© 2018 IJSRST | Volume 4 | Issue 2 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

Survey of Different Data Clustering Algorithms

N. Kavithasri*[1], R. Porkodi[2]
*[1]PG Student, Department of Computer Science, Bharathiar University, Coimbatore, Tamiul Nadu, India
[2]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamiul Nadu, India

ABSTRACT

Cluster is a group of objects that belongs to the same class. Clustering is widely used in diverse areas. There are number of clustering techniques available today. The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Data mining is mainly used in telecommunication industry used to identifying the telecommunication patterns, catch fraudulent activities, construct recovered use of source and obtain better value of service. This paper presents the study and analysis of five clustering algorithms namely Simple KMeans, Density Based clustering, Filtered Cluster, Farthest First, and Expectation Maximization for Individual household electric power consumption dataset. The performances of these algorithms are compared using the performance evaluation metrics namely Time taken to build, Number of cluster, and Number of cluster instances. The experimental results show that Filtered cluster, Simple KMeans and Farthest first produce better result than Expectation Maximization and Density Based.

Keywords: Clustering, Simple KMeans, Density Based clustering, Filtered Cluster, Farthest First, and Expectation Maximization.

I. INTRODUCTION

Data mining is the procedure of discovering patterns in massive data sets involving scheme at the relationship of machine learning, statistics and database systems. Data mining techniques is a necessary process where intellectual methods are applied to extract data patterns. It is a subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into a reasonable structure for further use [1]. Data mining is the study step of the "Knowledge Discovery in Databases" process, or KDD. There are many data mining techniques have been developing including sequential patterns, decision tree, association, classification, clustering and prediction. In association, a pattern is discovered depend on a connection between items in the similar transaction. Classification is used to classify each one item in a set of information into one of a predefined set of classes or groups. The prediction is one of a data mining techniques that find out the connection between independent variables and connection between dependent and independent variables. The clustering is grouped by the absorption of user.

Sequential patterns study is one of data mining technique that seeks to resolve or recognize likely patterns, regular events in transaction information over a business period. A decision tree is one of the majority used data mining techniques because its representation is easy to realize for users. Clustering is the undertaking of grouping objects in such a method that objects in the similar groups or more similar to each other than to those in other groups (clusters). It can be achieved by diverse algorithms that differ notably in their notion of what constitutes a cluster and how to powerfully find them. Popular conception

of clusters includes groups with small distances along with the cluster members, deeply areas of the information space, intervals or particular statistical distributions [2].

The section I discuss about the introduction of data mining and the clustering algorithm. Section II gives the analysis about literature survey. Section III explains the methods that are used clustering algorithms. The result and discussion are explained in section IV. Section V concludes this analysis work.

II. LITERATURE REVIEW

Indhirapriya .P, Dr.D.K. Ghosh (2013) presented a paper on clustering algorithms such as Soft Clustering, Neural Network Clustering and Genetic Based Clustering. They proposed the CURE algorithm suited for large dataset and produced accurate result as compared to other algorithm [3].

Archana Singh (2016) presented a paper on the social media analytics by understanding, analyzing the social media data for university students that support a broad variety of users and types of social media. The KMeans algorithm was used for grouping the cluster based on given values [4].

Amel Grissa Touzi et.al (2012) proposed the new software using Cluster Knowledge Discovery in Databases (KDD) and Classification knowledge Discovery in database (KDD). It concluded that Clustering Knowledge Discovery is suitable for larger dataset but the software contains more complication [5].

Namarata S Gupta et.al (2015) presented a paper on various clustering techniques name namely partitioning, density based, hierarchical, grid based, model based, constraint based technique along with their speciality, advantages and disadvantages [6].

Jaison B, Kumar T (2017) reviewed on understanding of stability clustering in Vehicle Ad hoc Network (VANET). In that review, Clustering was mainly depending on Cluster Head (CH). The VANET provided more stable, secure network structure and required efficient mobility prediction method. The mobility aware clustering in VANET facilitates clearly explained how network forms cluster and elect their CHs for creating more stable network [7].

Sukhvir Kaur (2016) presented a paper on clustering techniques namely partitioning, density based, hierarchical, grid based technique along with their speciality, advantages and disadvantages [8].

Jasmine Irani et.al (2016) presented the distance metric of similar cluster, pattern matching for similar cluster and negative data. This paper described some important distance measures with formula they are Euclidean Distance, Jaccard distance, Manhattan distance and Minkowski distance and it finally need for pattern matching for similar cluster and challenges of clustering of negative data [9].

