International Journal of Current Trends in Engineering & Technology
Received 7 June 2019, Received in revised form, 22 July 2019, Accepted 29 July 2019, Available online 7 August 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 04 (July- August, 2019)

Improved Performance of Dataset Classification Using K-Means Clustering Method and PVFCA

Tejaswini Priya[1], Surendra Chadokar[2]
Department of CSE, LNCTS, Bhopal, India
[1]priyatejaswini7@gmail.com, [2]surendrachadokar1984@gmail.com

***Abstract**: Data mining has created an excellent progress in recent year but the matter of missing data has remained an excellent challenge for processing algorithms. It's an activity of extracting some useful information from an enormous information or data base, by utilization any of its techniques. Processing is functioning to search out information out of data and presenting it throughout a kind that is simply understood to humans. data processing is that the notion of all ways that and techniques which allow analyzing very large info sets to extract and resolve previously unknown structures and relations out of such giant a lot of details. In studied of information mining applications in health care. Particularly, it discusses data processing and its applications among health care in major areas like the analysis of treatment effectiveness, management of health care data processing has been used intensively and extensively by several organizations. In health care, data processing is changing into progressively standard, if not more and more essential. Data processing applications will greatly profit all parties concerned within the health care business and therefore the class evolution and cluster techniques on the concept of algorithms that's want to predict antecedently unknown category of objects. Varied efforts are created to reinforce the performance of the K-means cluster formula. Throughout this paper we've been briefed among the sort of a review the work administered by the various researchers' exploitation K-means cluster. They have mentioned the restrictions and applications of the K-means cluster formula additionally. Determination these issues is that the topic of the numerous recent analysis works. They'll be doing a review on k-means cluster algorithms. To explain the objects a lot of exactly it has to be outlined by all the attainable and purposeful size of clusters. Projected vector performs cluster rule to decrease the amount of clusters information and will increase finding similarity kinds of information in clusters. Therefore discovering the distribution pattern of information becomes tough using projected vector perform cluster algorithmic program. This has necessitated searching for significant cluster in dataset analysis. Projected vector performs cluster algorithmic program to decrease error in cluster. Implementation using Mat laboratory tool and overcome challenges in varied application of information mining and implementation of cluster methods.*

***Keywords:** Data Mining, Classification, Clustering, Data Mining Techniques, K-mean methodology, Intelligent Data Analysis.*

I. INTRODUCTION

Data mining is that the method of extracting patterns from information. Data processing is seen as associate more and more necessary tool by modern business to rework information into business intelligence giving associate informational advantage. It's presently utilized in a good vary of identification practices, like promoting, police work, fraud detection, and scientific discovery. A primary reason for victimization data processing is to help within the analysis of collections of observations of behavior. Associate inescapable reality of information mining is that the sub-sets of data or information being analyzed might not be representative of the full domain, and thus might not contain samples of bound vital relationships and behaviors that exist across other parts of the domain. The aim of information mining technique is to mine information from an oversized data set and build over it into a reasonable sort for supplementary purpose. Processing is in addition known as the data discovery in databases (KDD). Technically, information is that the methods of finding patterns among vary of fields in massive electronic data service. It's the foremost effective technique to differentiate between data and information. Processing consists of extract, transform, and payload dealing data onto the data warehouse system, Store and manages data rations the data during a very information system. However data cannot be used directly. Its real value is anticipated by extracting information useful for decision support. In most areas, data analysis was historically a manual technique. Once the dimensions of data manipulation and exploration go on the way facet human capabilities, people explore for computing technologies to change the strategy. Information is method of extraction, transformation and loading of knowledge to from information or warehouse system. Storing and managing data provide access to data analyst [1]. The goal of cluster is to cluster information points that are close (or similar) to every alternative establish such groupings (or clusters) in an unsupervised manner. Varied definitions of a cluster will be developed, reckoning on the target of cluster. Generally, one could settle for the read that a cluster may be a cluster of objects that are a lot of kind of like one another each alternative than to members of other clusters. The term "similarity" ought to be understood as mathematical similarity, measured in some well-defined sense. In metric areas, similarity is usually outlined by means that of a distance norm. Distance will be measured among the information vectors themselves, or as a distance from a knowledge

