International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Survey on Various Clustering Techniques in Data Mining and its Comparative Study for Sundry Data

Nausheen Naaz[1], N. K. Gupta[2]
[1, 2]*Department of Computer Sciences, SHUATS.*

Abstract: ***The amount of data that is being generated in today's world is far more than we can handle or imagine, almost every single interaction leaves a trail that somebody somewhere captures, stores and analyses. This large amount of data has gone beyond human-sense capabilities and its becomes almost impossible to detect any patterns between them, here data mining comes into the picture, it automates a part of detection of interpretable patterns making it easy to handle and use it wherever require according to our need. Clustering algorithms can be used to find natural grouping and exploring data that is generated. It is useful for data preprocessing step to identify homogeneous group of data on which supervised models can be build. In this paper, a review of clustering and its different techniques in data mining along with a comparative studies among clustering algorithms, is covered.***

Keywords: Data Mining, Clustering, K-means, Fuzzy C Means, DBSCAN, Hierarchal Algorithms.

I. INTRODUCTION

A. Data Mining

Data mining refers to extracting or mining knowledge from large amounts of data. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. Data mining involves effective data collection and warehousing as well as computer processing. For segmenting the data and evaluating the probability of future events, data mining uses mathematical algorithms. It is also known as Knowledge Discovery in Data (KDD). 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.

1) The main purpose of data mining process is to discover those records of information and summarize it in a simpler format so it becomes easy to use when its needed again , others are as follows
a) It increases customer loyalty.
b) It identifies hidden profitability.
c) Minimizes clients involvement.
d) Customer satisfaction.
e) Uncovering trends and patterns of markets.
2) The key properties of data mining are as follows
a) Automatic discovery of patterns
b) Prediction of likely outcomes
c) Creation of actionable information

Focus on large datasets and databases

3) The Benefits of Data Mining are
a) Predict future trends, customer purchase habits
b) Help with decision making
c) Improve company revenue and lower costs
d) Market basket analysis
e) Fraud detection
4) The Scope of Data Mining are
a) Data mining process the work in such a manner that it allows business to more proactive to grow substantially.
b) It optimizes large database within the short time and works business intelligence which is more vital to organizational growth.
c) It represents the data in some logical order or may be in pattern to identify the sequential way of processing of data.

d) Brings a genetic way of classification of different sets of data items to view the data in the quick glance.
5) The characteristics of Data Mining are:
a) Prediction of likely outcomes.
b) Focus on large datasets and database.
c) Automatic pattern predictions based on behavior analysis.
To calculate a feature from other features.
6) The Barriers for Data Mining are :
a) User privacy/security.
b) Amount of data is overwhelming.
c) Great cost at implementation stage.
d) Possible misuse of information.
e) Possible in accuracy of data.
7) The six common Tasks involve in data mining are as follows :
a) *Anomaly/Outlier Detection:* The identification of unusual data records, that might be interesting or data errors that require further investigation.
b) *Association Rule Learning:* Searches for relationships between variables. It is for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness.
c) *Clustering:* It is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters).
d) *Classification:* It is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data.
e) *Regression:* It is a data mining technique used to predict a range of numeric value /*continuous values*, given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.
f) *Summarization:* It is a way of providing a more compact representation of the data set, including visualization and report generation

B. Clustering

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity , but are very dissimilar to objects in other clusters. A cluster**[1]** of data objects can be treated as one group. While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups. The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

1) *Requirements of Clustering*
a) *Scalability:* We need highly scalable clustering algorithms to deal with large databases.
b) *Ability To Deal With Different Kinds Of Attributes:* Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
c) *Discovery of Clusters With Attribute Shape:* The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
d) *High Dimensionality:* The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
e) *Ability To Deal With Noisy Data:* Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
f) *Interpretability:* The clustering results should be interpretable, comprehensible, and usable
2) *Benefits of Clustering*
a) Increased resource availability
b) Strategic resource usage
c) Increased performance.
d) Greater scalability
e) Simplified management

