International Journal of Advanced Research Methodology in Engineering & Technology,
Volume 1, Issue 1, December 2016

SPARSE HIGH DIMENSIONAL DATA OBJECT CLUSTERING ALGORITHM

R.Pushpalatha[1], Dr.K.Meenakshi Sundaram[2]

1 Ph.D. Research Scholar in Computer Science, Erode Arts and Science College, Erode and Assistant Professor, Department of Computer Science, Kongu Arts and Science College (Autonomous), Erode.
(Email Id.: rpljour@gmail.com)

2 Associate Professor, Department of Computer Science, Erode Arts and Science College, Erode.
(Email Id.:lecturerkms@yahoo.com)

ABSTRACT

Data mining (DM) is used for extracting useful and non-trivial information from large amount of data. Cluster analysis is used to form logical group of similar data, widely applied in many practical applications such as weather forecast, share trading, medical data analysis, aerial data analysis, etc., Clustering in data mining is an unsupervised learning model. Clustering techniques in handling high dimensional data is bit complex due to intrinsic sparsity of high dimensional nature of data. However, existing methods to prune irrelevant clusters were based on spectral clustering algorithm and graph-based learning algorithm,whose lack of sparsity and polynomial time complexity. In this paper to cluster sparsely distributed high dimensional data objects, Fuzzy Relational Scattered Distance based Clustering (FRSDC) technique is developed. The main objective of the FRSDC technique is to specify the clustering data points over sparsely distributed data within limited processing time. Fuzzy relational in FRSDC technique calculate the geometric median of sparsely distributed high dimensional data and determine clustering objects to be placed on each cluster. Initially, FRSDC identify the geometric median of similar sparse data and then the non selected sparse data objects appropriate the relational fuzziness across data points, reducing subspace of data objects in clustered plane. Next, Scattered Distance measures the distance of geometric median (i.e.,) inner object and similar object position (i.e.,) outer object and computes the probability distribution function while performing clustering. Finally, Scattered Distance with grid form is used to compute the area of the cluster in FRSDC and therefore obtains the clustered sparse data. The space complexity of each algorithm is analyzed and the results are compared with one another. By comparing the result of this technique, it was found that the results obtained are more accurate, easy to understand and above all the time taken to cluster the data was substantially low in FRSDC technique than the state-of-art methods.

Keywords: Data mining, Cluster analysis, Fuzzy relational scattered distance, Performance of FRSDC.

1.INTRODUCTION

With the growing interest in the information technology, the size of data is also getting increased including financial institutions, electricity board, educational department and so on. Therefore, there arises a great demand to cluster the data and use them according to their requirements. Many literary works have therefore contributed a lot in this area. Kernelized Group Sparse graph (KGS-graph) was introduced in [1] to express the contextual information of a data manifold. KGS-graph successively preserves the properties of sparsity and locality simultaneously and demonstrates the effectiveness. However, sparse graph construction does cannot satisfy the locality constraint and does not merge nonzero coefficients locality and sparsity. Clustering Algorithms for Probabilistic Graphs (CA-PG) [2] addressed the problem of clustering correlated probabilistic graph to improve clustering efficiency. However some recent

studies [3] [4] have pointed out that the multi-dimensional and inter dimensional space can also be correlated to reduce the computational complexity. Different clustering technique may produce different results. This research work extends the work of two of the distance based clustering techniques as stated. With these discussions, the next section discusses the proposed technique Fuzzy Relational Scattered Distance based Clustering with the objective of clustering for sparse high dimensional data objects. The rest of the paper is structured as follows. Section 2 presents some of the related works and their approaches. In Section 3, the basics of Fuzzy Relational Scattered Distance based Clustering with the algorithm are described. Section 4 discusses about the experimental setup of the proposed technique. Section 5 covers the experimental studies, results, and a brief discussion. Finally, Section 6 presents the conclusions of the research work.

