iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** IV **Month of publication:** April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4482

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: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

A Comprehensive Study about various Clustering Techniques

Dr. V. Kavitha[1], Mrs. S. Subhasini[2], Mr. Srikanth[3]
[1]Professor, [2]Assistant Professor, [3]PG Student, PG & Research Department of Computer Applications (MCA), [2]Department of BCA, Hindusthan College of arts and science,

Abstract: Clustering has been utilised in numerous kinds of research areas, which is one of the unsupervised learning method. This concept tries to discover some patterns and distributions in unlabelled data sets. This paper analysis the fundamental concepts of clustering techniques and its necessities. Some of the clustering techniques like k-means clustering technique, EM Clustering technique and DBSCAN clustering techniques are discussed. Moreover the general kinds of clusters are also discussed.

Keywords: Unsupervised learning, Anomaly Detection, Overlapping

I. INTRODUCTION

Data mining encompasses the clustering, anomaly detection, association rule learning, classification, regression and summarization. Clustering is a most significant problem that has been increased in recent years. The concept of clustering hazards has been addressed in most of the contexts and through researchers in numerous disciplines. This indicates its usefulness and appeal as one of the procedures in exploratory data analysis.

One of the data mining concept of clustering techniques aims at partitioning a set of data objects in classes such that data objects that belong to the identical class are more alike than data objects that belong to different classes. These kinds of classes are referred as clusters and their number might be preassigned or can be a parameter to be determined by the special techniques. Cluster analysis is the arrangement of a collection of patterns which referred as vector of measurements, or a data point in a multidimensional space into similarity based clusters. Data clustering has its roots in number of research areas including machine learning, data mining, statistics and biology.

Conventional clustering techniques can be categorised into two major categories like partition and hierarchical clustering. The number of clusters need not be defined in prior in hierarchical clustering, and obstacles due to making initialisation and local minima does not arise.

Since hierarchical clustering techniques consider only local neighbours in each step, they couldn't associate a previous knowledge regarding the size or global shape of the clusters. As a final outcome they cannot keeps individual overlapping clusters. Additionally, hierarchical clustering is not dynamic and data points committed to a declared cluster in the early stages that cannot move any different cluster.

Prototype based partition clustering techniques can be classified into two categories namely fuzzy clustering and crisp clustering. The former one of fuzzy clustering means each and every data object point belongs to every cluster to a certain degree, and the later one of crisp clustering refers that each data object point based to only one static cluster. Moreover fuzzy clustering techniques can deal with overlapping cluster boundaries. Partitional clustering techniques are not static and the data objects can migrate from one cluster to another one.

They can associate knowledge according to the size or shape of clusters through using relevant prototypes and distance measures. The main drawbacks of the partition approach are the trouble in setting the number of clusters and clusters and sensitivity to noise.

Clustering is a most significant task in data analysis and data mining applications. It is the assignment of combining a set of similar data objects.

So that data objects in the identical group are more related to each other than to those in other cluster groups. Cluster is an ordered list of data which have the similar characteristics. Cluster analysis is the process which is used to discover similarities among data based to their characteristics found in the data set and clustering similar and same data sets into cluster groups. Basically formation of clusters is an unsupervised learning procedure. A fine clustering process will make high superiority clusters with high quality.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

