# Insights to Existing Techniques of Subspace Clustering in High-Dimensional Data

Radhika K R, Pushpa C N, Thriveni J, Venugopal K R
Department of Computer Science and Engineering,
University Visvesvaraya College of Engineering,
Bangalore, India.
radhika@bmsit.in

**Abstract—** With an increasing attention towards the significance of increasing data sizes, the area of clustering is still under the focus of the researchers. There has been extensive research work from more than a decade in this regards and still the problems of understanding the data with a shape of business logic is yet to be seen. Hence, this manuscript gives the snapshots of the problems associated with high-dimensional data and cluster analysis that finally leads to discussion of Subspace clustering problem. It has been seen that large scale of research attempt is dedicated more towards subspace clustering process and quite less for dimensionality reduction process. The manuscript reviews the existing technique and draws a research gap of the existing literatures. Finally, the manuscript highlights a conceptual framework in order to overcome the problems associated with the subspace clustering problem

**Keywords-**component; High-Dimensional Data, Clustering, Subspace Clustering, Cluster Analysis, Dimensionality reduction.

## 1 INTRODUCTION

In the present era of digitization, majority of the users are continually moving on the pervasive computing that has significantly influenced the telecommunication section and social networking section. Owing to advancement in the communication technology, now storage is not a big problem as the data is growing from mere megabytes to petabytes just within a matter of a few seconds [1]. Although cloud offers a better storage privilege, but cloud not offer analysis effectively. Data analytics is still under research and development. However, in the area of datamining, such issues are continuously considered as critical problems where the probable solution lies in cluster analysis [2]. In easier manner, it can be said that cluster analysis assist in making the messy data to a meaningful data that will be easy to analyse. It attempts to discover a significant group that posses a similar characteristic. A conventional clustering technique considers entire dimensions of a data for the purpose of maximum utilization of knowledge discovery [3]. Although, for smaller amount of data, there are no issues which always arise when the size of the data increases. At present, almost all the data being captured for analysis are of high-dimensional data. Evolution of such data may be from various resources e.g. social network, medical science, education, etc.

One of the essential charecteristics of high-dimensional data is its nature of inherent irrelevancy in its dimensions that can often obfuscate the existing clustering technique by concealing the clusters with a corrupt or incomplete or noisy data [4]. Another significant problem arises when masking of the clusters arises due to locations of each object which are equidistant from each other [5]. A remedy for such problem is found as feature selection techniques for enhancing quality of clusters. A feature selection technique explores the possibility of dimensional subset for carrying out clustering by eliminating unnecessary and inappropriate dimensions. The entire dataset is analysed using this technique.

There are large variations of applications for subspace clustering. Essentially, subspace clustering can be suitably applied for exploring the relationship among various perspectives. Usually, it is used more on bioinformatics [6] and text-mining [7]. Formulations of the data hierarchies are one of the challenging tasks and are being under research with an aid of subspace clustering technique [8]. Usage of subspace clustering over micro-array DNA technology assists us to understand the medical data more in-depth, where knowledge discovery is done exploring and evaluating the significant data patterns. Contribution of clustering algorithm is quite significant in medical sector as it precisely assist to parameterize the specific disease more vividly. In this regards, there is one common phrase called as "Curse of Dimensionality", which is still an open problem with high-dimensional data [9]. Due to increased dimensionality of the data as well as size of micro-array, the task of knowledge discovery in mining becomes more complex for conventional clustering techniques. Therefore, this paper gives a snapshot of clustering and reviews some of the existing subspace clustering algorithms and dimensionality reduction techniques.

Organization: Section 2 discusses about the evolution and fundamentals of High-Dimensional data followed by essentials of cluster analysis in Section 3. Section 4 discusses about the problem of subspace clustering techniques with discussion of existing research work in Section 5.

Section 6 discusses about the research gap followed by Conclusion and Future Work in Section 7.