2.RELATED WORKS

Motivated by the fact that different groups of points assume some cluster relationship among the data objects, a significant amount of research has been elaborated upon similarity measure, which aims at discovering the similarity between pair of objects. In [9], multiple view points were considered during clustering by applying optimization algorithm resulting in improving the clustering results in accuracy. Clustering by Discrimination Information Maximization (CDIM) [10] to produce high quality clusters was designed using Domain Relevance (DR) and Domain Consensus (DC). A similarity model with Ant Colony Optimization was introduced in [11] to improve cluster accuracy. In [12], an enhanced K-Means clustering algorithm was presented to improve the cluster quality by applying subspace clustering. A mechanism for analyzing the impact parameter for subspace clustering was presented in [13]. Extreme Learning Machine (ELM) on high dimensional data was analyzed in [14]. Cluster analysis has been used in different areas to assign characteristics of an observation into clusters so that those data object in the same group are more similar than when compared to the presence of data object in other groups. A comparison study on similarity and dissimilarity measures in clustering continuous data was investigated in [16]. In [17], Constraint Partitioning K-Means algorithm was designed to produce good and accurate clusters using Principal Component Analysis. In [18], a classification algorithm for high dimensional data was applied to handle large scale high dimensional data Analysis. Clustering of high dimensional data with the aid of local input space histograms was presented in [19]. In [20], ensemble learning was applied for high dimensional data to reduce computational complexity involved during clustering.

3.FUZZY RELATIONAL SCATTERED DISTANCE BASED CLUSTERING

Fuzzy Relational Scattered Distance based Clustering (FRSDC) for sparsely distributed data is introduced by specifying the cluster data points within limited processing time. We begin by describing the overall architecture, followed by problem statement for FRSDC for clustering data points and then present the technique FRSDC. Figure 1 shows the architecture of Fuzzy Relational Scattered Distance based Clustering.

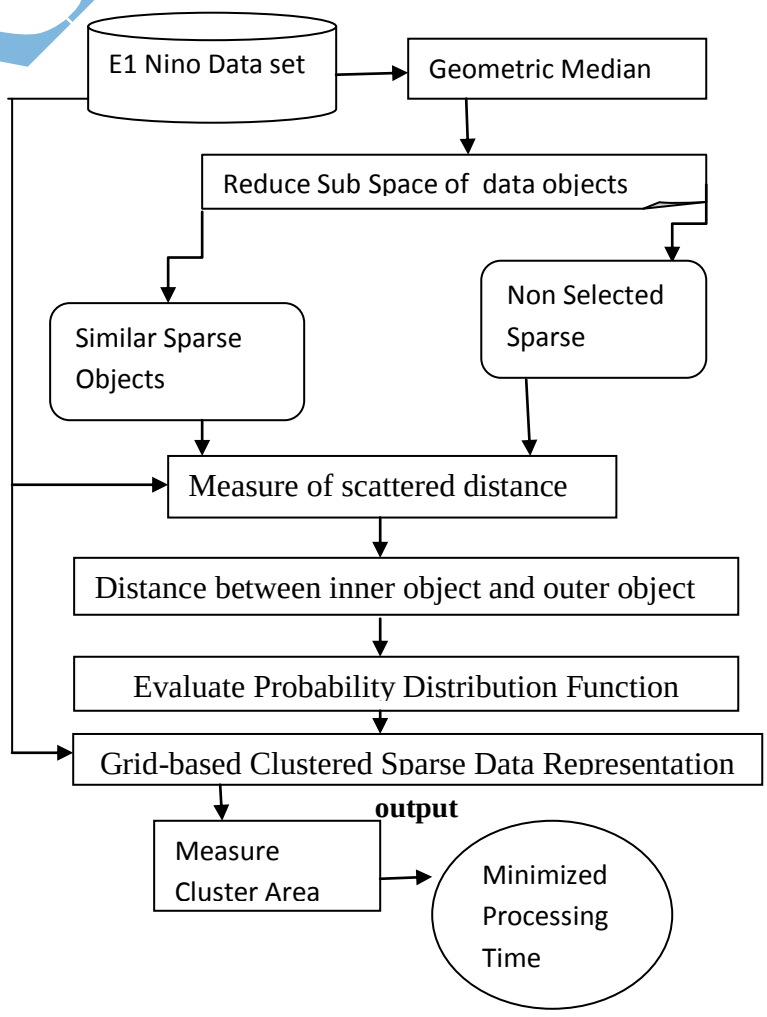

Figure 1 Fuzzy Relational Scattered Distance based Clustering

As shown in the figure, E1 Nino dataset is given as input to the FRSDC technique. To start with, geometric median is calculated to specify the clustering data points for the corresponding sparsely distributed high dimensional data. This geometric median on similar sparse data (air temperature, sea surface temperature, subsurface temperature) and non selected sparse data objects (date, latitude, longitude) are evaluated. Next, scattered distance measure is used to measure the distance between the inner and outer objects respectively. Finally, a grid form is applied to measure the area of the cluster and to determine the clustered sparse data.

