Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 7, July 2019, pg.152 – 160

Clustering Techniques of Data Mining- A Review

Ranbir Gagat
Ranbirgagat1234@gmail.com
Punjabi University, Patiala, Punjab, India

Guide Asst. Prof. Sikander Singh
Cheemasikander8@gmail.com
Punjabi University, Patiala, Punjab, India

***Abstract:** Data mining is the approach which is applied to extract useful information from the raw data. The technique of clustering, the similar and dissimilar type of data are clustered together to analyze complex data. The previous times, various types of clustering have been proposed for the efficient data analysis. In this paper, density-based clustering and their techniques have been reviewed and compared in terms of various parameters.*

***Keywords:** Hierarchical clustering, partitional clustering, Density-based clustering, Grid – based clustering.*

1. Introduction

Extraction the hidden predictive information from the huge databases is known as data mining. Companies or organizations have been able to focus and retrieve the information from their data warehouses as per the requirement. Data mining has been utilized successfully in the large number of companies .The companies that were involved here at first were mainly the information-intensive industries including the financial services as well as direct mail marketing. Now, a large data warehouse is managed as per the relationships with the users. For the success of data mining, there are two important factors which are; a huge, properly integrated data warehousing and the properly defined understanding of the business process as per which the data mining is applied. It assignment of the objects into specific groups known as clusters in such a manner that the objects of one cluster are more alike than the other objects present in different clusters is known as the clustering technique . This is no particular algorithm with the clustering method. There are certain algorithms which help in performing various tasks within this process. Its algorithms chosen are based on the methods they utilize for computing or identifying the cluster. On the basis of a specific property the gathering of similar image pixels within a cluster is known as clustering technique. Its clusters formed here show high intra-cluster similarities where as low inter-cluster similarities. There are various categories in which the clustering techniques are classified. Categories are explained below:

i) Hierarchical clustering: The basis of the closer relation of pixels which are near to each other than the ones that are far, the clusters are formed, and the technique is known as hierarchical clustering. The basis of the distance of pixels from each other, the clusters are formed within these algorithms. The data representation is done here in the form of a tree. Within the representation and the complete data set is denoted by a root node and data points present in it are represented as leaf node.

- **Agglomerative:** There is a bottom-up approach utilized within these methods where the individual data points are to be started with and further the clusters are merged for creating a tree-like structure. On the basis of the merging of clusters, various choices are possible. The basis of quality and efficiency, the various tradeoffs are provided. There are various examples such as single-linkage, all-pairs linkage, centroid-linkage, and sampled-linkage clustering in which these methods are utilized. There is a utilization of the shortest distance amongst a pair of points within the single-linkage clustering. The average of all pairs is utilized in the all-pairs linkage.
- **Divisive:** The purpose of partitioning the data points into a tree-like structure, a top-down mechanism is utilized within these methods. The partitioning at each step can be done by utilizing any flat clustering algorithm. In the terms of hierarchical structure of the tree as well as the level of balance within various clusters, the divisive partitioning is allows flexibility.

ii) Partitional clustering: The pixels or data points are separated into numerous partitions known as clusters within the partitional clustering algorithms. The data is partitioned into single partition within the partitional clustering instead of representing the data into nested like in hierarchical clustering. The data in which the representation in the form of tree is not possible opts for the partitional clustering.

iii) Density- and Grid-Based Methods

The two closely related classes are the density and grid based techniques. Here, the data space is explored at higher levels of granularity. In terms of number of data points in certain defined volume of its locality or in terms of smother kernel density estimate, the density at a specific point within the data space is defined. At certain level of granularity and the post-processing phase, the data space is explored. The dense regions of the data space are put together within an arbitrary shape. A grid-like structure is formed using the individual regions of the data space within the grid-based techniques of the specific class of density-based methods. As it is easy to put the various dense blocks within the post-processing phase, the grid-based structures are easy to be implemented. Within the high-dimensional methods, those grid-like techniques are also utilized as the lower dimensional grids help in defining the clusters on the subsets of dimensions. The data space is explored at higher level of granularity within these methods which proves to be beneficial. Thus, the complete shape of data distribution is utilized for reconstruction.

