International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-2, Issue-7,July 2016
ISSN: 2395-3470
www.ijseas.com

Different versions of K-Mean Clustering in complete set of numerical data points

S.Gokila[1], K. Ananda Kumar [2] , A. Bharathi[3]
[1] Research Scholar, Bharathiar University, Coimbatore, India
[2] Bannariamman Institute of Technology, Erode, India
[3] Bannariamman Institute of Technology, Erode, India

Abstract

The traditional K-Mean algorithm under went many versions of changes in its each stages of working procedure in finding cluster, patterns and outlines in given input data set. The enhancements are done in centroid fixing, finding similarity or dissimilarity among the clusters, reallocation of data points into other set of clusters, evaluating and re-evaluating new centroid. Each of these versions contains its uniqueness either in performance, accuracy and convergences. This paper presents the study report of above mentioned enhancements in traditional K-Means clustering algorithms. This study report will be useful for the researching to kick start the work on k-Means form the final version.

Keywords : *K-Means, Centroid, Distance, Outline, Project Space*

1. Introduction

Data mining is a technology automates the process to discover interesting and sensitive patterns from the large collection of data set. It enables the human understandability of discovering patterns and scalability of techniques[3]. The data mining techniques used to do either descriptive mining (describe general properties - clustering) or predictive mining (attempt to predict based on inference of data - classification) on large volume of data.

Cluster analysis is a explore the structure of data. Core Cluster analysis is a clustering. Clustering analysis in a data is a unknown label class (unsupervised) [2]. So it is learned by observation not learned by example [1]. Clustering divide the data set into classes using the principle of “Maximum intra class similarity and Minimum inter class similarity”. It doesn’t have any assumption about the category of data. The basic clustering techniques are Hierarchical, Partitioning, Density based, Grid based and Model based clustering. Some sort of measure that can determine whether two objects are similar or dissimilar is required to add them into particular class. The distance measuring type varies for different attribute type. Clustering can also used to detect outline in data which may occur due to human error or some abnormal events occurred while creating data set [1]. Cluster work well on scalable, heterogeneous and high dimensional data set. In all the clustering algorithms user defined parameters are given as input to find either similarity, dissimilarity among clusters and for root attribute of cluster and for maximum or minimum number of clusters.

The partitioning based algorithm divides the data set into cluster based on specific prototypes. One of the partition based clustering algorithm is K- Means clustering algorithm.

1.1 K-Means Clustering

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. [4] The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields. The traditional K-Means clustering algorithm works on the data set with the attributes of numerical types[1,2,3]. Clustering divide the data set into classes using the principle of “Maximum intra class similarity and Minimum inter class similarity”. This

principle is evaluated in cluster quality checking process.

1.1.1 Traditional K- means algorithm

The step wise working procedure of traditional K-Means Algorithm is:

Pre Requisition:
D = {d1, d2, d3, ..., dj, ..., dn } // Set of n data points.
k Number of desired clusters
Ensure: A set of *k* clusters.

Step1: Arbitrarily choose *k* data points from *D* as initial centroids (C_i).
where i = 1 to K.

Repeat
Step 2 : Calculate distance between d_j and C_i.
where j = 1 to n
Step 3: Assign each point *d*j to the cluster which has the closest centroid.
Step 4: Calculate the new mean for each cluster treat that as new centroid.
Step 5: Evaluate the cluster quality.
Step 6: If the convergence met stop.
Step 7: Else Repeat Step 2 to Step 6
Until

The data set (D) and the number of clusters (K) are the two mandatory inputs for K-Means algorithm. The initial centroids of each cluster are predicted or selected using threshold base evaluation or random selection respectively. In second step the distance between data points and the centroid using is evaluation using any of distance finding method like, Euclidean Distance, Manhattan or City Block, Chebysev or Maximum value distance, MinKowski, Jaccard, Dice's Co-efficient, Russell/Rao, Bit Vector, Hamming Distance – String/ Text, Cosine, Correlation Distance based the data type of attributes in data set.

The data point with minimum distance with the centroid is allocated to respective cluster to stick on the clustering principle. At the end of assigning all the instance to be in cluster with minimum distance(Dis_i). The new mean of all the K number of clusters are calculated in Step 4. The new mean is the centroid of respective cluster.

