A REVIEW ON K-mean ALGORITHM AND IT'S DIFFERENT DISTANCE MATRICS

Rashmi Sindhu[1], Rainu Nandal[2], Priyanka Dhamija[3], Harkesh Sehrawat[4], Kamaldeep

Computer Science and Engineering,

University Institute of Engineering and Technology,

Maharshi Dayanand University, Rohtak, Haryana, India

1. rashmisindhu88@gmail.com
2. rainu_nandal@yahoo.com
3. pd9493@gmail.com
4. sehrawat_harkesh@yahoo.com

***Abstract* – Data mining is a process of extracting desired and useful information from the pool of data. Clustering in data mining is the grouping of data points with some common similarity. Clustering is an important aspect of data mining. It simply clusters the data sets into given no. of clusters. Various no. of methods have been used for the data clustering among which K- means is the most widely used clustering algorithm. In this paper we have briefed in the form of a review work done by different researchers using K-means clustering algorithm. We have also analysed different distance metrics used by them for distance evaluation.**

Keywords: **K-means clustering, clusters, data points, data mining, Euclidian, Manhatten, Minkowski.**

I. INTRODUCTION

Data mining is the process of extracting information from a data set and transforms it into an understandable structure for future use. Data stored in a Data Warehouse is of a wide range that may or may not contain the relevant information that the user desire to use. Searching some information from the wider data space involves analysing of large data which probably may result in degrading of its efficiency. Thus to make the best use of the data of the data warehouse ,a tool called the data mining is introduced so that it can provide the required information to its user from the data pool.

Now a day, data mining is used in various fields just because of its efficiency in searching the required information needed by the user in a short period of time and with more accuracy. Data mining has broadened its area of implementation that is why it has become a topic of concern for the researchers.

In data mining there are six classes of tasks that are common: -

Detection of anomaly: - Identification of data records that are unusual or errors in data that require further investigation.

Learning of associative rule: - It searches for the relationships between the variables.

Clustering: - It is the process of discovering groups and different structures in the data.

Classification of data sets: - Process of generalizing the structure that is already known so that it can be applied to new data.

Data regression: - Process of attempting a function that will help in modeling the data with least error.

Summarization: - It helps in providing more compact representation of data set.

Role of Clustering in data mining is that it provides the user with the benefit to understand that in a data set how we can do natural grouping of a structure. Use of Clustering can be done either as a tool that can stand alone to go in detail in the distribution of data or as a pre-processing step for any other algorithm.

Clustering is the most essential part of Data mining. Clustering is a technique in which we group different data based on their similarities and dissimilarities from the other. Data point in one data set will posses similar properties as compare to the other. Clusters can differ from each other in terms of shape size and density. The similarity between the objects is calculated by the use of a similarity function. This is mainly useful for organising documents, to improve recovery and support browsing.

There are several clustering techniques that have been used to carry out the clustering process.

Centriod –based technique: - In this type of grouping method a vector of values is used for referring every cluster. Every object becomes part of that cluster whose value difference is minimal, as compared with some other cluster. Drawback with this method is that no. of clusters should be predefined.

Distributed –based: - This type of methodology combines those objects whose value belongs to the same distribution. This type of technique needs a model which is complex but well defined to interact in a better way with real data.

Connectivity-based: - Every object in data set is related to its neighbours, which will depend on the degree of the relationship of the distance between data set and its neighbour. Clusters are created with the objects that are nearby, and can be described as a maximum distance limit.

Density- based: - In this technique clusters are created on the basis of the high density of members of a data set, in a determined location.

Among all the data clustering algorithms K-means is so popular because it is the simplest and the widely used algorithm for clustering of the data sets. It is so because it uses unsupervised learning technique to solve the well known issues of the clustering. It is also well suited for the large data sets.

K-mean algorithm is the most widely used algorithms for cluster analysis of large no. of data sets. K-mean algorithm is a type of Centriod-based clustering technique. K-mean clustering algorithm is efficient in generating clusters for many applications. In K-mean algorithm centroids are used to represent a cluster which is basically the mean of the points of the cluster.

