RESEARCH HUB – International Multidisciplinary Research Journal (RHIMRJ)
Research Paper
Available online at: www.rhimrj.com

Survey on Different Distance Metrics in K-Means Algorithm

[1]Miss. Kothariya Arzoo
Student, Computer Engineering,
C.U. Shah College of Engineering and technology,
Surendranagar, Gujarat, India

[2]Asst. Prof. Kirit Rathod
Asst. Prof., Computer Engineering
C.U. Shah College of Engineering and Technology,
Surendranagar, Gujarat, India

Abstract: ***Clustering is process of grouping similar data points or objects into same group. The k-means clustering is popular for cluster analysis in data mining. It is unsupervised because the points have no external classification. Distance metrics are used to find similar data objects on the basis of distance between data points and center points. so, distance metrics are very important element in k-means algorithm and play vital role in k means clustering. There are many distance metrics are available. In this paper we will do a review on k-means algorithm with different distance metrics.***

Keyword: Clustering, K-Means, Distance Metrics

I. INTRODUCTION

Clustering [1] involves the task of dividing data points into similar clusters so that items in the same cluster are as similar as possible and items in different cluster are as dissimilar as possible. K-Means Algorithm is very Simple algorithm that is used in data mining [2]. Clustering is an unsupervised method [3]. The main aim of clustering is to divide a dataset into some 'clusters' such that data into one cluster sharing similar properties while data in different cluster showing different properties from data in another cluster. Depending on the data and the application, different types of similarity measures may be used to identify classes, where the similarity measure controls how the clusters are formed Unsupervised learning means there will be no training data and testing data instead the learning is by observation. Any cluster should exhibit two main properties; low inter-class connection and high intra-class connection [3]. Clustering algorithms are mainly divided into two types based on developed cluster properties: hierarchical and partitional [4]. K-means is very popular partitional algorithm [5]. There are many variants of K-Means Algorithm [4]. Variants of k means on the basis of centroid initialization, distance metrics, variants for improving k- means accuracy etc.

II. ALGORITHM: K-MEANS

K-Means is one of the simplest unsupervised learning algorithms. In general, the algorithm accepts some initial parameters to determine the initial number of clusters and then randomly locates the cluster centers in the multidimensional feature space The K-Means [4] is one of the famous partition clustering algorithm [5]. K-Means Algorithm processes as: 1) initialization by setting center points (or initial centroids) with a given K, 2) Dividing all data points into K clusters based on K current centroids, and 3) updating K centroids based on newly formed clusters. It is clear that the algorithm always converges after several iterations of repeating steps 2) and 3). As a result of this we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an *objective function [8]*, in this case a squared error function [7]. The objective function

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m} (x_{ik} - y_{ik})^2}$$

A graphical representation [7] of the K-Means algorithm is

  ISSN: 2349-7637 (Online)