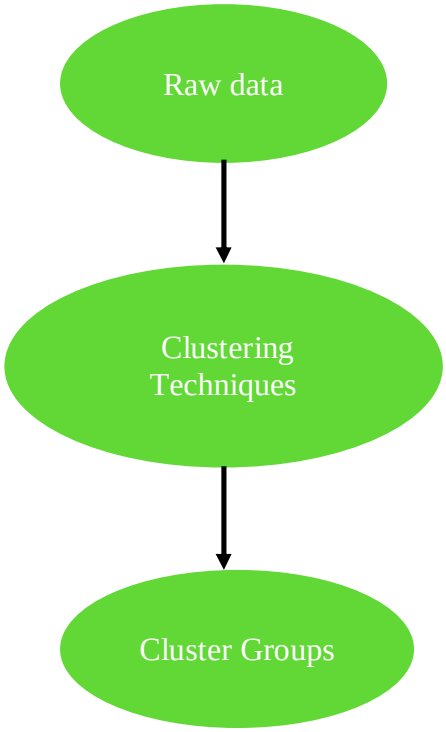


Figure 1 Clustering Stages

Clustering is an unsupervised learning process. A good clustering method will produce high superiority clusters with elevated intra cluster similarity and minimal inter cluster similarity. The superior result of the clustering based on equal similarity measure utilised through the method and its development. The superiority cluster technique is also measured through its ability to discover some or every hidden patterns. Similarity of the group can be calculated through the distance calculation. Clustering process requires some of the significant requirements. Those requirements are scalability, ability to handle with various kinds of cluster attributes, ability to deal dynamic kinds of data, Find out clusters with arbitrary shapes, least requirements for knowledge domain to determine input parameters, Handled with outlier and noise data, Insensitive input records, Association of user specified constraints, high dimensionality, usability and interpretability. Various kinds of data that are utilised for cluster analysis are binary variables, mixed data variables, nominal, ratio ordinal variables and interval scaled variables. Figure 1 depicts about the various stages of clustering process. In that the first process of raw data is taken for grouping process. And continuously, the next process of cluster implementation is executed. Generally, many kinds of clustering techniques are available in data mining research field. Finally, the fine tuned clusters are received through the relevant clustering techniques.

II. TYPES OF CLUSTERING TECHNIQUES

Generally, numerous kinds of clustering techniques available. Any type of clustering technique is to be considered that must be comes under the following general clustering categories. Figure 2 depicts about various kinds of clusters.

A. Nearest Neighbour Cluster

Nearest Neighbour Cluster is also referred as Contiguous Cluster. A cluster is a set of data objects like the data object in a specific cluster is closer to many data objects with in the cluster than to any data object that is not in the cluster.

B. Well-Separated Cluster

A cluster group is a set of data objects like any data object that is present in a cluster is very close to every other data object within the cluster than to any other object which is not appear in the cluster groups.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

