I.J. Mathematical Sciences and Computing, **2019, 3, 37-48**
Published Online July 2019 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2019.03.04

Available online at http://www.mecs-press.net/ijmsc

Mining Maximal Subspace Clusters to deal with Inter-Subspace Density Divergence

[1*]B.Jaya Lakshmi, [2]K.B.Madhuri

[1*]*Assiociatet Professor, Department of IT, GVP College of Engineering (A), Andhra Pradesh, 530048, India*
[2] *Professor, Department of IT, GVP College of Engineering (A), Andhra Pradesh, 530048, India*

Received: 06 March 2019; Accepted: 13 June 2019; Published: 08 July 2019

Abstract

In general, subspace clustering algorithms identify enormously large number of subspace clusters which may possibly involve redundant clusters. This paper presents Dynamic Epsilon based Maximal Subspace Clustering Algorithm (DEMSC) that handles both redundancy and inter-subspace density divergence, a phenomenon in density based subspace clustering. The proposed algorithm aims to mine maximal and non-redundant subspace clusters. A maximal subspace cluster is defined by a group of similar data objects that share maximal number of attributes. The DEMSC algorithm consists of four steps. In the first step, data points are assigned with random unique positive integers called labels. In the second step, dense units are identified based on the density notion using proposed dynamically computed epsilon-radius specific to each subspace separately and user specified input parameter *minimum points*, τ. In the third step, sum of the labels of each data object forming the dense unit is calculated to compute its signature and is hashed into the hash table. Finally, if a dense unit of a particular subspace collides with that of the other subspace in the hash table, then both the dense units exists with high probability in the subspace formed by combining the colliding subspaces. With this approach efficient maximal subspace clusters which are non-redundant are identified and outperforms the existing algorithms in terms of cluster quality and number of the resulted subspace clusters when experimented on different benchmark datasets.

Index Terms: Subspace Clustering, Maximal subspace Clusters, Inter-Subspace Density Divergence, Dynamic Epsilon, Density Notion.

* Corresponding author.
E-mail address: meet_jaya200@gvpce.ac.in

1. Introduction

Clustering is a process of automatically finding groups of similar data points for a given data set. Finding the clusters in the high dimensional datasets is an important and challenging data mining problem. An object may share a subset of attributes, A with one group of objects while the same object shares another subset of attributes, A' with another group of objects [1,2]. Hence, a subspace is defined by a subset of attributes and provides a basis for clustering the objects into groups called subspace clusters denoted by $<C_i, A_j>$ where C_i is the set of objects constituting the i^{th} cluster in the j^{th} subspace defined by the subset of attributes A_j.

The number of possible subspaces increases exponentially with the dimensionality of dataset, which in turn results in enormously large numbers of subspace clusters affecting the interpretability of results negatively [3,4]. For n-dimensional data, the possible number of subspaces are 2^n-1.

Subspace clustering [5] is the process of identifying clusters with objects similar in various subsets of attributes defining subspaces. A subspace cluster is denoted with two dimensions <C, A> representing the set of objects grouped into the cluster and the set of attributes defining the subspace.

1.1 Need for Subspace Clustering

The real-world data often consists of descriptions of complex data objects each of which is described in terms of a large set of attributes or variables [6]. Availability of data in abundance calls for efficient algorithms to analyze the data for pattern extraction. This is a challenging task [7,8,9]. Data mining functionalities such as cluster analysis become more complex as the number of dimensions increase [10]. The distance between data points may not be properly discriminated in a high dimensional space. This is referred to as curse of dimensionality.

1.2 Subspace Clustering

Subspace Clustering is an extension to traditional clustering that seeks to find clusters in different subspaces of a dataset. Often in high dimensional data, some dimensions may be irrelevant and this may mask the true clusters which are hidden in subspaces.

1.3 Inter-Subspace Density Divergence

Most of the subspace clustering techniques adopt density based clustering methods. The natural and arbitrary shaped clusters are identified using density based clustering algorithms and they do not ask for number of clusters as input parameter. These algorithms are also relatively insensitive to outliers. DBSCAN [11] is one of the density based clustering algorithms which uses two input parameters, namely τ and *epsilon* in order to find density connectivity among the objects and form dense clusters. As shown in Figure 1 when the data objects are uniformly spread among different clusters in a subspace, DBSCAN produces good results with appropriate input parameters. DBSCAN is adopted in SUBCLU which is a successful subspace clustering algorithm. SUBCLU [12] algorithm applies DBSCAN to find clusters in (k+1)-dimensional subspace by further partitioning the objects of clusters found in k-dimensional subspaces. The clustering when formed using the same parameter at higher dimensional subspaces may result in noises. Due to the concept of *Inter-Subspace Density Divergence*, [13,14] the objects in subspaces of higher dimensions are expected to be spread farther away which calls for subspace dimensionality specific parameter setting.