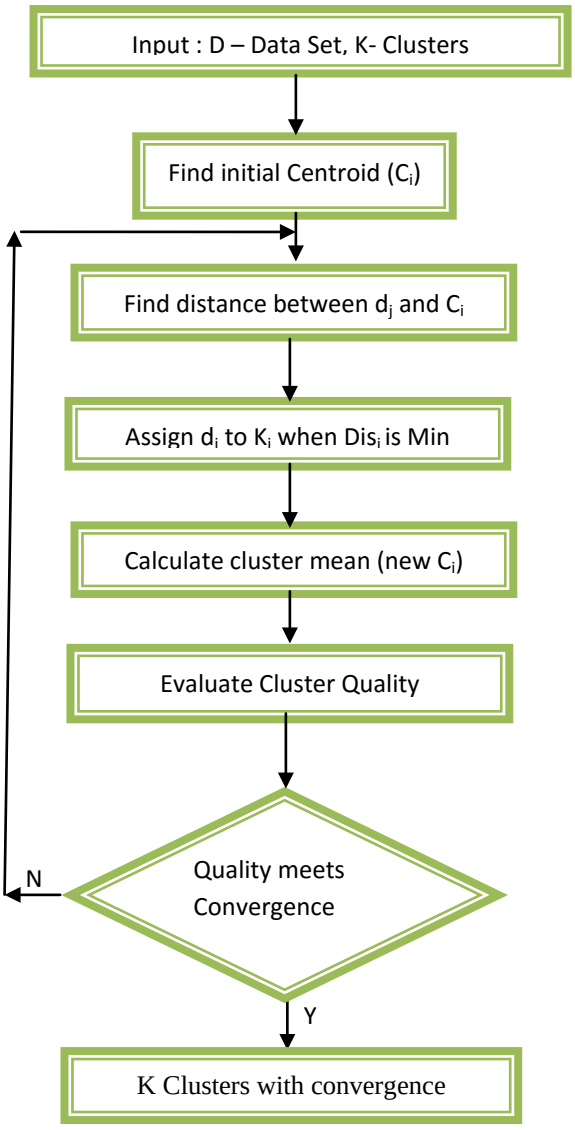

Fig 1 : K – Means algorithm

The Step 5 evaluate the cluster quality using either of Squared Error Criterion, Absolute Error Criterion, Distance between clusters. The time, memory usage, scalability, noise data handling, data order independency or arbitrary shaped cluster are other criteria to evaluate the clustering algorithm performance[3]. The algorithm get into next

iteration when the cluster quality is low. The Fig 1 gives the clear vision on algorithm working procedure.

In a research person perspective they can work on each step of algorithm to improve the performance. Like wise K-Mean enhance into many versions and still it required many areas to be treat with. This paper guide the researchers to work on next version of the K-Mean algorithm. In this perspective remaining section are organized as related survey of k-mean, analysis of k-mean's different versions and conclusion

2. Related Work

The traditional K-Means clustering algorithm works on the data set with the attributes of numerical types[1]. It is used in many domains like, image processing, linear data study, encryption, climate data analysis[4,15, 18]. The k-means clustering algorithm implemented on weather data to ensemble four different seasons of data point to study the climate[16]. All of these usage will directly applies the traditional k-means. Each of the perspective differs in terms of required out come of the algorithm. So the k-mean transformed into many versions based on the required output. The experimental study used the algorithm as such with different distance measuring strategies. The enhancement introduced in automation of the algorithm, which try to eliminate the user input like threshold value select the centroid and the number of clusters required [6,7]. The paper analysis the each level of enhancements in the traditional k-means algorithm.

3. K-Means Version Analysis

The K-Means clustering algorithm is very simple effective procedure in finding patterns in numeric data set. The contribution of computer and information technology researchers is to give the optimal K-means algorithm to apply the same on different domains. Whichever the domain applying the traditional k-mean is, the far most requirement from the domain expert to give the numeric data set. Even there are some advanced k-means like fuzzy k-mean which can do the fuzzy set on categorical data.

3.1 Enhancement in Distance calculation

Based on the domain the k-mean used for different output. In which the attribute type may vary like, complete numerical, binary value, ordinal number, cardinal value and categorical value. Due to this purpose the enhancement made in equations used in distance calculation between centroid and selected data point. In [5] complete set of numerical data is used to get the distortion in data point. The better result produced with Euclidean distance.

The expected outcome of each method differs. One choose the best quality cluster in distance concern, one in sensitivity of pattern, other may for response time. In [9] the expectation of research is to find the running time of algorithm based on the distance equation used. The distance equation CityBlock method outperform in computation time of algorithm. The approach produces better result in one aspect may perform low in other. In [10] the ChebyShev distance measured applied on Iris data set produce good accuracy but more iteration than other method, ChebyShev produced good accuracy on flower data set also when compared with Euclidean distance and Manhattan distance. It guides to select the method based on the requirement.

