International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

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

Premsagar Dandge[1], Aruna Gupta[2]

[1]Department Computer Engineering, Jayawantrao Sawant College of Engineering, Hadapsar Pune-28, Savitribai Phule Pune University, Pune, India

[2]Professor, Computer Engineering, Jayawantrao Sawant College of Engineering, Hadapsar Pune-28, Savitribai Phule 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 this algorithm are Euclidian distance, Manhattan distance etc. which are best suited for numeric data like geometric coordinates. These distance metrics does not given full proof results 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 considered while clustering. The k-mean algorithm needs the information about number of clusters present in the dataset in advance before proceedingfor cluster analysis. We will be using a different technique for finding out the number of clusters which is based on the data density distribution.Also 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 considering the density distribution which ensures the even distribution of initial seed selection. This will reduce the overall iteration required for convergent solution.***

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

1. Introduction

The most popular clustering technique is k-means algorithm. It groups down the data points which are near to each other. It is a simple and less complex technique. Even if total number of data points in dataset and number of clusters are kept same, sometimes it yields bad result for multiple runs. The various algorithms available are based on data partition or data hierarchy. Few other algorithms are combination of both partition and hierarchy. Data partition clustering algorithms are easy to understand hence more popular. In k-means algorithm the initial seeds are selected at random. If the selected seeds are less than the actual number of cluster than the wrong results are populated. The other way possibility of selecting more number of seeds also populates wrong results. Hence the number of seeds and number of clusters present should match for proper results. The proposed technique ensures the proper selection of seeds for clustering. It selects the seeds considering the overall data density distribution across the dataset. It takes care for seed selection within same cluster. It also prevents the seed selection amongst outliers.

K-means algorithm requires number of cluster value to proceed for the clustering. The algorithms available in market for calculating number of cluster uses k-means algorithm internally. K-means algorithm is repeatedly invoked for all possible values. The value which gives more homogeneity is selected and used for clustering. As multiple tries are required to get the intended value, it is less efficient method for calculating the number of clusters. This introduces the need for an efficient technique to determine number of clusters present in the dataset. The proposed technique, the number of clusters present in the dataset is calculated using data density distribution.

Categorical data has specific enumeration values e.g. cricket match types. In k-means algorithm with distance measures like Euclidian distance are used for calculating the distances between the data objects. The possible values of cricket match types are T20, ODI & Test. Each match type has its own characteristics and properties. It is difficult to find the similarity between cricket matches on the basis of Euclidean distance. Hence a need for effective distance metric for clustering is required. The proposed technique uses probability density and dynamic attribute weight for comparison of data.

2. Related Work

Hong Jia, Yui-ming Cheung has proposed a new distance metric which can used for clustering categorical data since the distance metrics like Euclidian distance cannot be used for clustering of categorical data. The distance metrics like Euclidian distance, Manhattan distance are used when the data point are numeric in nature. One option available is to convert the categorical data into numeric data and then use the distance metrics available for numeric data. The reason for not using this method is that the information associated with the categorical value is ignored. Also the relationship between the data is not revealed. This makes it important to have a separate distance metric for categorical data. This paper has introduced new distance metric for categorical data viz. frequency probability based distance metric and dynamic attribute weight based distance metric. In frequency probability the distance between two elements

from a database is calculated on the basis of similarity of objects considering each attribute value. The difference between the objects is the sum of probabilities of the attribute to have similar value provided they already have different values. If an attribute has similar value for both the elements then its contribution to the overall difference is zero.

In real life scenario while comparing two objects few attributes has given special grade i.e. it has more weightage that other attributes. If two data objects have different values then its contribution to the whole distance is the probability that these two objects have similar value for a given dataset. In other words for two objects having different values for an attribute then its contribution to the overall distance is inverse to the probability that these two objects have different values in a given dataset. This metric is based on the attribute value pair and not the general information of the attribute hence it has better adjusting ability. This method highlights the matching and not matching between the data objects. It also avoids the denomination of values with high frequency. The weightage of every attribute is based on the probability. If two attributes has same value then the weightage the attribute is probability of having different values. In other case, if the two attributes has different values then the weightage is probability of having similar values. The calculated weight of each attribute is used along with the distance calculated in previous method. This method assigns larger weights to the attributes with mismatching values as they provide more information. [1]

Greg Hamerly& Charles Elkan has published a paper for finding k value which is the input required for k-means clustering algorithm. A wrong k value selection leads to the improper data clustering. When k value taken into consideration is less than the number of clusters the centroid is placed somewhere else than the expected center. Same is the case when k value is more than the number of actual cluster. In this case same cluster consists of two or more centers. When clustering a dataset, the right number k of clusters to use is often not obvious and choosing k automatically is hard algorithmic problem. This paper has proposed an improved algorithm for learning k value while clustering. This paper has explained the G-mean algorithm for finding the k-value. This algorithm starts with a small k value which is kept on increasing after each iteration. In each iteration the center is splinted into two clusters for the clusters which appears not getting properly distributed as per the Gaussian distribution. The advantage of this method is that k-means algorithm implicitly assumes that the data points in each cluster are spherically distributed around the center. The Gaussian distribution is valid for covariance matrix assumption. The problem with this method is that it calculates the squared sum multiple times to calculate the number of cluster. Hence more number of calculations will be required which are proportional to the number of iteration. Also this algorithm tends to estimate more number of clusters than the actual clusters present. This algorithm gives wrong cluster value when the data is Gaussian distributed. [2]

David Arthur and Sergei Vassilvitskii have published a paper which describes the importance of careful seed selection in k-means algorithm. This paper describes the technique for seed selection which gives improved results for k means algorithm on execution. The approach of this algorithm is to spread out the initial cluster centers from each other over the common technique of random seed selection. It first selects the centers uniformly at random from the data set. For each point in the dataset its distance from already selected centers is calculated. A new data point is selected at random as a new center using the weighted probability distribution where a point x is chosen with probability proportional to squared distance. The point that is far away from all selected centers has high probability to be chosen as center. This method introduces a probabilistic seed selection. This papers also talks about wrong seeds selection and its impact on clustering. [3]

Li Xinwu has published a paper on text clustering based on improved version of k-means clustering algorithm. The original algorithm selects the initial seeds and then the iteration procedure is applied on it. Different set of initial seeds gives different results all together. The problem occurred because of random seed selection can be retrieved by selecting the initial seeds properly. In the search process the data should be undistorted and should be able to reflect the original data distribution through the random data selection. The sampled data and original data are clustered using k –means algorithm and very little change in the final cluster is found. Hence, the sampling method is more suitable for the selection of the initial cluster centers i.e. seeds. In order to reduce the sampling effects on the selection of the initial seeds, the sample set extracted each time should be able to be captured into the memory, and best is done to make the sum of the sample sets extracted number of times equivalent to the original data set. Each extracted data which is sampled is then clustered using the algorithm and one group of cluster center is generated. The number of samples produces those many clusters and then the comparison of clustering cost function values is conducted for those many cluster centers, and one group of minimum cluster center is given as the optimal initial cluster center. In this paper the technique uses a different method for seed selection. As this algorithm is specific to the text based clustering, while selecting the seed it is compared with the previously selected seed. If the seed selected in current iteration and seed selected in previous iteration are similar then the seed selected in current iteration is rejected. Seed selection iteration is then continued till the time where different type of seed is selected.

K-means is a commonly used for clustering algorithm in field of data mining across different disciplines in the past fifty years. However, k-means heavily depends on the position of initial centers, and the chosen starting centers randomly may lead to unwanted quality of clustering. Motivated by this, this paper proposes an optimized k-means clustering method along with optimization principles. Since k-means chooses initial centers randomly, it is difficult to avoid choose outliers or points that are too close to each other. In a situations of mistaken merging or dividing always occur once two seeds are chosen in a cluster and the cluster is far from other points. The risk of random initialization increases significantly especially for unbalanced. Normal k-means algorithm chooses k points as a cluster randomly. In

this method there is probability of two or more points falling into the same cluster. Hence to reduce the probability of points falling into the same cluster, the algorithm given in this method selects more number of centers k* over actual k. the number of initial seeds are more hence there is a less chance that multiple center belongs to same cluster corresponding to actual k value. Selecting more centers also introduces some distortion into the calculations. The distortion added in previous step need to be removed by some post processing. Merging top n clusters: This method is used to calculate the actual k clusters from the K* clusters calculated in previous steps. Distance between each cluster is calculated and two clusters with minimum distance are merged into one cluster. This merging process is repeated till the time K* becomes actual k or skewed data distribution. [4]

NoureddineBouhmala has published a paper to analyze the fitness of Euclidian distance metric for clustering. Cluster analysis is an important concept of data mining for discovering the similar groups. The distance between the data points is calculated and the solution is calculated which will try to improve the homogeneity inside the cluster and heterogeneity between the clusters. It is widely used distance metric for clustering. The problem with this distance metric is that it does not capture the quality of the clustering making it less suitable for increasing homogeneity inside the cluster and heterogeneity between the clusters. [5]

3. Proposed Work

This paper introduces a method to calculate number of clusters in given dataset. It targets the clustering of categorical data. The data from available sources is imported to the system and stores in database. It also provides the import of selective attributes from the source. The area between the data boundaries is divided into unit blocks. The data from data set is allocated to the blocks as per the values associated with it. The similarity between the objects is calculated using frequency probability and dynamic attribute weight. After dividing all the points, number of points per block is calculated say absolute value. Each block will transmit some of its weight to neighboring block say relative weight. The block which is surrounded with more number of points will have more weight. This relative weight will indicate the density of the block. As outliers lies far away from the main clusters, they will not contribute to the cluster density hence will have min impact. The high and low values of block density will be used to determine the number of clusters. A seed is selected from each cluster with high density. This ensures the proper seed selection. Also it prevents seed from getting selected out of outliers. This method is much more efficient than the random seed selection.

4. Architectural View

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

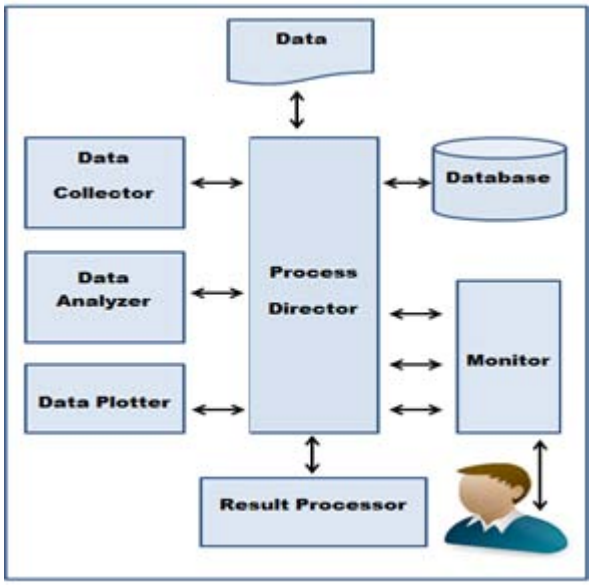

Figure 1: System Architecture

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 system 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.
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.

Sr No.	Paper	Technique / Algorithm	Advantages	Disadvantage
1	A New Distance Metric for unsupervised learning of categorical data	Probability density Dynamic attribute weight	The new distance metric is good for categorical data.	For numeric data Euclidian distance performs better than new distance metric.
2	Learning the k in k-means.	Iterative cost calculation for possible values of number of clusters.	Simple and easy method for finding number of clusters.	Iterative nature takes more time for completion. More unsuccessful attempts are involved. This algorithm tends to provide more number of clusters. Not good for Gaussian distributed data.
3	K-means++ The Advantages of Careful Seeding	Kmean ++ algorithm to find initial seeds.	Better than original k-means algorithm. Importance of proper seeding.	It works on probabilistic seed selection method. First seed is randomly selected hence chances of selecting an outlier.
4	Research on Text Clustering Algorithm Based on Improved K-means	Probability based initial seed selection.	The technique is better than random seed selection.	This method is more suitable for text based clustering.
5	K*-Means: An Effective and Efficient K-means Clustering Algorithm	Probabilistic k-means algorithm.	It choses more number of clusters and selects the best suitable clusters.	More complex operations are required for merging the clusters.
6	How Good Is The Euclidean Distance Metric For The Clustering Problem	Comparing the data objects based on the distance between them	Best suitable and widely used distance metric for numeric data.	Not suitable for categorical data where each categorical value has an implicit property associated with it.

5. Conclusion

This survey paper covers different algorithms and techniques which are available for use. It lists down the areas where an improvement is required from categorical data analysis perspective. The method proposed in the paper tries to overcome the problems by combining available techniques along with new techniques. It improves the quality keeping the properties of categorical data. The complexity of proposed technique is better than the available methods. The total iteration required in k-means will get reduced.

References

[1] Yiu-ming Cheung, Hong Jia and Jiming Liu 2016, A New Distance Metricfor Unsupervised Learning of Categorical Data in IEEE Transactions OnNeural 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 ofCareful Seeding.
[4] Li Xinwu Research2010, Text Clustering Algorithm Based on Improved Kmeansin International Conference On Computer Design And Applications(ICCDA 2010).
[5] NoureddineBouhmala, How Good Is The Euclidean Distance MetricFor The Clustering Problem in 2016 5th IIAI International Congress onAdvanced Applied Informatics in 2016.
[6] Jianpeng Qi, Yanwei Yu*, LihongWang,andJinglei Liu 2016 K*-Means:An Effective and Efficient K-means Clustering Algorithm in IEEEInternational Conferences on Big Data and Cloud Computing (BDCloud)in 2016

Author Profile

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