is an probabilistic seed selection method. It also provides the details around effect of wrong seed selection on clustering. Also the performance of clustering is improved by this method. [3]

Li Xinwu2010, has published a paper for clustering algorithm. It is an optimized version of k-means algorithm. K-means selects the initial centers and proceeds with the calculations. The change in seeds results into altogether new results. The problem of random seed selection can be overcome by proper seed selection. This paper talks about taking the samples of data before the actual seed selection. The sample selection speeds up the procedure as the number of data points decreases. This method also takes care of seeds getting selected from same cluster. The same process is continues to select all the required seeds. [4]

NoureddineBouhmala2016, has published a paper to describe the behavior of Euclidian distance metric. It also contains the details around the impact of Euclidian distance on clustering. Clustering is important for finding out the similar type of objects. The distance between the data points is the measure for the homogeneity of the data which belongs to the same cluster. This metric is used in most of the cases where numeric data comes into the picture. It does not retain the quality of values within same type. [5].

Jianpeng Qi, Yanwei Yu*, LihongWang,andJinglei Liu 2016 has published a paper around k-means algorithm. The behavior of kk-means algorithm depends on the nature of initial seeds which are selected for clustering. This paper proposes a optimized version of the existing algorithm. There is always a probability of two centers getting selected from the

same cluster. To reduce this effect extra seeds are selected. The extra seeds are then slowly reduced to actual cluster numbers by merging the recalculated cluster data. [6]

III. SYSTEM OVERVIEW

This paper introduces a technique for finding the number of clusters in a given data set which is different than the other traditional methods available for the same purpose. The proposed method mainly focuses on the clustering of categorical data types. The data from available source is imported to the system and stored in the suitable format. It also provides an option for selection of important attributes and rejecting the rest. The data is sampled and plotted across the boundary limits present in thedata set. The whole plane is divided into small blocks which are filled with the data points from the data set. Density of each block is calculated by considering number of elements which falls in the particular block. The data block frequencies are arranged in descending order. The nonzero values are then analyzed using elbow method to get the point which gives more output with minimum number of clusters. The turning point of curve determines the number of clusters present in the dataset.

The data blocks with higher density are considered as the potential cluster heads. These cluster heads can be detected by traversing the block densities. A data point can be selected as a seed from each data block. This may give slightly different seed than the actual seed but still much more efficient than the random seed selection. The seeds are selected from each block till the point it meets number of clusters calculated earlier. This maintains the number of iterations constant which is better than the random method. While performing the data clustering, time required for clustering, sum of all distances within the cluster and number of iterations are noted down to compare it with random method. The results of comparison between frequency based and random method are explained in performance and result section in detail.

IV. ARCHITECTURAL VIEW

The architecture diagram of the system shown below helps us to know the system.

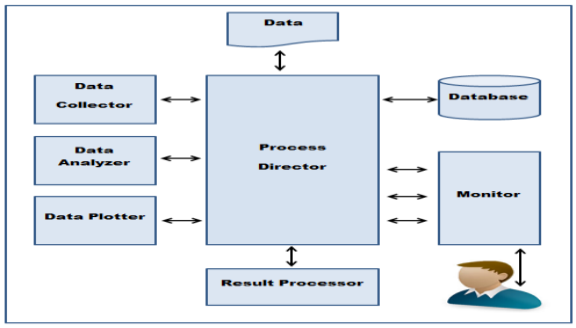


Figure 1:- System Architecture

The components from the system with details are as follow

1. Data
It is the categorical input which can be in any form. The source of the data is known to the user.

2. Monitor
It is the interface provided to the user for systems interactions. From this interface user can perform multiple operations. Also the status and output of the result is displayed on the monitor. User has to provide input data source form this interface.

3. Database
This component is the external entity which is connected to the system over network. It stores the input data as well as the processing data in it. This entity is optional depending on amount of data taken into consideration for clustering the categorical data.

4. Process Director
This component acts as a coordinator between all the system components. It keeps the track of overall process. It also provides the status details to monitor so that user can watch it on display.

5. Data Collector
It takes the details for data input form the monitor and connects to the source. It imports the data and stores it into the database. It takes the input in multiple format and stores processing friendly format.

6. Data Analyzer
It uses the input data and calculates the probability density. It also calculates dynamic attribute weight. Using both the fields distance between data objects is calculated.

7. Data Plotter
It divides the data plane into unit blocks. It takes the input from data analyzer on that basis distributes the data to the blocks. It also calculates the data density of each block. The relative weight for each block is also calculated.

