ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

A Subset Sparse Based Subspace Clustering on High Dimensional Data

[1] Prathima V.R, [2] Dr. Fayaz.K
[1] Research Scholar, Rayalaseema University, Kurnool, Andhra Pradesh, India1
[2] Research Supervisor, S.K University, Anantapur, Andhra Pradesh, India

Abstract- **Data mining and machine learning has been an important research topic in recent years. The topic of sparsely distributed data and the various issues related with curse of dimensionality of high-dimensional data, make most of the traditional clustering algorithm, lose action in high dimensional space. Therefore, clustering of data in high dimensional space is becoming the hot research areas. Clustering in data mining can be used as a data exploration or future prediction tool. With the advent of raise in huge data or high dimensional data such as DNA arrays, Images or GPS data, bag-of-words document representation etc., the goal of clustering is to group multiple data points in such way that they can be represented more efficiently for better understanding of the data. In this context we study the pitfalls of high dimensional data clustering concepts and algorithms are discussed then we study the SSC, SSSC, SMRS algorithms. This paper offers a subset based algorithm for automatically determining the optimal number of clusters on high dimensional data. The Main aim of this paper is to design an algorithm with reasonable complexity which computes representatives and clustering high-dimensional data accurately. In this paper we have made the following contributions i.e. designing of algorithm which used the divide-and-conquer strategy, which can able to compute the representatives within reasonable time and this algorithm is named as Hierarchical Sparse representatives.**

Index Terms— clustering, high dimensional data, hierarchical sparse representative, SSSC.

I. INTRODUCTION

KDD is the process of automatically searching massive volumes of data for patterns using primary tools such as classification, association rule mining, clustering and so on. Data mining is really a complex topic and has its association with multiple core fields such as computer science and adds special value to rich seminal computational techniques ranging from statistics, information retrieval, machine learning and pattern recognition. Data mining techniques are the result of a continual process of research and product development. Historical data depicts about the beginning of the business data being first stored on computers with continued improvements in data access, and more recently being used technologies allows users to navigate through their data in real time. Data mining takes this evolutionary process beyond the data access and navigation to proactive information delivery. Data mining is ready for application in the business community due to the support of three technologies that are now sufficiently mature, i.e., peta bytes of data collection, Powerful and effective multiprocessor computers and data mining algorithms. Data mining techniques can be implemented on already existing software and hardware platforms to enhance the value of existing information resources. Upon implementation on high performance client/server or parallel processing computers, data mining tools can analyse larger databases to deliver answers to lots of organizational queries.

The concept clustering is a challenging task for researchers in data exploration and raises many research challenges. The clustering is a challenging and difficult task because of the following reasons. Every researcher has to address these tasks as specified above and he should have through knowledge in order to handle these tasks. (a) in hierarchy of clusters we may get multiple clustering, (b) there is a chance getting many shapes, (c) we may get many data points (d) many features we may get, (e) we may not get well separated clusters, (f) the obtained data points may belong to multiple clusters, (g) outliers may present (h) insufficient of data or data is inadequate or missing. The algorithms designed in this area may solve some problems but no algorithm is there to solve all the problems. Now a day's lot of commercial databases are growing at infinite rates. A recent survey carried out by META Group revealed that 40% of respondents are beyond the peta byte level, while 60% expect to be there by second quarter of 2018.Some of the industries like retail, these numbers can be much larger. The real need for improved computational engines can now be utilized in a cost-effective manner with parallel multiprocessor computer technology with distributed scenario. Data mining algorithms includes techniques that have existed for at least few years, but have only recently been implemented as scalable, distributed, mature, reliable and understandable tools that consistently outperform older statistical methods. The remainder of this paper is organized as follows. Section II presents the Review of Literature followed by Section III which provides the details about the Existing Methods and Approaches, Section 4 Provides the Description of algorithm and new algorithm is presented in Section V. The conclusions are given in Section VI.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

II. RELATED WORK

The most traditional clustering approaches use distance or similarity between data points. The general assumption is that the data points which are far away or not in similar are less likely to belong to the same cluster than the data points that are near and similar. Traditionally the main aim of clustering is to divide the data points into groups with minimal within-group distances and maximal between-group distances [3]. Many researchers have different interpretations of dividing the groups based on distance or similarity. These interpretations have led to different kind of traditional clustering algorithms. Centroid-based clustering algorithms interpret the within-group distances as the distance to the centroid and the between-group distance as the distance between the centroid. Distribution-based clustering algorithms additionally differentiate in which direction the distance is measured. The most popular Density-based clustering algorithms group dense regions and consequently grouping data points with low distances connectivity clustering algorithms group two or more data points that are similar. These algorithms have their own pitfalls while functioning on high dimensional data.

