Subspace Clustering using CLIQUE: An Exploratory Study

Jyoti Yadav, Dharmender Kumar

Abstract— **Traditional clustering algorithms like K-means, CLARANS, BIRCH, DBSCAN etc. are not able to handle higher dimensional data because of the many issues occurred in high dimensional data e.g. "curse of dimensionality", "irrelevant dimensions", "distance problem" etc. To cluster higher dimensional data, density and grid based, both traditional clustering algorithms combined and let to a step ahead to the traditional clustering i.e. called subspace clustering. This paper presents an important subspace clustering algorithm called CLIQUE, which deals with all the problems ensued in clustering high dimensional data. CLIQUE find clusters of arbitrary shape in large dimensional data by taking grid size and a density threshold value as a user input. It starts process of finding clusters at a single dimension and then proceeds upward to the higher dimensions. In this paper, CLIQUE is compared with the other traditional clustering algorithms to measure its performance in terms of accuracy and time taken, in high dimensional space.**

Index Terms— **CLIQUE, Apriori approach, Subspace Clustering, Alternative Subspace Clustering.**

I. Introduction

CLIQUE (Clustering in QUEst) is a bottom-up subspace clustering algorithm that constructs static grids. It uses apriori approach to reduce the search space, described in section II. CLIQUE is a density and grid based i.e. subspace clustering algorithm and find out the clusters by taking density threshold and number of grids as input parameters [18]. CLIQUE operates on multidimensional data by not operating all the dimensions at once but by processing a single dimension at first step and then grows upward to the higher one. The clustering process in CLIQUE involves first dividing the number of dimensions into non-overlapping rectangular units called grids according to the given grid size and then find out the dense region according to a given threshold value. A unit is dense if the data points in this are exceeding the threshold value. Then the clusters are generated from the all dense subspaces by using the apriori approach.

Finally CLIQUE algorithm generates minimal description for the clusters obtained by first determines the maximal dense regions in the subspaces and then minimal cover for each cluster from that maximal region. It repeats the same procedure until covered all the dimensions [10].

Characteristics of CLIQUE

- CLIQUE allows finding clusters of arbitrary shape.
- CLIQUE is also able to find any number of clusters in any number of dimensions and the number is not predetermined by a parameter.
- Clusters may be found in any subspace means in a single or overlapped subspace.
- The clusters may also overlap each other meaning that instances can belong to more than one cluster.

II. APRIORI APPROACH

According to apriori approach, if a *k* dimensional unit is dense then all its projections in *k-1* dimensional space are also dense means a region that is dense in a particular subspace must create dense regions when projected onto lower dimensional subspaces. CLIQUE confines its search for dense units in high dimensions to the intersection of dense units in subspaces, because CLIQUE is based on Apriori property.

III. SUBSPACE CLUSTERING

Subspace clustering is the next step of traditional clustering, it solve the problems of clustering high dimensional data by combining density and grid based traditional clustering methods and seeks to find clusters in different subspaces within a dataset [8]. It mine clusters in high dimensional data by dividing the whole dataset into grids that is called subspaces [14]. The goal of the subspace clustering is to detect the most relevant subspace projections for any object in the database. Any cluster is then associated with a set of relevant dimensions in which this pattern has been discovered [15]. Subspace clustering automatically identifying subspaces [3] of a high dimensional data space that allow better clustering of the data points than the original space. The subspace is not necessarily (and actually is usually not) the same for different clusters within one clustering solution. The key issue in subspace clustering is the definition of similarity taking into account only certain subspaces [12]. For a clear understanding of subspace clustering, let's have a look on the fig. 1 shown above:

Manuscript received Oct, 2013.

