A DETAILED STUDY ON CLUSTERING TECHNIQUES AND TOOLS FOR DATA MINING

Sachin Sirohi[3], Naveen Kumar[2], Anuj Kumar[3]

[1]M. Tech Student, JPIET, Meerut, Uttar Pradesh
[2]Assistant Professor, JPIET, Meerut, Uttar Pradesh
[3]Assistant Professor, UCER College, Allahabad, Uttar Pradesh

---

***Abstract**: Extraction of useful information from huge amount of data is known as data mining also known as knowledge discovery in database (KDD). There are so many sources that generates data in a very large amount like social networking sites, camera, sensors etc. This is the main reason that data mining is increasing rapidly. This paper presents a survey of clustering techniques and tools used for data mining. Classification is a supervised learning technique in which it identifies the class of unknown objects whereas clustering is an unsupervised learning. Clustering is the process of partitioning a set of data objects into subsets. Objects with in a cluster are more similar and dissimilar to other clusters. The similarity between objects are calculated using various distance measures like Euclidean distance, Manhattan distance, cosine etc.*

Keywords- Data Mining, Machine Learning, Classification, clustering algorithms, Supervised, Unsupervised Learnng

I. INTRODUCTION

Data mining plays a very important role for finding the frequent data pattern from internet, data set, data warehouse, data mart etc. Data mining, also called as data archeology, data dredging, data harvesting, is the process of extracting hidden knowledge from large volumes of raw data and using it to make critical business decisions. Data mining is used in various applications like finance, marketing, banking, credit card fraud detection, whether prediction.

Data mining helps to extract hidden patterns and make hypothesis from the raw data. Data mining process has mainly 7 steps as Data integration, data cleaning, data selection, data transformation, data mining, pattern evaluation and knowledge representation [1]. This process is shown in Fig-1.

Data Cleaning: Data in the real world is dirty, means incomplete, noisy and inconsistent data. Quality decisions must be based on quality data. So, before performing the analysis on the raw data, data cleaning is performed, which includes the following tasks:

Filling missing values.

Smooth noisy data and remove outliers by using algorithms like Binning algorithm. Resolve inconsistencies.

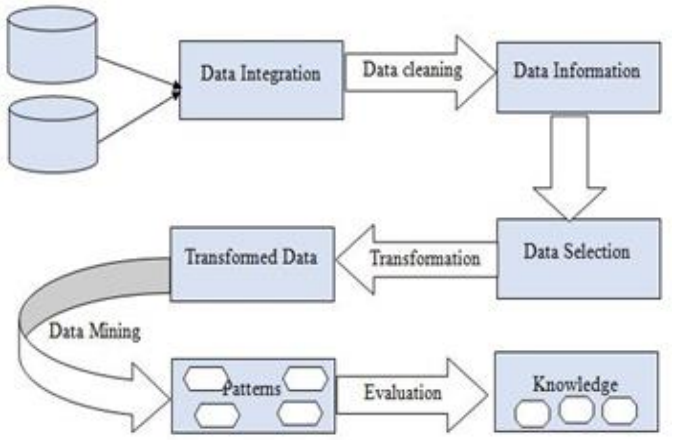

Figure 1: Data Mining Process

Data Integration: where multiple heterogeneous data sources may be combined.

Data Selection: Where task relevant data are selected from data warehouse or any other data sources including www, RDBMS etc.

Data Transformation: In data transformation, the data are transformed into format appropriate for data mining. For ex: An attribute data may be normalized

So as to fall between a small range 0 to 1. It includes the following tasks:

Smoothing: which works to remove noise from the data. Such techniques include binning, regression and clustering.

Aggregation: Various aggregation operations such as mean and median are applied to the data. For ex: the daily sales data may be aggregated.

Normalization: where the attribute data are scaled so as to fall within a small specified range, such as 0 to 1.

Data Mining: It is the process of extraction of interesting information or patterns from data in large database is known as data mining.

Pattern Evaluation: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns.