3.1.Problem statement

Let us consider the problem of identifying representatives from a collection of 'd' dimensional data points '$P = p_1, p_2, \dots, p_n$' in data space 'S' from ordered domains '$D = \{d_1, d_2, \dots, d_n\}$', where '$p_i = p_{i1}, p_{i2}, \dots, p_{in}$'. Let '$r(p_i, p_j)$' represents the dissimilarity between data points 'p_i' and 'p_j'. Therefore the 'nth' component of 'p_i' is obtained from domain 'd_n'. Let us further assume that we are given a set of nonnegative dissimilarities '$d_{ij}, where\ i, j = 1,2, \dots n$' between every pair of data points 'i' and 'j' respectively. The dissimilarity 'd_{ij}', indicates how well the data point 'i' is suited to be a representative of the data point 'j'. Let '$Y = y_1, y_2, \dots, y_n, y_i \in P$', where '$Y$' is a subset of '$P$'.

3.2.Geometric Median Based Fuzzy model

The main objective of FRSDC is to specify the clustering data points over the sparsely distributed data within the limited processing time. Initially, Geometric Median Based Fuzzy model identifies the geometric median of similar sparse data and then the non selected sparse data objects to appropriate the relational fuzziness across data points. The non selected sparse objects in Geometric Median Based Fuzzy model combined with more similar geometric median reduce the subspace of data objects in the clustered plane.

In this section Geometric Median-based Fuzzy model is designed in which the fuzzy relational classifies a data point, with the cluster centroid to be closest to the data point of membership. The membership is specifically required when the boundaries among the clusters are not well separated and restricts the processing time. At the same time, the membership assists in locating more advanced relations between a given object and the disclosed clusters. The membership function in Geometric Median Based Fuzzy with minimization function 'K_{min}', 'Y' being a subset of 'P'is as given below.

$$K_{min}(Y, P) = \sum_{i=1}^{n} \sum_{j=1}^{m} d_{ij}^{r}\, Dis_{ij}$$

Finally the geometric median of similar sparse data and non selected sparse objects helps in reducing the subspace of the data objects in the clustered plane. Figure 2 shows the algorithmic description of Geometric Median Based Fuzzy model.

Input: data points '$P = p_1, p_2, \dots, p_n$', data space 'S', ordered domains '$D = \{d_1, d_2, \dots, d_n\}$',
Output: Differentiates selected and non selected sparse objects reducing space complexity
1: Begin 2: For each 'n' data points 3: Measure membership function 4: Measure geometric median of similar sparse data 5: Measure non selected sparse objects 6: End for 7: End

Figure 2 Geometric Median Based Fuzzy algorithm

For a given E1 Nino dataset and for each 'n' dimensional data points, the Geometric Median Based Fuzzy algorithm measures the membership function based on the fuzzy partition matrix and the distance measure formula. Followed by this, the similar sparse data is obtained through geometric median by minimizing the sum function. Finally, the non selected sparse objects are evaluated to obtain the relational fuzzy.

3.3.Scattered Distance

Once, the similar selected and non selected sparse objects are differentiated by applying Geometric Media, a distance metric between data points is used to measure the distance between the inner and outer object. In the proposed work, a measure of Scattered Distance is applied. The Scattered Distance measures the distance of the geometric median (i.e.,) inner object and the similar object position (i.e.,) outer object.

$$T = W(CS) + B(CS)$$

Here '$W(CS)$' symbolizes the within cluster scatter whereas '$B(CS)$' symbolizes the between cluster scatter. The geometric median (i.e.,) inner object or within cluster scatter and outer object or between cluster with similar object position is mathematically given as below.

$$W(CS) = \sum_{Cl(i=K)}(p_i - r_k)^2 \quad (1)$$

$$B(CS) = \sum_{Cl(i=K)}(r_k - m)^2 \quad (2)$$

From (1) and (2), the geometric median (i.e.,) inner object or within cluster scatter and outer object or between cluster with similar object position is mathematically obtained. Given two uncertain objects 'p' and 'r', their corresponding probability distributions.Figure 3 shows the probability distributions of two uncertain objects 'p' (in red colour) and 'r' (in violet colour).