Researchers have been working on subspace clustering from past 1990 onwards and it is one of the young areas where we can find several research gaps. Later it was in 1998, some significant studies took place where the researchers aimed to solve some of the issues towards scalability of data, mainly worked towards the extraction of clusters which seems to be hidden inside the subspaces the critical high dimensional data. According to our survey there are two distinct categories namely cell based subspace clustering and density based subspace clustering. Cell based subspace clustering partitions the data space for efficient detection of dense grid cells with the bottom up fashion, whereas density based subspace clustering generally represents clusters as dense areas separated by sparsely populated areas. Researchers' contributions are discussed in the following section. Agarwal et al., proposed the method which addresses subspace clustering problem by introducing CLIQUE algorithm for identifying the subspace of the dense cluster. The results of CLIQUE (Clustering in Quest) was compared with already existing clustering algorithm BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) and DBSCAN (Density Based Spatial Clustering of Applications with Noise) to visualize better scalability with increase of dimensionality of the data. On the similar work the next significant research was seen in the year 2004. The author Baumgartner et al., proposed the technique of k-nearest neighbor approach with the technique to explore the interesting cluster behavioral traits within the subspace and named as SURFING (Subspace Relevant for Clustering). Their algorithmic design is considered as parameter less and adopted ranking mechanism to explore the best cluster with the subspaces. Their proposed algorithm was compared with previous algorithm CLIQUE on gene dataset and the obtained results were found to outperform CLIQUE with respect to data quality.The authors Gan and Wu proposed an algorithm SUBCAD for minimizing the objective function required for clustering. The authors have used the method of separation and compactness for representing the subspace of every cluster using iterative methods and then experimented over Wisconsin breast cancer data, soybean data, and congressional voting data.

Almost the same related problem was addressed in the research work carried out by Kailing et al., in same year of 2004. The authors have represented an algorithm with a new name and called it as SUBCLU which means Subspace Clustering using density-based approach. Their work had adopted the architecture of DBSCAN with density-connectivity features for randomly identifying location and shape of the clusters within a subspace. SUBCLU was then compared with CLIQUE on same said dataset and the experimental results were found to be moderately efficient.

Fewer studies were carried out during the end of year of 2004 up to 2008 and hence there was a gap in the milestone of subspace clustering. In 2009, another significant research work has been surfaced by Muller et al., focusing on almost similar problem as discussed above.

MINECLUS is a cell-based approach whereas PROCLUS is clustering based approach. Their studies also found that existing technique of SUBCLUE and CLIQUE is less compatible with high-dimensional data. In the year 2010,

Sembiring et al., introduced the projected clustering approach where subspace clustering enumerated clusters of objects in all subspaces of a dataset. Proposed method tends to produce many overlapping clusters.

In their study they discussed existing projected and subspace clustering algorithms, experimented and analyzed three clustering algorithms namely PROCLUS,P3C and STATPC and found out performance of PROCLUS is better in terms of time of calculation with the production of least number of unstructured data whereas STATPC outperforms PROCLUS and P3C w.r.t cluster points focusing on relevant attributes.

The experiments were carried out using weka tool. Researchers have left the scope towards the study and analysis of cell based and density based approaches for larger datasets.

Yet another significant study was introduced by Tatu et al., in 2012 by presenting an algorithm called as ClustMails. Their completely new and unique approach as compared to all the works carried out before 2012 depends on adoption of visual analysis approach using Weka.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

The outcome of their study was compared with VISA (Visualization framework). Designed by Assent et al.,. It is purely the visualization user interface which helped to carry out the experimental analysis. The lacunas of past UI had a new development with a better version of visualization based approach contributed through the work of Liu et al., in 2014. Their study uses subspace clustering without any consideration of single manifold data. Said 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.

The topic of subspace clustering is still under the eye of researchers. There are 12 significant research works which have been witnessed in the year 2015 alone. Chakra borty and Roy have adopted k–means clustering technique along with implementation of fuzzy clustering approach.

Their study was implemented on Matlab and was compared with the traditional k–means clustering.Later on Chang et al., have carried some investigation on spectral clustering technique and proposed an optimization technique on convex formulations. Their study was experimented over JAFFE dataset, UMIST face dataset, Bin Alpha dataset, USPS dataset, and YaleB dataset.

