International Journal of Innovative Research in Technology & Science

ISSN: 2321-1156 Volume VIII Issue IV, June-July 2020

Optimizing Initial Cluster Center Based on Data Analysis Using K-Means Clustering Algorithm and PAVSM

Chetali Makode[1], Kedar Nath Singh[2]; *Department of CSE; TITS, Bhopal, India; [1]cdmakode@gmail.com, [2]cseknsingh@gmail.com*

ABSTRACT*: As a partition primarily based bunch algorithmic rule, K-Means is wide utilized in several areas for the options of its efficiency and simply understood. However, it's documented that the K-Means algorithmic rule could get suboptimal solutions, depending on the selection of the initial cluster centers. During this paper, they propose a projection-based K-Means data format algorithmic rule. The planned algorithmic rule initially uses a typical mathematician kernel density estimation technique to search out the extremely density information areas in one dimension. Then the projection step is to iteratively use density estimation from the lower variance dimensions to the upper variance ones till all the scale is computed. Experiments on actual datasets show that our technique will get similar results compared with different typical strategies with fewer computation tasks this paper reviews numerous strategies and techniques utilized in literature and its benefits and limitations, to research the more would like of improvement of the k-means algorithmic rule. Planned algorithmic rule (PAVSM) increased information analysis.*

KEYWORDS: Initial Centroids, Clustering, Data mining, Data sets, clustering Technique, K-means clustering, unsupervised learning.

I. INTRODUCTION

Data mining is outlined because of the method of extraction of hidden data from giant volumes of data. Data processing has been outlined because of the nontrivial extraction of previously unknown and probably helpful data from the information held on in exceeding information. Data processing is employed to find data out of information and presenting it in an exceeding type that's simply understood to humans. Data processing is that the notion of all ways and techniques which permit analyzing giant information sets to extract and see previously, Unknown structures and relations out of such large tons of details. data mining or data methods is that the process of extracting data from giant information sets through the utilization of algorithms and techniques drawn from the sector of Statistics, machine learning and information base management systems, ancient information analysis strategies usually involve manual work and interpretation of information that's slow, costly and extremely subjective. Data processing popularly referred to as data discovery in giant information, enables firms and organizations to form calculated selections collecting, accumulating, and analyzing and accessing company information. It uses a type of tool like question and coverage tools, analytical process tools, and call network (DSS) tools [1]. Fayyad health care databases have a large quantity of information however but, there's an absence of effective analysis tools to find the hidden data. Acceptable computer-based data and/or call support systems will facilitate physicians in their work to counsel less costly therapeutically equivalent alternatives. The efficient and correct implementation of an automatic system desires a comparative study of varied techniques offered. during this paper, they present a summary of the present analysis being dole out victimization the information mining techniques for the identification and numerous diseases, light vital problems and summarizing the approaches in an exceedingly set of learned lessons [2]. Data mining method within the data Discovery in information method includes some steps leading from information collections to a variety of new data. It consists of the subsequent steps as shown in figure 1.

Data cleaning:-It is additionally referred to as the info cleansing, it's innovated that noise information and relevant information off from the gathering.
Information integration:-At this stage, multiple information sources, typically heterogeneous, could also be combined in exceedingly common supply.

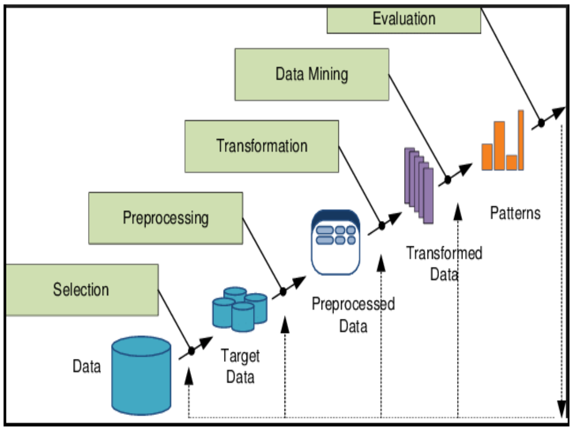

Figure 1 use full information to find the process in DM

Data selection:-At this step, the information relevant to the analysis is set on and retrieved from the information collection.
Data transformation:-It is additionally referred to as information consolidation, it's innovated that the chosen information is reworked into forms acceptable for the mining procedure.

