A Survey on Various Clustering Algorithms

Gurpreet Kaur[#1], Nidhi Bhatla[*2]

[#]*M. Tech, Research Scholar*
Department of Computer Science Engineering,
RIMT-IET, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India.
[1]gurpreetdhaliwal25@gmail.com

[*]*Assitant Professor*
Department of Computer Science Engineering,
RIMT-IET, Mandi Gobindgarh, Fatehgarh Sahib, Punjab, India.
[2]engineernidhi@yahoo.com

***Abstract*— The management of large amount of information becomes a big challenge for many areas such as business, marketing, medical science etc. The data mining provides the better solution for this problem. The text mining, which is application of data mining, provides different methods such as classification, clustering etc. for extracting the important information from unstructured text documents or data. The clustering is a technique which group similar data objects into one group or cluster. The requirements of the clustering algorithms are ability that how it deals with noisy data, insensitivity to the order of input data, scalability etc. There are different clustering models: connectivity based model, centroid based model, distribution model, density model. This paper presents different clustering algorithm such as k-means, Y-means, DBSCAN algorithm, Fuzzy c mean, Hierarchical clustering and EM clustering algorithm. This paper focuses on the working, performance, efficiency and merits and demerits of different algorithms and carefully analyses these algorithms to find out the further scope of research in clustering algorithms.**

***Keywords*— Data mining, Text mining, Clustering, k-means, Hierarchical clustering, DBSCAN clustering algorithm.**

I. INTRODUCTION

In today's World discovering patterns and trends from large databases becomes challenging issues as the amount of stored information has been enormously increasing day by day. The management or storage and extraction of useful information from unstructured data becomes a problem for many areas such as business, universities, research institutes, government funding agencies, and technology intensive companies. Data mining provides a solution for this problem.

A. Data Mining

Data mining emerged in 1980 for creating the useful information. Data mining is used to extract useful patterns and previously unknown trends from the large databases. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Data mining has various techniques such as classification, clustering, decision trees, neural networks etc. The applications of data mining are text mining and web mining. In the process of data mining, before applying any technique pre-processing of data is required, this is the part of knowledge discovery process (KDD). The important steps of KDD steps of process are:

• In the **first** phase, the expectation and business objectives are defined.

• In the **second** phase, the above defined objective helps in the selection of data from the data ware house. Then that data is pre-processed in order to improve quality of data.

• In the **third** phase, the algorithm of data mining is selected and applied to the data which was prepared in second step. This is a vital step, of knowledge discovery process. The relationship and patterns would be the output of this phase.

• In the **fourth** phase, according to the objectives, valid patterns are found by analysing the patterns and relationships.

• The **last** phase is the visual representation of the knowledge discovered. These results can be stored, assembled which can be used to improve the business.

B. Text Mining

In the modern life, text is a common vehicle for exchange of information. Extracting of useful information from large text is very challenging issue. So there arises a need of text mining.

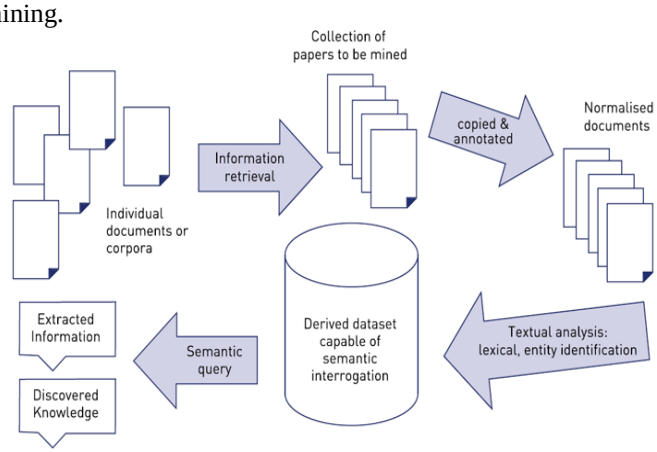

Fig. 1 Text Mining Process

Text mining is the process of extracting important knowledge or patterns from the unstructured text that are from no. of sources. Text mining is a new and exciting research area that tries to take the challenge and produce the intelligence tool. The tool is a text mining system which has the capability to analyse large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information. Text mining is the use of automated methods for exploiting the enormous amount of knowledge available in text document. The main stages of text mining as shown in Fig. 1 are:

1) Document pre-processing: The pre-processing of the document is done to represent the documents in such a way that their storage in the system and retrieval from the system become very efficient. To reduce the length or dimension of the document two methods are used:

- *Filtering*: It is a process for the removal of the words which do not provide any useful or relevant information. Stop word filtering is a standard filtering method. We can remove the words like prepositions, articles etc.
- *Stemming*: stemming is the process for reducing inflected or derived words to their stem, base or root form. A stemming algorithm reduces the words fishing, fished, fisher to the root word fish.

2) Text mining technique is applied: In this stage the text mining algorithm such as classification, clustering, summarization, natural language processing, information extraction is applied.

3) Text Analysis: In this stage the outputs which are obtained from the previous steps are analysed using various tools such as link discovery tool such that user gets important information to achieve the perspective.

II. CLUSTERING

A. Clustering

Clustering is a unsupervised learning which provides a important role in a business environment. Clustering means that grouping of similar types of objects into one cluster. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. The objects can be physical like a student or can be an abstract such as behaviour of a student, marks. There are number of clustering algorithms which can be categorised using cluster models. Also there are numbers of cluster models, some of them are mostly used, as shown in Fig. 2, which are [9]:

1) Connectivity models: Hierarchical clustering is based on this model. It connects objects based on their distance and form clusters.

2) Centroid model: In this each cluster is represent by a single mean vector or a central vector. For example: k-means clustering.

3) Distribution model: In this model clusters are define as objects belonging to the same distribution. In this statistical distribution is used. For example Expectation-maximization (EM) algorithm.

4) Density model: In this, clusters are defined as connected higher dense regions than the reminder of the data space. For example- DBSCAN algorithm.

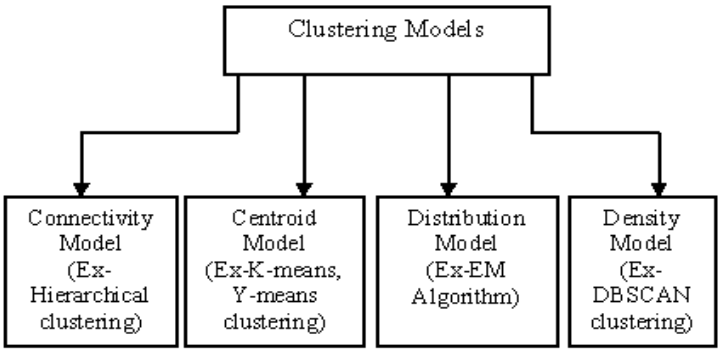

Fig. 2 Clustering Models

B. Applications of clustering

There are many areas where clustering can be applied and helps. The common areas where clustering can be applied are given below [9]:

1) Business and Marketing: The clustering provides useful information from surveys and trends. In marketing it helps by grouping shopping items.

2) World Wide Web: In www using clustering, recognitions of communities in social network, grouping of search results etc. can be done.

3) Social science: Clustering helps in finding the areas in crime analysis, where similar type of crimes mostly occurs.

4) Educational Institutes: Clustering helps in making groups of students in educational institutes.

5) Computer science: In computer science clustering can be applied in image segmentation, for software evaluation etc.

III. CLUSTERING ALGORITHMS

A. K-Means Algorithm

K-means clustering is a popular method for cluster analysis in data mining. In this method, n observations are partitioned into k clusters, where k is the number of clusters

defined by the users but the value of k is fixed. In clustering process, first of all centroid of the each cluster is selected then on the basis of selected centroid, data points having minimum distance from the given cluster are assigned to the particular cluster. Its main steps are [8]:

Let a document set D (d1, d2, d3……dm).

