

Fuzzy Relational Scattered Distance Based Clustering Method for Sparsely Distributed High Dimensional Data Objects

R.Pushpalatha, K. Meenakshi Sundaram



Abstract : Clustering is one of the most significant ideas in data mining. It is an unsupervised learning model. Clustering technique in handling high dimensional data is more complex due to intrinsic sparsity nature of high dimensional data. Though, existing methods to reduce immaterial clusters were based on spectral clustering algorithm and graph-based learning algorithm, whose lack of sparsity and polynomial time complexity compromises their efficiency when applied to sparse high dimensional data. This paper concentrates to cluster the sparsely distributed high dimensional data objects. Fuzzy Relational Scattered Distance Based Clustering (FRSDBC) method is developed with three models such as Geometric Median Based Fuzzy model, Scattered Distance measure model, Grid based clustered sparse data representation model. Geometric Median Based Fuzzy model calculates the geometric median of similar sparse data and then the non similar sparse data objects to fitting the relational fuzziness across data points. It involves in the subspace reduction of data objects. Scattered Distance measure model is used to measure the distance between the inner and outer object. Grid based clustering is used to calculate the area of the cluster in FRSDBC method. The main idea of the FRSDBC method is to clustering data points over sparsely distributed data within limited processing time. The Clustering Time, Clustering Accuracy and Space Complexity of each method is analyzed. The result of the FRSDBC method is compared with other techniques, the results obtained are more accurate, easy to understand and the clustering time was substantially low in FRSDBC method. It is widely used in many practical applications such as weather forecast, share trading, medical data analysis and aerial data analysis.

Keywords: Data mining, Geometric median based Fuzzy Concept, Scattered Distance measurement, Graph-Based Learning.

I. INTRODUCTION

Data mining is used to extract the useful and non-trivial information from large amount of database. The logical group of similar data is known as Clustering. It is widely applied in many practical applications such as weather forecast, share trading, medical data analysis, aerial data analysis, etc., Clustering in data mining, which evaluate inherent combination of data analysis and text documents.

Manuscript published on January 30, 2020.

* Correspondence Author

Dr.R.Pushpalatha*, Assistant Professor, Department of Computer Science, Kongu Arts and Science College (Autonomous), Nanjanapuram, Erode, Tamil Nadu, India. E-mail: rpljour@gmail.com

Dr.K.Meenakshi Sundaram, Associate Professor, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, Tamil Nadu, India. E-mail: lecturerkms@yahoo.com

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

With the growing interest in the information technology, the size of data is also getting increased including financial institutions, electricity board and educational department and so on. It 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. The rest of the paper is structured as follows. In Section 2, some of the Related Works and their approaches are explained. Section 3 describes the basics of Fuzzy Relational Scattered Distance Based Clustering for sparse high dimensional data objects with the algorithm. In Section 4 talks about the experimental setup of the proposed technique. Section 5 contains the experimental studies, results and a brief discussion. Finally, Section 6 shows the conclusions of the research work.

II. RELATED WORKS

Kernelized Group Sparse graph (KGS-graph) was introduced in [1] to state the contextual information of a data manifold. KGS-graph successively protects the properties of sparsity and locality concurrently. However, sparse graph construction does not merge nonzero coefficients locality and sparsity. Clustering Algorithms for Probabilistic Graphs (CA-PG) [2] addressed the trouble of clustering correlated probabilistic graph. It improves the clustering efficiency. However some recent studies [3], [4] pointed out that the multi-dimensional and inter dimensional space can also be correlated to decrease the computational complexity. Sparse Subspace Clustering [5] handle data points using sparse optimization program to gather the clustering of data into subspaces aiming at reducing the outliers.

A survey of graph clustering [6] was investigated to measure the cluster quality for a exact seed vertex by local calculation. A novel density function [7] was applied to characterize the points with geodesic distance to improve clustering efficiency. Another method called density conscious subspace clustering [8] was designed to reduce the subspace during clustering using divide-and-conquer technique. 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 intended 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 improved K-Means clustering algorithm was presented to improve cluster quality by applying subspace clustering. A mechanism for analyzing the impact parameter for subspace clustering was presented in [13]. The Extreme Learning Machine (ELM) on high dimensional data was examined in [14]. Cluster analysis are used in different areas to allocate characteristics of an observation into clusters as a result those data object in the same group are more similar,when compared to the presence of data object in other groups.

