ISSN 2320-9984

Volume 5 Issue 2 | June 2017

ISSN: 2320-9984 (Online)

International Journal of Modern Engineering & Management Research

Website: www.ijmemr.org

Improved Centriod Selection Process for K-Mean Clustering Algorithm in Data Mining

Ruchita Tiwari
M. Tech. Research Scholar
Shri Ram Group of Institutions
Jabalpur (M.P.), [INDIA]
Email: ruchitatiwari21@gmail.com

Anupam Choudhary
Lecturer,
Kalaniketan Polytechnic College
Jabalpur (M.P.), [INDIA]
Email:chowdharyanupam7@yahoo.com

Sapna Choudhary
Assistant Professor,
Department of Computer Science and Engineering.
Shree Ram Group of Institutions
Jabalpur (M.P.), [INDIA]
Email: choudharysapnajain@gmail.com

Abstract—*In statistic and data mining, k-means clustering is well known for its efficiency in clustering large data sets. The aim is to group data points into clusters such that similar items are lumped together in the same cluster. In general, given a set of objects together with their attributes, the goal is to divide the objects into k clusters such that objects lying in one cluster should be as close as possible to each other's (homogeneity) and objects lying in different clusters are further apart from each other.*

However, there exist some flaws in classical K -means clustering algorithm. According to the method, first, the algorithm is sensitive to selecting initial Centroid and can be easily trapped at a local minimum regarding to the measurement (the sum of squared errors) used in the model. And on the other hand, the K-means problem in terms of finding a global minimal sum of the squared errors is NP-hard even when the number of the cluster is equal 2 or the number of attribute for data point is 2, so finding the optimal clustering is believed to be computationally intractable.

In this dissertation, to solving the k-means clustering problem, we provide designing a Centroid selection in kmean, *which in this algorithm we consider the issue of how to derive an optimization model to the minimum sum of squared errors for a given data objects. We introduce the variant type of k-means algorithm to guarantee the result of clustering is more accurate than clustering by basic k-means algorithms. We believe this is one type of k-means clustering algorithm that combines theoretical guarantees with positive experimental results.*

Keywords:—*Kmean, Centroid, cluster, data objects, Optimization.*

1.INTRODUCTION

The history of extraction of patterns from data is centuries old. The earlier method which has been used is Bayes' theorem (1700s) and regression analysis (1800s). [1] In the field of computer technology, using the ever growing power of computers, we develop an essential tool for working with data. Such as, it is being able to work with increasing size of the datasets and complexity. And also an urgent need to further refine the automatic data processing, which has been aided by other discoveries in computer science, means that our ability for data collection storage and manipulation of data has been increased. Among these discoveries of importance, according to Wikipedia, are the neural networks, cluster analysis, genetic algorithm (1950s), decision trees (1960s) and support vector machines (1990s).

Historically the field of finding useful patterns in data has a myriad of names including but not limited to; Data mining, Knowledge Extraction, Information discovery, data archaeology and data pattern processing. Statisticians use the term of Data mining and also is very popular in the field of databases. The terms of knowledge discovery in databases was introduce at the first KDD workshop in 1989 (Piatetsky-Shapiro 1991) which has been used in the AI and machine- learning fields. [5]

As definition, Data mining or important part of Knowledge Discovery in Database (KDD), used to discover the most important information throughout the data, is a powerful new technology. Across a myriad variety of fields, data are being collected and of course, there is an urgent need to computational technology which is able to handle the challenges posed by these new types of data sets.

The field of Data mining grows up in order to extract useful information from the rapidly growing volumes of data. It scours information within the data that queries and reports can't effectively reveal.

As we mentioned earlier, the integral part of knowledge discovery in database (KDD) is data mining, which in our view, KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. The KDD role is to convert raw data into suitable information as shown in figure 1.1:This process contains a series of transformation steps, from data pre-processing to data mining results.