3) Types Of Clustering Models

a) *Connectivity Models:* These models are based on the notion that the data points closer in data space exhibit more similarity to each other than the data points lying farther away. These models are very easy to interpret but lacks scalability for handling big datasets. Examples ,Hierarchical clustering algorithm[2]and its variants.

b) *Centroid Models:* These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters. In these models, the no. of clusters required at the end have to be mentioned beforehand, which makes it important to have prior knowledge of the dataset. These models run iteratively to find the local optima. Example, K-Means and FCM clustering algorithm[3].

c) *Distribution Models:* These clustering models are based on the notion of how probable is it that all data points in the cluster belong to the same distribution. These models often suffer from overfitting. Example, Expectation-maximization algorithm[4] which uses multivariate normal distributions.

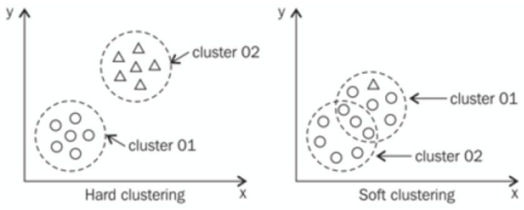
d) *Density Models:* These models search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster. Examples ,DBSCAN and OPTICS Algorithm[5].

4) Types of Clustering

In broad sense, clustering can be divided into two subgroups

a) *Hard Clustering:* In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.

b) *Soft Clustering:* In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.



5) Applications of Clustering

a) Market research, pattern recognition, data analysis, and image processing.

b) In the field of biology, earth observation, geographic location.

c) In classifying documents on the web for information discovery.

d) In security, business intelligence, and Web search.

6) Barriers of Clustering

Clustering is unsupervised process so it tends to segment elements in irrelevant class. So this is the only disadvantage of clustering.

The orthogonal aspects with which clustering methods can be compared are as follow:

a) *Partitioning Criteria:* All the objects are partitioned so that no hierarchy can exists among the clusters. So that, all the clusters are at the same level conceptually. Alternatively, other methods partition data objects hierarchically, where clusters can be formed at different semantic levels.

b) *Separation of Clusters:* Some methods partition data objects into mutually exclusive clusters. In some other situations, the clusters may not be exclusive, that is, a data object may belong to more than one cluster. For example, when clustering documents into topics, a document may be related to multiple topics. Thus, the topics as clusters may not be exclusive.

c) *Similarity Measure:* Some methods determine the similarity between two objects by the distance between them. Such a distance can be defined on Euclidean space,a road network, a vector space, or any other space. In other methods, the similarity may be defined by connectivity based on density or contiguity, and may not rely on the absolute distance between two objects.

d) *Clustering Space:* Many clustering methods search for clusters within the entire given data space. These methods are useful for low-dimensionality data sets. With high dimensional data, however, there can be many irrelevant attributes, which can make similarity measurements unreliable. Consequently, clusters found in the full space are often meaningless.

II. COMPARATIVE STUDY OF VARIOUS CLUSTERING ALGORITHM.

A. Case 1: Banking Dataset

Data mining is becoming strategically important area for many business organizations including banking sector. It is a process of analyzing the data from various perspectives and summarizing it into valuable information so it becomes important to understand which algorithm or techniques works best banking data .In**[6],**in this paper the authors focuses on six types of clustering techniques : k-Means Clustering, Hierarchical Clustering, DBScan clustering, Density Based Clustering, Optics , EM Algorithm and tested each one of them using Weka Clustering Tool on a set of banking data related to customer information. After analyzing the results of the algorithms and running them under different factors and situations. They concluded The performance of K-Means algorithm is better than Hierarchical Clustering algorithm. K-Means algorithm is faster than other clustering algorithm and also produces quality clusters when using huge dataset .