8. Result Processor
It traverses the output of data plotter around the data density and determines required parameters. The required parameters are number of clusters in the given dataset and seed selection.

The proposed system can be compared with existing system in following areas

a) Seed selection: Normal K-means algorithm selects initial seeds at random and then performs further calculations. The

ISSN No: - 2456 - 2165

seeds selected may belong same cluster hence takes more time for convergence. Also there is a possibility that an outlier gets selected as a seed which againgives unwanted results. The proposed system will select the seeds more precisely and overcomes the problems introduced due to random seed selection

b)Categorical Data: In existing system, categorical data is first converted into numerical data and then clustering is performed over it. This does not consider the meaning associated with categorical data. The proposed system will preserve the implicit meaning of categorical data

c) K-value calculations: In existing system the k value is determined by calculating squared sum of the elements again and again for all possible cluster values. This approach is less efficient. The proposed solution will determine the k values using block density of the data set in a more efficient way.

d)Iterations required: In existing system, due to random seed selection approach more number of iterations are required for convergent solution. The proposed solution will precisely select the initial seed considering data density and relative weight which will reduce the number of iterations required for convergence

e)Distance Metric: Existing system uses Euclidean distance for measuring the distance between elements. Euclidean distance is good for numerical data but gives bad results for categorical data. The proposed system will use another distance metric which is based on probability frequency and dynamic attribute weight. It gives better results for categorical data.

V. RESULT AND PERFORMANCE

Medium load calculations are performed using the proposed approach to compare it with traditional approach. There are mainly two approaches while comparing the results. One approach is to find out how good the k value selection is calculates. Traditional method to calculate the k value is to calculate sum of all distances between elements and its respective clusters centroids and then finding the point which is most efficient while balancing out the cluster number and sum. This approach involves many more calculation. Whereas in new approach, number of clusters are calculated using the block density term used in the earlier sections. This method reduces the number of calculations required. A data series block and its density are analyzed and a point with maximum frequency and minimum clusters is calculated. The same number can be used as cluster number for k means clustering.

The second approach for comparison is seed selection. In traditional method the seed are selected in random fashion. The random selection may lead to selecting outlier as a centroid, multiple seed selection from the same cluster etc. For categorical data it is not possible to differentiate to points having same value. In that case a new centroid value needs to be selected. The new frequency based approach is efficient when it comes to seed selection which is different from each other and evenly distributed.

A) Sum of distance from centroid

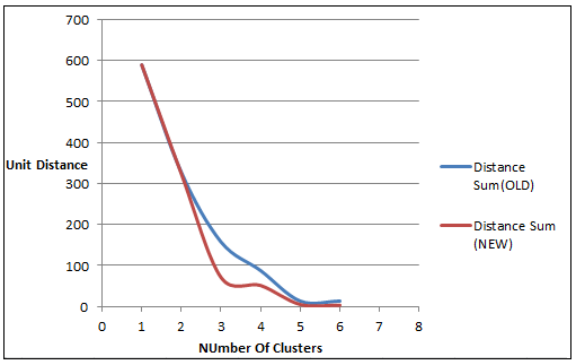
Following table provides more details around how the proposed and old method behaves in real time. The dataset used for the analysis is Credit Card application data.The distance calculated here is using frequency based distance metric

#Clusters	Distance Sum(OLD)	Distance Sum (NEW)
1	589.524	589.524
2	326.73248	323.1865
3	157.47209	70.84742
4	87.48878	51.239746
5	13.071791	4.901921
6	13.071791	3.02679477

Table 1: Distance Sum Comparison

The graph for comparison between distance sums calculated using random seed selection with frequency based seed selection.

It is clear from the below graph that the distance calculated using new seed selection is more efficient since it has a straight line till the point which has balanced cluster and sum values. The sum calculated using random seed selection is more than the new method and it takes time to optimize the cluster and sum values.



Graph 1: Distance Sum Comparison

.

B) Time taken for clustering

Time taken by clustering depends on the implementation and hardware on which the program is running. While comparing the random seed selection method with frequency based seed selection, both the implementations were run on same system to keep the results consistent. The hardware used was 2.53GHz CPU and 4GB RAM. In random

ISSN No: - 2456 - 2165

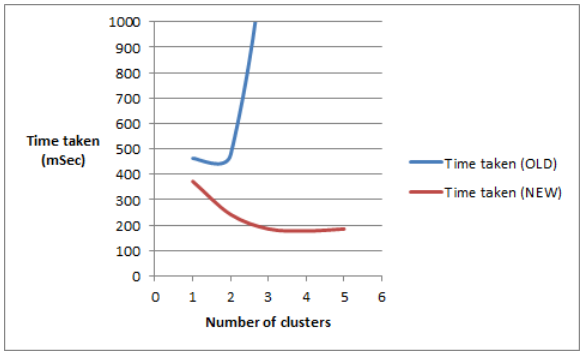
seed selection if a wrong seed is picked up then an extra iteration is required to take the solution towards convergence.

