Brainwave: A Multidisciplinary Journal (ISSN: 2582-659X), Vol. 1, No. 1, April 2020, pp. 28-35

A Comparative Study of Clustering Models in Data Mining

Prasanjeet Chaakraborty[#*]
itsprasanjeet2016@gmail.com

Sohini Sen[#]
ssohini50@gmail.com

Sourav Maity[#]
souravmaity2533@gmail.com

[#]*Student, Dept. of Computer Science and Engineering, Brainware University, West Bengal, India*
**Corresponding Author*

Abstract

Data is a treasure-house in today's world. Relentlessly huge amount of information is generated by different organizations and people. The key to handle this data is categorizing them into groups, partition or cluster based on some similarity among them. Data mining plays a significant role in potential segmentation, customer segmentation etc. It is a type of unsupervised learning used in data analysis to find some unrevealed pattern hidden in data. Clustering is a mining technique used to position the data element into their appropriate groups. It is the task of partitioning data(objects) into same class. The data are more similar to each other in one class than those present in other cluster. This paper makes a comparative study of different clustering algorithms based on performance, approach, its advantage, disadvantage and examples.

Keywords: Data mining, data clustering ,Soft Clustering, Hard clustering ,k-Mean, Fuzzy C-Mean, Time complexity.

Introduction

Data mining is incredibly popular. Data processing is that the process of finding anomalies, patterns and correlations within large data sets to increase revenues, reduce costs, improve customer relationships, reduce risks and many more . It is also referred to as "knowledge discovery in database"[1-5]. Interest has been generated to surface the importance of knowledge mining. It is important to handle explosively growing data from terabytes to more bytes [6]. It involves data assortment and data accessibility through mechanized information assortment devices, database frameworks. There are significant assets of copious information like [7] –

• Business: Web, O-business, Exchanges, stocks, so forth.

• Science: Remote detecting, Bioinformatics, Scientific reenactment.

• Society and everybody: news, computerized cameras, YouTube.

Data mining is non-trivial (knowledge isn't obvious, it's implicit and it's in-built data) process of extracting interesting previously unknown pattern which brings out the invention and potentially useful pattern or knowledge from huge amount of knowledge. It uses past data rather than existing data [8-10]. The disciplines of knowledge mining are explained through a chart in Figure 1.

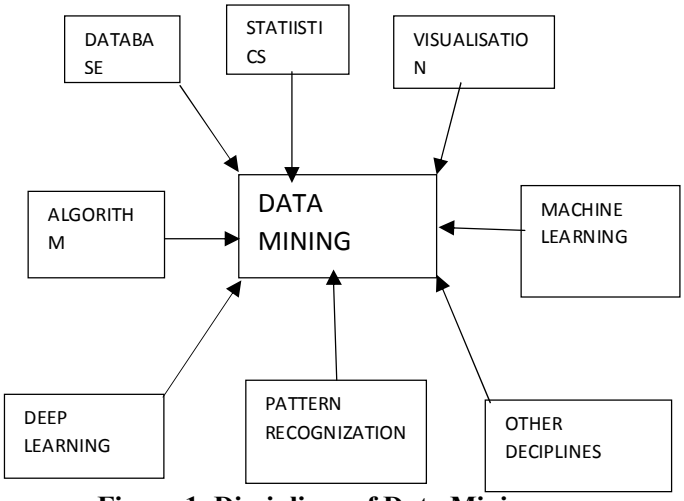


Figure 1. Disciplines of Data Mining

Data mining can be implemented on various datasets viz. Database oriented data streams and sensor data, Times series data, Structure Data, Multilink data, Multimedia Database, Text database, Serial networks and the world wide web. Types of data or pattern that will be discovered as knowledge can be [11]:

- Multidimensional Concept Description: characterization and discrimination. e.g. dry vs. wet regions.
- Frequent Pattern: Correlation us casually. e. g. Shoe and socks.
- Classification or prediction: e.g. Classify countries based on climates or Classify cars based on mileage.

This variation of classifications problem is called expiratory analysis. In cluster analysis, groups are not previously defined. In Cluster analysis [12-14]:

- Class label is unknown.
- Maximizing intra-class similarity and minimizing inter-class similarity.