Jyoti Yadav*, Department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar, Haryana*

Dharmender Kumar*, Department of Computer Science & Engineering, Guru Jambheshwar University of Science & Technology, Hisar, Haryana*

Fig. 1: 2-D space with Subspace Clusters

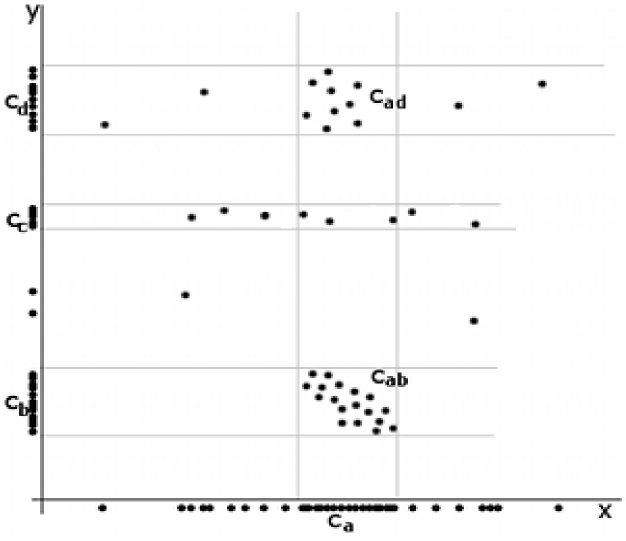


Fig. 1 contains a two dimensional space where a number of clusters can be identified in different subspaces. The clusters $\mathbf{c_a}$ (in subspace{x}) and $\mathbf{c_b}$, $\mathbf{c_c}$, $\mathbf{c_d}$ (in subspace {y}) can be found. C_c can’t be considered a cluster in a two-dimensional (sub) space, since it is too sparsely distributed in the x axis. In two dimensions, the two clusters C_{ab} and C_{ad} can be identified [23].

The term “subspace clustering” in a narrower sense does also relate to a special category of clustering algorithms in axis-parallel subspaces. Another family of clustering in axis-parallel subspaces is called “projected clustering”. Mostly subspace clustering and projected clustering both terms were used in parallel for development of further approaches [17]. Projected clustering algorithms are restricted to disjoint sets of objects, while subspace clustering algorithms might report several clusters for the same object in different subspace projections [16].

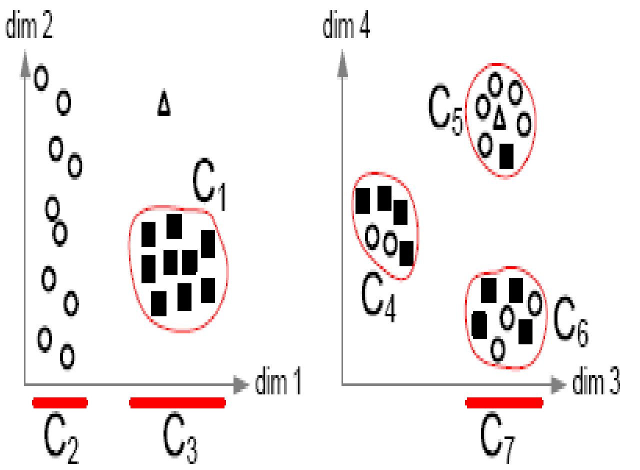
There is one other method, which is derived from the subspace clustering i.e. “Alternative subspace clustering”.

IV. ASCLU: ALTERNATIVE SUBSPACE CLUSTERING

Alternative subspace clustering ASCLU [27] is used to find out that information which is hidden in subspace clustering if the user wants something more than the results obtained from subspace clustering e.g. in economics a customer can potentially belong to several customer groups. If the user is not satisfied with the existent knowledge either because it does not meet his application needs or because he assumes there must exist more of those clusters in the data, then he aims for an alternative. So in such type of cases rather than finding the clusters in whole subspace again, the user prefers to find some useful clustering information from the already existing clustering. Then in this case alternative subspace clustering comes into play.

This problem statement of alternative subspace clustering is best explained by the fig 2, shown above.

Fig. 2: Alternative subspace clustering



The problem statement of alternative subspace clustering i.e. if given an already known subspace clustering which contain many subspace clusters, then the aim of alternative subspace clustering is to determine a meaningful subset of all possible subspace clusters such that the new subset differs from the input clustering such that suppose, we have a subspace cluster C, which contain set of objects O and set of dimensions S i.e. C = (O, S). The objects O are similar within the relevant dimensions S while the dimensions S are irrelevant for the cluster. In alternative subspace clustering, our task is to identify another subspace cluster C’ within the database that differs from the given one.