Li et al., contributed towards the adoption of Gaussian regression technique carrying out clustering in high dimensional data sets considering the noise aspects. Their study experiments were carried out over Hopkins 155 dataset and AR dataset (along with Yale and MINST dataset) to prove positive effect of grouping on clusters.

Peng et al., adopted the technique of thresholding based on ridge regression method. Segmentation based approach was seen with the work of Wang and Fu where the authors have contributed towards subspace clustering based on sparsity factor.

Petu-Khov and Kozlov et al.,contributed a greedy approach on subspace clustering for partial data. Wei et al.,[29] had investigated various segmentation techniques of subspace clustering. Sparsity on subspace clustering is emphasized in the research work carried out by Wang et al.,inspired by the past method one more unique research work was proposed by Wang and Zhu, here the authors have contributed an algorithm for noise-free as well as noisy data with a property embedded as self-expressive property with a Bayesian framework using Dirichlet process in principal component analysis. Authors have supported their technique with the help of motion segmentation. A Survey conducted byYang et al., and he presented a technique for clustering of subspaces, in their work information of knowledge is monitored for mechanizing convex optimization problem and the experiments conducted was much better than the previous approaches. K-nearest neighbor technique called as hubness was carried out by Tomasev et al., which emphasized the usage of significant subspace clustering characteristics. One more researcher Yin et al, proposed the multi-view clustering approach for sparse subspace representation.

III. EXISTING METHODS

Subspace clustering approaches can be grouped into statistical, algebraic, iterative and spectral based clustering methods. The detailed descriptions of the methods are provided below

i. Statistical

Mixtures of Probabilistic Principal Component Analysis (MPPCA), Multistage Learning (MSL) adopts Guassian approach towards distribution of data inside each subspace using Expectation Maximization (EM) algorithm. The main disadvantage of this method is they are sensitive to initialization parameter and they need to know the number and dimensions of subspaces. Robust statistical approaches Random Sample Consensus (RANSAC) [20], tries to accept a subspace of dimension d to arbitrarily chosen subsets of d points until the number of inliers is large enough. The inliers are then discarded, and the process is repeated to find a second subspace, and so on. RANSAC will be able to deal with noise and outliers, and does not need to know the number of subspaces. Here, the dimensions of the subspaces must be known. The time complexity of the algorithm increases exponentially in the dimension of the subspaces.

Agglomerative Lossy Compression (ALC) is an theoretic information related statistical method which checks for the segmentation of the data that helps in minimizing the coding length needed to fit the points with a combination of degenerate Gaussians up to a given level of distortion. Due to NP-hard situation, a suboptimal solution is found by first assuming that each point forms its own group, and then iteratively merging pairs of groups to reduce the coding length. This approach handles noise and outliers in the data. The number of subspaces determined by the algorithms is dependent on the choice of a distortion parameter, also there is a lack of theoretical proof for the optimality of the agglomerative algorithm.

ii. Spectral Clustering methods

Popular methods under this category are Local spectral clustering approach with the methods such as Local Subspace Affinity (LSA), Locally Linear Manifold clustering (LLMC), and Spectral Local Best-fit Flats (SLBF) which uses local information around each point to build a similarity between pairs of points. The division of the data is then obtained by applying spectral clustering to the similarity matrix. Related methods have disadvantages

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

in dealing with points which are near the intersection of two subspaces, because the neighborhood of a point can contain points from different subspaces. Spectral clustering is sensitive to the right choice of the neighborhood size to compute the local information at each point. Global spectral clustering approach tries to resolve the issues by building better similarities between data points using global information. The complexity of building the multi-way similarity tends to grow exponentially with the dimensions of the subspaces, practically; a sampling strategy is employed to reduce the computational cost. Advances in sparse and low-rank recovery algorithms, Sparse Subspace Clustering (SSC), Low-Rank Recovery (LRR), Low-Rank Subspace Clustering (LRSC) algorithms obtains the clustering problem as one of finding of sparse or low-rank representation of the data in the dictionary of the data itself, global optimization algorithm is then used to build a similarity graph from which the segmentation of the data is obtained. The advantages of these methods with respect to most of the state-of-the-art algorithms are that they can handle noise and outliers in data, and that they do not need to know the dimensions and, in principle, the number of subspaces a priori.

iii. Iterative

Iterative approaches are K-subspaces [12][13] and median K-flats [14] which alternates between assigning points to subspaces and fitting a subspace to each cluster. The major disadvantages of such approaches are they generally require know the number and dimensions of the subspaces, and they are sensitive to initialization parameter.

iv. Algebraic

Algebraic approaches are the factorization techniques [15] which find an early segmentation by thresholding the entries of a similarity matrix built from the factorization of the data matrix. These methods are likely to be correct when the subspaces are independent, but fails when the assumption is taken in to consideration. This method is sensitive to noise and outliers in the data. Commonly Generalized Principal Component Analysis (GPCA) fits the data with a polynomial type whose gradient at a point gives the normal vector to the subspace containing that point. Generalized PCA deals with subspaces of different dimensions, also it is more sensitive to noise and outliers and its complexity increases exponentially in terms of the number and dimensions of subspaces.

