ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(48), DOI: 10.17485/ijst/2016/v9i48/106000, December 2016* ISSN (Online) : 0974-5645

An Efficient Incremental Density based Clustering Algorithm Fused with Noise Removal and Outlier Labelling Technique

Pooja Yadav, Anuradha and Poonam Sharma

Department of Computer Science, The North Cap University, Gurgaon – 122017, Haryana, India;
27pooja1992@gmail.com, anuradha@ncuindia.edu, poonamsharma@ncuindia.edu

Abstract

Objectives: Due to advancements in storage technology, every moderate to large sized organization is keeping huge amount of multi facet data which is growing very fast. To deal with such enormous data, we need an efficient data analysis technique like classification and clustering. **Methods/Statistical Analysis:** Processing high dimensional data sets in the presence of noise and outliers can degrade the performance of any kind of data analysis task. The situation can even worse; if we are going for unsupervised classification (i.e. clustering).In this paper, we proposed a new method for incremental density based clustering for high dimensional data set with reasonable speed up. The proposed method fused with noise removal and outlier labeling technique is inspired from famous box plot method. **Findings:** The performance analysis of fusion is done on five high dimensional data sets taken from University California Irvine (UCI) repository along four cluster evaluation metrics (F-Measure, Entropy, Purity and Speed Up).The produced clustering results confirm the effectiveness of proposed fusion. **Application/Improvements:** The proposed technique can be refined by hybridizing it with some meta-heuristic technique for stock exchange application.

Keywords: Box Plot, Density Based Clustering, DBSCAN, , Entropy and Incremental Partitioning

1. Introduction

The objective of clustering process is to find similarity patterns among a given population so that similar elements can be bounded together. The manifold application of clustering can be seen from search engine optimization to market analysis through bio-informatics. Finding patterns of interest in the presence of noise and random variation is difficult to investigate. Due to unsupervised nature, there are many inherited challenges associated with clustering process, which can be categorized as: 1) What should be the ideal choice for the number of clusters 2) How to find correlation and noise present in the data set. 3) How to deal with high dimensional data set in a dynamic environment etc.

Data classification is also important data analysis technique but performs poor in the presence of random variability and noise. The unsupervised nature of clustering process makes it far superior than other state of art classification methods. A glimpse of few challenges and improvements in the field of data classification can be referred from references[1-3].

Incremental density based clustering algorithms faces a different set of performance issues that includes dynamically updating clusters with change in original data set. But, the presence of noise and outliers can ruin the clustering process. Apart from this, these algorithms are time consuming and suitable for real time data analysis. The major challenge lies in migrating objects from one cluster to other as new objects are introduced in the system. Detailed review of famous clustering techniques with their pros and cons can be found in many recent review articles[4].

To reduce the search space and facilitating incremental update in clusters, a new density based clustering algorithm[5] is proposed. Their algorithm utilizes DBSCAN algorithm to separate out dense regions and how to upgrade them in dynamic environment. At later stage, to

* *Author for correspondence*

refine the performance some neighborhood clusters are merged together to speed up the clustering process. But, there are few limitations that can be identified: 1) Lack of suitable noise removal and outlier detection technique 2) Not very much suitable for high dimensional data set with missing values 3) The merge process to combine dense regions is done simply on local information associated within given partition. The authors claimed the speed up by a factor 3.2 in comparison to existing incremental density based clustering techniques. So, here in this paper, we modified DBSCAN algorithm by fusing it with a new noise removal and outlier labeling method. The basic assumptions at various stages are taken from original readings.

2. Literature Review and Related Work

The task of unsupervised classification to separate out similar from dissimilar is known as clustering. Numerous authors have presented various tools and techniques for efficient clustering. Each of them has contributed in their own way to explore some new set of research dimensions. Clustering can be broadly categorized as static or dynamic clustering. In static clustering a fixed number of clusters are generated by partitioning a given fixed data set, but in dynamic environment clusters are being reorganized to deal with expanding data set.K-Means[6], PAM[7], BIRCH[8], and CLARANS[9]are the popular static clustering methods which are able to deal with random shape and sizes of clusters. But, these can perform poor by stucking into local optimum. The next breed of clustering methods is hierarchical clustering with single and multi-linkage connectivity. CURE[10] and Chameleon[11] are one of best performing hierarchical clustering algorithms but helpless to deal with object migration in dynamic environment and high computational complexity. The third clustering generation is density based clustering[12], DBSCAN[4], SSN[4] and OPTICS[4] are popular density based algorithm which achieves quality clustering results by forming some dense regions of data objects separated by low density regions.