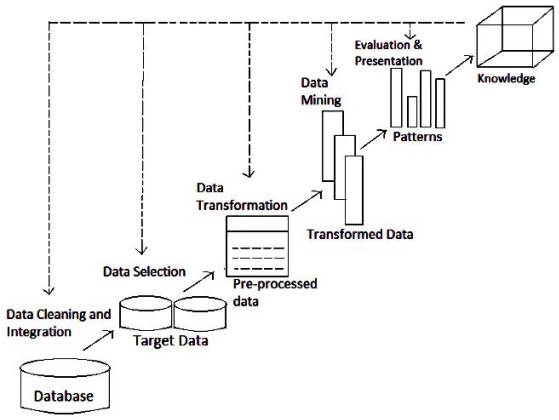


Figure 2. Phases of Data Mining

1.1 Clustering

Clustering analysis [1] has been an emerging research issue in data processing for its various applications [1]. Recently the extensive use of clustering algorithm on image processing, computational biology, mobile communication, medicine and economics has caused the recognition of this algorithm. Main problem with the info clustering algorithm is that it can not be standardized. Many clustering algorithms are proposed at this point. Clustering may be a process in which partitions of a given data set into homogeneous group supported given features, specified similar objects. Similar objects are kept in a group whereas dissimilar objects are in numerous groups[15-16]. It is a kind of unsupervised learning problem. It deals with finding pattern in an exceedingly large collection of unlabeled data. There are some advantages like –

- Scalability: - exceptionally versatile grouping calculation is required to manage huge databases.
- Ability to bargain various properties: - Algorithm must be appropriate for application on any form of information, for instance, interim based, straight out and paired information.
- Attribute Shape: - It must be suited identifying groups of self-assertive shape.
- High dimensionality: - grouping calculation should not exclusively have the choice to handle low-dimensional information yet additionally high dimensional space.

There also are some disadvantages:[2]

- Now days it's comprehensively utilized in statistical surveying, design acknowledgment, information examination and film handling.
- It's likewise utilized for "Location of charge plate misrepresentation".
- It helps within the recognizable proof of house in a very city as indicated by house type, esteem and geographical area.

Partition clustering-

I. Hard clustering
II. Soft clustering

The case for soft clustering techniques is employed to cluster the info into fuzzy sets (a set which has membership values)[4]. During this, each data is related to appropriate membership values. This is often the foremost natural way of clustering. The membership value 0 to 1 is assigned to things . The case of hard computing techniques involves the common method of getting used to K-mean where K is that the number of cluster we use for the dataset. This method applied to investigate the info and treats observation of the info on the situation where object is predicated on and therefore the distance between various computers file input points. During this paper initially the detailed discussion is being tried on each of these two algorithms presented. At the moment comparatively study is completed between them experimentally. The result and the conclusion are employed for necessary comparison.

1.1.1 Clustering Models

- **Connectivity models:** As the name recommends, these models [3] depend on the idea that the information nearer in information space show more likeness to one another than the information lying more remote away [4]. These models can follow two methodologies. In the main

methodology, they start with characterizing all information being focused into independent groups and then accumulating them as the separation diminishes. In the subsequent methodology, all information focuses are delegated a solitary group and afterward divided as the separation increments. These models, however, need versatility for dealing with enormous datasets [6]. Instances of these models are various leveled grouping calculation and its variations.

For Ex- hierarchical algorithm and its variants.

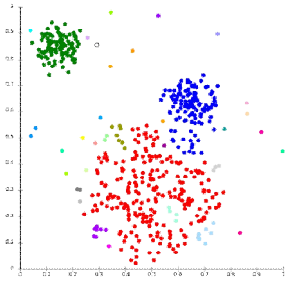


Figure 3: Connectivity-based Clustering

- **Centroid models:** These are iterative bunching calculations in which the idea of comparability is inferred by the closeness of an information point to the centroid of the clusters[5]. K-Means bunching calculation is a famous calculation that falls into this classification. In these models, the number of groups required toward the end must be referenced, which makes it essential to have earlier information on the dataset [9]. These models run iteratively to locate the neighborhood optima.

