International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | Apr -2017 www.irjet.net p-ISSN: 2395-0072

K-Means algorithm with different distance metrics in spatial data mining with uses of NetBeans IDE 8.2

Ms. Kothariya Arzoo[1], Asst. Prof. Kirit Rathod[2]

[1] *M. Tech student, Computer Engineering, C.U. Shah College of engineering and technology, Gujarat, India*
[2] *Asst. Prof , Computer Engineering, C.U. Shah College of engineering and technology, Gujarat, India*

Abstract -*Data mining is a process of finding useful information from large database. Clustering is a process of grouping the same characteristics elements in one group (cluster) and while distinct characteristics elements in different group (cluster).K-Means is very simple and very popular clustering technique. In this paper we will do the experiments with spatial data mining. Spatial data mining is the application of data mining. In spatial data mining spatial or geographic dataset is used. Distance metrics play very important role in clustering technique. In this paper we will do the experiments with the NetBeans IDE 8.2 and taking spatial data from Indian government website. This paper includes implementation analysis on k means clustering with different distance metrics with taking spatial database.*

Key Words: *Clustering, Spatial Data mining, K-Means, Distance Metrics*

1. INTRODUCTION

Data mining [1] [2] involves process of finding useful information. Data mining consists of steps of a) data cleaning, b) data selection, c) data integration, d) data mining, e) pattern evaluation and finally f) knowledge representation [2]. Nowadays large amount of data generating day by day but this all data are not useful for us. So. Data mining is very essential for us in daily life. The data Clustering [3] involves the process of dividing the same data in one cluster and distinct data in different cluster so that inter cluster property should high and intra cluster property should low. These two are the clustering property. Clustering is the popular research topic nowadays. Classification of Clustering algorithm as: partitioning based clustering, hierarchical based clustering, grid based clustering etc [4]. A partitional clustering a simply a division of the set of data objects into non-overlapping subsets such that each data object is in exactly one subset. • A hierarchical clustering is a set of nested clusters that are structured as a tree. Clustering process involves feature selection, clustering algorithm, cluster validation, result interpretation, knowledge [5]. K-means [6] [7] is very simple clustering algorithm in data mining. Clustering is unsupervised technique [3] [4] such that there is no any test data and training data are available by which we can predict our result. Clustering is totally unsupervised method for data mining. There are many research and review paper available for k means and its variants [6]. There are many parameters for modifying k means algorithm based centroid initialization, distance metrics, improving accuracy. There are many limitations of k means algorithm [8] like handling empty cluster, outlier detection, distance metrics, number of cluster predefined, cluster center chosen randomly etc. Spatial data mining [9] [10] [11] requires specific resources to get the spatial database in specific format. The application covered by spatial data mining are geomarketing, environmental studies, risk analysis, and so on. The aim of clustering is to automatically find groups of instances that are similar to each other. For example, in classroom there is cluster of students with similarities in their birth date is in same month. Using GIS, the user can query spatial data and perform simple analytical tasks using programs or queries. However, GIS are not designed to perform complex data analysis or knowledge discovery [12]. There are many distance metrics [13] [14] [15] are available like Euclidean, Manhattan (city block), Chebychev, Minkowski etc.

1.1 K-Means Algorithm

The **K-means** [6] [7] algorithm involves randomly selecting **K** initial centroids or mean where **K** is a user **defined** number of desired clusters. For each of the object the distance is calculated between center points and data points and with minimum distance data, cluster is generated. This data points are far from another cluster or group. This computation is stop when no center points do not move any more. K means algorithm work as follows:

1. Initialization by setting initial centroids with a predefined k.
2. Cluster or group the data points in given k clusters.
3. Assign data or objects to nearest cluster center as per distance function.
4. When all objects are assigned recalculate or update the position of k centroids.
5. Repeat step 3 and 4 until the centroids no longer move.