Existing subspace clustering algorithms like DENCOS [14], SCHISM [15] etc. that handle *density divergence* are grid based approaches wherein the density threshold of the grid cell in different subspaces is determined in terms of dimensionality of the subspace. Such approaches impose the same density threshold for all subspaces of equal dimensionality. However, the authors propose to apply tailor-made density thresholds for the individual subspaces based on the spread of the data in terms of the distance estimated in the projected

subspaces. Accordingly, the density thresholds in a two-dimensional subspace, formed by attributes A1A2 is likely to be different from that of A2A3 or A1A4, though all of them are two-dimensional.

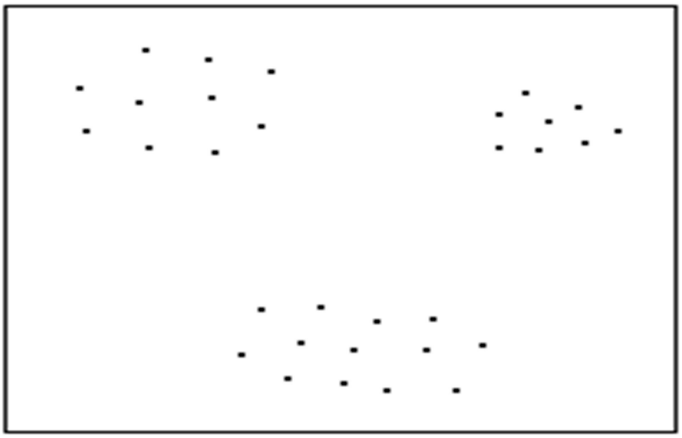

Fig. 1 A subspace with uniform spread of data objects

The spread of the data in the subspaces of different dimensionalities may vary. The Euclidean distance between a pair of objects (X, Y) represented in k-dimensional space as X=<x1, x2…xk> and Y= <y1, y2…yk> is estimated as $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ which is additive in nature with increase in dimensions. This results in increased distance between the same pair of objects when considered in higher dimensional subspaces, leading to reduced density in higher dimensional subspaces referred to as density divergence. In other words, the average density of objects decreases with increase in the dimensionality of the subspaces. For example, consider a dataset with A1, A2, A3 as attributes. A1A2 subspace may contain denser clusters compared to A1A2A3 subspace. The authors propose to capture the variation in density with appropriate variation in the value of epsilon while keeping the minimum points, τ as constant.

1.4 Estimation of Density Threshold

The conventional density based subspace clustering algorithms form clusters based on the density threshold represented by two user defined parameters, <*epsilon, τ* >. In order to handle inter-subspace density divergence automatically, the proposed algorithm, DEMSC algorithm has to use multiple density thresholds. For simplicity without loss of generality, in DEMSC algorithm the value of τ is fixed while changing the value of *epsilon* to represent varied densities.

The clustering process proceeds by dynamically computing epsilon value for each k-d subspace based on the average spread of the data using the following equation (1). In this paper, k-d subspace refers to the set of k-dimensional subspaces denoted by S_K while s_k is one among them. τ is one of the user defined input value provided to the algorithm with which core point identification is made. A core point is the data object which has at least τ number of data objects within $epsilon^{s_k}$ - neighborhood of it.

$$epsilon^{s_k} = \frac{\tau * maximum\_distance^{s_k}}{|D|} \quad (1)$$

Where, | D | is the total number of data objects in the dataset D. $\text{maximum\_distance}^{s_k}$ is the distance between farthest pair of data objects in the projected k-d subspace. In equation (1), epsilon^{s_k} is proportional to the $\text{maximum\_distance}^{s_k}$ which increases with the number of dimensions of the subspace. $\text{maximum\_distance}^{s_k}$ is divided by the size of the dataset to estimate the average density of the data objects in a k-dimensional subspace. This is multiplied with τ to obtain an estimate of the threshold, epsilon which is the radius of the neighborhood for the given k-d subspace.