In [15], Markov Chain Monte Carlo algorithm describes a strategy for prediction based on same data. In [16], the comparison study of similarity and dissimilarity measures in clustering continuous data was studied.

III. FUZZY RELATIONAL SCATTERED DISTANCE BASED CLUSTERING METHOD

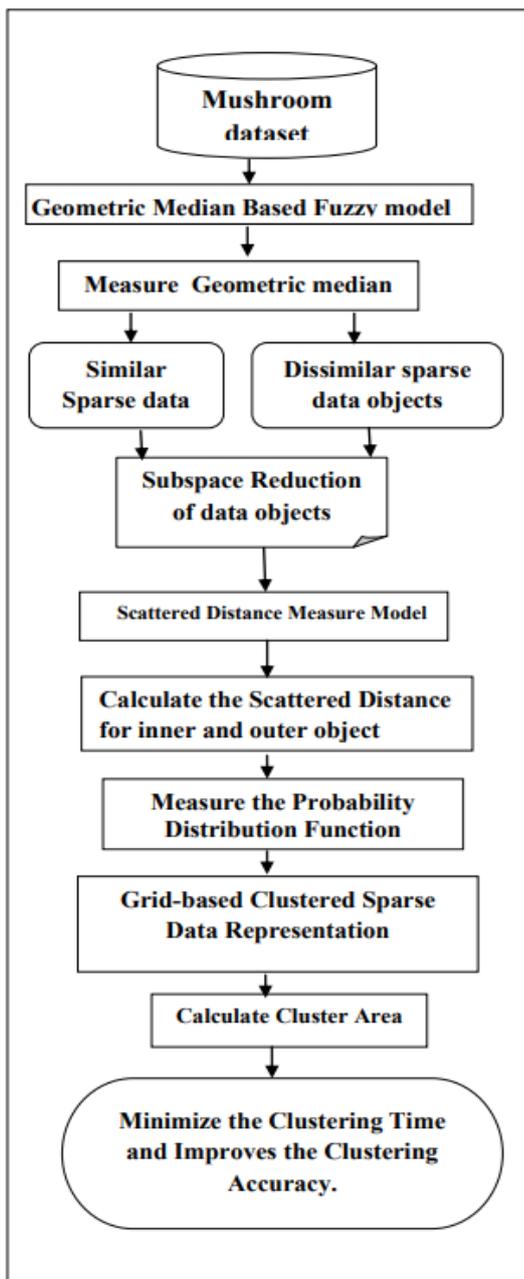


Fig. 1 Architecture of Fuzzy Relational Scattered Distance Based Clustering Technique

FRSDBC method, evaluate the geometric median of sparsely distributed high dimensional data and to determine clustering objects to be placed on each cluster. Initially, FRSDBC identify the geometric median of similar Fuzzy Relational Scattered Distance Based Clustering (FRSDBC) method is introduced for sparsely distributed high dimensional data to cluster the data points within the limited processing time.

sparse data objects and then the dissimilar sparse data objects to appropriate the relational fuzziness across data points, reducing subspace of data objects in clustered plane.

Next, Scattered Distance measures the distance of geometric median of inner object (similar object) and outer object (Dissimilar 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 FRSDBC and therefore obtains the clustered sparse data. The Clustering Accuracy and Clustering Time of each algorithm are examined and the outcomes are compared with one another.

Compare the result of this technique, it was found that the results gained are more accurate, easy to recognize and the time taken to cluster the data was significantly lesser in FRSDBC method than the state-of-art methods. Figure 1 demonstrates the architecture of Fuzzy Relational Scattered Distance Based Clustering method.

In the fig. 1, Mushroom dataset is considered as an input of FRSDBC method. Calculate the geometric median for the sparsely distributed high dimensional data objects. To identify the similar sparse data and non similar sparse data objects with the help of the geometric median.

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.

A. Problem Statement

Consider the 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 ‘ $r(p_i, p_j)$ ’ symbolizes the dissimilarity between the data points ‘ p_i ’ and ‘ p_j ’.

