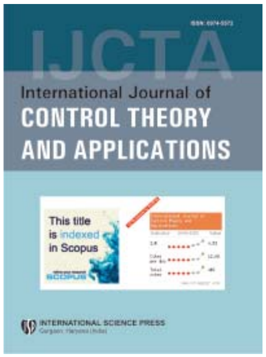


International Journal of Control Theory and Applications

ISSN : 0974–5572

© International Science Press

Volume 9 • Number 42 • 2016

Design a New Model for Enhanced Subspace Clustering of High – Dimensional Data

J. RamaDevi[a] and M. Venkateswara Rao[b]

[a]*Research Scholar, Department of Computer Science and Engineering, GITAM Institute of Technology, GITAM University, Visakhapatnam, Andhra Pradesh, India.*

[b]*Professor. Department of Information Technology, GITAM Institute of Technology, GITAM University,Visakhapatnam, Andhra Pradesh, India.*

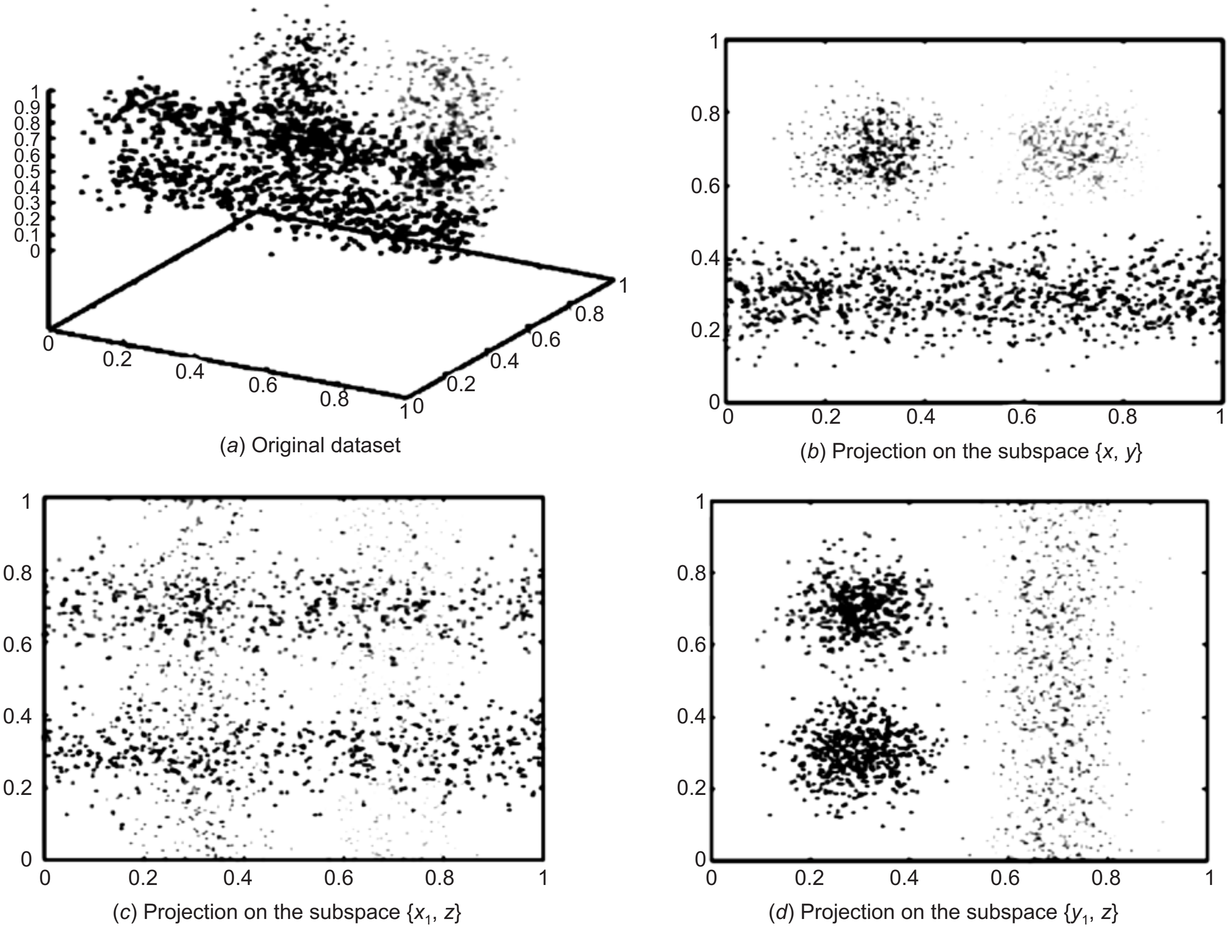
Abstract: Identifying clusters in the high dimensional data is an essential and challenging data mining problem. Clusters of objects in all subspaces of a dataset, that is different subset of dimensions called subspace clustering. In particular, high dimensional data sets can be well understood by clustering it in its subspaces. Many of existing subspace clustering algorithms desire several database scans and produce a more number of redundant subspace clusters in growing exponential data. To solve this problem, we propose a new model for discovering maximal subspace clusters and requires only n database scans for n dimensional data sets. In this paper a new model Enhanced subspace clustering *i.e* "EnSubClu" is designed by two phases. We first, find the dense units in different subspaces. Second, generate the maximal subspace clusters w.r.t. the objects in identified subspaces. Many big application areas like social networking, computer vision, biology, financial and sales analysis are maintained high dimensional data which is supported the proposed new enhanced subspace clustering algorithm.