The various clustering algorithms available on the base of the various methods as present .A few of these clustering algorithms are explained below:

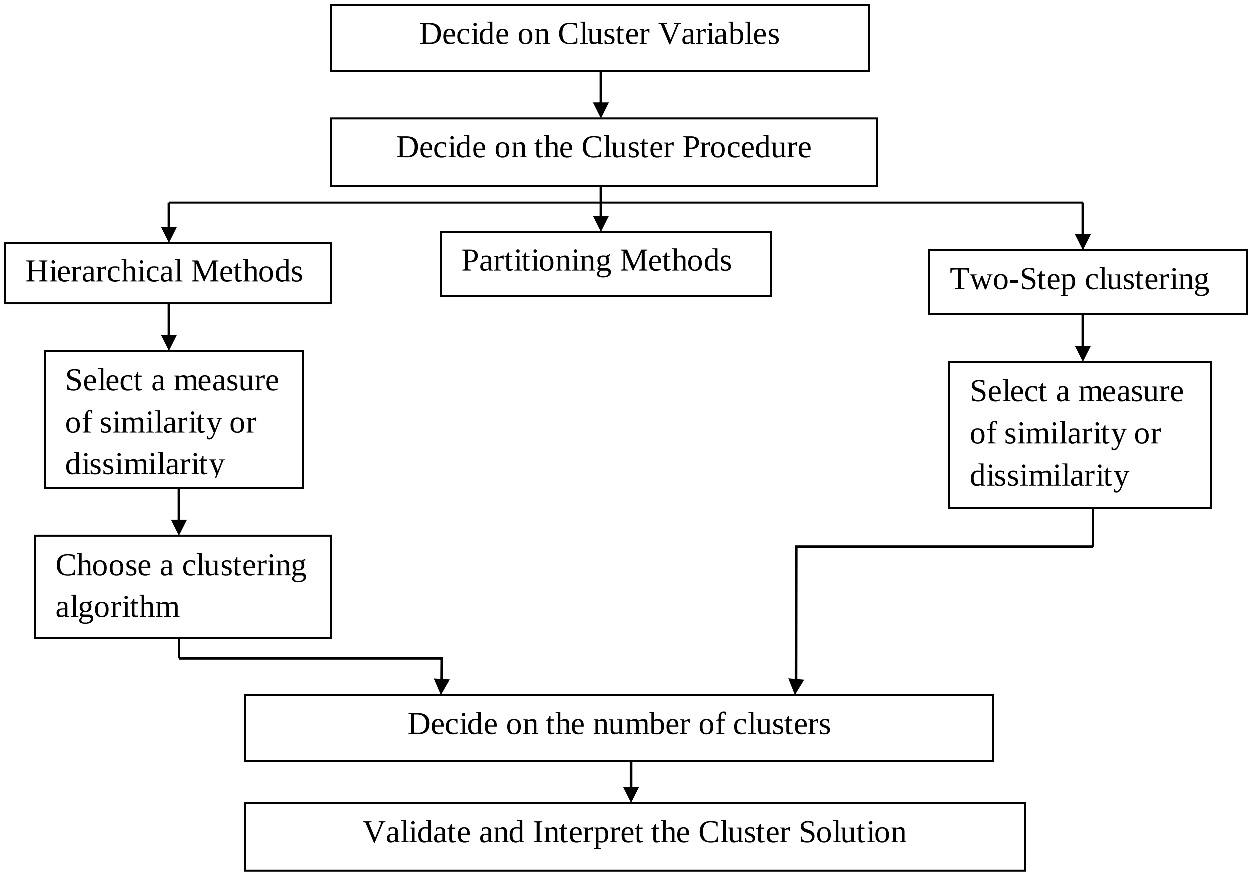
a. K-means Clustering Algorithm: Its process which executes the square error criteria is known as the k-means algorithm. It is very easy as compared to the other algorithms. The numbers of partitions to be provided are initially defined within this algorithm. There is a random initialization of the cluster centers which are required by the predefined number of clusters present. Further, for each data point, one cluster that is nearest is assigned to it for which there is a need to re-estimate the cluster centers as well as the new centroid.

b. N-cut Clustering Algorithm: The hierarchical divisive clustering process in which the tree structure form is utilized for representing the clusters in known as the N-cut technique. The nodes of tree are formatted within the groups or clusters.

c. Mean Shift Clustering Algorithm: It provided data set is clustered by connecting each point with a peak of the data set's probability density. Its process is known as Mean shift algorithm. The spherical window of the radius r at certain data point is utilized for computing the related peak.

