IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

(EDSFCA): Efficient Document Subspace Clustering in High-Dimensional Data using Fast Clustering Algorithm

Radhika K R[1], Pushpa C N[1], Thriveni J[1], Venugopal K R[2]

[1]Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore, India
[2] Vice-Chancellor, Bangalore University, Bangalore, India

[*]*Corresponding Author: radhika@bmsit.in*

DOI: https://doi.org/10.26438/ijcse/v7i2.10101015 | Available online at: www.ijcseonline.org

Accepted:15/Feb/2019, Published: 28/Feb/2019

Abstract— In the contemporary age of digitization, majority of the users are constantly moving on the prevalent computing in the area of telecommunication and social networking. The data may be produced from several resources from an individual to organization level. The existing data mining techniques are not suitable, due to the features of non structured and semi-structuredness in data which leads to dimensionality problems. To overcome these problems, an Efficient Document Subspace Clustering in High Dimensional Data using Fast Clustering Algorithm (EDSFCA) is proposed. This method performs Datamining techniques like preprocessing and removing of corrupted and repetative data from the subspace clusters. The twitter data is taken as an input and is divided into clusters in order to provide a characteristic of high-dimensional data. This information is organized arbitrarily in subspace clusters and then segmentation is done on data points. The EDSFCA approach does the cluster analysis of datasets in smallest period of time.

Keywords—: Data Mining, Fast Clustering Algorithm, High Dimensional Data, Subspace Clustering.

I. INTRODUCTION

Now a day's the data size is growing rapidly, this will not only lead to the issues in size but also arise problems like processing and analysis of the data [1]. Data Mining techniques give solutions for analyzing such large amount of data. Abstracting data from a dataset and renovating it into an comprehensible structure is called Data Mining. This technique collects raw data at an exploring rate, these data may be of different types like text data, images, audios and videos and store all data in the database or in a data warehouse and the objective is to expose the hidden patterns and trends.
The medicinal science, educational centers, social networks and many further networking application fields are the numerous assets for the growth of data. The evolution in the massive data doesn't come across any subjects in storage, but all kinds of complications starts sprouting up when it comes to the analysis of the data to abstract exact set of knowledge. However, if the dimensionality is very high, it carries a superior deal of problems and this becomes a challenging circumstances for taking out the clustering process.

The following are the common characteristics of data mining:

- The data is generated from various resources like from companies, individual organizations, educational centers and business. For example, in the business and companies, every day a huge amount of data being generated, this leads to increasing the size of the data through various resources,
- The generated data may be incomplete, redundant and noises. As the data are generated from different sources, it may be incomplete or the data generated may be not in an understandable form.
- Every day the size increases, so the velocity of data is also a problem in Data Mining.

The Data Mining technique has advantages and also has some disadvantages they are given below:

There are many organizations taking the benefit of data mining like banking, educational centers, business and markets and individual organization. The disadvantages are like it contains a huge data, the storage and analysis leads to some disadvantages. The security is also one of the major disadvantages in Data Mining.

The Fast Clustering algorithm [2] is best applied for finding similar clusters because it gives better competence in collecting large data sets. Clustering is the collection of identical clusters and is the

foremost operation in data mining. The utmost divergent feature of Data Mining is that it deals with very large dataset ranging from Gigabytes to even Terabytes, so the algorithm used should be accessible.

Motivation: Technology is growing day by day so the data generated from different sources and also it is not in structured form, it contains duplicated data, data may be corrupted, may contain noisy. So this may be difficult in the data analysis phase for various applications. In the present day, storing the data generated from different sources is not a big issue but processing and analyzing of that data is the major problem. In the present day the data is generated from various sources, it is difficult to understand and to do analysis as it contains more noise and redundancies

Contribution: This work helps to overcome the problems faced during clustering and segmentation of data. This helps in many applications like pattern recognition, image processing and machine learning. An Efficient Document Subspace Clustering in High Dimensional Data using Fast Clustering Algorithm (EDSFCA) has been designed to easier the process of analysing and processing.

Organization: This work provides the details about the Data Mining clustering and describes about the present subspace clustering algorithms and reduction of dimensionality techniques in Section II. Section III briefs about the Background study. Section IV explains the System Architecture. Results and performance analysis are discussed in Secton V, Conclusions and Future Work is given in VI.