Following table gives more details about the comparison between times taken. Time taken for clustering is more in traditional method of clustering in which the seed are selected randomly. Frequency based seed selection takes less time for clustering since it does not involve the multiple iterations required for convergence.

#Clusters	Time taken (OLD)	Time taken (NEW)
1	463	372
2	477	242
3	1638	185
4	21644	177
5	813635	185

Table 2: Time taken comparison

Following is the graph for above data presented in above table



Graph 2: Time taken comparison

In above graph it is clear that frequency based seed selection takes less time than time taken by random seed selection.

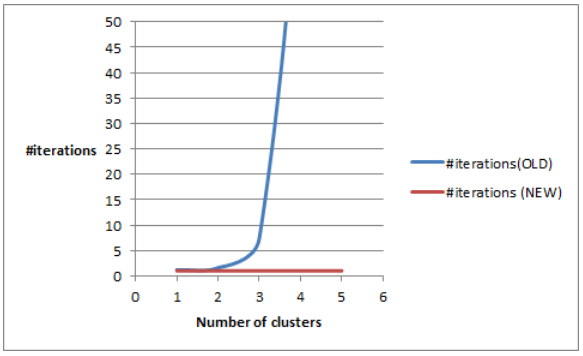
B) Iterations required for convergence
Following table provides more details around how the proposed and old method behaves when it comes to number of iterations required for convergence.

#Clusters	#iterations(OLD)	#iterations (NEW)
1	1.2	1
2	1.6	1
3	7.4	1
4	105.5	1
5	3834	1

Table 3: Iterations required comparison

The number of iterations required in case of random seed selection increases considerably with the number of clusters present in the dataset. Random selection gives comparatively better result when the data is evenly distributes but not in the case when these is a frequency biased distribution.

Below is the graph for comparison between iterations required while using random seed selection with frequency based seed selection.



Graph 3: Iterations required comparison

From the graph it is clear that new method of seed selection takes a contact number of iterations for convergence i.e. 1iteration.
From above comparisons between random method of seed selection and new frequency based seed selection, it can be concluded that that frequency based seed selection is an efficient method for cluster the categorical data.

VI. CONCLUSION

This paper has proposed a method to calculate k value required in k means clustering algorithm using block frequency method over traditional trial and error method. This reduces the overhead of running k means algorithm multiple times just to find the optimum number of clusters present in the dataset. This paper has also provided a method to select the seeds efficiently using frequency of elements over traditional method of random seed selection. The new seed selection selects the points smartly to avoid clash with points already selected as centroids. Also the new seed selection method is implemented and the results are explained in statistically as well as graphically. The analysis of the comparative results confirms that frequency based seed selection is better than random seed selection.

References

[1] Yiu-ming Cheung, Hong Jia and Jiming Liu 2016, A New Distance Metric for Unsupervised Learning of Categorical Data in IEEE Transactions On Neural Networks And Learning Systems, Vol. 27, NO. 5 on MAY 2016.

[2] Charles Elkan and Greg Hamerly, Learning the k in k-means

[3] Sergei Vassilvitskii and David Arthur, k-means++: The Advantages of Careful Seeding.

[4] Li Xinwu Research 2010, Text Clustering Algorithm Based on Improved Kmeans in International Conference On Computer Design And Applications (ICCDA 2010).

[5] NoureddineBouhmala, How Good Is The Euclidean Distance Metric For The Clustering Problem in 2016 5th IIAI International Congress on Advanced Applied Informatics in 2016.

[6] Jianpeng Qi, Yanwei Yu*, LihongWang,andJinglei Liu 2016 K*-Means: An Effective and Efficient K-means Clustering Algorithm in IEEE International Conferences on Big Data and Cloud Computing (BDCloud) in 2016.

Premsagar S Dandge is currently pursuing M.E (Computer) from Department of Computer Engineering, JSPM's JayawantraoSawant College of Engineering, Pune, India. Savitribai Phule Pune University, Pune, Maharashtra, India - 411007.

Prof.A.K.Guptais currently working as Professor with Department of Information Technology, JSPM's JayawantraoSawant College of Engineering, Pune, MH, India.