R.Kabilan, Dr.N. Jayaveeram (2015) discussed different data mining techniques. It would help to evaluating all possible software services on the cloud computing by using clustering technique. This paper presented KMeans algorithm is more efficient algorithm as compare to remaining algorithm and it is suitable for large database [10]. Eric Sanjuan (2005) presented the context of querying a scientific textual database, the overlap of terms and cluster labels with the keywords selected by human indexers as well as set of possible queries based on the clustering output. This paper used only minimal linguistic resources to extract terms and relate them with Lexical operations using some operations [11].

Divya Tomar, Sonali Agarwal (2013) presented different data mining techniques used in healthcare sector and they used KMeans clustering, Hierarchical

Clustering and Density Based Clustering. These clustering techniques are used for particular disease and it produce accurate result depends on dataset [12].

III. METHODOLOGY

The proposed research methodology consists of three phases as shown in Fig.1. The first phase is pre-processing. The second phase is clustering data mining in that the five algorithms namely Makedensitybased, Simple KMeans, Expectation Maximisation, Farthest First Clustering, Filtered Clustering are used. The last phase is used to evaluate the performance of the clustering algorithms using different evaluation metrics.

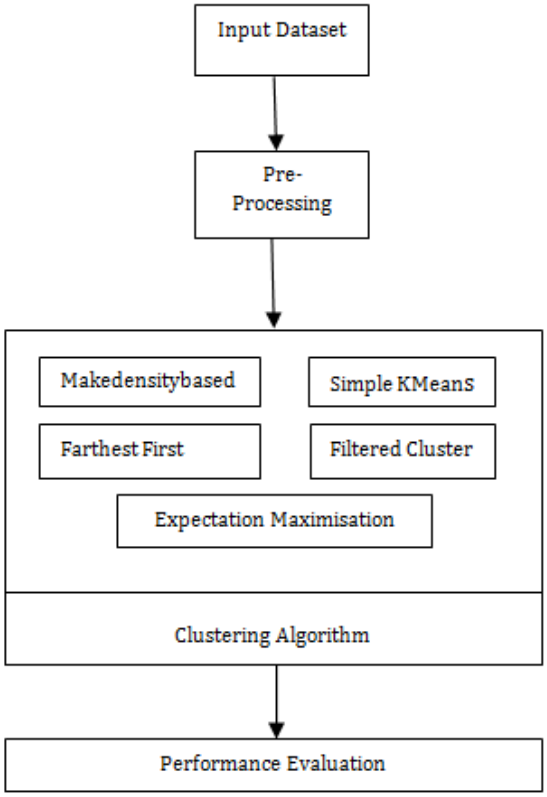

Figure 1. Methodology

A. Clustering Algorithms

In cluster analysis is primarily to separation the position of data into groups based on resemblance of data and then allocate the labels to the groups. A cluster of similar data objects can be treated as one group. A good clustering method will produce high quality clusters with high cluster similarity. The clusters contain different types of cluster.

i) Well-Separated Clusters: - A cluster is a position of points such that one point in a cluster is nearer (or more related) to each and every one other point in the cluster than to whichever point not in the cluster.

ii) Center-Based:- An entity is more close up to or related to the cluster in which it resides then the other clusters then it is called as center-based cluster.

iii) Contiguous Cluster: – A cluster is a set of points such that a point in a cluster is closer to one or more other points in the cluster than to any point not in the cluster.

iv) Density based Cluster: - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density. This type of cluster used only when the clusters are irregular or intertwined and when noise and outliers are present [13].

There are many algorithms for clustering in data mining in which five clustering algorithms such as Simple KMeans, Density Based clustering, Filtered Cluster, Farthest First, and Expectation Maximization are chosen for experimental study that are explained in next paragraphs.

1)Make Density Based Clustering Algorithm: Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-based spatial clustering of applications with noise (dbscan) is most widely used density based algorithm. It uses the concept of density reach-ability and density connectivity [2].

2)Simple KMeans: Simple KMeans clustering is a means of vector quantization, originally from signal processing. The most important aim of the algorithm is to partition n observations into k clusters in which each observation belongs to the cluster with the

nearest mean, allocation as a prototype of the cluster. The algorithm has a movable relationship to the k-nearest neighbour classifier, a popular machine learning technique for classification that is frequently confused with k-means because of the k in the name [2].

3) Expectation Maximization: An EM algorithm is an iterative method to find out maximum likelihood or maximum a posterior estimates of parameters in statistical models, where the form depends on unnoticed latent variables. This algorithm iteration alter between performing an expectation (e) step, which creates a task for the expectation of the log-likelihood, evaluated using the present estimate for the parameters and maximization (m) step, which calculate parameters maximizing the expected log-likelihood found on the step [2].