Therefore the ‘ n th’ component of ‘ p_i ’ is obtained from the domain ‘ d_n ’. Let us assume that given a set of nonnegative dissimilarities ‘ d_{ij} , where $i, j = 1, 2, \dots, n$ ’ between every couple of data points ‘ i ’ and ‘ j ’ correspondingly. The dissimilarity ‘ d_{ij} ’, indicates how the data point ‘ i ’ is suited to be the data point ‘ j ’. Let ‘ $Y = y_1, y_2, \dots, y_n, y_i \in P$ ’, where ‘ Y ’ is a subset of ‘ P ’.

B. Geometric Median Based Fuzzy Model

The main objective of FRSDBC is to identify the clustering data points over the sparsely distributed data within the limited processing time. Initially, Geometric Median Based Fuzzy model recognize the geometric median of similar sparse data and then the non similar 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 to reduce the subspace of data objects in the clustered plane.

In this section the Geometric Median Based Fuzzy model is designed in which the fuzzy relational classifies a data point, with the cluster centroid to be nearest 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.

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

Where ' $[d_{ij}]$ ' is the fuzzy partition matrix of size ' $n * m$ ' and ' $d_{ij} \in [0, 1]$ ' symbolizes the membership coefficient of ' i th' data point in the ' j th' cluster and ' r ' representing the fuzzification parameter. With the minimization function obtained, the Geometric Median Based Fuzzy model calculates the geometric median of sparsely distributed high dimensional data object to determine the clustering objects to be placed on each cluster, and is used for improving the undesirable effect of outliers.

Now consider a discrete set of ' d ' dimensional data points ' $P = p_1, p_2, \dots, p_n$ ' in data space ' S ' (i.e. sample points) is the point minimizing the sum of distances to the sample points. For a given set of ' n ' data points, ' p_1, p_2, \dots, p_n ' with each ' $p_i \in D$ ', then the geometric median of the sparsely distributed high dimensional data object to determine the clustering objects to be placed on each cluster is as given below.

$$GM = argmin \sum_{i=1}^n ||p_i - q||_2, q \in D \quad (2)$$

Refer (2), ' $argmin$ ' specifies the value of the argument ' q ' which minimizes the sum. In this case, it is the point ' q ' from where the sum of all Euclidean distances to the ' p_i ' is minimum. From equation(2), the geometric median of similar sparse data object is obtained. The similar sparse data object is obtained, then the non selected sparse data objects is evaluated as given below.

$$\sum_{i=1}^n \frac{p_i - 1}{||p_i - q||} \quad (3)$$

$$where \ q = \frac{(\sum_{i=1}^n \frac{p_i}{||p_i - q||})}{(\sum_{i=1}^n \frac{1}{||p_i - q||})}, p_i \neq q \quad (4)$$

From (4), ' q ' symbolizes the non selected sparse data objects. Finally the geometric median of similar sparse data and dissimilar sparse data objects helps in reducing the subspace of the data objects in the clustered plane. Fig. 2 shows the algorithmic description of Geometric Median Based Fuzzy model.

For the selected Mushroom 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. Next, the similar sparse data object is obtained through geometric median by minimizing the sum function. Finally, the non selected sparse objects are measured to obtain the relational fuzzy.

Input: d Dimensional Data Points ' $P = p_1, p_2, \dots, p_n$ ' in Data Space ' S ' and ordered domains ' $D = \{d_1, d_2, \dots, d_n\}$,
Output: Identify the selected and non selected sparse data objects. Reducing the subspace in clustering plane and reduce the space complexity.
1: Begin
2: For each ' n ' data points
3: Measure the membership function using minimization function (1)
4: Calculate the geometric median of similar sparse data objects using (2)
5: Evaluate the non selected sparse data objects using (4)
6: End for
7: End

Fig. 2 Algorithm for Geometric Median Based Fuzzy Model

C. Scattered Distance Measure Model

Once, the similar selected and non selected sparse objects are differentiated by applying Geometric Median Based Fuzzy Model, a Scattered Distance measure model is used to measure the distance between the inner and outer object.

