International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 1 | ISSN : 2456-3307

Comparison of Clustering Algorithm

R. Indhu[1], R. Porkodi[2]

[1]PG Scholar, Department of Computer Science, Bharathiar University, Coimbator, Tamilnadu, India
[2]Assistant Professor, Department of Computer Science, Bharathiar University, Coimbatore, Tamilnadu, India

ABSTRACT

Clustering is a technique used in data mining that groups similar objects into one cluster, while dissimilar objects are grouped into different clusters. Distributed data mining allows for access to volumes of data that are housed at several different company sites or at various organizations. Extremely complicated algorithms are formed to recover the essential data anyway of where it is stored so that it can be useful to a particular data model that will distribute the accurate knowledge and information. The objective of this paper is to perform a comparative analysis of four clustering algorithms namely K-means algorithm, Hierarchical algorithm and Density based algorithm and Expectation maximization algorithm. These algorithms are compared in terms of efficiency and accuracy and observed that K-means produces better results as compared to other algorithms.

Keywords: Clustering, K-means algorithm, Hierarchical algorithm, Expectation and maximization algorithm and Density based algorithm.

I. INTRODUCTION

Data mining refers to extracting information from large amounts of data, and transforming that information into an understandable and meaningful structure. Data mining is an essential step in the process of knowledge discovery from data (KDD).Data mining is used in various techniques they are classification, sequential pattern discovery, clustering, regression, association rule discovery, outlier detection, etc. Data mining is a multi-stage process data is mined by going through various phases, as Data selection retrieves the data from the database that are related to the analysis task. In Preprocessing, data are cleaned and/or integrated. Data transformation, transforms data into appropriate form for mining, by applying summarization or aggregation functions [1].

Data mining is an essential step where intelligent methods are performed in order to extract useful patterns and knowledge. Interpretation/evaluation identifies patterns that representing knowledge based on some measures. In data mining, mining of data can be done using two learning approaches Supervised and Unsupervised learning. Clustering is an unsupervised learning in data mining applications. Clustering is the task of grouping a set of objects in such a way that objects in the cluster is more similar to each other than to those in other clusters. Clustering techniques have numerous applications in various fields including, artificial intelligence, pattern recognition bioinformatics, segmentation and machine learning [2].

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

II. LITERATURE SURVEY REVIEW

Dr.D.K.Ghos, et.al [3] presented a paper on clustering algorithms such as 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.

Alex. A, et.al [5] presented paper to demonstrate different data mining techniques in healthcare sector. Kmeans clustering and Hierarchical Clustering are the two techniques used for particular disease and it produces accurate result depends on the dataset.

Meenu Sharma [4] presented a paper to perform analysis of different diabetic patient data. The clustering techniques used for comparison are kmeans, hierarchical, density based etc. Kmeans algorithm has better accuracy when compared with other algorithms.

Mythili, et.al [6] presented a paper which provides an overview of algorithms along with their advantages and disadvantages. The different clustering methods that have been studied are partitioning clustering, hierarchical clustering, density based clustering.

Tamilkili.M [7] presented a paper on various clustering techniques namely partitioning, density based, hierarchical, based, model based, constraint based technique along with their specialty, advantages and disadvantages.

Amandeep Kaur Mann [8] discussed the different data mining techniques used in cloud computing. It would help to evaluating all possible software services on the cloud computing by using clustering technique. This paper determines that the Kmeans algorithm is more efficient algorithm as compared to remaining algorithms and it is suitable for large database.

Madura phatak.et.al [9] proposed the new software using Cluster Knowledge Discovery in Databases and Classification knowledge Discovery in database (KDD). It concluded that Clustering Knowledge Discovery is suitable for larger dataset but the software contains more complication.

Mihika shah.et.al [10] presented a paper that discussed the various types of algorithms like k-means clustering algorithm. This paper provides a broad survey of the most basic techniques such as hierarchical and partition algorithm.

Unnati.R.et.al [11] presented a paper that demonstrates the distance metric of similar cluster, pattern matching for similar cluster and negative data. This paper described some important distance measures with formula such as euclidean distance, Manhattan distance and Murkowski distance.

Sukhvir Kaur [12] presented a paper on clustering techniques namely partitioning, density based, hierarchical, grid based technique along with their specialty, advantages and disadvantages.

III. METHODOLOGY

The proposed research methodology consists of three phases as shown in Figure1. The first phase is pre-processing and the second phase is clustering in data mining in which four algorithms namely kmeans, hierarchical, make density based and expectation maximization clustering are used. The last phase is used to evaluate the performance of clustering algorithms using different evaluation metrics.

Clustering is an important technique in data mining and it is the process of partitioning data into a set of clusters such that each object in a cluster is similar to another object in the same cluster and dissimilar to every object not in the same cluster. Dissimilarities and similarities are assessed based on the attribute standards describing the items and frequently involve distance measures [13].

Clustering analyses the data objects without consulting a known class label. This is because class labels are not known in the first place, and clustering is used to find those labels. Good clustering exhibits high intra-class similarity and low inter-class similarity, that is, the higher the similarity of objects in a given cluster, the better the clustering. The superiority of a clustering algorithm depends equally on the similarity measure used by the method and its implementation.

There are many algorithms for clustering in data mining in which four clustering algorithms namely Simple KMeans, hierarchical clustering, density Based clustering, and Expectation Maximization are chosen for experimental study that are explained in next paragraphs.

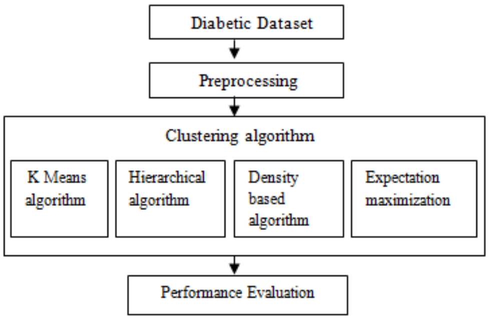

Figure 1. Methodology

1. Clustering Algorithms

1.1 K-Means Algorithm

K-means clustering is a process of vector quantization, initially from signal processing, that is well-liked for cluster analysis in data mining. K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Verona cells. K-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes [14]. The algorithm has a loose relationship to the k-nearest neighbor classifier, a popular machine learning technique for classification that is often confused with k-means because of the k in the name. One can apply the 1-nearest neighbor classifier on the cluster centers obtained by k-means to classify new information into the presented clusters. This is known as nearest centric classifier or Rocchio algorithm.

1.2 Hierarchical Algorithm

Hierarchical Clustering method merged or splits the similar data objects by constructing hierarchy of clusters also known as dendogram. Hierarchical Clustering method forms clusters progressively [15]. Hierarchical Clustering classified into two forms: Agglomerative and Divisive algorithm.

Agglomerative clustering: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy

Divisive clustering: This is a "top down" approach. This clustering observations start in one cluster, and splits are performed recursively as one moves down the hierarchy [16].

1.3 Density Based Clustering

A cluster is a dense region of points that is separated by low density regions from the tightly dense regions. This clustering algorithm can be used when the clusters are irregular It finds core objects i.e. objects that have dense neighborhoods [17]. It connects core objects and their neighborhoods to form dense regions as clusters. Clusters are formed as maximum sets of density connected points and can detect noise and used when outliers are encountered.

1.4 Expectation Maximization

In algorithm can be used to generate the best hypothesis for the distributional parameters of some multi-modal data. Expectation Maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates among the stage an expectation (E) step, which creates a function for the expectation of the log-likelihood calculate using the present estimate for the parameters, and a maximization (M) step, which calculate parameters maximizing the expected log likelihood found on the *E* step [18].

IV. RESULT AND DISCUSSION

4.1 Dataset Description:

The Table 1 shows the Diabetic data set with 391 instances and 7 following attributes with nominal value are considered: ct, bpan, bpan2, bwad, bwan, bwan2 ,bmad, bman, bman2 has been used for analysis diabetic patient data due to its proficiency of disease.