V. RELATED WORK

Most traditional clustering methods like k-means [26], CLARANS [25], OPTICS [20], BIRCH [21], STING [22], CURE [24] etc. have very high computational complexities. They are not suitable for clustering high dimensional datasets [6]. For high dimensional datasets, neither partitioned nor hierarchical, full-dimensional clustering algorithms can find meaningful clusters. This is due to the effects of high dimensionality because of that one single object contains a number of dimensions which results in poor quality of clustering; “distance problem” i.e. distance between objects becomes indistinguishable when the number of attributes becomes high, and the average density of candidate data points for a cluster, anywhere in the large dimensional data space, is likely to be low, that’s why these approaches missed out meaningful clusters. Also in high dimensional data, many of the dimensions are irrelevant. These irrelevant dimensions confuse clustering algorithms by hiding clusters in noisy data [19]. For large spatial databases, there is a need of such kind of clustering algorithm which provides solution to the problems occurred in clustering high dimensional data [11]. This type of clustering algorithms should be able to discover clusters of complex shape on large databases. All the well-known traditional clustering algorithms provide no solution to these requirements. The solution of these problems is subspace clustering which is the next step of traditional clustering that seeks to find clusters in different subspaces within a dataset [1]. It mine subspace clusters in high dimensional data.

DENCOS (DENsity Conscious Subspace clustering) can discover the clusters in all subspaces with high quality by identifying the regions of high density and consider them as clusters [2], [13]. Subspace clustering has many algorithms [4]. CLIQUE is the first subspace clustering algorithm. The CLIQUE algorithm finds the crowed region from the multidimensional database and discovers the patterns. If the unit is dense then it proceeds to form a cluster. The outlier detection of clusters and finalize about the noisy data also an important thing in high dimensional data sets. To do so, clusters are analyzed with respect to positive and negative objects in CLIQUE by intra-cluster similarity of clusters with respect to the occurrence of positive and negative objects through RandIndex [5]. The redundant objects are eliminated from the region by matrix factorization and partition method [7]. A method for speeding-up the step of CLIQUE algorithm is described in [9].

VI. PSEUDOCODE OF CLIQUE

```
CLIQUE has main three steps:
  1) Identification of subspace that is dense.
    A) Finding of dense units:
         • Firstly find the set D1 of all one dimensional dense units
         • K = 1
         • While $D_K$! = Ø do
         • K = K+1
         • Find the set $D_K$ which is the set of all the k-dimensional dense units whose all lower dimensional projections i.e. (k-1), belong to $D_K$-1
         • End while
    B) Finding subspaces of high coverage.
  2) Identification of clusters.
    For each high coverage subspace S do
       • Take the set of all dense units i.e. E in S
       • while E! = Ø
       • m=1
       • Select a randomly chosen unit u from E
       • Assign to Cm, u and all units of E that are connected to u
       • E = E – Cm
       • End while
     End for.
  3) Then generate minimal cluster descriptions
    For each cluster C do
    1st stage:
       • C = 0
       • While C! = Ø
             ➢ x = x+1
             ➢ choose a dense unit in C
             ➢ For i = 1 to L
           Unit proceeds in both the directions along the i-th dimension, trying to cover almost every unit in C (boxes that are not belong to C should not be covered).
             ➢ End for
             ➢ Represent the set I containing all the units covered by the above procedure
             ➢ C = C – I
       • End while
    2nd stage:
       • Remove all covers from the units covered by another cover.
```

VII. PROBLEM STATEMENT

Suppose we have data of “employ” relation that has three attributes namely salary, vacation and age. These attributes contains following instances or tuples:

Table I: Data of “employ” relation

Age (in year)	Salary (in Rs.)	Vacation (in week)
20	10,000	1
25	20,000	2
30	30,000	3
35	40,000	4
40	50,000	5
45	60,000	6
50	70,000	7

Use the CLIQUE algorithm to cluster this “employ” relation, which contains the 3 attributes, described above with their instances.