The scatter distance criteria are obtained from the scatter matrices, that shows the inner object (within cluster scatter) and the outer object (between cluster scatter). Scattered distance measure evaluates the inner object for two data points ' p_i ' and ' p_j ' where the data points are represented in two different clusters ' $Cl(i = K)$ ' and ' $Cl(j = K)$ ' is as given below.

$$T = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m Dis_{ij} \quad (5)$$

$$T = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m Dis(p_i, p_j) \quad (6)$$

$$T = \frac{1}{2} \sum_{k=1}^K \sum_{Cl(i=K)} (\sum_{Cl(j=K)} Dis(p_i, p_j) + Clj \neq K Dis p_i, p_j) \quad (7)$$

$$T = W(CS) + B(CS) \quad (8)$$

Refer (8), ' $W(CS)$ ' symbolizes the within cluster scatter whereas ' $B(CS)$ ' symbolizes the between cluster scatter. The geometric median of inner object or within the cluster scatter and outer object or between the cluster scatter of similar data object position is mathematically calculated as follows.

$$W(CS) = \sum_{Cl(i=K)} (p_i - r_k)^2 \quad (9)$$

$$B(CS) = \sum_{Cl(i=K)} (r_k - m)^2 \quad (10)$$

Consider two uncertain objects ‘*p*’ and ‘*r*’ and their corresponding probability distribution estimates ‘ $D(p || r)$ ’. Figure 3 shows the probability distributions of two uncertain objects ‘*p*’ (red colour) and ‘*r*’ (violet colour).

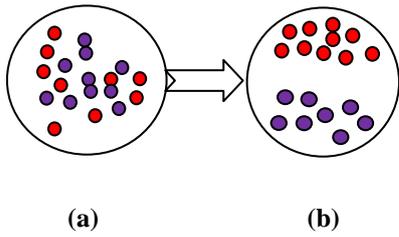


Fig. 3 Probability distributions and cluster

The distribution difference between inner object and outer object cannot be extracted by geometric median. The proposed work uses probability distributions. In fig. 3 (a), the locations of the two object (denoted by violet and red color) overlapping different distributions are presented, whereas in fig. 3 (b) the two objects have different locations. FRSDBC method computes the probability distribution function which is significant characteristics of uncertain objects while performing the clustering step. The probability density functions between ‘ p_i ’ and ‘ r_k ’ is measured as given below,

$$D(p || r) = \sum_{i,k \in P} f(x) \log \left(\frac{p_i}{r_k} \right) \quad (11)$$

Refer (11), every point in the data space ‘*S*’ indicates a probability density function and in this way different groups are mapped into different clusters. This in turn reduces the clustering time.

D. Grid-based Clustered Sparse Data Representation

Different groups mapped into different clusters are partitioning it into a grid. Scattered Distance with grid form is used to calculate the area of the cluster in FRSDBC method. Every dimension is partitioned into ‘2’ portions. So, ‘*n*’ data is separated into ‘2*n*’ cells and each cell possessing equal size. The central points of cells in the grid are considered as values in the ordered domain. The probability of an object in the cell in a grid is the sum of the probabilities of all its sample points of this cell. Given the clusters ‘*n*’ for a dataset, ‘*n*’ probability density function is produced with single uniform distribution ‘ $(n - 1) / 2$ ’ distributions with unlike distance measures. For each distribution, the proposed method generated a group of samples that identifies clustering data points, every one forms one clustering object. Therefore, clustering objects in the similar group are sampled from the identical probability density function. In this manner, the Grid form in FRSDBC decides the clustered sparse data and therefore considerably improves the clustering accuracy.

IV. EXPERIMENTAL SETTINGS

In this section, assess the efficiency of the proposed FRSDBC method and compare it with the two existing methods of Kernelized Group Sparse graph (KGS-graph) [1] construction method and Clustering Algorithms for Probabilistic Graphs (CA-PG) [2]. The proposed method uses mushroom dataset to conduct experiments.

The mushroom dataset contains 22 attributes and 8124 records. Every record demonstrates the physical qualities of

a single mushroom. A categorization label of poisonous (“p”) or not poisonous or edible (“e”) is supplied with each and every record. The numbers of edible mushrooms in the dataset are 4208 and poisonous mushrooms in the dataset are 3916.

The mushroom data is nominal. Instead of telling each mushroom with a string the proposed method explains it with a binary vector (0 to 1) where ‘1’ indicates the attribute value for each mushroom and ‘0’ stands for the non-attribute values. The attributes have definite length and only certain possibilities are existing for describing the attribute, i.e. for attribute one, Cap-Shape, there are 6 options, Bell (B), Conical (C), Convex (X), Flat (F), Knobbed (K), Sunken (S). The feature flat (f) can be considered as the binary string 000100. The flat feature has number four in the attribute string row. The binary vector explaining the mushrooms is of the length 126; it includes all possible features for the mushrooms.