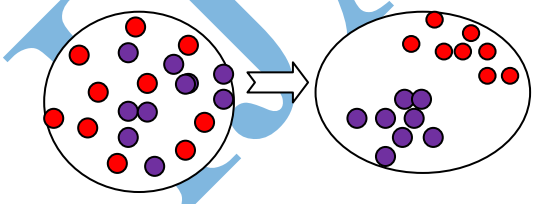

Figure 3 Probability distributions and cluster

The distribution difference between inner object and outer object cannot be extracted by geometric median. The proposed work uses probability distributions. In figure 3 (a), the locations of the two object (denoted by violet and red color) overlapping different distributions are presented, whereas in figure 3 (b) the two objects have different locations.

3.4.Grid-based Clustered Sparse Data Representation

Different groups mapped into different clusters are discretized by partitioning it into a grid. Scattered Distance with grid form is used to compute the area of the cluster in FRSDC technique. Every dimension is partitioned into '2' portions. So, an 'n' data is divided into '$2n$', cells with each cells possessing equal size. The central points of cells in the grid are used as values in the ordered domain. The probability of an object in a cell in the grid is the sum of the probabilities of all its sample points in this cell.

Given the clusters 'n' for a dataset, 'n' probability density function is generated with one uniform distribution '$(n-1)/2$ 'distributions with different distance measures. For each distribution, the proposed technique generated a group of samples that specifies clustering data points, each of which forms one clustering object. Therefore, clustering objects in the same group are sampled from the same probability density function. In this way, the Grid form in FRSDC determines the clustered sparse data and therefore significantly improves the clustering accuracy.

4.EXPERIMENTS

In this section, we evaluate the effectiveness of our proposed technique FRSDC and compare it with the existing methods Kernelized and Group Sparse graph (KGS-graph) [1] construction method and Clustering Algorithms for Probabilistic Graphs (CA-PG) [2]. The proposed technique uses E1 Nino dataset from UCI repository to conduct experiments.

This E1 Nino dataset from Tropical Atmosphere Ocean (TAO) array was developed by the international Tropical Ocean Global Atmosphere (TOGA) program. The TAO array comprises of 70 moored buoys spanning the equatorial Pacific that not only measures the oceanographic but also the surface meteorological variables for improved

detection, understanding and prediction of seasonal-to-inter annual climate variations originating in the tropics, most notably those related to the El Nino/Southern Oscillation (ENSO) cycles.

The moorings were developed by National Oceanic and Atmospheric Administration's (NOAA) Pacific Marine Environmental Laboratory (PMEL). Each mooring measures air temperature, relative humidity, surface winds, sea surface temperatures and subsurface temperatures down to a depth of 500 meters and a few a of the buoys measure currents, rainfall and solar radiation.

5. DISCUSSION

We evaluate the performance of FRSDC technique based on the parameters data objects, clustering time, clustering accuracy and space complexity and compare the results with both (KGS-graph) [1] and (CA-PG) [2]. We use dense E1 Nino Dataset for this performance evaluation. Experimental evaluation is done on FRSDC technique with E1 Nino dataset extracted from UCI repository.

5.1 Scenario 1: Clustering time

With the first set of experiments, we evaluate the impact of Clustering time with different number of data objects or attributes with different range of sizes. Clustering time analysis for FRSDC technique is done on the basis of the number of data objects that are considered for clustering and how much time is taken by this whole process. Clustering time is the time taken to cluster with respect to number of data objects taken for experimentation.

$$CT = No.of\ data\ objects * Time\ for\ clustering \quad (3)$$

From (3), clustering time 'CT' is measured in terms of milliseconds. Table 2 tabulates the results for the clustering time and number of data objects accessed respectively.