## 2 HIGH DIMENSIONAL DATA

With the proliferation of mobile networks as well as cloud

computing, the usages of the various applications as well as user based have exponentially increased. Such increase in the user base has finally resulted in the evolution of massive data that is critically required to be analysed. Hence, such a massive set of data that exponentially increases with evolution of time can be termed as high-dimensional data. At present, it is quite a challenging task to understand and effectively have a clear picture of high-dimensional data in the various area of engineering. Majority of the existing techniques are usually found to depend on the techniques of learning using manifold approaches that essentially develops a single embedding of the information or use to select subspace to explore the original attributes along with their corresponding subsets for showcasing the structure [11].

The growth of massive data doesn't encounter any issues in storage, but all sorts of problems starts surfacing up when it comes to the analysis of the data to extract a specific set of knowledge. However, it poses a greater deal of problem if the dimensionality of the data is very high and it poses a challenging situation for carrying out the clustering task.

### 2.1 Significance of Dimensionality of Dataset

The concept of the dimensionality of the data plays a critical role in the area of machine learning techniques. There are various applications and domains at present where massive forms of data are being generated. Some examples are-data generated from sensor network, social network, electronic medical records of healthcare industry, database of e-governance frameworks etc. Such data are continuous growing data, where the dimensionality keeps on increasing. A significant amount of data is being generated from the genetic engineering, micro-biology, etc., where DNA sequence yields a massive number of information in one sequence itself. In case, a conventional clustering technique is used for studying such high-dimensional data, the frequency vector of a specific word is almost equivalent to the size of the dictionary. Such facts will also mean equivalent study from social network analysis as well other management techniques in the dimensionality study. The technique to deal with high-dimensional data is to perform precise identification of the clusters that depends on object similarities evaluated in terms of distance function. Owing to the massively growing dimensions, conventional clustering algorithms are not directly applicable to solve the problem of clustering. Hence, dimensionality of the data plays a significant role in engineering and science.

### 2.2 Issues with High-Dimensional Data

Here are lists of critical issues owing to the high-dimensional data with respect to clustering.

1. With the increment in the size of the dimensionality of the data, the space volume of the data too increases rendering the existing data to be sparse. High-dimensional data also results in failure to explore the precise clusters during cluster analysis.
2. High-dimensional data renders Euclidean distance meaningless as owing to the concentration effects, the distance usually converges at a given points of a dataset.
3. One of the biggest problems in high-dimensional data is the insufficiency of the global filtering process of the different subspaces for different clusters. It is also called as relevancy problems of the local feature.
4. There is ahigher possibility of correlation of attributes for the massive quantity of the attributes. This fact may result in formation of the clusters in random position in the subspaces.
5. There is a greater deal of feasibility for higher dimensional data to possess irrelevant characteristics. These facts have significant influence in data analysis.

Hence, due to high-dimensional data, it is not possible to extract the logical clusters from the subspaces as there is a strong relationship established to show that distance is directly proportional to the dimensionality of an object. Another significant problem that surfaces is failure in identifying the specific patterns owing to inappropriate dimensions. The most frequently adopted Principle Component Analysis used for minimizing the dimensionality is also not sufficient. This is because of the reason that Principle Component Analysis can minimize the space of the high-dimensional data for homogeneous objects. However, it is most likely that high-dimensional data in real-time will possess heterogeneous data groups that cannot be processed using conventional Principle Component Analysis. Hence, it is very important to emphasize the problem and investigate on various factors responsible for clustering problems in high-dimensional data. The next section brief about the standard technique of cluster analysis that is also found to be adopted by various existing researchers.