International Journal of Current Trends in Engineering & Technology
Received 7 June 2019, Received in revised form, 22 July 2019, Accepted 29 July 2019, Available online 7 August 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 04 (July- August, 2019)

vector to some prototypal object (prototype) of the cluster. The prototypes are typically not better-known beforehand, and are wanted by the cluster algorithms at the same time with the partitioning of the information. The prototypes could also be vectors of an equivalent dimension because the information objects; however they'll even be outlined as "higher-level" geometrical objects, like linear or nonlinear subspaces or functions. The concept of information grouping, or cluster, is easy in its nature and is about to the human approach of thinking; whenever we are conferred with an oversized quantity of information, we tend to typically tend to summarize this large range of information into a little range of teams or classes so as to further facilitate its analysis. Cluster analysis is done by finding similarities between information in line with the characteristics found within the information and grouping similar information objects into clusters. Cluster is AN unsupervised learning method. Cluster is beneficial in many searching pattern analysis grouping deciding and machine learning things together with data processing, document retrieval image segmentation and pattern classification. The term cluster is used in many analysis communities to explain ways for grouping of unlabelled information. These communities have completely different terminologies and assumptions for the parts of the cluster method and also the contexts during which cluster are used. Samples of cluster are crusty and cluster genes, market segmentation. Typical pattern cluster activity involves the subsequent steps pattern illustration, Definition of pattern proximity live acceptable to the information domain, Clustering or grouping information abstraction if required and assessment of output if required. Pattern illustration refers to the quantity of categories the quantity out there of accessible patterns and also the number sorts and scale of the options available to the cluster algorithmic rule Feature choice is that the method of distinguishing the foremost effective set of the initial. Feature extraction is that the use of 1 or additional transformations of the input options to supply new salient options. Either or each of those techniques is accustomed acquire an acceptable set of options to use in cluster. Pattern proximity is typically measured by a distance operate outlined on pairs of patterns cluster. Information abstraction is that the method of extracting an easy and compact illustration of information set. Within the cluster context a typical information abstraction could be a compact description of every cluster sometimes in terms of cluster prototypes or representative patterns like the centroid[2].K-Means cluster K-means cluster is most generally used cluster rule that is employed in several areas like data retrieval, pc vision and pattern recognition. K-means cluster assigns n information points into k clusters in order that similar information points are classified along. It's a reiterative methodology that assigns every purpose to the cluster whose centroid is that the nearest. Then it once more calculates the centroid of those teams by taking its average. The rule the essential approach of K-means cluster [4].

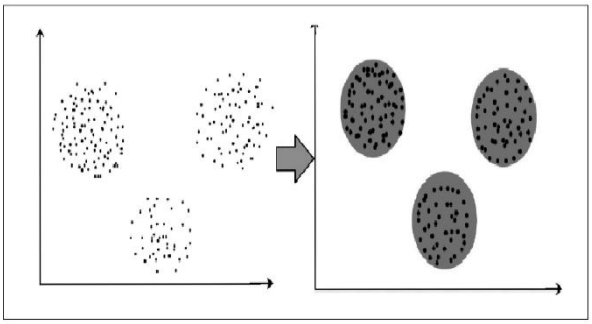

Figure1 clustering processes get three clusters

Graphical illustration for operating of K-means rule. Within the start there are 2 sets of objects. Then the centroids of each set are determined. In keeping with the centroid once more the clusters are shaped that gave the various clusters of dataset. This method repeats till the simplest clusters are achieved K-means bunch technique is wide used cluster rule, which is most well-liked bunch rule that's utilized in scientific and industrial applications. it's a technique of cluster analysis that is employed to partition N objects into k clusters in such the simplest way that every object belongs to the cluster with the closest mean. The normal Means rule is extremely easy

1. Choose the worth of K i.e. Initial centroids.
2. Repeat step three and four for all information points in dataset.
3. Realize the closest purpose from those centroids within the Dataset.
4. Kind K clusters by assignment every purpose to its nighest centroid.
5. Calculate the new world centroid for every cluster.

Properties of k-means rule

1. efficient whereas process giant information set.
2. It works only on numeric values.
3. The shapes of clusters are lent form.