Data mining:-It is that the crucial steps within which clever techniques are applied to extract patterns probably helpful.

Pattern evaluation:-In this step, strictly interesting patterns representing data are known supported given measures.

Data representation:-It is that the final innovate that the discovered data is visually painted to the user. This essential set uses a visual image technique to assist users perceives and interprets the information mining result [3].

Data Mining Clustering Techniques

Clustering is finding teams of objects specified the objects in one cluster are kind of like each other and different from the objects in another cluster. The cluster may be thought of as the foremost vital unsupervised learning technique. A cluster may be thought of as the foremost vital unattended learning technique thus as each different drawback of this type. It deals with finding a structure in an exceedingly collection of unlabelled information. Cluster is the method of organizing objects into teams whose members are similar in away. Cluster analysis has been widely employed in several applications like business intelligence image pattern recognition net search biology and security. In business intelligence cluster may be wont to organize an oversized range of shoppers into teams wherever customers among a bunch share similar characteristics. This facilitates the event of business ways for increased client relationship management. In the image, the recognitions cluster may be wont to discover clusters or subclasses in a written character recognition system. Suppose we have a knowledge set of written digits wherever every digit is labeled as either 1 2, 3, and so on. Note that there may be an oversized variance within the means during which individuals write constant digit. Take the quantity a pair of, for instance, .some individuals could write it with a little cycle at the left bottom half, whereas another might not. They will use the cluster to see subcategories for every of that represents a variation on the means during which a pair of may be written. Victimization multiple models supported the subclasses will improve overall recognition accuracy [4]

Clustering Algorithms:-Clustering may be a data processing technique to cluster similar information into a cluster and dissimilar information into totally different clusters. 1) K-Mean algorithmic rule:-K-mean is a repetitious cluster algorithm during which items are affected among sets of clusters unit the required set is reached. As such, it's going to be viewed as a kind of square error algorithmic rule, though the convergence criteria needn't be outlined supported the squared error. A high degree of similarity among components in clusters is obtained, whereas a high degree of similarity among components in clusters is obtained whereas a high degree of difference among components in several clusters is achieved at the same time. Sets of the algorithm: a) initial it selects the initial k prototypes at random. b) The square error criterion is used to see the cluster quality. c) In every iteration, the paradigm of every cluster is recomputed to be the cluster mean. d) The fundamental version of k- suggests that it doesn't embody any sampling techniques to scale to very large database**s** [5].

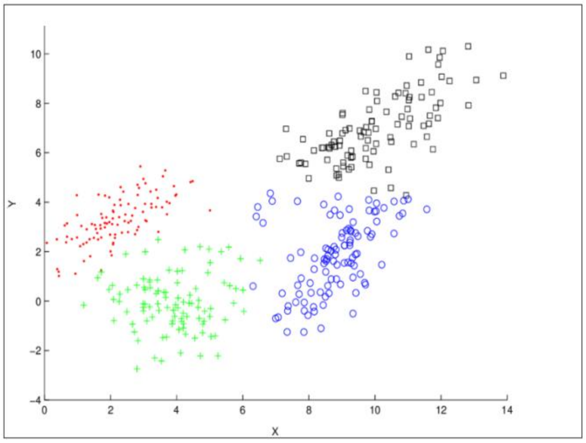

Figure 2 clustering

II.RELATED WORK

Shi et al. [6]. Aiming at the problems of the classical data classification method, this paper proposes a method using a genetic algorithm and K-means algorithm to classify data. To improve the effectiveness of data analysis, considering that the classical K-means algorithm is easy to be influenced by the initial cluster center with random selection, this paper improves the K-means algorithm by using the method of optimizing the initial cluster center. This paper first uses the sorted neighborhood method (SNM) to preprocess the data, and then the K-means algorithm is used to cluster data. To improve the accuracy of the K-means algorithm, this paper optimizes the initial cluster center and unifies the genetic algorithm for the data dimensionality reduction. The experimental results show that the proposed method has higher classification accuracy than the classical data classification method has.

Kolçe et al. [7] In this paper we present an overview of the current research being carried out using the data mining techniques for the diagnosis and prognosis of various diseases, highlighting critical issues and summarizing the approaches in a set of learned lessons. The goal of this study is to identify and evaluate the most commonly used data mining algorithms on medical databases. The following algorithms have been identified: Decision Trees (DT's) C4.5 and C5, Support Vector Machine (SVM), Artificial neural networks (ANNs) and their Multilayer Perception model, Naïve Bayes, Logistic Regression, Genetic Algorithms (GAs) /

Evolutionary Programming (EP), Fuzzy Rules. Analyses show that it is very difficult to name a single data mining algorithm as the most suitable for the diagnosis and/or prognosis of diseases. At times some algorithms perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms together results in more effective.