Diagrammatic representation of k means algorithm as follows [17]:

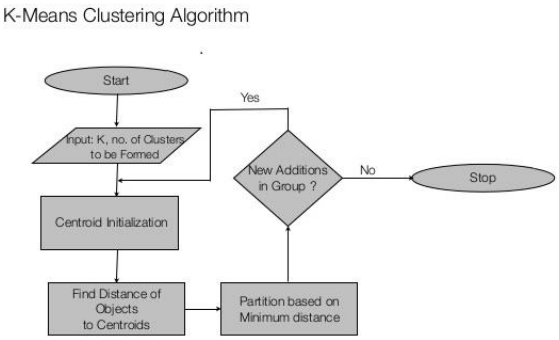

Fig-1: Flowchart of k means algorithm

1.2 DISTANCE METRICS

Distance metrics [13] [14] [15] play very important role to measure similarities between objects. Distance metrics are used to find how far the data points to the center points and by this distance the algorithm can compute which data objects in one cluster while another is in different cluster. Many different distance metrics are available like Euclidean, Manhattan (City Block) [16], Chebychev etc. In general, squared Euclidean distance metrics is used in k-means algorithm but here we will do the experiments of different distance metrics in k means algorithm using spatial database. The overview of different distance metrics are as follows:

1) Euclidean Distance :

The Euclidean distance [13] 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.

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

2) Manhattan(City Block)

Manhattan distance [16] is also named as city block distance because it is a distance the car would drive in a city put out in square blocks like Manhattan.

The formula for calculate Manhattan distance is as follows:

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

Manhattan distance is also known as L1 distance. The distance between two points is the absolute difference between the points. Absolute value distance gives more robust result whereas Euclidean influenced by unusual values.

3) Chebychev Distance

Chebychev distance (Tchebychev) [15] is the distance between two vectors is the greatest of their differences along any coordinate dimension. It is named after Pafnuty Chebychev.

The formula for calculate Chebychev distance is as follows:

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

Chebyshev distance is also known as maximum value difference. It is also known as chessboard distance as it can be illustrated on real number plane as the number of moves needed by chess king to travel from one point to another point.

4) Minkowski Distance:

Minkowski distance [14] is defined as generalization of both Euclidean and Manhattan distance metric. The formula for calculate Minkowski distance is defined as follows:

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

Different names for the Minkowski distance or Minkowski metric arise form the order:

- $\lambda = 1$ is the Manhattan distance. Synonyms are L_1-Norm, Taxicab or City-Block distance. For two vectors of ranked ordinal variables the Mahattan distance is sometimes called Footruler distance.
- $\lambda = 2$ is the Euclidean distance. Synonyms are L_2-Norm or Ruler distance. For two vectors of ranked ordinal variables the Euclidean distance is sometimes called Spearman distance.
- $\lambda = \infty$ is the Chebyshev distance. Synonym are L_{max}-Norm or Chessboard distance.

The Minkowski distance is often used when variables are measured on ratio scales with an absolute zero value

1.3 SPATIAL DATA MINING

Spatial data mining [8] [9] is the process of discovering useful, interesting, non trivial pattern from large spatial database. Spatial data mining (SDM) consists of extracting knowledge, spatial relationships and any other properties which are not explicitly stored in the database. SDM is used to find implicit regularities, relations between spatial data and/or non-spatial data [11]. Spatial data mining is the discovery of interesting the relationship and characteristics that may exist implicitly in spatial databases. There are many spatial data mining application like risk analysis, earthquake analysis, flood analysis, in medical research, in army center etc. if in any place earthquake is placed then spatial data mining is very useful to measure damage causes because of earthquake. In astronomy also spatial data mining is very useful. So, spatial data mining is data mining application which is related to geometric or spatial information and it is used to mine the spatial data.

- Spatial Data Mining (SDM) [8] [9] is a process of discovering trends or patterns from large spatial databases that hold geographical data. Objects in

space such as roads, rivers, forests, deserts, buildings, cities etc., are stored in spatial database.

- Geographic data collection devices connected to global positioning system (GIS) receivers allow field researchers to collect unprecedented amounts of data. Position alert devices such as cell phones, in-vehicle navigation systems and wireless Internet clients allow tracking of single movement behaviour in space and time [11].