## 3 CLUSTER ANALYSIS

The prime task of the cluster analysis is to extract the meaningful information from the set of information. It can also be represented as a process of segregating core data to meaningful clusters (or groups) that assist in better knowledge discovery. Therefore, the cluster analysis can be adopted for the purpose of grouping the related information for extensive exploration of various targets (proteins, genes etc.) with equivalent operations and to furnish a clustering of spatial positions that are susceptible to natural calamities. In many studies e.g. [11], it was seen that cluster analysis can also assists in compressing data and precisely identifying the nearest neighbour locus. The area of cluster analysis is also studied in biology, social science, statistics, psychology, pattern recognition, machine learning, retrieval of information, and data mining operations. While discussing cluster analysis, the term "object" is frequently used in the literature that signifies number of events or a set of observation. Therefore, the operation of cluster analysis usually groups such objects existing in the information that illustrates the objects along with establishing relationship between numbers of objects. While doing so, it is only checked if similar objects in multiple groups have similar characteristics. The clustering is optimal for greater similarity extent within the groups.

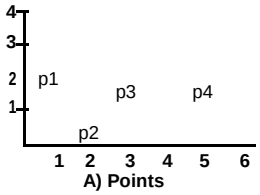
Analysing a cluster can be quite constructive and valuable in many ways, but it has contradictory sides. Cluster analysis

is all about studying and grouping the data based on certain unique characteristics. However, in real-time, the occurrences and evolutions of the data are quite different. A simple social networking site alone can generate multiple forms of text data, image data, and video data, which can also be called as unstructured data [12]. Understanding such complicated data is extremely essential as with normalizing the stream of massive data, it cannot be subjected to datamining algorithms. Hence, what is important is getting the precise data for data analysis and not the cluster analysis. However, cluster analysis is just a starting point of data analysis that can solve various complicated issues in data engineering. The significance of the data analysis can be understood by the problem when there is a need of undertaking a decision for a specific feature to be adopted for mapping the objects in statistics, datamining, and pattern recognition. The basic operations of the cluster analysis are done by selection of an appropriate features and moves ahead. The common techniques (Figure.1) adopted while performing clustering are discussed as below:

1. Data Matrix: Data matrix is a form of representation of the data in the form of matrix with specific rows and columns where the data are represented in the form of vector or points in a 2-3 dimensional space with specific set of attribute [11].

2. Proximity Matrix: While performing cluster analysis, the system considers genuine (original) information as data matrix where different forms of clustering techniques are also applied with various forms of matrix (dissimilar/similar). For easiness in usage, both the forms of dissimilar or similar matrix are called as proximity matrix [11].

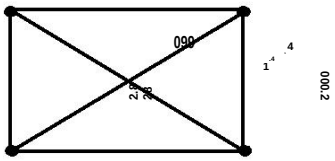
3. Proximity Graph: A weighted graph is defined by the proximity matrix, where the nodes represent data points to be clustered and the proximities between the data points are represented as weighted edges. Such forms of the graphs are usually directed graph that corresponds to a proximity matrix of asymmetric nature [11].




A) Points

| point | x | y |
|---|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

B) Data Matrix




| | p1 | p2 | p3 | p4 |
|---|---|---|---|---|
| p1 | 0.000 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0.000 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0.000 | 2.000 |
| p4 | 5.099 | 3.162 | 2.000 | 0.000 |

Figure 1 Clustering Techniques

## 4 PROBLEM OF SUBSPACE CLUSTERING

The problem of the subspace clustering has been vividly discussed by Vidal [12]. Accordingly, the mathematical representation of the subspaces was shown as,

$$S_i = \{x \in R^D : x = U_i + U_{iY}\} \quad (1)$$

In the above equation, S represents affine space, D is the maximum dimension, μ represents random data point existing in the subspace, Y represents low dimension, and U represents a basis function. Figure. 2 is the sample data points on the plane for the study. The prime target of the subspace clustering is to identify the number of the subspaces along with their dimensions and segmentations of the points.