- Firstly choose k-data points as initial centroids.
- Then Find out the distance between each d ε D and the chosen centroid.
- Assign d to the closest cluster.
- Recompute the centroid until it becomes stable.

In [1], their work is for classification and clustering of research proposals and reviewers. Their main focus is on assigning the appropriate proposal group to the appropriate reviewer. For classification of research proposals and reviewers, the C4.5 decision tree algorithm and for clustering, k-means algorithm is used. Using k-mean algorithm with the help of ontology, the proposals and reviewers are grouped according to their similarities in each discipline area. This work efficiently classifies the research area and easily groups the reviewers and proposals.

In [2], their research has proposed a new algorithm enhanced k-means. Their algorithm reduces the number of iterations, by calculating initial centroids. Their work proceeds in two steps. In the first step, keeping cluster size fix and initial clusters are formed by splitting the input array of element into sub-arrays. In the second step the size of the cluster vary and finalised clusters are formed using initial clusters as input. The centroids of initial clusters are calculated, then the distance between centroid and data element is calculated. The data elements having equal or less distance stay in the same cluster otherwise put them into their relevant cluster. In last comparison between basic k-means algorithm and their enhanced k-means algorithm is done. The comparison shows that their enhanced k-mean algorithm reduces the number of iterations and improves the elapsed time than basic K-means algorithm.

1) Advantages: The k-means algorithm takes less computation time than hierarchical clustering for number of variables. K-mean algorithm can produce tighter clusters than hierarchical clustering, if the value of k is small.

2) Disadvantages: It is very difficult to predict the value of k. This algorithm cannot work well with global cluster. This algorithm cannot work efficiently with different size and different density.

B. Y-Mean Algorithm

Y-mean is an autonomous clustering algorithm which means it has an ability to autonomously decide the number of clusters. This algorithm produce the self- defined number of clusters based on the statistical nature of the data than the user define constant as in the case of k-mean algorithm.

In [3], their research addresses the comparison between k-means and y-means algorithm as shown in Fig. 3. For comparison they used iris flower's dataset and cluster the different items between three cluster based on the species of iris flowers namely Iris virginica, Iris setosa and Iris versicolor. Their research concludes that Y-mean algorithm do the clustering of data sets in few iteration than k-means algorithm and also improves the average run time.

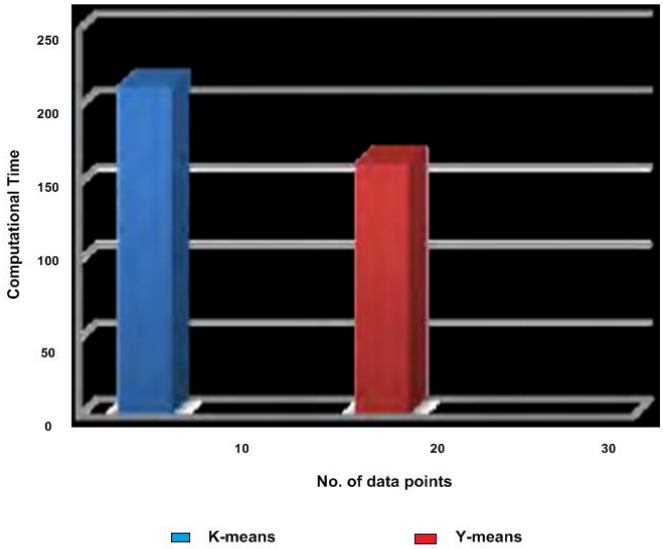

Fig. 3 Comparison between K-means and Y-means algorithms

1) Advantages: The y-mean algorithm removes the drawback dependency on initial centroids of k-means clustering algorithm. Y-mean has an ability to decide the number of cluster automatically as in k-mean user has to decide the value of k (number of cluster).

2) Disadvantages: This algorithm cannot show the dependency relationship between different data points that belongs to different clusters.

C. DBSCAN Algorithm

The density-based spatial clustering of application with noise is a density based clustering algorithm. It identifies clusters using one input parameter by analysing the local density of the database elements from large spatial data sets. The DBSCAN algorithm put the nodes into separate clusters using density distribution of these nodes. These separated clusters define different classes [6]. In this minimal knowledge of the domain is required. The DBSCAN can also determine that which information we can classified as noise or outliers. The working process of this algorithm is quick and performs very well with the size of the database – almost linearly.

In [7], their research is on document clustering using ontology with concept weighting. The clustering is done using k-means and DBSCAN respectively and compares the results of clustering. The clustering is done in three steps: pre-processing of document, on the basis of ontology compute the

concept weight, documents clustering using concept weight. In the research comparison between two ontology based algorithms DBSCAN and k-means is done on the basis of F-measure and accuracy. F-measure is calculated using precision and recall value. The precision and recall values are derived using four values: true positive (TP), false positive (FP), true negative (TN) and false negative (FN). The comparison show that the DBSCAN algorithm produce better clustering results than k-means clustering using the concept weight in ontology.

In [11], their research combines the two algorithms namely Multi-type Feature Co-selection for Clustering algorithm (MFCC) and DBSCAN clustering algorithm. The MFCC algorithm uses the feature selection. The DBSCAN algorithm is used to cluster high density regions. The main advantages of DBSCAN algorithm that it works time efficiently and have the ability to cluster the arbitrary shapes whereas MFCC can effectively reduce the noise features. The two algorithms are merged to improve the performance and result of clustering. For the modified algorithm selection score is:

$$fss(\varepsilon,minpts)=sf\ (\ fss\ (p_{i,c},\ \varepsilon,minpts))$$

$$\text{i=1}\quad \text{c=1}$$

Where:

SF – selection function
fss – feature selection score which is used to select best center point from the given feature space
ε – maximum distance between two samples
minpts – minimum number of points exists in the ε neighborhood
npts – neighborhood points
c – cluster
p – size of database

For best precision metrics, the measures like F-measure and time precision taken from each fss criteria. F-measure is computed from the harmonic mean of vocabulary terms (P) and total terms(R). The P & R terms are defined by each fss criteria. Accuracy is calculated by dividing the correctly classified testing documents to the total number of testing documents. DBSCAN is better in the terms of time factor and also there is no need to specify the number of clusters.

1) Advantages: In this algorithm, user has not required to specify the number of clusters. This algorithm can build arbitrary shaped clusters and find the clusters that are surrounded completely by the different clusters.

2) Disadvantages: The quality of this algorithm depends upon the distance measure. It cannot work efficiently if there is large difference in densities.

D. Hierarchical Clustering

Hierarchical clustering is a method in which hierarchy of clusters are generated. This method belongs to connectivity based model. In this clusters are formed continuously [12]. It consists of two categories:

1) Bottom Up hierarchical clustering method: This category is also known as agglomerative approach. In this each document is considered as a single cluster and combines the clusters repeatedly on the basis of similarity until a single cluster is achieved [8].

2) Top Down hierarchical clustering method: This category is also known as divisive approach. In this one cluster is divided into multiple clusters. If any document has less similarity to that cluster then put into another cluster. This approach is less used in applications because the number of computations is more and it is complex in nature [8].

In [14], in their research hierarchical method is represented which relates with partition method such as K-means, K medoids etc. and with density based methods such as DENCLUE which is centre-defined. The research concludes that hierarchical method is better than both K-means and K-medoids such as it reduce outlier influence. Hierarchical method has less complexity than partition based algorithm and have less storage requirements than density based method. The implementation is done on stream, multimedia and spatial data using P-trees.

In [15], the new method Clustering based SVM (CB-SVM) is proposed in which SVM means support vector machine is a method doing classification and regression analysis. SVM cannot work efficiently for very large data sets. Their work merges a clustering method with SVM to effectively execute for large data sets. CB-SVM using limited amount of resources available generates SVM boundary for large data sets on the basis of hierarchical clustering in which division can be performed for finding the high quality boundaries for SVM. When high classification accuracy is generated for large data sets on real and synthetic data sets then CB-SVM method is very scalable.