II. RELATED WORK

Radhika K R et al., [3] explains that the data may be produced from different resources this increases the size of the data, there by problems in storage. Not only in terms of storage, processing and analyzing also makes it difficult as the size increases. The existing Data Mining techniques did not fulfill the required results, so the authors proposed an effective method to overcome these problems by eliminating the redundant and corrupted data from the database in order to do clustering.

M Verleysen et al., [4] focused on the curse of dimensionality and empty space phenomenon while designing the neural networks. The data generated from various sources were very complex, those data points or the objects which have a number of variables, referred as high dimensional data. In neural network it is very difficult to handle such high dimensional data. The authors discussed the practical problem that arises in both the high dimensional and low dimensionality of data.

The Fast Greedy algorithm was discussed by Petukhov et al., [5] it is very efficient in subspace clustering, where the data are mainly related to a low dimensional. The greedy algorithm has some disadvantages like, if there are some missing entries in the data set on a particular location, or if the rate of error is more in the data set, in such conditions the algorithm fails to give better performance. To overcome from these problems authors have proposed Fast Greedy Sparse Subspace Clustering method. In this method the subspace clustering is divided into two parts (i) Pre-processing and (ii) Examine the graphs for clusters founded on the adjustments of sparse subspace clustering calculation using greedy algorithm.

The Bayesian Nonparametric Subspace Cluster Method considers the number and dimension of every subspace from the experimental data. Y Amardeep Kaur et al., [6] concentrated on discovering all the clusters in all possible subspaces and also finding the dimension of the considered subspaces and quantity of clusters concealed in indivisual subspace.

J Wei et al., [7] discussed about the difficulty of clustering partial data drawn from the union of subspaces. To overcome from this problem a Low-Rank Matrix completion method is applied. Authors proposed two approaches (i) Sparse Subspace Clustering to obtain sparse representation of data. (ii) Suitable kernel matrix for missing entries.

Singh Vijendra et al., [8] discussed Density Based Clustering which also affected by the Density Divergence problem and the accuracy of clustering. So clusters constructed using relative region densities and by applying the Density Based Subspace Clustering algorithms. Lance Parson el al., [9] discuss algorithms pertaining to subspace clustering along with their characteristics.
Kumutha et al., [10] have discussed Comparative Empirical Evaluation using data sets taken from UCI ML repository and also discussed about various solutions for high Dimensional Data Clustering. Subspace clustering helps to solve the problems of High dimensional Data Clustering. Sunita Jahirabadkar et al., [11] proposed a proper selection of clustering technique to suit a particular application.by considering features such as intersecting/non-coinciding, axis parallel and so on.

Singh Vijendra el al., [12] proposed a robust Multi Objective Subspace Clustering (MOSCL) algorithm for the problems of high dimensional clustering. It works in two phases: in first phase subspace relevance analysis done and in second phase subspaces are discovered in dense region using MOSCL method. Hans peter Kriegel et al., [13] designed a framework built on an effectual filter refinement design that scales at most quadratic with respect to the dimensionality of any subspace clustering.

Yining wang el al., [14] proposed a simple post processing procedure for graph connectivity problem which gives the exact clustering of sparse subspace clustering (SSC) for subspaces of dimension greater than 3. J wei el al.,[15] have discussed performance comparison of different subspace segmentation algorithms and also explained other views and other conservative

approaches that can be applied in this field. C Giraud Taylor et al., [16] explained Mathematical foundations of High-Dimensional Statistics.

R Agarwal et al., [17] proposed Automatic Subspace Clustering of High Dimensional Data to identify the dense clusters in subspace of maximum dimensionality. This method finds the accurate clusters in large High Dimensional dataset. Nenad Tomašev et al., [18] discussed about hubness and its negativity of the present clustering algorithms by quality of hub points decreasing the distance between-clusters. Here an attempt is made by S wang et al., [19] to cluster high dimensional data using online data stream technique. They have used an online technique to partition the data as dense data space and conserved the superset of these data. Then applied proposed method in order to find subspace in high dimensional data of any random shape.

M C Tsakiris et al., [20] discussed about the problems which occur with the linear subspace problems. For any algebraic geometric methods subspaces should be of equal dimension. They have considered the union of subspaces as an alternative for data points. They proposed an algorithm to make algebraic variety into subspaces in general which consists of different dimensions. E C Ozan et al., [21] conferred Vector quantization method for Approximation Nearest Neighbour search which allows fast and additional perfect retrieval on publicly available datasets. For developing Hyper Spectral Images (HIS) subspace clustering is the vigorous method which is generally used, but this avoids the in depth exploration of spatial data. Han Zhai et al., [22] suggested l_2-norm regularized algorithm which takes the advantage of retrieving spatial spectral data confined in HIS.