A new subspace clustering algorithm, Dynamic Epsilon based Maximal Subspace Clustering Algorithm (DEMSC) is proposed to deal with redundancy and the phenomenon of inter-subspace density divergence. The DEMSC algorithm is capable to identify improved quality dense subspace clusters which are maximal.

2. RELATED WORK

Kailing et al. [12] have identified the drawbacks of grid-based subspace clustering methods and proposed SUBCLU (density connected SUBspace CLUstering), a basic subspace clustering algorithm which follows bottom-up approach and density connectedness as the clustering notion. The concept of density connectedness [11] is used to explore the hidden dense subspace clusters. The algorithm begins by forming single-dimensional subspace clusters using a density-based clustering method, DBSCAN [11]. Each of them are further partitioned into higher-dimensional subspace clusters, thus following a hierarchical method in partitioning the subspace clusters. To avoid exploration of the entire exponential search space, the property of anti-monotonicity in density connectivity is used to efficiently prune the candidate subspaces. Since density notion is applied in the algorithmic process, arbitrarily shaped and positioned dense subspace clusters are identified. The greedy approach mines all the dense regions hidden in subspaces of high dimensional data.

Maria et al. [16] have considered streaming data for subspace clustering which is a quite challenging task as data gets evolved time to time. The authors have used a sliding window protocol model to form the clusters incrementally. The clusters continuously and gradually get updated with the changing time series data. Based on pair-wise stream similarities in a particular subspace projection, the proposed algorithm mines the maximal subspace clusters. The pruning criteria (i.e. cluster pruning, dimension pruning and stream pruning) that is proposed have reduced the search space to a significant extent. The incremental clustering that aims to identify maximal subspace clusters is proved to be efficient for both static data and time series data.

Support and Chernoff-Hoeffding bound-based Interesting Subspace Miner, SCHISM [15] mines interesting and maximal subspace clusters considering density divergence of subspaces. The algorithm considers a subspace to be interesting if it contains data points higher than the expected number i.e. user defined threshold. As CLIQUE [17], SCHISM is also a grid based subspace clustering algorithm that models Chernoff-Hoeffding bound-based threshold function which would be either constant or monotonically decreasing or monotonically increasing based on the dimensionality of the subspace. There are three main steps in the algorithm. In step 1, the dataset is discretized and represented in the vertical format to enable faster computation. In step 2, the algorithm follows depth first approach with backtracking in finding maximal interesting subspaces. An interestingness measure is used to prune the search space. In final step of the algorithm, each data point is either assigned to the most similar maximal interesting subspace or labelled as an outlier. As SCHISM is a static grid based algorithm, the limitations of such methods are still existent in the algorithm.

CLICKS [18] is an effective subspace clustering algorithm that mines subspace clusters in categorical datasets which was proposed by Mohammed et al. The categorical dataset is modelled as k-partite graph [19] in which vertex set is attribute values and is partitioned into k disjoint sets where the edges in different partitions are connected based on their relationship. Thus, the dataset is represented in a more compact way which would be effective when larger datasets are dealt. The attributes are also ranked using the concept of connectivity. The maximal k-partite cliques [19] in the graph are identified since they are analogous to the subspace clusters. Finally, the support of the candidate cliques is verified in the original dataset to ensure that the maximal cliques are dense and the overlapping cliques could be merged to form larger cliques.

Assent et al. [20] have proposed an algorithm called INdexing Subspace Clusters with in-process-removal of

redundancY (INSCY) that handles redundancy issue and eliminates it in an efficient way in top-down fashion. Instead of generating the subspace clusters and then removing the redundant ones, this approach finds only those subspace clusters which are non-redundant. For this to be accomplished, it uses a special index called SCY-tree which stores regions that are likely to hold subspace clusters. INSCY algorithm follows depth first approach, the maximal high dimensional clusters are included first in the result set before comparing them with the lower dimensional subspace clusters in the same region which can be pruned later if redundant. However, globally there might be some redundant clusters in the result set.

Yi-Hong et al. [21] have proposed a new algorithm called Non-Redundant Subspace Cluster mining (NORSC) for mining concise and non-redundant subspace clusters without loss of information and achieves maximum coverage of data points. NORSC handles two problems: information overlapping and data coverage. Since NORSC is a grid based subspace clustering algorithm, the feature space is divided into grid cells with fixed interval [22]. The NORSC algorithm is divided in two steps. In the first step, the maximal dense units are extracted. A dense unit is defined as a grid cell that consists of at least threshold number of data points. A maximal dense unit is the dense unit that is not dense while being extended into higher dimensional feature spaces. In the second step, the non-extensible data units for the given data point can be identified by intersecting the d-dimensional unit in the subspace where the data point lies and the maximal dense units. Later, the non-redundant clusters are incrementally discovered on different cardinalities of subspace.