K-means is that the most typically used partitioning rule in cluster analysis owing to its simplicity and performance. However it's some restrictions once coping with terribly giant datasets owing to high machine quality, sensitive to outliers and its results depends on initial centroids that are hand-picked at random. Several solutions are planned to enhance the performance of K-Means. However nobody gives a world resolution. A number of planned algorithms are quick however they fail to keep up the standard of clusters. Some generate clusters of excellent quality however they're terribly pricey in term of machine quality. The outliers are major drawback which will result on quality of clusters. Some rule only works on only numerical datasets [5].

II.RELATED WORK

Yan Zhu et al. [6] has proposed a new method in which clustering initialization has been done using clustering

exemplars produced by affinity propagation. They have also minimized the total squared error g the clusters.

S.Poonkuzhali et al. [7] propose a framework for an effective retrieval of medical records using data mining techniques. Their work focuses on retrieval of updated, accurate and relevant information from Medline datasets using Machine Learning approach. The proposed work uses keyword searching algorithm for extracting relevant information from Medline datasets and K-Nearest Neighbor algorithm (KNN) to get the relation between disease and treatment

G. Liu et al. [8] has presented a general K-means clustering to identify natural clusters in datasets. They have also shown high accuracy in their results

S. K. Wasan et al. [9] examine the impact of data mining techniques, including artificial neural networks, on medical diagnostics. They identify a few areas of healthcare where data mining and statistics can be applied to healthcare databases for knowledge discovery.

A K Dogra et al. [10] Data mining has made a great progress in recent year but the problem of missing data has remained a great challenge for data mining algorithms. It is an activity of extracting some useful knowledge from a large data base, by using any of its techniques. Data mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. Data mining is the notion of all methods and techniques which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. This paper studied the classification and clustering techniques on the basis of algorithms which is used to predict previously unknown class of objects.

K. B. Sawan et al. [11] existing K-means clustering algorithm has a number of drawbacks. The selection of initial starting point will have effect on the results of number of clusters formed and their new centroids. Overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has been proposed in the paper along with the modified K-Means algorithm to overcome the deficiency of the classical K-means clustering algorithm. The new method is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm. The improved version of the algorithm uses a systematic way to find initial centroid points which reduces the number of dataset scans and will produce better accuracy in a smaller number of iteration with the traditional algorithm. The method could be computationally expensive if used with large data sets because it requires calculating the distance of every point with the first point of the given dataset as a very first step of the algorithm and sort it based on this distance. However this drawback could be taken care by using multi-threading technique while implementing it within the program. However further research is required to verify the capability of this method when applied to data sets with more complex object distributions.

Junatao Wang et al. [12] propose an improved means algorithm using noise data filter in this paper. The shortcomings of the traditional k-means clustering algorithm are overcome by this proposed algorithm. The algorithm develops density-based detection methods based on characteristics of noise data where the discovery and processing steps of the noise data are added to the original algorithm. By pre-processing the data to exclude these noise data before clustering data sets the cluster cohesion of the clustering results is improved significantly and the impact of noise data on k-means algorithm is decreased effectively and the clustering results are more accurate

D. T. Pham et al [13] has worked on the number of k used in K-means clustering. They have concluded different number of clusters for different datasets.

M. P. Sebastian et al. [14] proposes k-means algorithm, for different sets of values of initial centroids, produces different clusters. Final cluster quality in algorithm depends on the selection of initial centroids. Two phases include in original k means algorithm: first for determining initial centroids and second for assigning data points to the nearest clusters and then recalculating the clustering mean

G. Sahoo et al. [15] focused on K-Means initialization problems. The K-Means initialization problem of algorithm is formulated by two ways; first, how many numbers of clusters required for clustering and second, how to initialize initial centers for clusters of K-Means algorithm. This paper covers the solution for of the initialization problem of initial cluster centers. For that, a binary search initialization method is used to initialize the initial cluster points i.e. initial centroid for K-Means algorithm Performance of algorithm evaluated using UCI repository datasets.

III. Simulation Environment MATLAB

The Performance analysis of MATLAB version 2012 i.e. used for this thesis Implementation of information mining provides processor optimized libraries for quick execution and computation and performed on input cancer dataset. It uses its JIT (just in time) compilation technology to supply execution speeds that rival traditional programming languages. It may also additional advantage of multi core and digital computer computers,

International Journal of Current Trends in Engineering & Technology
Received 7 June 2019, Received in revised form, 22 July 2019, Accepted 29 July 2019, Available online 7 August 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 04 (July- August, 2019)

MATLAB give several multi threaded algebra and numerical operate. These functions automatically execute on multiple process thread during a single MATLAB, to execute quicker on multicore computers. During this thesis, all increased efficient information retrieve results were performed in MATLAB (R2012a). MATLAB is that the high-level language and interactive environment utilized by a lot of engineers and scientists worldwide. It lets the explore and visualize concepts and collaborate across totally different disciplines with signal and image process, communication and computation of results. MATLAB provides tools to accumulate, analyze, and visualize information, modify you to induce insight into your information during a division of the time it'd take exploitation spreadsheets or traditional programming languages. It may also document and share the results through plots and reports or as printed MATLAB code.

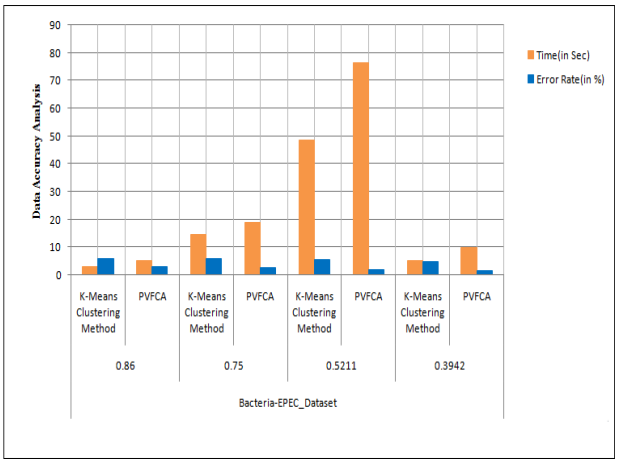

Figure 2 Results analysis based on Bacteria-EPEC Dataset

IV. Results Analysis

In research in area of data mining-based clustering method using improve performance of dataset classification method using k-means clustering and find minimum error of medical health care dataset analysis and exceptional best solution.

(a) Results analysis based on bacteria-EPEC dataset used and set random values like (0.86, 0.75, 0.5211, 0.3942).K-means clustering method more error rate but PVFCA less error. K-means clustering method time take minim as compare to PVFCA. K-Means clustering method and PVFCA compare analysis. Our proposed method error minimization more and average time. Overall our proposed method gets best output. Results graph analysis based on bacteria-EPEC dataset show in figure2 above, PVFCA is improved as compare to k-means clustering method ,here k-means clustering method copy data or redundancy is more but PVFCA remove copy data then gets minimum redundancy.

(b) Results analysis based on Cream-yeast dataset used and set random values like (0.681, 0.816, .0251, 0.754).K-means clustering method more error rate but PVFCA less error. K-means clustering method time take minim as compare to PVFCA. K-Means clustering method and PVFCA compare analysis. Our proposed method error minimization more and average time. Overall our proposed method gets best output. Results graph analysis based on Cream-yeast dataset show in figure3 above, PVFCA is improved as compare to k-means clustering method ,here k-means clustering method copy data or redundancy is more but PVFCA remove copy data then gets minimum redundancy.

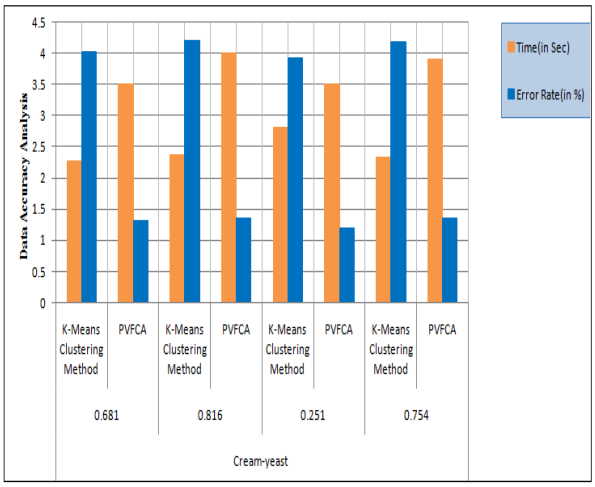

Figure 3 Results analysis based on Cream-yeast Dataset

IV. CONCLUSION

In Improved performance of dataset classification victimization k-means cluster technique and planned vector perform cluster algorithmic program. In study area of data mining aim at the matter of the data classification methodology of the classical K-means rule, this paper proposes the plan of action of optimizing the initial cluster center to enhance the K-means rule, and exploitation the genetic rule to clean the data. The experimental results show that the projected methodology is further correct than the classical data classification methodology. Analyses show that it's very robust to decision one processing rule as a result of the simplest suited to the identification and/or prognosis of all diseases. Reckoning on concrete things, someday some algorithms perform on top of others, but there is a unit cases once a mixture of the only properties of variety of equivalent algorithms results less complicated. Knowledge analysis, they have created an analysis on work distributed by fully completely different researcher's exploitation K-means cluster approach. They in addition mentioned the evolution, limitations and applications of K-means cluster rule. It's determined that much improvement has been created to the operative of K-means rule inside the past years. Most work distributed on the event of potency and accuracy of the clusters. This field is sometimes open for

Received 7 June 2019, Received in revised form, 22 July 2019, Accepted 29 July 2019, Available online 7 August 2019
www.ijctet.org, ISSN: 2395-3152
Volume: 05, Issue: 04 (July- August, 2019)

enhancements. Setting applicable initial kind of clusters is sometimes a tough task. At the tip it's everywhere that although there has been created millions of work on K-means cluster approach. These approaches are with success applied in several areas. Knowledge analysis must compare these completely different techniques and higher perceive their strengths and limitations. a selected technique is appropriate for a selected distribution of information. Information analysis cannot expect that one variety of cluster approach is appropriate for all sorts of information or maybe for all high dimensional data. Several problems like quantifiability to massive information sets, independence of order of input, corroborative cluster result area unit resolved to a lot of extent. Information analysis must specialize in strategies which might provide USA lead to a way that is simple to interpret. Result obtained ought to be in a very manner which might additionally provide U.S.A. some conclusion and information concerning data distribution. It ought to any recommend USA on however the clusters obtained is useful for varied applications. K-mean information average error rate rely on incorrect information in clusters however PVFCA is minimum average error rate rely on correct information in clusters. Our planned technique (PVFCA) overcome error information and additionally known as best solution.

REFERENCES

[1]. E. A. Khadem, E. F. Nezhad, M. Sharifi, "Data Mining: Methods & Utilities", Researcher,5(12):47-59. (ISSN: 1553-9865), 2013.

[2]. Fayyad, U. M. , Piatetsky-Shapiro, G., Smyth, P., Uthurusamy , R. G. R.: Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, Menlo Park, CA.,1996.

[3]. E. Kijsipongse, S. U-ruekolan, "Dynamic load balancing on GPU clusters for large-scale K-Means clustering, " 2012 IEEE International Joint Conference on Computer Science and Software Engineering (JCSSE), vol., no., pp.346, 350, May 30 2012-June 1 ,2012.

[4]. M. Li and al. "An improved k-means algorithm based on Map reduce and Grid", International Journal of Grid Distribution Computing, 2015.

[5]. Saroj, Tripti Chaudhary, "Study on Various Clustering Techniques", International Journal of Computer Science and Information Technologies, Volume 6, Issue 3, 2015.

[6]. Y. Zhu, J. Yu, C. Jia, "Initializing K-means Clustering Using Affinity Propagation, " Ninth International Conference on Hybrid Intelligent Systems, 2009. HIS '09. vol.1, no., pp.338, 343, 12-14 Aug. 2009.

[7]. S. Poonkuzhali, T. Sakthimurugan, An Effective Retrieval of Medical Records using Data Mining Techniques, International Journal Of Pharmaceutical Science And Health Care. ISSN: 2249-5738. 2(2), pp 72-78, 2012.

[8]. G. Liu ,Y. Sun; K. Xu, "A k-Means-Based Projected Clustering Algorithm, " 2010 Third International Joint Conference on Computational Science and Optimization (CSO), vol.1, no., pp.466, 470, 28-31 May 2010.