Shulin Wang et al., [23] proposed a new clustering algorithm Nonnegative Matrix Factorization (NMF) for getting good clustering quality by building good affinity matrix. Junjian Zhang et al., [24] recommended two steps to solve the problem of high dimensionality by building affinity matrix by making use of self-expressiveness model in the first step and Low Ranked Structured Sparse Subspace Clustering (LRS3C) in order to get the low-rank representation based subspace clustering in the second step.

Fast Greedy Sparse Subspace Clustering (FGSSC) algorithm are suggested by Alexander Petukhov et al., [25] to work with noisy data and which are having high rate of error. Ran He et al., [26] elaborated about how to handle outliers by correntropy sparsity-induced measures in order to get combination of many subspaces in the presence of outliers. Yifan Fu et al., [27] addressed the issue related to high dimensional data if it contains noisy or redundant data and also suggested Tensor Low-Rank Representation (TLRR) to build robust subspace segmentation from corrupted data.

III. BACK GROUND STUDY

The high dimensional data comes from many of the sources like social networking, medical science and education. These data which are generated from different sources are in the form of unstructured and semi structured data and existing data analytics tools are not the feasible one because of dimensionality of the data. So Efficient Document Subspace Clustering (EDSC) is one of the methods which is existing in order to remove redundant data and also to form the discriminate segmentation points to detect the dimensionality of unseen subspaces in the clusters. The system also gives the specific count of clusters hidden in the subspaces.
The existing algorithms like Redundancy check in subspaces and algorithm for segmentation of data points are used in order to remove the redundancy and also to get the data points. In the proposed method a Fast Clustering Algorithm is applied in order to improve the clustering process.

Table I. Algorithm for Redundancy check in subspaces

```
Begin
        Cost D [x1,x2,x3 …xn ],  nϵδi
        if D ⩾ 1
                Remove xi
        else, go for δj , i< j
end
```

Here the above algorithm finds the cost-factor D which is used as a weighing attribute in order to find the redundancy and also to remove
the same.

Table II. Algorithm for finding segmentation of data points in subspace

```
Begin
        1.  I={Ik|kϵΔ}, where k=1, 2, ….., m
        2. J = I1 ∩ I2 ∩ ……Ik
        3. α = maxarg(D)
        4.  for α=1 to λn
        5. call algorithm for Redundancy check in
           subspaces
        6. Find Y (Δ1, Δ2, …. Δm)
        7. Show Y
end
```

In Table II, first it calculates the set I which are in high dimensional data in order to get the original data by using intersection operation. The redundancy can be removed by calling the redundancy removal algorithm and data segments are estimated and stored in Y.

IV. PROPOSED SYSTEM

The proposed system consists of different phases like, a collection of High Dimensional Data, Preprocessing, Redundancy Check,

segmentation of data points in subspace clustering [3]. Fig. 1 gives the design of the proposed system. The Subspace Clustering is done using Fast Clustering Algorithm and the final results are obtained. In the proposed system architecture first step is to collect High Dimensional Data that are generated from various resources like considering the social networking data set. The data generated from different twitter accounts are collected and formed a excel document and that is deposited in the database. The database holds the twitter data as high dimensional data.

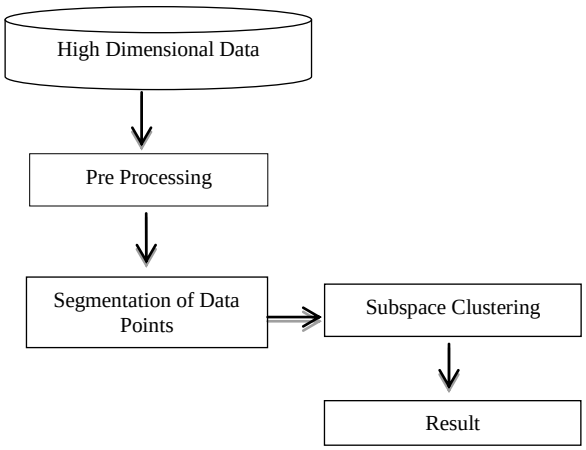

Fig. 1: System Architecture of Document subspace clustering

The proposed system improves the cluster analysis in various aspects like the speed of the clustering process, the accuracy of the clustering. It also performs the common operations of data mining, removes the redundancies and also the pre-processing to remove noise. The Fast Clustering Algorithm helps to improve the speed of the clustering and also the processing capability of the datasets is improved. The segmentation of data points will be done in the subspace clusters using the segmentation of data points. The speed of the subspace clustering process was improved considerably.