With the increase in dimensionality of the databases, the problem of subspace clustering becomes more complex [23]. Amardeep and Amitav [24] have developed an algorithm called subscale that needs only k database scans for k-dimensional data. The algorithm determines maximal subspace clusters using a novel idea of computing signatures for each single-dimensional subspace clusters that are hashed to hash table. Initially, each data point in the database is assigned with a random integer. The dense units are identified in each single-dimensional projection of the database based on epsilon-radius ϵ and a user defined density threshold. The signature of each dense unit is computed by the summation of the random integers assigned to the data points contained in the dense unit. The collisions of signatures of the dense units are helpful in ultimately identifying the maximal subspace clusters without the generation of the candidates. The algorithm is proved for its parallelism, performance and scalability.

To handle the issues of exponential search space and the density divergence that are inherent in high dimensional datasets, Hans-Peter et al. [25] have proposed a framework called FIlter Refinement Subspace clustering (FIRES). The algorithm does not follow level wise exploration of the search space and no more candidates are used in determining higher dimensional subspace clusters. To speed up the process, the cluster approximations with maximum dimensionality are determined. FIRES algorithm mainly consists of three steps: (i) Using any clustering notion, single-dimensional clusters which are referred to as base clusters are generated. (ii) The maximal subspace clusters are approximated by merging the most similar candidate base clusters. (iii) The cluster approximations are refined by removing noise leading to true subspace clusters. FIRES algorithm is proved to be efficient when compared to traditional algorithms, CLIQUE [17] and SUBCLU [12].

Jian et al. [26] have proposed a model called Maximal Subspace Clustering (MSC) to mine subspace clusters hidden in maximal subspaces based on density and references [27]. The references are grouped to indicate that each one provides the basic information of a cluster. The shape and trends of data are captured by finding all references and later the data points are mapped to the references. The database is scanned only once to generate the enumeration tree and from which the maximal subspace clusters are mined. The property of monotonicity is used to prune the enumeration tree of subspaces. The simple set intersection operation of subspaces is used to generate subspace clusters. The MSC algorithm is scalable to high-dimensional datasets.

3. PROPOSED WORK

The main objective of Dynamic Epsilon based Maximal Subspace Clustering Algorithm (DEMSC) algorithm is to find maximal subspace clusters by finding the dense units in the subspaces. A cluster is said to be a maximal if there is no other cluster $C_j = (P, S')$ such that $S' \supset S$. A subspace which is maximal for a certain

group of points might not be maximal for another group of points.

A core point along with its neighbours forms a Core Set (CS). If |CS| is the number of data points in a Core-Set, then ($|^{CS}|C\tau+1$) number of dense units are formed from each Core Set. As an initial step, Core Sets are identified in each single dimensional subspace using dynamically computed epsilon given in Equation (1) for each subspace and minimum points, τ. A hash table (hTable) is created. For each Core Set, dense units are identified. For each dense unit, its signature (H_i) is computed by calculating the sum of the labels of each object and are hashed into hTable. If the signature of the dense unit of a subspace collides with that of the dense unit of other subspace, this implies that the same dense unit exists in both subspaces. The colliding subspaces against each signature form an entry in the hash table and union of colliding subspaces are formed and density-reachable dense units form maximal clusters. Repeat the process in all single dimensional subspaces. The block diagram of the DEMSC algorithm is given figure 2.

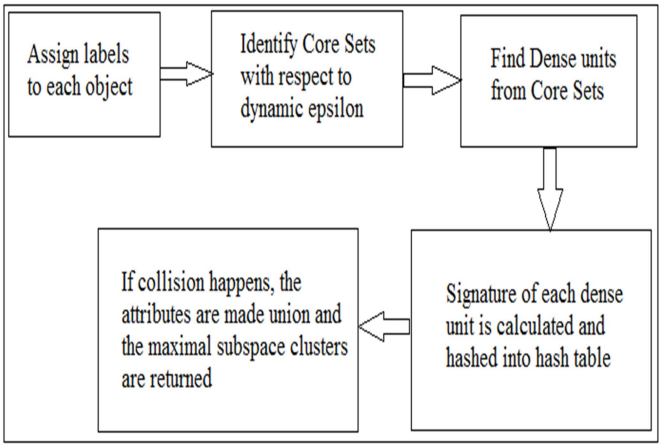

Fig. 2 Block Diagram of DEMSC Algorithm

3.1 Algorithm: Dynamic Epsilon based Maximal Subspace Clustering Algorithm (DEMSC)

STEP-1: Consider a set P consisting of positive and unique integers. P= {p_1, p_2…p_n}, assign to each data object.