Du et al. [8]. As a partition-based clustering algorithm, K-Means is widely used in many areas for the features of its efficiency and easily understood. However, it is well known that the K-Means algorithm may get suboptimal solutions, depending on the choice of the initial cluster centers. In this paper, we propose a projection-based K-Means initialization algorithm. The proposed algorithm first employs a conventional Gaussian kernel density estimation method to find the high-density data areas in one dimension. Then the projection step is to iteratively use density estimation from the lower variance dimensions to the higher variance ones until all the dimensions are computed. Experiments on actual datasets show that our method can get similar results compared with other conventional methods with fewer computation tasks.

Satu et al. [9] Protein localization prediction is a computation approach to predict where a protein resides in a cell. Accurate localization of proteins is needed to provide physiological substance for their function and aberrant localization of protein causes pathogenesis of various human diseases. E. Coli and Yeast is a unicellular organism and different proteins allocate in their cell. If those proteins are dislocated, then these cause various infections that affected the human body adversely. So, the objective of this work is to classify proteins into different cellular localization sites based on amino acid sequences of E. Coli bacterium and Yeast. In this experiment, we collect the dataset of E. Coli and Yeast from the data repository and preprocessed it for further processing.
Then we train our dataset with several data mining classification algorithms and artificial neural networks. After classifying both datasets, we compare accuracies among different classifiers and try to find the best classifiers for Protein localization sites prediction of E. Coli and Yeast dataset.

Hareesha K. et al. [10] present a modified K-means algorithm to improve the cluster quality and to fix the optimal number of clusters. As the input number of clusters (K) given to the K-means algorithm by the user. But in the practical scenario, it is very difficult to fix the number of clusters in advance. The method proposed in this paper works for both the cases i.e. for a known number of clusters in advance as well as an unknown number of clusters. The user has the flexibility either to fix the number of clusters or input the minimum number of clusters required. The new cluster centers are computed by the algorithm by incrementing the cluster counter by one in each iteration until it satisfies the validity of cluster quality. This algorithm will overcome this problem by finding the optimal number of clusters on the run.

Bapusaheb B et al. [11the K means clustering algorithm which mainly based on initial cluster centers. In this paper, K means clustering algorithm designed in such a way that the initial centroids selected using the Pillar algorithm. Pillar algorithm effectively chooses the initial centroids and improves the accuracy of clusters. However, the proposed algorithm has an outlier problem that leads to reduced performance. So there is a need to choose the appropriate parameter in data set for outlier detection mechanisms. An improvement in the pillar algorithm is done and the number of distance calculations reduced for the previous initial centroids neighbors and used for the next step of iterations which causes to increase in the computational time. The experimental results show that the use of a pillar algorithm with a change improved the solution.

Shi Na et al. [12] present the analysis of shortcomings of the standard k-means algorithm. K-means algorithm has to calculate the distance between each data object and all cluster centers in iteration. This repetitive process affects the efficiency of the clustering algorithm. An improved k-means algorithm is proposed in this paper. A simple data structure is required to store some information in iteration which is to be used in the next iteration. The computation of distance in iteration is avoided by the proposed method and saves the running time.

Canlas et al. [13] The successful application of data mining in highly visible fields like e-business, marketing, and retail has led to the popularity of its use in knowledge discovery in databases (KDD) in other industries and sectors. Among these sectors that are just discovering data mining are the fields of medicine and public health. This research paper provides a survey of current techniques of KDD, using data mining tools for healthcare and public health. It also discusses critical issues and challenges associated with data mining and healthcare in general. The research found a growing number of data mining applications, including analysis of health care centers for better health policy-making, detection of disease outbreaks and preventable hospital deaths, and detection of fraudulent insurance claims.

Kamaljit Kaur et al. [14] found that the K-Means algorithm has two major limitations 1. Several distance calculations of each data point from all the centroids in iteration. 2. The final clusters depend upon the selection of initial centroids. This work improves the k-Means clustering algorithm designed in MATLAB and the datasets from the UCI machine learning repository used. The initial centroids not selected randomly. By using a new approach good clustering results obtained. The new method of selection of initial centroid is better than selecting the initial centroids randomly.