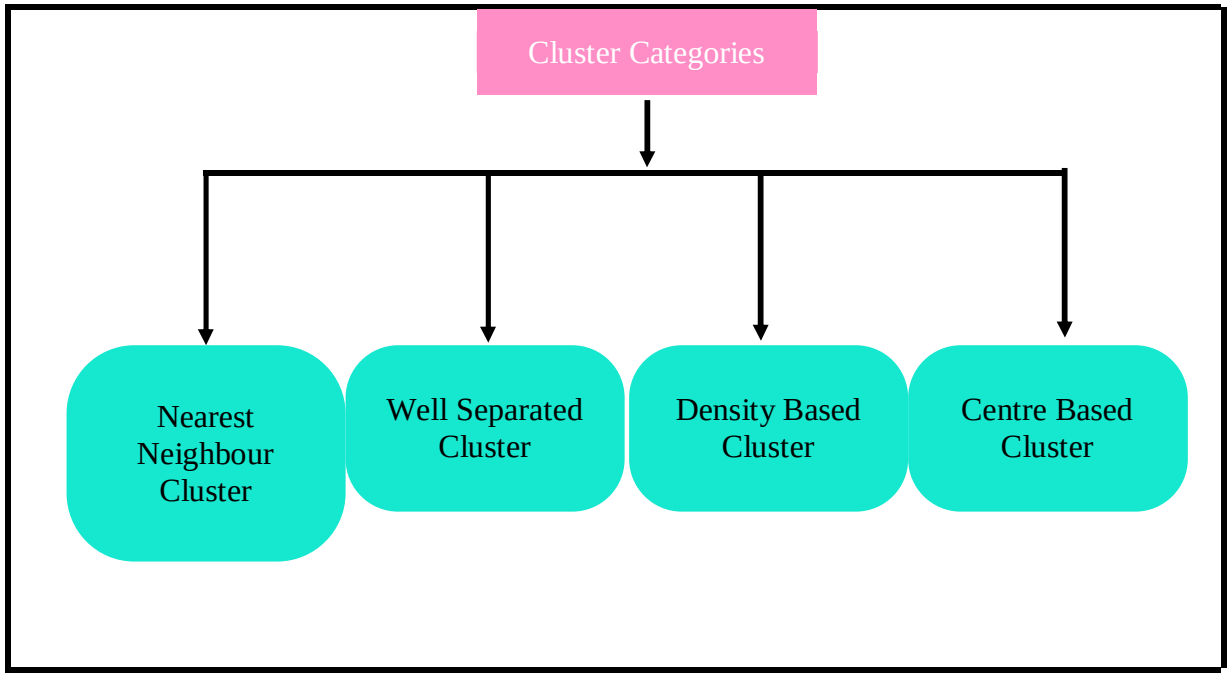


Figure 2 Cluster Categories

1) *Density Based Cluster:* A cluster is dense region of points, separated by a low density regions, from other regions with high density is referred as Density based cluster.
2) *Center based Cluster:* A cluster is a set of objects such that an object in a cluster is closer to the centroid of a cluster, than the centroid of any other cluster.

III. VARIOUS CLUSTERING ALGORITHMS

A. K-mean clustering

K-means is one of the significant partition based clustering technique which can be easily developed and most efficient one in terms of the execution time. This cluster data objects are grouped basis of minimising the sum of squared distances among the data items and the specified centroid data points. A centroid is referred as centre of mass of a geometric data object of constant density. In k-means clustering algorithm each cluster's centre point is referred through the calculated mean value of the data objects in the cluster. K-means clusters requires the specified number of clusters and the syntactic or real time data set. Finally it produced set of k clusters. Figure 3 depicts about the execution of k-means clustering technique.

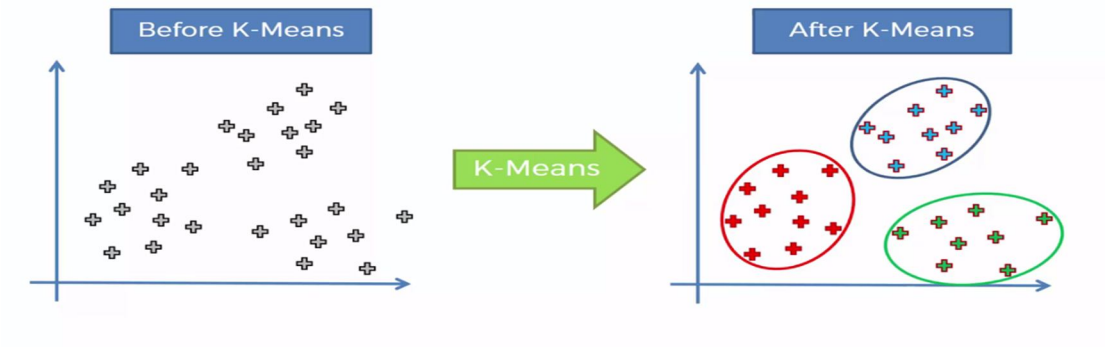


Figure 3 k-means cluster

B. Methodology of K-means Cluster

1) Arbitrarily choose k data objects from dataset as the initial cluster centre.
2) Repeat the process.
3) Reassign each data object to the cluster to which the data object is most similar based on the mean value of the data objects inside the cluster.
4) Update the cluster means continuously.
5) until no change

C. Expectation Maximization Clustering

The EM (expectation maximization) clustering technique is almost identical to the K-Means clustering technique. The EM clustering technique extends this basic approach of k-means clustering in the following ways. Instead of assigning samples to clusters to increase the differences in means for continuous variables, the EM clustering technique calculates probabilities of cluster memberships based on one or more probability distributions. The goal of the EM clustering technique is to maximize the overall probability or likelihood of the data, given the (final) clusters. Figure 4 EM clustering technique explains about the functionality of the algorithm. Through this clustering technique input data objects are taken as input samples. Then the technique of filtering is

Input sample Filtering Samples

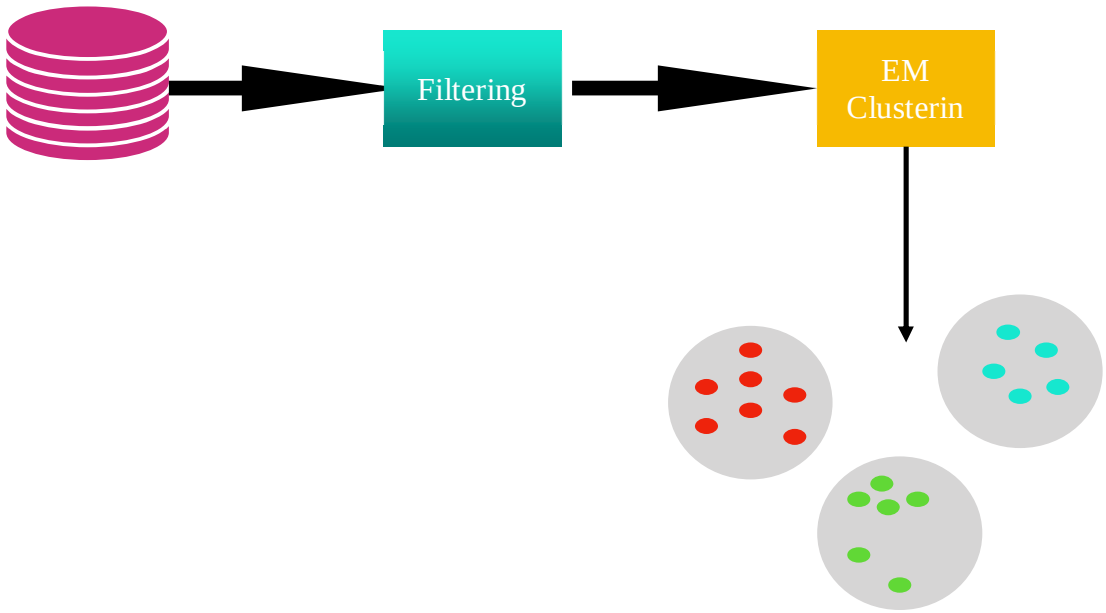


Figure 4 EM Clustering Technique

used to preprocess the input data from the database. Using EM Clustering techniques the data objects are grouped according the their similarities. Hence, through the below said example three various clusters are formed. So that the first group contains only red colour data objects, second group contains blue data objects and finally the last group contains green colour data objects. Like this the EM Clustering technique is processed.

D. DBSCAN Clustering

DBSCAN is a density based clustered algorithm similar to mean-shift, but with a couple of notable advantages. It is a non identical type of clustering technique with some unique benefits. As this clustering technique focuses more on the density and proximity of observations to generate clusters. This is totally different technique from other clustering techniques. This clustering technique based on the observation, it becomes a part of group represented by nearest centroid. This can discover outliers, it means that the observations which could not based on any other clusters. Moreover it can identify the number of clusters and also useful with unsupervised learning of the data objects.

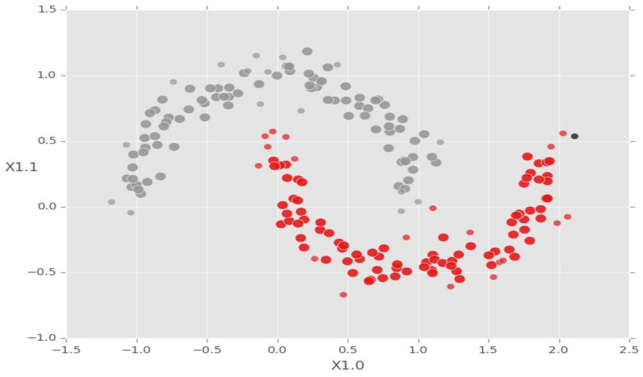


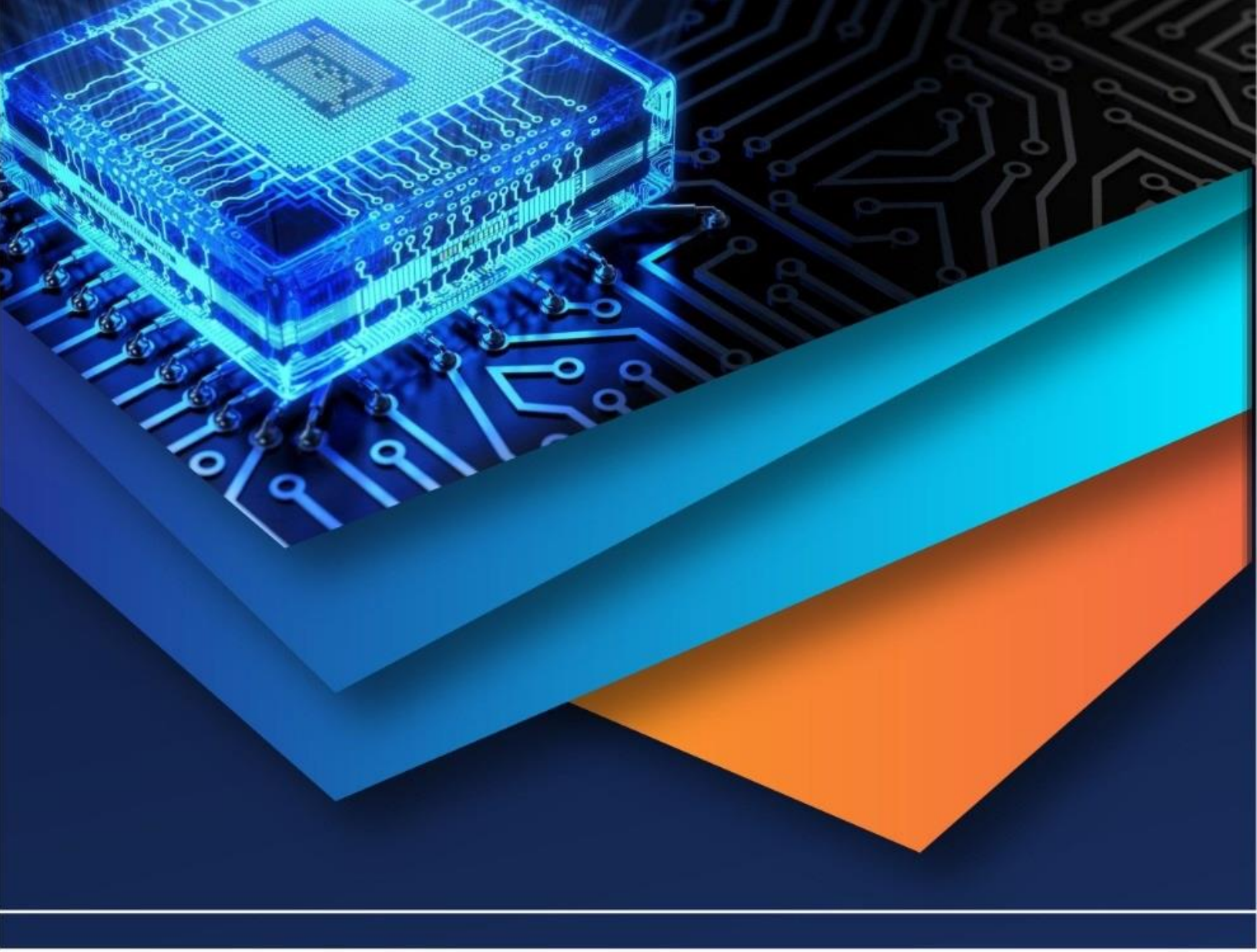
Figure 5 DBSCAN Clustering

IV. CONCLUSION

Clustering is one of the most important grouping technique which is used to discover the similar unlabelled data sets. For that many number of clustering techniques available in the data mining research fields. This survey discussed about the significance of clustering techniques and various kinds of techniques are discussed. Moreover some of the important clustering techniques like k-means clustering technique, EM Clustering technique and DBSCAN clustering techniques are discussed. This will help the researchers to select the real time obstacles. Moreover the general category of clustering describes in what way the clusters are generated. This will help to lead the researchers to choose the exact clusters inspite of the real time obstacles.

REFERENCES

[1] Pantelis n.Karamolegkos, Charalampos Z.Patrikakis Nikolaos D.Doulamis Panagiotis, “An Evaluation Study of Clustering Algorithms in the Scope of user Communities Assessment” Computers $ Mathematics with Applications, Elsevier, Vol No 58, issue no 8, October 2009, Pages 1498 - 1519.
[2] Man Abdel - Maksoud, Mohammed Elmogy, Rashid Al-Awadi, “Brain Tumor Segmentation Based on a Hybrid Clustering Technique”, Egyptian Informatics Journal, Vol No 16, Issue no 1, March 2005, Pages 1 - 81.
[3] Madjid Khalilian, Norwati Mustapha, Data Stream Clustering: Challenges and Issue, Proceedings of the International Multi conference of Engineers and Computer Scientists 2010 Vol No1, IMECS 2010,March 17-19 2010.
[4] Maryam Mousavi1 , Azuraliza Abu Bakar, and Mohammadmahdi Vakilian, “Data Stream Clustering Algorithms: A Review”, International Journal of Advance Soft Computer Applications Vol o 7, Issue No 3, November 2015, ISSN 2074-8523.
[5] Jose R. Fernandez,” A Framework and Algorithm for Data Stream Cluster Analysis”, International Journal of Advanced Computer Science and Applications, Vol No 2, Issue No11, Pages 87, 2011.
[6] Twinkle B Ankleshwaria, Twinkle B Ankleshwaria, Mining Data Streams: A Survey, International Journal of Advance Research in Computer Science and Management Studies, Vol No 2, Issue No 2, Feb 2014, ISSN: 2321-778.
[7] Amineh Amini, Teh Ying Wah, “Density Micro-Clustering Algorithms on Data Streams: A Review”, Proceedings of the International MultiConference of Engineers and Computer Scientists 2011 Vol No 1, IMCES 2011, March 16-18, 2011.
[8] Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., de Carvalho, A. C. P. L. F., and Gama, J, “Data stream clustering: A survey”, ACM Computing Surveys, Vol No 46, Issue No1, Article 13, October 2013, Pages 31.
[9] DoniaAugustine, “A Survey on Density based Micro-clustering Algorithms for Data Stream Clustering”, International Journal of Advanced Research in Computer Science and Software Engineering Research, Vol No 7, Issue No 1, January 2017.
[10] Dure Supriya Suresh, Prof. Wadne Vinod, “Survey Paper on Clustering Data Streams Based on Shared Density between Micro-Clusters”, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Vol No 04 ,Issue No 01, January 2017.
[11] Amini A, Wah TY, Saboohi H, “On density-based data streams clustering algorithms: A survey”,Journal of Computer Science and Technology, Pages 116–141, January 2014, DOI 10.1007/s11390-013-1416-3.
[12] Safal V Bhosale, “A Survey: Outlier Detection in Streaming Data Using Clustering Approache”, International Journal of Computer Science and Information Technologies, Vol No 5, 2014, 6050-6053 ISSN 0975 - 9646.
[13] Prashant V. Desai, Vilas S. Gaikawad, “Novel approach for data stream clustering through micro-clusters shared Density”,International Journal of Computer Sciences and Engineering Volume-5, Issue-1 E-ISSN: 2347-2693.
[14] M.S.B.PhridviRaj, C.V.GuruRao, “Data Mining - Past, Present and Future - A Typical Survey on Data Streams”, Elsevier Procedia Technology”, Vol No 12, 2014, Pages 255 - 263.
[15] Yisroel Mirsky, Bracha Shapira, Lior Rokach, and Yuval Elovici, “pcStream: A Stream Clustering Algorithm for Dynamically Detecting and Managing Temporal Contexts”, Springer International Publishing Switzerland 2015, PAKDD 2015, Part II, LNAI 9078, pp. 119–133, 2015. DOI: 10.1007/978-3-319-18032-8_10.
[16] Shufeng Gong, Yanfeng Zhang, Ge Yu1, “Clustering Stream Data by Exploring the Evolution of Density Mountain”, PVLDB, 11(4) 2017. DOI: 10.1145/3164135.3164136.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY