RMSC: Robust Modeling of Subspace Clustering for 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*— Subspace clustering is one of the active research problem associated with high-dimensional data. Here some of the standard techniques are reviewed to investigate existing methodologies. Although, there have been various forms of research techniques evolved recently, they do not completely mitigate the problems pertaining to noise sustainability and optimization of clustering accuracy. Hence, a novel technique called as Robust Modeling of Subspace Clustering (RMSC) presented to solve the above problem. An analytical research methodology is used to formulate two algorithms for computing outliers and for extracting elite subspace from the high-dimensional data inflicted by different forms of noise. RMSC was found to offer higher accuracy and lower error rate both in presence of noise and absence of noise over high-dimensional data.**

Keywords*— *Accuracy, High Dimensional data, Noise. Grouping, Performance, Subspace Clustering.

I. INTRODUCTION (*HEADING 1*)

With the evolution of pervasive and ubiquitous computing, there is a growth in the generation of data from every part of the world. Such form of growing size of data is impossible to be stored in physical servers and hence cloud-based storage is the appropriate place to reside [1]. The unique applications conceptualized using the technologies like Artificial intelligence and Machine Learning to the specific domains of Bioinformatics, Computer Vision, Signal Processing, etc., is pervasive to the high dimension data [2]. The applications of subspace clustering of high dimension data generally includes a) Retrieving Feature of a rigidly moving object in a video b) Identifying face images under varying illumination c) Multiple instances of a hand-written digit. d) Image representation and compression e) Image segmentation, Motion segmentation and temporal video segmentation. In all the above applications, Subspace clustering denotes the problem of separating data according to their basic subspaces The existing methods/ algorithms are categorized into four types namely Iterative, Algebraic, Statistical, and Spectral. Such forms of data are too much bigger in size and hence are called as high-dimensional data. Unfortunately, in such form of high-dimensional data, the distance and locations of the data points are more disperse and scattered in order to take the shape of sparsity [3][4][5]. Therefore, clustering is the only alternative technique to solve such classification problem. This problem becomes much worst when there is displacement of any data point between two different spaces of higher dimensional data [6][7]. The dimensionality varies as there are large pixel quantities in images, huge number of frames in videos and very-very high diversified attributes associated with web logs and text document. The processing of these high dimensioned data poses huge requirements of computational time complexities as well as memory usage. Additionally, inadequate sample and presence of noise leads to falsifications of analysis due to dimensionality. The hardness of the process can be normalized by exploring the low-dimension structures namely subspace.

Another bigger problem is that such forms of massive data are often found to possess significant amount of noises [8]. Although, there has been various techniques for noise-related problems in high-dimensional data [9], but a closer look into such work will show that enough importance has not been laid towards minimizing algorithm complexities. Presence of any form of noise will potentially affect the clustering accuracy.

This is one of the toughest research challenges as it is quite challenging to discretize real noise from certain legitimate numerical value. Moreover search-based optimization will fail as forming of fitness function is not feasible if the location of cluster is unknown.

Therefore, this paper presents a simple technique that performs robust and cost-effective modelling of subspace clustering. In section 2 we provide an overview of related work which done in this area. Section 3 highlights about the problem identification followed by brief highlights of proposed research methodology in Section 4. Section 5 outlines the algorithm implementation followed by result analysis in section 6. Finally, summary of the paper is briefed in Section 7.

II. RELATED WORK

This section discusses about the prior studies towards subspace clustering over high dimensional data. Our earlier review work [10] has already discussed about various techniques and approaches towards similar issues as well as highlighted some of the open research issues. We have also presented a sub-clustering technique over document form in high-dimensional data [11].