IV. DETAILS OF ALGORITHMS

A. Sparse Subspace Clustering

As defined in the previous section the self-expressiveness property advises that if a data point is linearly represented with less number of data points as possible, then these data points lie in the same subspace. Hence, if we minimize the number of data points used for each reconstruction, i.e. the l0-norm, gives a strong indication on which data points belongs in the same subspace and in the same cluster. However, an l0 minimization is a combinatorial problem and thus NP-hard. The algorithm sparse subspace clustering was created by Elhamifar and Vidal [21]. This sparse subspace clustering algorithm uses an l1-norm instead of a l0 norm. The l1-norm is the tightest convex relaxation of the l0-norm and is known to have similar sparse solutions [13]. Consequently, the sparse results can be computed efficiently. In this algorithm 4.1 minimizing the reconstruction error where the noisy data present simultaneously and minimizing the sum of the coefficients i.e. the l1-norm and this is also known as least absolute shrinkage and selection operator or popularly known as lasso. The Lasso can be calculated by using convex optimization and hence finding minimalcoefficients c* for each data point Yi becomes feasible:

Input: A set of points $\{y_i\}_{i=1}^{N}$ lying in a union of n linear subspaces $\{S_i\}_{i=1}^{n}$

1. Solve the sparse optimization program in Equation 1.6.
2. Normalize the columns of C as $c_i \leftarrow \frac{c_i}{\|c_i\|_\infty}$
3. Form a similarity graph with N nodes representing the data points. Set the weights on the edges between the nodes by $W = |C| + |C|^T$
4. Apply spectral clustering to the similarity graph

Output: Segmentation of the data: $Y_1, Y_2, \ldots, Y_n$

Algorithm 4.1: Sparse Subspace clustering

Together the coefficients c* create a square matrix C that defines how each data point is expressed as a linear combination of others (Y= YC). Before using this coefficient matrix as an affinity matrix in spectral clustering it is adapted in the following way, be invariant of the norm of the data point, the coefficients are normalized by such ci ← c i / ‖ci‖∞ that the affinity matrix is not dominated by the data points that are furthest from the origin. Then the affinity matrix W is constructed by making the normalized coefficients symmetrical W= | C | + | C | T.

In non-zero components c* for the data points which are both close and in the same subspace because the coefficients C can be used in affinity matrix. Therefore, C is the weighting of a connectivity graph which connects every data point to the other nearby data points in the same subspace.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

At first we need to compute the Eigenvectors and the items of these eigenvectors are to be clustered by using the popular k-means algorithm. In the work "Sparse subspace clustering: algorithm, theory, and applications "Elhamifar and Vidal [13] showed that SSC can certainly successfully be used for subspace clustering by using the l1 norm and the results in sparse coefficients can be divided into groups using spectral clustering.
In order to verify this theoretical result in reality, Elhamifar and Vidal [15] created linear subspaces with different cosine similarities and different number of data points per subspace. The results are shown in Figure 4.1.

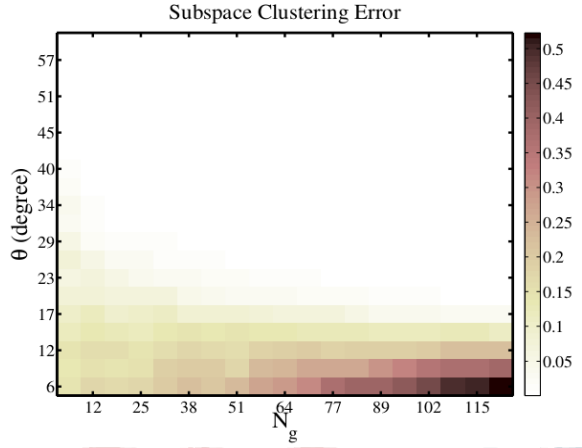

Figure 4.1: Clustering error for different numbers of data points per subspace and cosine angles (low degree is high similarity) between the subspaces. Clustering is not successful when either the number of data points per subspace is to low or cosine similarity is too high .

