International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Analyzing the Best Distance Measure for K-Means Algorithm

B Jayanthy[1], Dr.C.Senthamarai[2]

Research Scholar, Department of Computer Applications, Government Arts College (Autonomous), Salem, Tamilnadu, India [1]

Assistant Professor, Department of Computer Applications, Government Arts College (Autonomous), Salem, Tamilnadu, India [2]

ABSTRACT: K-Means is a very simplest algorithm to efficiently handle the large dataset using clustering technique. In K-Means, Distance measure decides the performance of the algorithm to partition the dataset. In real word, many distance measure techniques are available. Euclidean distance is the most common distance measure technique used in K-Means algorithm. But performance of K-Means algorithm not only depends on distance measure, but also depends on the accurate section of distance measure that applied on particular data and also on researcher's question. In this paper, different distance measures are applied on different dataset to find out the appropriate distance measures for particular kind of dataset.

KEYWORD: Partition Method, K-Means, Clustering, Distance Measures.

I. INTRODUCTION

Clustering algorithm[1] is one of the method used to classify the given dataset into groups based on similarity exists between data by using distance measure technique. Clustering algorithm is classified into two types.

1. Partitioned Method
2. Hierarchical Method.

In hierarchical method[2-4] the given dataset is decomposed into different hierarchy of cluster. Two types of hierarchical method,

a. Agglomerative method
b. Divisive method, are available.

Agglomerative method[2-4] merges lower hierarchy cluster into higher and higher and finally results one large cluster. Divisive method[2-4] divides the largest cluster and finally results different hierarchy of cluster s. Partitioned method[1-4] partition the given dataset into different clusters based on distance measure technique. The most famous partition method is K-Means algorithm. In this paper, we discussed about different distance measurement techniques that applied on K-Means. Here we implement K-Means using different distance measures that applied on Breast Cancer, Heart Disease, Thyroid Disease, Forest Fires and Wholesale customers dataset. Based on the experiment results, we drawn out the conclusion about the K-Means performance with respect to different distance measurement techniques.

II. K-MEANS ALGORITHM

K-Means[2] is one of the most popular algorithm. K-Means partitioned the data into K number of clusters where the data object inside the same cluster has high similarity and data object of different cluster has high difference. The main concept of K-Means is depends on selection of number of clusters K and proper selection of initial seeds as centroids[2] for each K cluster. Centroid plays an important role in K-Means to cluster the datasets into different cluster. To identify which dataset belongs to which cluster distance measure techniques are used to find out distance of each object from each cluster centroid. And finally we place data object to anyone of the cluster using distance. Mostly, we place the nearest distance value dataset to cluster. At second iteration, we recalculate the new centroid values for each cluster and place object based on new centroid values. If any changes occurs in placement of data object into cluster when compared to first iteration, then we will go for next iteration. Finally we stop the iteration when there is no more changes in the placement of data objects in clusters.

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

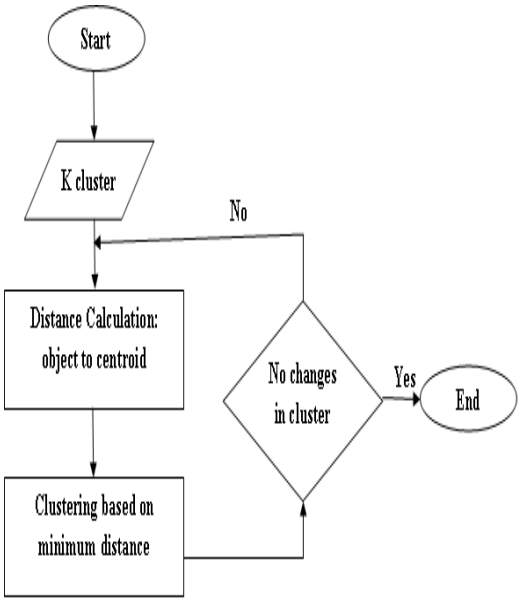
Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Algorithm:

1. Select K
2. Select K initial seed and put into K cluster
3. Assign initial seed as centroid value for K cluster
4. Calculate distance of each data object from K cluster using distance measure technique
5. Assign data object into K cluster based on step 4
6. Find out new centroid value for each cluster
7. Repeat Step 4
8. If no more changes stop else goto Step 4

Figure(1)-K-Means Algorithm



III. DIFFERENT DISTANCE MEASURES

3.1 Euclidean Distance:

Euclidean distance[5] is a basic distance measure for K-Means. It is calculated the square root of the sum of differences between each point. Another name for Euclidean distance is L2 Norm of a vector. When the input variable are similar in type or when to find out distance between two points Euclidean distance method is used.

