ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

A comparative study of K Means Algorithm by Different Distance Measures

Kahkashan Kouser[1], Sunita[2]

Assistant Professor, Dept. of Cambridge Institute of Technology, Ranchi, India[1,2]

ABSTRACT: This document gives formatting instructions for authors preparing papers for publication in the Proceedings of an International Journal. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text. Clustering is very important research areas in the field of data mining. In simple words, clustering is a division of elements into different groups. Data are grouped into clusters in such a way that elements of the same group are similar and those in other groups are dissimilar. K-Means algorithm is an important method for finding clusters .Its implementation is very simple and fast execution. In this paper the K-Means clustering algorithm is applied on flower data set is studied .here various distance function such as Euclidean distance and Manhattan distance ,Chebyshev distance function is used for analyzing the result of number of iterations, Overall Accuracy, Mean absolute error

Keywords: Simple K Means, Euclidean Distance, Manhattan Distance, Chebyshey distancence

I. INTRODUCTION

A. *K-means Clustering* [1]

The process of dividing a set of physical or abstract object into classes of similar objects is called clustering. A cluster is a collection of data object that are similar to one another within the same cluster and are dissimilar to the object in other cluster. An important and commonly used partitioning method is K Means. The K Mean partitions a set of N objects into K cluster. Cluster similarity is measured in regard to the mean value of the object in a cluster, which can be viewed as the cluster centroid or center gravity.Another definition of clustering is " it is the process of organizing elements into groups whose elements are similar in some way"[2]. So cluster is a collection of elements whichare "similar" between them and are "dissimilar" to the elements belonging to other clusters. Unlike classification, in which elements are assigned to predefined classes, clustering does not have any predefined classes. The main advantage of clustering is that interesting patterns and structures can be found directly from very large data sets with not any priori knowledge about the clusters. The quality of a clustering method depends on:

- The similarity measure used by the method and its Implementation.
- Its efficiency to discover some or all of the hidden patterns.
- How to define and represent the chosen cluster .

B. Measurement of Distance Between Objects And Means

An important component of a clustering algorithm is the distance measure between data points. The problem arises from the mathematical formula that are used to combine the distances between the single components of the data feature vectors into a unique distance measure that can be used for clustering purposes: different methods leads to different clustering. The most popular distance measure are Euclidean distance Manhattan distance ,Chebyshev distance function.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

II. RESEARCH METHODOLOGY

A. *Distance metrics overview*[3]

Euclidean distance is the most commonly used – it calculate the root of square differences between coordinates of two objects.

$$D_{XY}=\sqrt{\sum_{k=1}^{m}(x_{ik}-x_{jk})^2}$$

Manhattan distance or city block distance represents distance between two points in a city road grid. It computes the absolute differences between coordinates of two objects:

$$D_{XY}=\sum_{k=1}^{m}|x_{ik}-x_{jk}|$$

Chebyshev distance is also known as Maximum value distance. It calculate absolute magnitude of the differences between coordinates of two objects:

$$D_{XY}=\max_k|x_{ik}-x_{jk}|$$

Minkowski distance have metric distance:

$$D_{XY}=(\sum_{k=1}^{d}|x_{ik}-x_{jk}|^{1/p})^p$$

Note that when p=2, it represent the Euclidean distance. When p=1 it represent city block distance. Chebyshev distance is a special case of Minkowski distance with p=∞ (taking a limit).

In order to show different metrics the following two points are used: point Phas coordinate (1, 2, 3, 4) and point Q has coordinate (5, 6, 7, 8).
For example, the Euclidean distance between point P and Q is:

$$D_{pq}=\sqrt{(2-3)^2+(3-5)^2+(4-7)^2+(5-9)^2}\quad=5.5$$

The Manhattan distance between point Pand Q is:

$$D_{pq}=|2-3|+|3-5|+|4-7|+|5-9|\quad=10$$

The Chebyshev distance between point P and Q is:

$$D_{pq}=max\{|2-3|,|4-7|,|5-9|\}=max\{1,2,3,4\}=4$$

The Minkowski distance of order 3 between point P and Q is:

$$D_{pq}=(|2-3|^3+|3-5|^3+|4-7|^3+|5-9|^3)^{1/3}=4.6$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

Table1
Proximity measures and their applications

Measure	Metric	Examples and applications
Euclidean distance	Yes	K-means with its variants
Manhattan distance	Yes	Fuzzy ART, clustering algorithms
Chebyshev distance	Yes	Fuzzy C-means clustering
Minkowoski distance	Yes	Fuzzy C-means clustering

B. K-MEANS[4]

K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. It classifies a given set of *n* data objects in *k* clusters, where *k* is the number of desired clusters and it is required in advance. A centroid is defined for each cluster. All the data objects are placed in a cluster having centroid nearest (or most similar) to that data object. After processing all data objects, k-means, or centroids, are recalculated, and the entire process is repeated. All data objects are bound to the clusters based on the new centroids. In each iteration centroids change their location step by step. In other words, centroids move in each iteration. This process is continued until no any centroid move. As a result, *k* clusters are found representing a set of *n* data objects. An algorithm for k-means method is given below.

Algorithm

Input : '*k*', the number of clusters to be partitioned; '*n*', the number of objects.
Output: A set of '*k*' clusters based on given similarity function.
Steps:
i) Arbitrarily choose '*k*' objects as the initial cluster centers;
ii) Repeat,

a. (Re)assign each object to the cluster to which the object is the most similar; based on the given similarity function;
b. Update the centroid (cluster means), i.e., calculate the mean value of the objects for each cluster;

iii) Until no change.

The reason behind choosing k-means algorithm[5]
:

- Its time complexity is O(nkl), where n is the number of patterns, k is the number of clusters and l is the number of iteration taken by algorithm to coverge.
- Its space complexity is O(k+n).It requires additional space to store the data matrix.
- It is order independent; for a given initial seed set of cluster centers, it generates the same partition of the data irrespective of the order in which the patterns are presented to the algorithm

Similar to other clustering algorithms, k-means clustering has many drawbacks[2]

- Number of clusters, k, must be known in advance.
- It is difficult to find out the contribution each attribute makes to the grouping process, since it is assumed that each attribute has the equal value.
- By using the same data, we may never know the real cluster. If the number of data is a few, by inputting data in a different order, a result may be a different cluster.
- In case there are not many numbers of data, the cluster will be significantly determined by the initial grouping.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

- Weakness of arithmetic mean is not robust to outliers. As a result, the centroid may be pulled away from the real data by very far data.
- It is sensitive to initial condition, since different initial condition may lead to different result of cluster. The algorithm may be trapped in the local optimum.

As a result one gets a circular cluster shape which is based on distance

III. PROPOSED WORK

A. Materials and methods

The purpose of the experimental part was to test the operation of the k-means algorithm by applying different metrics. Three different metrics have been chosen: Euclidean distance, Manhattan distance and Chebyshev distance . In the course of the experiments in order to determine clusters by k-means clustering algorithm sequentially all three metrics have been used. The results obtained have been analyzed and the clustering correctness has been tested.

During the experiment the well-known Fisher's IRIS data set was employed , containing three species classes of 50 elements each: Setosa, Versicolor and Virginica. four attributes: SL - sepal length, SW - sepal width, PL - petal length, PW - petal width. The implementation is carried in 'C' language And Confusion matrix is used for comparison of the results.

TABLE 2

Experiment result of ,overall accuracy, number of iteration ,mean absolute error in Euclidean distance function , Manhattan and Chebyshey distance function

For data size 10

	Euclidean	Manhattan	Chebyshev
Overall Accuracy	80%	70%	100%
No. of iteration	100	50	150
Mean absolute error	0.16	0.6	0.12

TABLE 3

Experiment result of ,overall accuracy, number of iteration, mean absolute error in euclidean distance function , manhattan and chebyshey distance function

For data size 50

	Euclidean	Manhattan	Chebyshev
Overall Accuracy	84%	70%	86%
No. of iteration	20	10	30
Mean absolute error	0.2	0.3	0

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

IV. CONCLUSION

Cluster analysis groups various objects based on their similarity. Clustering analysis is the pivot for data mining. K-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem .The purpose of this experiment is to find the effect of distance functions on clustering. Euclidean distance function and Manhattan distance Chebyshev distance function were used to see this effect. The overall accuracy of Chebyshev is as greater as compared to Euclidean distance function and Manhattan distance while the number of iteration in Chebyshev distance function are greater as compared to Euclidean distance function and Manhattan distance function. Different approaches were used to measure the distance among various data objects which is the most significant step of creating cluster. So special consideration should be given to choose distance function and it should be chosen according to dataset and number of cluster.

REFERENCES

[1]Rekha Awathi, Anil K Tiwari,Seema Pathak "Empirical evaluation on k Means clustering with effect of distance function for bank database" in IJITR ,pp2013,233-235 vol -2,2013

[2] T. Velmurugan and T. Santhanam Department of Computer Science, DG Vaishnav College, Chennai, India"Computational Complexity between K-Means and K-Medoids Clustering Algorithms for Normal and Uniform Distributions of Data Points"

[3]Peter Grabusts "The choice of matrics for clustering algorithms",8th International Scientific and Practical conference, vol-2,2011.

[4]Shalini S Singh N C Chauhan" K-means v/s K-medoids: A Comparative Study" in National Conference on Recent Trends in Engineering & Technology.

[5] **Amit Singla and Mr. Karambir**" Comparative Analysis & Evaluation of Euclidean Distance Function and Manhattan Distance Function Using K-means Algorithm" in International Journal of Advanced Research in Computer Science and Software Engineering.

[6] Margaret H. Dunham "Data Mining Introductory and advanced topics" 2009.

BIOGRAPHY

Kahkashan kouser is Assistant Professor in the computer science Department, of Cambrigde Institute of Technology, Ranchi . She received Master of Computer Science (M.Tech) degree in 2013 from Birla Institute of Technology , India. Her research interests are Data mining ,Algorithms etc.

Sunita is Assistant Professor in the computer science Department, of Cambrigde Institute of Technology, Ranchi . She received Master of Computer Science (M.Tech) degree in 2013 from Birla Institute of Technology , India. Her research interests are Data mining ,Algorithms etc.