Huang Xiuchang et al. [15] focused on the problem of user behavior pattern analysis, which has the insensitivity of numerical value, uneven spatial and temporal distribution characteristics strong noise. The traditional clustering algorithm does not work properly. This paper analyses the existing clustering methods, trajectory analysis methods, and behavior pattern analysis methods, and combines the clustering algorithm into the trajectory analysis. After modifying the traditional K-MEANS clustering algorithm, the new improved algorithm designed which is suitable to solve the problem of user behavior pattern analysis compared with traditional clustering methods on the basis of the test of the simulation data and actual data, the results shows that the improved algorithm was more suitable for solving the trajectory pattern of user behavior problems.

III. MATLAB TOOL

The Performance analysis of MATLAB (R2013a) 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 add the advantage of multi-core and digital computer computers, MATLAB gives several multi-threaded algebras and numerical operate. These functions automatically execute on multiple process threads during a single MATLAB, to execute quicker on multicore computers. During this thesis, all increased efficient information retrieve results were performed in MATLAB (R2013a).MATLAB is the high-level language and interactive environment utilized by a lot of engineers and scientists worldwide. It lets them explore and visualize concepts and collaborate across totally different disciplines with signal and image processes, 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. MATLAB (matrix laboratory) could be a multi-paradigm numerical computing scenario and fourth-generation programming language. It's developed by maths work; MATLAB permits matrix strategy, plotting of operates and knowledge, implementation of the rule, construction of user interfaces with programs. MATLAB is meant in the main for mathematical computing; an optional toolbox uses the MuPAD symbolic engine, permitting access to symbolic computing capabilities.

IV. RESULT ANALYSIS

In research in the field of data mining based on optimizing the initial cluster center based on data analysis using the k-means clustering algorithm and PASVM. Find minimum error of medical dataset (E. Coli and Yeast) analysis and the best possible solution. Result analysis Histoplasmosis_Yeastdataset, Yeastbacteria_dataset, Ecolibacteria_datasetand Etec_Ecolidataset analysis time select any random values set. The previous method (PAKMC) is more error rate and compare to the present method (PAVSM). Previous scheme (PAKMC) estimation time obtain minimization and compare to the present method (PAVSM). The present method (PAVSM) is better as compare to the previous method (PAKMC) because data idleness is more but the present method (PAVSM) is minim idleness, but error the lowest amount. The present method (PAVSM) gets fine data in Histoplasmosis_Yeastdataset, Yeastbacteria_dataset, and Ecolibacteria_datasetandEtec_Ecolidataset.

(i) Results Graph based on Etec_Ecolidataset: Result in analysis etec_ecoli dataset analysis time select any random values set. The previous method (PAKMC) is more error rate and compare to the present method (PAVSM).

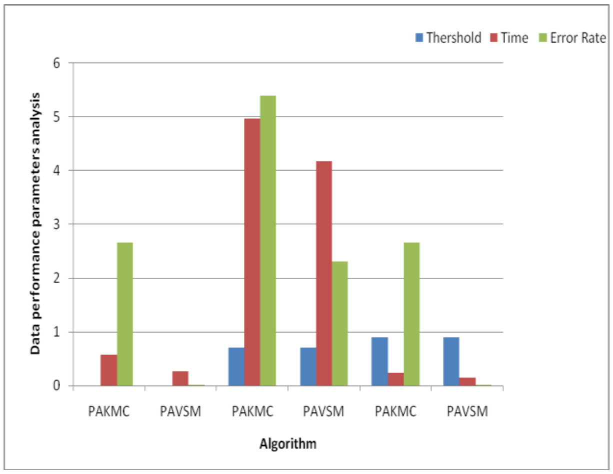

Figure 3 Results Graph based on Etec_Ecolidataset

The previous method (PAKMC) time takes minimization and compares it to the present method (PAVSM). The present method (PAVSM) is better as compare to the previous method (PAKMC) because data idleness is more but the present method (PAVSM) is minim idleness, but error the lowest amount. The present method (PAVSM) gets fine data in etec_Ecolidataset. It is shown in figure 3.