B. Case 2 : Iris Dataset

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. he use of this data set in cluster analysis however is not common, since the data set only contains two clusters with rather obvious separation. In **[9]** , this paper is regarding the comparison of K-Means Clustering and CLARA Clustering on Iris Dataset, which are using Euclidean distance and Manhattan Distance as a dissimilarity measure respectively. They concluded that CLARA (Clustering Large Applications) Clustering using Manhattan distance is better than K-Means Clustering with Euclidean distance. Further in **[7]**, the authors has focuses on a comparative study on K-Means, K-Means++ and Fuzzy C-Means by passing sorted and unsorted data in all three algorithms utilizing MATLAB software tool on iris dataset and concluded that the number of iterations decremented which greatly affected the cluster performance in case there is an increment in the number of data points. Moreover, passing the sorted data in all the three algorithms reduced the elapsed time to a greater extend. As the data points are sorted, the fluctuations of cluster centre is reduced & hence effecting the no. of iterations and time complexity. Additionally the total sum of distance is minimized which further amends the performance.

C. Case 3: Student Dataset

Student Data provides detailed reports all aspects of student-related data including student demographics, Student Assessment results, Completion, graduation, and dropouts, AP and IB, college admissions testing ,reports for graduation, dual credit, and high school to college, and grade-level retention . For a Educational institution data for student as well as teachers and staff are very essential and have great significance so its becomes important to understand which algorithm work better for this dataset. In **[10]** ,in this paper the author compare four clustering Algorithms which are k-Means, k-Medoids, Fuzzy C Means (FCM) and Expectation Maximization (EM) on student dataset. The clustering algorithms was evaluated using execution time, purity and NMI. The result shows that FCM and EM algorithm performs well compared with other two clustering algorithms. In **[12] ,**the authors have collected real dataset containing 666 instances with 11 attributes from the Common Entrance Examination (CEE) data of a particular year for admission to medical colleges of Assam, India conducted by Dibrugarh University in which they have tried to find out the association rules using the data and Various clustering , classification methods were also used to compare the suitable one for the dataset. The data mining tools applied in the educational data were Orange, Weka and R Studio. They concluded that PAM and K-means clustering performs better than hierarchical clustering with silhouette with 0.54 and number of clusters is three.

D. Case 4 : Seed Dataset

Agriculture field is considered to be the backbone of India and also the oldest profession adopted by the mankind. Many issues like pests, unbalanced climate conditions and other agricultural conditions were affecting the agricultural output. Utilizing the recent technologies **[11]** like computers and highly designed equipment in the agriculture, creates the pleasant scenario, upgrading the crop production and also comes forward in aide to one of the oldest inventions of human race. In **[8]** ,in this paper the authors represents a study on different clustering techniques that are incorporated on the seed data sets to enhance the clustering approach based on the various parameters like area, perimeter, compactness, length and width of the kernel, asymmetric coefficient and length of the kernel groove been proposed in which various parameters are passed into K-Means, Fuzzy C-Means, hierarchical and model based clustering algorithms. This cluster based approach using R-Programming is used for framing a merged cultivable holding committed to particular nourishment grains, vegetables, foods grown from the ground cultivation crops.

E. Case 5 : DNA analysis Dataset

Every single strand of DNA consists of 10 sequences of nucleotides**[14].** These sequences cannot be separated or randomly arranged because each sequence of DNA contain certain genomic encoding. RNA- type viruses are able to alter the patterns of infected DNA, which is one way for such a virus to defend itself. In **[13]**,the authors has proposed a new hybrid clustering method that combines K-means, fuzzy C-means, and hierarchical clustering to predict the direction of DNA mutation trends. They have combined these three different approaches in a hybrid clustering method and tested it on two data sets of 1000 isolated positive hepatitis C virus (HCV)-infected and non-infected DNA strands with 37 HCV primers. They concluded that the hybrid clustering method is a combination of three clustering methods that achieves improved performance by exploiting the advantages of these methods for emphasizing three different aspects of data: hierarchical clustering for obvious clustering, K-means clustering for exclusive clustering, and FCM clustering for overlapping clustering.

III. CONCULSION

In this review paper, we have come cross the gaining basic knowledge on data mining and its of one its task named as clustering. We have seen the need of data mining and also different types of clustering models and methods. Apart from this, the paper covers seventeen years of literature review on different researches on comparison and comparative study between different types of clustering algorithms and analyzing which algorithm works better for which kind of data so it becomes easy for us, as a user to better understand the need for these algorithms and also to get knowledge for application of these algorithm to get better and faster results.

