A Novel Dencos Model For High Dimensional Data Using Genetic Algorithms

T. Vijayakumar[1], V.Nivedhitha[2], K.Deeba[3] and M. Sathya Bama [4]

[1]*Assistant professor / Dept of IT, Dr.N.G.P College of Engineering & Technology*
[2]*Assistant professor / Dept of CSE, Akshaya College of Engineering & Technology*
[3]*Assistant professor / Dept of CSE, Akshaya College of Engineering & Technology*
[4]*Assistant professor / Dept of CSE, Akshaya College of Engineering & Technology*

***ABSTRACT* -** *Subspace clustering is an emerging task that aims at detecting clusters in entrenched in subspaces. Recent approaches fail to reduce results to relevant subspace clusters. Their results are typically highly redundant and lack the fact of considering the critical problem, "the density divergence problem," in discovering the clusters, where they utilize an absolute density value as the density threshold to identify the dense regions in all subspaces. Considering the varying region densities in different subspace cardinalities, we note that a more appropriate way to determine whether a region in a subspace should be identified as dense is by comparing its density with the region densities in that subspace. Based on this idea and due to the infeasibility of applying previous techniques in this novel clustering model, we devise an innovative algorithm, referred to as DENCOS(DENsity Conscious Subspace clustering), to adopt a divide-and-conquer scheme to efficiently discover clusters satisfying different density thresholds in different subspace cardinalities. DENCOS can discover the clusters in all subspaces with high quality, and the efficiency significantly outperforms previous works, thus demonstrating its practicability for subspace clustering. As validated by our extensive experiments on retail dataset, it outperforms previous works. We extend our work with a clustering technique based on genetic algorithms which is capable of optimizing the number of clusters for tasks with well formed and separated clusters.*

Key words: *data clustering, subspace clustering, data mining, density divergence problem*

I. INTRODUCTION

Among recent studies on high-dimensional data clustering, subspace clustering is the task of automatically detecting clusters in subspaces of the original feature space. Most of previous works take on the density-based approach, where clusters are regarded as regions of high density in a subspace that are separated by regions of lower density.

However, a critical problem, called "the density divergence problem" is ignored in mining subspace clusters such that it is infeasible for previous subspace clustering algorithms to simultaneously achieve high precision. "The density divergence problem" refers to the phenomenon that the cluster densities vary in different subspace cardinalities.

Due to the loss of the distance discrimination in high dimensions, discovering meaningful, separable clusters will be very challenging, if not impossible. A common approach to cope with the curse of dimensionality problem for mining tasks is to reduce the data dimensionality by

DOI : 10.5121/ijcseit.2012.2104

using the techniques of feature transformation and feature selection. The feature transformation techniques, such as principal component analysis (PCA) and singular value decomposition (SVD), summarize the data in a fewer set of dimensions derived from the combinations of the original data attributes.

The transformed dimensions have no intuitive meaning anymore and thus the resulting clusters are hard to interpret and analyze. They also reduce the data dimensionality by trying to select the most relevant attributes from the original data attributes. Motivated by the fact that different groups of points may be clustered in different subspaces, a significant amount of research has been elaborated upon subspace clustering, which aims at discovering clusters embedded in any subspace of the original feature space. The applicability of subspace clustering has been demonstrated in various applications, including gene expression data analysis, E-commerce, DNA microarray analysis, and so.

To extract clusters with different density thresholds in different cardinalities is useful but is quite challenging. Previous algorithms are infeasible in such an environment due to the lack of monotonicity property. That is, if a k-dimensional unit is dense, any (k −1)-dimensional projection of this unit may not be dense. A direct extension of previous methods is to execute a subspace algorithm once for each subspace cardinality k by setting the corresponding density threshold to find all k-dimensional dense units.

However, it is very time consuming due to repeated execution of the targeted algorithm and repeated scans of database. Due to the requirement of varying density thresholds for discovering clusters in different subspace cardinalities, it is challenging for subspace clustering to simultaneously achieve high precision and recall for clusters in different subspace cardinalities.