2. PROPOSED ALGORITHM

Methodology used in simple k-means

The formal algorithm is:

a) Select cluster center randomly.

b) Calculate the distance between data points and center points using Euclidean distance formula.

c) Data points are assigned to cluster with minimum distance.

d) Repeat step b and c until centroid do not change.

In K-Means algorithm [6] [7], we calculate the distance between each point of the dataset to every centroid initialized. Based on the values found, points are assigned to the centroid with minimum distance.
The main goal of proposed algorithm is to analyze how different distance metrics perform in k means clustering algorithm.
In my proposed algorithm we have modified distance function formula in basic k means algorithm and analyze the result. I merge the two equation of Euclidean distance and Chebychev distance and apply that distance formula in basic k means algorithm.

1) Euclidean distance formula:

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

2) Chebychev distance formula:

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

Proposed Equation as follows:

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

So, the algorithm as follows:
1. Select 'c' cluster centers randomly.
2. Calculate the distance between each data point and cluster centers using the new distance metric as follows:

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

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:

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

where, 'ci' denotes the number of data points in ith cluster.
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.

Table -1: Comparison table of old algorithm and new algorithm

Old Algorithm	New Algorithm
• In old algorithm only Euclidean distance measure is used to find distance between center points and data points.	• In new algorithm the new distance measure is used. • In new algorithm there are merging of two distance measure: Euclidean distance and Chebychev distance. And develop new distance measure. • This new distance measure is used to find distance between data points and center points.

3. EXPERIMENTS AND RESULT ANALYSIS

Software requirements:
To do the experiments I used NetBeans IDE 8.2. Some Applications of NetBeans are listed here:

1) best support for latest java technologies
2) fast and smart code editing
3) easy and efficient project management
4) rapid user interface development
5) write bug free code

For more information about NetBeans please visit the website www.netbeans.org .

Fig-2: Screenshot of NetBeans IDE 8.2

3.1 SPATIAL DATA

Spatial data [12] also known as geographical data. It is the data that identifies geographical location of features and boundaries on earth, ocean etc. spatial data are data that have spatial component it means that data are connected to a place in the earth. Any data that is dependent on its location is called spatial data even human's weight is also called spatial data because it is dependent on its location. Non spatial data is that information which is independent of all geometric consideration. For example human's age, height etc. are non-spatial data.

Examples of non-spatial data

- Names, phone numbers, email addresses of people

Examples of Spatial data

- Census Data
- NASA satellites imagery - terabytes of data per day
- Weather and Climate Data
- Rivers, Farms, ecological impact
- Medical Imaging

In my experiments I used data of rainfall_area-wt_India_1901-2015. I downloaded this data from the website of www.data.gov.in. This is the India's most reputable website for geographic dataset. Data are freely available on this website. Another website for geographic dataset is www.bhuvan.nrsc.gov.in , www.diva-gis.org etc. for download the data from the bhuvan site you have to first register on the website and then you can login and download the data.

3.2 EXPERIMENTS

We do all experiments in NetBeans IDE 8.2 and the screenshot of output of the algorithm as follows:

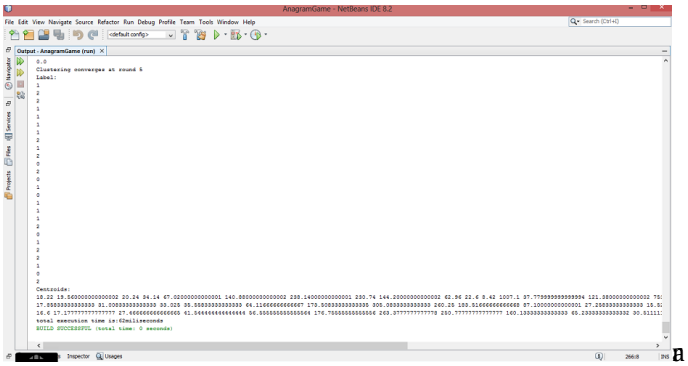

Fig-3: Screenshot of NetBeans IDE, where k means performed with Euclidean distance metric