For Ex- K – means algorithm is one of popular example of this algorithm

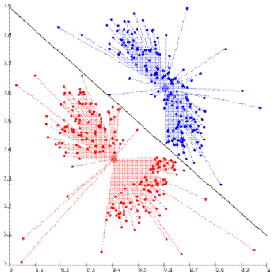


Figure 4: Centroid based Clustering

- **Distribution models**: These grouping models depend on the idea of how plausible is it that all information focusing in the bunch have a place with a similar appropriation (For instance: Normal, Gaussian). These models frequently experience the ill effects of over fitting [6]. A

well-known case of these models is Expectation-expansion calculation which utilizes multivariate typical dispersions.

For Ex- *Expectation-maximization algorithm* which uses multivariate normal distributions is one of popular example of this method.

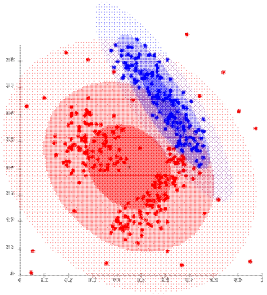


Figure 5

- **Density Models:** These models look at the information space for regions of fluctuated thickness of information focusing into the information space [10]. It separates different diverse thickness locales and allots the information focusing inside these areas in a similar group. Well known instances of thickness models are DBSCAN and OPTICS.

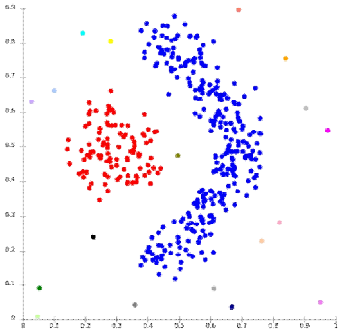
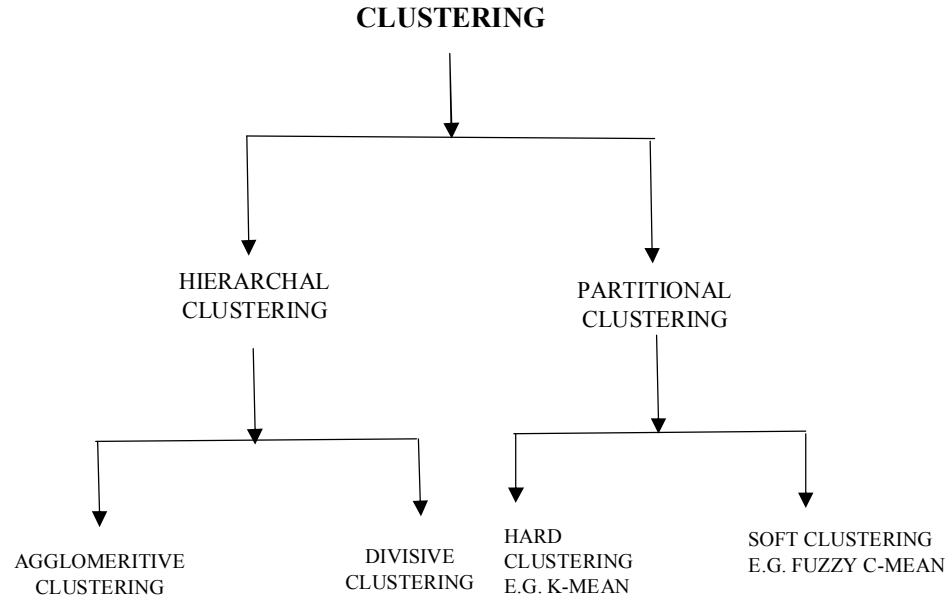


Figure 6: Density based Clustering



- Hierarchal Clustering: - This category is an example of cluster analysis to generate a sequence of nested partition which can be demonstrated as tree or to say the hierarchy of cluster as clusterdendogram[1]. Hierarchal trees can provide a view of data at different level of abstraction. In case the hierarchy is compared with the trees, the lowest level visualizes as a leaves of the trees and highest level as a root [5]. Each point possesses in the leaf node its own cluster whereas the root contains all points in one cluster. The dendrogram can be cut at intermediate level meaningful cluster. It can be divided as follows

  I. Agglomerative clustering: - It is a bottom–up approach. A structure that is more informative than the unstructured set of clusters returned by flat clustering.
  II. Divisive Clustering: - It is a top-down approach. This algorithm also does not require pre specifying the number of clusters [6].