The complexity of SSC is O (tN3D), where t is the number of iterations used in the minimization, N the number of data points and D the number of features. On real data sets SSC performs well.

Sl no	Data Set	Clustering Error	Median
1	Hopkins-155	2:18%	0:00%
2	Extended Yale B dataset	4:31%	2:50%

Table 4.1: SSC State-of-the-art performance

The above results show that SSC is an algorithm that has state-of-the-art performance. However, SSC is not suitable for larger data sets because its complexity is cubic in the number of data points: O (tN3D). The following section discusses a method that improves the complexity of SSC.

B. Scalable Sparse Subspace Clustering Algorithm

The complexity of algorithm (SSC) is improved by Scalable sparse subspace clustering (SSSC) algorithm which is specified in [16].

In SSC computing a linear representation for all data points over all data points, whereas SSSC uses a subset and this subset is clustered with SSC and then the model is used to cluster the other data points.

Input: A set of points $\{y_i\}_{i=1}^{N}$ from a union of n subspaces $\{S_i\}_{i=1}^{n}$
Input: The ridge regression parameter λ
1. Select p data points from Y denotes by $X = (x_1, x_2, \ldots, x_p)$
2. Perform SSC over X
3. Calculate the linear representation of out-of-sample data $\bar{X}$ over X by $C_i^* = (X \top X + \lambda I)^{\cdot} X \top \bar{X}$
4. Calculate the normalized residual of $\bar{x}^i \in \bar{X}$ over all classes by $r_j(\bar{x}_i) = \frac{\|\bar{x}_i - x\delta_j(c_i^*)\|_2}{\|\delta_j(c_i^*)\|_2}$)
5. Assign $\bar{x}_i$ to the class which produces the minimal residual by identity$(\bar{x}_i) = \text{argmin}_j r_j(\bar{x}_i)$

Output: Segmentation of the data: $Y_1, Y_2, \ldots, Y_n$

Algorithm 4.2: Scalable Sparse Subspace Clustering

We can call this approach as an "out-of-sample approach", because, part of the data is not in the sample that is used to build a first model. In this approach data is treated as new data that arrives after the initial model building. In order to perform this, SSSC takes a fixed number of data points from the data set and applies SSC to it. At least di points from each subspace Si are required to get a block sparse coefficient matrix. The out-of-sample data is considered after the group labels for each data point are known. Regularized linear regression points are created by a linear representation of all the out-of-sample data points over the in-sample data points. This type of second optimization has a much lower complexity than the default SSC optimization. Here only the in-sample data points are used as a dictionary and but not all the data points.

C. Sparse Modeling Representative Selection

As discussed in the above sections a variation of SSC called sparse modeling representatives selection (SMRS) is an algorithm used for selecting representatives, in other words SMRS are also called examples from a dataset. The Representatives of a summary of the dataset, i.e. every data point is similarity one of the representatives. In data discovery and data reduction the concept of representatives are very much useful. In reality SMRS is certainly able to select representatives from complex data. Using video frames as data points, [20] showed that SMRS selected one or more representatives from each scene of the video.The results from [20] show that SMRS is successful at selecting

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 6, June 2018

representatives. However, its complexity is just as large as SSC: O (tN3D).

V. PROPOSED APPROACH

In this paper we have proposed an hierarchical sparse representative (HSR) algorithm which splits the computation of SMRS into parts which are already defined in Algorithm 4. With the proposed algorithm as shown in table 5 instead of splitting the computation of SMRS on the whole dataset, it is separately applied on two or more parts of the dataset. The representatives from the parts are added together. However the process can be repeated with only the found representatives, hence a hierarchical divide-and-conquer strategy is applied.

In HSR algorithm in order to achieve the required results, we require two parameters: the maximum number of representatives called Nrep and the branching factor h. For each recursion, the SMRS is applied on each of the h parts. The representatives from each of the parts are combined and HSR and is applied again if there are more representatives than Nrep. Using more parts reduces the computation load since SMRS is applied on a smaller dataset and also, increasing the maximum number of representatives reduces the computational load because HSR keeps applying SMRS until there are fewer representatives than the maximum.

For the sake of empirical tests however we used only one recursion, Hence SMRS was applied on the parts and then applied once more on the representatives. In this way the results were very similar to the ones obtained by SMRS. Consequently, the parameter Nrep was not used. Furthermore, for all experiments in this research the branching factor will be h=2.