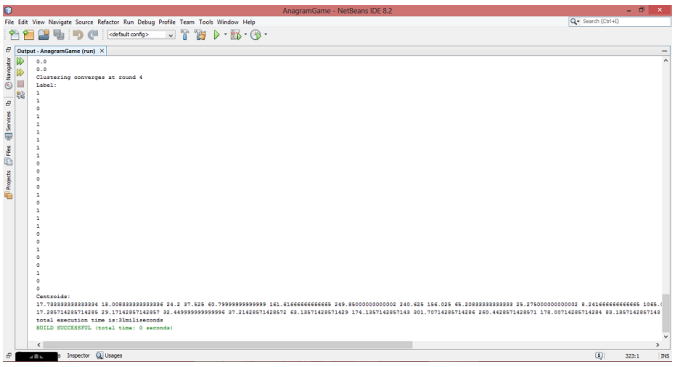

Fig-4: Screenshot of NetBeans IDE, where k means performed with Chebychev distance metric

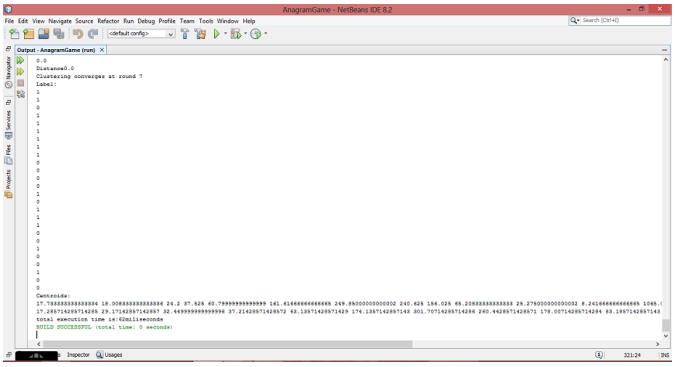

Fig-5: Screenshot of NetBeans IDE, where k means performed with Minkowski distance metric with p=4

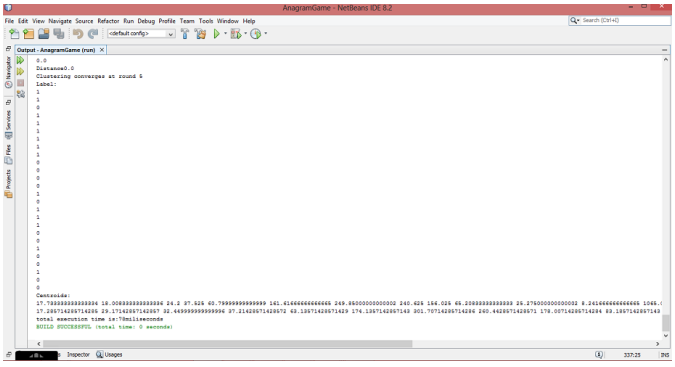

Fig-6: Screenshot of NetBeans IDE, where k means performed with Minkowski distance metric with p=5

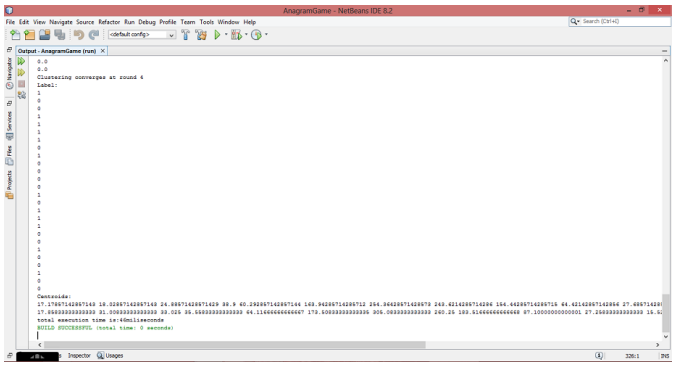

Fig-7: Screenshot of NetBeans IDE, where k means performed with Manhattan distance metric