The most recent work carried out by Zhai et al. [12] has presented a unique sparse sub-clustering technique for addressing the constraint of local averaging in hyperspectral images. The technique introduces a scalable l1-normalization approach that uses conventional technique adopting data related to spatial spectrum.. Hence, its applicability is

978-1-5090-6367-3/17/$31.00 ©2017 IEEE

restricted to controlled research environment itself. Peng et al. [13] have addressed the time complexity associated with the spectral clustering for the purpose of obtaining similarity graph. The technique uses an integrated clustering mechanism for varied forms of data and outliers. The problem pertaining to grouping followed by clustering of subspaces is investigated by He et al. [14]. A different form of conceptualized objective function has been presented along with half quadratic optimization. The technique has used Parzen window estimation for managing the outliers. Li et al. [15] have investigated towards low rank-based approach in subspace clustering that is capable of extracting the impartial data from low to high dimensional spaces.

that may pose a significant problem in inferring the outcome of some new feature in view of domain context.

An optimization technique is presented where data points play a crucial role in linear combination of sparse matrix. Similar type of work is also carried out by Fang et al. [16] towards low rank representation. A semi-supervised algorithm is implemented and presented to carry out subspace clustering using harmonic function and Gaussian field.

Hence, it can be seen that there are various technique towards subspace clustering in present time. Brief highlights of some more approaches and research techniques are tabulated in Table I.

TABLE I. SUMMARY OF EXISTING TECHNIQUES

Methods	Work and Reference	Limitations
Iterative	K-subspace[17] Predictive [18]	The number and dimension of the subspace should be known, Sensitive to initialization
Algebraic	Factorization [19][20][21] Geometric [22]	Fails when assumption of independent subspace is violated. Sensitive to noise and Outlier into the data Complexity is exponential to the number of subspace dimension.
Statistical	Principal Component Analysis[23] Multistage learning[24][25] Robust Statistical Approach [26] agglomerative lossy compression [27]	The number and dimensions of the subspaces should be known Sensitive to initialization. Dimension of subspace need to be known and equal Complexity is exponential to the number of subspace dimension There is No Theoretical proof for the optimality of the agglomerative algorithm.
Spectral	locality-constrained linear coding [28], Spectral curvature clustering [29] Sampling, Sparse & Low Rank : [30][31]	Difficult to deal near the intersection Sensitive to right choice of neighbourhood. Dimension of subspace need to be known and equal Complexity is exponential to the number of subspace dimension they can handle noise and outliers in data They do not need to know the dimension (No of sub-spaces a priori)

III. PROBLEM IDENTIFICATION

From the prior section, it can be seen that majority of the existing techniques for solving problems pertaining to subspace clustering are not without flaws. Apart from there, the other frequently used techniques of subspace clustering are i) projection-based, ii) grid-based, iii) pattern based, and iv)bipartite based. The projection-based clustering techniques are responsible for generating subspace clusters created from partitioning of dataset.

Pattern-based techniques are responsible for exploring specific forms of pattern of interest depending on certain specific fitness condition. All the techniques related to transforming features primarily try to overview the dataset in the form of minimal size of dimension. This is carried out by constructing various permutations of original parameters. Such techniques are found to have good result for exploring the hidden pattern within the dataset. Unfortunately, all these techniques make use of recording the distances among the objects. This makes the existing subspace clustering technique less effective especially in the case of presence of maximal irrelevant parameters (which are obviously hidden within certain subspaces owing to noises). It was also seen that existing techniques make use of integration of original feature

On the other hand, the mechanism of selection of feature chooses only the appropriate dimension in order to expose the cluster of objects corresponding to the subset of their parameters. The problems pertaining to subspace clustering is still an open ended and calls for serious investigation. The next section briefs a technique for enhancing subspace clustering accuracy.

IV. PROPOSED METHODOLOGY

Data points are collected from the high-dimensional data for performing subspace clustering along with its associated dimensions. The prime research goal is implement a novel sparse subspace clustering technique to make it resilient against adverse condition of noise. The technique also uses threshold for inliers in order to ensure minimal error occurrences during subspace clustering.

We introduce two different algorithms focusing on outlier computation that has the capability to deal with different forms of noises as well as perform extraction of elite group (or subspace). Two different test-parameters of noise were used for testing the sustainability. We associate robustness in

clustering design as it can enable the process to withstand higher accuracy.