Some measures says what to use and what to not. The distance equations have to get selected based on the type of data. The Euclidean distance is not a accurate method of distance measuring on binary data set [11]. The table1 summarizes the enhancements in distance measurements in clusters.

3.2 Enhancement in Centroid

Many research results produced many versions of K-Means with the enhancement in any one part of algorithm. From the first step of

centroid selection to algorithm optimization. This can be achieved by picking one data instance randomly or using threshold [5]. The final clustering result of the kmeans

Table 1: Distance measure

Reference No	Distance Measured	Data Set	Prediction
[5]	Euclidean, Manhattan and Minkowski	Synthesis data	Euclidean performs best in finding distortion.
[8]	Projected Space	Iris, Wine Data	PCA with Gaussian width produced good result minimizing clusters.
[9]	Euclidean distance Cosine distance City Block distance	Iris, Wine Data	City Block outperform in computation time.
[10]	Euclidean Distance Manhattan Distance ChebyShev distance	Iris	ChebyShev produce good accuracy with more iterations.
[11]	Eculidean distance	dichotomous variables	Not correct one for binary data set.
[12]	Euclidean distance Manhattan distance, ChebyShev distance	Flower Data set	ChebyShev produce good accuracy.

clustering algorithm greatly depends upon the correctness of the initial centroids, in traditional algorithm it is selected randomly [7]. In one enhancement divide the data point into 100 parts. Select one data point from data set, find the distance between selected one and other data point in set. The data point as statistic and density as 1 If the distance meet the threshold. For non statistic data point find the density value. Select the data point with density greater than threshold value as initial centroid of cluster [6]. This method is partially automated for centroid calculation, but the initial data set is picked in random basis. In [7] for each data point distance is calculated from origin, obtained distance values are sorted the the data points accordance with the distances also sorted. Then partition the sorted data points into k equal sets, In each set, middle point has taken as the initial centroid. This method is independent of threshold value.

Table 2 summarizes the enhancements in centroid selection.

Table 2: Centroid Selection

Ref.	Method	Data Set	Prediction
[[7]	No threshold value	Iris New Thyroid Echocardiogram	Cluster accuracy
[6]	Random pick and threshold	Iris, Wine	Cluster quality

3.3 Enhancement in Iteration Reduction

The k-mean work well in projected space to cluster the data. This projection is made to reduce the number of clusters and to reduce the iteration[8]. The K-Mean clustering applied after the projection of data set give the better result in iteration and also in the accuracy of cluster[13]. The projected space clustering work on categorical data set to identify the necessary by giving some weight to the attribute. The k-mean works on weighted projected data set[14, 16].

4. Conclusion

The K-mean clustering is the simplest algorithm in finding a pattern in data set of many domain. The compute research perspective is to produce a unique optimum algorithm suitable for all the domain. For this proposal the researcher get updated with all the levels of improvements in algorithm. This study analysis support to learn the enhancements in k-mean algorithm. The centroid selection of initial step decides the entire performance of algorithm. Which under went upmost automation with out the user input like threshold value. In another optimization required in existing logic in terms of time, usage and quality. Deciding authority of all these factor are in the hand of domain in which it is applied. As the domain data type varies the accuracy depends on the distance measure used. Even the accuracy of some distance measure is high the iteration also high. In order to reduce that the data are projected with weight based on the importance. The k-mean applied on the projected space data set to reduce the iteration and to reduce the computation time.

References

[1] Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques”, India, MK Publications, 2006.

[2] K. P. Soman, Shyam Diwakar, V. Ajay, “Insight into Data Mining Theory and Practice” – India, PHI Learning , 2014.

[3] Vikram Pudi, P. Radha Krishna, “Data Mining”, India, Oxford Higher Education, 2011.

[4] T V Rajini kanth, V V SSS Balaram and N.Rajasekhar “Dhinaharan Nagamalai, ANALYSIS OF INDIAN WEATHER DATA SETS USING DATA MINING TECHNIQUES”, Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, Volume 1, Issue 2, 2014, pp. 89–94.

[5] Archana Singh, Avantika Yadav, Ajay Rana, “K-means with Three different Distance Metrics” International Journal of Computer Applications”, Volume 67, No.10, 2013, pp: 13-17.

[6] Chunfei Zhang, Zhiyi Fang , “An Improved K-means Clustering Algorithm”, Journal of Information & Computational Science, Volume 10, No. 1, 2013, pp :193-199.