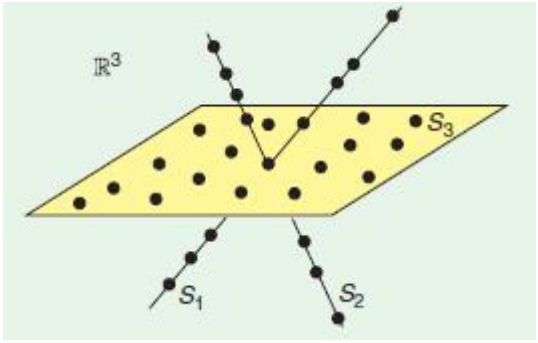



Figure 2 Sample Data Points [12]

In this regard, the authors [12] have discussed about the evolution of Principle Component Analysis that is a representation of a problem when the subspaces quantity is equated to 1 and such issue minimized to explore a basis function, a vector, a low dimensional representation, and a dimension. Although there were enough studies conducted over Principle Component Analysis, it is now obsolete in the area of high-dimensional data analysis. However, when the number of data points are more than one, various forms of problems surfaces up

1. There is a potentially strong relationship between model estimation and data segmentation. Therefore, it is possible for someone to fit in a unit subspace if the data segmentation is known for each set of points using Principal Component Analysis. The similar aspects happen vice-versa. But, in reality if there is a need for solving a problem, then neither subspace parameters are known nor segmentation of the data.
2. Theory and practical study say that information about the data distribution is usually not known in subspace. Although, subspace clustering problem minimizes if each centers of the clusters is populated with data distribution for different subspaces. Problems occur when the data distribution occurs randomly.
3. There is a strong feasibility that orientation as well as position of respective subspaces ccan be relatively random order.
4. Subspace clustering cannot be effectively carried out in case of missing data or noisy data. Alt hough there is a pre-processing operators, but still

they are not applicable in unstructured data to a larger extent.

## 5 EXISTING RESEARCH WORK

This section discusses about the existing recent attempts that has been focused towards subspace clustering problems and understanding its criticality with respect to high-dimensional data. The study of this section is done by going through 97 research articles available in the existing literatures. Studies using similar kinds of techniques or the studies using similar enhancement criteria of the problems are excluded from discussion and finally short listed 20 significant studies categorized into following sections

### 5.1 Studies towards Subspace Clustering

Research work toward subspace clustering dates back in 1990 where the frequently adopted clustering techniques for study was found to be hierarchical and partitional [13]. It was in 1998, a significant studies took place to address problems of scalability of data and extraction of clusters hidden inside the subspaces of the critical high-dimensional data.

The work presented by Agrawal et al., [14] addresses such problems by introducing CLIQUE algorithm for identifying the subspace of the dense cluster. The outcome of CLIQUE (Clustering in Quest) was compared with existing clustering algorithm BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) and DBSCAN (Density Based Spatial Clustering of Applications with Noise) to see better scalability with increase of dimensionality of the data. The next set of significant research work on the similar problem of subspace clustering was seen in the year 2004.

Baumgartner et al., [15] have adopted the technique of k-nearest neighbor approach along with the technique to explore the interesting cluster behavioral traits within the subspace. The authors have named it as SURFING that stands for Subspace Relevant for Clustering. The design of the algorithm is done as parameterless and uses ranking mechanism to explore the best cluster with the subspaces. The algorithm was compared with previous algorithm CLIQUE on gene dataset and found to outperform CLIQUE with respect to data quality.

Gan and Wu [16] proposed an algorithm called as SUBCAD for reducing the objective function required for clustering. The authors have used separation and compactness for representing the subspace of every cluster using iterative methods experimented over Wisconin breast cancer data, soybean data, and congressional voting data. Similar problem is also addressed in the work carried out by Kailing et al., [17] in same year of 2004. The authors have presented an algorithm called as SUBCLU which means Subspace Clustering using density-based approach. The work adopts the underlying architecture of DBSCAN with density-connectivity features for randomly identifying location and shape of the clusters within a subspace. SUBCLU was compared with CLIQUE on same dataset to find SUBCLU the appropriate technique for exploring clusters. The duration of end of year of 2004 till 2008-year end was found with less significant studies. In 2009, another significant research work has been surfaced by Muller et al., [18] focusing on similar problem.