A naive method to examine all regions to discover the dense regions, we devise an innovative algorithm, referred to as “DENsity COnscious Subspace clustering” (abbreviated as DENCOS), to efficiently discover the clusters satisfying different density thresholds in different subspace cardinalities. In DENCOS, the mechanism of computing the upper bounds of region densities to constrain the search of dense regions is devised, where the regions whose density upper bounds are lower than the density thresholds will be pruned away in identifying the dense regions.

We compute the region density upper bounds by utilizing a novel data structure, DFP-tree (Density FP-tree), where we store the summarized information of the dense regions. In addition, from the DFP-tree, we also calculate the lower bounds of the region densities to accelerate the identification of the dense regions. Therefore, in DENCOS, the dense region discovery is devised as a divide-and-conquer scheme. At first, the information of region density’s lower bounds is utilized to efficiently extract the dense regions, which are the regions whose density lower bounds exceed the density thresholds. Then, for the remaining regions, the searching of dense regions is constrained to the regions whose upper bounds of region densities exceed the density thresholds.

II. RELATED WORK

The problem of finding different clusters in different subspaces of the original input space has been addressed in many papers. Subspace Clustering is a very important technique to seek clusters hidden in various subspaces (Dimensions) in a very high dimensional database. There are very few approaches to Subspace Clustering. These approaches can be classified by the type of results they produce.

The first class of algorithms allows overlapping clusters, i.e., one data point or object may belong to different clusters in different projections e.g. CLIQUE, ENCLUS, MAFIA, SUBCLU and FIRES. The second class of subspace clustering algorithms generate non-overlapping clusters and assign each object to a unique cluster or noise e.g. DOC and PreDeCon.

The first well-known Subspace Clustering algorithm is CLIQUE, CLUstering in QUEst. CLIQUE is a grid-based algorithm, using an apriori-like method which recursively navigates through the set of possible subspaces. A slight modification of CLIQUE is the algorithm ENCLUS, Entropy based CLUStering. A more significant modification to CLIQUE is MAFIA, Merging of Adaptive Finite IntervAls, which is also a grid-based but uses adaptive, variable sized grids in each dimension.

The major disadvantage of all these techniques is caused by the use of grids. Grid-based approaches are based on positioning of grids. Thus clusters are always of fixed size and depend on orientation of grid. Density based Subspace Clustering is one more approach. The first of this kind, DOC proposes a mathematical formulation regarding the density of points in subspaces. But again, the density of subspaces is measured using a hypercube of fixed width w, so it has the similar problems. Another approach SUBCLU (density connected SUBspace CLUstering) is able to effectively detect arbitrarily shaped and positioned clusters in subspaces. Compared to the grid-based approaches SUBCLU achieves a better clustering quality but requires a higher runtime.

SURFING is one more effective and efficient algorithm for feature selection in high dimensional data. It finds all subspaces interesting for clustering and sorts them by relevance. But it just gives relevant subspaces for further clustering. The only approach which can find subspace cluster hierarchies is HiSC. However it uses the global parameters such as Density Threshold (μ) and Epsilon Distance (ε) at different levels of dimensionalities while finding subspace clusters. Thus its results are biased with respect to the dimensionality.

III. PROBLEM STATEMENT

Problem Statement:-

Given the unit strength factor α and the maximal subspace cardinality k_{max}, for the subspaces of cardinality k from 1 to k_{max}, find the clusters in which each is a maximal set of connected dense k-dimensional units whose unit counts are larger than or equal to the density threshold τ_k.

Theorem 1:-

Subspace clusters have the downward closure property on the attribute set.

Theorem 2:-

Subspace clusters have the downward closure property on the object set.

Theorem 3:-

Given a set of subspace clusters *G*={<*Ci*, *Si*>}, construct an object set *C* and an attribute set *S* by *C*=∩*Ci*, *S*=∪*Si*, then <*C*, *S*> is a subspace cluster if |*C*|≥*coverage* · *n*.