So the data space for this data would be 3-dimensional with dimensions - vacation, age, salary. As data is distributed among these dimensions, the first step is to divide the dimensions by an equal interval, called grid size. Let’s take grid size 7 and threshold value 3. Firstly, the CLIQUE algorithm finds dense region in two dimensions at a time and then by combining these, finally produce clusters in all three dimensions as shown in fig. 3(a), 3(b) and 3(c) given above.

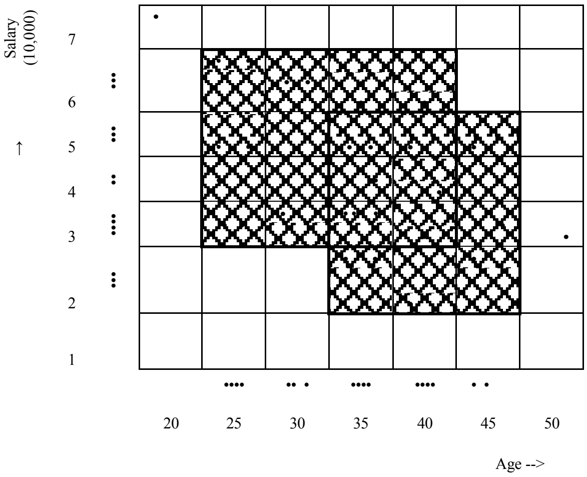


Fig. 3(a): Identification of clusters along (age, salary) plane

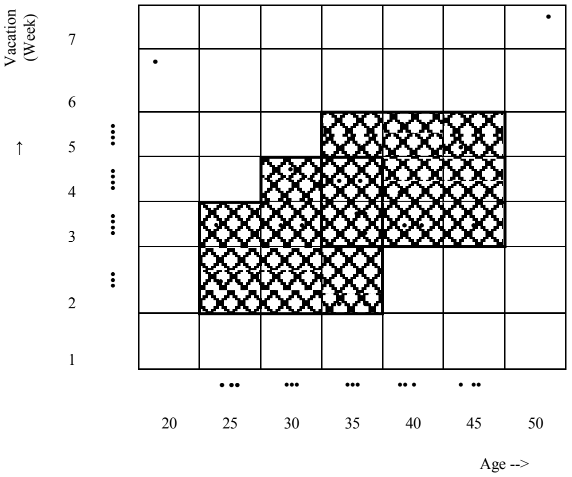


Fig. 3(b): Identification of clusters along (age, vacation) plane

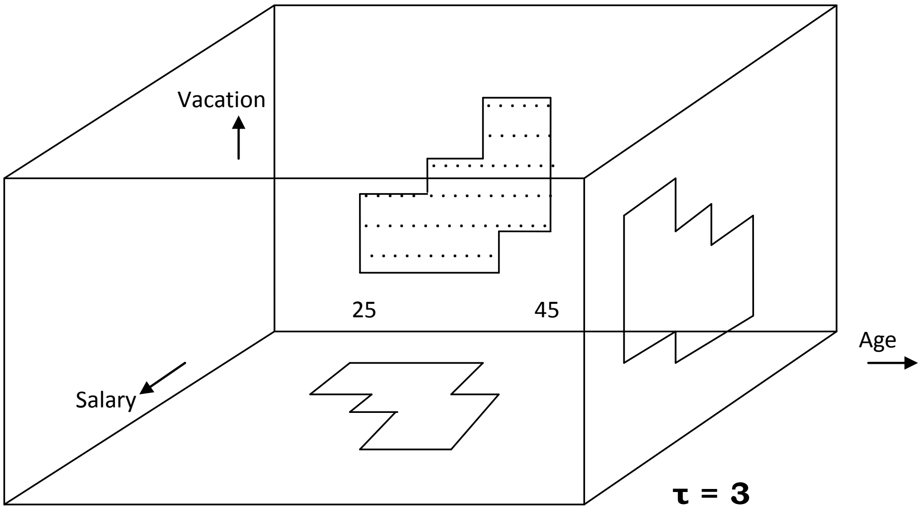


Fig. 3(c): Final clusters in three dimensional (age, salary, vacation) space

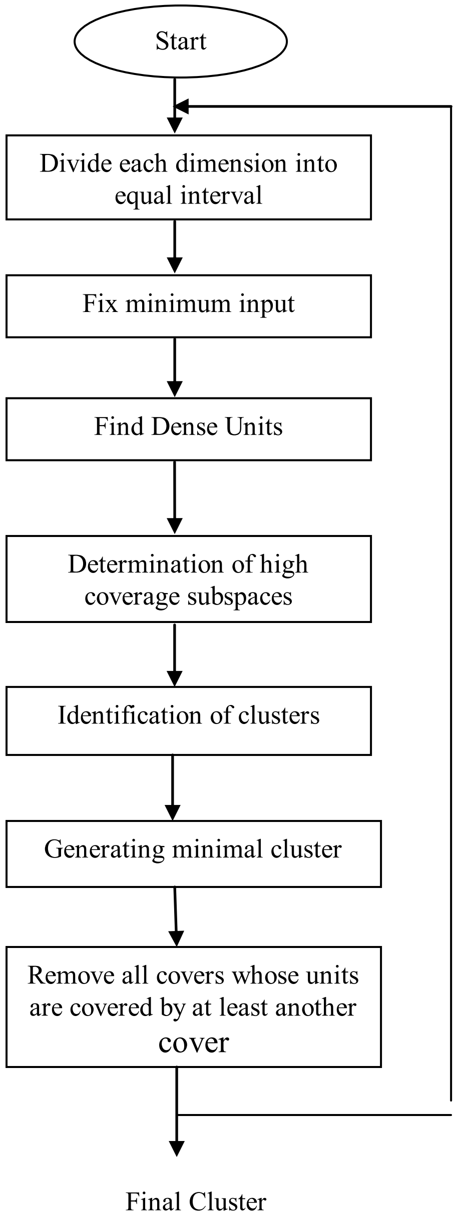
We can also extend the dense region in the vacation-age plane inwards and the dense regions in the salary-age plane upwards.

The 3-dimensional dense units exist in a candidate search space that is obtained by the intersection of these two spaces. Then determine the dense region in the salary-vacation plane and form an extension of the subspace that represents these dense regions.

VIII. FLOW DIAGRAM

The flow diagram of CLIQUE algorithm is shown in the fig. 4.

Fig. 4: Flow diagram of CLIQUE algorithm


Start
Divide each dimension into equal interval
Fix minimum input
Find Dense Units
Determination of high coverage subspaces
Identification of clusters
Generating minimal cluster
Remove all covers whose units are covered by at least another cover
Final Cluster

IX. EXPERIMENTAL RESULTS

Performance of clique in terms of time taken and accuracy is calculated and compared with the other traditional clustering algorithms.