Table 2 Tabulation for clustering time

Data objects (attributes)	Clustering time (ms)		
	FRSDC	KGS-graph	CA-PG
2	3.9	5.2	5.8
4	6.8	8.9	9.4
6	11.3	13.5	14.3
8	14.8	17.9	18.9
10	12.2	15.2	16.7
12	18.3	21.4	24.9

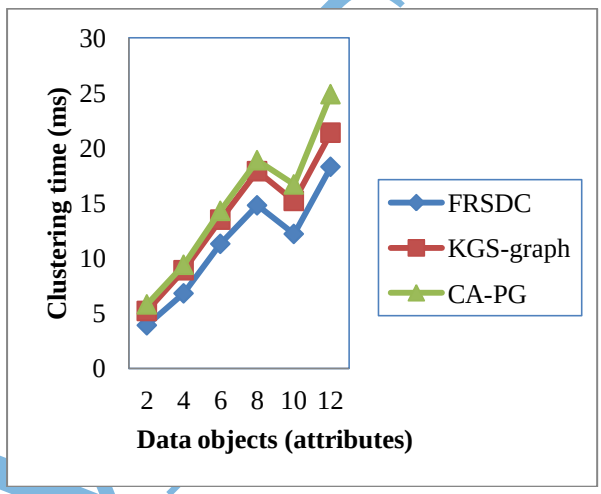

Figure 4 Comparison of clustering time

As shown in the figure 4, for 2 data objects, the time taken to cluster using FRSDC technique is 39ms. On the other hand with the same number of data objects, the clustering time was observed to be 5.2ms when applied with KGS-graph and 5.8ms when applied with CA-PG. The corresponding recordings are shown in the figure given above.

5.2.Scenario 2: Cluster accuracy

With the second set of experiments we evaluate the impact of cluster accuracy to the data objects considered in terms of iterations being performed. Clustering algorithms effectiveness is measured in terms of cluster accuracy.

$$A = \frac{No.of\ data\ objects\ correctly\ clustered}{No.of\ iterations} * 100 \quad (5)$$

From (5), the cluster accuracy 'A' is measured with respect to the number of data

objects or data points '*P*' considered as input. In the experimental settings, when the number of iterations was considered as 5, a total number of 18 data objects were used. By applying FRSDC, 15 data objects were correctly clustered, 14 data objects correctly clustering using KGS-graph and finally 13 data objects correctly clustered using CA-PG. For each technique, we report the cluster accuracy in processing each round based on the data objects being clustered. The result is shown in table 3.

Table 3 Tabulation for cluster accuracy

No. Of iterations	Cluster accuracy (%)		
	FRSDC	KGS-graph	CA-PG
5	85.32	79.28	74.38
10	89.18	83.16	78.06
15	92.31	86.29	81.19
20	88.14	82.12	76.02
25	91.28	86.26	81.16
30	93.21	85.19	80.09
35	95.86	89.84	83.74