1.1 Density-based spatial clustering of applications with noise (DBSCAN):

Various types of clustering techniques developed here are partitioning, hierarchical, density, grid, model, and constraint based. On the basis of notion of density, the density based method works. There is a difference between the clusters formed in thick regions as well as thin regions. The objective here is to increase the recognized clusters until the density in the neighborhood is higher than the threshold value. For the purpose of finding the arbitrary shaped clusters and differentiating the noise from huge spatial databases, the Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is utilized. There are two parameters in this algorithm. It is Epps (radius) and the MinPts (minimum points-a threshold). The basis of center-based approach, this method is based. Here, the density is estimated for a specific point within the dataset.


Decide on Cluster Variables
Decide on the Cluster Procedure
Hierarchical Methods
Partitioning Methods
Two-Step clustering
Select a measure of similarity or dissimilarity
Select a measure of similarity or dissimilarity
Choose a clustering algorithm
Decide on the number of clusters
Validate and Interpret the Cluster Solution

Objective of cluster analysis is to identify groups of objects (in this case, customers) that are very similar with regard to their price consciousness and brand loyalty and assign them into clusters. After the having decided on the clustering variables (brand loyalty and price consciousness), we need to decide on the clustering procedure to form our groups of objects. This step is crucial for the analysis, as different procedures require different decisions prior to analysis.

1. Clustering applications:

- The Intermediate Step for other fundamental data mining problems: The solution to most of the data mining issues for instance classification is done through the summarization of data which is mainly known as the clustering. For various types of application-specific organizations, the less information related to data is helpful.

- The Collaborative Filtering: The summarization of closely related users is done through the collaborative filtering techniques. The collaborative filtering is done using the ratings which are given by the various users towards each other. This helps in providing certain recommendations as per the requirements to enhance them.

- The Customer Segmentation: The collaborative filtering is similar to this method as there are groups which involve similar clusters within the data. The only difference here is that the arbitrary attributes related to the objects are utilized here for clustering rather than the rating information.

- The Data Summarization: There are various dimensionality reduction methods which provide the clustering techniques. These techniques help in providing data summarization which further helps in providing compact data representations. These representations help in providing usage in various applications which is easier.

- The Dynamic Trend Detection: There are various dynamic as well as streaming algorithms which are utilized in order to detect data in various applications which involve dynamically clustered data. Various patterns of changes are performed here. For instance, the multidimensional data, text streams, trajectory data, etc. With the help of clustering methods, the key trends as well as events in data are identified.

- The Multimedia Data Analysis: The multimedia data involves the images, audio, video and various types of documents. There are huge applications such as recognition of similar snippets of music, or pictures are involved for the recognition of similar segments. There are various types of data and it might also involve the multimodal representation in various instances.

- The Biological Data Analysis: Due to the evolvement of human genome as well as various kinds of gene expression data, the biological data is very important. The sequences or networks can be formed for the purpose of structuring the biological data. Better ideas for providing new trends related to data are done using the clustering algorithms.

- The Social Network Analysis: For determining the important communities within the network, the structure of social network is utilized. Within the community detection there is a better understanding of the community structure within the network, which helps to introduce it in the social network analysis. The social network summarization also utilizes the clustering technique which is used in various applications. There are also applications related to clustering within the social network summarization.