1) Advantages: There is an embedded flexibility related to the levels of granularity. It is applicable to any type of attributes. It handles every type of similarity easily [12].

2) Disadvantages: The termination criteria are unclear. Once the clusters are constructed hierarchical approach do no visit again for the purpose of improvement.

E. EM Algorithm

The expectation-maximization (EM) algorithm is used for finding maximum a posteriori or maximum like estimates of parameters by number of iteration in statistical model. The EM performing an expectation (E) step, by using the current estimate for the parameter in which the expectation of the log-likelihood is calculated. In maximization step, the parameters for the expected log-likelihood found on the E step are

evaluated. The result of the parameters which are estimated is then used for the latent variables in the next E step [16].

- *Expectation:* The missing labels are estimated by fix the model.
- *Maximization:* It finds the model of expected log-likelihood of the data by fix the missing labels.

In [16], there research proposed a new method for clustering HMM models. This model assumes that generation of signal is done by the process double embedded stochastic. This model is based on probabilities. In the research, a hierarchical EM algorithm is used for clustering various HMMs (VHEM-H3M). This algorithm cluster the HMM models using the idea of distribution they represent. The represented input HMMs are modelled by small mixture of novel HMMs which are estimated and maximizing the log-likelihood of virtual samples which are generated from the input HMMs . In this way the research provides a novel HMMs cluster centres. The efficiency of the algorithm is great for clustering HMMs model than other existing algorithms.

1) Advantages: This algorithm is numerically stable with EM iteration which increases the likelihood. It has suitable global convergence which fits under any general conditions. EM algorithm is easily implements, analyses, and computed because it acquires small storage space and is easily implemented. If the monotone increases in likelihood with number of iterations, to monitor programming and convergence errors is easy. The number of iterations needed for the EM algorithm is more in number in contrast with other existing procedures but for this the cost per iteration is low. This algorithm provides estimation for missing data.

2) Disadvantages: It does not provide the estimation automatically of the parameters estimated for the co-variance matrix. Sometimes it takes more time for the convergence. The E-steps and M-steps are analytically intractable for the number of problems.

F. Fuzzy C-mean Algorithm

This algorithm is same as k-means algorithm because in this the value of C (number of cluster) has to be defined by the user. Fuzzy C-mean is a technique in which clustering is done by grouping datasets into n clusters. In this every data point belongs to every cluster with a high degree of belonging (connection) to that cluster and other which have low degree of belonging to that cluster lies far away from the centre of a cluster [10].

In [4], their research work compares the two algorithms the k-means and Fuzzy C-mean algorithms on basis of efficiency of clustering output. For their research they use UCI Machine Learning Repository. UCI is a collection of database which is mostly used for research. The dataset of Iris plant is used. The implementation is done in the MATLAB. The function kmeans is used for clustering the dataset into k clusters and similarly the fcm function is used for fuzzy c mean clustring. The fcm function also returns the membership grades for each data point. At last their work concludes that the k-mean algorithm takes a less execution time than fuzzy c mean algorithm. The fuzzy c-mean algorithm is mainly used for finding functional dependencies and association rules.

In [5], their research also compares the two algorithms fuzzy c mean and entropy based fuzzy clustering (EFC) using four data sets IRIS, OLITOS, WINES and psychosis. The EFC algorithm calculate the entropy values of data points and select the cluster centre from the data points whose entropy value is minimum. In the research work they proposed the new methods in EFC algorithm for clustering. In last Self Organising Map (SOM) is used for visualisation by mapping clusters into 2-D. The research concludes that for some dataset the FCM algorithm perform faster and for other the methods of EFC algorithm perform faster. It means the performance of the algorithm depends upon the dataset.

1) Advantages: This algorithm is an unsupervised technique. It has ability to show the dependency relationship between the different points that belongs to different clusters.