$$D(x.\ y) = (\sum(x_i - y_{i)})^{2})^{1/2}$$

3.2 Manhattan distance:

Also called as City block distance[5]. It is a L1 norm of the difference vector. It is the distance travel along coordinates only. It is calculated as the sum of absolute distances between two points.

$$D(x,\ y) = \sum| x_i - y_i|$$

3.3 Pearson Correlation Distance:

Pearson Correlation Distance[5] measures the degree of a linear relationship between two profiles.

$$D(x,\ y) = \frac{-\ \sum\ (x_i - x)\ \ (y_i - y)}{(\sum(x_i - x)^2 \sum (y_i - y)^2)^{1/2}}$$

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

3.4 Jaccard Distance:

Jaccard Distance[5] is used to find the similarity of two sets. Jaccard similarity is defined as the intersection of sets divided by their union.

$$J(x, y) = \frac{|x \cup y|}{|x \cap y|}$$

3.5 Chebychev distance

Also called the Lα norm of the difference vector[5]. It is the maximum of differences between x and y in any dimension.

$$D(x, y) = \text{Max } |x_i - y_i|$$

IV. EXPERIMENTS

For doing our experiments we have selected the following datasets.

1. Breast Cancer
2. Heart Disease
3. Thyroid Disease
4. Forest Fires
5. Wholesale customer

These dataset are gathered from UCI Repository data [6]. Experiments are done in MATLAB.

Table 1: Breast Cancer Dataset

Dataset characteristics	:	Mutivariant
Attribute characteristics	:	Real
Associated Tasks	:	Classification
Number of instances	:	286
Number of attribtes	:	9

Table 2: Heart Disease Dataset

Dataset characteristics	:	Mutivariant
Attribute characteristics	:	Categorical
Associated Tasks	:	Classification
Number of instances	:	160
Number of attribtes	:	5

Table 3: Thyroid Disease Dataset

Dataset characteristics	:	Mutivariant
Attribute characteristics	:	Real, Categorical
Associated Tasks	:	Classification
Number of instances	:	7200
Number of attributes	:	21

|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Table 4: Forest Fires Dataset

Dataset characteristics	:	Mutivariant
Attribute characteristics	:	Real
Associated Tasks	:	Regression
Number of instances	:	517
Number of attribtes	:	13

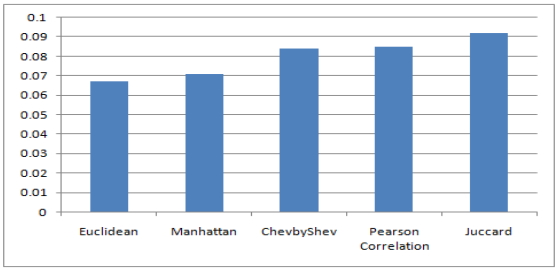
Table 5: Wholesale Customer Dataset

Dataset characteristics	:	Mutivariant
Attribute characteristics	:	Real
Associated Tasks	:	Classification, Clustering
Number of instances	:	440
Number of attributes	:	8

Time with respect to different distance measures for dataset

Dataset	Euclidean	Manhattan	ChevbyShev	Pearson Correlation	Juccard
Breast Cancer	0.067	0.071	0.084	0.085	0.092
Heart Disease	0.071	0.078	0.087	0.089	0.076
Thyroid Disease	0.064	0.069	0.076	0.079	0.089
Forest Fires	0.081	0.082	0.092	0.089	0.096
Wholesale customers	0.072	0.087	0.074	0.086	0.089

Figure 2. Comparison of different distance measure for Breast Cancer dataset in terms of computation time.



|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

Figure 3. Comparison of different distance measure for Heart Disease dataset in terms of computation time.

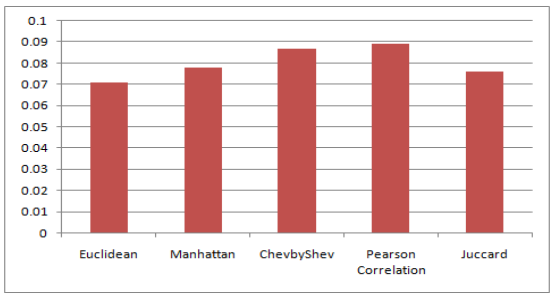


Figure 4. Comparison of different distance measure for Thyroid disease dataset in terms of computation time.