(ii)Results Graph based on Histoplasmosis_Yeastdataset: Result analysis Histoplasmosis_Yeastdataset analysis time select any random values set. The previous method (PAKMC) is more error rate and compare to the present method (PAVSM). The previous method (PAKMC) time takes minimization and compares it to the present method (PAVSM). The present method (PAVSM) is better as compare to the previous method (PAKMC) because data idleness is more but the present method (PAVSM) is minim idleness, but error the lowest amount. The present method (PAVSM) gets fine data in Histoplasmosis_Yeastdataset. It is show figure 4in below graph shows.

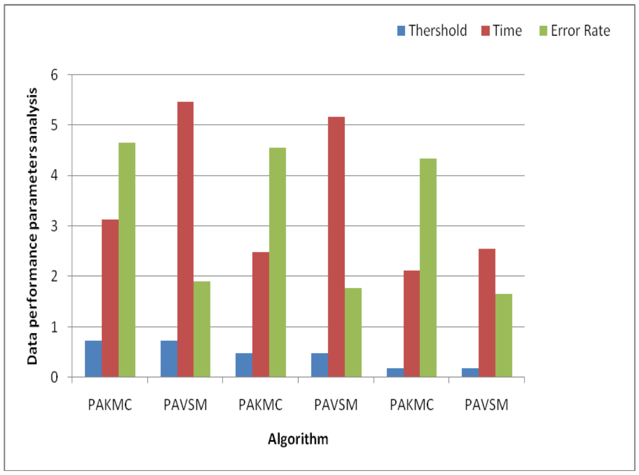

Fig4. Results Graph based on histoplasmosis_Yeastdataset

V. CONCLUSION

Proposed algorithm using vector space method (PAVSM) partitioning based clustering algorithm required to define the number of final clusters (k) beforehand. PAKMC needs more computation time. PAVSM algorithm seems to be superior to the PAKMC algorithm. In PAKMC apriority specification of the number of clusters. Euclidean distance measures can unequally weight underlying factors. PAVSM is fast, robust, and easier to understand. The result analysis dataset is distinct in form of clusters and well-separated data analysis in clusters from each other. It is also observed PAKMC produces close results to PAVSM clustering but it still requires more computation time than PAVSM clustering. PAVSM is better than the PAKMC algorithm; k-means cluster techniques of information mining are analyzed. This work shows that there are many strategies to enhance the cluster with totally different approaches. Numerous cluster techniques are reviewed that improve the existing algorithmic program from a different perspective. Some limitations of the existing algorithmic program are going to be eliminated in future work. These methods are going to be helpful in the extraction of helpful data victimization clusters from large information. It removes the limitation of the K-means bunch algorithmic program and offers correct end in less time therefore they can say it's extremely efficient than normal K-means cluster algorithmic program and quality of cluster additionally improved. Our analysis of various K-means approaches, they conclude that it's higher than the ancient K-means cluster algorithmic program.PAVSM Clustering algorithm is an efficient algorithm for large data and reduces computation time as compared to the PAKMC algorithm. Moreover, this PAVSM algorithm is fast, easier to understand, and identifies minimization values of the dataset and optimal solution.

REFERENCES

[1]. Prajapati. D, Prajapat. J, "Handling missing values: Application to University Data Set", August 2011.

[2]. J.Han and M. Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann Publishers, 2nd ed., New Delhi, 2006.

[3]. Grabmeier. J, Rudolph. A "Technique of Clustering Algorithms in Data Mining", Data Mining and Knowledge Discovery,2002.

[4]. Tayel, Salma, et al. "Rule-based Complaint Detection using RapidMiner", Conference: RCOMM 2013, At Porto, Portugal, Volume: 141- 149,2014.

[5]. Margaret H. Dunham, "Data Mining Introductory and Advanced Topics", Dorling Kindersley Pvt.Ltd.India,Sixth Edition,2013.

[6]. Shi, Haobin, and Meng Xu. "A Data Classification Method Using Genetic Algorithm and K-Means Algorithm with Optimizing Initial Cluster Center." 2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET). IEEE, 2018.

[7]. Kolçe, Elma, and Neki Frasheri. "A literature review of data mining techniques used in healthcare databases." ICT innovations,2012.

[8]. Du, Wei, "A new projection-based K-Means initialization algorithm." 2016 IEEE Chinese Guidance, Navigation, and Control Conference (CGNCC). IEEE, 2016.