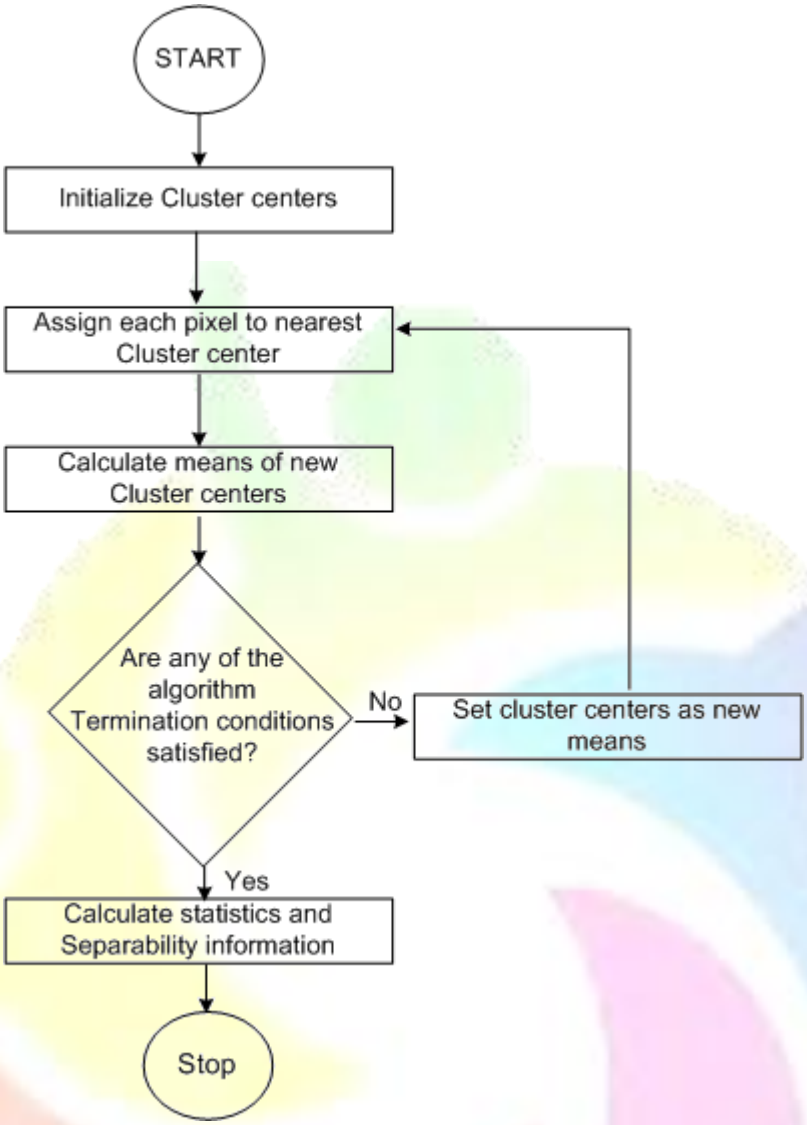

Chart -1: Flowchart of K-Means Algorithm

III. DISTANCE METRICS

Distance metrics [9] [11] [12] plays a very important role to measure the similarities among the data items.
A metric function or distance function is a function which defines a distance between elements/objects of a set [4]. As we know, distance between two points can be computed with different techniques available, so, our main aim is to pick up a proper technique from the available ones.

3.1 Euclidean Distance

The Euclidean distance [11] between two points, *a* and *b*, with *k* dimensions is calculated as : The Euclidean distance or Euclidean metric is the ordinary distance between two points that one would measure with a ruler. It is the straight line distance between two points.

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m} (x_{ik} - y_{ik})^2}$$

One advantage of this method is that the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers [13].

ISSN: 2349-7637 (Online)

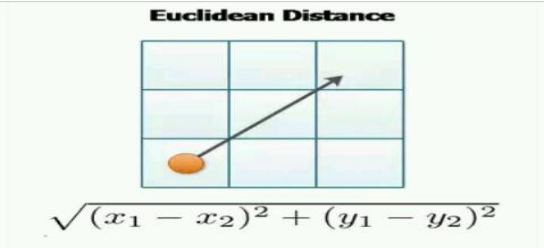

Fig -1: Euclidean Distance

K-Means Algorithm using Euclidean Distance [11]:

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1. Select 'c' cluster centers randomly.
2. Calculate the distance between each data point and cluster centers using the Euclidean distance metric as follows

$$Dist_{xy} = \sqrt{\sum_{k=1}^{m} (x_{ik} - y_{ik})^2}$$

3. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. New cluster center is calculated using:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{1}^{c_i} x_i$$

5. The distance between each data point and new obtained cluster centers is recalculated.
6. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

3.2 Manhattan Distance (City Block)

The *City block distance* [14], also known as Manhattan distance, absolute distance. It represents distance between points in a city road grid. It examines the absolute differences between coordinates of a pair of objects. $Dist_{xy} = |x_{ik} - y_{ik}|$

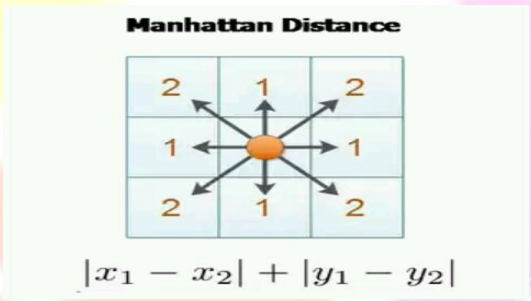

Fig-2 : Manhattan Distance

K-Means Algorithm using Manhattan Distance [11]

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1. Select 'c' cluster centers randomly.
2. Calculate the distance between each data point and cluster centers using the Euclidean distance metric as follows

$Dist_{xy} = |x_{ik} - y_{ik}|$

3. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. New cluster center is calculated using:

$$v_i = \left(\frac{1}{c_i}\right) \sum_{1}^{c_i} x_i$$

5. The distance between each data point and new obtained cluster centers is recalculated.
6. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

3.3 Chebychev Distance

Chebychev Distance [12] is also known as maximum value distance and is computed as the absolute magnitude of the differences between coordinate of a pair of objects. Which is defined as:

$$max_k\ |x_{ik} - y_{ik}|$$

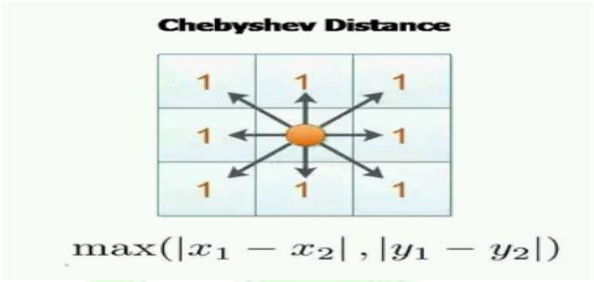

Fig- : Chebychev Distance

K-Means Algorithm using Manhattan Distance [11]

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1. Select 'c' cluster centers randomly.
2. Calculate the distance between each data point and cluster centers using the Euclidean distance metric as follows

$$Dist_{xy} = max_k\ |x_{ik} - y_{ik}|$$

3. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. New cluster center is calculated using:

$$v_i = \left(\frac{1}{c_i}\right)\sum_{1}^{c_i} x_i$$

5. The distance between each data point and new obtained cluster centers is recalculated.
6. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

Inspite of these many other distance metrics are available. Some of these are as follows:

1) Minkowski Distance defined as:

$$Dist_{xy} = \left(\sum_{k=1}^{d} max\,|x_{ik} - x_{jk}|^{\frac{1}{p}}\right)^{p}$$

2) Cosine similiarity distance defined as:

$$Dist_{xy} = \frac{\left(\sum_{k=1}^{d} x_i\ y_i\right)}{\sqrt{\sum_{k=1}^{d} x^2{}_i}\ \sqrt{\sum_{k=1}^{d} y^2{}_i}}$$

3) Canberra distance defined as:

$$Dist_{xy} = \sum_{k=1}^{d} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

IV. LITERATURE REVIEW

4.1 A.K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, vol. 31, pp. 264-323, Sep. 1999.

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis.

4.2 Venkatadri.M, Dr. Lokanatha C. Reddy. A Review on Data mining from Past to the Future, International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011

Data and Information or Knowledge has a significant role on human activities. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Hence, this paper discusses the various improvements in the field of data mining from past to the present and explores the future trends.

ISSN: 2349-7637 (Online)

4.3P.Berkhin.(2001)"Survey of Clustering Data Mining Techniques" [Online]. Available:http://www.accure.com/products/rp_cluster_review.pdf.

Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis.

4.4 Aakansha Chaudhry, K-Means and its variants, International Journal of Innovative Research in computer and communication Engineering,Vol 4, Issue 1, 2016

K-means is most widely used method of clustering. It randomly select k objects which act as initial centroids for k clusters and then iteratively assign rest of the objects to these k clusters on the basis of similarities between the objects of clusters. The goal of this survey is to study research and development work done on k-means clustering method.

4.5 Kehar singh, dimple malik, Naveen Sharma.Evolving Limitations in k-means algorithm in data mining and their removal. IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011

K-means is very popular because it is conceptually simple and is computationally fast and memory efficient but there are various types of limitations in k means algorithm that makes extraction some what difficult. In this paper we are discussing these limitations and how these limitations will be removed.