STEP-2: CS be a Core-Set such that each object is within computed $epsilon^{S_k}$ distance in subspace S_k containing at least minimum points (τ).

$$epsilon^{S_k} = \frac{minpts * maximum\_distance^{S_k}}{|\,D\,|}$$

All possible dense units are identified using the formula $|cs|_{C_{\tau+1}}$. Appropriate value of epsilon is dynamically computed using equation (1) depending on the spread of the data in a given subspace.

STEP-3: Create a hash table. For each dense unit, compute its signature (H_i) by calculating the sum of the labels of each data object in the dense unit.

If the signature of the dense unit of a subspace collides with that of the dense unit of other subspace, this implies that the same dense unit exists in both subspaces. Store the colliding dimensions against each signature form an entry in the hash table.

Repeat the process for all single dimensional subspaces.

STEP-4: Obtained dense units are processed to create density-reachable maximal clusters. The performance of Dynamic Epsilon based Maximal Subspace Clustering Algorithm (DEMSC) mainly depends on generated dense units in single dimension.

4. RESULT ANALYSIS

To evaluate the performance of DEMSC Algorithm, experimentation is done on different numeric benchmark datasets from UCI machine learning repository [28]. The characteristics of the benchmark datasets in terms of number of data objects, number of attributes and number of classes are given Table 1.

Table 1. Characteristics of benchmark datasets

Dataset	#data objects	#attributes	#classes
Seed data	210	7	3
Image Segmentation data	2310	19	7
Bank Authentication	1372	4	2
Wine Quality data	4398	12	4
Glass data	214	10	7
Breast Tissue data	106	9	6

The proposed algorithm is compared with SUBCLU, a density based subspace clustering algorithm and SCHISM, a grid based maximal subspace clustering algorithm in terms of cluster quality metrics, Purity and Silhouette Coefficient and number of subspace clusters obtained.

Purity is an external evaluation criteria of cluster quality. It is defined as percentage of total objects which were classified correctly [29]. The range of purity is [0, 1]. The comparison of SUBCLU, SCHISM and DEMSC algorithms in terms of purity are depicted as bar chart in Figure 3. The actual values are shown in Table 2. The best purity values obtained by the proposed DEMSC algorithm are shown in bold.

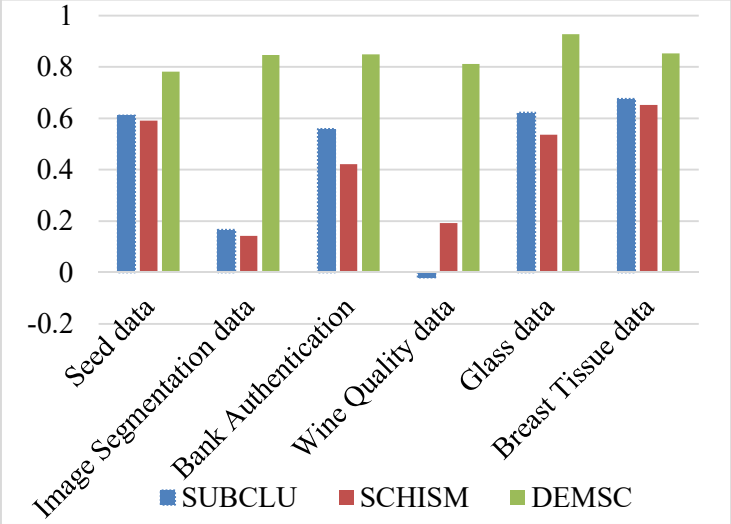

Fig. 3 Comparison of SUBCLU, SCHISM and DEMSC algorithms in terms of Purity on different datasets

Table 2. Comparison of SUBCLU, SCHISM and DEMSC algorithms in terms of Purity

Purity			
Dataset	SUBCLU	SCHISM	DEMSC (Proposed)
Seed data	0.729	0.736	**0.923**
Image Segmentation data	0.152	0.544	**0.922**
Bank Authentication	0.790	0.726	**0.927**
Wine Quality data	0.417	0.502	**0.892**
Glass data	0.796	0.753	**0.897**
Breast Tissue data	0.737	0.669	**0.916**