The design and development of the proposed system is carried out using analytical research methodology. The major purpose of the proposed system is to offer increased clustering accuracy for a high-dimensional data shown in Fig. 1.

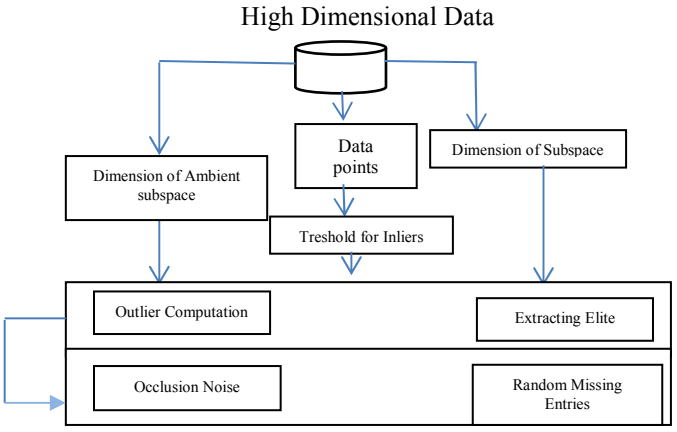

Fig. 1 Architecture of the Proposed RSCM

V. ALGORITHM IMPLEMENTATION

The proposed system is basically implemented on the basis of two simple algorithm i.e. i) Algorithm for outlier computation in subspace clustering and ii) Algorithm for extracting elite group in subspace clustering. Basically the algorithm targets to minimize error that assists in accomplishing higher accuracy rate. The next section outlines the algorithm implemented for achieving the research goal.

A. Algorithm for outlier computation in subspace clustering

This algorithm is responsible for computing the amount of outliers that may be present in one cluster (or group). The algorithm takes the input of S (Data points), d (dimension of subspace to find), g (number of subspace), t (threshold), which after processing yields an output of ol (outlier). The basic steps of the algorithm are as follows:

Algorithm for outlier computation in subspace Clustering

Input: S (Data points), d (dimension of subspace to find), g (number of subspace), t (threshold),

Output: ol (outlier)

Start

1. Init S, d, g, t
2. **for** (i=1:g)
3. **if** (size(ol)<4)
4. **for** (j=1:i-1)
5. $\delta=I(S)-b.b^{-1}$
6. $dist=\sum(\delta.S)^2$
7. **end**
8. $min\_dist=arg_{min}(dist)$
9. [v, o]sort(min_dist)
10. olorder(1:4)
11. **end**

end

The algorithm initially computes the size of data points S and saves in map of K by N matrix. The outliers are considered to be between 1 to N while we create a three dimensional matrix b of dimension equivalent to K, d, and certain test value. For all the number of subspaces (Line-2), we check if the outliers are less than 4 as a constraining factor. A recursive function is carried out to check all the sub-spaces (Line-4) in order to compute a matrix δ that is formed by product of an identity matrix I and product of b and b-1 (Line-5). Distance between the subspaces are computed using similar matrix δ and data points as shown in Line-6 of the algorithm. Minimum distance from every iterations of the clustering is recorded in min_dist matrix (Line-7) that is further subjected to sorting (Line-8). Finally order is used for evaluating the total number of outliers. Not only the outliers, the algorithm also assist in retaining the error entropy to as low as possible and this positively affect the accuracy during subspace clustering.