Table 1 contains the ten attributes out of twenty two attributes that shows the mushroom explanation in the dataset.

Table- I: List of Ten Attributes of Mushroom Dataset

Attrib ute no.	Attribute name	Possibilities
1	Cap-Shape	Bell (B), Conical (C), Convex (X), Flat (F), Knobbed (K), Sunken (S)
2	Cap-Surface	Fibrous (F), Grooves (G), Scaly (Y), Smooth (S)
3	Cap-Color	Brown (N), Buff (B), Cinnamon (C), Gray (G), Green (R), Pink (P), Purple (U), Red(E), White (W), Yellow (Y)
4	Bruises	Bruises (T), No (F)
5	Odor	Almond (A), Anise (I), Creosote (C), Fishy (Y), Foul (F), Musty (M), None (N), Pungent (P), Spicy (S)
6	Gill-Attachment	Attached (A), Descending (D), Free (F), Notched (N)
7	Gill-Shaping	Close (C), Crowded (W), Distant (D)
8	Gill-Size	Broad (B), Narrow (N)
9	Gill-Color	Black (k), Brown (N), Buff (B), Chocolate (H), Gay (G), Green(R), Orange (O), Pink (P), Purple (U), Red (E), White (W), Yellow (Y)
10	Stalk-Shape	Enlarging (E), Tapering (T)

V. RESULTS AND DISCUSSIONS

Evaluate the performance of FRSDBC method based on the parameters such as clustering time and clustering accuracy, compare the results with both (KGS-graph) [1] and (CA-PG) [2]. For performance evaluation the Dense Mushroom Dataset is used as an input. Mushroom dataset is extracted from UCI repository. Experimental evaluation is done on FRSDBC method.

A. Impact of Clustering Time

Clustering time is defined as time taken to cluster the data objects with respect to number of data objects taken for experimentation.



$$CT = \text{No. of data objects} * \text{Time for clustering} \quad (12)$$

From equation (12), clustering time ‘CT’ is measured in terms of milliseconds. With the calculated time, the effectiveness of the technique is evaluated. If the clustering time is lesser, then the method is said to be more efficient. Table-II tabulates the results for the clustering time and number of data objects accessed respectively.

Table-II: Clustering Time

No. of Data objects	Clustering time (ms)		
	Proposed Method	Existing Methods	
		FRSDBC	KGS-graph
2	3.5	4.1	4.7
4	6.8	7.4	8.1
6	9.1	9.7	10.2
8	5.3	5.9	6.4
10	14.8	15.4	16.3
12	17.5	18.1	18.9
14	20.3	21.4	22.7

Since similar and non similar sparse data objects are extracted separately and then the distance between them is measured using Scattered Distance model in the proposed FRSDBC method. This makes the clustering time reduced when compared to the two other methods, where clustering is made on the entire data objects. The second observation is that while all the three techniques perform similar for two to eight data objects, FRSDBC method outperforms the other two techniques for larger values of ‘P’. Therefore an improvement was observed using KGS-graph by 8.11% and 16.60% over the two existing techniques.

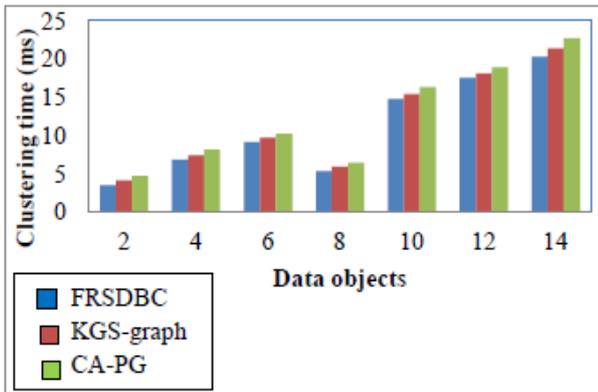


Fig. 4: Comparison of Clustering Time

B. Impact of Clustering Accuracy

Evaluate the impact of clustering accuracy of the data objects. Clustering algorithms effectiveness is calculated in terms of cluster accuracy. A clustering accuracy is defined as the number of data objects are correctly clustered to the total number of iterations. It is mathematically expressed as

$$CA = \frac{\text{No. of data objects correctly clustered}}{\text{No. of Iterations}} * 100 \quad (13)$$

From (13), the cluster accuracy ‘CA’ is measured with respect to the number of data objects or data points ‘P’

considered as input. If the clustering accuracy is higher, then the method is said to be more efficient. It is measured in terms of percentage (%).