Fig. 5: Execution time comparison between CLIQUE and other clustering algorithms

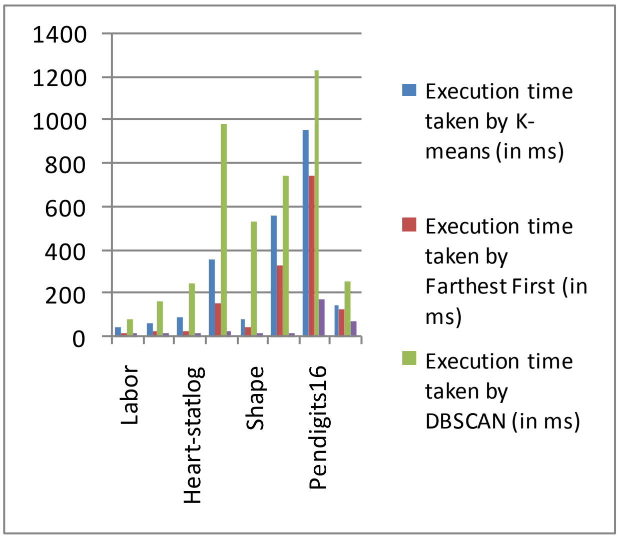
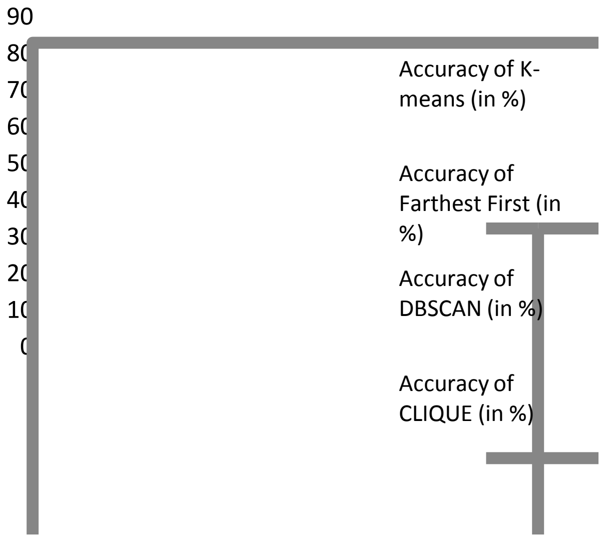


Figure 6: Accuracy of CLIQUE and other clustering algorithms



X. CONCLUSION

CLIQUE is a subspace clustering algorithm and it is compared with the other traditional clustering algorithms. After comparing CLIQUE with K-means, Farthest first and DBSCAN, it is concluded that CLIQUE is better than these algorithms in terms of both execution time and accuracy when forming almost same number of clusters.

It is also concluded that CLIQUE finds clusters of arbitrary shape and is able to find any number of clusters in any number of dimensions while the number is not pre-determined by a parameter.

XI. FUTURE WORK

In future, performance of CLIQUE can be done much better by using adaptive grids instead of fix grid size. Also, instead of using a user defined threshold value, threshold of each grid can be computed automatically by making histograms of each grid.

REFERENCES

[1] Lance Parsons, Ehtesham Haque, Huan Liu, "Subspace Clustering for High Dimensional Data: A Review", *ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets*, vol. 6, no.1, pp.90-105, June 2004.

[2] 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.

[3] Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan, "Automatic subspace clustering of high dimensional data for data mining applications", *ACM SIGMOD international conference on Management of data*, vol.27, no.2, pp.94-105, June 1998.

[4] Anne Patrikainen and Marina Meila, "Comparing Subspace Clusterings", *IEEE Transactions on knowledge and data engineering*, vol.18, no.7, pp.902-916, July 2006.

[5] Raghunath kar , Minakshi Sahu , Tarini Charan Tripathy, "Intra cluster similarity of clique by positive and negative objects", *Int'l journal of Science Engineering and advance Technology (IJSEAT),* vol.1, no.1, pp. 21-24, June 2013.

[6] MR Ilango and Dr V Mohan, "A survey of Grid based clustering algorithms", *Int'l Journal of Engineering Science and Technology*, vol. 2, no.8, pp. 3441-3446, 2010.

[7] Raghunath Kar, Dr. Susant Kumar Dash, "Analysis of clique by matrix factorization and partition methods", *IJCSMS International Journal of Computer Science and Management Studies,* vol.11, no.3, pp. 9-16, October 2011.

[8] Nitin Agarwal, Ehtesham Haque, Huan Liu, and Lance Parsons, "Research Paper Recommender Systems: A subspace clustering approach", *In Proc. of the 6th Int'l Conf. on Advances in Web-Age Information Management*, pp.475–491, 2005.

[9] J Pawar, P R Rao, "An attribute based storage method for speeding up clique algorithm for subspace clustering", *10th International Database Engineering and Applications Symposium (IDEAS '06),* ISSN**:** 1098-8068, pp. 309-310, December 2006.

[10] Raghunath Kar and Susant Kumar Dash, "A study on high dimensional clustering by using clique", *Int'l Journal of Computer Science & Informatics*, vol.1, no.2, pp. 22-25, 2011.