[9]. S.K. Wasan, V. Bhatnagar , H. Kaur, The Impact Of Data Mining Techniques On Medical Diagnostics, Data Science Journal, Volume 5, pp. 119-126, 2006.

[10]. A K Dogra, TanujWala, "A Review Paper on Data Mining Techniques and Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4 Issue 5, May 2015.

[11]. Kedar B. Sawant, "Efficient Determination of Clusters in K-Mean Algorithm Using Neighborhood Distance "International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015.

[12]. Junatao Wang, Xiaolong Su, An Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 may, (pp. 44-46), 2011.

[13]. D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of K in K-means clustering", Proc. ImechE Vol. 219 Part C: J. Mechanical Engineering Science, IMechE 2005.

[14]. M. P. Sebastian, K. A. Abdul Nazeer, ,"Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering Vol I WCE 2009, July 1 - 3, 2009, London, U, 2009.

[15]. Yugal Kumar and G. Sahoo, "A New Initialization Method to Originate Initial Cluster Centers for K-Means Algorithm", International Journal of Advanced Science and Technology Vol.62, 2014.

[16]. Koh, Hian Chye, and Gerald Tan. "Data mining applications in healthcare." Journal of healthcare information management19, no. 2 (2011): 65.

[17]. Kesavaraj, Gopalan, and Sreekumar Sukumaran. "A study on classification techniques in data mining." In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pp. 1-7. IEEE, 2013.

[18]. Maulik, Ujjwal, and Sangha mitra Bandyopadhyay. "Performance evaluation of some clustering algorithms and validity indices." IEEE Transactions on pattern analysis and machine intelligence 24, no. 12 :1650-1654, 2002.

[19]. artigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28, no. 1: 100-108, 1979.

[20]. Kanungo, Tapas, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE

Transactions on Pattern Analysis & Machine Intelligence 7: 881-892, 2002

[21]. Arthur, David, and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding." In Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms, pp. 1027-1035. Society for Industrial and Applied Mathematics, 2007.

[22]. Vora, Pritesh, and Bhavesh Oza. "A survey on k-mean clustering and particle swarm optimization." International Journal of Science and Modern Engineering 1, no. 3 : 24-26, 2013.

[23]. Park, Hae-Sang, and Chi-Hyuck Jun. "A simple and fast algorithm for K-medoids clustering." Expert systems with applications 36, no. 2 : 3336-3341, 2009.

[24]. Johnson, Stephen C. "Hierarchical clustering schemes." Psychometrika 32, no. 3 : 241-254, 1967.

[25]. Zhao, Ying, George Karypis, and Usama Fayyad. "Hierarchical clustering algorithms for document datasets." Data mining and knowledge discovery 10, no. 2: 141-168, 2005

[26]. Day, William HE, and Herbert Edelsbrunner. "Efficient algorithms for agglomerative hierarchical clustering methods." Journal of classification 1, no. 1 : 7-24, 1984.

[27]. Murtagh, Fionn. "A survey of recent advances in hierarchical clustering algorithms." The Computer Journal 26.4 : 354-359, 1983.

[28]. Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim. "CURE: an efficient clustering algorithm for large databases." In ACM Sigmod Record, vol. 27, no. 2, pp. 73-84. ACM, 1998.

[29]. Madan, Siddharth, and Kristin J. Dana. "Modified balanced iterative reducing and clustering using hierarchies (m-BIRCH) for visual clustering." Pattern Analysis and Applications 19, o. 4 (2016): 1023-1040.

[30]. Guha, Sudipto, Rajeev Rastogi, and Kyuseok Shim. "ROCK: A robust clustering algorithm for categorical attributes." Information systems 25, no. 5 : 345-366, 2000.

[31]. Karypis, George, Eui-Hong Sam Han, and Vipin Kumar. "Chameleon: Hierarchical clustering using dynamic modeling." Computer 8: 68-75, 1999.

[32]. Elhamifar, Ehsan, and Rene Vidal. "Sparse subspace clustering: Algorithm, theory, and applications." IEEE transactions on pattern analysis and machine intelligence 35, no. 11: 2765-2781, 2013.