2. Literature Review

Research of the client classification and prediction in commercial banks based on Naive Bayesian classifier is proposed in this paper that accommodates the uncertainty inherent in predicting client conduct. With the wild rivalry in the domestic and global business, the Customer Relationship Management (CRM) has turned out to be one of matters of concern to the enterprise. The study will help the company to break down and forecast client's pattern of consumption, and the premise of personalized marketing services and management. A study on the speech articulation trouble symptoms of PD influenced people and attempt is proposed in this paper .To formulate the model on the behalf of three data mining methods. It is concluded that LR model indentified people with PD all the more effectively then Tree and SVM classifiers on the behalf of discussed performance matrices. An evolutionary approach is shown in this paper. For extracting a model of flood prediction from hydrological data observed timely on water heights in a river watershed. An evolutionary algorithm is involved to permit choosing the best sets, juries of classifiers, of such variables as predictive variables. The experiments have demonstrated that this difference could be important, either with our stochastic method or with classical classification methods. The classifiers utilized for the automatic detection of the disease are evaluated in this paper. Utilizing the data mining methods. This work particularly puts concentrate on the classification techniques to accurately order the disease associated with the retina based on the features extracted from retinal images through image processing techniques. A training accuracy of cent percent is accomplished by a couple of classifiers while the prediction accuracy remains at 76.67. The results demonstrate that a training accuracy of 100% can be accomplished by a couple of classifiers and a prediction accuracy 76.67%. A framework is proposed in this paper of methodology of DBSCAN algorithm with the integration of fuzzy logic. The extent to which an object has a place with a particular cluster will be resolved utilizing membership values. The improved version of DBSCAN algorithm will be the hybridization of DBSCAN algorithm with fuzzy if-then rules.The continuous image super-pixel segmentation method is proposed in this paper with 50fps by utilizing the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm. A quick two-stage framework is adopted. The experimental results demonstrate that our constant super-pixel algorithm (50fps) by the DBSCAN clustering outperforms the state-of-the-craftsmanship super-pixel segmentation methods as far as both accuracy and efficiency. An efficient approach for clustering analysis is proposed in this paper to detect embedded and nested adjacent clusters utilizing idea of density based notion of clusters and neighborhood difference. This experimental result that recommended that proposed algorithm is more effective in detecting embedded and nested adjacent clusters thought about both DBSCAN and ENDBSCAN without including any additional computational complexity.

The Comparison of various existing techniques is given below:

Table 1: Author Analysis

Authors' Names	Year	Description	Outcome
Gao Hua	2011	The research of the client classification and prediction in commercial banks based on Naive Bayesian classifier is proposed in this paper that accommodates the uncertainty inherent in predicting client conduct.	Its study will help the company to break down and forecast client's pattern of consumption, and the premise of personalized marketing services and management.
Geeta Yadav, Yugal Kumar, G. Sahoo	2012	The study on the speech articulation trouble symptoms of PD influenced people and attempt is proposed here to formulate the model on the behalf of three data mining methods.	It is concluded that LR model indentified people with PD all the more effectively then Tree and SVM classifiers on the behalf of discussed performance matrices.
Wilfried Segretier, Manuel Clergue, Martine Collard, Luis Izquierdo	2012	Evolutionary approach is shown in this paper for extracting a model of flood prediction from hydrological data observed timely on water heights in a river watershed.	Its experiments have demonstrated that this difference could be important, either with our stochastic method or with classical classification methods.
Dr.R. Geetha Ramani, Lakshmi.B, Shomona Gracia Jacob	2012	Its classifiers utilized for the automatic detection of the disease are evaluated utilizing the data mining methods.	Its results demonstrate that a training accuracy of 100% can be accomplished by a couple of classifiers and a prediction accuracy of 76.67%.
Saefia Beri, Kamaljit Kaur	2015	The framework of methodology of DBSCAN algorithm is proposed in this paper with the integration of fuzzy logic.	The improved version of DBSCAN algorithm will be the hybridization of DBSCAN algorithm with fuzzy if-then rules.
Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao	2016	Its continuous image super-pixel segmentation method is proposed in this paper with 50fps by utilizing the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm.	The experimental results demonstrate that our constant super-pixel algorithm (50fps) by the DBSCAN clustering outperforms the state-of-the-craftsmanship super-pixel segmentation methods as far as both accuracy and efficiency.
Nagaraju S, Manish Kashyap	2016	The efficient approach for clustering analysis is present in this paper to detect embedded and nested adjacent clusters utilizing idea of density based notion of clusters and neighborhood difference.	Experimental results that recommended that proposed algorithm is more effective in detecting embedded and nested adjacent clusters thought about both DBSCAN and ENDBSCAN without including any additional computational complexity.