Keywords: *Data Mining, High dimensional data, Subspace clustering, New enhanced subspace clustering.*

1. INTRODUCTION

Clustering is a data mining problem, which finds dense regions in high dimensional data space. Traditional clustering algorithms were constructed to produce clusters in the full dimensional space [1,2]. But as the growing of dimensionality of data tends to few dimensions become irrelevant to some clusters. These traditional clustering algorithms are inefficient for increasing volume of data. The high dimensional data also suffers due to the curse of dimensionality[3][4]. The following fig 1 Aggarwal R et al.[5] describes the high dimensional data attributes.

The problem of identifying clusters present in the subspaces of a high dimensional data space that allows quality clustering of the data objects than the full space called subspace clustering. More number of dimensions are not easy to handle. So the complexity increases exponentially with the dimensionality. As n – dimensional data can have up to $2^n – 1$ possible subspaces are require n number of data scans. So these state -art-of subspace clustering algorithms computationally expensive. Thus, we propose a new generic model enhanced subspace clustering, which eliminates the redundant subspaces and reduce the database scans.



(*a*) Original dataset

(*b*) Projection on the subspace {*x*, *y*}

(*c*) Projection on the subspace {x_1, *z*}

(*d*) Projection on the subspace {y_1, *z*}

Figure 1: Existence of redundant attributes and Existence of irrelevant attributes

1.1. Motivating examples

In social networks the detection of communities having identical interests can support both target marketers[30] and sociologists. Gunneman et al.[6] proposed subspace clustering on social networking graphs for community detection. Another important area of subspace clustering is radio astronomy[7], clusters of galaxies the origin of universe theories. In sales analysis: by recognizing the different subspace clusters that exist in large amount of sales data, we can detect which of the different attributes are related. This can be useful promoting the sales and in planning the inventory levels of various products. Another area of subspace clustering is web text mining through document clustering.Li et al.[8] proposed iterative subspace clustering for text mining. In biology[9][10], Eren et al.[11] have compared the performance of related subspace clustering approaches in micro array data.

