iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5040

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Improving Performance of Elastic K-Means Clustering using similarity Measures

Yogita K. Patil[1], Mr. Sandip S. Patil[2]
[1, 2]*Department of Computer Science and Engineering SSBT's College of Engineering and Technology, Kaviyatri Bahinabai Chaudhari N.M.U, Jalgaon [M.S], India*

Abstract: ***Clustering assign a proper membership to the data. Clustering has numerous applications inmarket research, pattern recognition, data analysis and image processing. Standard K-means clustering uses crisp membership to assign the data to a single cluster only. In real world data, some noise or ambiguity is present, so k-means clustering is not able to handle those data and it generates wrong membership for the clusters. In the proposed system, Elastic k-means clustering is used for creating flexible membership. For that purpose elastic k-means clustering uses vectorization and similarity measures for improving the membership and also handles the problem of the noisy data. Unstructured documents are used for experiment and the experimental results shows that, the proposed system gives better accuracy than existing one.***
Keywords: Clustering, Elastic k-Means, Ambiguity

I. INTRODUCTION

Clustering is one of the most popular unsupervised classification technique. In the process of clustering the data can be divided on the bases of centroid, distribution and densities etc. The clustering is the method of identifying similar group of data in a dataset. In simple words, the aim is to isolate groups with similar forms and assign them into clusters. Clustering can be divided into two subgroups i.e. hard clustering and soft clustering [1]. In hard clustering, each object is assign to exactly one cluster completely. The cluster is do not overlap means each element either belongs to one and only one cluster or not. Hard clustering is also called as crisp clustering [2] and also referred as fuzzy clustering or soft K-means. Second important type of clustering is soft clustering. In soft clustering, each data point can belongs to more than one cluster. Means clusters may overlap. In hard clustering, each data point can have membership to multiple clusters these membership coefficients define strictly from 1 or 0 means the range from 1 to 0. Following types of clustering are the also subtypes of clustering base on hard and soft clustering. Partitioning and hierarchicalmethods are main group of clustering. Also density-based methods, model-based clustering and grid-based methods are important categories of clustering [3].

A. Partition Based Clustering

Partitioning Clustering This method decomposes a dataset into a set of disjoint clusters. K-means algorithm is partitioning base algorithm [4].

B. Hierarchical clustering

Hierarchical clustering is set of nested clusters and it is organized as tree like structure. Hierarchical clustering is work on different levels of checking the dissimilarity on each level. Agglomerative and divisive are the algorithms of Hierarchical clustering.

C. Density-based clustering

Density-based clustering is use for finding nonlinear shape structure base on density. Density-based clustering uses the concept like density reachability and connectivity. Density-based clustering is able to identify noise data while clustering.

D. Model-based clustering

Model-based clustering the data is generated by usinga mixture of probability distribution in each component that represents different clusters.

E. Grid-based clustering

Grid-based clustering is efficient in mining large multidimensional data. Using thisclustering the data space is partition into finite number of cells to form a grid structureand then form cluster belongs to those cells in the grid structure.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Following are the few soft clustering algorithms:

1) *Fuzzy C Means Clustering:*Fuzzy c-means(FCM) clustering works on assigningmembership to each data point belongs to each cluster center on the basis of distancebetween data point and center of cluster. if data is more near to the cluster centerthen the membership correspond to the particular cluster center is also more. Fuzzy C-means clustering gives the best result for overlapped dataset.Fuzzy C-means clusteringis comparatively better than k-means algorithm [5].
2) *Elastic K-means Clustering:* Elastic K-means clustering is work as fuzzy clustering but, elastic k-means clustering ismore effective as compare to fuzzy clustering. Elastic k-means is the synthesis between thefuzzy logic and clustering which is the requirement of modern computing. The aim of elasticclustering is to model the ambiguity within the unlabeled data objects efficiently. Every dataobject is assigned a membership to represent the degree of belonging to a certain class. Therequirement that each object is assigned to only one cluster is relaxed to a weaker requirementin which the object can belong to all clusters with a certain degree of membership.Elastic k-means algorithm using posterior probability withsoft capability where each data point can belong to multiple clusters fractionally and showthe benefit of proposed Elastic K-means.[6] In elastic k-means clustering with posterior probability attributes are used for proper clustering. The results of EKM clustering shows that belongness of clusters are increases but, it cannot handle the noise in dataset.Fuzzy clustering algorithm to improve of Dataobjects membership. The Proposed Fuzzy K-Means clustering assigns membership to anobject inversely related to the relative distance of the object to cluster prototype.[23] Fuzzyclustering uses membership values to assign data objects to one or more clusters. Themembership values indicate the strength of the association between that data object anda particular cluster. Fuzzy K-Means approach takes less execution time and requires lessmemory than that of hard K-Means.

II. OBJECTIVES

To develop a system to handle the problem of noisy data and scattered clusters by assigning data points to each several possible clusters and also provide exibility in the membership for clustering.To improve the accuracy of clustering in term of precision, recall and f-score.

III. METHODOLOGY

Proposed EKM clustering uses posterior probability with soft capability where each data point can belong to multiple clusters fractionally and also use vector and similar data. EKMclustering algorithm, have more fuzziness as compare to hard clustering algorithms. EKM clustering with vector data relaxes the requirement by providing gradual membership which is suited for the real world problems. It solves the problem of belongness of clusters means, each data point can belong to multiple clusters fractionally and show the benefit of proposed Elastic K-means clustering algorithm.

A. Proposed System Architecture

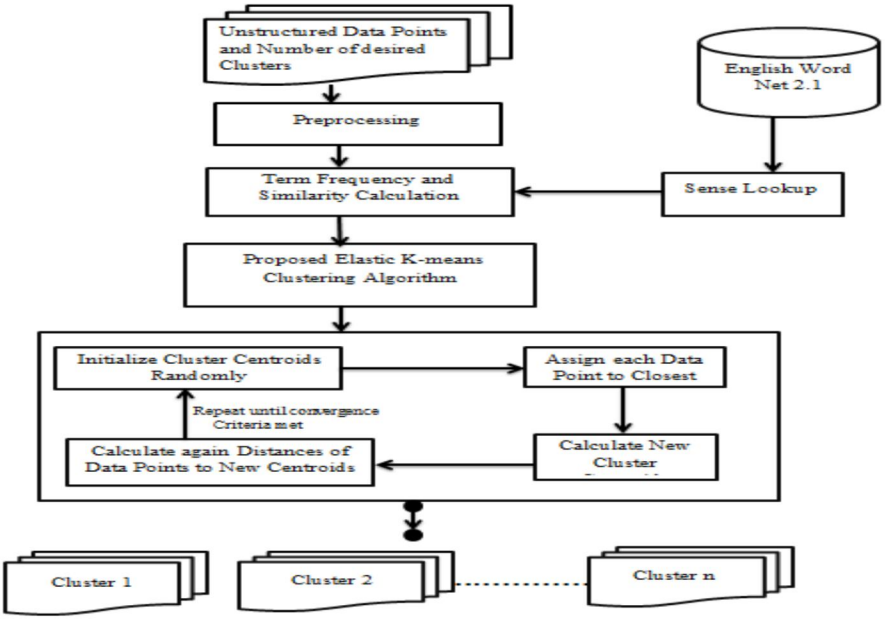


Fig.1 Architecture of Proposed System

For clustering Elastic k-means algorithm is used. Set of documents is taken as a input to the system and also takes the number of cluster for grouping of given set. The data preprocessing consist of stemming, removal of stop words, tokenization. The next step is features extraction be normalized into system acceptable format. In the preprocessing stage, the document mainly includes identification of sentence or boundary word, stop word elimination, stemming procedure and features extraction and synonyms generation. Preprocessing of data includes the removal of hyperlinks as well as numbers. Additional punctuation, lengthening words extraction,replacement before tokenization and negation is a type of expression that can shift sentiment polarity in the text, which needs to be taken into consideration in sentiment analysis. Postagging and removal of pronouns, prepositions, conjunctions and punctuation. All above preprocessing steps can be summarized as:

1) *Tokenization:* The tokenization is required for separating for separating from the sentences and produces the tokens. The tokens act as a particular word which is extracted from the sentence. Generally in any language, the sentence boundary is identifiedwhen stop word occurs;similarly, the words are separated by using space. The special characters such as,./?[]:;()=+- are removed.
2) *Stemming:* The goal of stemming is to reduce inflectional forms and sometimes the related forms of a word to a common base form. For instance: am, is, are, be, dog, dogs, dog's, dogs'= dog.The result of this mapping of text will be something like:The dog's legs are different color=the dog leg is differ color. However, the two words differ in their syntax. Stemming usually refers to do the things properly with the use of a vocabulary and morphological analysis of words,normally aiming to remove inflectional endings only.
3) *Stop word Elimination:* The stop words are natural language words which has less meaning in natural language processing. The stop word are raw data or provides no useful information which have no any value for that purpose it has to be removed. It can be removed by using the stop word removal algorithm. The stop words list is provided as in the text file, the input sentences words is compared to the stop word list, and remove it if it is present in stop word list.
4) *Synonyms Generation:* Synonyms generation is used to similar words detection by using word net 2.1 dictionary. In proposed system, clustering is done on the basis of similar words hence; similarity generation step is added into process. Synonyms areextracted from the wordnet. The wordnet contains the several thousand of sysnets which is a rich source of synonyms.

B. Algorithm 1

1) Input: D=d_1, $d_{2.....}d_n$ ||Set of data points. K=||Number of desired Clusters
2) Output= Clustered Data points
3) Find attributes from all data points
4) Combine attributes from all data points and remove duplicate attributes.
5) Calculate the weights of the attributes using term frequency and inverse document frequency.
6) Choose random two points and set as initial centroids.
7) Calculate distances of data points to all initial centroid and assign each data point tothe closest centroid's cluster
8) For each cluster j(1 6 j 6 k) recalculate the positions of centroids.
9) Calculate distance of data points to new centroids and assign to cluster which has minimum distance from centroid.
10) Go back to step 7 unless the centroids are not changing.
11) Make a recursive call to step 7 to further divide the biggest cluster,until the convergencecriteria met.

IV. EXPERIMENTAL RESULTS

The proposed system is implemented with the use of variant of JAVA module development kit with version 7 and MySQL database. The proposed experiment is performed on Reuter dataset. Generally, performance of clustering is evaluate with the help of following metrics such as, accuracy, precision, recall and f-score. Precision and Recall are defined in terms of a set of retrievedsamples of anomalies and a set of relevant samples of anomalies. For retrieved and relevantsamples of anomalies all the following rates is calculated.

1) *F1-Score or F- measure***:** F1-Score is a measure of a test's accuracy. It considers both the precision and recall of the test to compute the score: p is the number of relevant retrieved records divided by the total number of all retrieved records and r is the number of relevant retrieved records divided by the total number of relevant records that should have been returned.

$$\boldsymbol{F1 - Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall}}$$

Table 1 shows Experimental results consists the values of Precision, Recall and F-Measure with respect to proposed system, considered sample datasets as Numbers of documents. For both of the algorithm, reuter dataset has been used. The proposed experiment is performed on set of 5 numbers of random total sample records for finding relevant and retrieved records from that. Experimental result shows the effectiveness of proposed system.

Fig. 2 shows the f1-score of 10 to 60 document set. EKM Algorithm (using vector data), 50 documents has highest f1-score i.e. 0.918367, while 10 documents have lowest f1-score i.e. 0.8 compare to all.EKM Algorithm(using vector and similarity data), 50 documents has highest f1-score i.e. 0.938776, while 10 documents have lowest f1-score i.e. 0.8 compare to all. Retrieved sense of 50 documents is most relevant and the probability gives true positive rate. It contains highest f1-score with EKM(Using vector and similarity data). F1-score is a measure of a test's accuracy. From the above two performance matrices such as precision and recall, we calculate F-measure. Performance of proposed system is higher as compare to existing.

TABLE I
F1-Score Using Ekm Clustering And Ekm Clustering (Using Vector And Similarity Data)

No. of Reviews	f1-Score of EKM (using vector data)	F1-Score EKM (using vector and similarity data)
10	0.8	0.8836
20	0.871795	0.871795
30	0.881356	0.896552
40	0.88	0.923077
50	0.918367	0.938776
60	0.904348	0.932203

In proposed system, similarity ensures are used with vectorization hence, performance of proposed system is increased. It is compared basedon precision, recall, F-score and finally accuracy of the proposed system. F-measure valueschanges according to the precision and recall values are change. Average Precision values for EKM Algorithm(using vector data) and EKM(using vector and similarity data) is 0.90and 0.92 respectively. Precision is the probability that a retrieved documents of a group ofrelevant documents. As EKM Algorithm(using vector data) degrades the precision and leadsto increase the false negative rate and false positive rate. This is also true for reverse case ofhigh precision. Average Recall values for EKM Algorithm(using vector data) and EKM(usingvector and similarity data) is 0.85 and 0.88 respectively. The F-measure values are calculatedfrom the precision and recall measures. EKM Algorithm(using vector data) is 0.87 and EKMalgorithm (using vector and similarity data) is 0.90. Results of experiment shows that theprecision values of EKM algorithm (using vector and similarity data) increases and getbetter accuracy in terms of the clustering of documents because, It uses similarity measureswith vectorization.

The overall comparative analysis of both clustering algorithms The proposed system gives better results that existing one, becausevector and similarity data relaxes the requirement by providing gradual membership whichis suited for real world problems and it uses posterior probability with soft capability whereeach data point belong to multiple clusters fractionally and show the benefit of proposed elastic K-means clustering.

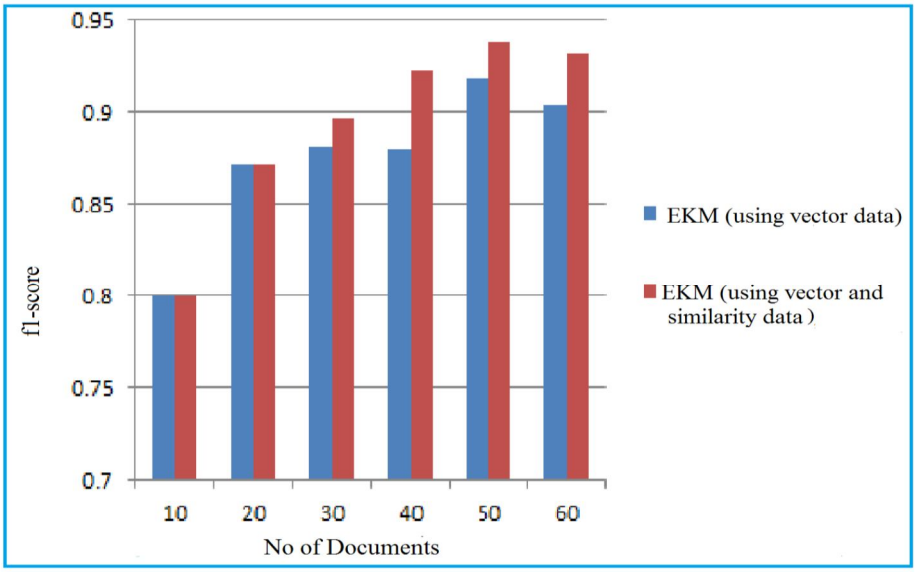


Fig.2 F1-Score using EKM Clustering and EKM Clustering (using vector and similarity data)

V. CONCLUSION

Document clustering is an application of cluster analysis to textual document. Clustering helps in organizing document and retrieving fast information or filtering information. The proposed system is work on elastic k-means algorithm(using vector and similarity data). The proposed system overcomes the problem of fuzziness or non-overlapping of cluster. EKM algorithm is use for removing the noisy data or outlier in pre-processing. For calculate the performance matrices the two terms are used i.e. precision and recall. Average Precision values of EKM(using vector and similarity data) are improved into proposed system also the recall values are improved. On the basis of precision and recall result of proposed system is automatically improved. According to the performance measures we can conclude that, the proposed EKM(using vector and similarity data)gives better results as compared to existing system.

As the part of future scope, document clustering is extended to categorical clustering for getting better results in term of higher information retrieval by using different genetic algorithms.

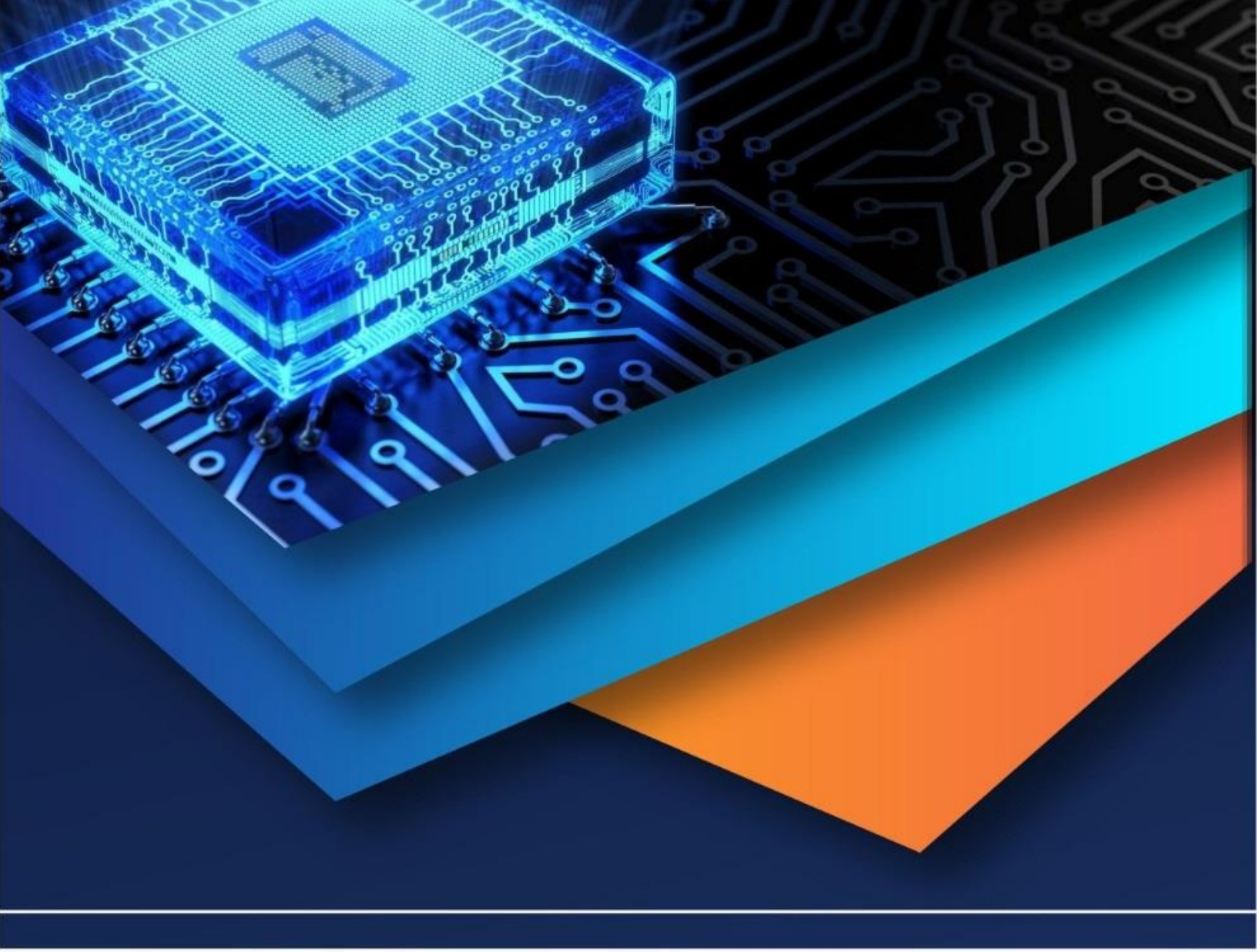
VI. ACKNOWLEDGEMENT

I present my sincere thanks to Prof. Dr. Kishor S. Wani, Principal for moral support andproviding excellent infrastructure in carrying out the project work. I am very thankful toProf. Dr. Girish K. Patnaik, guide and HOD of Computer Engineering for his cooperationand valuable guidance during the work.I express my gratitude towards Mr. Sandip S. Patil, project guide whose experienced guidance becomes very valuable for me.I would like to thank all teaching and nonteaching staff in computer department whohelped me in my work. I would like to thank my classmates who helped me for projectwork,also thanks all those people who helped me in anyway what so ever at some point intime. Last but not least, I would like to thank my parents without whose cooperation thiswon't be possible.