4.6 Dibya Jyoti Bora,Dr. Anil Kumar Gupta. Effect of Different Distance Measures on the Performance of K-Means Algorithm: An Experimental Study in Matlab, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 2501-2506

K-means algorithm is a very popular clustering algorithm which is famous for its simplicity. Distance measure plays a very important rule on the performance of this Algorithm. In this paper an experimental study is done in Matlab to cluster the iris and wine data sets with different distance measures and thereby observing the variation of the performances shown.

4.7 Archana singh, Avantika Yadav,Ajay rana. K-Means with three different distance metrics, International journal of computer Applications, (0975 – 8887) Volume 67– No.10, April 2013

In this paper, the results obtained by implementing the k-means algorithm using three different metrics Euclidean, Manhattan and Minkowski distance metrics along with the comparative study of results of basic k-means algorithm which is implemented through Euclidian distance metric for two-dimensional data, are discussed.

4.8 B.S.Charulatha, Paul Rodrigues, T.Chitralekha, Arun Rajaraman Member IEEE A Comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms, International Journal of Emerging trends & technology in computer science(IJETTCS)- Special Issue ISSN -2278 6856

Fuzzy clustering is a broad classification of clustering methods .They are helpful when there exists a dataset with sub groupings of points having indistinct boundaries and overlap between the clusters. This paper reviews FCM with five distance metrics can be used with fuzzy clustering. They are the Euclid, Manhattan, Canberra, TChebychev and Cosine or angular. Performance of the metrics are presented and compared.

V. CONCLUSIONS

The K-Means algorithm is very famous partitional algorithm. K-Means Algorithm is robust simple and easy to understand. This algorithm fails for non-linear dataset. As per all the paper I have observed that distance measures play very important role in k-means clustering. Euclidean distance works good where all dimensions are properly scaled. The city block distance is same as you move in a city where you have to move around the buildings instead of going straight through. As per the paper[9] they have found that city block distance shows better performance for both the datasets in terms of less computation time. As per the paper [10] they have found that the distortion in k-means using Manhattan distance metric is less than that of k-means using Euclidean distance metric. The K-means, which is implemented using Euclidean distance metric gives best result and K-means based on Manhattan distance metric's performance, is worst. So, Manhattan Distance is based on absolute value distance as opposed to squared error distance. I conclude that Extensive exploration on the distance metrics needs to be done on various data sets on K-Means algorithms.

REFERENCES

1. A.K. Jain, M. N. Murty, P. J. Flynn, "Data Clustering: A Review", ACM Computing Surveys, vol. 31, pp. 264-323, Sep. 1999.
2. Venkatadri.M, Dr. Lokanatha C. Reddy. A Review on Data mining from Past to the Future, International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011

3. P.Berkhin.(2001)"Survey of Clustering Data Mining Techniques" [Online]. Available:http://www.accure.com/products/rp_cluster_review.pdf.
4. J. Han , M. Kamber, Data Mining, Morgan Kaufmann Publishers, 2001.
5. Lloyd., S. P. "Least squares quantization in PCM". IEEE Transactions on Information Theory 28 (2), 1982, pp. 129–137
6. Aakansha Chaudhry, K-Means and its variants, International Journal of Innovative Research in computer and communication Engineering,Vol 4, Issue 1, 2016
7. nptel.ac.in/courses/105104100/lectured_28_4.htm
8. www.tutorialspoint.com/Data_Mining
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, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (2), 2014, 2501-2506
10. Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar & M.Inayat Khan, "Data Mining Model for Higher Education System ",European Journal of Scientific Research, ISSN 1450-216X Vol.43 No.1 ,2010, pp.27
11. Archana singh, Avantika Yadav,Ajay rana. K-Means with three different distance metrics, International journal of computer Applications, (0975 – 8887) Volume 67– No.10, April 2013
12. B.S.Charulatha, Paul Rodrigues, T.Chitralekha, Arun Rajaraman Member IEEE A Comparative study of different distance metrics that can be used in Fuzzy Clustering Algorithms, International Journal of Emerging trends & technology in computer science(IJETTCS)-Special Issue ISSN -2278 6856
13. Sanjay Chawla, Aristieds giones.k-means--: A unified approach to clustering and outlier detection
14. ple.revoledu.com/kardi/tutorial/Similiarity/CityBlockDistance.html

  ISSN: 2349-7637 (Online)