Subspace clustering is important in computer vision and image problems. Example, moving objects[12,13] and identification of faces. In all of these applications described useful knowledge is unrevealed in the underlying subspaces of the data. There are two main techniques in the literature to share with subspace clustering top down and bottom up approaches. Only the user determines the number of clusters and relevant subspaces [14,15].This approach is not applicable to automatically detect all possible clusters in all subspaces. So the other technique for subspace clustering is the bottom-up hierarchical method located on Apriori principal [16].This approach has the capability of generating all hidden clusters in their relevant subspaces.[17-18].Corresponding Apriori principal of downward closure property of the dense regions in higher dimensional subspaces are close to each other. As the following fig 2 [19], the one-dimensional clusters from subspaces ({1},{3} and {4} are combined to find the clusters in the two dimensional subspaces ({1,3},{3,4} and {1,4}) and finally the three-dimensional space {1,3,4} . Anyhow, several combinations of the lower dimensional clusters over the step-by-step bottom-up clustering process. Even though these algorithms look likely for detecting all possible subspace clusters, they fail to scale due to the exponential search space [19, 20]. Detection of more redundant trivial clusters and an excessive number of database scans leads to inefficiency in the traditional bottom-up subspace clustering algorithms. However sophisticated indexing structures have been used for these look-ups, even it remains one of the main inefficiencies in the subspace clustering techniques yet. In this paper we present a new model enhanced subspace clustering algorithm that focus to remove these inefficiencies.

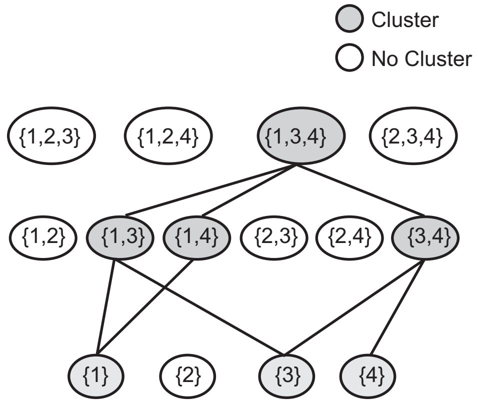


Figure 2: Shows the Bottom-up clustering

This paper will be coordinated into a few portions. Initially we will present current literature related to our approach. Next discuss the propose work methodology in new enhanced subspace clustering algorithm. The proposed algorithm and designing models were discussed in followed sections and concluding remarks.

2. BACKGROUND AND LITERATURE REVIEW

The traditional clustering techniques use the whole data to detect full dimensional clusters. DBSCAN [2] is a well known full dimensional clustering technique. However,the curse of dimensionality implies that the data loses in high dimensional space[3,4].so these full dimensional clustering approaches are not capable to find any useful clusters with the increase in dimensionality of the data. Another important technique to handle with the high dimensionality is reduce the number of dimensions by eliminating the irrelevant dimensions .Eg : PCA (Principal component Analysis). If no cluster structure was detected in the original space [21]. No new clusters in the transformed dimensions will be detected also, dimensionality reduction is not attainable[17]. Subspace clustering is a classic problem where one is liable points in a high dimensional feature space and the significance of the local relevance over the data w.r.t. the subset of dimensions. that has lead to the arrival of subspace clustering algorithms.[20,9,2]. There are two main techniques in to share with subspace clustering top down and bottom up approaches.**Top down approaches** are PROCLUS [14] and FINDIT[15] use projected clustering for high dimensional data.These algorithms fail to find all maximal clusters and not applicable to automatically detect all possible clusters in all subspaces. **The bottom up approach** is based on the downward cluster feature of the Apriori principal which was used for frequent item set mining [16]. In this principal, a set of points from n- dimensional space when projected onto a lower (n-1) dimensional is not dense. These algorithms can detect all possible clusters in all subspaces but they breakdown to scale with dimensions[17-18, 22, 23]. Agrawal et.al [17] were first to propose the grid-density based algorithm in their famous CLIQUE algorithm to detect the subspace clusters in the data. All of these approaches MAFIA[24], ENCLUS [22] are the generation of (n-1) dimensional units before n- dimensional units ,leads to inefficiency as lot of duplicate clusters and generate trivial clusters. Bottom- up clustering approaches, SUBCLU[25] evaluates the clusters higher dimensional clusters utilizing the lower dimensions clusters. SUBCLU generates all lower dimensional trivial clusters and fails to satisfy the generating maximal subspace clusters. Kriegel et al. Proposed FIRES[18] which determines 1-dimensional histograms called base clusters. These base clusters are merged by intersecting points to find maximal subspace clusters, but it requires multiple database scans. Assent et al. proposed INSCY [23] for subspace clustering which is an extension of SUBCLU. They use a SCY-tree called index structure which can be traversed in depth first order and generate the high dimensional clusters. So it requires multiple database scans and generate all intermediate trivial clusters .The multi-dimensional index structure introduces computational cost as well as inefficiency. All of these existing subspace clustering algorithms required multiple database scans and generate more trivial subspace clusters [27] with expensive costs. In the next section, we propose “EnSubClu” : A new Enhanced Subspace Clustering design model and methodology.

3. PROPOSED DESIGN MODEL

We propose a generic new model enhanced subspace clustering “Ensubclu” to find non trivial subspace with minimal cost clusters and requires only n database scans for *n* – dimensional data sets. We present the details of “EnSubClu” algorithm and its design models. This approach follows two phases. In phase1, the “EnSubClu” aims to extract the subspace clusters by finding the dense units in the relevant subspaces of given high dimensional dataset.

Based on the monotonicity of the Apriori principal, a group of dense points in an n –dimensional space S is dense in all the lower dimensional projectional of zone.

In preference, if we have dense sets of a given data. These 1-dimensional common points sufficiently will point us to dense points in the higher dimensional subspaces.. Based on this criteria proof, we develop our model to effectively and find the maximal clusters in all possible subspaces of a high-dimensional data set. Here we observe the our approach state model in fig 3.

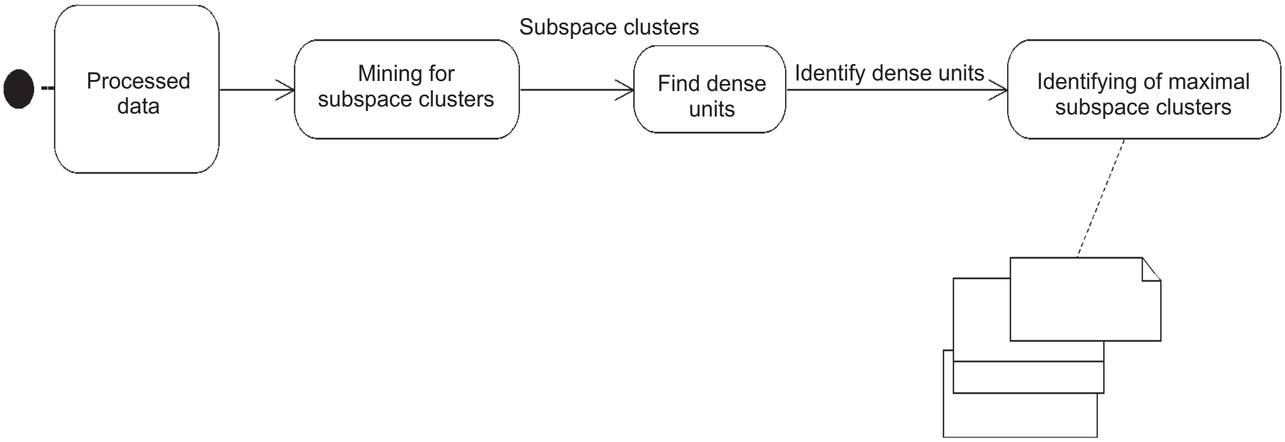

Processed data
Mining for subspace clusters
Subspace clusters
Find dense units
Identify dense units
Identifying of maximal subspace clusters

Figure 3: Proposed pipeline model

Definitions and problem : We assume Db = {p1, p2,--------, pm}, is a database of m points in a n-dimensional space, where each point is a n-dimensional vector, $s.t.$ pi = {pi^1, pi^2, p, ---, pi^n}. A subspace S is subset of full attribute set D_s, $s.t$ an k-dimensional subspace is denoted as S = {d1, d2,----, dk} where d_i € D_s and 1< = k <= n. A subspace s^l is the projection of a high dimensional subspace S. if $s^l c$ S. In this ,we retrieve the definition of density from DBSCAN[2] which is based on € and ÷ parameters $s.t.$, a point is dense if it has at least ÷ points within € distance.

3.1. Phase I: Finding dense units

3.1.1. Here we follow the observations

Observation 1: At least ÷ +1 density connected points from a dimension d_i also exist in the single dimensions *dj*,---- *dr*. So these points will form a set of dense points. S = {*di*, *dj*,---- *dr*} in the maximal subspace. We observe the fig 4 , shows the activity diagram of finding dense units and observe the step by step flows in the design model.

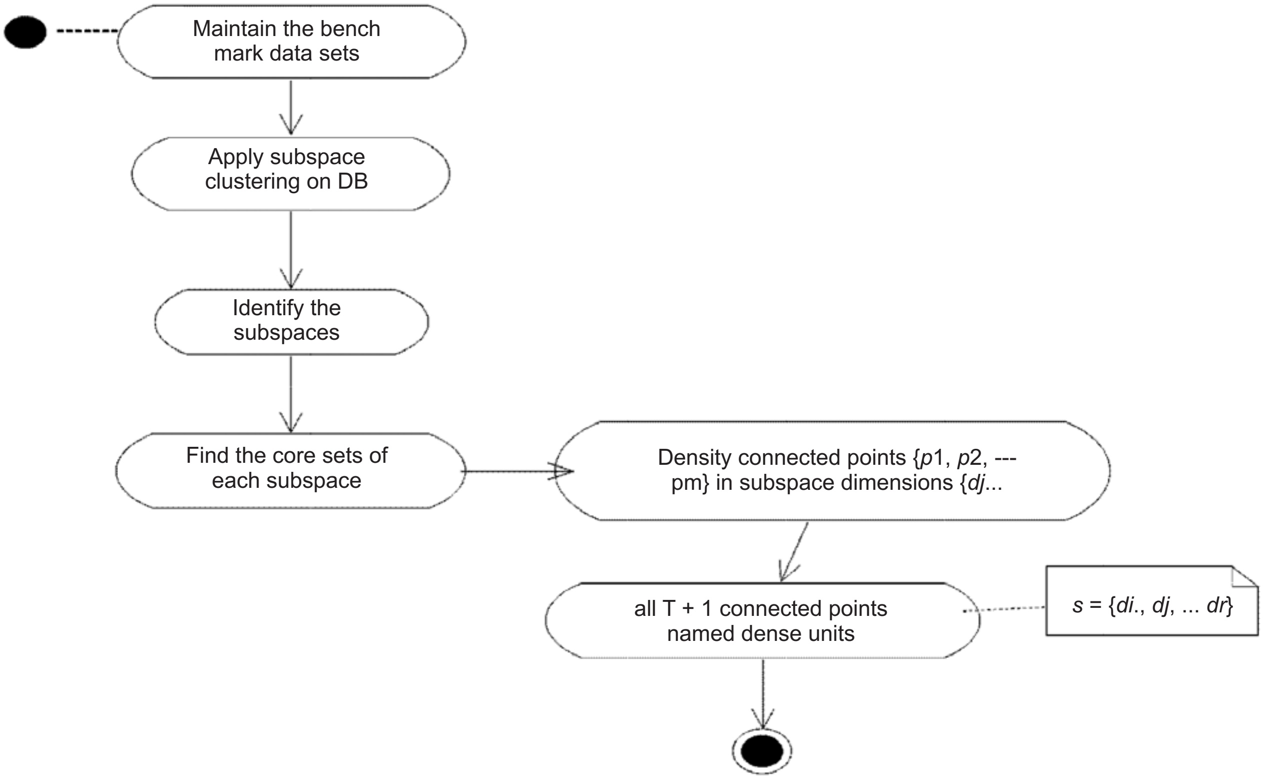

Maintain the bench mark data sets
Apply subspace clustering on DB
Identify the subspaces
Find the core sets of each subspace
Density connected points {*p*1, *p*2, --- pm} in subspace dimensions {*dj*...
all T + 1 connected points named dense units
s = {*di*., *dj*, ... *dr*}

Figure 4: Flow of core sets in dense units

Observation 2: Following the observation 1, In order to test if two dense units are same, we propose a noval idea "Adding Signatures" to each of these 1-D dense units and observed in the following fig 5. The functionality is,comparing the individual points among all dense units.The dense unit contains same points or not. Finally we can hash the signatures by extendable hashing datastructures and resulting collisions will give us to maximal subspace dense units. Our noval proposal for adding signatures to the dense units is adopted by the work in number theory by Erdos and Lehner [26]. We observe that if sum (U1) = sun (U2) = L,then U1 and U2 are same with an extremely high probability, otherwise L is large. Thus ∦ = ÷ + 1, the two dense units U1 and U2 will contain the same points with very high probability, if sum(U1) = sum(U2), mentioned this sum is very large.

3.1.2. Phase II: Identification of maximal subspace clusters

We can hash their signatures sums to extendable hash table as shown in fig 6. If all the sums collide these dense units are same with high probability. Thus, after hashing the dense units of collisions in all dimensions generate dense units in the appropriate maximal subspaces .We can integrate these dense units to obtain final clusters in their respective signatures. Extract subspace clusters by finding the dense units in the relevant subspaces of given data set. We now have dense units in all maximal subspaces. We process these to create density reachable sets and hence maximal clusters. We apply DBSCAN in each found subspace for clustering process which € and ÷ parameters can be maintained differently.