REFERENCES

[1] Chris Fraley And Adrian E. Raftery:How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis:The Computer Journal, Vol. 41, No. 8, 1998.
[2] West, J. D., Wesley-Smith, I., &Bergstrom, C. T. (2016). A Recommendation System Based On Hierarchical Clustering Of An Article-Level Citation Network. Ieee Transactions On Big Data, 2(2), 113–123.Doi:10.1109/Tbdata.2016.2541167 .
[3] Nurhayati, T. S. Kania, L. K. Wardhani, N. Hakiem,Busman,HarisMaarif:Big Data Technology For Comparative Study Of Kmeans And Fuzzy C-Means Algorithms Performance:International Conference On Computer And Communication Engineering(Iccce) , 978-1-5386-6992-1/18/$31.00 ©2018 Ieee
[4] Maya R. Gupta And Yihua Chen:Theory And Use Of The Em Algorithm:Foundations And Trendsr In Signal Processing Vol. 4, No. 3 (2010) 223–296.
[5] Hari Krishna Kanagala And Dr. V.V. Jaya Rama Krishnaiah(2016): A Comparative Study Of K-Means, Dbscan And Optics: International Conference On Computer Communication And Informatics (Iccci-2016), Jan. 07 – 09, 2016.
[6] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta (2012) / International Journal Of Engineering Research And Applications (Ijera) Issn: 2248-9622 Www.Ijera.Com Vol. 2, Issue 3, May-Jun 2012, Pp.1379-1384.
[7] Akanksha Kapoor And AbhishekSinghal(2017) : A Comparative Study Of K-Means, K-Means++ And Fuzzy C- Means Clustering Algorithms: 3rd Ieee International Conference On "Computational Intelligence And Communication Technology" (Ieee-Cict2017).
[8] DrMadhaviGudavalli, Vidyasree P, S Viswanadha Raju (2017): Clustering Analysis For Appropriate Crop Prediction Using Hierarchical, Fuzzy C-Means, K-Means And Model Based Techniques: International Journal Of Advance Engineering And Research Development Volume 4, Issue 11, November -2017.
[9] TanviGupta, Supriya P. Panda(2018) : A Comparison Of K-Means Clustering Algorithm And Clara Clustering Algorithm On Iris Dataset: International Journal Of Engineering &Technology, 7 (4) (2018) 4766-4768.
[10] K. Govindasamy, T.Velmurugan(2018) : Analysis Of Student Academic Performance Using Clustering Techniques: International Journal Of Pure And Applied Mathematics Volume 119 No. 15 2018, 309-323
[11] Viswanadharaju, S., Vidyasree, P., Gudavalli, M., &Sekhar, B. C. (2016, November). Minimum Cost Fused Feature Representation AndReconstruction With AutoencoderIn Bimodal Recognition System. In Proceedings Of The International Conference On Big Data And Advanced Wireless Technologies (P. 17). ACM.
[12] SadiqHussain,RashaAtallah, AmirrudinKamsin, And Jiten Hazarika: Classification, Clustering And Association Rule Mining In Educational Datasets Using Data Mining Tools: A Case Study: Springer International Publishing AG, Part Of Springer Nature 2019 R. Silhavy(Ed.): CSOC 2018, AISC 765, Pp. 196–211, 2019.
[13] Berlian Al Kindhi, Tri Arief Sardjono, Mauridhi Hery Purnomo, Gijbertus Jacob Verkerke, Hybrid K-Means, Fuzzy C-Means, And Hierarchical Clustering For DNA Hepatitis C Virus Trend Mutation Analysis, Expert Systems With Applications (2018), Doi: Https://Doi.Org/10.1016/J.Eswa.2018.12.019
[14] Francisco, D., Guillermo, B., Ricardo, L., Antonio, R., Michael, H., And José, L. O. (2014). DNA Clustering And Genome Complexity. Computational Biology AndChemistry, 53:71–78.