Table-III: Cluster Accuracy

No. of Iterations	Cluster accuracy (%)		
	Proposed Method	Existing Methods	
		FRSDBC	KGS-graph
5	89.14	85.14	75.32
10	91.35	84.30	74.25
15	94.18	89.13	79.08
20	90.32	85.27	75.22
25	92.16	87.11	77.06
30	94.25	89.20	79.15
35	95.23	90.18	80.13

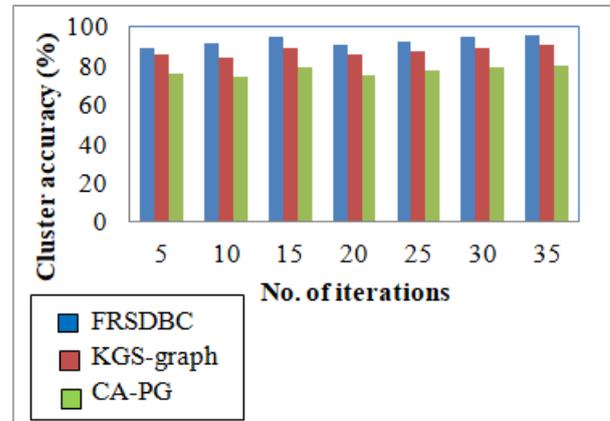


Fig. 5: Comparison of Cluster Accuracy

In the experimental settings, when the number of iterations was considered as 5, a total number of 18 data objects were used. By applying FRSDBC, 16 data objects were correctly clustered, 15 data objects correctly clustering using KGS-graph and finally 13 data objects correctly clustered using CA-PG. The result is shown in Table-III. FRSDBC method is compared with both KGS-graph and CA-PG. Comparatively the growth rate of FRSDBC observed to be very high 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. B, C, X, F, K, S for Cap-Shape representation and F, G, Y, S for Cap-Surface representation) were mapped into different clusters (i.e. Cap-Shape, Cap-Surface) by partitioning it into a grid. Therefore, the clustering accuracy of FRSDBC was improved by 5.61% when compared to KGS-graph and 16.46% when compared with CA-PG.

VI. CONCLUSION

This paper deals with the problem of identifying and evaluating the clustering objects to be placed on each cluster for sparsely distributed high dimensional data objects. In the newly designed FRSDBC method, the first model Geometric Median-based Fuzzy model identifies the similar and non similar data objects based on Geometric Median.

It reduces the sub spacing of data objects. The proposed Scattered Distance Measure Model measures the distance between and within cluster scatter to improve the efficiency of clustering time. The probability distribution function is used to map the different groups into different clusters. Different groups mapped into different clusters are partitioning it into a grid. Scattered Distance with grid form is used to calculate the area of the cluster. In particular, the introduced Fuzzy Relational Scattered Distance Based Clustering method is used to cluster the sparsely distributed data points within the limited processing time using grid-based clustered sparse data representation. The proposed method experimentally proved its efficiency is better in terms of clustering accuracy and clustering time.