MINECLUS is a cell-based approach, PROCLUS is clustering based approach. The study also found that existing technique of SUBCLUE and CLIQUE is less compatible with high-dimensional data in extensive study. In the year 2010, Sembiring et al., [19] have introduced a projected clustering technique using Weka. The authors have performed a comparative analysis with PROCLUS and P3C with respect to accuracy and relevancy. Another significant study was introduced in 2012 by Tatu et al., [20] by presenting an algorithm called as ClustMails. This is a completely new and unique approach as compared to all the works carried out before 2012 and the uniqueness lies in adoption of visual analysis approach using Weka. The outcome of the study was compared with VISA which is a type of subspace cluster visualization framework designed by Assent et al., [21].

However, better version of visualization-based approach was witnessed in the work of Liu et al., [22] in 2014. The study uses subspace clustering without any consideration of single manifold data. The technique evaluates both internal dimensionality as well as the linear basis of a subspaces discovered from subspace clustering. The evaluation is carried over MNIST dataset.

Most recently, the topic of subspace clustering is still under the scanner of researchers. 12 significant research works have been witnessed in 2015. Chakraborty and Roy [23] have adopted k–means clustering technique along with implementation of fuzzy clustering approach. The study is implemented on Matlab and is compared with k–means clustering. Chang et al., [24] have investigated on spectral clustering technique and proposed an optimization technique on convex formulations. The study was experimented over JAFFE dataset, UMIST face dataset, BinAlpha dataset, USPS dataset, and YaleB dataset.

Adoption of Gaussian regression technique was done by Li et al., [25] for performing subspace clustering considering noise aspects. The study was carried out over Hopkins 155 dataset and AR dataset (along with Yale and MINST dataset) to show positive effect of grouping on clusters. Adoption of thresholding based technique using ridge regression technique was seen in work of Peng et al., [26].

Another unique implementation of segmentation based approach was seen in work of Wang and Fu [27], where the authors used subspace clustering based on sparsity factor. Petukhov and Kozlov [28] have presented a greed approach on subspace clustering for partial data. Wei et al., [29] have investigated various segmentation techniques of subspace clustering.

Consideration of sparsity on subspace clustering is emphasized in the work of Wang et al., [30]. The authors have presented an algorithm for noise-free as well as noisy data with SEP (Self-expressive property) charecteristics. Another unique work is proposed Wang and Zhu [31] towards the same problems. The authors have introduced a Bayesian framework using Dirichlet process in principle component analysis. The work is found to support its technique towards motion segmentation. Yang et al., [32] have presented a technique for clustering of subspaces. The authors have used a technique where the information of the knowledge is monitored for mechanizing convex optimization problem. Tomasev et al., [33] have emphasized the usage of significant charecteristics of k-nearest neighbour technique called as hubness.

The study conducted by Yin et al. [34] has introduced multi-view clustering approach for sparse subspace representation.

### 5.2 Studies towards Dimensionality Reduction

In the area of datamining, clustering plays a significant role. However, owing to curse of dimensionality, the concept refers high dimensional data as useless. Hence, subspace clustering attempts to solve this problem by extracting relevant and precise information from the subspaces in clusters. Another technique to address this problem is to work on reducing the dimensionality of the clusters. Works done by Assent et al., [35] have presented a technique for detecting the non-redundant clusters for the purpose of minimizing the result set. Another unique work of nonlinear dimensionality reduction was carried out by Goh [36] using Riemannian manifold.

Niu et al., [37] have investigated spectral clustering approach that involuntarily learns the pertinent dimensions. The Table 1 gives the List of existing Subspace Clustering Techniques. Gunnermann et al., [38] have investigated on KNIME framework which is used for advanced data mining and incorporated a subspace clustering algorithm.