Knowledge representation: Various visualization and knowledge representation techniques are used to present the extracted knowledge to the user.

The rest of the paper is organized as follows. The section II provides a brief overview of the literature survey related to the clustering and classification learning algorithms. The section III explain about various clustering algorithms and the section IV provides the overview of data mining tools. Finally, section V concludes the work and provides future work.

II. RELATED WORK

A lot of researchers have implemented various data mining approaches in order to solve the various problems related to forecasting and analysis.

Nisha et al. [2] [2015] presented comparative study of clustering techniques. Clustering algorithms like K-means, K-medoid, Agglomerative are used for segmentation. Clustering is an unsupervised learning in which no target classes are defined. Algorithms learns from their experience and divide the data set into many clusters on the basis of similarity between data objects. Objects in a cluster have higher similar as comparison to other clusters.

Anna L. et al. [3] [2016] presented a literature of machine learning algorithms used for cyber security intrusion detection. There are three main types of cyber analytics in support of IDSs: misuse-based (sometimes also called signature based), anomaly-based, and hybrid. Misuse-based techniques are designed to detect known attacks by using signatures of those attacks. They are effective for detecting known type of attacks without generating an overwhelming number of false alarms. They require frequent manual updates of the database with rules and signatures. Misuse-based techniques cannot detect novel (zero-day) attacks.

Yong wang et al. [4] [2011] presented a survey of various data mining tools used for real world projects to perform prediction, analysis tasks and many more. This paper presented a comparative study between open source and commercial data mining tools. As it reports there have been more than 600 data mining software's in the world, and based on a tree hierarchy relationship with links it builds a classification framework to category all the software's into 21 groups as Classification software, Clustering and Segmentation software, Social Network Analysis, Link Analysis, and Visualization software, Statistical Analysis software, Text Analysis, Text Mining and Information Retrieval (IR), Visualization software, Web Analytics software, Web Usage Mining, Agents, Association rules and market basket analysis, Audio and Video Mining, Bayesian and Dependency Networks, BI (Business Intelligence), Database and OLAP software, Data Transformation, Data Cleaning, Data Cleansing, Deviation and Fraud Detection.

Maitri P. Naik et al. [5] [2015] presented a survey on document clustering. Clustering is the process of partitioning a set of data objects into subsets. It is commonly used technique in data mining, information retrieval, and knowledge discovery for finding hidden patterns or objects from a data of different category. Text clustering process deals with grouping of an unstructured collection of documents into semantically related groups. A document is considered as a bag of words in traditional document clustering methods; however, semantic meaning of word is not considered. Thus, more informative features like concept weight are important to achieve accurate document clustering and this can be achieved through semantic document clustering because it takes meaningful relationship into account. This paper highlights major challenges in traditional document clustering and semantic document clustering along with brief discussion.

Ashish Dutt et al. [6] [2016] presented a systematic review on educational data mining. Presently educational institutions compile and store huge volumes of data such as student enrolment and attendance records, as well as their examination results. Mining such data yields stimulating information that serves its handlers well. Rapid growth in educational data points to the fact that distilling massive amounts of data requires a more sophisticated set of algorithms. This issue led to the emergence of the field of Educational Data Mining (EDM). Traditional data mining algorithms cannot be directly applied to educational problems, as they may have a specific objective and function. This implies that a preprocessing algorithm has to be enforced first and only then some specific data mining methods can be applied to the problems. One such preprocessing algorithm in EDM is Clustering.

Hussain Ahmad Madni et al. [7] [2015] presented a survey on data mining techniques and applications. Data mining also known as knowledge discovery in database (KDD). It is defined as the process of extracting useful information from huge amount of data. In this paper various data mining techniques like classification and clustering are discussed. Classification is a twostep process. In the first step training is done and in next step testing of the system is performed. In classification, an object is classified into the predefined class. Whereas, in clustering the data set is partitioned into the clusters on the basis of similarity.

