

EDSC: Efficient Document Subspace Clustering Technique 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—With the advancement in the pervasive technology, there is a spontaneous rise in the size of the data. Such data are generated from various forms of resources right from individual to organization level. Due to the characteristics of unstructured or semi-structuredness in data representation, the existing data analytics approaches are not directly applicable which leads to curse of dimensionality problem. Hence, this paper presents an Efficient Document Subspace Clustering (EDSC) technique for high-dimensional data that contributes to the existing system with respect to identification by eliminating the redundant data. The discrete segmentation of data points are used to explicitly expose the dimensionality of hidden subspaces in the clusters. The outcome of the proposed system was compared with existing system to find the effective document clustering process for high-dimensional data. The processing time of EDSC for subspace clustering is reduced by 50% as compared to the existing system.

Keywords: Cluster Analysis, High-Dimensional Data, Subspace Clustering.

I. INTRODUCTION

In the present world, various objects can be electronically illustrated with “High Dimensional Data” (HDD) e.g. audio signals, pictures, videos, text documents, fingerprints, and hyper spectral images etc., [1]. The high dimensional linear and generalized linear models are considered as more flexible generalized additive models and together of them cover a very high range of applications viz. data analysis / mining, astronomy, pattern recognition and climatic condition. A conventional clustering technique is the feasible technique to investigate data and it considers entire dimensions of a data for the purpose of maximum utilization of knowledge discovery [2]. Evolution of such data may be from various resources e.g. social network, medical science, education, and a lot of other networking application domains. The process of data analysis is not new and it is the only best and cost effective technique to explore the unique knowledge from the complex data. In the present time, the data is quite a complex and applying existing mining techniques are not feasible.

A major concern is the dimensionality of data [3]. A feature selection technique explores the possibility of dimensional subset for carrying out clustering by eliminating unnecessary and inappropriate dimensions. Subspace clustering is one of such technique that positions its search operation and generates information about the clusters present in multiple subspaces in overlapping conditions [4][5].

Motivation: With the advancement in the pervasive technology, there is a spontaneous rise in the size of the data. Such data are generated from various forms of resources right from individual to organization level. Unfortunately, owing to the inherent characteristics of unstructured or semi-structuredness in data representation, the existing data analytics approach is not directly applicable leading to evolution of curse of dimensionality problem.

Contribution: This paper presents a technique called as Efficient Document Subspace Clustering (EDSC) mechanism for high dimensional data that contributes to the existing system with respect to identification followed by elimination of redundant data. The discrete segmentation of data points are used to explicitly expose the dimensionality of hidden subspaces in the clusters.

Organization: This paper gives the information about the clustering and reviews some of the existing subspace clustering algorithms and dimensionality reduction techniques in Section II. Section III highlights the problem of subspace clustering techniques and Section IV explains the proposed system. Section V, performance analysis is discussed in Section VI followed by Conclusions and Future Work in Section VII

II. RELATED WORK

This section discusses about the recent work being done in the area of subspace clustering of high-dimensional data. Yang et al., [6] proposed two new methodologies for

subspace grouping and completion. The first sums up the inadequate subspace clustering algorithm with the goal that it can get data from observed entries. The second one calculates a suitable portion of a network by expecting an irregular model for the missing entries and gets the sparse representation from this part.

Wang et al., [7] presented an innovative nonparametric Bayesian subspace grouping model that interprets the quantity of subspaces and the measurement of every subspace from the collected information.

Wang et al., [8] has determined the graph-connectivity issue that inconveniences the hypothesis of Sparse Subspace Clustering (SSC). Wei et al., [9] concentrates on the execution of distinctive subspace cluster algorithms by taking care of subspace cluster issues and perspectives. Petukhov et al., [10] introduced the Fast Greedy Sparse Subspace Clustering technique (FGSSC), which is a greedy approach methodology based adjustment of the SSC calculation.

III. PROBLEM IDENTIFICATION

Several algorithms are available for clustering to semi-automate or automate the clustering procedure. Different kinds of algorithms are subjected to database, may produce different clusters with different answers. Each cluster completes procedures by using different algorithm with own complexity, error, resources, run time, frequency etc. The main issue is that the clustering output or results depend on type of database used. The number of clustering problems are stated below.

- **Addressing:** Clustering techniques do not address all the requirements simultaneously (and concurrently).
- **Database:** The output of clusters always depends on type of database used. As dimension and size of database increases, the handling becomes difficult.
- **Time Complexity:** While dealing with the large number of data items, the dimension will be problematic due to time complexity.
- **Effectiveness:** The method effectiveness depends on the definition of clustering.
- **Distance:** Defining is required when measurement of obvious distance doesn't exist. In multi-dimensional spaces defining is always not easy. Distance measures like Manhattan, maximum distance measurement and Euclidian are required for numerical attributes.
- **Number of Clusters:** The cluster number identification is very difficult, if the class label number is not known. A careful cluster number analysis is required or heterogeneous tuples may found. As a result tuples may get broken into many similar or merged tuples.

- **Clustering algorithm:** The clustering algorithm results can be interpreted in many ways.

Working in higher dimensional data leads to lists of critical issues with respect to clustering.

- With the increment in the size of the dimensionality of the data, the volume of the data too increases rendering the existing data to be forcible sparse.
- High-dimensional Data also results in failure to explore the precise clusters during cluster analysis.
- 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.
- There is a higher possibility of correlation of an attributes for the massive quantity of the attributes. This fact may result in formation of the clusters in the random position in the subspaces.

IV. PROPOSED SYSTEM

The prime aim of the proposed study is to present a solution for subspace clustering problem in high dimensional data. The data accomplished from social networking can be considered as high dimensional data, which can be represented by a complex data matrix form for assisting in cluster analysis. The proposed system was evaluated over textual data that is considered as high-dimensional data. Hence, the proposed system is termed as Efficient Document Subspace Clustering (EDSC) for high dimensional data

The considered input of the proposed system is both unstructured and semi-structured type. The proposed system performs the common operation of datamining e.g. preprocessing the noise and elimination of redundant data from the subspace. The data considered as an input is categorized to clusters, which are then arranged in subspace clusters randomly in order to give a shape of high-dimensional data. The system is followed by segmentation of the data points and identification of precise number of clusters present in the study leading to generation of candidate subspace.

The system architecture of the proposed EDSC is shown in Fig.1. The top layer of the architecture is the database of high-dimensional data, which is self-created in the study and is considered as an input for the technique. The second layer is responsible for empirically mapping with the subspace clustering from the input database. The third layer of the proposed study is divided into two blocks, where the

first block processes dimensions using a simple and novel mathematical approach and the second block performs pre-processing followed by segmentation of the input database. Finally, the model is subjected for evaluation.

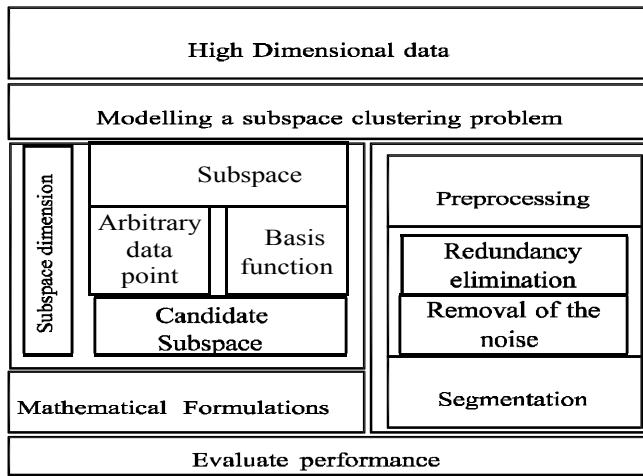


Fig. 1 Efficient Document Subspace Clustering System Architecture

Therefore, the contributions of the proposed EDSC system are as follows e.g. i) modeling a subspace clustering problem, ii) performing preprocessing of high dimensional data, iii) performing segmentation of the clusters, iv) generation of weighing cost to locate the hidden clusters in subspace. Finally, the proposed system applies segmentation of Data Points algorithm for subspace clustering to explore the number of subspaces along with the dimensional information and segmentation of the high dimensional data points.

V. RESEARCH METHODOLOGY

The proposed system is designed using empirical research methodology, which permits to take help of scientific methods for performing modelling of problem followed by exploration of solutions. For simplicity in the DSC model implementation, we consider to represent the subspaces mathematically as then removed from the subspace clusters and hence narrowing of the search space is done iteratively. The identification of the subspace clusters hidden in the high dimensional data is done by segmentation process of the proposed system. Consider a set S containing a data from T₁ to T_n i.e. S={S_k|kεΔ}, where k=1, 2,, m and formulate a data matrix for this. Applying the set theory, the proposed system finds the genuine data points from all the subspaces.

$$SI = \{x \in R : x = \mu_I + U_i Y\} \quad (1)$$

Equation (1) represents subspace in the forms of variable S and x is candidate subspace that is equivalent to real-time R to the power of minimal subspace dimension D.

The variable z is an arbitrary data point existing in the subspace, while Y and U represent low dimension and basis function respectively. The prime intention of the proposed EDSC system is to identify the number of the subspaces along with their dimensions and segmentations of the points. The proposed EDSC system considers the integration of the linear subspaces for extraction of the data points mathematically represented as,

$$\rho_I = 3\gamma I \quad (2)$$

In the Equation (2), ρ_I is the variable that represented the data points in the subspace clusters while γ represents the updated variable representation of data point ρ_I . Hence, the problem is encountered for optimally retain the relationship for the Euclidean data structure. This condition is the exact condition of the high-dimensional data that renders Euclidean distance meaningless. Therefore, the proposed system is more focused on internal geometric structure of the subspaces rather than identifying the Euclidean structure.

$$\varphi = \varphi_1 + \varphi_2 + \varphi_3 + \dots + \varphi_n \quad (3)$$

The Equation (3) shows the mathematical representation of high dimensional data input depicted by variable φ . The sub-variables $\varphi_1, \varphi_2, \dots, \varphi_n$ represents the components of high- dimensional data in proportion of n numbers. The initial task of the proposed system is to perform redundancy elimination along with removal of noise.

$$\varphi_i = \Delta \quad (4)$$

Equation (4) shows the new variable Δ , which represent a particular set of text in one document, and φ_i represents the i th subspace that posses this textual content. A set of condition is written for identifying the missing terms or special characters from the document Δ and then the system performs removal of the noise.

TABLE I. ALGORITHM FOR REDUNDANCY CHECK IN SUBSPACES

Algorithm - 1: Redundancy Check in Subspaces

Begin

Cost C [t₁, t₂, t₃,t_m], mεΔ_i

if C ≥ 1

Remove t_i

else, go for Δ_j, i < j.

End

The system evaluates the cost-factor C responsible for increasing the weighing attribute of the text and performs computation of the occurrences of each cluster. Any data (t₁, t₂, t₃,t_m) that posses lower values of cost will mean that they posses insignificant informative words. Such textual data are

$$G = S_1 \cap S_2 \cap \dots \cap S_k \quad (5)$$

After the segmentation of the data points are carried out, the proposed system recursively performs identification of the redundant data for enhancing the preciseness of cluster positions.

TABLE II. ALGORITHM FOR SEGMENTATION OF DATA POINTS IN SUBSPACE

Algorithm-2: Segmentation of Data points

Begin

1. $S = \{S_k | k \in \Delta\}$, where $k=1, 2, \dots, m$
2. $G = S_1 \cap S_2 \cap \dots \cap S_k$
3. $\rho = \arg_{\max}(C)$ flag
4. for $\rho=1$ to ϕ_n
5. call algorithm for Redundancy check in subspace
6. Estimate $\delta (\Delta_1, \Delta_2, \dots, \Delta_m)$
7. Display δ

End

The Algorithm-2 is an essential part of performing subspace clustering mechanism by the proposed system. This algorithm initially computes the set S residing in the internal structure of the high-dimensional data, where the objective function is to get the genuine or error free data i.e. G . G will represent the data pertaining to the actual subspace that was initially missing owing to curse of dimensionality problem in high-dimensional data. It is now evaluated by performing intersection operation used for flagging the subspace positions for identifying the redundant and non-redundant subspaces. The latter is retained in the clustering process while the redundant data is removed from the data matrix. Hence, we use maximum argument to represent the recursive process. A loop is being formulated from the initial position to maximum position of the subspaces where position is identified from the variable ϕ .