The quality of subspace clusters is also analysed in terms of Silhouette Coefficient(SC) which is applicable to unsupervised datasets also. Silhouette Coefficient [29] estimates the quality of a cluster in terms of the cohesion among the cluster members and separation of the cluster to its closest cluster. The value of SC ranges between -1 to +1 and higher values of SC implies better quality clusters. When SC value of a cluster is equal to 1, it implies that the cluster members are highly compact in nature and far away separated from the members of other clusters. This case is preferred as it is an indication for best clustering. When SC value is -1, it indicates that the objects of a cluster is more close / similar to the objects of other clusters rather than its cluster members. This is an indication for poor clustering and should be avoided.

The comparison of SUBCLU, SCHISM and the proposed DEMSC algorithms in terms of Silhouette Coefficient when experimented on different benchmark datasets is depicted as bar chart in Figure 4. The actual values of Silhouette Coefficient obtained by resulting clustering solution by different algorithms are tabulated in Table 3 and the best values are shown in bold. It is proved that DEMSC algorithm has produced improved quality clusters when compared to SUBCLU and SCHISM.

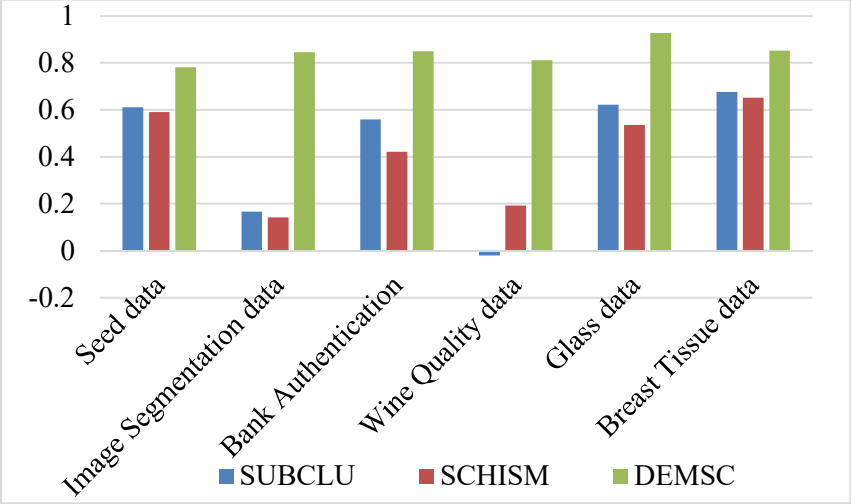

Fig. 4 Comparison of SUBCLU, SCHISM and DEMSC algorithms in terms of Silhouette Coefficient on different datasets

Table 3. Comparison of SUBCLU, SCHISM and DEMSC algorithms in terms of Silhouette Coefficient

Silhouette Coefficient			
Dataset	SUBCLU	SCHISM	DEMSC (Proposed)
Seed data	0.611	0.591	**0.782**
Image Segmentation data	0.166	0.142	**0.846**
Bank Authentication	0.559	0.422	**0.849**
Wine Quality data	-0.02	0.192	**0.812**
Glass data	0.622	0.536	**0.927**
Breast Tissue data	0.676	0.652	**0.853**

The number of subspace clusters resulted by the clustering solution obtained by different algorithms when implemented on different benchmark datasets are tabulated in Table 4.

It is proved that the proposed DEMSC algorithm has produced most concise and compact set of clusters indicated by minimum number of obtained subspace clusters which is easy for interpreting the clustering results. Minimum number of subspace clusters are resulted with minimal loss of information and at the same time retaining the quality of the subspace clusters.

Table 4. Comparison of SUBCLU, SCHISM and DEMSC algorithms in terms of No. of obtained subspace clusters

Number of obtained subspace clusters			
Dataset	SUBCLU	SCHISM	DEMSC (Proposed)
Seed data	6,686	2,934	106
Image Segmentation data	93,58,769	45,184	11
Bank Authentication	4,754	1,992	34
Wine Quality data	915,09,366	4,81,937	549
Glass data	17,135	16,592	384
Breast Tissue data	10,049	10,542	273

5. Conclusion

Existing subspace clustering algorithms generate enormously large number of subspace clusters after exploring a large number of subspaces which results in high complexity. This paper discusses the proposed algorithm, namely Dynamic Epsilon based Maximal Subspace Clustering Algorithm (DEMSC) to extract a compact set of maximal subspace clusters with minimal redundancy. To efficiently deal with inter-subspace density divergence, the DEMSC algorithm dynamically computes the density threshold in terms of epsilon specific to each subspace separately. It is proved that the DEMSC algorithm obtained high quality subspace clusters when compared to SUBCLU and SCHISM algorithms in terms of high purity and Silhouette Coefficient and minimal number of resulted subspace clusters.