Input: A set of data points $Y = \{y_i\}_{i=1}^{N}$ from a union of n subspaces $\{S_i\}_{i=1}^{n}$
Input: Maximum number of representatives N_{rep}
Input: Branching factor **h**
$r^{out} = \{1,2, \ldots, N\}$
while length (r^{out}) $>N_{rep}$ **do**
  Randomly divide the dataset Y into h parts: $Y_1, Y_2, \ldots, Y_h$
  $r = SMRS(Y_i)$
  $r^{out} = \{ r^{out}_{ri} \mid ri \in r \}$
  $Y = \{Y_{ri} \mid ri \in r \}$
End while
Output: Representatives : $r_1, r_2, \ldots, r_k$ with $k<N_{rep}$

Algorithm 5: Hierarchical Sparse Subspace Clustering

VI. CONCLUSION

Bridging gap between software and technology development, and social need is a great challenge. One of the main obstacles is the underlying metaphor of delivery and assumption of social scientist, potential of presenting relevant ideas and empirical analysis to meet the expectation that need to many practical insights. Complexity has become ubiquitous part of modern life. Complexity is the heart of many systems, from those that send astronauts into space to office desktop computers. The technologies that surround us embody complexity in both there form and function. While, to certain degree, these complexities cannot be avoided, it can be better managed, and must be if we are to develop systems that will allow users to have high levels of situation awareness when working with these systems. The journey into the world of algorithms begins with some preparation and back ground information. A typical algorithm takes a system from one state to another, possibly transitioning through a series of intermediate stage along the way. The presented paper provided a group of algorithms meant for subspace clustering with existing methods and a new approach was introduced. Algorithms shall be tested with the chosen application that is scalable in nature. Research work is attempted to conduct experiments with bench mark datasets which are readily available in the form of UCI machine learning repository over internet. Synthetic and real time data sets shall be tested and comparison will be carried out for experiments.

REFERENCES

[1] Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in High dimensional space. In Proceedings of the 8th International Conference on Database Theory, pages 420{434, Berlin. Springer.

[2] Aggarwal, C. C. and Reddy, C. K. (2013). Data clustering: algorithms and applications. CRC Press.

[3] Basri, R. and Jacobs, D. (2003). Lambertian reectances and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence

[4] Tomasi, C. and Kanade, T. (1992). Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision.

[5] Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: algorithm, theory, and applications. IEEE transactions on pattern analysis and machine intelligence.

[6] Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: algorithm, theory,and applications. IEEE transactions on pattern analysis and machineintelligence.

[7] Vidal, R. (2011). A tutorial on subspace clustering. IEEE Signal Processing Magazine.

[8] Costeira, J. and Kanade, T. (1998). A multibody factorization method for independently moving objects. International Journal of Computer Vision.

[9] Vidal, R., Ma, Y., and Sastry, S. (2005). Generalized principal component analysis (GPCA). IEEE transactions on pattern analysis and machine intelligence.

[10] Mustafa, N. H. (2004). k-Means Projective Clustering. In Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems.

[11] Fischler, M. a. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM.

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

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

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

[15] 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.

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

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

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

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

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

[21] 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.

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

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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

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

[32] 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.

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

[34] Chen, J. and Yang, J. (2014). Robust subspace segmentation via low-rank representation.IEEE Transactions on Cybernetics.

[35] Elhamifar, E. and Vidal, R. (2009). Sparse subspace clustering. In IEEE Conference on Computer Vision and Pattern Recognition, pages. IEEE.

[36] Elhamifar, E. and Vidal, R. (2010). Clustering disjoint subspaces via sparse representation. In IEEE International Conference on Acoustics Speech and Signal Processing, pages 1926{1929.

[37] Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: algorithm, theory, and applications. IEEE transactions on pattern analysis and machine intelligence.

[38] Donoho, D. L. (2006). For most large underdetermined systems of equations, the minimal L1-norm near-solution approximates the sparsest near-solution. Communications on Pure and Applied Mathematics,

[39] Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: algorithm, theory, and applications. IEEE transactions on pattern analysis and machine intelligence, 35(11):2765{81.

[40] Peng, X., Zhang, L., and Yi, Z. (2013a). Inductive sparse subspace clustering. Electronics Letters.

[41] Elhamifar, E., Sapiro, G., and Vidal, R. (2012). See all by looking at a few:Sparse modeling for finding representative objects. In Proceedings of the IEEEComputer Society Conference on Computer Vision and Pattern Recognition,pages 1600{1607.