IV. Conclusion

In this paper, different clustering methods which can be used for large databases are discussed such as k-means, y-means, fuzzy c mean, DBSCAN, hierarchical clustering etc. Different methodologies and the associated parameters for clustering algorithms are described. The k-means, Y-means and fuzzy c means algorithms depends upon the centroids. The k-means and y-means algorithm not show the dependency relationship between different clusters but fuzzy c mean algorithm show the relationship between clusters. In both K-means and fuzzy c means the number of clusters has to be defined by the user. DBSCAN algorithm is depend upon the density of the database and can build arbitrary shaped cluster. The DBSCAN, there is no requirement that user specify the number of clusters. In hierarchical clustering tree like structure of cluster is made using bottom up or top down approaches. The hierarchical clustering is applicable for any attribute and can handle any type of similarity. EM algorithm comes under the distribution based model. In this, the EM iteration which increases the likelihood is reliable and stable under different general conditions. The cost per iteration is low in this algorithm and also used to provide estimation for missing data. This paper discusses the general working and results on the basis of parameters such as accuracy, execution time, efficiency etc.

Acknowledgement

The authors wish to thank the reviewers and editors for their suggestions and constructive comments that help in bringing out the useful information and improve the content of paper.

REFERENCES

[1] Preet Kaur and Richa Sapra "*Ontology Based Classification and Clustering of Research Proposal and External Research Reviewers*", International Journal of Computers & Technology, Volume 5, No. 1, May -June, 2013, ISSN 2277-3061.

[2] 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 12 (7): 959-963, 2012, ISSN 1990-9233.

[3] V.Leela, K.Sakthi Priya, R.Manikandan "**A** Comparative Analysis Between K-Mean and Y-Means Algorithms in Fisher's Iris Data Set", published in International Journal of Engineering and Technology (IJET), Vol 5 No 1 Feb-Mar 2013.

[4] Soumi Ghosh and Sanjay Kumar Dubey "*Comparative Analysis of K-means and Fuzzy C-Means Algorithms* ", International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013.

[5] Subhagata Chattopadhyay, Dilip Kumar Pratihar and Sanjib Chandra De Sarkar, "*A Comparative Study Of Fuzzy C-Means Algorithm and Entropy Based Fuzzy Clustering Algorithms*", Computing and Informatics , Vol. 30, 2011, 701–720.

[6] Henrik Bäcklund, Anders Hedblom and Niklas Neijman "*A Density-Based Spatial Clustering of Application with Noise*" Linköpings Universitet , TNM033 2011-11-30.

[7] V.Sureka and S.C.Punitha **"***Approaches to Ontology Based Algorithms for Clustering Text Documents* **"** Int.J.Computer Technology & Applications, Vol 3 (5), 1813-1817 (2002).

[8] Divya Nasa, "*Text Mining Techniques- A Survey*", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 2, Issue 4, April 2012, ISSN: 2277 128X pp. 50-54.

[9] http://en.m.wikipedia.org/wiki/cluster_analysis.

[10] http://en.m.wikipedia.org/wiki/fuzzy_clustering.

[11] K.Parimala and Dr. V.PalaniSamy, "*Implementation of DB-Scan in Multi-Type Feature CoSelection for Clustering*", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 1, No 2, January 2013 ISSN (Print): 1694-0784 | ISSN (Online): 1694-0814.

[12] Amandeep Kaur Mann and Navneet Kaur, " *Survey Paper on Clustering Techniques* ", International journal of Science, engineering and technology research, Volume 2, Issue 4, April 2013, ISSN:2278-7798.

[13] Anne Denton, Qiang Ding, William Perrizo and Qin Ding, " *Efficient Hierarchical clustering of large data sets using P-trees*".

[14] Hwanjo Yu, Jiong Yang and Jiawei Han, "*Classifying large datasets using SVM with Hierarchical Cluster*".

[15] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, "*A Comparative Study of Various Clustering Algorithms in Data Mining*", International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com Vol. 2, Issue 3, May-Jun 2012, pp.1379-1384.

[16] Emanuele Coviello, Antoni B. Chan and Gert R.G.Lanckriet, "*The Variational hierarchical EM algorithm for clustering hidden Markov models*".