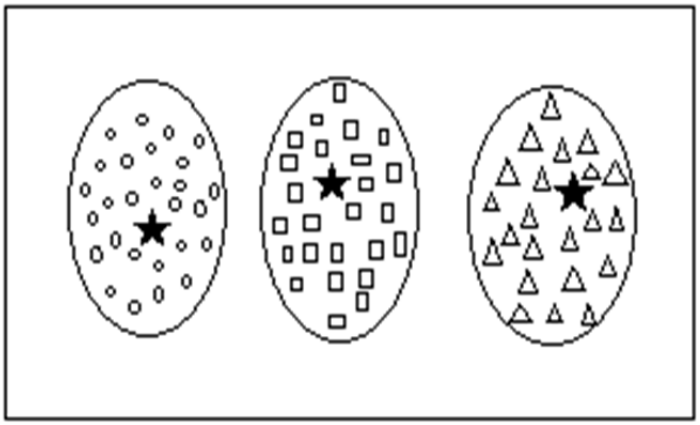

Fig 1. Example of Cluster formation.

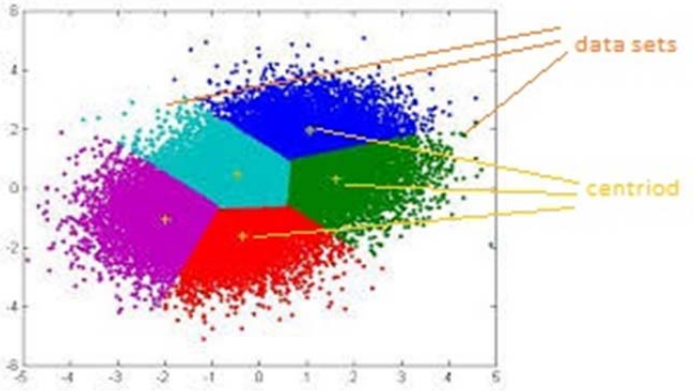

Fig 2. Example of Data sets with different Centriod.

These figures show the working of K-mean algorithm on the given data set. In this algorithm the entire process is carried out in 2 phases, in the first phase the Centriod is chosen randomly and assigning data objects to the nearest cluster and then recomputed the cluster Centriod, and in the second phase the above process is continued until the clusters become stable.

II. LITRATURE REVIEW

Narender Kumar et al in 2013.Explains K-mean algorithm by taking example of LIC policy plan and implemented it using K-mean(1). Finds the group of people who belongs to same criteria. He considered first three people in the list as the three seeds for finding out the cluster using. After that he compute the distance using the attributes and uses the sum of differences. Based upon the value of these distances each person is allocated to the nearest cluster. Now compare the mean of the cluster to recompute the distance, when the cluster shows that the clusters have not changed specify them as the final cluster.

Miss Vrindakhairnar and Miss Sonal Patil in 2016 describes the algorithm by randomly selecting initial cluster Centriod from data item (2).They divided K-mean into two phases, first phase is to randomly select initial Centriod with better accuracy and assign each data point to which it is the most similar cluster, the second phase is to calculate the new mean for each cluster. The complexity of K-mean algorithm is O (nkt) where *n* is the no. of objects is no. of iterations, and *k* is the no of clusters. Also explains the drawbacks of K-mean algorithm and has made the study of various techniques of modified K-mean algorithm.