The important research point is to investigate the challenges associated with dynamic cluster update, computational complexity and the presence of noise and outlier. Recently, a kernel based clustering[13] method to deal with dynamic information is discussed. The modification is done through popular K-Means algorithm. A bottom up hierarchical clustering[14] was proposed to rebuild concepts by rearranging object ordering. An incremental dynamic update clustering method[15] to support self-assertive shape and sizes was found to have remarkable performance. A pair wise document similarity clustering method[16] was proposed by forming similarity histograms. The key idea was to incorporate statistical significance of attributes[5].

Various authors have modified k-means algorithm to make it suitable for multiple applications[18-23].All the existing incremental density based clustering techniques works around following four stages: incremental partitioning of dataset, incremental DBSCAN of partitions, incremental merging of dense regions and noise removal and outlier labeling.

- Incremental partitioning of dataset
- Incremental DBSCAN of partitions
- Incremental merging of dense regions
- Noise removal and outlier labeling

3. Incremental DBSCAN Algorithm

Incremental DBSCAN algorithm is density based clustering algorithm that can distinguish discretionary shaped clusters. While static DBSCAN is connected to static datasets in which the presence of all items is required before running the calculation, incremental DBSCAN works by preparing objects as they come and overhaul/ make clusters as required. Given two parameters Eps and Minpts:

Definition 1: The Eps neighborhood of an object p is given by $N_{Eps}(p)$, is characterized by

$$N_{Eps}(p) = \left\{ \frac{q}{dis(p,q)} \leq Eps \right\}. \quad (1)$$

Definition 2: An object p is specifically density reachable from object q if p € $N_{Eps}(q)$ and $|N_{Eps}(q)| \geq Minpts$ (i.e. q is a center). (2)

Definition 3: A object p is density reachable from an object q in the event that there is a chain of items $p_1, p_2, \ldots, p_n$ with the end goal that p_{i+1} is straightforwardly density reachable from p_i and $p_1 = p$ and $p_n = q$.

Because of the density based nature of the incremental DBSCAN algorithm, the addition or deletion of an object influences just the items inside a specific neighborhood (Figure 1). Influenced objects are potentially the objects that may change their cluster participation after insertion/

deletion of an object p and they are characterized as the objects in the $N_{Eps}(p)$ additionally all different objects that are density reachable from objects in $N_{Eps}(p)$.

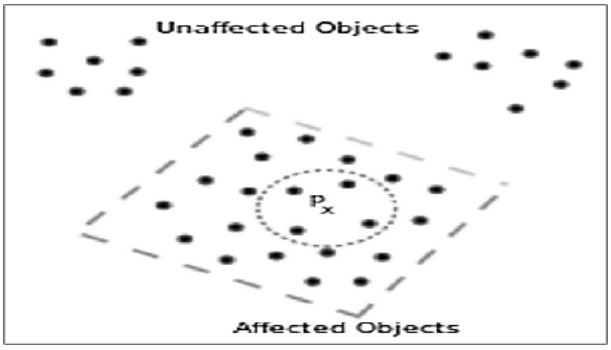

Figure 1. Affected Objects due to Insertion/Deletion of bject X.

3.1 Proposed Method

The main aim of our proposed work is to provide noise removal and outlier labeling for high dimensional data sets. In 2015, an incremental density based clustering algorithm[17]was proposed to incrementally make and update clusters in datasets. But the authors have not proposed any suitable technique for noise removal and outlier labeling. The efficiency of any clustering algorithm gradually decreases with the increase in dimensionality. So, in this paper we proposed a new method to detect noise and outliers in more efficient manner.