B. Algorithm for Extracting Elite group of Subspace Cluster

This algorithm is mainly responsible for extracting elite group (or subspace) while performing subspace clustering. The input for the algorithm is d1 (dimension of subspace), d2 (dimension of ambient subspace), η (number of point in each group), g (number of group), t (threshold for inliers), S (data point), which after processing generates the outcome of G (elite group). The essential steps of the algorithm are as follows:

Algorithm for Extracting Elite group of Subspace Cluster

Input: d_1 (dimension of subspace), d_2 (dimension of ambient subspace), η (number of point in each group), g (number of group), t (threshold for inliers), S (data point)

Output: G (elite group)

Start

1. init d_1, d_2, η, t
2. $g=arg_{max}(Label)$
3. gr=Algo-1(S, d_1, g, t)
4. G=em(Label, gr)
5. $\theta^- =\sum(Label-G)/length(Label)$
6. $\theta^+=1-\theta^-$

End

The initialization of the algorithm calls for defining the dimensions of the subspaces as well as dimensions of ambient space mainly (Line-1). It also uses data points and threshold. We create a null matrix Stack and Label that is used for stacking the data for transforming multiple variables into one variable for assisting in converging the search space during

subspace clustering in high dimensional data. The total number of subspaces (or groups) is considered to be maximal number of Labels, while we consider an inliers threshold of 0.01 (Line-2). The first algorithm for outlier computation is then applier on the input arguments in order to obtain subspace gr (Line-3). The extracted subspace gr is then subjected to processing for obtaining elite maps em. The method em is responsible for permuting the matrix Label in order to obtained highly optimized subspace. Design of the em method is carried out using a squared cost matrix, optimal assignment, and cost of optimal assignment in such a way that the cost is highly reduced for all feasible assignments of data points. Therefore, a superior version of the subspace is obtained. The final part of the algorithm is to compute missing rate of clustering and thereby find accuracy. We use the matrix Label and finally extracted subspace for computing missing rate of clustering θ- (Line-5) that ultimately assists us to compute the accuracy level θ+ (Line-6). Hence, irrespective of possessing the data points from different number of subspaces in high dimensional data, the algorithm make use of the concept to lower the error in entropy and thereby makes the system robust. This algorithm can help in mining numerical data, which is quite a challenging problem till now. Therefore, a precise extraction of elite subspace is obtained by this algorithm.

VI. RESULT ANALYSIS

The proposed system implements the algorithm on Yale face database that possess approximately 16128 images pertaining to 28 subjects using 64 different conditions of illumination and 9 different poses. For effective analysis, the proposed system RMSC is compared with work done by Song et al. [32] towards sparse subspace clustering. The work carried out by Song et al. [32] have used Hermitian positive definite embedded over Hilbert space in order to develop a sparse subspace clustering. We do fine tune the work of Song et al. by implementing the same dataset of Yale on our performance parameter of accuracy and error. The obtained outcome is tested in presence of most frequently used occlusion and random noise in subspace clustering.

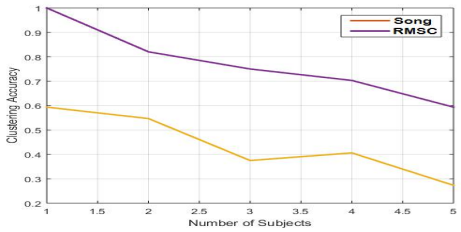

Fig. 2 Analysis of Clustering Accuracy with no Occlusion

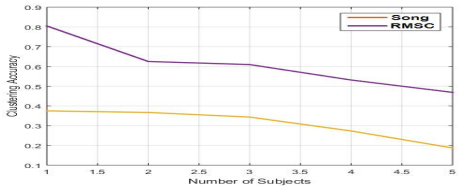

Fig. 3 Analysis of Clustering Accuracy with Occlusion

Fig. 2 and Fig. 3 shows that the proposed system offers significantly good level of clustering accuracy both in absence of occlusion noise as well as in presence of 20% of occlusion noise. Hence, the proposed system can be easily used in high-dimensional data where extent of occlusion noise is always there. Most importantly, the level of noise doesn't really effects the accuracy level of proposed RMSC. The main reason for this outcome is that Song et al. [32] have used an integrated technique of classification as well as weighted joint coding in order to obtain better contextual data from subspaces. However, it leads to minimization of iterative speed for which reason the search space cannot be much optimized over the progressive subspaces. This leads to lowered clustering accuracy as compared to proposed RMSC where we use a technique that fundamentally increases iterative speed along with identification of outliers. This phenomenon leads to increase in accuracy level.