3. Conclusion

Work, it has been concluded that density-based clustering is the most efficient type of clustering to analyze similar type of data. The DBSCAN is the algorithm of density based clustering which is applied to cluster the input data according to their density. In future, the DBSCAN algorithm will be improved to increase accuracy of clustering and reduce execution time.

References

[1] F. Messina, G. Pappalardo, D. Rosaci, C. Santoro, and G. Sarne, “A trust-aware, self-organizing system for large-scale federations of utility computing infrastructures,” 2016, Future Generation Computer Systems, vol. 56, pp. 77–94.

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” 1996, 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231.

[3] C. P. McQuellin, H. F. Jelinek, and G. Joss, “Characterisation of fluorescein angiograms of retinal fundus using mathematical morphology: a pilot study,” 2002, 5th International Conference on Ophthalmic Photography, Adelaide, p. 152.

[4] T. Y. Wong, W. Rosamond, P. P. Chang, D. J. Couper, A. R. Sharrett, L. D. Hubbard, A. R. Folsom, and R. Klein, “Retinopathy and risk of congestive heart failure,” 2005, Journal of the American Medical Association,vol. 293, no. 1, pp. 63–69.

[5] M.M. Fraz, P. Remagnino, A. Hoppe, B. Uyyanonvara, A.R. Rudnicka, C.G. Owen and S.A. Barman “Blood vessel segmentation methodologies in retinal images – A survey,” 2012, computer methods and programs in biomedicine.

[6] K. Buhler, P. Felkel and A.L. Cruz, “Geometric methods for vessel visualization and quantification – a survey,” 2003, Geometric Modelling for Scientific Visualization, pp. 399–421.

[7] Huan Yu, Wenhui Zhang,” DBSCAN Data Clustering Algorithm for Video Stabilizing System”, 2013, International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC).

[8] R. Giunta, F. Messina, G. Pappalardo, and E. Tramontana, “Providing qos strategies and cloud-integration to web servers by means of aspects,” 2015, Concurrency and Computation: Practice and Experience, vol. 27, no. 6, pp. 1498–1512.

[9] Gao Hua,” Customer Relationship Management Based on Data Mining Technique-Naive Bayesian classifier”, 2011, IEEE.

[10] Geeta Yadav, Yugal Kumar, G. Sahoo," Predication of Parkinson's disease using Data Mining Methods: a comparative analysis of tree, statistical and support vector machine classifiers", 2012 National Conference on Computing and Communication Systems (NCCCS).

[11] Wilfried Segretier, Manuel Clergue, Martine Collard, Luis Izquierdo," 2012, WCCIIEEE World Congress on Computational Intelligence.

[12] Dr.R. Geetha Ramani, Lakshmi.B, Shomona Gracia Jacob," Data Mining Method of Evaluating Classifier Prediction Accuracy in Retinal Data", 2012, IEEE.

[13] Saefia Beri, Kamaljit Kaur," Hybrid Framework for DBSCAN Algorithm Using Fuzzy Logic", 2015 1st International conference on futuristic trend in computational analysis and knowledge management (ABLAZE).

[14] Jianbing Shen, Xiaopeng Hao, Zhiyuan Liang, Yu Liu, Wenguan Wang, and Ling Shao," Real-time Superpixel Segmentation by DBSCAN Clustering Algorithm", 2016, IEEE.

[15] Nagaraju S, Manish Kashyap," A Variant of DBSCAN Algorithm to Find Embedded and Nested Adjacent Clusters", 2016, 3rd International Conference on Signal Processing and Integrated Networks (SPIN).