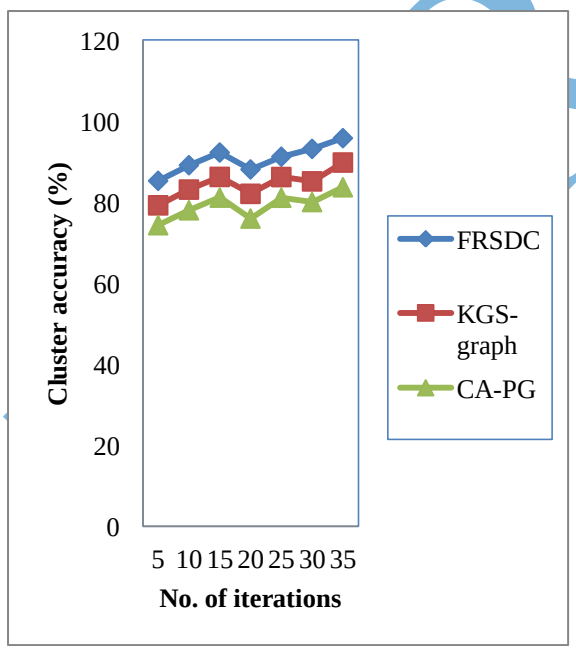

Figure 5 Comparison of cluster accuracy

For almost all the cases as shown in figure 5, the FRSDC technique outperforms both KGS-graph and CA-PG. While for all the techniques, the cluster accuracy increases with the increase in the number of iterations applied, comparatively the growth rate of FRSDC is observed to be very high. At the same time, the cluster accuracy is not linear because the data objects considered during each iteration, vary with each data object or attribute (i.e. zonal winds, meridional winds) has different number of possibilities. But comparatively was observed to be higher using FRSDC technique because of the application of grid-based clustered sparse data representation that efficiently partitions the data objects into a grid. At the same time, different groups (i.e. date, latitude, longitude) for location identification and air temperature, sea surface temperature and subsurface temperatures (for temperature representation) were mapped into different clusters (i.e. weather-location, weather-temperature) by partitioning it into a grid. Therefore, cluster accuracy when applied with FRSDC was improved by 7% compared to KGS-graph and 13% compared to CA-PG.

5.3.Scenario 3: Space complexity

In the third set of experiments, we study the impact of space complexity for clustering with respect to the data objects using FRSDC, KGS-graph and CA-PG. In order to measure the efficiency of clustering, the space complexity has to be measured. Space complexity refers to the complexity involved during clustering or in other words the computational space required during the clustering of data objects.

$$SC = Memory\ Space \left(\frac{clustering\ objects\ to\ be\ placed\ on\ each\ cluster}{No.of\ data\ objects}\right) * \quad (6)$$

From (6), the space complexity for clustering is measured in terms of Mega Bytes (MB). Figure 6 shows the results. We compared the effects of varying the number of data objects across all parameter variations using experiments conducted and stored in Table 4.

Table 4 Tabulation for space complexity

Data objects (attributes)	Space complexity (MB)		
	FRSDC	KGS-graph	CA-PG
2	13	19	23
4	24	30	34
6	39	45	49
8	45	51	54
10	62	68	72
12	73	79	83
14	88	94	97

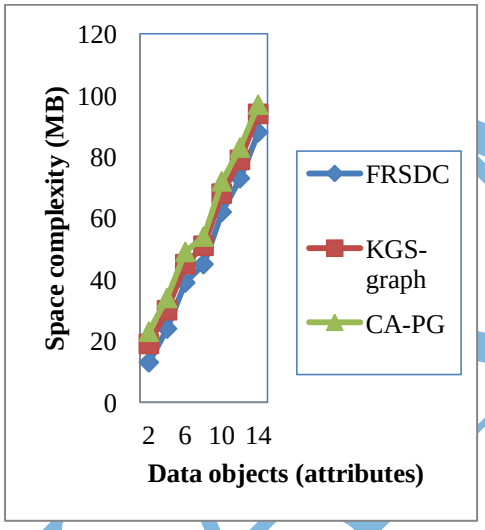

Figure 6 Comparison for space complexity

Figure 6 shows the comparison for space complexity with respect to the varying data objects (observation, year, month, date, latitude, longitude, zonal winds, meridional winds, relative humidity, air temperature, sea surface temperature, sub surface temperature). The significant observation from the figure given above is that the memory space for clustering is directly proportional to the number of data objects used. Therefore though major deviations are not being observed, but comparatively the FRSDC technique proved to be better, since FRSDC uses the geometric median for sparsely distributed high dimensional data. As a result, the FRSDC technique takes into account the geometric median of similar sparse data and non selected sparse data objects. This adversely has an effect in memory space for clustering where fuzzy relational measure is employed to measure the relative distance values. With relative distance measure values, the space complexity for clustering using FRSDC technique is reduced by 18% compared to KGS-graph and 29% compared to CA-PG.

6.Conclusion

In this paper we introduced the problem of identifying and evaluating the clustering objects to be placed on each cluster for sparsely distributed high dimensional data. A new Geometric Median-based Fuzzy model by seamlessly similar and non similar data objects based on Geometric Median is designed. Also an efficient Scattered Distance measure the distance between and within cluster to improve the efficiency of clustering time is proposed. In particular, the Fuzzy Relational Scattered Distance based Clustering (FRSDC) technique is introduced and showed how it is used to specify cluster data points within the limited processing time using grid-based clustered sparse data representation. We experimentally studied our technique, which proved its efficiency in terms of clustering accuracy, clustering time and space complexity for clustering.