REFERENCES

1. (Vancouver reference style), Ali Seyed Shirkorshidi, Saeed Aghabozorgi, Teh Ying Wah, (2015, December), "A comparison study on similarity and dissimilarity measures in clustering continuous data", Plos One, Volume 10 (Issue 12), pp. 1-20.
2. Aloysius George, (2013, September), "Efficient high dimension data clustering using constraint-partitioning k-means algorithm", The International Arab Journal of Information Technology, Volume 10 (No.5).
3. (Citation Style), Asim Roy, (2015), "A classification algorithm for high-dimensional data", Elsevier, Procedia Computer Science, Volume 53, pp. 345-355.
4. Dongjin Lee and Junho Shim, (2015), "Impact parameter analysis of subspace clustering", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, Volume 2015, pp. 1-6.
5. Duc Thang Nguyen, Lihui Chen and Chee Keong Chan, (2012, June), "Clustering with multiviewpoint-based similarity measure", IEEE Transactions on Knowledge and Data Engineering, Volume 24 (Issue 6), pp. 988-1001.
6. (Citation Style), Ehsan Elhamifar, and René Vidal, (2013, November), "Sparse subspace clustering: algorithm, theory, and applications", IEEE Transactions on Pattern Analysis & Machine Intelligence, Volume 35 (Issue 11), pp. 2765-2781.
7. (Vancouver reference style), Jessie J. Hsu¹, Dianne M. Finkelstein¹, David A. Schoenfeld, (2015, November), "Outcome-driven cluster analysis with application to microarray data", Plos One, Volume 10 (Issue 11), pp. 1-11.
8. (Citation Style), Jochen Kerdels, Gabriele Peters, (2015, December, 2), "Analysis of high-dimensional data using local input space histograms", Elsevier, Neuro computing, Volume 169, pp. 272-280.
9. (Citation Style), Malik Tahir Hassan, Asim Karim, Jeong-Bae Kim, Moongu Jeon, (2015, September, 20), "CDIM: Document clustering by discrimination information maximization", Elsevier, Information Sciences, Volume 316, pp. 87-106.
10. Ramzi A. Haraty, Mohamad Dimishkieh, and Mehedi Masud, (2015), "An enhanced k-means clustering algorithm for pattern discovery in healthcare data", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, Volume 2015, pp. 1-12.
11. Rashmi Paithankar and Bharat Tidke, (2014, December), "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), pp. 1-9.
12. Satu Elisa Schaeffer, (2007, August), "Graph clustering", Elsevier, Computer Science Review, Volume 1 (Issue 1), pp. 27-64.
13. (Reference Style), Sharadh Ramaswamy and Kenneth Rose, (2011, June), "Adaptive cluster distance bounding for high-dimensional indexing", IEEE Transactions on Knowledge and Data Engineering, Volume 23 (Issue 6), pp. 815-830.
14. Thenmozhi Srinivasan and Balasubramanie Palanisamy, (2015, April), "Scalable clustering of high-dimensional data technique using spcm with ant colony optimization intelligence", Hindawi Publishing Corporation, The Scientific World Journal, Volume 2015, pp. 1-6.
15. (Citation Style), T. Velmurugan, (2014, June), "Performance based analysis between k-means and fuzzy c-means clustering algorithms for connection oriented telecommunication data", Elsevier, Applied Soft Computing, Volume 19, pp. 134-146.
16. (Citation Style), Yang Wang, Xiaodi Huang, Lin Wu, (2013, January, 20), "Clustering via geometric median shift over Riemannian manifolds", Elsevier, Information Sciences, Volume 220, pp. 292-305.

AUTHORS PROFILE



Dr. R. PUSHPALATHA, works as an Assistant Professor in Department of Computer Science at Kongu Arts and Science College (Autonomous), Nanjanapuram, Erode, Tamil Nadu, India. She has 16 years of teaching experience. She received her B.Sc degree from Bharathiar University, Coimbatore, M.C.A degree from Bharathidasan University, Trichy, M.Phil and Ph.D. degrees from Bharathiar University, Coimbatore. She has published a book titled "Mathematical Foundations of Computer Science" under Partridge India publications and published more papers in international journals indexed by Scopus. She has presented more papers in national and international conferences. Her research experience is in the field of Data Mining and Networking. She has received Best Teacher Award from the college Management for three consecutive years. She has received Research Award from the Secretary and Correspondent of Erode Arts and Science College (Autonomous), Erode.



Dr. K. MEENAKSHI SUNDARAM, works as an Associate Professor in Department of Computer Science at Erode Arts and Science College (Autonomous), Erode, Tamil Nadu, India. He has 31 years of teaching experience. He produced 45 M.Phil Research Scholars and 2 Ph.D. Research Scholars in Computer Science. Also 6 Ph.D Research Scholars on progress. He conducted 7 Ph.D Public Viva-Voce Examination as examiner. He serves as a member in Board of studies of various universities and Autonomous colleges. He published 45 papers in international journals and 60 papers in National and International conferences. His area of interest is Data Mining and Software Engineering.