[11] Michael Steinbach, Levent Ertoz, and Vipin Kumar, "The challenges of clustering high dimensional data", Springer Berlin Heidelberg, pp. 273-309, 2004.

[12] Hans-Peter Kriegel, Peer Kroger, Matthias Renz, Sebastian Wurst, "A generic framework for efficient subspace clustering of high-dimensional data", *In proc. 5th IEEE Int'l conf. on Data Mining (ICDM),* pp. 250-257, November 2005.

[13] Karin Kailing, Hans-Peter Kriegel, Peer Kroger, "Density-connected subspace clustering for high-dimensional data", *In proc. of 4th SIAM Int'l Conf. on Data Mining,* pp. 246-257, 2004.

[14] Guanhua Chen, Xiuli Ma, Dongqing Yang, Shiwei Tang, Meng Shuai "Mining representative subspace clusters in high-dimensional data", *Sixth Int'l Conf. on Fuzzy Systems and Knowledge Discovery,* vol.1, pp.490-494, August 2009.

[15] Hans-Peter Kriegel, Peer Kroger and Arthur Zimek, "Clustering high-dimensional data: a survey on subspace clustering, pattern-based clustering and correlation clustering", *ACM Trans. on Knowledge Discovery from*

Data, vol.3, no.1, pp.1-58, March 2009.

[16] Gabriela Moise, Jorg Sander, "Finding non-redundant, statistically significant regions in high dimensional data: A novel approach to projected and subspace clustering", In *Proc. of the 14th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining*, pp.533-541, August 2008.

[17] Rahmat Widia Sembiring, Jasni Mohamad Zain, Abdullah Embong, "Clustering high dimensional data using subspace and projected clustering algorithms", *Int'l journal of computer science & information Technology (IJCSIT)*, vol.2, no.4, pp.162-170, August 2010.

[18] Sunita Jahirabadkar and Parag Kulkarni, "ISC – Intelligent subspace clustering, a density based clustering approach for high dimensional dataset", *World Academy of Science, Engineering & Technology*, no.31, pp. 69-73, July 2009.

[19] H. Y. Shum, K. Ikeuchi, and R. Reddy, "Principal component analysis with missing data and its application to polyhedral object modeling", *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol.17, no.8, pp.854–867, September 1995.

[20] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jorg Sander, "OPTICS: Ordering Points To Identify the Clustering Structure", *In Proc. ACM SIGMOD'99 Int'l. Conf. on Management of Data,* vol.28, no.2, pp.49-60, June 1999.

[21] Tian Zhang, Raghu Ramakrishnan and Miron Livny, "BIRCH: A new data clustering algorithm and its applications", *In Proc. of the ACM SIGMOD Int'l Conf. on Management of data,* vol.25, no.2, pp.103-114, June 1996.

[22] W Wang, J Yang and R Muntz, "STING: A statistical information grid approach to spatial data mining", *In Proc. of 23rd Int. Conf. on VLDB*, pp.186-195, February 1997.

[23] F. de la Torre and M. Black, "A framework for robust subspace learning", *Int'l Journal of Computer Vision* (*IJCV)*, vol.54, no.1, pp.117–142, November 2002.

[24] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, "CURE: An efficient clustering algorithm for large databases", In *Proc. of ACM SIGMOD Int'l conf. on Management of data*, vol.27, no.2, pp. 73-84, June 1998.

[25] Raymond T. Ng and Jiawei Han, "CLARANS: A method for clustering objects for spatial data mining", *IEEE Transactions on Knowledge and Data Engineering*, vol.14, no.5, pp.1003-1016, September/October 2002.

[26] Tajunisha and Saravanan, "Performance analysis of k-means with different initialization methods for high dimensional datasets"*, International Journal of Artificial Intelligence & Applications (IJAIA),* vol.1, no.4, pp.44-52, October 2010.

[27] Hans-Peter Kriegel and Arthur Zimek, "Subspace clustering, Ensemble clustering, Alternative clustering, Multiview clustering: What can we learn from each other", *In Proc. 1st Int'l workshop on discovering, summarizing and using multiple clusterings,* 2010.