This research work could be extended to apply on datasets of complex data and mission values. Fuzzy memberships could also be used while forming clusters in subspaces.

Acknowledgement

We would like to thank Dr.M.Shashi, Professor, Department of Computer Science and Systems Engineering,

Andhra University, Visakhapatnam for the support in the development of this research work.

References

[1] Assent I, Krieger, M¨uller, E and Seidl T. DUSC: Dimensionality unbiased subspace clustering. Proceedings International Conference on Data Mining 2007; 409–414.
[2] Lakshmi BJ, Shashi M and Madhuri K B. A Rough Set Based Subspace Clustering Technique for High Dimensional Data. Journal of King Saud University-Computer and Information Sciences [Online] 2017.
[3] Hans-Peter Kriegel, Peer Kro¨ Ger and Arthur Zimek Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. ACM Transactions on Knowledge Discovery from Data, 2009; 3(1).
[4] Jaya Lakshmi B, Shashi M and Madhuri KB. Summarization of Subspace Clusters based on Similarity Connectedness. International Journal of Data Science. 2018;3(3): 255-265,
[5] Jaya Lakshmi B, Shashi M and Madhuri K.B. Automation of Power Transformer Maintenance through Summarization of Subspace Clusters. Journal of Engineering Science and Technology. 2018;13(11):3610-3618 .
[6] Sahil Raj and Tanveer Kajla. Sentiment Analysis of Swachh Bharat Abhiyan. International Journal of Business Analyics and Intelligence, 2015;3(1): 32-38.
[7] Lifei Chen, Shengrui Wang, Kaijun Wang and Jianping Zhu. Soft subspace clustering of categorical data with probabilistic distance. Pattern Recognition. 2016; 51: 322-332.
[8] Wang Lijuan, Hao Zhifeng, Cai Ruichu and Wen Wen. “Enhanced soft subspace clustering through hybrid dissimilarity”. Journal of Intelligent and Fuzzy Systems, 2015; 29(4):1395-1405.
[9] Zhaohong Deng, Yizhang Jiang, Fu-Lai Chung, Hisao Ishibuchi, Kup-Sze Choi and Shitong Wang. Transfer Prototype-Based Fuzzy Clustering. IEEE Transactions on Fuzzy Systems. 2016b; 24(5).
[10] Feiping Nie, Dong Xu, Ivor Wai-Hung Tsang and Changshui Zhang. Flexible Manifold Embedding: A Framework for Semi-Supervised and Unsupervised Dimension Reduction. IEEE Transactions on Image Processing. 2010; 19(7)1921-1932.
[11] Ester M, Kriegel H, Sander J and Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. International Conference on Knowledge Discovery and Data Mining. 1996; 169-194.
[12] Kailing K, Kriegel H and Kroger P. Density Connected Subspace Clustering for High - Dimensional Data. International Conference on Data Mining. 2004; 246-256.
[13] Huirong Zhang, Yan Tang, Ying He, Chunqian Mou, Pingan Xu and Jiaokai Shi. A novel subspace clustering method based on data cohesion. model Optik- International Journal for Light and Electron Optics. 2016;127(20): 8513–8519.
[14] Yi-Hong Chu, Jen-Wei Huang and Kun-Ta Chuang. Density Conscious Subspace Clustering for High-Dimensional Data. IEEE Transactions on Knowledge and Data Engineering. 2010; 22(1):16-30.
[15] Sequeira and Zaki M. SCHISM: A New Approach for Interesting Subspace Mining. The Proceedings of the Fourth IEEE Conference on Data Mining. 2004; 186–193.
[16] Maria Kontaki, Apostolos N. Papadopoulos and Yannis Manolopoulos. Continuous subspace clustering in streaming time series Information Systems. 2008; 33: 240–260.
[17] Agrawal R, Gehrke J, Gunopulos D and Raghavan P. Automatic subspace clustering of high dimensional data for data mining applications. Proceedings of the 1998 ACM SIGMOD international conference on Management of data.1998: 94-105.
[18] Mohammed J Zaki, Markus Peters, Ira Assent and Thomas Seidl. CLICKS: An effective algorithm for mining subspace clusters in categorical datasets. Data & Knowledge Engineering. 2007; 60: 51–70.
[19] MARY JEYA JOTHI, R. AND EBIN EPHREM ELAVANTHINGALN. ANALYZING THE REGULARITY OF COMPLETE K-PARTITE GRAPH USING SUPER STRONGLY PERFECT