Theorem 4:-

For all subspace clusters in *derive*(*G*), the representative subspace cluster <*C*, *S*> of *G* has the largest set of *rep*(<*C*, *S*>).

Theorem 5:-

In path removal technique, the two steps of the path reconstruction process, i.e., path exclusion and path reorganization, can correctly prepare the paths for performing the path removal.

Definition: Density thresholds:-

Let τ_k denote the density threshold for the subspace cardinality k, and let N be the total number of data points. The density threshold τ_k is defined as:

$$\tau_k=\alpha \frac{N}{\delta^k}$$

IV. DENCOS TECHNIQUE

In DENCOS, we model the problem of density conscious subspace clustering to a similar problem of frequent itemset mining. By regarding the intervals in all dimensions as a set of unique items in frequent itemset mining problem, any k-dimensional unit can be regarded as a k-itemset, i.e., an itemset of cardinality k. Thus, to identify the dense units satisfying the density thresholds in subspace clustering is similar to mine the frequent itemsets satisfying the minimum support in frequent itemset mining. However, our proposed density conscious subspace clustering problem is significantly different from the frequent itemset problem since different density thresholds are utilized to discover the dense units in different subspace cardinalities, and thus, the frequent itemset mining techniques cannot be adopted here to discover the clusters.

The monotonicity property no longer exists in such situations, that is, if a k-dimensional unit is dense, any *(k - 1)* - dimensional projection of this unit may not be dense. Therefore, the a-priori-like candidate generate-and-test scheme, which is adopted in most previous subspace clustering works, is infeasible in our clustering model. Besides, the FP-tree-based mining algorithm, FP-growth proposed to mine the frequent itemsets, cannot be adopted to discover the clusters with multiple thresholds. However, because the units in the enumeration of the combinations of the nodes in the path are of various cardinalities, they may not all be dense units in this problem, thus showing the infeasibility of the FP-growth algorithm in our subspace clustering model. For this challenge, we devise the DENCOS algorithm in this paper to efficiently discover the clusters with our proposed density thresholds.

Algorithm:

Step 1: Start

Step 2: Select the dataset

Step 3: Discover the dense units by Density FP-Tree

Step 4: Group the connected dense units into clusters

Step 5: Compute the lower bounds and upper bounds of the unit counts for accelerating the dense unit discovery from the DFP Tree

Step 6: Mine the dense units using the divide and conquer scheme

Step 7: End

DENCOS is devised as a two-phase algorithm comprised of the preprocessing phase and the discovering phase. The preprocessing phase is to construct the DFP-tree on the transformed data set, where the data set is transformed with the purpose of transforming the density conscious subspace clustering problem into a similar frequent itemset mining problem. Then, in the discovering phase, the DFP-tree is employed to discover the dense units by using a divide-and-conquer scheme.

1. Preprocessing Dataset

In this phase, we first transform the data set by transforming each d-dimensional data point into a set of d one-dimensional units, corresponding to the intervals within the d dimensions it resides in. the DFP-tree is constructed to condense the transformed data set. In this paper, we devise the DFP-tree by adding the extra feature in the FP-tree for discovering the dense units with different density thresholds. In this paper, we propose to compute the upper bounds of unit counts for constraining the searching of dense units such that we add extra features into the DFP-tree for the computation. The DFP-tree is constructed by inserting each transformed data as a path in the DFP-tree with the nodes storing the one-dimensional units of the data. The paths with common prefix nodes will be merged and their node counts are accumulated.

2. Generate and discover Inherent Dense Units

In this discovering stage, we consider to utilize the nodes satisfying the thresholds to discover the dense units. For the nodes with node counts satisfying the thresholds for some set of subspace cardinalities, we will take their prefix paths to generate the dense units of their satisfied subspace cardinalities. However, a naive method to discover these dense units would require each node to traverse its prefix path several times to generate the dense units for the set of satisfied subspace cardinalities. In this paper, we have explored that the set of dense units a node requires to discover from its prefix path can be directly generated by utilizing the dense units discovered by its prefix nodes, thus avoiding the repeated scans of the prefix paths of the nodes. Therefore, by a traversal of the DFP-tree, we can efficiently discover the dense units for all nodes satisfying the thresholds.

3. Generate and discover Acquired Dense Units

In this stage, for the nodes whose node counts do not exceed τ_k, we take the nodes carrying the same one dimensional unit together into consideration in discovering the k-dimensional dense units. Note that the surplus count SC_u^k is the maximal possible unit count of the units that can be generated from the prefix paths of the nodes inN_u^k, which is the case when a unit can be derived from all these paths so that this unit has the unit count equal to the summation of the node counts of the nodes inN_u^k, i.e., SC_u^k . Clearly, if SC_u^k < τ_k, there will be no k-dimensional dense units that can be discovered from the prefix paths of the nodes in N_u^k, such that we need not apply the discovery process on N_u^k, to explore the k-dimensional dense units.

4. Implement path removal techniques

In path removal technique, the two steps of the path reconstruction process, i.e., path exclusion and path reorganization, can correctly prepare the paths for performing the path removal.

In the step 1, path exclusion, the paths in PathList cannot exist in current u'_hconditional pattern base if they do not contain the set of one-dimensional units in I, i.e., u'_h. Because u'_h conditional pattern base is constructed by extracting the prefix paths of the u'_h nodes in the DFP-tree TreeB, the paths in PathList which do not contain u'_h cannot exist in u'_h conditional pattern base.

The step 2, "path reorganization," in the reconstruction process is to reorganize the remaining paths in PathList to the form they should appear in u'_h conditional pattern base. Note that u'_h conditional pattern base is the set of prefix paths of u'_h nodes in TreeB.

5. Genetic Optimization Algorithm

We enhance our research with Genetic optimization algorithm for generate better dense region and also to discover the clusters in different subspace cardinalities. The genetic algorithm is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover. The genetic algorithm gives better accuracy than previous work.

Algorithm:

Step 1: Start

Step 2: Generate initial population of candidate solutions

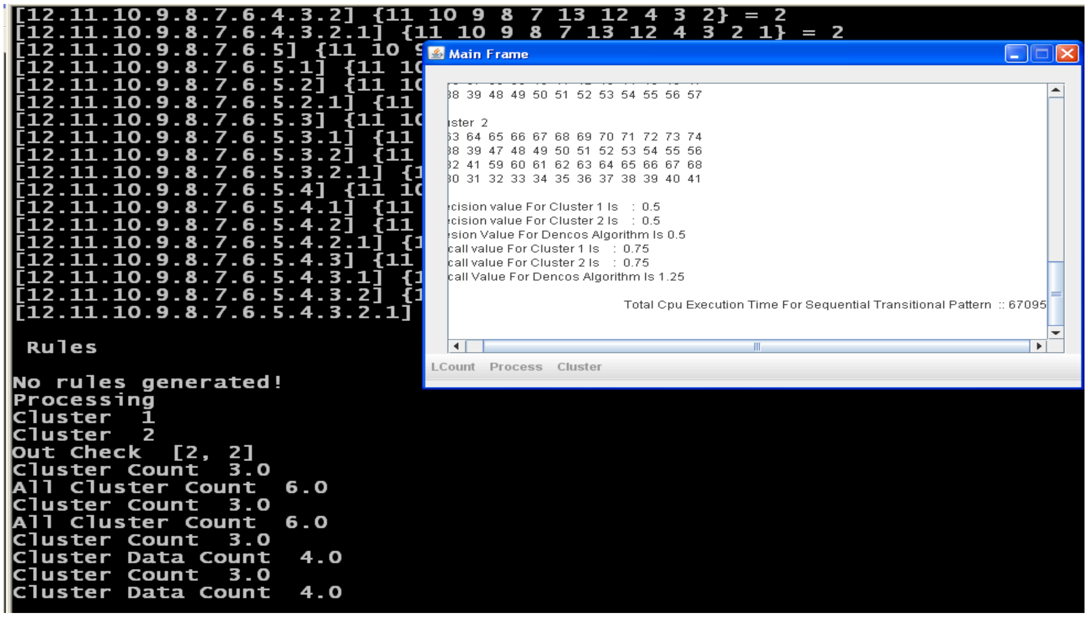
Step 3: Apply fitness function to population members

Step 4: Choose the fittest member to form the new population

Step 5: Apply genetic operators and generate new population

Step 6: Stop

V. EXPERIMENTAL RESULTS

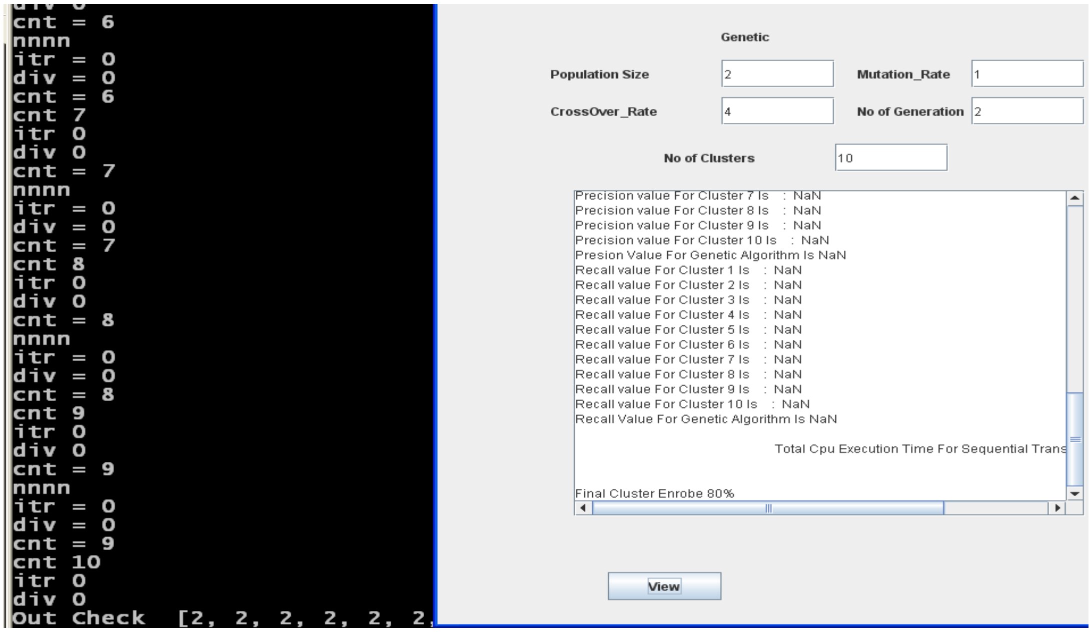


```
[12.11.10.9.8.7.6.4.3.2] {11 10 9 8 7 13 12 4 3 2} = 2
[12.11.10.9.8.7.6.4.3.2.1] {11 10 9 8 7 13 12 4 3 2 1} = 2
[12.11.10.9.8.7.6.5] {11 10 9
[12.11.10.9.8.7.6.5.1] {11 10
[12.11.10.9.8.7.6.5.2] {11 10
[12.11.10.9.8.7.6.5.2.1] {11
[12.11.10.9.8.7.6.5.3] {11 10
[12.11.10.9.8.7.6.5.3.1] {11
[12.11.10.9.8.7.6.5.3.2] {11
[12.11.10.9.8.7.6.5.3.2.1] {1
[12.11.10.9.8.7.6.5.4] {11 10
[12.11.10.9.8.7.6.5.4.1] {11
[12.11.10.9.8.7.6.5.4.2] {11
[12.11.10.9.8.7.6.5.4.2.1] {1
[12.11.10.9.8.7.6.5.4.3] {11
[12.11.10.9.8.7.6.5.4.3.1] {1
[12.11.10.9.8.7.6.5.4.3.2] {1
[12.11.10.9.8.7.6.5.4.3.2.1]

 Rules

No rules generated!
Processing
Cluster  1
Cluster  2
Out Check  [2, 2]
Cluster Count  3.0
All Cluster Count  6.0
Cluster Count  3.0
All Cluster Count  6.0
Cluster Count  3.0
Cluster Data Count  4.0
Cluster Count  3.0
Cluster Data Count  4.0
```

```
38 39 48 49 50 51 52 53 54 55 56 57

ister 2
63 64 65 66 67 68 69 70 71 72 73 74
38 39 47 48 49 50 51 52 53 54 55 56
32 41 59 60 61 62 63 64 65 66 67 68
30 31 32 33 34 35 36 37 38 39 40 41

ecision value For Cluster 1 Is   :  0.5
ecision value For Cluster 2 Is   :  0.5
sion Value For Dencos Algorithm Is 0.5
call value For Cluster 1 Is   :  0.75
call value For Cluster 2 Is   :  0.75
call Value For Dencos Algorithm Is 1.25

                         Total Cpu Execution Time For Sequential Transitional Pattern  :: 67095
```

Figure 1. Outcome of DENCOS Algorithm

Figure 1 and 2 shows the results of DENCOS and enhanced DENCOS with GENETIC algorithms.



```
div 0
cnt = 6
nnnn
itr = 0
div = 0
cnt = 6
cnt 7
itr 0
div 0
cnt = 7
nnnn
itr = 0
div = 0
cnt = 7
cnt 8
itr 0
div 0
cnt = 8
nnnn
itr = 0
div = 0
cnt = 8
cnt 9
itr 0
div 0
cnt = 9
nnnn
itr = 0
div = 0
cnt = 9
cnt 10
itr 0
div 0
Out Check  [2, 2, 2, 2, 2, 2,
```

Genetic

Population Size: 2
Mutation_Rate: 1
CrossOver_Rate: 4
No of Generation: 2
No of Clusters: 10

```
Precision value For Cluster 7 Is   :  NaN
Precision value For Cluster 8 Is   :  NaN
Precision value For Cluster 9 Is   :  NaN
Precision value For Cluster 10 Is   :  NaN
Presion Value For Genetic Algorithm Is NaN
Recall value For Cluster 1 Is   :  NaN
Recall value For Cluster 2 Is   :  NaN
Recall value For Cluster 3 Is   :  NaN
Recall value For Cluster 4 Is   :  NaN
Recall value For Cluster 5 Is   :  NaN
Recall value For Cluster 6 Is   :  NaN
Recall value For Cluster 7 Is   :  NaN
Recall value For Cluster 8 Is   :  NaN
Recall value For Cluster 9 Is   :  NaN
Recall value For Cluster 10 Is   :  NaN
Recall Value For Genetic Algorithm Is NaN

                         Total Cpu Execution Time For Sequential Trans

Final Cluster Enrobe 80%
```

Figure 2. Outcome of Genetic Algorithm

Here we compare the DENCOS with respect to the enhanced DENCOS with GENETIC approach for the Density Conscious Subspace Clustering for High-Dimensional Data. We compare the two algorithms with two constraints (i.e.) Execution Time and Clustering Performance.

The results are tabulated below:

<table>
<tr><th colspan="5">TIME Comparison</th></tr>
<tr><th>Dataset</th><th>DENCOS</th><th>CLIQUE</th><th>SUBCLU</th><th>Enhanced DENCOS with GENETIC</th></tr>
<tr><td>Adult</td><td>470</td><td>3797</td><td>1436900</td><td>205</td></tr>
<tr><th colspan="5">Performance Comparison</th></tr>
<tr><th>Dataset</th><th>DENCOS</th><th>CLIQUE</th><th>SUBCLU</th><th>Enhanced DENCOS with GENETIC</th></tr>
<tr><td>Adult</td><td>78%</td><td>76%</td><td>70%</td><td>80%</td></tr>
</table>

Table 1: Experimental results – comparison

As validated by our extensive experiments on retail dataset, enhanced DENCOS with GENETIC algorithm can discover the clusters in all subspaces with high quality, and the efficiency outperforms previous works using DENCOS.

VI. CONCLUSION

In this research, we devised a novel subspace clustering model to discover the subspace clusters. We noted that previous works lack of considering the critical problem, called “the density divergence problem,” in discovering the clusters, where they utilize an absolute density value as the density threshold to identify the dense regions in all subspaces. Therefore, as shown in the conducted experiment results, previous works have the difficulties in achieving high qualities of the clusters in all subspaces. In view of the density divergence problem, we identify the dense regions (clusters) in a subspace by discovering the regions which have relatively high densities as compared to the average region density in the subspace. Therefore, in our model, different density thresholds will be utilized to discover the clusters in different subspace cardinalities. As shown by our experimental results, DENCOS can discover the clusters in all subspaces with high quality, and the efficiency of DENCOS significantly outperforms previous works, thus demonstrating its practicability for subspace clustering.

VII. REFERENCES

[1] C.C. Aggarwal, A. Hinneburg, and D. Keim, “On the Surprising Behavior of Distance Metrics in High Dimensional Space,” Proc. Eighth Int’l Conf. Database Theory (ICDT), 2001.

[2] C.C. Aggarwal and C. Procopiuc, “Fast Algorithms for Projected Clustering,” Proc. ACM SIGMOD Int’l Conf. Management of Data, 1999.

[3] C.C. Aggarwal and P.S. Yu, “The IGrid Index: Reversing the Dimensionality Curse for Similarity Indexing in High Dimensional Space,” Proc. Sixth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining, 2000.

[4] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications," Proc. ACM SIGMOD Int'l Conf. Management of Data, 1998.

[5] I. Assent, R. Krieger, E. Muller, and T. Seidl, "DUSC: Dimensionality Unbiased Subspace Clustering," Proc. IEEE Int'l Conf. Data Mining (ICDM), 2007.

[6] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is Nearest Neighbors Meaningful?" Proc. Seventh Int'l Conf. Database Theory (ICDT), 1999.

[7] A. Blum and P. Langley, "Selection of Relevant Features and Examples in Machine Learning," Artificial Intelligence, vol. 97, pp. 245-271, 1997.

[8] M.-S. Chen, J. Han, and P.S. Yu, "Data Mining: An Overview from Database Perspective," IEEE Trans. Knowledge and Data Eng., vol. 8, no. 6, pp. 866-883, Dec. 1996.

[9] C.H. Cheng, A.W. Fu, and Y. Zhang, "Entropy-Based Subspace Clustering for Mining Numerical Data," Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 1999.

[10] H. Fang, C. Zhai, L. Liu, and J. Yang, "Subspace Clustering for Microarray Data Analysis: Multiple Criteria and Significance," Proc. IEEE Computational Systems Bioinformatics Conf., 2004.

[11] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[12] Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang, "Density Conscious Subspace Clustering for High-Dimensional Data," IEEE Transactions On Knowledge And Data Engineering, VOL. 22, NO. 1, pp. 16 – 30, JANUARY 2010.

[13] Rama.B, Jayashree.P, Salim Jiwani, "A Survey on clustering:Current status and challenging issues," (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 09, 2010, pp - 2976-2980.

[14] K. Mumtaz and Dr. K. Duraiswamy, "A Novel Density based improved k-means Clustering Algorithm – Dbkmeans," (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 02, 2010, pp - 213-218.