Surender Kumar and Nancy Uses various normalization techniques for data mining (3). These are: -

- Min-max normalization: - This type of normalization is used for performing a transformation that is linear on the original data. Let us assume that we have a minimum and maximum values of an attribute that is min(A) and max(A), A, then min-max normalization maps a value, (vi) in this range [new min(A), new max(A)] by complexity

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- Score Standard Normalization: - Normalization criterion is based on the values of an attribute, A, which depends on the mean and standard deviation of A. A value, v(i), of A is normalized to v(i) by

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}$$

- Decimal Scaling: - In this Normalization technique the decimal point of the values for an attributes A are moved, then no. of decimal points moved depends on the maximum positive values of A. A value v(i), of A is normalized to v(i)' by calculating the given formula where *j* is the smallest integer such that max |v(i)|<1

$$v_i' = \frac{v_i}{10^j}$$

Jyoti yadav and Monika Sharma in 2013 has introduced 3 main methods to measure the distance between cluster Centriod and data items (4). These are: -

I. Euclidean method: -

$$d(i,j) = \sqrt{\left(\sum_{i=0}^{n}(x_i - y_i)^2\right)}$$

II. Manhatten method: -

$$d(i, j)= [|X_{i1} –X_{j1}| + |X_{i2} – X_{j2}| +……+ |X_{in}-X_{jn}|]$$

III. Minkowski method: -

d(i, j) = [(|Xi1-Xj1|p+|Xi2 – Xj2|p+……+|Xin- Xjn|p)1/p]

Describes K-mean algorithm in detail and also specify its drawbacks. Introduces a new algorithm which is just the advancement of general K-mean algorithm. This algorithm does not require to fix the value of K initially.

Akanksha Choudhary in 2016 Introduced variants of K-mean algorithm(5). States that to generate accurate initial Centriod points rather than picking them randomly use dissimilarity matrix to create Huffman tree. Output of Huffman tree is taken as the initial Centriod. Also gives an option of selecting Centriod acc to which pick up the initial Centriod randomly and the remaining Centriod are selected as the data point that has the greatest minimum distance to the previously selected Centriod. Various distance matrices for measuring the distance between the cluster Centriod and data points. Uses three different distance matrices namely Euclidean, Manhatten, Minkowski and have found that Euclidean distance metric gives the best result.

Archna Singh et al in2013 elaborated the distance matrices in detail (6). Implement the K-mean algorithm using these different distance matrices.

a) Euclidean distance: -

$$d(i,j) = \sqrt{\left(\sum_{i=0}^{n}(x_i - y_i)^2\right)} \quad (4)$$

New cluster centre is calculated using: -

$$Vi = (1/ci) \sum_{1}^{ci} xi$$

b) Manhatten distance: -

$$d(i, j) = |X_{i1} - X_{j1}| + |X_{i2} - X_{j2}| + \ldots\ldots + |X_{in} - X_{jn}| \quad (4)$$

New cluster centre is calculated using:-

$$Vi = (1/ci) \sum_{1}^{ci} xi$$

c) Minkowski distance: -

$$d(i, j) = [(|Xi1 - Xj1|p + |Xi2 - Xj2|p + \ldots\ldots + |Xin - Xjn|p)1/p] \quad (4)$$

New cluster centre is calculated using: -

$$Vi = (1/ci) \sum_{1}^{ci} xi$$

III. GENERAL APPROACH

The K-means algorithm is one among the simplest and popular classical methods for clustering which is quite easy to implement. These types of classical methods can be used if and only if the data related to all the objects is at one location i.e. in the main memory. K-means clustering algorithm was first time used by James MacQueen in 1967. Stuart Lloyed first proposed the basic original algorithm as a technique for signal processing, but it was not published until 1982.
The name K-means is given to this method because each cluster, say as K clusters is represented by the value of mean of all the data objects within it, the mean value is called as centroid. Another name given to this method is the Centriod method. It is called so because in every step of assigning, the point of centroid for each cluster is considered that it is known and each point which is still remaining is assigned to that cluster whose Centriod is closest to it. Once we have allocated all the data points, then we will again calculate the centroids of the cluster simply by computing the mean and the process of assigning points to those clusters that are nearest is repeated till they have no change in the positioning of the clusters.

The K-means method uses the Euclidean distance metric for calculating the distance between the centroid and the data point, that work well with the clusters that are compact in size. But instead of using the Euclidean distance, the Manhatten distance metric can also be used and this method is called the median-K method. The median- K method tries to remove the drawback of K-means method as it can be less sensitive to outliers.

The algorithm for K-means can be explained as: -

1. Let K be the number of clusters.
2. Choose K points which will be taken as centroids for the K clusters. These points can be chosen on the random basis. This criterion is valid unless the user has some detailed view into the data.
3. Calculate the distance by using Euclidian formula/ Manhatten distance formula for each data point in the set from each of the chosen centroids.

4. Each object is assigned to its nearest cluster based upon the value of the calculated distances in the above step.
5. Calculate the cluster centroids again by calculating the means of the attribute values of all objects in each cluster.
6. Now, analyse that whether the membership of the cluster is changed or unchanged, if it is unchanged then this is the last step otherwise go to step 3.

To illustrate how the k-means algorithm works, we had considered the data set shown below which is consisting of the scores for two variables on each of seven particulars:

Theme	A	B
a	1.0	1.0
b	2.0	1.5
c	4.0	3.0
d	7.0	5.0
e	5.0	3.5
f	5.0	4.5
g	4.5	3.5

We will group this data set into two clusters. The very first step for finding an initial partition, let the two values of A & B individuals furthest apart by the use of Euclidean distance metric, define the means of initial cluster giving:

	particular	Value of Mean (centroids)
Group 1	a	(1.0, 1.0)
Group 2	d	(7.0, 5.0)

The individuals that are left yet are examined in a defined sequence and will be allocated to those clusters to which their Euclidean distance is minimum, from the mean of cluster. The mean vector is calculated every time a new particular is added. This will lead in the below series of steps:

	1 Cluster		2 Cluster	
Steps	Particular	Value of Mean (Centriod)	particular	Value of Mean (Centriod)
1	a	(1.0, 1.0)	d	(7.0, 5.0)
2	a, b	(1.5, 1.2)	d	(7.0, 5.0)
3	a, b, c	(2.3, 1.8)	d	(7.0, 5.0)
4	a, b, c	(2.3, 1.8)	d, e	(6.0, 4.2)
5	a, b, c	(2.3, 1.8)	d, e, f	(5.7, 4.3)
6	a, b, c	(2.3, 1.8)	d, e, f, g	(5.4, 4.1)

Now here we analyse that the partitioning done initially has changed, and the clusters are now having the below characteristics:

	Particular	Value of Mean (Centriod)
Cluster 1	a, b, c	(2.3, 1.8)
Cluster 2	d, e, f, g	(5.4, 4.1)

But still we are not sure about each individual that it has been assigned to the right cluster. Therefore, we will now compare distance of each individual to its own cluster mean and to that of the opposite cluster. And we conclude:

Particular	Distance to mean (Centriod) of 1 Cluster	Distance to mean (Centriod) of 2 Cluster
a	5.4	1.5
b	4.3	0.4
c	1.8	2.1
d	1.8	5.7
e	0.7	3.2
f	0.6	3.8
g	1.1	1.8

So, we analysed that only particular 3 is nearer to the calculated mean of the opposite cluster (Cluster 2) than its own (Cluster 1). We can also say that, distance of each individual to its own cluster mean should be smaller that the distance to the other cluster's mean. Thus, we will relocate the individual 3 to Cluster 2 which results in the new partition:

	Particular	Value of Mean (Centriod)
Cluster 1	a, b	(1.5, 1.3)
Cluster 2	c, d, e, f, g	(5.1, 3.9)

Eg: illustrate working of K-means algorithm (15)

The process of relocating will continue from this new partition until no more relocation occurs. In the above example each individual is now nearer to its own cluster mean than that of the other cluster thus the iteration stops, after leaving the above partitioning as the final solution for the cluster formation.

It is also possible that the k-means algorithm is not able find a final solution. In that case it is preferred to stop the algorithm after execution of prior chosen no of iterations.

IV. METHODOLOGY

Waikato environment for knowledge analysis (WEKA) is a collection of open source software for data mining developed at the University of Waikato in New Zealand. It was developed in the mid – 1990s and now is used all over the world.
WEKA provides a collection of software. It allows data to be uploaded from files, URLs, and databases. It provides a graphical interface for each of the following: - (17)

- Data pre- processing – allow data to be uploaded and transformed.
- Classification and regression – allow data to classified using supervised classification.
- Clustering – this is for unsupervised classification.
- Attribute selection
- Visualization – this provides the user to generate colour-coded scatter plots
- Experimentation.

Figure 3: - WEKA INTERFACE (16)

WEKA is freely available for download thus it has become one of the most widely used data mining systems. WEKA also became one of the most desired medium for data mining research and helped to enhance it by providing many powerful features to all.[7]

Features of WEKA due to which it has been used widely are: -

i. It is an open source, freely available and provides better user interaction.
ii. Many different algorithms can be implemented in it for data mining and machine learning.
iii. It can be used for algorithm implementation by those also who have lesser knowledge about WEKA.
iv. WEKA is quite flexible when used for scripting experiments.
v. It is platform independent data mining tool

5. PROPOSED METHOD

We will try to apply various distance evaluation metrics on the dummy data prepared for the mobile network companies and their users. We will try to form clusters using K-means algorithm for various mobile networks, and then calculate the distance of the user form the nearest cluster mean of that network.

After applying various distance metrics on the dummy data of the network we will also try to compare the different metrics so that we can reach to a decision that which distance evaluation metrics is more efficient for the data points, so that they can be assigned to the nearest cluster. By this clustering we will able to analyse that in a particular locality how many users use a particular network for mobile communication.

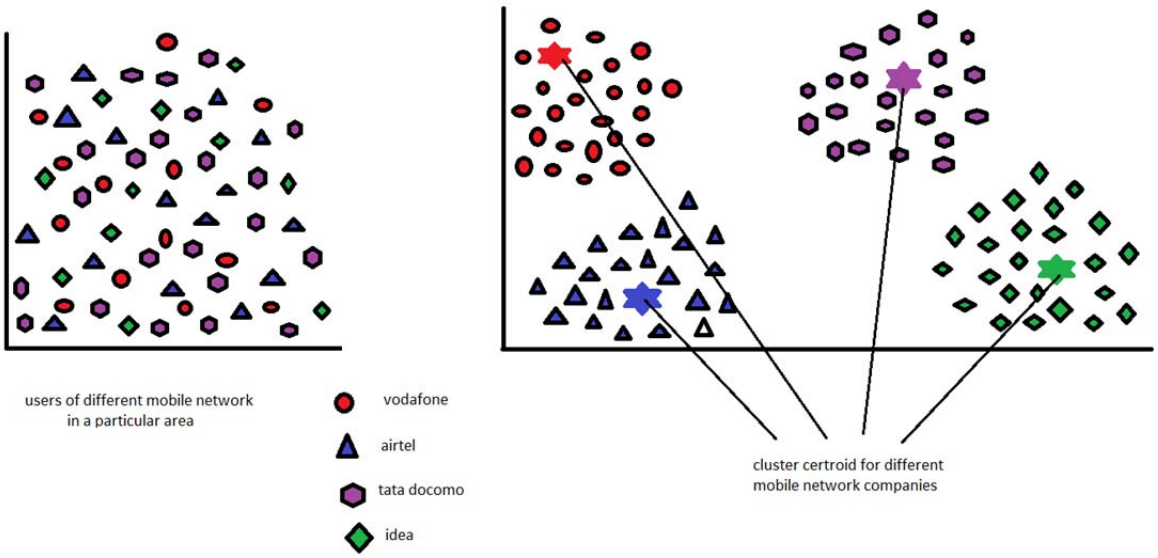

Fig 4. Example showing proposed method

6. CONCLUSION

In this paper we have analysed the work of different researchers made by them on the clustering of the data using the K-means algorithm. We also had survey on the different distance metrics used by them for the evaluation of the distance between the centre and the data points. At the end we conclude that a lot of improvement had been made in the working of the K-means algorithm and this field is still having a scope for more improvements and enhancements.