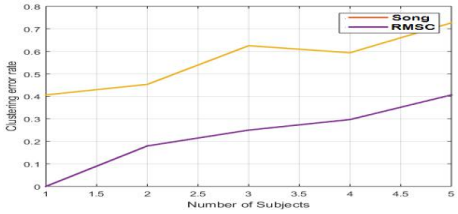

Fig. 4 Analysis of Clustering Error Rate with no Occlusion

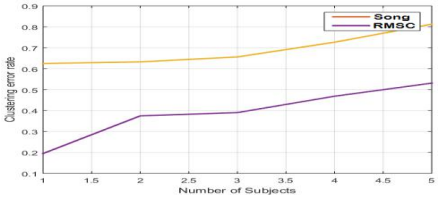

Fig. 5 Analysis of Clustering Error Rate with Occlusion

In order to validate the outcomes, we also check the error rate both in absence of occlusion Fig. 4 and in presence of 20% occlusion Fig. 5. Apart from this, we also compute the time complexity of both algorithms over similar simulation environment. The algorithm processing time of proposed RMSC is 85% faster as compared to work carried out by Song et al. [32]. Hence, proposed system offer cost effective subspace clustering technique for high dimensional data.

VII. CONCLUSION

Performing subspace clustering is not a simple task and its accuracy depends on many intrinsic and extrinsic factors in high dimensional data. After reviewing the existing system associated with subspace clustering, we find that still the problems related to noise and error-based sustainability has to be addressed. Therefore, we formulated a very simple approach where we define two different algorithms for computing outliers as well as for computing the most efficient subspaces from the high dimensional data. In order to testify the effectiveness of proposed system, we apply two different types of noise to see the extent its affects the outcome of clustering accuracy. Interestingly, we find that clustering

accuracy stay unaffected both in presence and absence of noise. Our future work will be to design an integrated technique for performing subspace clustering using subspace learning and subspace clustering approach.

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] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity: The Lasso and Generalizations, CRC Press, 2015

[4] M. Pourahmadi, High-Dimensional Covariance Estimation: With High-Dimensional Data, John Wiley & Sons, 2013

[5] C. Giraud, Introduction to High-Dimensional Statistics, CRC Press, 2014

[6] F. Aleskerov, B. Goldengorin, P. M. Pardalos, Clusters, Orders, and Trees: Methods and Applications, Springer, 2014

[7] J. Wang, Geometric Structure of High-Dimensional Data and Dimensionality Reduction, Springer Science & Business Media, 2012

[8] J. Aluya, The Influences of Big Data Analytics, Author House, 2014

[9] P. L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing Data: provable guarantees with nonconvexity", The Annals of Statistics, Vol. 40, No. 3, 1637–1664, 2012

[10] Radhika K R, Pushpa C N, Thriveni J, Venugopal K R, "Insights to Existing Techniques of Subspace Clustering in High-Dimensional Data",International Journal of Scientific & Engineering Research, Volume 6, Issue 11, November-2015

[11] Radhika, K.R. and Pushpa, C.N. and Thriveni, J. and Venugopal, K.R. "EDSC: Efficient document subspace clustering technique for high-dimensional data", International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), 11-13 March 2016, Bangalore.

[12] H. Zhai, H. Zhang, L. Zhang, P. Li and A. Plaza, "A New Sparse Subspace Clustering Algorithm for Hyperspectral Remote Sensing Imagery," in IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 1, pp. 43-47, Jan. 2017.

[13] X. Peng, H. Tang, L. Zhang, Z. Yi and S. Xiao, "A Unified Framework for Representation-Based Subspace Clustering of Out-of-Sample and Large-Scale Data," in IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 12, pp. 2499-2512, Dec. 2016.

[14] R. He, L. Wang, Z. Sun, Y. Zhang and B. Li, "Information Theoretic Subspace Clustering," in IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 12, pp. 2643-2655, Dec. 2016.