At first, the algorithm requires characterizing of K partitions before operating. The algorithm considers the primary K objects to be centroids of the K partitions are long as the separations between them are bigger than Eps. For the subsequent objects, the algorithm incrementally partitions the dataset to wipe out the versatility issue of checking the whole search space for finding the area of the new inserted objects. Taking after allotting the new object its closest partition, the algorithm creates/updates dense regions in this partition as per the incremental DBSCAN algorithm. After that, the algorithm incrementally consolidates the dense regions of various partitions in light of a given network measure to create/update the conceivable last cluster. At last, the algorithm marks outliers also and delete noise to produce the final cluster. Exploratory comes about demonstrate that the proposed calculation accelerates the clustering handle with a component up to 3.2 contrasted with important incremental clustering algorithms. Figure 2, highlights the detailed clustering framework used for experimental setup.

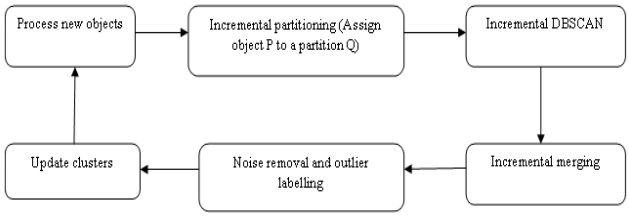

Figure 2. Detailed Clustering Framework.

3.2 Incremental Partitioning of Dataset

Incremental partitioning of items is expected to lessen the search space down object's neighborhood. Given another object p, the area is dictated by filtering objects of the closest partition instead of the whole dataset (search space). The partitioning algorithm divides the dataset into a predefined number of partitions and appoints the object p to the partition with the closest centroid. To start with, because of the dynamic way of the dataset, old items may change their positions after the inclusion of new objects; hence, the apportioning calculation ought to have the capacity to identify the current questions that turned out to be nearer to different centroids after inclusion of new question. Second, it gives what is called ''stability''; the partitioning algorithm achieves stability by using a learning rate that rots with time so that objects are reliably allotted to centroids and the centroids are changing with a little resistance. The consequence of the dividing stage is finding the closest centroid for the new protest p notwithstanding overhauling places of current centroids with regard to the utilized learning rate.

3.3 Incremental DBSCAN of Partitions

Incremental DBSCAN algorithm is utilized to upgrade dense regions in partitions. A dense region can be characterized as a set of objects that can frame a last cluster or be a piece of a last cluster.. Given the new object p, the addition module of DBSCAN (incDbscanAdd) is utilized to locate a dense locale at the closest partition that the object p can join. For the old objects which changed their partitions, the deletion module (incDbscanDel) is utilized to expel them from their old dense regions and addition module (incDbscanAdd) adds them to their new dense regions. The objective of this stage is upgrading dense regions in various partitions with the progressions happened in the partitioning stage.

3.4 Incremental Merging of Dense Regions

To frame the last arrangement of clusters, an extra stage is

required to combine dense regions of various partitions. Given two regions A and B in two unique partitions, A is converged to B if the between network IE(A,B). Figure 3,clarifies the possibility of the between network between two dense regions. Expect there is an imaginary edge between two border objects in two distinctive dense regions if the separation between these objects is not exactly the Eps characterized by the incremental DBSCAN algorithm, then the interconnectivity between dense regions A and B is characterized as IE (A,B) where

$$IE(A,B) = \frac{N_{ab}}{(N_a + N_b)/_2} \quad (3)$$

Where, N_a and N_b are the quantity of edges that associate visitor objects in areas A and B individually and Nab is the number of edges that associate visitor objects from dense regions A and to objects in dense region B.

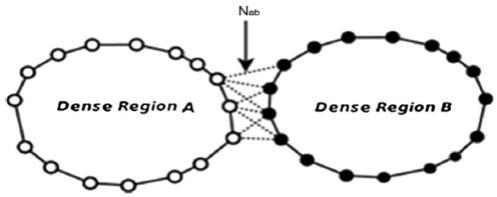

Figure 3. Interconnectivity between Two Regions.

3.5 Noise Removal and Outlier Labeling

A last step is taken to eliminate noise from the dataset. The proposed algorithm can recognize and take out the noise where noise objects are neither one of the classifieds objects (not relegated to a cluster) nor reachable from any center objects in the clusters. As Figure 4 illustrating that noise objects are not from the classified objects and even they are far away from center objects in the clusters. So the outliers are labeled in the mentioned figure. In any case, objects whose separation to any cluster is not exactly Eps are considered as exceptions and they are named by the mark of the closest cluster.