The preprocessing step cleans the data by removing noisy data and the next step is removing redundancy. After these two steps the data is formatted into an understandable and structure format and the segmentation of data points are done using segmentation of data point's algorithm. Then the subspace clusters are formed using the Fast Clustering Algorithm

The whole work is divided into 3 modules, those are given below.

- Preprocessing and Redundancy Check.
- Segmentation of Data Points.
- Subspace Clustering using Fast Clustering Algorithm

The Fast Clustering Algorithm is given in Table III, where it groups the related data by finding the relation between all the data points and forms subspace clusters and for each clusters one object will be allocated to find nearest mean of the next clusters in the dataset.

Table III. Algorithm Fast Clustering Algorithm

Input: Let the set data points be $Y = \{y_1, y_2, y_3, \ldots\ldots, y_n\}$
Output: Let the set of centers be $Z = \{z_1, z_2, \ldots\ldots, zc\}$

1: Select centers of the cluster 'c' Randomly.
2: Calculate the distance between each point of data and centers of clusters.
3: The points of data are allocated to the center of cluster whose distance is smallest from the center of the cluster to all the cluster centers.
4: Find the fresh cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' denotes the number of data points in i^{th} cluster.
5: calculate the distance between each data point and fresh cluster centers.
6: if no data point was reallocated
Then
stop
else
goto step 3
end

The Fast Clustering Algorithm works by grouping the related objects into clusters: Algorithm consists of following steps

1. It finds the relation between each object in the subspace clusters, the relation between each object can be done by tokenization. The tokenization is the process where it calculates the frequency of distribution of words in the particular data set and form the data matrix.

2. Group the related data into one cluster. Clusters are created by k-means function based on the frequency of distribution of words.

3. Select all the Subspace Clusters in the dataset by calculating the center of the cluster whose distance from the cluster center is smallest compared to all other cluster centers. .

4. Allocate an object to each cluster and recalculate the cluster center.

The data points segmentation process is completed at this stage. The planned system executes the method recursively for identifying redundant data by improving the preciseness of cluster positions. The proposed system highlights about the important part of performing subspace clustering mechanism.

The performance of the proposed EDSFCA algorithm is appraised by considering intersection operation that is carried out for finding the subspace positions and also for categorizing the redundant and non-redundant subspaces. In this process the subspace positions are reserved while the redundant data is deleted from the data matrix. The determined argument is used to denote the repetitive process.

To represent the initial position to maximum position of the subspace a loop is formulated. The location is recognized from the variable φ.

V. PERFORMANCE ANALYSIS

The specification of the information considered and execution time for making clusters is given in Table IV.

The performance of the work is measured in terms of time complexity and is compared with the Sembiring approach, EDSC approach proposed. The processing time vary as the input file size varies, so when the size of the input file is big then the time complexity increases

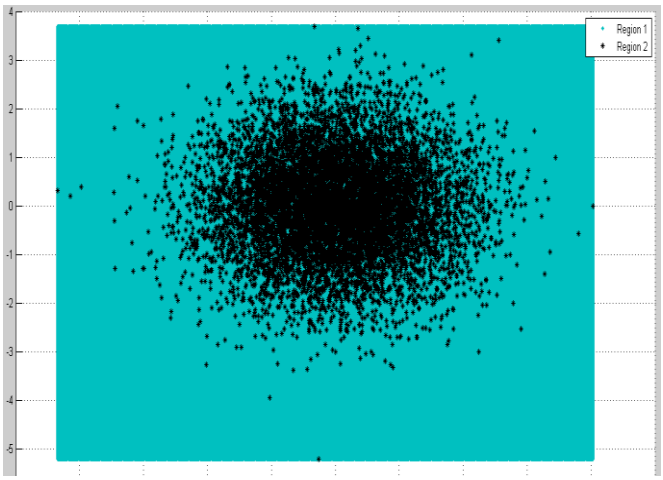

Fig. 2: The Scattered Twitter Data

The Fig. 2 shows the scattered twitter data where all the data are combined after applying the Fast Clustering Algorithm. The related data are grouped by forming clusters and the data which are not related do not form a cluster. In the Fig. 3 the region 1 shows the location allotted for the dataset and region 2 shows the location where the data points are present. The clustering is done based on the segmentation of data points. The clustering process reflects complete magnitudes of a data for the persistence of maximum exploitation of knowledge detection. Clustering groups the interrelated documents for browsing, this helps in data compression and effectively discovers the nearby neighbors of points.