[15] C. G. Li and R. Vidal, "A Structured Sparse Plus Structured Low-Rank Framework for Subspace Clustering and Completion," in IEEE Transactions on Signal Processing, vol. 64, no. 24, pp. 6557-6570, Dec.15, 15 2016.

[16] X. Fang, Y. Xu, X. Li, Z. Lai and W. K. Wong, "Robust Semi-Supervised Subspace Clustering via Non-Negative Low-Rank Representation," in IEEE Transactions on Cybernetics, vol. 46, no. 8, pp. 1828-1838, Aug. 2016.

[17] E. C. Ozan, S. Kiranyaz and M. Gabbouj, "K-Subspaces Quantization for Approximate Nearest Neighbor Search," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1722-1733, July 1 2016.

[18] B. McWilliams and G. Montana, "Subspace clustering of high-dimensional data: a predictive approach", Data Mining and Knowledge Discovery, Vol. 28, No. 3, pp. 736-772, 2014

[19] B. Wang, R. Liu, C. Lin and X. Fan, "Matrix Factorization with Column L0-Norm Constraint for Robust Multi-subspace Analysis," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, 2015, pp. 1189-1195.

[20] X. Li, W. Wang, A. Razi and T. Li, "Nonconvex Low-Rank Sparse Factorization for Image Segmentation," 2015 11th International Conference on Computational Intelligence and Security (CIS), Shenzhen, 2015, pp. 227-230.

[21] Shulin Wang, Fang Chen and Jianwen Fang, "Spectral clustering of high-dimensional data via Nonnegative Matrix Factorization," 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, 2015, pp. 1-8.

[22] M. C. Tsakiris and R. Vidal, "Abstract algebraic-geometric subspace clustering," 2014 48th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, 2014, pp. 1321-1325.

[23] Z. Fan et al., "Modified Principal Component Analysis: An Integration of Multiple Similarity Subspace Models," in IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 8, pp. 1538-1552, Aug. 2014.

[24] T. Xu and W. Wang, "Methods for learning adaptive dictionary in underdetermined speech separation," 2011 IEEE International Workshop on Machine Learning for Signal Processing, Santander, 2011, pp. 1-6.

[25] J. Zhang, C. G. Li, H. Zhang and J. Guo, "Low-rank and structured sparse subspace clustering," 2016 Visual Communications and Image Processing (VCIP), Chengdu, 2016, pp. 1-4.

[26] P. A. Traganitis, K. Slavakis and G. B. Giannakis, "Sketch and Validate for Big Data Clustering," in IEEE Journal of Selected Topics in Signal Processing, vol. 9, no. 4, pp. 678-690, June 2015

[27] E. Elhamifar and R. Vidal, "Sparse Subspace Clustering: Algorithm, Theory, and Applications," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 11, pp. 2765-2781, Nov. 2013.

[28] Peihua Li, Xiaoxiao Lu and Qilong Wang, "From dictionary of visual words to subspaces: Locality-constrained affine subspace coding," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 2348-2357.

[29] D. Park, C. Caramanis, and S. Sanghavi, "Greedy subspace clustering", In Advances in Neural Information Processing Systems, pp. 2753-2761. 2014.

[30] Y. Fu, J. Gao, D. Tien, Z. Lin and X. Hong, "Tensor LRR and Sparse Coding-Based Subspace Clustering," in IEEE Transactions on Neural Networks and Learning Systems, vol. 27, no. 10, pp. 2120-2133, Oct. 2016.

[31] J. Shen and P. Li, and H. Xu, "Online low-rank subspace clustering by basis dictionary pursuit", arXiv preprint arXiv:1503.08356, 2015

[32] H. Song, W. Yang, N. Zhong and X. Xu, "Unsupervised Classification of PolSAR Imagery via Kernel Sparse Subspace Clustering," in IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 10, pp. 1487-1491, Oct. 2016.