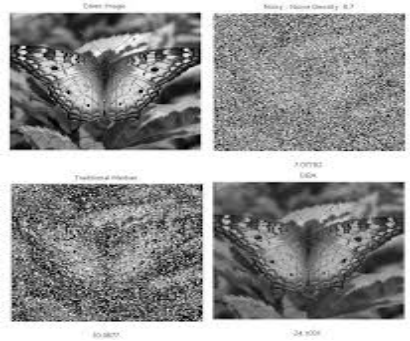

Figure 4. An illustrative Example Showing Points of Interest for Noise Removal and Outlier Labelling.

Explained below is the detailed algorithm for incremental clustering fused/embedded with noise removal and outlier labeling method.

Input: Set of new objects X, centroids

Output: Clusters after removing noise and outlier labeling

1: A:- List of object that may change their centroids
2: N:- List of updated dense regions
3: For each x_i in X do
4: q: nearest_centroid (p_i)
5: A: update_centroids (p_i, q)
6: Add xi to A
7: **end for**
8: For each z_i in A do
9: q_n: z_i new_centroid
10: q_o: z_{iold}–centroid
11: Apply incDbscanDel to remove z_i from q_o
12: Apply incDbscanAdd to insert z_i to q_n
13: Add updated dense regions to N (step 11 and 12)
14: **end for**
15: For each n_i in N do
16: For each n_j in N and i ≠ j do
17: If inter_connectivty (n_i,n_j) > α merge
18: merge(n_i, n_j)
19: **IQR=T3-T1**
20: Inner fences=[T1-1.5IQR, T3+1.5IQR]
21:Outer fences= [T1-3IQR, T3+3IQR]
22: end if
23: end for
24: end for

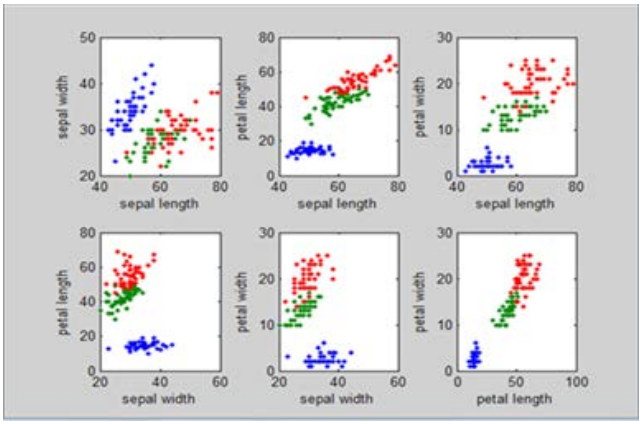

Figure 5. Snapshot of Cluster Formation on IRIS Data Set Fused with Proposed Method.

Figure 5 is a snapshot of cluster formation on iris

data set after applying proposed noise removal and outlier detection method. Outliers may cause a negative effect on data analysis such as regression (a measure of the relation between the mean value of one variable and corresponding value of other variables). There are some methods of outlier labeling, such as Standard deviation and Tukey method(Boxplot). In this paper, we use the boxplot method to detect noise and outliers.

The standard deviation and boxplot methods are easy to use, and are quite reasonable when the data distribution is symmetric and mount-shaped such as normal distribution.

3.6 Boxplot Method

It makes no distributional assumption, also it does not depend on a standard deviation or mean. In this method, we need to calculate the lower quartile(t1) and upper quartile(t3) and IQR(inter-quartile range) is the interval between t1 and t3(T3-T1).

This method is effective when working with large datasets that are normally distributed. It uses quartiles which are resistant to extreme values.

- IQR= T3-T1(distance between lower and upper quartiles).
- Inner fences are located at a distance 1.5IQR below T1 and above T3.[T1-1.5IQR, T3+1.5IQR]
- Outer fences are located at a distance 3IQR below T1 and above T3.[T1-3IQR, T3+3IQR]
- A value between the inner and outer fences is a possible outlier and an extreme value beyond the out fences has been neglected.