[7] Madhu Yedla, Srinivasa Rao Pathakota, T M Srinivasa “Enhancing K-means Clustering Algorithm with Improved Initial Center”, International Journal of Computer Science and Information Technologies, Volume 1, No. 2, 2010, pp : 121-125.

[8] Alissar NASSER, Denis HAMAD, Chaiban NASR, “K-means Clustering Algorithm in Projected Spaces”, IEEE Explore 9th International conference on information fusion, Volume 9, No. 1, 2006, pp: 1-6.

[9] Dibya Jyoti Bora, Dr. Anil Kumar Gupta, “Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab”, International Journal of Computer Science and Information Technologies, Volume 5, No. 2 , 2014, pp 2501-2506.

[10] Shraddha Pandit, Suchita Gupta, “A comparative study on distance measuring approaches for clustering”, International Journal of Research in Computer Science, Volume 2, No. 1 , 2011, pp: 29-31.

[11] Holmes Finch , “Comparison of Distance Measures in Cluster Analysis with Dichotomous Data”, Journal of Data Science, Volume 3, No. 1, 2005, pp :85-100.

[12] Kahkashan Kouser, Sunita, “A comparative study of K Means Algorithm by Different Distance Measures”, International Journal of Innovative

Research in Computer and Communication Engineering, Volume. 1, No. 9, 2013, pp :2443-2447.
[13] Ilango Murugappan," PCFA: Mining of Projected Clusters in High Dimensional Data Using Modified FCM Algorithm", The International Arab Journal of Information Technology, Volume. 11, No. 2, 2014, pp :168-177.
[14] M. Bouguessa, "Clustering categorical data in projected spaces", Data Min Knowl Disc, Volume 29, No. 1, 2015, pp :3-38.
[15] Muthukumar Arunachalam and Kannan Subramanian, "Finger Knuckle Print Authentication Using AES and K-Means Algorithm", The International Arab Journal of Information Technology, Volume 12, No. 6A, 2015, pp 642-649.
[16] Maheswari Neelam Harikrishnan, Dr. Bala Buksh, "Recent Trends in Correlation Clustering", International Journal of Scientific Engineering and Applied Science, Volume 2, No 2, 2016, pp: 244-248.
[17] Sarah N. Kohail, Alaa M. El-Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis (A case study for Gaza Strip)", International Journal of Information and Communication Technology Research, Volume 1, No. 3, 2011, pp 96-100.
[18] Aswini Gulhane, Shreyas Deshmukh, "Generalized method for image data clustering", International Journal of Scientific Engineering and Applied Science, Volume 1, No 2, 2015, pp:19-26.

S. Gokila received her B.Sc[CT & App Sci.]., M.C.A., from Bharathiyar University in the year 2001 and 2004 respectively and M. Phil., from Periyar Univerity in the year 2007. Currently pursuing Ph.D in Computer Science at Bharathiyar University. She is working as Asst. Professor in the Dept. of Computer Application, Shri Shankarlal Sundarbai Shasun Jain College for Women, Chennai, Inida. She qualified UGC NET examination in June 2015. She published research papers in various National and International Journals. And also published book "Advanced Java Programming". Her area of interest is objected oriented programming and data mining.

Dr. K. Ananda Kumar was born in Tamilnadu. He did his Bachelor's degree at Sri Ramakrishna Mission Vidhyalaya Coimbatore in B.Sc Physics under Bharathiar University, Coimbatore in the year 1995. He then completed his Post Graduate Degree at Kongu Nadu Arts and science in Computer Applications under Bharathiar University, Coimbatore in the year 1998. He finished his Doctoral degree in Computer Science Specializing in Data Mining in the year 2012. He is working as Associate Professor in the Department of Computer Applications, Bannari Amman Institute of Technology, Sathyamangalam, Erode, India. He has over 18 years of teaching experience. He published and presented more than 30 journals and 15 international conferences.

Dr. A. Bharathi was born in Tamilnadu. She did her Bachelor's degree at Kongu Engineering College, Perundurai in Computer Science and Engineering under Bharathiar University, Coimbatore in the year 1998. She then completed her Post Graduate Degree at Bannari Amman Institute of Technology in Computer Science and Engineering under Anna University, Chennai in the year 2007. She finished her Doctoral degree in Information and Communication Engineering Specializing in Data Mining in the year 2012. She is working as Professor in the Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, Erode, India. She has over 18 years of teaching experience. She published and presented more than 40 journals and 20 international conferences.