[9]. Satu, Md Shahriare, Tania Akter, and Md Jamal Uddin. "Performance analysis of classifying localization sites of protein using data mining techniques and artificial neural networks." 2017 International Conference on Electrical, Computer, and Communication Engineering (ECCE). IEEE, 2017.

[10]. Hareesha, K., Shafeeq A, Dynamic Clustering of Data with Modified K-Means Algorithm, International Conference on Information and Computer Networks, vol. 27,2012

[11]. Bapusaheb B. Bhusare, S. M. Bansode, "Centroids Initialization for K-Means Clustering using Improved Pillar Algorithm", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 4, April 2014.

[12]. Shi Na, Liu Xumin, Guan Yong, ìResearch on K-means Clustering Algorithm: An Improved Kmeans Clustering Algorithm, Intelligent Information Technology, and Security Informatics,2010 IEEE Third International Symposium on 2-4 April 2010(pp. 63-67)

[13]. Canlas, Ruben. D. "Data mining in healthcare: Current applications and issues." School of Information Systems & Management, Carnegie Mellon University, Australia (2009).

[14]. Kamaljit Kaur, Dr. Dalvinder Singh Dhaliwal, Dr. Ravinder Kumar Vohra, "Statistically Refining the Initial Points for K-Means Clustering Algorithm ", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 11, November 2013.

[15]. Huang Xiuchang, SU Wei, "An Improved K-means Clustering Algorithm", Journal Of Networks, Vol. 9, No. 1, January 2014.

[16]. Omran, Mahamed GH, Andries P. Engelbrecht, and Ayed Salman. "An overview of clustering methods." Intelligent Data Analysis 11, no. 6: 583-605, 2007.

[17]. Reynolds, Alan P., Graeme Richards, Beatriz de la Iglesia, and Victor J. Rayward-Smith. "Clustering rules: a comparison of partitioning and hierarchical clustering algorithms." Journal of Mathematical Modelling and Algorithms 5, no. 4: 475-504, 2006.

[18]. Khan, Shehroz S., and Amir Ahmad. "Cluster center initialization algorithm for K-means clustering." Pattern recognition letters 25, no. 11: 1293-1302, 2004.

[19]. Bachem, Olivier, Mario Lucic, Hamed Hassani, and Andreas Krause. "Fast and provably good seedings for k-means." In Advances in neural information processing systems, pp. 55-63. 2016.

[20]. Nayak, Janmenjoy, Bighnaraj Naik, and H. S. Behera. "Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014." In Computational intelligence in data mining-volume 2, pp. 133-149. Springer, New Delhi, 2015.

[21]. Schubert, Erich, and Peter J. Rousseeuw. "Faster k-medoids clustering: improving the PAM, CLARA, and CLARANS algorithms." In International Conference on Similarity Search and Applications, pp. 171-187. Springer, Cham, 2019.

[22]. Reynolds, Alan P., Graeme Richards, and Vic J. Rayward-Smith. "The application of k-medoids and pam to the clustering of rules." In International Conference on Intelligent Data Engineering and Automated Learning, pp. 173-178. Springer, Berlin, Heidelberg, 2004.

[23]. LIN, Kai-yan, Li-hong XU, and Jun-hui WU. "A fast fuzzy C-means clustering for color image segmentation." Journal of Image and Graphics 9, no. 2: 159-163, 2004.

[24]. Xin-Quan, C. H. E. N. "Feature-weighted fuzzy C clustering algorithm [J]." Computer Engineering and Design 22, 2007.

[25]. Ji, Zexuan, Yong Xia, Qiang Chen, Quansen Sun, Deshen Xia, and David Dagan Feng. "Fuzzy c-means clustering with weighted image patch for image segmentation." Applied soft computing 12, no. 6: 1659-1667, 2012.

[26]. Miyamoto, Sadaaki. "Different Objective Functions in Fuzzy c-Means Algorithms and Kernel-Based Clustering." International Journal of Fuzzy Systems 13, no. 2,2011.

[27]. Bandyopadhyay, Seema, and Edward J. Coyle. "An energy-efficient hierarchical clustering algorithm for wireless sensor networks." In IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428), vol. 3, pp. 1713-1723. IEEE, 2003.

[28]. Zhang, Tian, Raghu Ramakrishnan, and Miron Livny. "BIRCH: an efficient data clustering method for very large databases." *ACM sigmod record* 25, no. 2: 103-114, 1996.