REFERENCES

[1] Narender Kumar, Vishal Verma, Vipin Saxena "*CLUSTER ANALYSIS IN DATA MINNING USING K-mean METHOD*", International Journal of Computer Applications, volume 76, No.- 12, August(2013).
[2] Miss. Vrindakhairnar and Miss. Sonal Patil "*EFFICIENT CLUSTERING OF DATA USING IMPROVED K-means ALGORITHM: A REVIEW*", Imperial Journal of Interdisciplinary research, volume 2, Issue 1, ISSN- 2454-1362, (2016).
[3] Sudesh Kumar and Nancy *"EFFICIENT K-mean CLUSTERING ALGORITHUM FOR LARGE DATASETS USING DATA MINNING STANDARD SCORE NORMALIZATION*", International Journal on Recent and Innovation trends and Computing and Communication, volume 2, issue 10, ISSN – 2321-8169, 3161-3166.
[4] Jyoti Yadav and Monika Sharma "*A REVIEW OF K-mean ALGORITHM*". International Journal of Engineering Trends and Technology, volume 4, Issue 7, July (2013)
[5] Akanksha Choudhary "*SURVEY ON K-mean AND ITS VARIENTS*", International Journal of Innovative Research in Computer and Communication Engineering, volume 4, Issue 1, January (2016).
[6] Archana Singh et al "*K-MEANS WITH THREE DIFFERENT DISTANCE MATRICS*", International Journals of Computer Applications (0975-8887), volume 67, No.10, April(2013).
[7] Ketan Sarvakar "*A COMPARATIVE STUDY OF CLUSTERING DATA MINNING: TECHNIQUES AND RESEARCH CHALLENGES*", International Journal of Latest Technology in Engineering, management & Applied Science, volume 3, Issue 9, September (2014).
[8] Brijesh Kumar Bhardwaj and Saurabh Pal "*DATA MINING: A PRIDICTION FOR PERFORMANCE IMPROVEMENT USING CLASSIFICATION*", International Journal of Computer Science and Information Security, volume 9, No. 4, April (2011).
[9] Mihika Shah and Sindhu Nair "*A SURVEY OF DATA MINNING CLUSTERING ALGORITHMS*", International Journal of Computer applications (0975-8887), volume 128, No.-1, October (2015).
[10] Rishikesh Suryawanshi and Shubha Puthran "*A NOVEL APPROACH FOR DATA CLUSTERING USING IMPROVED K-means ALGORITHM*". International Journal of Computer Application (0975-887), volume 142, No.-12, May (2016).
[11] Sachine Shinde and Bharat Tidke "*IMPROVED K-MEANS ALGORITHM FOR SEARCHING RESEARCH PAPER*". International Journal of Computer Science & Communication Networks, volume 4(6),197-202 197, ISSN:2249-5789
[12] Oyelade O.J, Oladipupo O.O and Obagbuwa I.C "*APPLICATION OF K-MEANS CLUSTERING ALGORITHM FOR PREDICTION OF STUDENTS ACEDEMIC PERFORMANCE*". International Journal of Computer Science and Information Security, volume 7,_o. 1, (2010)
[13] Sharddha Sukla and Naggana S. "*A REVIEW ON K-MEANS DATA CLUSTERING APPROACH*", International Journal of Information & Computation Technology, volume 4, No.-17 (2014).
[14] Gopi Gandhi and Rohit Srivastava "*REVIEW PAPER: A COMPARITIVE STUDY ON PARTITIONING TECHNIQUES OF CLUSTERING ALGORITHMS*", International journal of Computer Applications (0975-8887), volume 87, No. - 9, February (2014).
[15] Step- by- Step clustering K-means example http://mnemstudio.org/clustering-k-means-example-1.htm.
[16] "Introduction to Data Mining with Case Studies", 3rd edition, G.K Gupta.