Analysis of Clustering Techniques in Data Mining

Maneeb Shahzad[1], Anurag Rana[2], Ankur Sharma[3]
[1]*M.Tech, Research Scholar, Arni University Kathgarh, Indora, (H.P) India*
[2]*Assistant Professor, Arni University Kathgarh, Indora, (H.P) India*
[3]*Associate Professor, Arni University Kathgarh, Indora, (H.P) India*

Abstract: The Data mining is the technology which is applied to extract useful information from the rough data. The clustering, classification and predictive analysis are the three domains of data mining. The clustering algorithms are broadly classified into hierarchal clustering, density based clustering etc. The k-means is the most efficient clustering algorithm of partitioned based clustering. In this paper, various variants of k-means clustering is reviewed and discussed in terms of description, outcome.

Keywords - *Clustering, Hierarchal, Partitioned, k-means.*

I. INTRODUCTION

In order to extract important information and the interesting patterns from the gathered data there was a need to develop an efficient mechanism. The process that helps in analyzing the data and extracting the necessary information from it with the application of certain tools is known as data mining. With the growth of technology, the numbers of applications that involve this process have been growing as well. In almost all the fields such as in marketing, medical fields, in education, in research and engineering fields this method has been involved [1]. In order to extract the required information the mining of data is done which is also known as the discovery of new knowledge in the databases. There are various types of information gathered within the systems. All this information needs to be stored within the proper locations in proper manner. For this there is a need to develop an organized database which can handle all such different types of information being received. In order to identify any frequent item set from the storage devices, the Knowledge Discovery Process is included. This process uses the association rule in order to extract the last frequently used set of item. From the data present in the databases, the KDD help in nontrivial extraction of implicit, new, as well as potentially useful information [2]. Data mining is originally a part of KDD which is also now used as a synonym. There are various steps followed in the case of knowledge discovery from databases. The steps begin from identifying the raw material and gathering it to form new important information.

There are various steps included in order to perform the knowledge discovery process within the databases [3]. The step-wise procedure is explained below:

i. Any type of data that includes noise within it or is irrelevant is eliminated which is also known as the data cleaning process.
ii. There is a combination of various types of data sources within the next step which is also known as the data integration step.
iii. Further, any type of data that is relevant to the analysis being made is extracted from the storage and the step is known to be data selection.
iv. On the basis of various types of data mining techniques, modifications are made in the data which is known as transform step [4].
v. With the help available methods, the interesting patterns are identified and extracted within the next step.
vi. The required patterns are then analyzed on the basis of various properties with the help of certain unique patters as well.
vii. The final step includes the discovered knowledge is provided to the user.

The grouping of data on the basis of their similarities into small groups known as clusters is known as data clustering method. The objects or clusters are generated here in which the similar data is placed into one cluster. The dissimilar data is placed within different clusters. In order to identify the similar objects within the process of knowledge discovery, this is the initial step. The grouping of data objects into set of disjoint classes which are known as clusters is known as the clustering mechanism. The objects within similar class are more like each other with respect to the objects present within different class. Numerous algorithms are applied in order to perform clustering [5]. The basic clustering algorithms being applied in current applications are:

a. Partitioning Methods:- The partitioning mechanism has the main objective of bringing together the samples with higher similarity together into the form of clusters and separating the ones that are dissimilar. Let k be the number of partitions that are required to be constructed [6]. An initial partitioning is generated with the help of this method and an iterative relocation technique is used in this method which helps in enhancing the partitioning method. The objects are moved from one group to the other in this process.

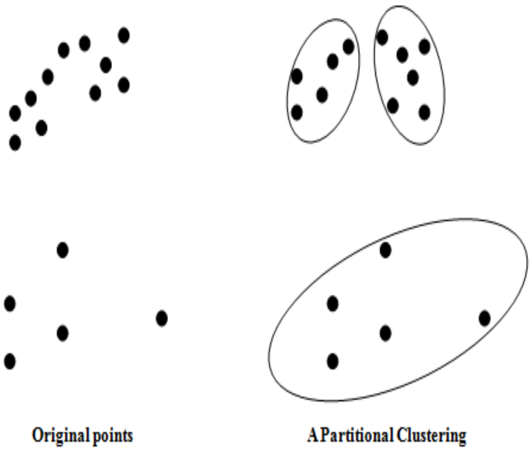

Fig. 1 Partitional Clustering

b. Hierarchical Methods:- There is a generation of hierarchical decomposition of the provided set of data objects through this method [7]. There are two classifications further in this method which is agglomerative and divisive.