4. Experimental Result and Discussion

This section presents detailed experimental analysis carried out to prove our proposed clustering technique better in comparison to other state of art methods used for high dimensional clustering. We have taken five high dimensional data sets from UC Irvine repository (refer Table 1) to test the performance in terms of F-Measure, Purity, Entropy and speed up. Our proposed noise removal and outlier labeling method is compared with an incremental density based clustering algorithm proposed by Ahmad M.Bakr, Nagia M.Ghanem, Mohamed A.Ismail[17]. Presented below is the brief discussion about evaluation metrics used for evaluating clustering results:

Table 1. Detailed Description of UCI Datasets

Dataset	No. of instances	No. of attributes	Data types
Adult	48842	14	Multivariate
Bank marketing(BM)	45211	17	Multivariate
Census income(CI)	48842	14	Multivariate
Chronic kidney disease(CKD)	400	25	Multivariate
Cylinder bands(CB)	512	39	Multivariate

Definition 3:F-measure (FM) is used to evaluate the accuracy of the final clusters. The two basic measures are recall and precision. FM ranges from 0 to 1.

$$FM = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (4)$$

Definition 4: Entropy is the measure of impurity. The clusters are compared with the true label data sets and the value of entropy is calculated. The lower entropy indicates better clustering and higher entropy means the clustering is not good.

$$E(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i) \quad (5)$$

The entropy function is at zero minimum when p=1 or p=0.

Definition 5: Purity is an evaluation criteria of cluster quality. It is defined as the percent of the total number of objects that were classified correctly in the unit range 0 to1.

As, it can be seen from Table 2 that the entropy measurements on all the data sets are lowered in comparison to existing technique proposed by Ahmad M.Bakr et. al. Only, CI data set is showing less margin gap (0.46) and maximum reduction is achieved by BM data set (0.513).

Table 2. Entropy Measurement Across All Datasets

UCI Data sets	Entropy (Proposed byAhmad M.Bakr et. al, 2015)	Proposed Method Entropy
Adult	0.856	0.515
BM	0.734	0.221
CI	0.877	0.829
CKD	0.501	0.141
CB	0.475	0.129

Table 3. Purity Measurement Across All Datasets

UCI Data sets	Purity (Proposed byAhmad M.Bakr et. al, 2015)	Proposed Method Purity
Adult	0.479	0.765
BM	0.485	0.892
CI	0.781	0.905
CKD	0.564	0.868
CB	0.765	0.814

Table III, highlights the purity statistics measured across all the datasets. CB data set is showing slight increase in purity (0.39), whereas maximum increase is found in BM data set (0.407). The rest of the data sets also showed a reasonable increase in purity in comparison to existing one. A reasonable increase in purity claims that our proposed method is able to capture noise present in the training data set.

Table 4. F-Measure (FM) Measurement Across All Datasets

UCI Data sets	F-Measure (Proposed by Ahmad M.Bakr et. al, 2015)	Proposed Method F-Measure
Adult	0.475	0.845
BM	0.431	0.847
CI	0.612	0.903
CKD	0.792	0.810
CB	0.552	0.717

Table IV presents the F-Measure values recorded for all the data sets. A high value of F-Measure proves better quality of clustering process. A significant improvement is found on all data sets except CKD data set. The maximum increase is observed in both Adult and BM data sets. The improvement in F-Measure proves our proposed method to be efficient in terms of noise removal and outlier labelling.

Table 5. Speed Up Measurement Across All Datasets

UCI Data sets	Time Taken (In Seconds) (Proposed by Ahmad M.Bakr et. al, 2015)	Time Taken by Proposed Method (In Seconds)
Adult	5476.9	2045.2
BM	5079.6	3276.6
CI	6990.7	4008.1
CKD	5688.9	3192.2
CB	7886.9	4214.5

Apart from entropy, purity and F-Measure, our proposed method is enable us to achieve good clustering results in reasonable time.Table V, highlights the time taken by our proposed approach in comparison to existing clustering technique.Adult and CB data sets are showing maximum reduction in computation time.