VI. PERFORMANCE ANALYSIS

Every data differs in their nature on the basis of their features, based on their features data are categorized as structure, semi-structure and unstructured data. The result analysis is performed on the High dimensional unstructured data.

Algorithms discussed in this proposed study has dual advantages i.e. i) its processing capability for massive datasets is reduced and ii) the accuracy of the subspace number identification is 98.56% for a massive dataset of 1 TB. The entire process of subspace clustering presented takes around 1-3 seconds to analyze Megabytes of data, 2-5 seconds for processing Gigabytes of data, and 90 seconds to 2 minutes to analyze Terabytes

of data. Hence, the execution time consumed by the proposed DSC system is within the tolerable limits even for mission critical applications. The biggest potential of the algorithm is correct identification of data points and identification of subspaces with their dimensionality data.

The outcome of the proposed study was compared with the work done by Sembiring et al., [11]. The authors have presented a subspace clustering algorithm using WEKA tool. The specifications of the data used and the processing time taken is exhibited in Table III.

TABLE III. RESULT ANALYSIS OF PROCESSING TIME

Database Size	PROCESSING TIME	
	Proposed System	Sembiring Approach
10 GB	0.211	0.414
20 GB	0.213	0.457
30 GB	0.214	0.614
40 GB	0.216	0.703
50 GB	0.216	0.815
60 GB	0.217	0.816
70 GB	0.301	0.818
80GB	0.399	0.818
90GB	0.411	0.829
100GB	0.487	0.899

From Table III it is evident that the amount of processing time increases with increase in the database size. However, the proposed system consumes almost half of processing time as compared to the conventional scheme [11].

The outcome of the study shows that proposed system consume lesser processing time by using the subspace clustering algorithm to yield the similar outcome for same test data as that considered for study in Sembiring et al., [11] approach.

VII. CONCLUSIONS

This paper has discussed one of the simplest approaches of solving curse of dimensionality in high-dimensional data. It was studied performing analysis over the high-dimensional data is a time consuming as well as resource consuming process. Moreover, high dimensional data has abundant level of redundancies, therefore performing any forms of processing will lead to incorrect segmentation of the data points or non-disclosure of the subspaces from the clusters. This problem was solved in this paper by introducing a simple technique of subspace clustering process. The outcome of the study shows that the proposed system is able to perform cluster analysis of massive dataset in least duration of time. Our future work is to develop a fast clustering

algorithm for high-dimensional data using optimization theory.

REFERENCES

- [1] M Verleysen. "Learning High-Dimensional Data", University atholique Louvain, Microelectronics laboratory, pp. 141-162, 2003.
- [2] Kogan, Jacob. "Introduction to Clustering Large and High-dimensional Data", Cambridge University, 2007.
- [3] C Giraud "Introduction to High-Dimensional Statistics", Taylor & Francis group, 2014.
- [4] R Agrawal, J Gehrke, D Gunopulos, and P Raghavan "Subspace Clustering Technique" Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. vol. 27, no. 2. 1998.
- [5] N Tomašev, M Radovanović, D Mladenović and M Ivanović, "Hubness- based Clustering of High-dimensional Data", In Partitional Clustering Algorithms, Springer International Publishing, pp. 353-386, 2015.
- [6] C Yang, D Robinson and R Vidal, "Sparse Subspace Clustering with Missing Entries", In Proceedings of the 32nd International Conference on Machine Learning, pp. 2463-2472, 2015.
- [7] Y Wang, C Edu, and J Zhu, "DP-Space: Bayesian Nonparametric Subspace Clustering with Small-variance Asymptotics", In Proceedings of the International Conference on Machine Learning (ICML). 2015.
- [8] Y Wang, Y-X Wang, and A Singh, "Clustering Consistent Sparse Subspace Clustering", arXiv preprint arXiv: 1504.01046, 2015.
- [9] J Wei, M Wang and Q Wu, "Study on Different Representation Methods for Subspace Segmentation", International Journal of Grid Distribution Computing, vol.8, no.1, pp.259-268, 2015.
- [10] A Petukhov and I Kozlov, "Greedy Algorithm for Subspace Clustering from Corrupted and Incomplete Data", IEEE Transaction on Information Security, 2015.
- [11] 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.