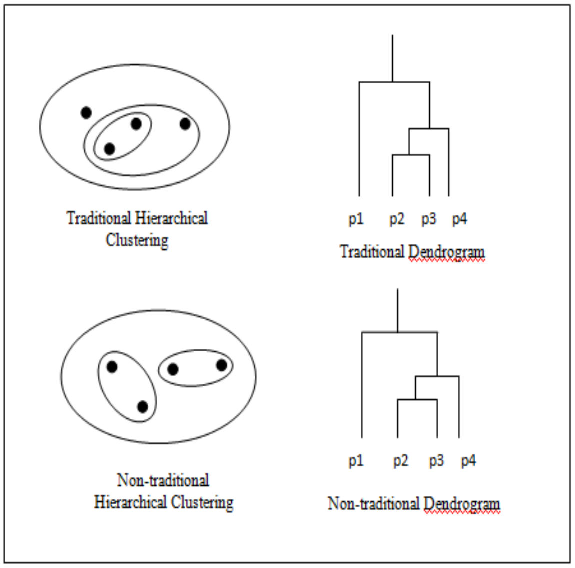
Traditional Hierarchical Clustering
p1 p2 p3 p4
Traditional Dendrogram
Non-traditional Hierarchical Clustering
p1 p2 p3 p4
Non-traditional Dendrogram

Fig. 2 Hierarchical Clustering

c. Density Based Methods:- On the basis of the notion of density, there are new methods introduced which can be utilized in cases where arbitrary shapes are involved. The arbitrary shape of the clusters is discovered with the help of this process along with the handling of noise present within the data. The scanning is done only once and the density parameters are required within it as well.

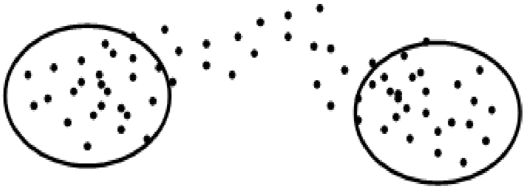

Fig. 3 Density based clustering

In order to extract the important information which can be useful, the raw data is analyzed. This process is known as data analytics. It is of two types which are [8]:

- ✓ **Classification:** The categorical class labels are predicted with the help of classification models and the continuous valued functions are predicted with the help of prediction models in this process.
- ✓ **Prediction:** The cost spent by the potential customers on the computer devices as per their occupation and income is predicted with the help of prediction model.

II. LITERATURE REVIEW

K. Rajalakshmi et al., represented an extremely fast growing field of medical [9]. The medical data mining are useful to produce optimum results on prediction based system of medical line. This paper analyzes various disease predictions techniques using K-means algorithm.

Oyelade et al., defined the ability of the student performance of high learning [10]. To analyze student result based on cluster analysis and use standard statistical algorithm to arrange their score according to the level of their performance. In this paper K-mean clustering is implemented to analyze student result. The model was combined with deterministic model to analyze student's performance of the system.

Bala Sundar V et al., examined the conclusion of the accuracy of the result by using k-mean clustering technique in prediction of heart disease diagnosis with real and artificial datasets [11]. Clustering is the method of cluster analysis which aims to cluster to partition into k clusters and each cluster has its observations with nearest mean. Each cluster assigned to the cluster k and started from random initialization. The proposed technique further divided into k groups. The grouping is done by minimizing the sum of squares of distances between data using Euclidean distance formula and the corresponding cluster centroid. The research result shows that the integration of clustering gives promising results with highest accuracy rate and robustness.

Daljit Kaur et al., explained that clustering is a division of data into groups of similar objects [12]. Each group consists of objects that are similar among them and dissimilar compared to objects of other groups. K-means algorithm is widely used for clustering data. But this algorithm is computationally expensive and quality of final results depends on the selection of initial centroid. This paper proposes a method to make the algorithm more efficient and effective. The proposed method decreases the complexity and effort of numerical calculation but it maintains the easiness of implementing k-means algorithm. It also solves the problem of dead unit.

Richa Sharma, et.al, surveyed on two different complex diseases which incorporates the coronary illness and in this survey paper the literary works of different writers are reviewed in field of medical data mining utilizing different classification and clustering techniques advance it is talked about that various tools are accessible for data preprocessing and classification [13]. This survey study reveals the importance of research in area of life debilitating disease diagnosis. It is discussed that one need to achieve for the precision of cent percent various investigates approximately comes to their target yet disease diagnosis suffers from high false alarm so there is a need to propose novel approach to reduce this false alarm rate which would help in early diagnosis of disease.

Sonali Shankar, et.al, proposed a study on the enormous data of 14000x5 of Harvard University online course. The performance metrics of registered students are discovered from different countries by means of K-mean clustering method. The paper expects to break down the performance of the students in light of different attributes with respect to their country [14]. The average performance of the students belonging to different countries is investigated in light of different attributes, for example, composed occasions, sections learned and number of days they interacted with the course. The attributes are subsequently compared with the average grades of students of respective countries and it is concluded that the grades are by all account not the only factor to represent the best possible understanding of the course.

Vadlana Baby, et.al, proposed in this paper an efficient distributed threshold privacy-preserving k-means clustering algorithm that uses the code based threshold secret sharing as a privacy-preserving instrument [15]. This protocol takes less number of iterations compare with existing protocols and it don't require any trust among the servers or users. The experiment results are likewise furnished alongside comparison and security analysis of the proposed scheme. It permits gatherings to collaboratively perform clustering and hence avoiding trusted outsider. The protocol is compared with CRT based clustering proposed. This algorithm does not require any trust among the servers or users and it give idealize privacy preserving of client data.