Table 1

ATTRIBUTE	DESCRIPTION	POSSIBLE VALUES
Ct	computed tomography	Numeric
Bpan	beta-propeller protein-associated neurodegeneration	Numeric
Bpad	Bipolar-affective disorder	Numeric
Bwad	Britishweightlifters associationfor disabled	Numeric
Bmad	behavioral medicine and addictive disorder	Numeric
Bpad	beta-propeller protein-associated disorder	Numeric
Bpan2	beta-propeller protein-associated neurodegeneration2	Numeric

4.2 Experimental Result

The Table 2 shows the result of performance evaluation of the clustering algorithm namely kmeans, hierarchical clustering, density based clustering, expectation maximization.

Table 2. Result of Performance Evaluation

Performance evaluation	kmeans	Hierarchical	Density based	Em
Number of cluster	2	2	2	5
Cluster instances	0(97%)	0(100%)	0(62%)	0(44%)
	1(3%)	1(0%)	1(38%)	1(66%)
Log likelihood	-	-	-62.7962	-49.095
Number of iteration	4	2	3	6
Sum of squared error	3811	-	47.583	-
Time taken	0.05	1.28	0.05	11.47

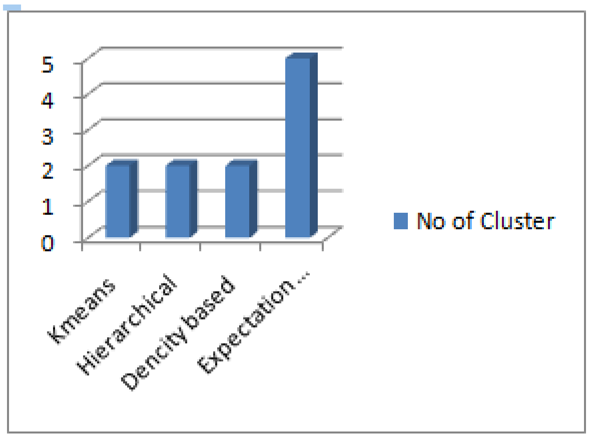

Figure 2. Number of Clusters Produced

The Figure 2 shows the number of clusters grouped between different algorithms. The expectation maximization algorithm grouped into 5 clusters and remaining algorithm mostly grouped into 2 clusters. The expectation maximization algorithms would be grouping into cluster slightly different from remaining algorithms.

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

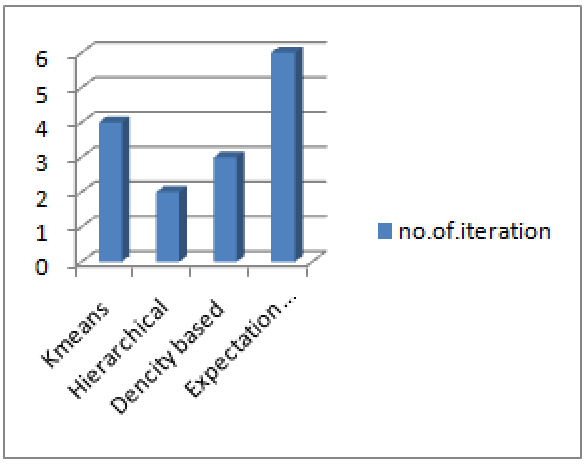

Figure 3. Number of Iteration Produced

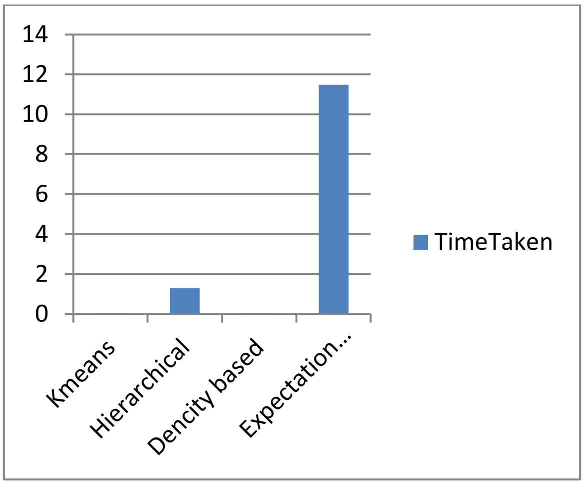

Figure 4. Time Taken to Clusters Produced

The Figure 4 shows the time taken for clustering between different algorithms. The expectation maximization clustering algorithm taken more time to compared remaining algorithms.