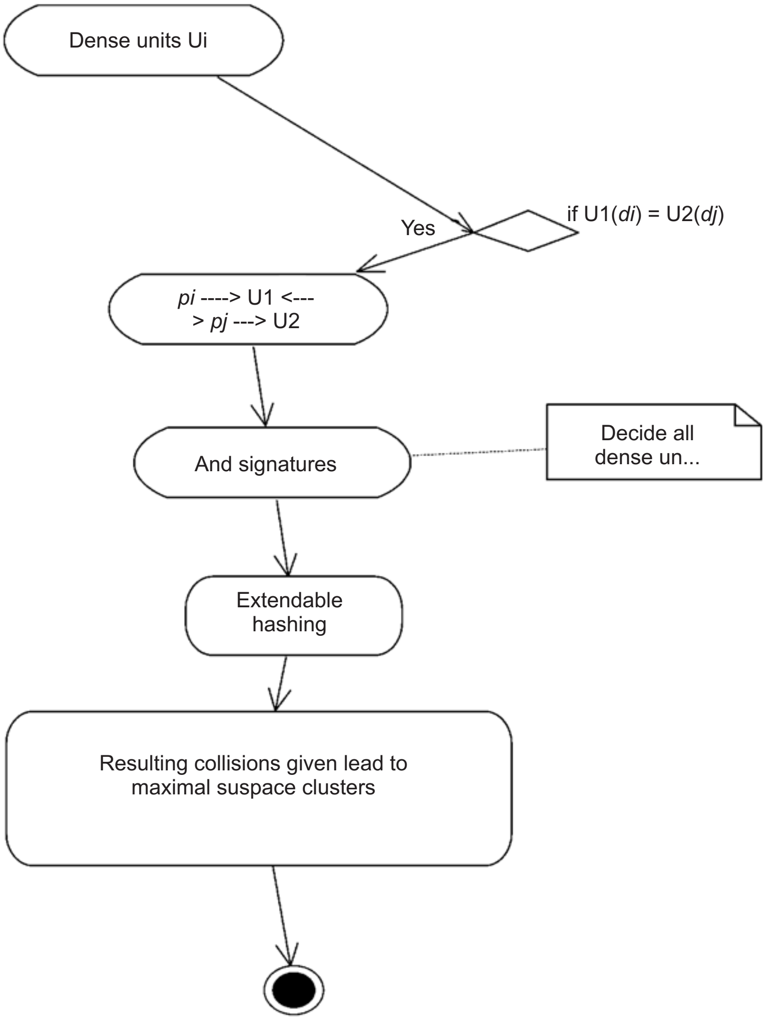


Figure 5: Identification of maximal subspace clusters

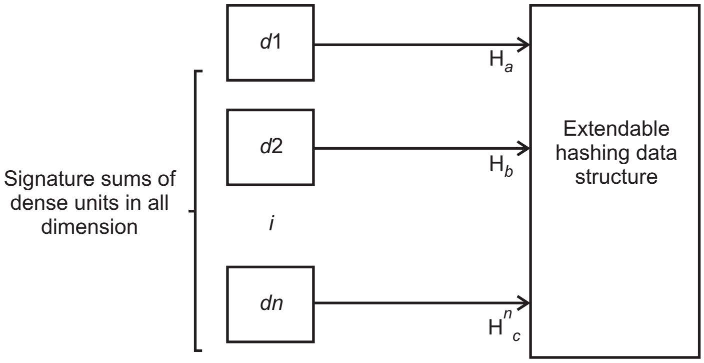

d1
d2
i
dn
Signature sums of dense units in all dimension
H_a
H_b
H^n_c
Extendable hashing data structure

Figure 6: Collisions among the signatures from different dimensions

In our design model , we analyse, how to identify the core sets in each dimensions and then produce all combinations of size ÷ + 1 as potential dense units and improve the significance of subspace clusters. Moreover, design the phases in terms of activity flows. However, one fine advantage of our approach is that the only time we need to scan the database set in the complete algorithm.

4. CONCLUSION

In this paper, we proposed the new enhanced subspace clustering algorithm “EnSubClu” which overcomes the problems of existing subspace clustering methods including scalability [29]w.r.t. data dimensionality and improving the clustering results[28]. This paper provides algorithm design models for finding dense units and providing maximal subspace clusters in high dimensional space. Further we will be describe the algorithm steps and implementing the algorithms for the above mentioned design models. Though,we will be evaluate efficient results on benchmark data sets in high dimensional data and visualizing the results.

REFERENCES

[1] Zhang T, Ramakrishnan R, Livny M ,“BIRCH: an efficient data clustering method for very large databases”, In: Proc. of the ACM SIGMOD international conference on management of data, vol. 1. ACM Press, USA. pp 103–114 , 1996.

[2] Ester M, Kriegel H, Sander J, Xu X ,”A density-based algorithm for discovering clusters in large spatial databases with noise”, Int Conf Knowl Discov Data Min 96(34):226–231 , 1996.

[3] Bellman RE ,“Adaptive control processes: a guided tour”, Princeton University Press, New Jersey,1961.

[4] Beyer K, Goldstein J , “When is nearest neighbor meaningful? Proc 7th Int Conf Database Theory. In: Database Theory –ICDT’99. Lecture Notes in Computer Science. Springer, Berlin Heidelberg Vol. 1540. pp 217-235,1999.

[5] Agrawal R, Gehrke J, Gunopulos D, Raghavan P (1998) Automatic subspace clustering of high-dimensional data for data mining applications. *In: Proceedings of the ACM international conference on management of data (SIGMOD),* pp 94–105.

[6] Günnemann S, Boden B, Seidl T, “ Finding density-based subspace clusters in graphs with feature vectors”,In: Data mining and knowledge discovery. Springer, US Vol. 25. pp 243–269,2012.

[7] Jang W, Hendry M , “ Cluster analysis of massive datasets in astronomy”, Stat Comput 17(3):253–262,2007

[8] Li T, Ma S, Ogihara M,“ Document clustering via adaptive subspace iteration”, In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. ACM, USA. pp 218–225,2004.