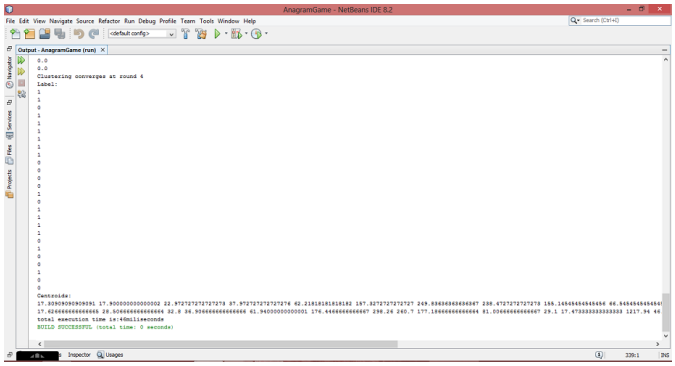

Fig-8: Screenshot of NetBeans IDE, where k means performed with new distance metric

3.3 COMPARISON OF DIFFERENT DISTANCE METRICS IN BASIS OF TIME

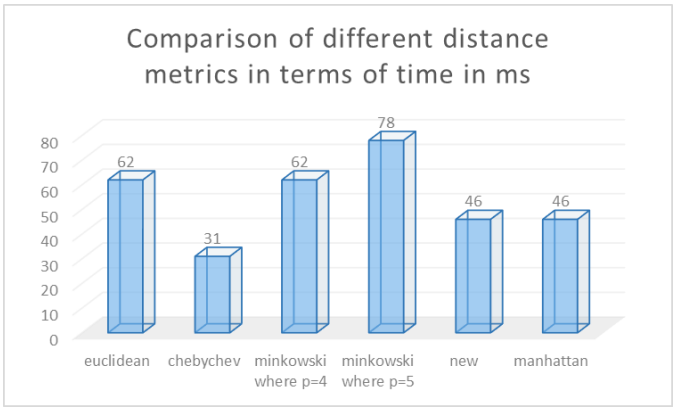

Chart -1: Comparison chart

4. CONCLUSIONS

From whole my experiments and literature review experience I conclude that, in general in all k means algorithm mostly Euclidean distance metrics is used. But we try different distance metrics in k means algorithm using spatial database and we have conclude that the new distance metrics that is described in proposed algorithm is better work than Euclidean distance and Minkowski distance with two different values in basis of time. While Chebychev distance metrics and Manhattan distance metrics take low time than new distance metrics.

REFERENCES

[1]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

[2]. J. Han , M. Kamber, Data Mining, Morgan Kaufmann Publishers, 2001.

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

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

[5] Rui Xu, Donald C. Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions on neural Networks, vol. 16, pp. 645-678, May 2005

[6]Aakansha Chaudhry,"Survey of K-Means and its variants", International Journal of Innovative Research in computer and communication Engineering,Vol 4, Issue 1, 2016

[7] Kapil Joshi, Himanshu Gupta, Prashant Chaudhry,Punit Sharma. Survey on different Enhanced k-means Clustering Algorithm. International journal of Engineering trends and technology (IJETT).

[8] Aakunuri Manjula, Dr.G.Narsimha. A REVIEW ON SPATIAL DATA MINING METHODS AND APPLICATIONS, International Journal of Computer Engineering and Applications, Volume VII, Issue I, Part II, July 14

[9] M.Hemalatha.M; Naga Saranya.N. A Recent Survey on Knowledge Discovery in Spatial Data Mining, IJCI International Journal of Computer Science, Vol 8, Issue 3, No.2, may, 2011

[10] Harvey J. Mailer, Jiawei Han, "GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY: AN OVERVIEW", GKD chapter 1 v8.

[11] 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

[12] Gangireddy Ravikumar, Mallireddy Sivareddy. An Effective analysis of spatial data mining methods using range queries. Journal of global research in computer science

[13] 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

[14] 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
[15] 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
[16]ple.revoledu.com/kardi/tutorial/Similiarity/CityBlockDistance.html
[17] www.slideshare.net/VaradMeru/kmeans-its-variants-and-its-applications