4)Farthest-First: The farthest-first traversal of a surrounded metric space is a series of points in the space, where the initial point is particular randomly and each succeeding point is as far as feasible from the set of previously-selected points. The similar idea can also be apply to a finite set of geometric points, by restricting the particular points to belong to the set or regularly by considering the finite metric space generated by these points. For a finite metric space or finite set of geometric points, the resultant sequence forms a permutation of the points, known as the greedy permutation [2].

5)Filtered Clustering: The filtered algorithm is used for filtering the data or pattern. In this the user provisions a model set of appropriate information. On the future of latest data they are comparing against the filtering profile and the data identical to the keywords is offered to the user [2].

IV. RESULT AND DISCUSSION

A. Dataset Description:

The Table 1 consists of Individual household electric power consumption Data Set [14] information and contains measurements of electric power consumption in one household with a one-minute sampling rate over a period of almost 4 years. Different electrical quantities and some sub-metering values are available.

Table 1. Household electric power consumption Data Set

ATTRIBUTE	DESCRIPTION
Date	Date in format dd/mm/yyyy
Time	Time in format hh:mm:ss
Global_active_power	Household global minute-averaged active power.
Global_reactive_power	Household global minute-averaged reactive power.
Voltage	Minute-averaged voltage.
Global_intensity:	Household global minute-averaged current intensity.
Sub_metering_1	Energy sub-metering no.1. It corresponds to the kitchen, containing mainly a dishwasher, an oven and a microwave.
Sub_metering_2	Energy sub-metering no.2. It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.
Sub_metering_3	Energy sub-metering no. 3 It corresponds to an electric water-heater and an air-conditioner.

B. Result of Clustering Algorithms

The Table 2 shows the result of performance evaluation of the clustering algorithms namely Simple

KMeans, Density Based clustering, Filtered cluster, Farthest First and Expectation Maximization.

Table 2. Result of Performance Evaluation

RESULT	DENSITY BASED	SIMPLE KMEANS	EM	FILTERED CLUSTER	FARTHEST FIRST
Number of cluster	2	2	8	2	2
Number of Instances	0(94%)	0(99%)	0(90%) to	0(99%)	0(99%)
	1(6%)	1(1%)	7(1%)	1(1%)	1(1%)
Number of Iteration	3	3	68	-	3
Log-Likelihood	-4.9	-	-3.6	-	-
Sum of squared errors	16241	16241	-	-	16241
Time Taken	0.3	0.2	1368.3	0.09	0.2

The Figure 2 shows the number of clusters grouped between different algorithms. The Expectation Maximisation algorithm grouped into 8 clusters and remaining algorithms mostly grouped into 2 clusters. The Expectation Maximisation algorithm would be grouping into cluster slightly different from remaining algorithms. .

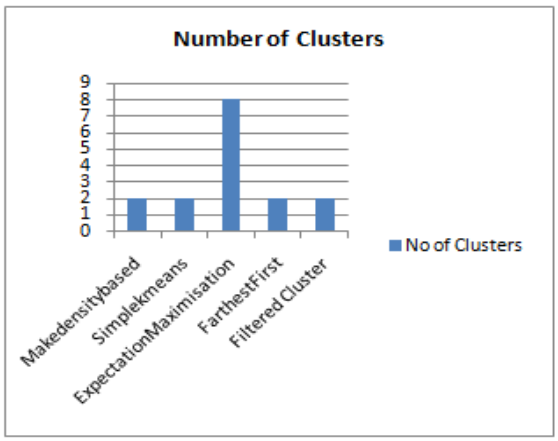

Figure 2. Number of Clusters Produced

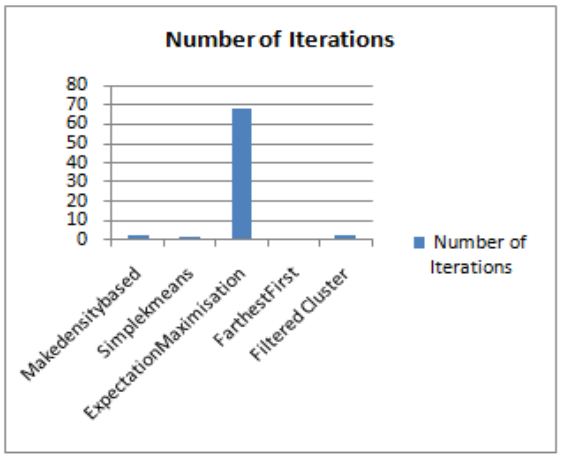

Figure 3. Number of Iterations Taken

The Figure 3 shows the numbers of iterations performed for grouping the cluster between different algorithms. Here Expectation Maximisation (EM) takes more iteration for grouping the cluster as compare to different algorithms. The remaining algorithms take similar number of iterations for grouping the cluster.

The Figure 4 shows the time taken for building the clusters between different algorithms. The Expectation Maximisation clustering algorithm taken more time to build compared to remaining algorithms.

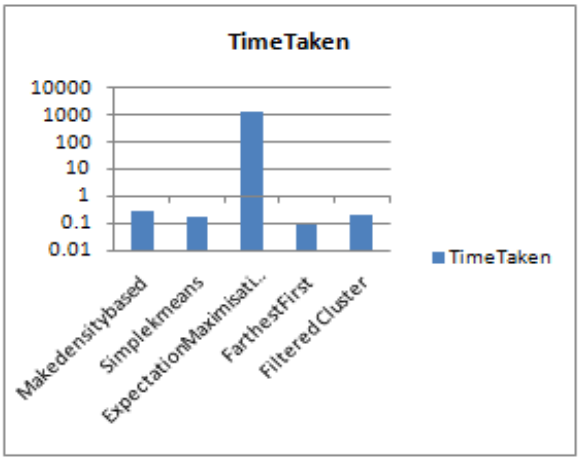

Figure 4. Time Taken to Build Clusters

The five clustering algorithms have been validated using the important metrics as shown in Table 2 and the result shows that filtered cluster, farthest first and simple kmeans algorithm gives better accuracy result than density based algorithm and EM algorithm.

V. CONCLUSION

The extensive survey has been conducted in clustering in data mining area and analysed various clustering algorithms used so far. This paper implemented the five clustering algorithms namely Simple KMeans, Density Based clustering, Filtered Cluster, Farthest First, and Expectation Maximization by using individual household electric power consumption dataset as experimental dataset and compared the performance of the above algorithms. The simple KMeans clustering algorithm is simplest algorithm as compared to other algorithms and its performance is better than EM and Make density based Clustering algorithm. Make Density based clustering algorithm is not suitable for data having very huge variations in density. EM algorithm takes more time to build cluster. Make Density based algorithm takes more or less fewer time to construct a cluster but it is not better than the Simple KMeans algorithm since density based algorithm have higher log likelihood value, if the value of log likelihood is high then it makes bad cluster. Finally, Based on the performance results, it is observed that simple KMeans, filtered cluster and farthest first provides better accuracy than other algorithms.

VI. REFERENCES

[1]. META Group application development strategies: "Data mining for data warehouses: uncovering hidden patterns".

[2]. http://www.zentut.com/data-mining/data-mining-techniques/

[3]. P. Indirapriya, Dr.D.K. Ghosh," A survey on different clustering algorithms in data mining technique", A survey on different clustering algorithms in data Mining technique.

[4]. Archana Singh, "Mining of social media data of university students", Springer.

[5]. Amel Grissa Touzi, Amira Aloui, and Rim Mahouachi, "Cluster_KDD: A visual clustering and knowledge discovery platform", Springer.

[6]. Namrata Gupta, Bijendra, S. Agrawal, Rajkumar M. Chauhan, "Survey on clustering techniques of data mining", American international journal of research in science, technology, engineering & mathematics.

[7]. Jaison B, Kumar T, "A review of stability aware clustering algorithm in vehicular ad hoc network" International conference on innovations in information embedded and communication systems.

[8]. Sukhvir Kaur, "Survey of different data clustering algorithms", International journal of computer science and mobile computing.

[9]. Jasmine Irani, Nitin Pise, Madhura Phatak, "Clustering techniques and the similarity measures used in clustering", International journal of computer applications.

[10]. R. Kabilan, Dr.N.Jayaveeran, "Survey of data mining techniques in cloud computing, International journal of scientific engineering and applied science.

[11]. Eric Sanjuan, "Query refinement through lexical clustering of scientific textual databases", Springer.
[12]. Divya Tomar, Sonali Agarwal, "A survey on data mining approaches for healthcare", International journal of bio-science and bio-technology.
[13]. Tryon, Robert C. Cluster analysis: Correlation profile and Optometric (factor) Analysis for the isolation of unities in mind and personality. Edwards brothers.
[14]. https://archieve.ics.uci.edu/ml/datasets/.
[15]. Prakash Singh, Aarohi Surya, "Performance analysis of clustering algorithms in data mining in weka", International journal of advances in engineering & technology.
[16]. Priyanka Sharma, "Comparative Analysis of Various Clustering Algorithms Using WEKA", International Research Journal of Engineering and Technology.
[17]. Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya, "Comparison the various clustering algorithms of weka Tools", International Journal of Emerging Technology and Advanced Engineering.
[18]. Anu Sharma, Dr. M.K Sharma & Dr. R.K Dwivedi, "Literature Review and Challenges of Data Mining Techniques for Social Network Analysis", Advances in Computational Sciences and Technology.
[19]. Sharmila, R.C Mishra, "Performance Evaluation of Clustering Algorithms", International Journal of Engineering Trends and Technology.