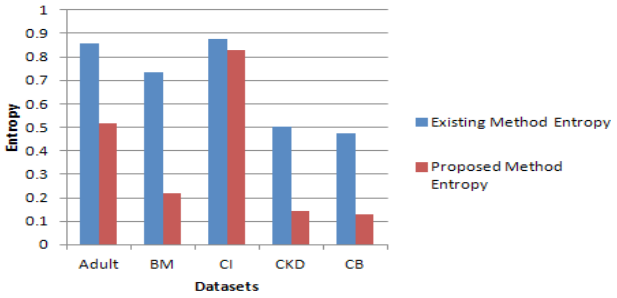

Figure 6. Entropy Comparison Across all Datasets.

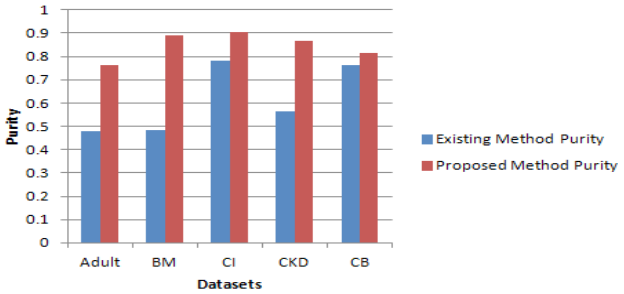

Figure 7. Purity Comparison Across all Datasets.

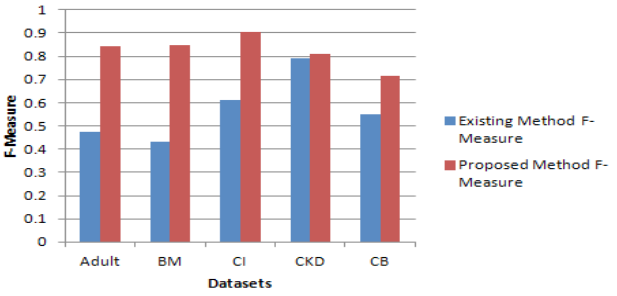

Figure 8. F-Measure (FM) Comparison Across all Datasets.

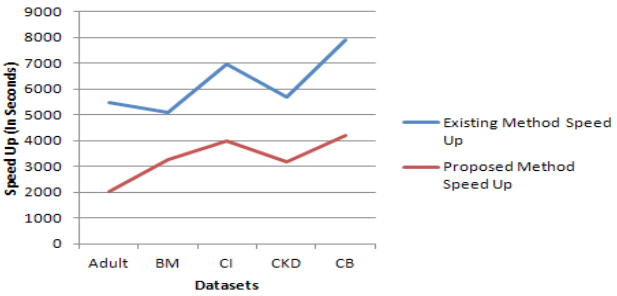

Figure 9. Speed Up Comparison Across all Datasets.

It can be easily observed from Figure 6 -Figure 9 that our proposed clustering method for noise removal and outlier labeling is well suited for high dimensional data sets and it outperforms the existing state of art methods.

5. Conclusion

In this paper, an incremental density based clustering algorithm is proposed which is fused with a suitable noise removal and outlier detection technique inspired from box plot method. The algorithm incrementally partitions the dataset to diminish the search space to every segment as opposed to examining the entire dataset. After that the algorithm incrementally structures and updates dense regions in every segment. After taking recognizing conceivable thick districts in every segment, the algorithm utilizes a between network measure to dense regions to frame the last number of clusters. The experimental analysis proves the effectiveness of clustering results by giving possible speed up.