REFERENCES

[1] Yuqiang Fang, Ruili Wang, Bin Dai, and Xindong Wu, "Graph-Based Learning via Auto-Grouped Sparse Regularization and Kernelized Extension", IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 1, January 2015, Pages 142-154.

[2] Yu Gu, Chunpeng Gao, Gao Cong, and Ge Yu, "Effective and Efficient Clustering Methods for Correlated Probabilistic Graphs", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 5, May 2014, Pages 1117-1130.

[3] T. Velmurugan, "Performance based analysis between k-Means and Fuzzy C-Means clustering algorithms for connection oriented telecommunication data", Elsevier, Applied Soft Computing, Volume 19, June 2014, Pages 134–146.

[4] Sharadh Ramaswamy and Kenneth Rose, "Adaptive Cluster Distance Bounding for

High-Dimensional Indexing", IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 6, June 2011, Pages 815-830.
[5] Ehsan Elhamifar, and Ren´e Vidal, "Sparse Subspace Clustering:Algorithm, Theory, and Applications", IEEE Transactions on Pattern Analysis & Machine Intelligence, Volume 35, Issue 11, Nov. 2013, Pages 2765-2781.
[6] Satu Elisa Schaeffer, "Graph clustering", Elsevier, Computer Science Review, Volume 1 Issue 1, August, 2007, Pages 27-64.
[7] Yang Wang, Xiaodi Huang, Lin Wu, "Clustering via geometric median shift over Riemannian manifolds", Elsevier, Information Sciences, Volume 220, 20 January 2013, Pages 292–305.

[8] Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang and Ming-Syan Chen, "Density Conscious Subspace Clustering for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, Volume 22, Issue 1, January 2010, Pages 16-30.
[9] Duc Thang Nguyen, Lihui Chen and Chee Keong Chan, "Clustering with Multiviewpoint-Based Similarity Measure", IEEE Transactions on Knowledge and Data Engineering, Volume 24, Issue 6, June 2012, Pages 988-1001.
[10] Malik Tahir Hassan, Asim Karim , Jeong-Bae Kim , Moongu Jeon, "CDIM: Document Clustering by Discrimination Information Maximization", Elsevier, Information Sciences, Volume 316, 20 September 2015, Pages 87–106.
[11] Thenmozhi Srinivasan and Balasubramanie Palanisamy, "Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence", Hindawi Publishing Corporation, The Scientific World Journal, Volume 2015, April 2015, Pages 1-6.
[12] Ramzi A. Haraty, Mohamad Dimishkieh, andMehediMasud, "An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, Volume 2015, Pages 1-2.
[13] Dongjin Lee and Junho Shim, "Impact Parameter Analysis of Subspace Clustering", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, Volume 2015, Pages 1-6.
[14] Zhiping Lin, Jiuwen Cao, Tao Chen, Yi Jin, Zhan-Li Sun, and Amaury Lendasse, "Extreme Learning Machine on High Dimensional and Large Data Applications", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2015, Pages 1-3.
[15] Jessie J. Hsu1, Dianne M. Finkelstein1, David A. Schoenfeld, "Outcome-Driven Cluster Analysis with Application to Microarray Data", Plos One, Volume 10, Issue 11, November 2015, Pages 1-11.
[16] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah, "A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data", Plos One, Volume 10, Issue 12, December 2015, Pages 1-20.
[17] Aloysius George," Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm", The International Arab Journal of Information Technology, Vol. 10, No. 5, September 2013.
[18] Asim Roy, "A classification algorithm for high-dimensional data", Elsevier, Procedia Computer Science, Volume 53, 2015, Pages 345–355.
[19] Jochen Kerdels , GabrielePeters, "Analysis of high-dimensional data using local input space histograms", Elsevier, Neuro computing, Volume 169, 2 December 2015, Pages 272–280.
[20] Rashmi Paithankar and Bharat Tidke, "A H-K Clustering Algorithm For High Dimensional Data Using Ensemble Learning", International Journal of Information Technology Convergence and Services (IJITCS), Volume 4, Issue 5/6, December 2014, Pages 1-9.