[9] Babu MM ,” Introduction to microarray data analysis”. In: Grant RP (ed). Computational genomics: Theory and application. Horizon Press, UK. pp 225–249,2004.

[10] Eisen MB, Spellman PT, Brown PO, Botstein D ,” Cluster analysis and display of genome-wide expression patterns “, Proc Natl Acad Sci 95(25):14863–14868 ,1998.

[11] Eren K, Deveci M, Kktun O, atalyrek mV,”A comparative analysis of biclustering algorithms for gene expression data”, Brief Bioinforma 14(3):279–292,2013.

[12] Ho J, Yang MH, Lim J, Lee KC, Kriegman D , “Clustering appearances of objects under varying illumination conditions”,In: Computer vision and pattern recognition, 2003. Proceedings. 2003 IEEE computer society conference on, vol. 1. IEEE. pp 1–11,2003.

[13] Tierney S, Gao J, Guo Y ,” Subspace clustering for sequential data. In: Computer vision and pattern recognition”, (CVPR), IEEE conference On. IEEE. pp 1019–1026, 2014.

[14] Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, Park JS,”Fast algorithms for projected clustering”, In: Proc. of the ACM SIGMOD international conference on management of data. ACM, USA. pp 61–72,1999.

[15] Woo KG, Lee JH, Kim MH, Lee YJ , FINDIT: “a fast and intelligent subspace clustering algorithm using dimension voting”, Inf Softw Technol 46(4):255–271 ,2004.

[16] Agrawal R, Mannila H, Srikant R, Toivonen H, Verkamo AI, " Fast discovery of association rules", Adv Knowl Discov Data Min 12:307–328,1996.

[17] Agrawal R, Gehrke J, Gunopulos D ,"Automatic subspace clustering of high dimensional data for data mining applications", In: Proc. of the ACM SIGMOD international conference on management of data. pp 94–105,1998.

[18] Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, Park JS " Fast algorithms for projected clustering", In: Proc. of the ACM SIGMOD international conference on management of data. ACM, USA. pp 61–72,1999.

[19] Aggarwal CC, Reddy CK , " Data clustering: algorithms and applications",Data Mining Knowledge and Discovery Series 1st. CRC Press ,2013.

[20] Parsons L, Haque E, Liu H " Subspace clustering for high dimensional data: a review", ACM SIGKDD Explor Newsl 6(1):90–105,2004.

[21] Joliffe IT ,"Principle component analysis", 2nd edn. Springer, New York,2002.

[22] Cheng CH, Fu AW, Zhang Y " Entropy-based subspace clustering for mining numerical data", In: ACM SIGKDD international conference on knowledge discovery and data mining. ACM, NY, USA. pp 84–93,1999.

[23] Assent I, Emmanuel M, Seidl T, "Inscy: Indexing subspace clusters with in-process-removal of redundancy", In: Eighth IEEE international conference on data mining. IEEE. pp 719–724,2008.

[24] Nagesh H, Goil S, Choudhary A ,"Adaptive grids for clustering massive data sets", Proc 1st SIAM Int Conf Data Min:pp. 1–17,2001 .

[25] Kailing K, Kriegel HP, Kroger P," Density-connected subspace clustering for high-dimensional data",In: SIAM international conference on data mining. pp 246–256,2004.

[26] Erdös P, Lehner J ," The distribution of the number of summands in the partitions of a positive integer", Duke Mathematical Journal 8(2):335–345,1941.

[27] Müller E, Günnemann S, Assent I, Seidl T, Färber I ,"Evaluating Clustering in Subspace Projections of High Dimensional Data", 2009.

[28] Sim K, Gopalkrishnan V, Zimek A, Cong G ,"A survey on enhanced subspace clustering", Data Min Knowl Disc 26(2):332–397,2012.

[29] Fan J, Han F, Liu H ,Challenges of big data analysis. National Science Review 1(2):293–314,2014.

[30] Vidal R, Tron R, Hartley R, "Multiframe motion segmentation with missing data using PowerFactorization and GPCA", Int J Comput Vis 79(1):85–105,2008.