However, the study is not witnessed with any extended outcome analysis. Study on dimensionality reduction using k-means clustering is quite frequently observed in the literature. A similar cadre of study was witnessed in the work done by George [39] where the technique was integrated with principle component analysis. The outcome of the study evaluated on Parkinson's dataset exhibits precise clustering. Heckel et al., [40] studied the impact of the dimensionality problem using arbitrary projection on space subspace clustering technique. Table 1 showcases the different forms of subspace clustering that were frequently investigated by various researcher in past decades.



TABLE 1 LIST OF EXISTING SUBSPACE CLUSTERING TECHNIQUES

| Techniques | Parallel Axis | Grid oriented | Overlapping | Search | | Noise resiliency | Free from data order | Free from dimension order | Random Subspace dimensionality |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Top-down | Bottom-up | | | | |
| CLIQUE | ✓ | ✓ | ✓ | × | ✓ | × | ✓ | ✓ | ✓ |
| DOC | ✓ | × | × | ✓ | × | ✓ | ✓ | ✓ | ✓ |
| SUBCLU | ✓ | × | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ |
| MAFIA | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ |
| PROCLUS | ✓ | × | | ✓ | × | ✓ | ✓ | ✓ | × |
| FIRES | ✓ | × | ✓ | × | × | ✓ | ✓ | ✓ | ✓ |
| DENCLU | ✓ | ✓ | ✓ | × | ✓ | × | ✓ | ✓ | × |
| DENCOS | ✓ | × | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ |
| INSCY | ✓ | × | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ |
| DUSC | ✓ | × | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ |
| OptiGrid | ✓ | ✓ | ✓ | × | ✓ | ✓ | ✓ | ✓ | ✓ |

RESEARCH GAP

The prior sections have discussed about the existing techniques exclusively seen for addressing subspace clustering and dimensionality reduction problems associated with high-dimensional data. From the study, it can be seen that there are various forms of the subspace clustering technique for the purpose of evaluating as well as exploring a better presentation of the various subspace clustering. However, there are effectiveness as well as limitations associated with almost all the existing systems.

1. The literatures have witnessed some of the subspace clustering algorithm that is highly dependent on positioning of grids e.g. CLIQUE, MAFIA, DENCLU, OptiGrid etc. In such approach, there is a higher possibility of an object to be missed even if it originally belongs to a cluster. Incorrect grid positioning principle will also lead to positioning of noisy data on grid resulting in degradation in analysis process.
2. Existing technique e.g. SUBCAD still has an open problem about the ambiguity of model selection as it is dependent on the cluster numbers as input. Therefore, still now, much clarity is not illustrated how to use the existing clustering techniques for selecting the input parameters into this subspace clustering algorithm is not addressed in any research work.

3. The concept of dimensionality reduction is also associated with various issues that are still unaddressed. It is quite difficult and challenging task to perform interpretation of the resulting cluster as attributes that are transformed doesn't possess much logical meaning. Majority of the dimensionality reduction techniques implemented till date are only focused on cluster analysis on a specific subspace only. In case the information of other objects is clustered separately that is different from existing subspace, then such information will be volatile and vanished.

## 6 CONCLUSION & FUTURE WORK

This paper has discussed the essentials of high-dimensional data and its critical problems of cluster analysis. Although subspace clustering is one of the biggest boon in datamining, still there lays a massive problem in extracting the clusters from the hidden subspaces. Various existing research techniques have been discussed in this paper pertaining to work done in subspace clustering as well as dimensionality reduction. Our future work direction will be to evolve up with a technique that can mitigate the problem of subspace clustering over high dimensional data. The secondary aim of the future study will be to extract precise clusters that are usually found hidden in subspaces of the high-dimensional data.