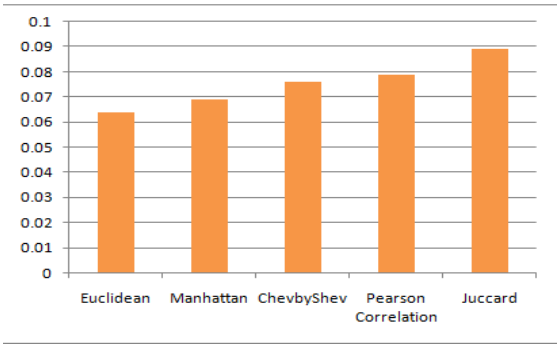


Figure 5. Comparison of different distance measure for Forest Fires dataset in terms of computation time.

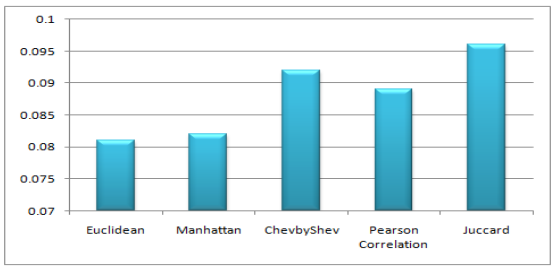
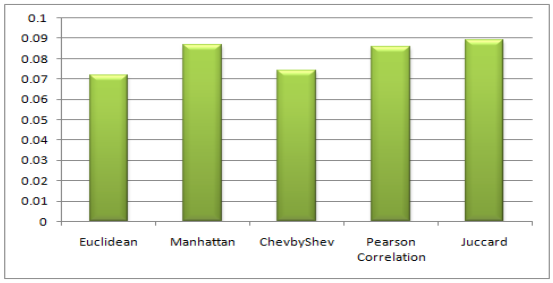


Figure 6. Comparison of different distance measure for Wholesale customers dataset in terms of computation time.



|| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512| || Volume 9, Special Issue 2, October 2020 ||

3rd International Conference on Emerging Trends in Science, Technology and Mathematics [ICETSTM 2020]

18th September 2020

Organized by

Department of Computer Science, Parvathy's Arts and Science College, Dindigul, Tamilnadu, India

V. CONCLUSION

Based on the experiments results, we have found that Euclidean Distance is best for health oriented data and climate based data and customer dataset. These experiments only aimed on five different datasets to observe the K-Means performance with different number of attributes. In our future experiments we will consider some more distance measures that applied on different kind of datasets to find out best alternative to Euclidean distance method.

REFERENCES

[1] Ziad Obermeyer, Ezekiel J.Emanuel, "Predicting the Future- Big Data, Machine Learning, and Clinical Medicine", The New England Journal of Medicine, Vol.375, No.13(2016), 1216-1219.
[2] Yong Gyu Jung, Min Soo Kang, Jun Heo, "Clustering performance comparison using K-means and expectation maximization algorithms", Biotechnology & Biotechnological Equipment, Vol. 28, No. S1(2014), 45-48.
[3] Min-Soo Kang, Yong-Gyu Jung, Du-Hwan Jang, "A study on the search of Optimal Aquaculture farm condition based on Machine Learning", The journal of The Institute of Internet, Broadcasting and Communication(IIBC) Vol. 17, No.2 (2017), 135-140.
[4] Wu, Xindong and Zhu, Xingquan and Wu, Gong-Qing and Ding, Wei," Data mining with big data", IEEE transactions on knowledge and data engineering, 26, 1, 97-107, IEEE, 2014
[5] www.datanovia.com
[6] www.achieve.ics.ci.edu

BIOGRAPHY

Mrs.B.Jayanthy, at present doing research in datamining. She worked as an Assistant Professor in the Department of Computer Science. She did Bachelors of Computer Science from Government Arts College for Women, Salem-8(Periyar University) on 2004. She did Master of Computer Science and M.Phil Computer Science from Periyar University. She did Bachelor of Education(B.Ed) from TamilNadu Teachers Education University. She published papers in National journals. She has 11 years of experience in teaching field. Her area of specialization is in Data mining, Machine Learning, Java. She guided M.Sc,MCA, M.Phil students for doing projects and research. She developed websites for Government colleges also.

Dr.C.Senthamarai, at present working as an Assistant Professor in the Department of Computer Application, Government Arts College(Autonomous), Salem-7. Before that she worked as an Assistant Professor in the Department of MCA, K.S.Rangasamy College of Technology, Tiruchengode. She did Bachelors of Physics and Master of Computer Application. She completed Doctor of Philosophy in Computer Science. She has more than 25 years of experience in teaching field and research field. She published many papers in International journals and National journals. Her research interest includes Grid computing, Cloud computing, Fuzzy Logic, Genetic Algorithm, Machine Learning, Computer Network and Data mining. She guides many research scholars too.