6. References

1. Anuradha, Gaurav Gupta. Fuzzy Decision Tree Construction in Crisp Scenario Through Fuzzified Trapezoidal Membership Function. Internetworking Indonesia. 2015 March; 7(2):21–8.
2. Anuradha, Gaurav Gupta. MGI: A New Heuristic for Classifying Continuous Attributes in Decision Trees. International Conference on Computing for Sustainable Global Development (INDIACom), 2014.p. 291––95.
3. Anuradha, Gaurav Gupta. A Self Explanatory Review of Decision Tree Classifiers. IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), Jaipur, India: May 09–11, 2014.p. 1––7.
4. Nagpal A,. Jatain A,. Gaur D. Review based on data clustering algorithms. IEEE Conference on Information & Communication Technologies (ICT). April 2013.p. 298–303.
5. Ahmad M. Bakr, Noha A. Yousri, Mohamed A. Ismail. Efficient incremental phrase-based document clustering. International Conference on Pattern Recognition ICPR. Nov2012.p. 517–20.
6. Zhe Zhang, Junxi Zhang, Huifeng Xue. Improved K-means clustering algorithm.Congress on Image and Signal Processing CISP. 2008 May; 5:169–72.
7. Li L, You J, Han G, Chen H. Double partition around medoids based cluster ensemble. International Conference on Machine Learning and Cybernetics. 2012 July; 4:p. 1390–94.
8. Du H, Li Y. An improved BIRCH clustering algorithm an application in thermal power. International Conference on Web Information Systems and Mining (WISM). 2010 Oct; 1:p. 53–6.
9. Ng RT, Han J. CLARANS: a method for clustering objects for spatial data mining. IEEE Trans. Knowl. Data Eng. 2002; 14 (5): 1003–16.
10. Guha S. Rastogi R. Shim K. CURE: an efficient clustering algorithm for large databases. Proceedings of the ACMSIGMOD International Conference Management of Data (SIGMOD'98), 1998 Oct;p. 73–84.
11. Karypis G, Eui-Hong H, Kuma V. Chameleon: Hierarchical clustering using dynamic modeling. Computer.1999; 32 (8): 68–75.
12. Kriegel HP, Pfeifle M. Effective and efficient distributed model-based clustering. Proceedings of the 5th International Conference on Data Mining (ICDM'05), 2005. 265–85.
13. Yu W. Qiang G. Xiao-Li L. A kernel aggregate clustering approach for mixed data set and its application in customer segmentation.International Conference on Management-Science and Engineering ICMSE.t 2006 Oct;p. 121–4.
14. Nafar Z, Golshani A. Data mining methods for protein–protein interactions. Canadian Conference on Electrical and Computer Engineering. CCECE. 2006 May;p. 991–4.
15. Nithyakalyani S, Kumar SS. Data aggregation in wireless sensor network using node clustering algorithms a comparative study. IEEE Conference on Information & Communication Technologies (ICT). 2013 April;p. 508–13.
16. Hammouda KM, Kamel MS. Efficient phrase-based document indexing for web document clustering. IEEE TransKnowledge and Data Eng. 2004 Oct;16(10): 1279–96.
17. Ahmad M. Bakr, Nagia M. Ghanem, Mohamed A. Efficient incremental density-based algorithm for clustering large datasets. Alexandria engineering journal.2015; 54 (4):1-8.
18. Suganya M. Nagarajan S.Message Passing in Clusters using Fuzzy Density based Clustering.Indian Journal of Science and Technology.2015 July;8(16): DOI:10.17485/ijst/2015/v8i16/61761
19. Gayathri S, Mary Metilda M, Sanjai Babu S.A Shared Nearest Neighbour Density based Clustering Approach on a Proclus Method to Cluster High Dimensional Data.Indian Journal of Science and Technology. 2015 Sep; 8(22): DOI:10.17485/ijst/2015/v8i22/79131
20. Pavel V, Skribtsov, Sergey O, Surikov, Michael A, Yakovlev. Background Image Estimation with MRF and DBSCAN Algorithms.Indian Journal of Science and Technology. 2015 Dec; 8(10): DOI:10.17485/ijst/2015/v8is(10)/85409
21. Razia Sulthana A, Ramasamy Subburaj.An Improvised Ontology based K-Means Clustering Approach for Classification of Customer Reviews.Indian Journal of Science and Technology. 2016 Apr; 9(15): DOI:10.17485/ijst/2016/v9i15/87328
22. Naveen A, Velmurugan T.Identification of Calcification in MRI Brain Images by k-Means Algorithm.Indian Journal of Science and Technology. 2015 Nov; 8(29): DOI:10.17485/ijst/2015/v8i29/83379
23. Dharmarajan A, Velmurugan T.Lung Cancer Data Analysis by k-means and Farthest First Clustering Algorithms. Indian Journal of Science and Technology.2015 July; 8(15):DOI:10.17485/ijst/2015/v8i15/73329