The research objectives to accomplish the above mentioned aim are as follows i) To carry out a study for the existing techniques of subspace clustering and dimensionality reduction and extract a research gap, ii) To design a technique for optimizing the subspace clustering using Laplacian graph for extracting both global and local manifold structures in high-dimensional data, iii) To extend the optimization of the subspace clustering using evolutionary approach in order to identify the precise number of clusters and their respective dimensions, iv) To evolve up with a mathematical modelling of sparse subspace clustering technique to make it resilient against non-Gaussian noise, v) To design an integrated technique for performing subspace clustering using subspace learning and subspace clustering approach, vi) To carry out comparative performance analysis of the outcome of the proposed system with the existing one.

## REFERENCES

[1] V. M. Schönberger, K. Cukier, "Big Data: A Revolution That Will Transform How We Live, Work, and Think, Houghton Mifflin Harcourt", Business & Economics, 2013

[2] M. R. Anderberg, "Cluster Analysis for Applications: Probability and Mathematical Statistics", Academic Press, Mathematics, 2014

[3] M. J. Zaki, W. Meira, "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2014

[4] E. Elhamifar, R. Vidal, "Sparse Subspace Clustering: Algorithm, Theory, and Applications", Arxiv, 2012

[5] L. Parson, E. Haque, H. Liu, "Subspace Clustering for high dimensional data: A review," ACM Special Issues of Imbalanced Dataset, vol.6, iss.1, pp.90-105, 2004.

[6] J. D. MacCuish, N. E. MacCuish, "Clustering in Bioinformatics and Drug Discovery", CRC Press, Mathematics, 2010

[7] J. L. Balcazar, F. Bonchi, A. Gionis, M. Sebag, "Machine Learning and Knowledge Discovery in Databases", Springer Science & Business Media, 2010

[8] C. C. Aggarwal, C. K. Reddy, "Data Clustering: Algorithms and Applications", CRC Press, Business & Economics, 2013

[9] W. B. Powell, "Approximate Dynamic Programming: Solving the Curses of Dimensionality", John Wiley & Sons, Mathematics, 2011

[10] M. Steinbach, L. Ertoz, and V. Kumar, "The Challenges of Clustering High Dimensional Data", Springer New Directions in Statistical Physics, pp 273-309, 2004

[11] N. Sawant, H. Shah, "Big Data Application Architecture Q&A: A Problem - Solution Approach", Apress, Computers, 2013

[12] R.Vidal, "Subspace Clustering", IEEE Signal Processing Magazine, March 2011.

[13] M. E. Celebi, "Partitional Clustering Algorithms", Springer-Technology & Engineering, 2014

[14] R. Agrawal, J. Gehrke, D. Gunopulos, "Automatic subspace clustering of high dimensional data for data mining applications". ACM_ Proceedings of the International Conference on Management of Data, Vol. 27. No. 2. 1998.

[15] C. Baumgartner, C. Plant, K. Kailing, H-P Kriegel, "Subspace selection for clustering high-dimensional data." Fourth IEEE International Conference, 2004.

[16] G. Gan, and J. Wu. "Subspace clustering for high dimensional categorical data." ACM SIGKDD Explorations Newsletter, Vol.6, No.2, pp.87-94, 2004.

[17] K. Kailing, H-P Kriegel, and P. Kröger. "Density-connected subspace clustering for high-dimensional data." Proceedings of International Conference on data Mining, Vol. 4. Pp.246-257, 2004.

[18] E. Muller, S. Gunnemann, I. Assent, T. Seidl, "Evaluating clustering in subspace projections of high dimensional data" ACM-Proceedings of the VLDB Endowment, Vol.2, No.1, pp.1270-1281, 2009.

[19] R.W. Sembiring, J M. Zain, and A. Embong. "Clustering high dimensional data using subspace and projected clustering algorithms", arXiv preprint arXiv: 1009.0384, 2010.