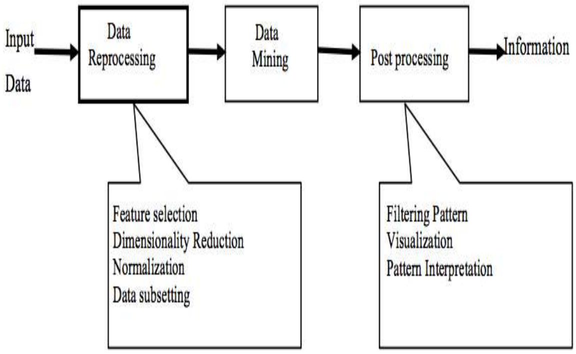


Figure 1 : The overall Steps of process of Knowledge Discovery in Database (KDD)

The type of stored data as **Input Data** involves flat files, spread sheets or relational tables, and may be residing in a centralized Data Repository 1 or distributed across multiple sites.

In order to transform raw data into the appropriate format, **Pre-processing** Phase has been done to subsequent analysis. This step includes fusing data from multiple sources, removing nose and duplicate observations to cleaning purpose, select relevant features and record to data mining task. This step is the most time-consuming and laborious because of the many different types of data.

The phase **Post-processing** ensures that only valid and useful results are integrated and incorporated into decision support system. The example of this step is visualization, that allows the analysts to explore the data and the data mining results from a variety of viewpoints. Statistical and hypothesis methods can also be applied during this step to eliminate incorrect data mining results.

If the learned patterns do not meet the desired standards, then it is necessary to re-evaluate and change the pre-processing and data mining. If the learned patterns do meet the desired standards then the final step is to interpret the learned patterns and turn them into knowledge.

For more, as result validation in post-processing, in the final step of knowledge discovery from database data mining algorithm verify the patterns produced in the wide data set. Of course all patterns found by the data mining algorithms are not necessarily valid. The main purpose for the algorithms is to find patterns in the training set which are not present in the general data set (the definition of over fitting). To do this purpose use attest set of data for evaluation on which algorithm was not trained. Then compare desire output with the result of the learned patterns are applied to this test set. As an example, in the field of distinguishing spam from legitimate e- mails would be trained on a set of sample e-mails that first trained, the learned patterns would be applied to the test set which had not trained, then the accuracy measure from the number of emails they correctly classify.

2. REALTED WORK

Cluster Analysis technique as a field grew very quickly with the goal of grouping data objects, based on information found in data and describing the relationships inside the data. The purpose is to separate the objects into groups, with the objects related (similar) together and unrelated with another group of objects. It is being applied in variety of science disciplines and has been studied in myriad of expert research communities such as machine learning, statistic, optimization and computational geometry. [8]

The following are some examples:

Biology. Biologists when they a long time ago created a taxonomy (hierarchical classification) made a form of clustering according to genus, family, species and so on But also recently they have applied clustering to analyze the myriad amount of genetic information, such as a group of genes that has similar functions.

Information Retrieval. The World Wide Web consists of billions of web pages that are accessed with the help search engine queries. Clustering can be used to create small clusters of search results.

Psychology and Medicine. Clustering techniques are used to analyse frequent conditions of an illness and identifying different subcategories. For example, clustering is used to identify different types of depression, and cluster analysis is used to detect patterns in the distribution/spread of a disease.

Business. In this field there exists a large amount of information on current and potential customers. Clustering helps to group customer activities, as previously mentioned in detail.

In a lot of research and in many applications, the cluster is not well defined. The figure 1.4 is for understanding this concept:

Assume we have twelve points and three different ways to dividing them into clusters. The figure represent that the definition of clustering is imprecise, grouping data depends on desired result and nature of data.

It is important to notice difference between discriminate analysis (supervised classification) and clustering (unsupervised classification). In clustering task, given a collection of unlabeled data, it should be grouped into more meaningful clusters. And also a label is assigned to each cluster. In contrast, in case of supervised classification, a collection of already labeled entities (called training set) are given.

When predefined labels are available for the data sets, new unlabeled patterns classify into one of the predefined groups associated with the labels. Typically, by using a training set it tries to find a classification scheme which can be used to label or predict new objects into the group. For this reason, cluster analysis refers to unsupervised classification. However the term of classification without any qualification within data mining refer to supervised classification. [9] And also, the articles *Segmentation* and *Partitioning* are used as an synonym of clustering. The terms partitioning sometimes refer to graph divided to sub graph and segmentation are used for dividing data into group using simple technique, such as grouping people based on their income.

Historically as Wikipedia references, the terms of K-means clustering algorithm was first developed by J. MacQueen (1967) and then the idea was followed by J. A. Hartigan and M.A.Wong around 1975. The standard algorithm as a technique for pulse-code modulation was proposed the by Stuart Lloyd in 1957, though it wasn't published until 1982.

The consideration of K-means was demonstrated as early as 1956 by Steinhaus [15]. A simple local heuristic for problem was proposed in 1957 by Lloyds [16]. The method represents that first step choosing k arbitrarily point as facilities. In each stage, assign each point X into cluster with closest facility and then computes the center of mass for each cluster. These centers of mass become the new facilities for the next phase, and the process repeats until the solution stabilizes.

In aspect of how the neighbours are computed for each centre there exists some different procedures for performing K-means:

- Lloyd's: Repeatedly applies Lloyd's algorithm with randomly sampled starting points.
- Swap: A local search heuristic, which works by performing swaps between existing centers and a set of candidate centers

This algorithm iteratively changes centers by performing swaps. Each run consists of a given number (max swaps) executions of the swap heuristic.

3. PROPOSED WORK AND RESULT

The K-means algorithm finds the predefined number of clusters. In the practical scenario, it is very much essential to find the number of clusters for unknown dataset on the runtime. The fixing of number of clusters may lead to poor quality clustering. The proposed method finds the number of clusters on the run based on the cluster quality output. This method works for both the cases i.e. for known number of clusters in advance as well as unknown number of clusters. The user has the flexibility either to fix the number of clusters or by input the minimum number of clusters required. In the former case it works same as K -means algorithm. In the latter case the algorithm computes the new clusters by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality threshold. The modified algorithm is as follows:

$$\text{Inter} = \min \{ \square m_k - m_{kk} \square \} \nabla \; k = 1, 2, \ldots\ldots\ldots\ldots K\text{-}1 \text{ and } \; kk = k+1, \ldots\ldots\ldots\ldots\ldots, K$$

Eq.2

$$\sqrt{\frac{1}{n-1}} \sum_{i=1}^{n} (x_{i} - x_{m})^2$$

Eq.3

Input:

- k: number of clusters (for dynamic clustering initialize k=2) Fixed number of clusters = yes or no (Boolean).
- D: a data set containing n objects.

Output:

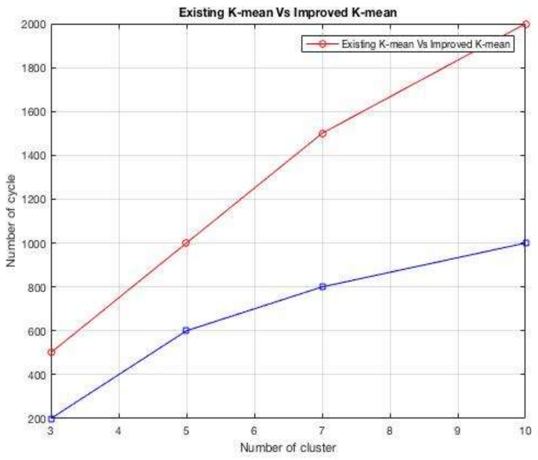
A set of k clusters.

Method:

1. Arbitrarily choose k objects from D as the initial cluster centers.
2. Repeat.
3. (re)assign each object to the cluster to which the object is most similar, based on the mean value of the objects in the cluster.
4. Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
5. until no change.
6. If fixed_no_of_clusters =yes goto 11.

7. Compute inter-cluster distance using Eq.2
8. Compute intra-cluster distance using Eq. 3.
9. If new intra cluster distance < old_intra_cluster distance and new_inter-cluster >old_inter_cluster distance goto
10. else goto 11.
11. k= k + 1 goto step 1.
12. STOP

This graph shows comparison of existing kmean algorithm and improved k mean algorithm based on number of cluster and number of cycle performed by algorithms.



4. CONCLUSION

In this Paper, the problem was to solve the K-clustering problem by introducing a clustering technique – Multilevel context of K-means. The problem in clustering as we notice in result part of chapter 5, is because of the nature of K-means method which in the first step the centroids are initialized randomly. Sometimes we have a poor clustering (some clusters don't have any member). The goal is clustering in the best behaviour, which should be to group similar data points as much as possible. But with basic K-means clustering this rarly is the case.

To optimize K-means, we propose an algorithm. Which in this method first take 2 data points of N data points randomly were selected, after calculate average of each this 2 data points, generate one new data point. Then we reduce the size of data point to N/2 and repeat this reduction until the number of data points in last reduction, are equal or greater than 10 % of N. Then we were running K-means algorithm of each layer and also in each layer, 10000 times by exchanging points between clusters and get the minimum SSE, we try to reach to optimal clustering.

There are several different ways to extend our results. First; the current model can deal with only a simple case of basic K-means clustering. The issue of how to deal with general constrained K-means clustering still remains open. It is worth mentioning that in this algorithm instead of choosing randomly each two point, another method for reduction can be used instead. It could be considered a research study in itself to find a method of choosing these two points.

RFERENCES:

[1] M. S. V. K. Pang-NingTan, "Data mining," in *Introduction to data mining*, Pearson International Edition, 2015, pp. 2-7.

[2] J. Peng and Y. Wei, "Approximating k-means-type clustering viasemi definite programming," *SIAM Journal on Optimization*, vol. 18, 2014.

[3] D. Alexander "Data Mining" [Online].Available: http://www.laits.utexas.edu/~norman/BUS.FOR/course.mat/Alex/.

[4] "What is Data Repository," Geek Interview 4 June2013.[Online]. Available: http://www.learn.geekinterview.com/data-warehouse/dw-basics/what-is-data-repository.html.

[5] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (2014), from Data mining to knowledge discovery in data base.

[6] M. S. V. K. Pang-NingTan, "Data mining," in Introduction to *data mining*, Pearson International Edition, 2015 pp. 8.

[7] M. S. V. K. Pang-NingTan, "Data mining," in *Introduction to data mining*, Pearson International Edition, 2014, pp. 7-11.

[8] Han, Jiawei, Kamber, Micheline. (2014) Data Mining: Concepts and Techniques. Morgan Kaufmann

[9] M. S. V. K. Pang-NingTan, "Data mining," in *Introduction to data mining*, Pearson International Edition, 2015, pp. 487-496.

[10] "An Introduction to Cluster Analysis for Data Mining," 2013.[Online]. Available: http://www.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.

[11] Joaquín Pérez Ortega, Ma. Del Rocío Boone Rojas, María J. Somodevilla García Research issues on, K-means Algorithm: An Experimental Trial Using Matlab.

[12] J. MacQueen, "Some Methods For Classification And Analysis Of Multivariate Observations," In proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 2015, pp. 281-297.

[13] M. S. V. K. Pang-NingTan, "Data mining," in *Introduction to data mining*, Pearson International Edition, 2012, pp. 496-508.

[14] Jain, A. K., Murty, N.M. and Flynn, P.J. (2015) Data Clustering: A Review.ACM Computing Surveys, Vol.31 No.3, pp. 264-323.

[15] V. Braverman, A. Meyerson, R. Ostrovsky, A. Roytman, M. Shindler and B. Tagiku, "Streaming k-means on Well-Clusterable Data," pp. 26-40, 2011.

[16] Stuart Lloyd. "Least Squares Quantization in PCM". In Special issue on quantization, IEEE Transactions on Information Theory, volume 28, PP. 129,137, 2014.