Sivaramakrishnan R Guruvayur et al. [8] [2017] provides a survey on Machine learning techniques for data mining. Throughout the year's data mining has delighted in enormous achievement, the application domains extended persistently yet the mining methods additionally kept up moving forward.

Various issues have developed and solution have found by data mining scientists. In any case, there are ranges and issues that still require consideration for future upgrades in this innovation. More research on the most proficient method to manage the social issue of in some cases, unconscious and unsuspecting people's security require to be conducted. Data mining procedures should accordingly develop to coordinate with this challenge.

Jinwook Seo et al. [9] [2006] We believe that the guiding GRID principles and, especially, the rank-by-feature framework can be useful to designers of other information visualization tools. Since it is difficult to conduct controlled experiments on complex tools that require substantial training and changes to analytic processes, we conducted three longitudinal case studies and an e-mail user survey.

A.Vinothini et al. [10] [2017] presented a review on machine learning methods for big data applications. Big data is described with 8 V's: Volume, Value, Veracity, Visualization, Variety, Velocity, Viscosity and Virality. Volume talks about the quantity of produced and warehoused data. Value refers to our ability to turn our data into value. Quality of captured data can vary significantly. Visualization tools present the information in a manner that is easy for the end user to understand and interpret. Data today comes in many different formats like structured data, unstructured data, semi-structured data, and even complex structured data. Velocity refers to the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. Viscosity measures the confrontation to flow in the volume of data. Virality describes how quickly information gets spread across network. Rate of speed is measured with respect to time

Paritosh Nagarnaik et al. [11] [2015] presented a survey on recommendation system. Recently several recommendation systems have been proposed, that are based on collaborative filtering, content based filtering and hybrid recommendation technique. Collaborative filtering technique (CF) is one of the most successful recommendation techniques to solve the scalability problem related to recommendation system and also providing better recommendation. In this paper cover all collaborative based recommendation techniques which are used for better recommendation. Also proposed new improve collaborative filtering technique using Hybrid recommendation which is combination of both K-mean algorithm and CHARM algorithm. This Hybrid recommendation method improves the prediction quality of recommendation system.

III. CLUSTERING ALGORITHMS:

Clustering is an essential technique in data mining in which a group of data objects is taken as input and as an output of number of clusters are obtained so that the objects in a group are more similar but are dissimilar to objects outside the clusters [12]. Clustering is also known as unsupervised learning because in clustering no target output is defined. Clustering algorithms are learning from their experience. Clustering is also used to detect outliers. An outlier is a value that do not belong to any cluster. Figure 2 shows outlier value and clusters.

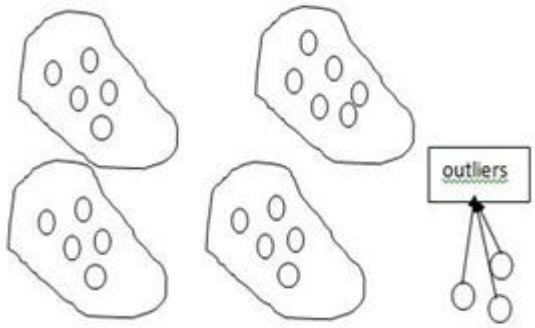

Figure 2: Clusters and Outliers

In various fields there are many benefits to be had from grouping similar objects. For example:

In a finance application we analyses the performance of companies by making clusters.

In a banking application we categorize the customers into various categories like profitable and non-profitable.

In a crime analysis application, we might look for clusters of high crimes such as burglaries or murders.

Types of Clustering Algorithms:

Figure 3 shows the various types of clustering algorithms:

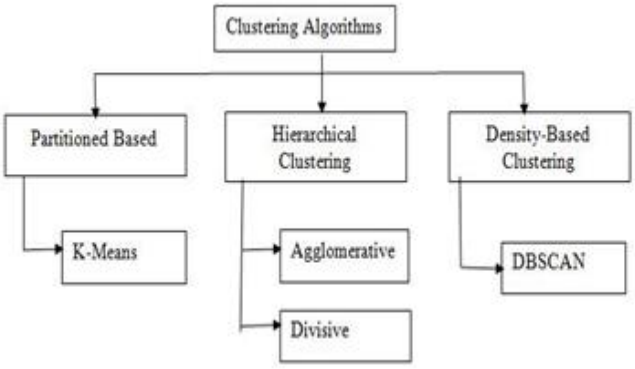

Figure 3: Types of Clustering algorithms

A. Partitioned based clustering:

In Partitional clustering a number of n objects data set is given as input and the data set is partitioned into the k cluster where k<=n. The k cluster satisfies the following two conditions: 1) Every object must belong to at least one cluster 2) The cluster must contain at least one object.

K-Means Algorithm:

K-means clustering is a method of cluster analysis which main aim to partition a set of n observations in to given k clusters. Figure 4 shows how the algorithm works?

Algorithm:

Input: Data set S having n observations and number of cluster K.

Output: K clusters

Step 1: The algorithm randomly selects k points as the initial cluster centroid.

Step 2: Calculate the distance of each point in data set S with every cluster centroid and assigned to the closed cluster.

Step 3: Recomputed the mean of each cluster.

Step 4: Repeat step 2 and 3 until the centroids no longer move.

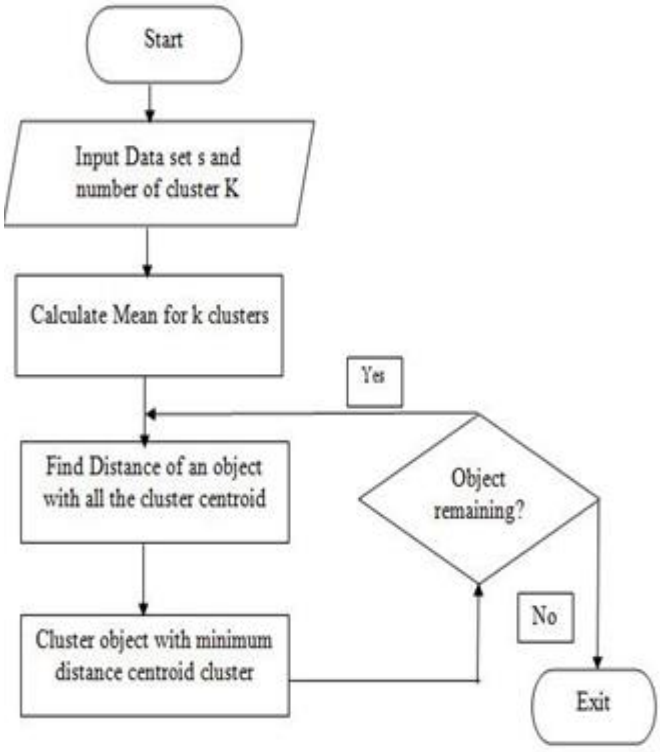

Figure 4: K-Means Flowchart

B. Hierarchical clustering:

Hierarchical clustering is a method of cluster analysis in which hierarchy of clusters is created in such a way that the data objects in clusters are decomposed based on some criteria. The clusters thus obtained in hierarchy are known as dendogram that shows how the clusters are related to each other [2]. There are mainly two approaches to generating a hierarchical clustering: Figure 5 shows how both the algorithms works?

Agglomerative:

In this algorithm starts with the points as individual clusters and at every step, merge the closet pair to form a cluster. This process is repeated until single cluster is formed having all the points.