[20] A. Tatu, L.Zhang, E.Bertini, "Clustnails: Visual analysis of subspace clusters." IEEE-Tsinghua Science and Technology, Vol.17, No.4, pp. 419-428, 2012.

[21] I. Assent, R. Krieger, E. Muller, T. Seidl, "VISA: visual subspace clustering analysis." ACM-SIGKDD Explorations Newsletter, Vol. 9, No.2, pp.5-12, 2007.

[22] S. Liu, B. Wang, J.J. Thiagarajan, "Visual Exploration Of High-Dimensional Data: Subspace Analysis Through Dynamic Projections". Technical Report UUSCI-2014-003, SCI Institute, University of Utah, 2014.

[23] S. Chakraborty, B. Roy. "Performance Analysis of Subspace Clustering Algorithms in Biological Data", International Journal of Advanced Research in Computer and Communication Engineering, vol.4, Iss.2, 2015

[24] X. Chang, F. Nie, Z. Ma, Y. Yang "A convex formulation for spectral shrunk clustering." arXiv preprint arXiv:1411.6308, 2014.

[25] B. Li, Y. Zhang, Z. Lin, H. Lu, "Subspace Clustering by Mixture of Gaussian Regression" Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[26] X. Peng, Y. Zhang, and H. Tang. "Robust subspace clustering via

thresholding ridge regression." AAAI Conference on Artificial Intelligence, 2015.

[27] J. Wang, FZ. Fu,"Online Motion Segmentation Based on Sparse Subspace Clustering." Journal of Information & Computational Science, vol.12, No.4, pp.1293-1300, 2015

[28] A. Petukhov, I. Kozlov, "fast Greedy Algorithm for Subspace Clustering from Corrupted and Incomplete Data." CoRR abs/1306.1716, 2013

[29] J. Wei, M. Wang, and Q. Wu. "Study on Different Representation Methods for Subspace Segmentation" International Journal of Grid and Distributed Computing, vol. 8, No.1, pp.259-268, 2015

[30] Y. Wang, Y-X Wang, and A Singh. "Clustering Consistent Sparse Subspace Clustering." arXiv preprint arXiv: 1504.01046, 2015.

[31] Y. Wang and J. Zhu. "DP-space: Bayesian Nonparametric Subspace Clustering with Small-variance Asymptotics," International Conference on Machine Learning, 2015.

[32] C. Yang, D. Robinson, and R. Vidal, "Sparse Subspace Clustering with Missing Entries." Proceedings of the 32nd International Conference on Machine Learning, 2015.

[33] N. Tomasev, M. Radovanovic, D. Mladenic, "Hubness-based clustering of high-dimensional data." Springer International Publishing of Partitional Clustering Algorithms., pp. 353-386, 2015.

[34] Q. Yin, S. Wu, R. He, L. Wang "Multi-view clustering via pairwise sparse subspace representation." Elsevier-Neurocomputing, vol.156, pp.12-21., 2015

[35] I. Assent, E. Muller, S. Gunnemann, "Less is more: Non-redundant subspace clustering." ACM-MultiClust-KDD, 2010.

[36] A. Goh, "Riemannian manifold clustering and dimensionality reduction for vision-based analysis." Springer-Machine Learning for Vision-Based Motion Analysis, pp.27-53 , 2011.

[37] D. Niu, J.G. Dy, and M.I. Jordan. "Dimensionality reduction for spectral clustering." International Conference on Artificial Intelligence and Statistics. 2011.

[38] S. Gunnemann, H. Kremer, R. Musiol, "A subspace clustering extension for the KNIME data mining framework," IEEE-Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on. IEEE, 2012

[39] A.George,"Efficient high dimension data clustering using constraint-partitioning k-means algorithm." International Arab Journal Journal of Information Technology, Vol. 10, No.5, pp.467-476, 2013.

[40] R. Heckel, M. Tschannen, and H. Bolcskei, "Subspace clustering of dimensionality-reduced data." IEEE International Symposium on Information Theory, 2014.