- Partition Clustering:-It is highly dissimilar to hierarchical approach which yields an incremental level of cluster with iterative fusion or divisions, partition clustering assigns a set of object into k-cluster which is not laid down as trees [11]. These clustering techniques algorithms are suitable choice for large data set and these algorithms are computational requirement.

Evaluation of Different Clustering Algorithm:

Clustering model	Example	Time complexity	Advantage	Disadvantage
Centroid	K-mean	$O(Knt_{dist})$ Where, K = number of clusters (centroids) n = number of objects t_{dist}=be the time to calculate the distance between two objects	It is very lucid and can sustain for large amount of data set	Being dependent on initial value
Connectivity	Hierarchical	$O(n^3)$ Where, N=number of data points	It allows to define exactly how each part is connected	Not work well in presence of noise outlier and not scalable
Density	DB-SCAN	O(n*logn)	Work well in presence of noise	Not handle the data point with varying densities
Distribution	Expectation maximization algorithm	O(n*k)	It is always guaranteed that likelihood will increase with each iteration.	It has slow convergence.

Conclusion

Cluster analysis is a very powerful technique throughout the data mining process and is essential for capturing data pattern. This paper compared and analyzed some highly popular clustering algorithms in which some can be scaled and some of the methods work best against noise in the data. This paper describes different clustering algorithm and compares its advantages and drawbacks. Extensive research and study has been carried out in the field of data mining. Real life examples such as Netflix, market basket analysis studies for business giants, biological breakthroughs using complex combinations of various algorithms resulting in hybrids, and subsequent cluster analysis will expose more complex database relationships and categorical analysis in the future.

Reference

[1] "A Detailed Study of Clustering Algorithms", Kamalpreet Bindra[1], AnuranjanMishra[2],
[2] "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms",SoumiGhoshVidyarthi.
[3] "Review based on Data Clustering Algorithms" ,Arpita Nagpal[1], Aman Jatain[2], Deepti Gaur[3]

[4] RuiXu, Donald Wunch, "survey of clustering algorithms", IEEE transactions on neural networks ,vol 16 no.3 may 2005.

[5] AmandeepKaur Mann, and NavneetKaur, "Survey Paper on Clustering Techniques", IJSETR: International Journal of Science,Engineering and Technology Research (ISSN: 22787798), vol. 2,Issue 4, April 2013.

[6] Ma hong ,kangjing,liuxiong "research on clustering algorithms of data streams",ICIME,the 2nd IEEE international conference .2010.

[7] J. Kleinberg, "An impossibility theorem for clustering," in *Proc. 2002Conf. Advances in Neural Information Processing Systems*, vol. 15,2002, pp. 463–470.

[8] Arun K. Pujari,Data mining techniques-a reference book ,pg. no.-114-147.

[8] Miao Guojun, LijunDaun , Wang Shi, ``principal and algorithm of data mining" published in tsinhua university press ,2007

[9] He, Z., Xu, X. and Deng, S., Scalable algorithms for clustering large datasets with mixed type attributes. International Journal of Intelligence Systems.2005 v20. 1077-1089

[10] A. Hinneburg and D. A. Keim.A general approach to clustering in large databases with noise. Knowledge and Information Systems, 5(4):387-415,2003

[11] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2001.

[12] S. Guha, R. Rastogi, and K. Shim. ROCK: a robust clustering algorithm for categorical attributes. I18n Proc. of the 15th Int'l Conf. on Data Eng., 1999.

[13] AbdellahIdrissi,HajarRehioui*An improvement of denclue algorithm for the data clustering*. Information & Communication Technology and Accessibility (ICTA), 2015 5th International Conference. IEEE Xplore 10 march 2016.

[14] Renato Cordeiro de amorium "Asurvey on feature weighting based k-means algorithms" Springer journal ,vol 33 ,issue 2,pp 210-242,july 2016.

[15] Guifenchen,Yuqinyang,hangcheng, "Analysis and research of k-means algorithm in soil fertitlity based on hadoopplatform",Springer, international conference on computer and computing technologies .pp304-312 . 2014

[16] CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling , George Karypis , Eui-Hong Han, Vipin Kumar IEEE Computer 32(8): 68-75, 1999