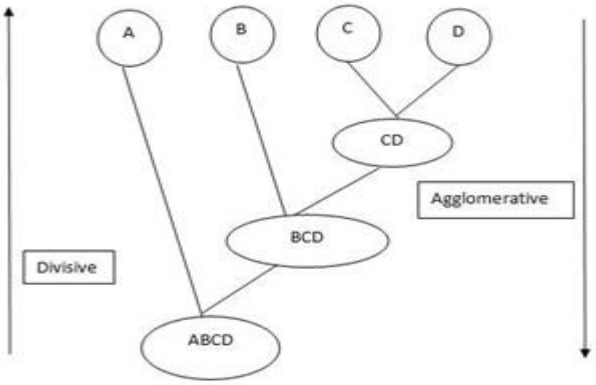

Figure 5: Agglomerative and Divisive Algorithms

Divisive:

In divisive, all of the objects are used to form one initial cluster. The cluster is split according to some principle, such as the maximum Euclidean distance between the closest neighboring objects in the cluster. This process is repeated until each cluster contains only a single object.

C. Density based clustering:

This technique of clustering is suitable for the arbitrary shaped clusters. Density based clustering helps us to separate the low dense regions of the clusters from the high dense regions. High dense regions of objects are combined together to form clusters. It deals with the noisy data and scans the whole data in only one scan [2].

DBSCAN (Density - Based Spatial clustering of Applications with noise):

DBSCAN is a density-based clustering algorithm. The algorithm grows regions with sufficiently high density into clusters, and

discover clusters of arbitrary shape in spatial databases with noise. It defines a cluster as a maximal of density-connected points. Two parameters are used:

Eps: It defines maximum radius of the neighborhood.

MinPts: It represents minimum number of points in an Eps-neighborhoods of that point.

A point is a core point it has more than a specified number of points (MinPts) within Eps. A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point. A noise point is any point that is not a core point or a border point.

Algorithm:

1. Select a point p.
2. Retrieve all points density reachable from p with respect to Eps and MinPts.
3. If p is a core point, a cluster is formed.
4. If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
5. Repeat this process until all of the points have been processed.

IV. Comparison of Different clustering algorithms:

This section discusses the comparison between various clustering algorithms with their advantages and disadvantages. Table I provides information about various algorithms.

Table I: Comparison of clustering algorithms

Comparison of clustering algorithms			
Algorithm	**Finding**	**Advantages**	**Disadvantages**
K-means	It gives accurate result when data is distinct and don't contain noisy value. It is simple and efficient. It is relatively fast	It product tighter clusters.	Difficult to predict K value. The clusters nonhierarchical and do not overlap.
Hierarchical clustering	Well suited for issues including point linkages, e.g. scientific classification trees.	No apriori information about the number of cluster is required.	Algorithm can never undo what was done previously.
Neural Networks	Works well with continuous values	Used to classify the pattern on untrained data.	Less interpretability
			It takes long training time
Density Based Clustering	Arbitrary shape clusters are formed and analyzed.	Work well with noisy data.	The algorithm is sensitive to the user defined parameter.

V. TOOLS Used for Data Mining:

This section provides the detail about data mining tools used for clustering. Also, it compares various data mining techniques which are used to perform analysis. Table II provides information about different tools.

Table II: Tools for Data mining

Tool for Data mining			
Tools Name	**Purpose of Tools**	**Open Source**	**Support for Clustering**
Weka [13]	Used for:- Data preprocessing, Classification, clustering, visualization, association rule mining and regression. –Also apply to big data	Yes	Yes
Rapid Miner[14]	It support all the phases of KDD	Yes	Yes
Matlab [15]	It is used for data and visualization	No	Yes
Standford Tokenizer[[16]	Used for text categorization and tokenization	Yes	Not Available
Apache OpenNLP[17]	It is an open source library and it mainly used for natural language processing.	Yes	Not Available
Lucene [18]	It is used for document preprocessing.	Yes	Not Available
Snowball stemming[19]	The Main aim of stemming is to reduce different grammatical forms.	Yes	Not Available

Conclusion:

In this paper, we have presented the survey of clustering algorithms and data mining tools used for analysis. There are mainly three types of clustering methods are discussed. In partition based clustering, the data set is divided into k clusters containing similar objects. In partitioned clustering, overlapping clusters are not allowed. In hierarchical clustering, overlapping clusters are allowed. In density based algorithms, an arbitrary shape clusters are formed. Density based method help us to differentiate between the low dense regions of the clusters from the high dense region. In future we will discuss more types of clustering algorithms.

