Mining Maximal Subspace Clusters for High Dimensional Data

Singampalli Roopesh Kumar[1*], Marlapalli Krishna[2] and Nalla Sai Sheetal Naidu[3]

[1] M.Tech Student, Department of CSE, Sir C R Reddy College of Engineering, Eluru.*
[2] Professor, Department of CSE, Sir C R Reddy College of Engineering, Eluru.
[3] Assistant Professor, Department of CSE, GITAM deemed to be University, Visakhapatnam.
[1] s.roopeshkumar@outlook.com, [2] marlapallikrishna@gmail.com, [3] sheetal.n93@gmail.com .*

Abstract: Clustering is a process of automatically finding similar data points in the space of dimensions or attributes for a given data set and finding the clusters in the high dimensional datasets is an important and challenging data mining problem. Data set can be better understood by clustering in its subspaces, a process called 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.

This paper presents a Dynamic Epsilon based Subscale Algorithm (DESS) dealing with a unique problem of mining maximal subspace clusters in high dimensional data. A maximal subspace cluster is defined by maximal number of attributes. The mining algorithm involves four steps. In the first step, data points are assigned with unique positive integers called labels. In the second step, dense units are created based on the density notion which considers an input parameter called minimum points (τ), within the proposed dynamically calculated epsilon radius. The epsilon value is considered dynamically based on the data distribution. In the third step, sum of the labels of each data object forming the dense unit is calculated to be its signature and is hashed into the hash table. If a dense unit of a particular subspace collides with that of the other subspace, then both the dense units exists in the subspace formed by the colliding subspaces with the high probability. In other words, if the dense unit exists in both subspaces and there is high probability that it exists in higher subspaces formed by the union of colliding subspaces.

This process is repeated in single dimensional space and the dense units are hashed in the hash table. Maximal dense units are formed based on the colliding dimensions against each signature in the hash table. Then density reachable sets are identified from the existing dense units and hence, maximal subspace clusters are generated. The experiments are done on benchmark numeric datasets which are taken from UCI machine learning repository. The proposed algorithm Dynamic Epsilon based Subscale Algorithm (DESS) considers dynamic epsilon and has produced better quality maximal subspace clusters when compared to existing (SUBSCALE) algorithm. The time taken is less compared to SUBSCALE algorithm. Purity of DESS algorithm has increased when compared to SUBSCALE algorithm. Number of subspace clusters are less compared to the SUBSCALE algorithm. The percentage of increase in purity is by 0.3%, and the percentage of execution time is decreased by 0.5 %, and the percentage of decrease in number of subspace clusters is 0.02% on an average given data distributions.

Keywords: Clustering, Dynamic Epsilon based Subscale Algorithm, Dynamic Epsilon, Subspace Clusters.

1. INTRODUCTION

Finding out of interesting (Important, indirect ,possible useful and previously unknown) designs or knowledge from enormous aggregate data is called as Data mining [4] .Data mining can be called with different names, namely Business intelligence, information harvesting, data archeology, data dredging etc.

1.1 Cluster Analysis:

The procedure of grouping set of physical objects into classes of similar objects is called clustering [4]. A cluster is a group of data objects that are similar to one another within the same cluster and dissimilar to the objects in other clusters. Finding similarities between data objects according to the characteristics found in the data and grouping similar data objects into clusters is called as cluster analysis.

1.2 Quality of clustering:

Similarity is expressed in terms of distance function. It is a separate "quality" function that measures the "goodness" of a cluster. The definitions of distance functions usually varies different for Boolean, interval-scaled ordinal ratio, categorical, and vector variables [4].

1.3 Requirements of clustering in data mining:

Incorporation of user-specified constraints, ability to deal with different types of attributes, ability to handle dynamic data, scalability, able to deal with noise and outliers[15],minimal requirements for domain knowledge to determine input parameters, usability insensitive to order of input records, high dimensionality and interpretability .

1.4 Problem Definition:

There are two main disadvantages in bottom-up subspace clustering algorithms: one is finding of redundant trivial clusters which requires excessive number of database scans. The second problem arises during the iterative bottom-up process. Combining lower dimensional candidate clusters, multiple database scans are required for determining the occupancy of each and every data point while merging these candidate clusters where, |DB| is the size of the dataset [7].

A new subspace clustering algorithm is introduced in order to overcome these both inefficiencies and has a high degree of parallelism is DYNAMIC EPSILON based SUBSCALEALGORITHM (DESS), helps in computing the maximal subspace clusters. In this approach assign signatures [19] to each 1-dimensional clusters such that their collisions will help in identifying the maximal subspace clusters without generation of intermediate clusters. DYNAMIC EPSILON based SUBSCALE ALGORITHM (DESS) requires only k-database scans to process a k-dimensional dataset and it is more scalable with the dimensions as compared to SUBSCALE algorithm.

2. LITERATURE REVIEW

The foremost data mining task is Clustering. Datasets are distinguished by high-dimensional infrequent data space. Often data sets fail to identify the significant clusters in traditional clustering algorithms. Sometimes datasets contains concealed clusters in various subspaces of the native feature space. Many applications like geography, molecular biology, finance and marketing produces enormous amounts of data which cannot be analyzed without the help of data mining methods. Analyzing the data for large amounts of data is bit difficult task. Consequently, subspace clustering concept has been recently adopted from clustering. The main aim of subspace clustering is repeatedly identifying subspaces of higher dimensional space in which clusters exist. Subspace Clustering identifies set of objects that are same in subspaces of high-dimensionality datasets. Subspace clustering is the task of finding all clusters in subspaces [7].

Based on this strategy of subspace clustering there are two approaches. The first approach of subspace clustering is top-down approach which finds initial clusters in full set of dimensions and evaluates the subspaces of each cluster, iteratively. The second approach of subspace clustering is bottom-up approach. In this approach dense regions are identified in low-dimensional spaces and these dense regions are combined to form clusters in higher dimensions.

Traditional clustering algorithms use the entire data space to find clusters in full-dimensional space. One of them is DBSCAN (Density Based Spatial Clustering of Applications with Noise) is a full-dimensional clustering algorithm, a point is said to be dense if it has τ or more points in its ε-neighborhood. But there arises a problem i.e. the Curse of dimensionality. Curse of dimensionality is defined as number of dimensions increases, some of the dimensions becomes irrelevant, so that the cluster obtained may not be meaningful in the higher-dimensionality. Different techniques are available to find clusters in high dimensionality i.e Principal Component Analysis (PCA). In this technique transform the high dimensional space into lower dimensional space. Principal Component Analysis does not change the original variance of the full-dimensional data during this transformation, suppose, if cluster was detected in the original dimensions, no clusters are not obtained in the transformed dimensions. The transformed (categorical) [20] dimensions lack the real meaning and it is difficult to interpret the clusters found in the new dimensions with respect to the original data space. Dimensionality reduction is not possible always. Sometimes user may remove the data which is irrelevant to form cluster. By doing this the user might miss useful data or may consider the noise data which does not form clusters in higher dimensionality [7] data.

There are two important properties need to considered in subspace clustering algorithms. One is data set points, participate to form clusters and the other is data points combine differently based on set of attributes.

3. EXISTING SYSTEM

In traditional clustering methods, finding of clusters is a bit difficult task. Some of the traditional algorithms are k-means, K-mediods etc. By using this algorithms clusters are obtained in full dimensional space. But they might not be true clusters because some of the clusters might miss due to latent data distribution. As numbers of dimensions increases finding the similarity between the points becomes difficult. This is the main disadvantage of traditional clustering approach. This is called as curse of dimensionality.

In order to overcome curse of dimensionality, different methods are introduced. One of the methods is dimensionality reduction. Mapping is done with the lower dimensions so that data can be better understood and working of it will be efficient. But there are two main disadvantages in it. One is transformed attributes often have no meaning anymore and thus leading clusters are hard to explain. Second, using dimensionality reduction techniques, the data is clustered only in a particular subspace. Third, dimensionality reduction does not give required results because of certain reasons. One of them as the data is interpreted in some other way rather than original form the data might be missed or wrongly interpreted.

Another approach is introduced i.e. projected clustering, while the dimensionality reduction is not flexible. Projected clustering is used for finding the clusters. Even though it finds clusters it suffers from the problem that the information of objects which are clustered differently in varying subspaces. So, the information might get lost. The above problem is illustrated in the following figure.

In figure 1 the problem using a feature space of four attributes A, B, C and D. The objects 1 and 2 in the subspace AB together with the objects 3 and 4, where as in the subspace CD they cluster with objects 5 and 6. Either the information of the cluster in subspace AB or in subspace CD will be lost.

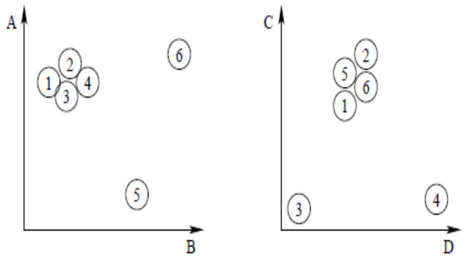

Fig 1: Projected Clustering drawbacks

To overcome problem of projected clustering approach mentioned above, a new approach is introduced i.e. SUBSCALE approach is introduced.

4. PROPOSED SYSTEM:

4.1 DYNAMIC EPSILON based SUBSCALE ALGORITHM (DESS):

Clustering is a process of finding clusters with respect to attributes. Different approaches are introduced to find the clusters. Traditional clustering algorithms are introduced to find the clusters. But there arise some disadvantages. As explained in SUBSCALE algorithm epsilon is user defined parameter .In order to overcome the disadvantages a new approach is introduced. DYNAMIC EPSILON based SUBSCALE ALGORITHM(DESS) introduced.

4.1.1 Preliminary Definitions:

Core Object: It is defined as an object which is within epsilon distance and τ neighbors. An epsilon distance is calculated by a formulae i.e maximum object value subtracted with minimum object value divided by total number of objects. τ is considered according to the project.

Dense Unit: A core object is said to be dense if it satisfies the τ and epsilon.

Signature Sum: Adding the label values of each object to get the signature of dense units.

Architecture Flow: The Block Diagram of DESS Algorithm is shown in Figure 2.

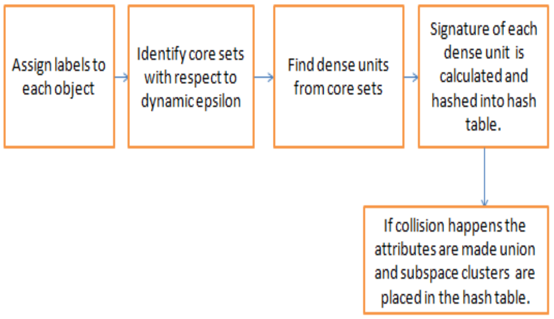
Assign labels to each object
Identify core sets with respect to dynamic epsilon
Find dense units from core sets
Signature of each dense unit is calculated and hashed into hash table.
If collision happens the attributes are made union and subspace clusters are placed in the hash table.

Fig 2: Block Diagram of DESS Algorithm

According to Apriori principle, a set of dense points in an k-dimensional space *S* is dense in all the lower dimensional projections of the space. In other words, having the dense sets of points in 1-dimensional projections of given data, then the common points among the 1-dimensional sets will lead to the dense points in the higher dimensional subspaces. Based on this approach, subscale algorithm efficiently finds the maximal clusters in all possible subspaces of a high-dimensional dataset. Density is based on two defined parameters *ε* and *τ*, a point is considered as dense if it has at least *τ* points within *ε* distance. A point P dense in a subspace *S*, if it has at least *n*eighbors within *ε* distance. These dense points can be easily connected to form a subspace cluster. Epsilon value is calculated dynamically in the algorithm. Epsilon formulae is represented in the following equation 1.

$$\varepsilon = Distance\ between\ the\ farthest \frac{points}{|D|} \ldots (1)$$

Where |D| represents total number of objects in a data set.

The theme of DESS (Dynamic Epsilon based Subscale algorithm) is to find maximal subspace clusters by finding the dense units in the applicable subspaces of given data distribution using certain parameters. The parameters are epsilon and minimum points. In Dynamic Epsilon based Subscale algorithm (DESS),epsilon value is calculated dynamically. Epsilon varies according to the data set distribution. The formula is defined as distance between the farthest points to the total number of objects of a data set. In the present algorithm epsilon value differs from one attribute to the other attribute. A dense unit can be defined as a core object which satisfies the τ and epsilon. Dense unit is represented with U.

Dense unit (U) is the smallest cluster. If |CS| is the number of dense points in a 1-D core-set CS then ($|CS|_{\tau+1}$) dense units from one such CS. Create a hash table (**hTable)**. For every dense unit, compute its **signature (H_i)** by calculating the sum of the labels of each object. If the signature of the dense unit (U) 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 which contain dense units in the maximal subspace.Repeat the process in all single dimensions. Obtained dense units are processed to create density-reachable maximal clusters. This is how dynamic epsilon subscale algorithm works.

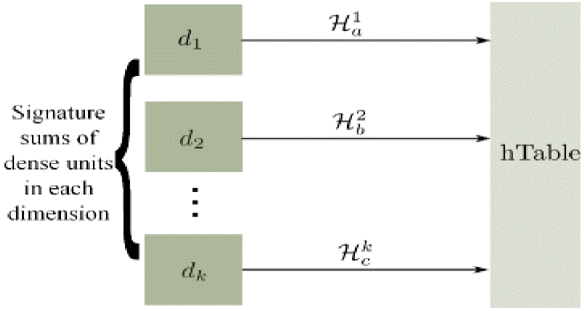

Fig 3: Collisions among signatures.

The above figure 3 represents signatures of different dimensions collide to identify the relevant subspaces. d_i is the ith dimension and H_{ix} is dense unit signature.

4.1.2 Algorithm Steps:

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

STEP-2: CS be a Core-Set such that each object is within dynamic ε distance containing at least minimum points (τ).

$$\varepsilon = \frac{Distance\ between\ the\ two\ \ farthest\ points}{|D|} \ldots 2)$$

Where |D| represents total number of objects in a data set. With the Core Object sets identify the dense units using the formula $cs_{C_{\tau+1}}$

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

- 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 which contain dense units in the maximal subspace.
- Repeat the process in all single dimensions.

STEP-4: Obtained dense units are processed to create density-reachable maximal clusters.

The performance of Dynamic Epsilon based Subscale algorithm (DESS) mainly depends on generated dense units in single dimension. Dense units which are generated depends on ε and size of data distribution. In the algorithm ε value is generated dynamically from the data set. If a small value is obtained for ε then the number of core sets will be less then automatically dense units will be less. A larger ε value increases the core sets and thus leading to increase in dense units. The number of subspace clusters varies according to the situation.

Information about the previous of latent data is not known, one dimension can have any number of dense units. Thus, for an expandable solution to subspace clustering through DESS, the system must be able to process and store more collisions of dense units very effectively. To identify collisions among dense units across multiple dimensions, hash table is needed big enough to hold the dense units in system memory [9].

Identification of maximal dense units across multiple dimensions involves matching of same signature using a hash tale.

5. RESULTS AND ANALYSIS

Dataset Description: 3 data sets from UCI machinery learning repository[21] are considered**.**

1. **Seed data** which comprised of three different varieties of wheat namely: Karma, Rosa, Candian.
   7 Attributes namely: length of kernel groove, perimeter, width kernel, length of kernel area, asymmetry coefficient, compactness, length of kernel.
2. **Image Segmentation data** comprised of six different varieties of classes.
   9 Attributes namely: region-centroid-col, region-centroid-row, pixel count, Short line density, vedge mean, intensity mean, hedge mean, exred mean, rawred mean, Value mean.
3. **Bank Authentication Data** which comprised of two varieties of classes are used one is industrial camera, and wavelet transform tool
   4 attributes namely: variance image, skewness image, entropy and curtosis.

Purity: Purity is an external evaluation criteria of cluster quality. It is defined as percentage of total number of objects that were classified correctly. Range of purity is between [0, 1].

Purity Formula:

$$Purity = \frac{1}{N\sum_{1=1}^{k} max_j \left|c_i \cap t_j\right|} \ldots\ldots\ldots\ldots..(3)$$

Where N = number of objects (data points), k= number of clusters, c_i is a cluster in C and t_j is the classification which has the max count for cluster $c_{i.}$.

The experiments are conducted on the following data sets and results are tabulated.

Table 1: Comparison of DESS and SUBSCALE with respect to execution time, number of sub space clusters, purity.

Data (#Objects, #Attributes, #classes)	DESS Algorithm			Subscale Algorithm		
	Execution Time	*Number Of Sub space Clusters obtained*	*Purity*	*Execution Time*	*Number Of Sub space Clusters obtained*	*Purity*
Seed Data (210, 7, 3)	16 min 26 sec	106	0.923	79 min 8 sec	109	0.90
Image Segmentation Data (400, 9, 6)	187 min	11	0.922	412 min 42 sec	21	0.91
Bank Authentication Data (1372, 4, 2)	512 min 23 sec	34	0.927	1023 min	43	0.89

- ➢ Comparison of DESS and SUBSCALE with respect to epsilon

By considering the dynamic epsilon time taken to complete the task is less compared to the static epsilon

Table 2: Comparison of DESS and SUBSCALE with respect to execution time

Number of objects	Execution Time	
	DESS Algorithm	*Subscale Algorithm*
50	2 min 6 sec	48min 21 sec
100	3 min 23 sec	72 min 2sec
150	7 min 19 sec	112 min 56 sec
200	17 min20 sec	157min 55 sec

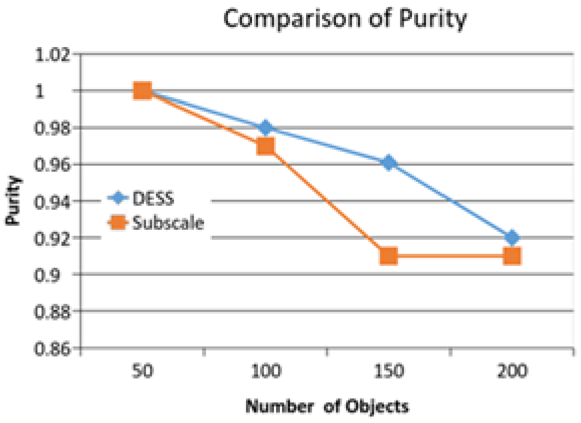

Fig 4: Comparison of DESS and SUBSCALE with respect to execution time

- Comparison of DESS and SUBSCALE with respect to purity

The efficiency of algorithm is more in dynamic way rather than in static way, the epsilon value is considered according to the user the purity cannot be obtained efficiently compared to dynamic value epsilon.

Table 3: Comparison of DESS and SUBSCALE with respect to Purity

Number of Objects	Purity	
	DESS Algorithm	Subscale Algorithm
50	1.0	1.0
100	0.985	0.97
150	0.961	0.91
200	0.92	0.91

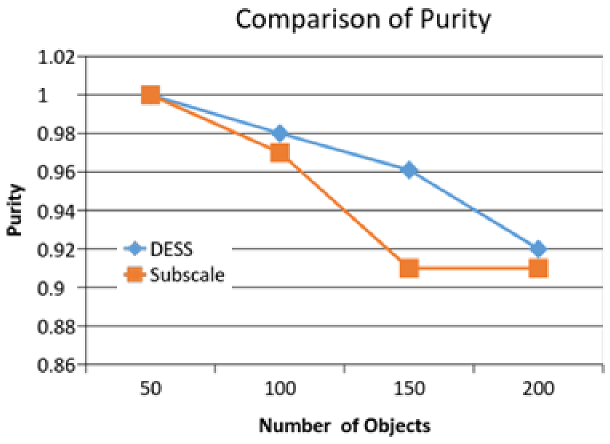

Fig 5: Comparison of DESS and SUBSCALE with respect to Purity

The number of subspace clusters obtained is more in static epsilon rather than dynamic epsilon.

Table 4: Comparison of DESS and SUBSCALE with respect to Number of subspace clusters

#Objects	Number of Subspace Clusters Obtained	
	DESS Algorithm	Subscale Algorithm
50	65	66
100	91	99
150	95	120
200	106	130

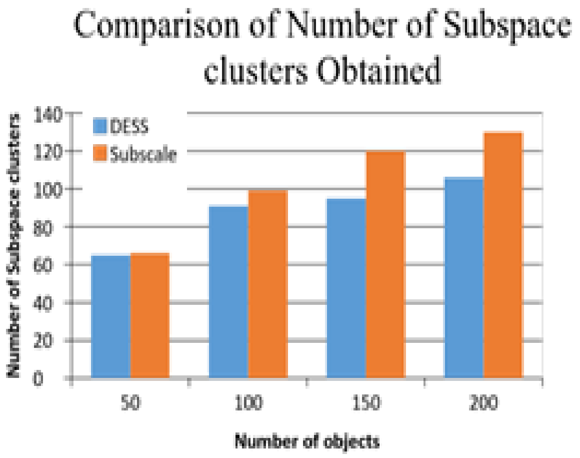

Fig6: Column Chart representation of number of subspace clusters with increase in number of objects.

6. CONCLUSION

Subspace Clustering is defined as finding the clusters with respect to subset of attributes. The main purpose of this algorithm is that finding clusters with respect to high dimensional data. If the number of dimensions increases identifying clusters becomes difficult. In order to overcome this problem Dynamic Epsilon based Subscale algorithm is proposed. It finds the clusters in all single dimensions, and finds the clusters in higher dimensional data with high probability. The execution time, purity, number of subspace clusters are the metrics used for comparison. While comparing the execution time. The time taken is less compared to SUBSCALE algorithm. Purity of DESS algorithm is increased when compared to SUBSCALE algorithm. Number of subspace clusters are less compared to the SUBSCALE algorithm. The percentage of increase in purity is by 0.3%, and the percentage of execution time is decreased by 0.5 %, and the percentage of decrease in number of subspace clusters is 0.02% on an average given data distributions.

REFERENCES

[1] C.C. Aggarwal, J.L. Wolf, P.S. Yu, C. Procopiuc, J.S. Park, “Fast algorithms for projected clustering”, *InProceedings of 1999 ACM-SIGMOD International Conference on Management of Data (SIGMOD’99), Philadelphia, PA,* pp. 61–72, June 1999.

[2] C.C. Aggarwal, P.S. Yu, “Finding generalized projected clusters in high dimensional spaces”, *In Proceedings of 2000 ACM-SIGMOD International Conference on Management of Data (SIGMOD’00), Dallas, TX,* pp. 70–81, May 2000.

[3] H.Nagesh, S.Goil, A.Choudhary, Mafia: “Efficient and scalable subspace clustering for very large datasets “, *Technical Report9906-010,* June1999.

[4] Jiawei Han, Micheline kamber : Data Mining: Concepts and Techinques, Second Edition(The Morgan Kaufmann Series in Data Management System)

[5] J.Yang, W.Wang, H.Wang, P.S.Yu, d-cluster: “Capturing subspace correlation in a large dataset”, *In Proceedings of 2002 International Conference on Data Engineering (ICDE’02),SanFransisco,CA,*April2002.

[6] K.-G.Woo,J.-H.Lee,M.-H.Kim,Y.-J.Lee, Findit: “A fast and intelligent subspace clustering algorithm using dimension voting”, *Inf.Software Technol.*Volume:46 Issue:4,pp.255–271, June 2004.

[7] L.Parsons, E.Haque, H.Liu, “Subspace clustering for high dimensional data”: a review, *SIGKDD Explor.Newsl.*Volume: 6Issue:1,pp.90–105,June 2004.

[8] Ming Hua , Jian Pei “Clustering in applications with multiple data sources—A mutual subspace clustering approach” *Neurocomputing,*Volume:92Issue :1,pp.133–144, September 2012.

[9] R.Rymon, “Search through systematic set enumeration”, *In Proceedings of 1992 International Conference on Principle of Knowledge Representation and Reasoning (KR’92),Cambridge, MA*, pp.539–550,1992.

[10] R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications”, *In Proceedings of 1998 ACMSIGMOD International Conference on Management of Data (SIGMOD’98), Seattle, WA*, pp. 94–105,June 1998.

[11] K. Kailing, H.P. Kriegel and P. Kroger,“Density-connected subspace clustering for high dimensional data”, *In Proceedings of the 4th SIAM International Conference on Data Mining, Orlando, FL,* pp. 46-257, (2004).

[12] Jyoti Yadav, Dharmender Kumar “(Subspace Clustering Using CLIQUE: An Exploratory study”, *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET),* Volume:3 Issue:2, pp. 372-378, (2014)

[13] Y.Chai,N.Cercone,and J.Han, “Attribute-oriented induction in relational Databases” G.Piatetsky-Shapiro and W.J.Frawley(eds.),Knowledge Discovery in Databases,pp.213-228.MIT Press,1991

[14] Z.A.Bakar, R.Mohamad,A.Ahamad,and M.MDeris “A comparative study for outlier detection techniques in data mining in data mining”. *In proc.*2006 IEEE *conf.Cybernetics andIntelligent Systems*,pp.1-6,*Bangkok,Thailand*,2006.

[15] Z.Borgelt and M.R Berthhold.Mining molecular fragments: “Finding relevant subspaces.”*In Proc.*2002 *Int.Conf.Data Mining(ICDM’02),*pp.211-218,Japan,Dec.2002.

[16] Z.Chakrabarti, R.Kumar,and A.Tomkins, “High dimension clustering techniques,”*In Proc.2006 ACM SIGKDDInt.Conf.KnowledgeDiscoveryinDatabases(KDD’06)* ,pp.554-560,Philadelphia,PA,Aug.2006.

[17] Z.Arabie,L.J.Hubert and G.De Soete.”*Clustering and classification*.”World Scientific,1996.

[18] Z. Boneh, B. Lynn and H. Shacham, Short Signatures from the Weil Pairing,*Proc. ASIACRYPT*, volume:22,Issue 48, pp. 514 ‐ 532, 2001.

[19] Xueqi Cheng, Liang Bai Jiye Liang and Huawei Shen,“An Optimization Model for Clustering Categorical Data Streams with Drifting Concepts”, IEEE Transactions on Knowledge and Data Engineering, Volume 28, Issue 11, pp 2871,2016.

[20] UCI Machine Learning Repository: Data Sets. https://archive.ics.uci.edu/ml/datasets.html.