GRAPHS. ONLINE INTERNATIONAL CONFERENCE ON GREEN ENGINEERING AND TECHNOLOGIES (IC-GET). 2015.

[20] Assent I, Krieger R, Muller E and Seidl T. INSCY: indexing subspace clusters with in-process-removal of redundancy.Proceeding of IEEE International Conference on Data Mining. 2008; 719–724.

[21] Yi-Hong Chu, Ying-Ju Chen, De-Nian Yang and Ming-Syan Chen. Reducing Redundancy in Subspace Clustering. IEEE Transactions on Knowledge and Data Engineering. 2009; 21(10): 1432-1446.

[22] Gabriela Moise, Arthur Zimek, Peer Kröger, Hans-Peter Kriegel and Jörg Sander. Subspace and projected clustering: experimental evaluation and analysis. Knowledge Information Systems. 2009; 21: 299–326.

[23] Brian McWilliams and Giovanni Montana. Subspace clustering of high-dimensional data: a predictive approach. Data Mining and Knowledge Discovery. 2014; 28:736–772.

[24] Amardeep Kaur and Amitava Datta. A novel algorithm for fast and scalable subspace clustering of high-dimensional data. Journal of Big Data. 2015;2(17).

[25] Hans-Peter Kriegel, Peer Kr¨oger, Matthias Renz and Sebastian Wurst. Generic Framework for Efficient Subspace Clustering of High-Dimensional Data. Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM'05). 2005; 250-257.

[26] JIAN YIN, ZHILAN HUANG AND JIAN CHEN. AN EFFECTIVE MAXIMAL SUBSPACE CLUSTERING ALGORITHM BASED ON ENUMERATION TREE. FOURTH INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS AND KNOWLEDGE DISCOVERY, HAIKOU, CHINA, 2007L; 572 - 576.

[27] Ma shuai, Wang tengjiao, Yang dongqing and Gao jun. A new fast clustering algorithm based on reference and density. International Conference on Web-Age Information Management (WAIM'03), Lecture Notes in Computer Science. 2003;214–225.

[28] Datasets from UCI Machine Learning Repository, [Online]. available at: http://archive.ics.uci.edu/ml

[29] Jaya Lakshmi B, Madhuri KB and Shashi M. An Efficient Algorithm for Density Based Subspace Clustering with Dynamic Parameter Setting. International Journal of Information Technology and Computer Science, 2017;6(4):27-33.

Authors' Profiles

B.Jaya Lakshmi received M.Tech. degree in Computer Science and Technology (Specialization-Artificial Intelligence & Robotics) from AU College of Engineering(A), Andhra University, Visakhapatnam in 2009. She was awarded Ph.D from JNTUK, Kakinada. Presently she is working as Associatet Professor in department of Information Technology at Gayatri Vidya Parishad College of Engineering(A), Visakhapatnam, Andhra Pradesh, India. Her research interests include Data Mining and Pattern Recognition. She published 9 research papers in International Journals. She obtained UGC minor research project No.F:MRP-4554/14(SERO/UGC) in 2014.

K.B. Madhuri received M.Tech. degree in Computer Science and Technology from Andhra University in 1999. She obtained Ph.D from JNTU, Hyderabad in 2009. Presently she is working as Professor and Head of the department in department of Information Technology at Gayatri Vidya Parishad College of Engineering (A), Visakhapatnam, Andhra Pradesh, India. Her research interests include Data Mining, Pattern Recognition, Data warehousing and RDBMS. She has guided one Ph.D scholar and currently guiding one Ph.D scholar. She published research papers in National and International Journals. She is a member of IEEE and associate member of Institute of Engineers (India).

How to cite this paper: B.Jaya Lakshmi, K.B.Madhuri,"Mining Maximal Subspace Clusters to deal with Inter-Subspace Density Divergence", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.5, No.3, pp.37-48, 2019. DOI: 10.5815/ijmsc.2019.03.04