IV. TABLE OF COMPARISION

Author	**Year**	Description	**Outcome**
K.Rajalakshmi et.al	2015	The medical data mining are useful to produce optimum results on prediction based system of medical line. The paper analyzes various disease predictions techniques using K-means algorithm. This	Data mining based on prediction system are reduces the human effects and cost effective one.
Oyelade et.al	2010	This paper defined the ability of the student performance of high learning. In the paper K-mean clustering is implemented to analyze student result.	K-mean clustering is implemented to analyze student result.
Bala Sundar V et.al	2012	This paper examined the conclusion of the accuracy of the result by using k-mean clustering technique in prediction of heart disease diagnosis with real and artificial datasets.	The research result shows that the integration of clustering gives promising results with highest accuracy rate and robustness.
Daljit Kaur et al	2013	This paper proposes a method to make the algorithm more efficient and effective.	The proposed method decreases the complexity and effort of numerical calculation but it maintains the easiness of implementing k-means algorithm.
Richa Sharma, et.al	2016	This paper surveyed on two different complex diseases which incorporates the coronary illness	It is discussed that one need to achieve for the precision of cent percent various investigations approximately comes to their target.
Sonali Shankar, et.al	2016	The paper expects to break down the performance of the students in light of different attributes with respect to their country.	It is concluded that the grades are by all account not the only factor to represent the best possible understanding of the course.
Vadlana Baby, et.al	2016	In this paper an efficient distributed threshold privacy-preserving k-means clustering algorithm that uses the code based threshold secret sharing as a privacy-preserving instrument	This protocol takes less number of iterations compare with existing protocols and it don't require any trust among the servers or users.

IV. CONCLUSION

In this paper, it has been concluded that clustering is the technique of data mining. The clustering techniques have been classified into hierarchal, partitioned and density based clustering. The k-means is the most efficient clustering algorithm which can cluster similar and dissimilar type of data and has been analyzed that it gave maximum accuracy. The accuracy of any algorithm is the ratio of number of points clusters correspond to total number of data points.

V. REFERENCES

[1]. QASEM A. AL-RADAIDEH, ADEL ABU ASSAF 3EMAN ALNAGI, "Prediction Stock Prices Using Data Mining Techniques", The International Arab Conference on Information Technology (ACIT'2013)

[2]. Siddheswar Ray and Rose H. Turi, "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation", School of Computer Science and Software Engineering Monash University, Wellington Road, Clayton, Victoria, 3168, Australia, 1999.

[3]. Azhar Rauf ,Mahfooz, Shah Khusro and Huma Javed "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, ISSN 1990-9233, Issue 12,vol. 7, pp 959-963,2012

[4]. Madhu Yedla, T M Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center", International Journal of Computer Science and Information Technologies, vol. 1 (2), pp 121-125, 2010

[5]. K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering , IWCE , vol. 1,July 1 - 3, 2009, London, U.K

[6]. Osamor VC, Adebiyi EF, Oyelade JO and Doumbia S "Reducing the Time Requirement of K-Means Algorithm" PLoS ONE, Issue 12, vol. 7, pp-56-62, Dec. 2012.

[7]. Azhar Rauf, Sheeba, Saeed Mahfooz, Shah Khusro and Huma Javed, "Enhanced K-Mean Clustering Algorithm to Reduce Number of Iterations and Time Complexity", Middle-East Journal of Scientific Research, ISSN 1990-9233, Issue 7,vol. 7, pp 959-963, 2012.

[8]. Kajal C. Agrawal and Meghana Nagori, "Clusters of Ayurvedic Medicines Using Improved K-means Algorithm," International Conf. on Advances in Computer Science and Electronics Engineering, 2013

[9]. K. Rajalakshmi, Dr. S. S. Dhenakaran, N. Roobin "Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2278 – 7798, Issue 7, vol. 4, pp 2697-2699, July 2015

[10]. Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, ISSN 1947-5500, Issue 1,vol. 7, pp 292-295, 2010

[11]. Bala Sundar V,T Devi, N Saravan, "Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888), Issue 7, vol. 48, pp 8, June 2012

[12]. Daljit Kaur and Kiran Jyot, "Enhancement in the Performance of K-means Algorithm", International Journal of Computer Science and Communication Engineering, ISSN: 2319-7080, Issue 1, vol. 2, pp 29-32, 2013

[13]. Richa Sharma, Dr. Shailendra Narayan Singh, Dr. Sujata Khatri," Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey", 2016, IEEE, 978-1-5090-0210-8

[14]. Sonali Shankar, Bishal Dey Sarkar, Sai Sabitha, Deepti Mehrotra," Performance Analysis of Student Learning Metric using K-Mean Clustering Approach", 2016, IEEE, 978-1-4673-8203-8

[15]. Vadlana Baby, Dr. N. Subhash Chandra," Distributed threshold k-means clustering for privacy preserving data mining", 2016, IEEE, 978-1-5090-2029-4