Grenze International Journal of Engineering and Technology, July Issue

A Survey Paper on Data Analysis by using Model K-Means Clustering

Sanchita Mondal[1] and Bichitrananda Patra[2]
[1]Institute of Technical Education and Research, Bhubaneswar, India
Email: sanchitamondal923@gmail.com
[2]Dept. of Computer Science and Engineering, Siksha O Anusandhan (Deemed to be) University, Bhubaneswar, India,Email: bichitranandapatra@soa.ac.in

***Abstract*—Clustering is an unsupervised machine learning technique that serves a gargantuan task in passing on the data sets into precise clusters depending on various convergence or divergence characteristics. It has a brawny prospective in health-related data analysis for programmed disease prophecy. K-means is a clustering scheme that is extensively used in various areas of machine learning. The objective of our paper is to upgrade an existing clustering algorithm, K-Mean. The model will be trained using Micro-array datasets and the testing will be done using WEKA, this is an open source application. Apparently, from innumerable biological experiments and various community researches, there has been upsurge in the amount and complexity of Micro-array datasets. A storehouse that contains Micro-array gene manifestation data is called a Micro-array database.**

***Index Terms*— clustering, k-means clustering, machine learning, health care.**

I. INTRODUCTION

One of the most fundamental and rousing field of research that serves intention of discovering momentous information from vast data sets is Machine Learning Technique. There are many benefits offered by Machine Learning Technique in the health care domain that includes decreasing the cost of existing therapeutic treatment for patients, the uniqueness of numerous patients, identifying causes of a range of medical disorder, and establishing the probable mechanism of medication. More so, it assist curative care analyzer, to forge proficient health related stratagem, erecting antidote advocacy managements, and blooming well-being account of independent personage.

The fundamental viewpoint of the clustering arrangement is depended on how we choose the clusters or the groups. One possible way might be that we do not have to choose at all as because we can organize these varied possibilities by pertaining all of them and then integrating their clustering outputs. Although, there is no potential technique to instinctively use in the fusion algorithms from categorization (supervised domain) to clustering (unsupervised domain). The patterns that are unlabeled in clustering has different data partitioning that might result in different data labeling, causing difficult correspondence problems, especially when the numbers of resulting clusters are different.

Grenze ID: 01.GIJET.6.2.505
© Grenze Scientific Society, 2020

II. CLUSTERING

Clustering[5] is used in assembling bulky data into clusters or groups that helps us to visualize the internal structure of the data. For example, there is some online shopping site where we can find variety of stuffs from electronics, clothing, books, grocery items, cosmetic items, accessories. Here in figure 2 describes how it looks after clustering is done.

Figure 1. Before clustering is not implemented

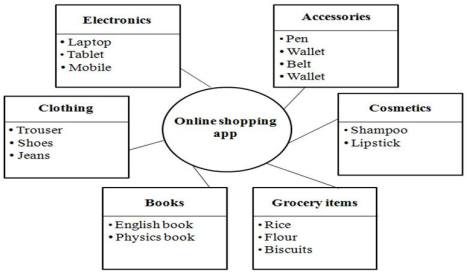

Figure 2. After clustering is done

A. Stages of Clustering

Various stages of clustering are mentioned as follows:

a. *Raw Data:* Raw data (which are not being processed yet) are collected from various sources on which we want to solicit various clustering algorithm.
b. *Clustering Algorithm:* A specific algorithm is selected according to our requirements and then that very algorithm is applied on the raw data that were being selected.
c. *Clusters:* After soliciting the selected clustering algorithm on the raw data, we acquire our clusters.

B.Types of clustering

a. Partitioning Method[1]:In the case of partitioning clustering method, the objects of the datasets are segregated into numerous subsets. Given some examples of the partitioning algorithms are K-means, PAM (*Partitioning AroundMedoids).*
b. Density-based Method[8]:In density-based clustering method, clusters are formed or the data spaces are partitioned by the density of the data point in a particular region.

Figure 3. Before and after applying partitioning clustering technique

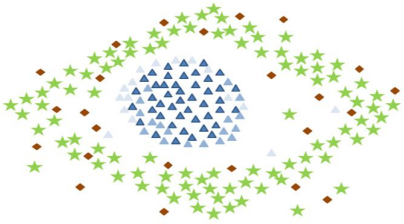

Figure 4.Density-based clustering technique

c. Hierarchical Method[9]: In the case of hierarchical clustering method, the objects of the datasets are segregated in the hierarchical fashion of clusters or groups. Agglomerative Hierarchical clustering algorithm (AGNES), Divisive Hierarchical clustering algorithm (DIANA) etc.
d. Grid-based method[1]: In grid-based clustering method, the object space is divided into fixed number of cells that forms the shape of a grid like structure. Clustering algorithm is STING (Statistical Information Grid).
e. Model-based clustering method[9]:Model-based clustering works on the concept of Probability Model which is a mathematical representation of any random occurrence of dataset. Each of the groups that would form will have different Probability Model.
f. Constraint-based method[1]: Constrained-based clustering method is a semi-supervised learning technique where amalgamation of small proportion of labeled data with a large proportion of unlabeled

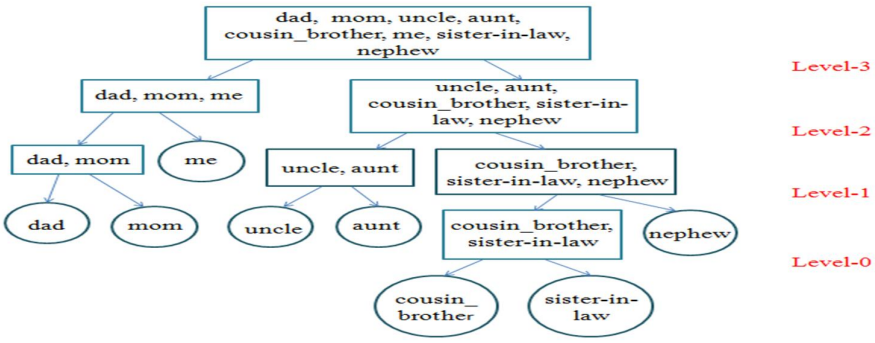

Figure 5.Hierarchical clustering method

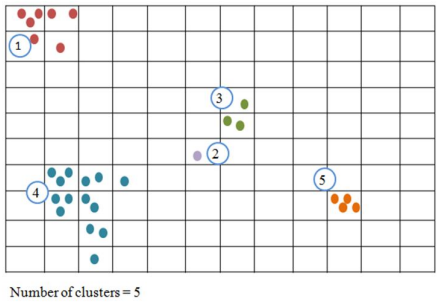

Figure 6. Grid-based clustering method

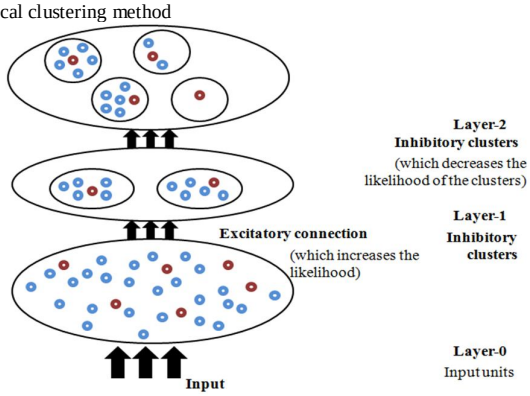

Figure 7. Model-based clustering technique

data occurs. *Constrained K-means* (COP-K-Means) algorithm is one of the common algorithms using this method.

C. Applications of clustering:

a. Sports: An athlete's performance can be modeled more effectively by using clustering.
b. Fraud Detection: By using clustering approach, we can detect credit card frauds faster.
c. Spam filter: E-mail content can be better classified using clustering and therefore spam detection would be easier.
d. Identifying fake news: Detection of fake news reduces misinformation risks and can be effectively carried out by applying cluster approach.
e. Stock Market prediction:NSE and BSE datasets for intraday trading can be facilitated by clustering algorithms.
f. Medical health: Various chronic diseases like breast cancer; brain tumor etc. It can be diagnosed with the help of clustering techniques.

III. K-MEANS CLUSTERING

K-Means[4] algorithm is a type of partition based clustering method which falls under the class of unsupervised learning techniques. It fragments a large of group data into K number of sub-groups of those data.[4] This algorithm is composed of two independent phases as discussed below:

a. *First phase:* In this phase, haphazardly K centroids or centers are chosen. The value of K should be permanent. It cannot be changed in the middle of the process.
b. *Second phase:* In this phase, every data point is assigned to the closest center or centroids.

To measure the distance between cluster centers or centroids and all data points, Euclidean distance is used. The Euclidean distance between any two points say point x and point y is the distance connecting point x and point y. The distance between x and y is same as the distance between y and x. The Euclidean distance between any random points x and y is given as below in equation (1):

$$\text{dist}(x, y) = \text{dist}(y, x) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (1)$$

A. Algorithm for K-Means

1. *Input:*Choose a database and select the value of K that is the number of clusters we want at the end.Let the database be D with n number of data objects.D = {d1, d2, d3, ….,dn}
2. *Output:* We will obtain an arrangement of K number of clusters.
3. *Algorithm*

(i) Randomly select the number of clusters, K.

(ii) Choose the centre or the centroids for K clusters. The initial values of the centres are selected arbitrarily.

(iii) Arrange all data objects to the closest cluster; this is determined with the help of Euclidean distance theory.

(iv) Again calculate the centre of the cluster. This is evaluated by taking the mean of the data objects present in each of the cluster individually. If there are n objects say x1, x2, x3, …., and then the mean is given in equation (2)

$$\text{Mean}(x) = \frac{\sum_{i=1}^{n}(xi)}{n} \quad (2)$$

(v) Repeat step (iii) and (iv) until convergence. This is basically an iterative technique.

B. Numeric example

Let the dataset be D = {2, 3, 4, 10, 11, 12, 20, 25, 30}. We need to find the clusters using K-mean algorithm.

(i) We know that K is the number of clusters we want, let K = 2 (selected randomly)

(ii) Now we need to choose the centres arbitrarily. Let the two centres be M1 = 4 and M2 = 12.

(iii) Let the two clusters be K1 and K2.

So, K1 = {2, 3, 4} and K2 = {10, 11, 12, 20, 25, 30}

(iv) The new centres values are M1 = 3 and M2 = 18

(v) Now we will repeat until convergence:

K1 = {2, 3, 4, 10}, K2 = {10, 11, 12, 20, 25, 30}

M1 = 3, M2 = 18

K1 = {2, 3, 4, 10}, K2 = {11, 12, 20, 25, 30}

M1 = 4.75 = 5, M2 = 19.6 = 20

As the centre values are not changing so we will end the process here and finally the required clusters are as follows:

K1 = {2, 3, 4, 10, 11, 12}

K2 = {20, 25, 30}

IV. Applications

a.Prediction of performance of players: Using K-means clustering it can be easily predicted how a particular team or player would play in various sports or matches like football, cricket and so on.

b. Fraudulent recognition: Fraud or false transaction can also be determined using this technique.

V. ADVANTAGES

a. K-means algorithm is the simplest iterative method and can most easily implementable.

b. Works well even on bulky datasets.

VI. Drawbacks

a. We need to choose the value of K manually.[1]

b. The major drawback is that the whole procedure depends on the initial value of K and the centroids.

VII. Overcome Problems With K

The main disadvantage of K-means clustering algorithm is selecting the value of K. There are few metrics that give us the clairvoyance about K. Two of them are discussed below:

A. Elbow method

Elbow method[6] provides us with the concept of a better K-number of clusters that would be hinged on the Sum of Squared distance (SSE) between data points and the respective allocated centroids of the clusters.
Here we will select the value of K as 3 as the graph starts to take the shape of elbow at K = 3.

B. Silhouette analysis

In Silhouette analysis[7] sample in the respective cluster we need to do the listed steps:

a. Calculate the average distance from each data points to all data points in the cluster where it is present, not in the other cluster (ai).
b. Determine the average distance from every data points present in that cluster to the closest cluster (bi).
c. Evaluate the coefficient with the given formula in the equation (3):

$$\frac{b^i - a^i}{\max(a^i, b^i)} \qquad (3)$$

d. The coefficient takes the values in between the range [-1, 1].
e. If the coefficient is 0 then the sample is extremely close to the adjacent clusters.
f. If the coefficient is 1 then the sample is enormously far from the adjacent clusters.
g. Hence, the value of the coefficients should be as large as possible so that it is nearer to 1 that would help to form good clusters.

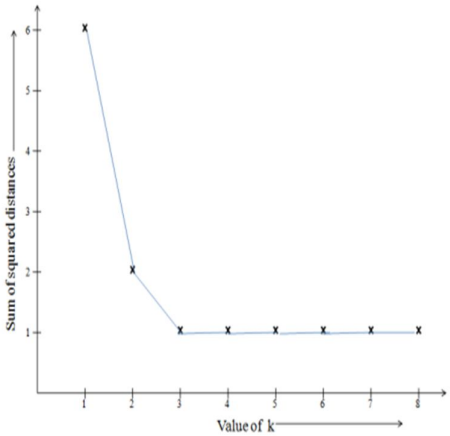

Figure 8. Finding the value of K using Elbow method

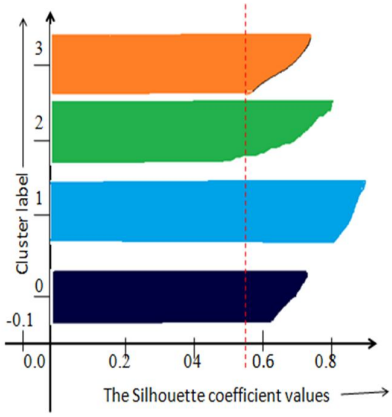

Figure 9. Silhouette analysis for K-means clustering on sample data with n cluster = 4
Here for n-clusters = 4, the average Silhouette score is 0.5882004012129721 (approximately)

VIII. Relationship Among Various Diseases And Their Prediction Using K-Means Clustering

K-means clustering techniques have been tremendously applied in the medical care department for trouble-free analysis and detection of diseases which results in delivering rapid, satisfactory. The table given below depicts the accuracy of detecting some of the diseases with the help of K-Means algorithm and combination of other algorithms with it.

TABLE I. ALGORITHM USED TO DETECT ILLNESS WITH THE PERCENTAGE OF ACCURACY
(Collected from International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015)

Algorithm	Accuracy	Disease
K-Means	98.24	Heart Disease
K-Means	78	Diabetics
K-Means (Attribute based)	80.198	Heart Disease
K-Means based MAFIA	74	Heart Disease
K-Means based MAFIA With ID3	85	Heart Disease
K-Means based MAFIA With ID3 and C4.5	92	Heart Disease
K-Means	78	Diabetics

IX. Conclusion

In this paper, we have anatomized in details about clustering and its types. We have also analyzed about K-Means clustering technique and also about its prime pitfalls. We have studied varied illness and determined ways that can be employed using K-Means algorithm solitarily or merged with some other algorithm that helped to reduce the cost of treatment and predict the disease accurately. As for further works, we plan to upgrade existing K-Means algorithm and also our model will be trained using different Micro-array datasets so that the performance of our model can be enhanced and that might help man-kind also in future.

Acknowledgement

I would like show my indebtedness towards my supervisor Dr. Bichitananda Patra. I am fortunate enough that I could work under his supervision. His exemplary guidance and constant encouragement were that stupendous that my profound gratitude would not be enough for him.

References

[1] K.Rajalakshmi, Dr.S.S.Dhenakaran and N.Roobini*"Comparative Analysis of K-Means Algorithm in Disease Prediction"*, International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015.

[2] KhedairiaSoufiane and Mohamed TarekKhadir*"A multiple clustering combination approach based on iterative voting process"*, Journal of King Saud University – Computer and Information Sciences, May 2019.

[3] VikasChaudhary, R.S. Bhatia and Anil K. Ahlawat*"A novel Self-Organizing Map (SOM) learning algorithm with nearest and farthest neurons"*, Alexandria Engineering Journal, September 2014.

[4] HailunXie, Li Zhang, CheePeng Lim, Yonghong Yu, Chengyu Liu, Han Liu andJulie Walters *"Improving K-means clustering with enhanced Firefly Algorithms"*, Applied Soft Computing Journal, December 2018.

[5] Ramzi A. Haraty, MohamadDimishkieh and MehediMasud ***"****An Enhanced k Means Clustering Algorithm for Pattern Discovery in Healthcare Data"*, International Journal of Distributed Sensor Networks, Volume 2015, Article ID 615740, 11 pages, December 2014.

[6] Bholowalia, Purnima, and Arvind Kumar. *"EBK-means: A clustering technique based on elbow method and k-means in WSN."* ,International Journal of Computer Applications 105, no. 9 ,2014.

[7] Wu, Di, and Ling Shao. *"Silhouette analysis-based action recognition via exploiting human poses."* Transactions on Circuits and Systems for Video Technology 23, no. 2 (2012): 236-243, IEEE

[8] Ye, Qixiang, Wen Gao, and Wei Zeng. *"Color image segmentation using density-based clustering."* In 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698), vol. 2, pp. II-401. IEEE, 2003.

[9] Chicco, Gianfranco, Roberto Napoli, and Federico Piglione. "*Comparisons among clustering techniques for electricity customer classification.*" Transactions on power systems 21, no. 2 (2006): 933-940, IEEE, 2006.