From the Table IV it is observed that time taken for making clusters by the proposed algorithm is less as compared to the Sembiring and EDSC approach because the the Fast clustering Algorithm which performs tokenization process where it calculates the frequency distribution of the word in the particular dataset and forms the data matrix. Based on the frequency of distribution of data points the clusters are formed. Here the comparison has done with the Sembiring, EDSC, EDSFCA approaches, from the comparison table the time taken by the proposed approach for the given dataset of size from 10MB to 50 MB the time taken is between 0.203 to 0.214 which is very less compared to Sembiring approach.

Table IV. Processing Time in milliseconds.

Data size	Time taken by Sembiring Approach	Time taken by EDSC	Time taken by EDSFCA
10MB	0.414	0.211	0.203
20MB	0.457	0.213	0.210
30MB	0.614	0.214	0.212
40MB	0.703	0.216	0.213
50MB	0.815	0.216	0.214

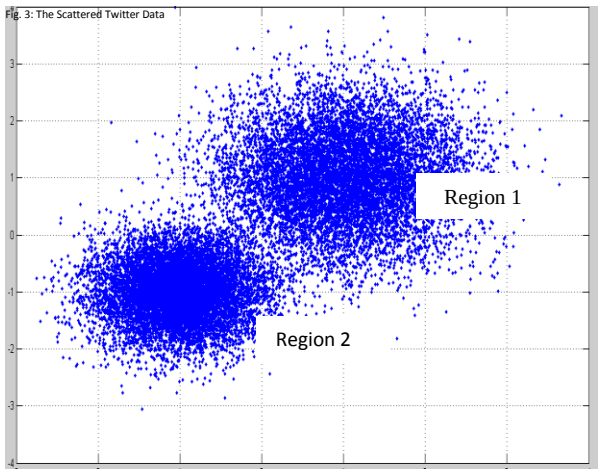

Fig. 3: The Clustered Twitter Data

VI. CONCLUSIONS

This work performs the preprocessing and removes redundancies in the data set and makes segmentation in subspace and efficient clusters using Fast Clustering Algorithm. This work solves the problems of clustering and segmentation efficiently. The result generated by the work is useful for many applications in Machine Learning and pattern recognition.

REFERENCES

[1] P. Buhlmann, S. van de Geer, "Statistics for High-Dimensional Data: Methods, Theory and Applications", Springer Science & Business Media, 2011

[2] V. B. Canedo, N. S. Marono, A. A. Betanzos, "Feature Selection for High-Dimensional Data", Springer-Computer, 2015

[3] Radhika K R, Pushpa C N, Thriveni J and Venugopal K R, "EDSC: Efficient Document Subspace Clustering Technique for High-Dimensional Data", In proceedings of International Conference on Computational Techniques in Information and Communication Technologies, Delhi , PP. 11-13, March 2016.