V. CONCLUSION

In this paper a comparative study has been performed on the analysis of four clustering algorithms: k-means, hierarchical, density based and expectation maximization clustering algorithms. The experimental result shows that is hierarchical clustering, producing the best cluster accuracy of 100% and the time taken is 1.28 seconds. The second best algorithm is the kmeans, producing the cluster accuracy of 97% and the time taken is 0.05 seconds. The third best algorithm is density based cluster, producing cluster accuracy 62% and the time taken is 0.05 seconds. The fourth best algorithm is Expectation maximization, producing cluster accuracy 44% and the time taken 11.47 seconds. The performance evaluation of an algorithm is mainly dependent on the type of the data set. The future work will be focused on using the other clustering algorithms of data mining.

VI. REFERENCES

[1]. https://en.wikipedia.org/wiki/Cluster_analysis

[2]. http://googleweblight.com.

[3]. Survey of different clustering algorithm in data mining technique in p.Indirapriya Dr.D.K.Ghos2 International Journal of modern engineering research (IJMER).

[4]. Clustering in data mining: A Brief Review in Meenu Sharma International Journal of Core Engineering & management (IJCEM)

[5]. A Survey of Evolutionary Algorithms for Clustering in Alex A.Freitas,and Andre Capone Leon f. de Carvalho IEEE Transactions on system, man, and Cybernetics part C:Application and review, vol.39.no.2, march 2009.

[6]. An analysis on Clustering Algorithm in Data Mining Mythili S1, Madhiya E2 International Journal of Computer Science and mobile Computing.

[7]. A Survey on Recent Traffic Classification Techniques Using Machine Learning Methods in M.Tamilkili journal of Advanced Research in Computer Science and Software Engineering.

[8]. Survey paper on Clustering Techniques in Amandeep Kaur Mann(M.TECH C.S.E)International journal of Science,Engineering and Technology Research(IJSETR).

[9]. Clustering Techniques and the Similarity Measures used in Clustering:A Survey Jasmine lrani Nitinpise Maduraphatak International of Computer Application(0975-8887)Volume 134-No.7,January 2016.

[10]. A Survey of Data Mining Clustering Algorithm in Mihika Shah Sindhu Nair International Journal of Computer Applications.

[11]. Implementing & Improvisation of K-means Clustering Algorithm in Unnati R. Raval1, ChaitaJani2 International Journal of Computer Science and Mobile Computing.

[12]. Survey of Different Data Cluster Algorithm in Sukhvir kaur Sukhvir kaur,International Journal of Computer science and Mobile Computing.

[13]. T.T. Nguyen, G. Armitage, A survey of technique or Internet traffic classification using machine learning, IEEE Commun. Surveys Tutor. 10 (4) (2008) 56-76.

[14]. J. Erman, A. Mahanti, M. Arlitt, I. Cohen, C. Williamson, Offline/realtime traffic classification using semi-supervised learning, Performance Evaluation 64 (9-12) (2007) 1194-1213.

[15]. Jun Zhang, Yang Xiang, Wanlei Zhou, Yu Wang, Unsupervised traffic classification using flow statistical properties and IP packet payload, Journal of Computer and System Sciences 79 (2013) 573-585.

[16]. JyotiYadav, Monika Sharma, A Review of K-mean Algorithm, International Journal of Engineering Trends and Technology (IJETT) - Volume 4 Issue 7- July 2013.

[17]. G. Sathiya and P. Kavitha, An Efficient Enhanced K-Means Approach with Improved Initial Cluster Centers, Middle-East Journal of Scientific Research 20 (4): 485-491, 2014.