REFERENCES:

1) H. Wahidad , L.V. Pey , N.K. Lee and O.L.Zhen, "Application of Data Mining Techniques for Improving Software Engineering ," International Conference on Information Technology, vol.5, pp. 1-5.

2) Nisha and Puneet jai kaur, "A Survey of Clustering Techniques and Algorithms," InternationalConference on Computing for SustainableGlobal Development (INDIACom), vol.2, pp. 304-307.

3) Anna L. Buczak, Member, IEEE, and Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, pp. 1153-1176.

4) Yong wang, Zhi-Gang Gu and Hao wang, "A Survey of Data Mining Softwares Used for Real Projects," IEEE International Workshop on Open-source Software for Scientific Computation, 2011, pp. 94-97.

5) Maitri P. Naik, Harshadkumar B. Prajapati and Vipul K. Dabhi, "A Survey on Semantic Document Clustering," IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) pp. 1-10

6) Ashish Dutt, Maizatul Akmar Ismail, and Tutut Herawan, "A Systematic Review on Educational Data Mining," IEEE Access 2017, Vol. 5, pp. 15991-16005.

7) Hussain Ahmad Madni, Zahid anwar and Munam ali Shah, "Data Mining Techniques and Applications – A Decade Review," 23rd International Conference on Automation and Computing (ICAC), pp. 1-7

8) Sivaramakrishnan R Guruvayur and Dr. Suchithra R, "A DETAILED STUDY ON MACHINE LEARNING ECHNIQUES FOR DATA MINING," International Conference on Trends in Electronics and Informatics ICEI

9) Jinwook Seo and Ben Shneiderman, "Knowledge Discovery in High-Dimensional Data: Case Studies and a User Survey for the Rank-by-Feature Framework, "IEEETRANSACTIONSON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 12, NO. 3, pp. 311-322.

10) A.Vinothini and .S.Baghavathi priya, "Survey of Machine Learning Methods for Big Data Applications," 2017 International Conference on Computational Intelligence in Data Science(ICCIDS), VOL. 1

11) Paritosh Nagarnaik and A. Thomas, "Survey on Recommendation System Methods," 2ND INTERNATIONAL CONFERENCE ON ELETRONICS AND COMMUNICATION SYSTEM (ICECS 2015),

12) G. Nizar , C. Michel , B. Nozha, "Unsupervised and Semi-supervised Clustering: a Brief Survey ,"INRIA Rocquencourt , B.P 105, France , pp. 1-12, 2005.

13) Weka: Data mining software in Java, Available on [URL:http://www.cs. waikato.ac.nzlml/weka/]. Accessed on: 30 march., 2018

14) RapidMiner Documentation, Available on [URL:http://rapidminer.com/documentation/], Accessed on: 30 march., 2018

15) Overview of matlab, Available on [URL:http://www.tutorialspoint. com/matlab/matlab_overview.htm], Accessed on: 30 March, 2018

16) Description of stanford tokenizer, Available on [URL:http://nlp. stanford.edu/software/tokenizer.shtml], Accessed on: 30 March, 2018

17) Description of Apache openNLP, Available on [URL:https://opennlp .apache.org/], Accessed on: 2nd Apr., 2018

18) Overview of Apache lucene, Available on [URL:http://lucene. apache.org/], Accessed on: 2nd Apr., 2018

19) Introduction of snowball stemmer, Available on [URL:http:// preciselyconcise.comla pis_and _ installations/snowball_ stemmer. php], Accessed on: 2nd Apr., 2018

20) M. Fionn, C. Pedro, "Methods of Hierarchical Clustering," CSIR , vol.1, pp. 1-21, May 3,2011

21) R. Yogita, Dr. R.Harish , "A Study of Hierarchical Clustering Algorithm," International Jouranl of Information and Computation Technology, ISSN 09742239, vol. 3, pp. 1225-1232, Nov 11, 2013.