[4] M Verleysen, "Learning High-Dimensional Data", University atholique Louvain, Microelectronics laboratory, pp. 141-162, 2003.
[5] A Petukhov and I Kozlov, "Greedy Algorithm for Subspace Clustering from Corrupted and Incomplete Data", IEEE Transaction on Information Security, 2015.
[6] Amardeep Kaur and Amitava Datta. "A Novel Algorithm for Fast and Scalable Subspace Clustring in High Dimensional Data", Journal of BigData, 2015.
[7] C Yang, D Robinson and R Vidal, "Sparse Subspace Clustering with Missing Entries", In Proceedings of the 32nd International Conference on Machine Learning, pp. 2463-2472, 2015.
[8] Singh Vijendra, "Efficient Clustering for High Dimensional Data:Subspace Based Clustering and Density Based Clustering", Information Technology vol. 10, pp. 1092-1105, 2011.
[9] Lance Parson, Ehtesham Haque and Huan Liu, "Subspace Clustering for High Dimensional Data: A Review", sigkdd Explorations, vol. 6, pp. 90-105, 2004.
[10] V. Kumatha and S. Palaniammal, "Evaluation of Subspace Clusteing of High Dimensional Data", International Journal of Computational Science and Applications", pp. 11-14, 2012.
[11] Sunita Jahirabadkar and Parag Kulkarni, " Clustering for High Dimensional Data:Density Based Subspace Clustering Algoriithms",International Journal of Coomputer Applications (0975-8887), vol. 63, pp. 29-35, 2013.
[12] Singh Vijendra and Sahoo Laxman, "Subspace clustering of High Dimensional Data: An Evolutionary Approach", Applied Computational Intelligence and Soft Computing", vol. 2013, Article ID 863146, pp. 12.
[13] Hans-peter Kriegel, Peer Kroger, Matthias Renz, Sebastian Wurst, "A Generic Framework for Efficient Subspace Clustering of High Dimensional Data", In Proceedings of 5th IEEE International Conference on Data Mining (ICDM), Houston, TX, 2005.
[14] Y Wang, Y-X Wang, and A Singh, "Clustering Consistent Sparse Subspace Clustering", Carnegie Mellon University, USA, arXiv preprint arXiv: 1504.01046, 2015.
[15] J Wei, M Wang and Q Wu, "Study on Different Representation Methods for Subspace Segmentation", International Journal of Grid Distribution Computing, Vol. 8, no.1, pp.259-268, 2015.
[16] C Giraud Taylor and Francis group, "Introduction to High-Dimensional Statistics", xv+252 pp. ISBN: 978-1-482-23794-8 2014.
[17] R Agrawal, J Gehrke, D Gunopulos, and P Raghavan ,"Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications". In Transaction of Data Mining and Knowledge Discovery, vol. 11, Issue. 1, pp. 5-33, 2005.
[18] N Tomašev, M Radovanović, D Mladenić and M Ivanović, "Hubness- based Clustering of High-dimensional Data", In Partitional Clustering Algorithms, Springer International Publishing, pp. 353-386, 2015.
[19] Shuyun Wang, Yingjie Fan, Chenghong Zhang, HeXiang Xu, Xiulan Hao and Yunfa Hu, " Subspace Clustering of High Dimensional Data Streams", In Proceedings of 7th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2008), pp. 14-16, Portland, USA.
[20] Manolis C. Tsakiris and René Vida, "Abstract algebraic-geometric subspace clustering", In 48th Asilomar Conference on Signals, Systems and Computers , EISSN: 1058-6393, 2-5 Nov. 2014, CA, USA.
[21] Ezgi Can Ozan and Serkan Kiranyaz, "K-Subspaces Quantization for Approximate Nearest Neighbor Search", In IEEE Transactions on Knowledge and Data engineering, Vol. 28, No. 7, pp. 1722-1733, 2016.
[22] Han Zhai, Hongyan Zhang, Liangpei Zhang, Pingxiang Li and Antonio Plaza, "A New Sparse Subspace Clustering Algorithm for Hyperspectral Remote Sensing Imagery", In proceedings of IEEE Geoscience and Remote Sensing Letters, vol. 14, Issue. 1, pp. 43 – 47, 2017.
[23] Shulin Wang, Fang Chen and Jianwen Fang, "Spectral clustering of high-dimensional data via Nonnegative Matrix Factorizationion", In proceedings of International Joint Conference on Neural Network (IJCNN), pp. 12-17, 2015, Ireland.
[24] Junjian Zhang, Chun-Guang Li, Honggang Zhang and Jun Guo, "Low-rank and structured sparse subspace clustering", n proceedings of Visual Communication and Image Processing (VCIP), pp. 27-30, 2016, China.
[25] Alexander Petukhov and Inna Kozlov, "Greedy algorithm for subspace clustering from corrupted and incomplete data", In proceedings of International Conference on Sampling Theory and Applications (SampTA), pp. 25-29, 2015, USA.
[26] Ran He, Liang Wang, Zhenan Sun, Yingya Zhang and Bo Li, "Information Theoretic Subspace Clustering", IEEE Transactions on Neural Networks and Learning Systems, vol. 27, Issue. 12, pp. 2643-2655, 2016.
[27] Yifan Fu, Junbin Gao, David Tien, Zhouchen Lin and Xia Hong, " Tensor LRR and sparse coding-based subspace clustering", In IEEETransactions on Neural Networks and Learning Systems, vol. 27, Issue. 10, pp. 2120-2133, 2016.