REFERENCES

[1] P. M. Jafar and R. Sivakumar, “A comparative study of hard and fuzzy data clustering algorithms with cluster validity indices”,International Conference on Emerging Research in Computing, Communication an Application (*ERCICA*), pp. 775-782, 2013.
[2] Chris and R. A. E, “How many clusters? Which clustering method?Answers via model-based cluster analysis”, the computer journal, vol.41, No-8, pp. 578-588, 1998.
[3] M. S. Yang, “A survey of fuzzy clustering”, Mathematical and Computer modeling, vol.18, No-11, pp. 1-16, 1993.
[4] S. Ayramo and T. Karkkainen, “Introduction to partitioning based clustering methods with a robust example”,Reports of the Department of Mathematical Information Technology; Series C: Software and Computational Engineering, pp. 1-36, 2006.
[5] T. T. Le and K. Gardiner, “Probability-based imputation method for fuzzy cluster analysis of gene expression microarray data”, IN Information Technology: New Generations (*ITNG*), 2012 Ninth International Conference, pp. 42-47, April 2012.
[6] Zheng and C. Ding, “Elastic k-means using posterior probability”, *PloS one 12(12)*,pp. 1-16, 14 Dec 2017.
[7] P. Rai and S. Singh, “A survey of clustering techniques”, International Journal of Computer Applications, No-12, vol. 7, pp. 1-5, October 2010.

[8] L. X. Shi na and G. Yong, \Clustering algorithm: An improved k-means clustering algorithm," Third International Symposium onIntelligent Information Tecnology ad Security Informatics, pp. 63-67, April 2010.
[9] H. Khanali and B. Vaziri, "A survey on clustering algorithm for partitioning method", International Journal of Computer Applications 155(4), p. 2025, December 2016.
[10] Chadha and S. Kumar, "An improved k-means clustering algorithm: a step forward for removal of dependency on k",International Conference on Reliability Optimization and Information Technology (*ICROIT*), pp. 136-140, 2014.
[11] Bhat, "K-medoidsclustering using partitioning around mediods for performing face recognition", International Journal of Soft Computing, Mathematics and Control 3(3), pp. 63-67, August 2014.
[12] R. Nagarajan, S. Anu, H. Nair, N. Puviarasan, and P. Aruna, "Document clustering using agglomerative hierarchical clustering approach (ahdc) and proposed tsgkeywordextraction method", IJRET: InternationalJournal on Research Engineering Technology 5(11), pp. 118-124, Nov 2016.
[13] P. Praveen, B. Rama, U. N. Dulhare, and T. Warangal, "A study on monothetic divisive hierarchical clustering method", International Journal of Advanced Scientific Technologies, Engineering and Management Sciences (*IJASTEMS-ISSN: 2454-356x*) Volume 3, Special Issue 1, pp. 369-372, March 2017.
[14] O. Kettai and F. Ramdani, "An agglomerative clustering method for large datasets", International Journal of Computer Applications, Volume 92(14), pp. 24-28, April 204.
[15] M. Reddy, M. Vivekananda, and R. Satish, "Divisive hierarchical clustering with k-means and agglomerative hierarchical clustering", International Journal of Computer Science Trends and Technology, Volume 5, Issue 5, pp. 6-11, Oct 2017.
[16] Nidhi and A. Patel, "An efficient and scalable density-based clustering algorithm for normalize data", Procedial Computer Science, Volume 92, pp. 136-141, 2016.
[17] A. Ram, S. Jalal, A. S. Jalal, and M. Kumar, "A density based algorithm for discovering density varied clusters in large spatial databases" International Journal of Computer Application, Volume 3, pp. 226-231, June 2010.
[18] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks", International Journal of Science 344, pp. 1492-1496, 27 June 2014.
[19] Y. Chen and L. Tu, "Density-based clustering for real tie stream data", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovering and Data Mining, pp. 133-142, 2007.
[20] M. Ilango and D. V. Mohan, "A survey of grid based clustering algorithm", International Journal of Engineering Science and Technology, Volume 2(8), pp. 66-68, 2010.
[21] A. Amini, T. Y. Wah, M. R.Biyani, and S. R. Ahabozorai, "A study of density based clustering algorithm on data streams", International Conference on Fuzzy System and Knowledge Discovery *(FSKD)*, pp. 1652-1656, 26-28 July 2011.
[22] F. Liu, C. Ye, and E. Zhu, "Accurate grid-based clustering algorithm with diagonal grid searching and merging", IOP Conference Series: Materials Science and Engineering*(ICAMMT)* 242, pp. 1-5, 2017.
[23] C. T. Baviskar and S. S.Patil, "Improvement of data objects membership by using fuzzy k-means clustering approach", International Conference on Computation of Power, Energy Information and Communication *(ICCPEIC)*, pp. 139-147, 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY