IJCSMS International Journal of Computer Science and Management Studies, Vol. 11, Issue 03, Oct 2011
ISSN (Online): 2231-5268
www.ijcsms.com

ANALYSIS OF CLIQUE BY MATRIX FACTORIZATION AND PARTITION METHODS

Raghunath Kar[1], Dr. Susant Kumar Das[2]

[1]Sr.Lecturer, Roland Institute of Technology, Berhampur, Orissa
karrajblocan@yahoo.com

[2]Reader ,Department of computer science,Berhampur University,Orissa
dr.dassusanta@yahoo.co.in

Abstract

In real life clustering of high dimensional data is a big problem. To find out the dense regions from increasing dimensions is one of them. We have already studied the clustering techniques of low dimensional data sets like k-means, k-mediod, BIRCH, CLARANS, CURE, DBScan, PAM etc. If a region is dense then it consists with number of data points with a minimum support of input parameter ø other wise it cannot take into clustering. So in this approach we have implemented CLIQUE to find out the clusters from multidimensional data sets. In dimension growth subspace clustering the clustering process start at single dimensional subspaces and grows upward to higher dimensional ones. It is a partition method where each dimension divided like a grid structure. In this paper the elimination of redundant objects from the regions by matrix factorization and partition method are implemented. The comparisons between CLIQUES with these two methods are studied. The redundant data point belongs to which region to form a cluster is also studied.

Keywords: CLIQUE, APRIORI, dense unit

I INTRODUCTION

CLIQUE clustering is a data mining problem which finds dense regions (collections of units) in a sparse multi-dimensional data set. The attribute values or points and ranges of these regions characterize the clusters. Data from a database or data warehouse having multiple dimensions are called attributes. Many clustering algorithms are good at handling up to three dimensions. We can really observe the clusters up to three dimensions. To find out the clusters from the high dimensional data sets can be highly skewed. We have taken the CLIQUE (Clustering in QUEst) algorithm to find out the clusters. CLIQUE automatically finds the dense units. The dense units are present in subspaces of the increasing dimensions. It scales linearly with the size of input and has good scalability as the number of dimensions in the data increased. The data are present in the different clusters are may be redundancy in nature. The redundancy of data becomes the cluster in large size. Unless the computational complexity growths by taking the redundant data. In this paper the matrix decomposition method is used to eliminate redundant data points. The QR-decomposition method and partition method used to find the clusters as like CLIQUE algorithm.

II CLIQUE OVERVIEW

A unit (cell) is a dense if the sum of total data points in a unit exceeds the input parameter. Clique partitions the *m*-dimensional data space into non-overlapping rectangular units. The dense units are identified from these units. The clusters are generated from all the subspaces of original data spaces, using a Apriori property. If a k- dimensional unit is dense, then so are its projections are in (k-1) - dimensional space. CLIQUE generates minimal descriptions over its data points as follows.

(i) It first determines the maximal dense regions over the data sets in the subspaces

(ii) Each cluster then determines the minimal cover from the maximal regions.

(iii) If the dimension increases the same procedure follows to find out the clusters from the highly density covered areas.

III PROBLEM STATEMENT

[A] QR DECOMPOSITION WITH GRAM-SCHMIDT

The QR decomposition (also called the QR factorization) of a matrix is a decomposition of the matrix into an orthogonal matrix and a triangular

matrix. The QR decomposition of a real square matrix A is a decomposition of A as $A = QR$. where Q is an orthogonal matrix (i.e. $Q^TQ = I$) and R is an upper triangular matrix. If A is nonsingular, then this factorization is unique. There are several methods for actually computing the QR decomposition. One of such method is the Gram-Schmidt process.

$$A = [x_1 \mid x_2 \mid \ldots \mid x_n]$$

$$Then$$

$$v_1 = x_1,\ u_1 = \frac{1}{\|v_1\|} v_1$$

$$v_2 = x_2 - \frac{x_2 . v_1}{v_1 . v_1} v_1$$

$$u_2 = \frac{1}{\|v_2\|} . v_2$$

.....

$$v_k = x_k - \frac{x_k . v_{k-1}}{(v_{k-1}).(v_{k-1})} . v_{k-1}$$

$$and$$

$$u_k = \frac{1}{\|v_k\|} . v_k$$

$$The\ resulting\ QR\ factorization\ is$$

$$A = [x_1 \mid x_2 \mid \ldots \mid x_n]\ is\ defined\ as$$

$$Q = [u_1 \mid u_2 \mid \ldots\ldots \mid u_n]$$

$$R = Q^T . A$$

$$The \|.\| is\ in\ the\ L_2\ form$$

[B] MATRIX PARTITION METHOD

One of the key things in linear algebra is the ability to split up a bigger matrix into smaller subsections, which is also known as partitioning a matrix. For an instance, we have a matrix A.

$$A = \begin{bmatrix} 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 4 & 5 & 1 \\ 2 & 3 & 4 & 1 & 2 \\ 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 2 & 4 & 1 \end{bmatrix}$$

Can be partitioned into its subsections based on the users discretion. Here is a partitioned Matrix:

$$A_{11} = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 3 & 4 \\ 2 & 3 & 4 \end{bmatrix} \quad A_{12} = \begin{bmatrix} 5 & 6 \\ 5 & 1 \\ 1 & 2 \end{bmatrix}$$

$$A_{21} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix} \quad A_{22} = \begin{bmatrix} 4 & 5 \\ 4 & 1 \end{bmatrix}$$

Finally $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$

[C] SUBSPACE CLUSTERING

CLIQUE was one of the first algorithms proposed that attempted to find clusters within subspaces (I.e. combination of units) of the dataset. As described above, the algorithm combines density (which satisfies minimum threshold value) and grid (each cell) based clustering and uses an APRIORI technique to find subspaces which is more dense upon data and clusterable. Once the dense subspaces are found they are sorted by coverage, in the dimensions where coverage is defined as the units of a two dimension dataset covered by the subspace. The subspaces with the maximal coverage are kept and the rest are pruned. The algorithm then finds adjacent dense units in each of the selected subspaces using a depth first search. Clusters are formed by combining these dense units using a greedy growth scheme. The algorithm starts with an arbitrary dense unit and greedily grows a maximal region in each dimension until the union of all the regions covers the entire cluster (maximal regions). The intersections of regions are not a cluster.

Redundant regions are (called overlapping of subspaces) removed by a repeated procedure where smallest redundant regions or a unit is having the data not satisfies with minimum threshold value are discarded. The hyper-rectangular clusters are then defined by a Disjunctive Normal Form (DNF) expression. A *region* is a set of axis parallel rectangular areas in *n*- dimensions. Clustering is expressed as the union of regions only. The region can also be expressed in the mean of DNF expressions as described in the Apriori property. We say that if a cluster R is over the region F, then R= R I F. Always the minimal description of a cluster is a non redundant covering of the cluster with maximal regions**.**

V RELATED WORK

The data points are present in a multidimensional database usually not in a uniform manner. The CLIQUE algorithm finds the dense units (crowed units) from the multidimensional database and discovers the patterns among dimensional axes. If the data points are present in a unit is dense, then the clusters are formed from these dense units. If no units or cells consisting with the minimum threshold value or data points, then it follows the following rules.

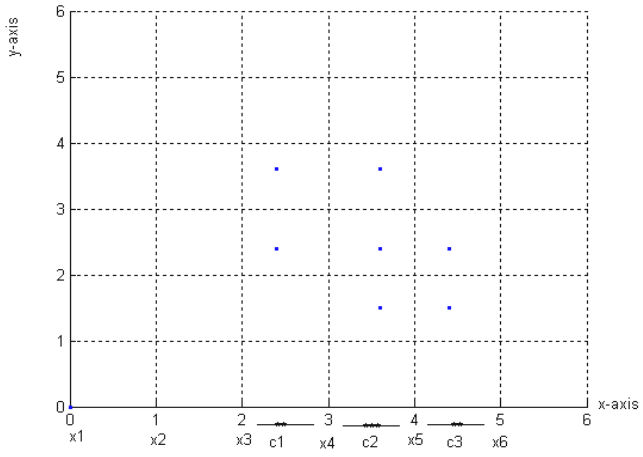

Figure 1

(i)Assume figure 1 let ø=2, project along Y-axis and count number of data elements from the cells. Here the number of data points in the individual cells is only one. It is not satisfied with the value of ø, and then count the total number of data points along the Y-axis represented as $\{(x_3 \rightarrow x_4) = 2, (x_4 \rightarrow x_5) = 3, (x_5 \rightarrow x_6) = 2\}$. These are greater than or equal to ø. So the connected regions are forming individual clusters as c_1, c_2 and c_3 as shown in the figure 2.

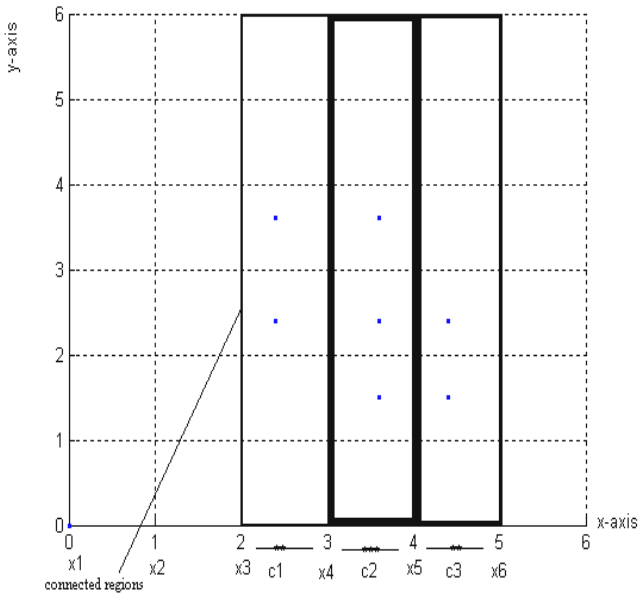

Figure 2

If the cells are already satisfied with the minimum threshold value then we follow the following logical rules to form clusters. For example it was observed that.

TABLE 1 PRODUCT TABLE

Occupation	Age	Product purchased
Student	15-30	laptop
employee	20-25	printer

Age(Y,"15-30") ∧ occupation(Y,"student") ⟹ buys(Y,"laptop") Eq (1)

Association rules that involve two or more dimensions or predicates can be referred to as multidimensional association rules. The above rule is containing three predicates (age, occupation, buys) where each one occurs once. No repetition of predicates is present here. Hence multidimensional association rules without repeated predicates are called inter dimensional association rules. The

equation (1) is one of the DNF expressions for the table 1.

DNF expression for clustering

(i) The dense region has been shaded. $c_1 \curlyvee c_2$ is a cluster.

(ii) c_1 or c_2 independently are the maximal region contained in this cluster.

(iii) Where $c_1 \curlywedge c_2$ is not a maximal region.

(iv) The minimal description for this cluster in the DNF expression as in the figure 3 is

$((2 \leq x \leq 4) \wedge (2 \leq y \leq 4)) \vee ((3 \leq x \leq 5) \wedge (1 \leq y \leq 3))$

If we apply the density based method for CLIQUE then the connected adjacent cells are the maximum region in these specified dimensions. So the two regions are defined in the figure 3 as c_1 and c_2. The DNF expression for these two regions are as

$c_1 = ((2 \leq x \leq 4) \wedge (2 \leq y \leq 4))$

$c_2 = ((3 \leq x \leq 5) \wedge (1 \leq y \leq 3))$

The DNF expression for the cluster which is formed from the regions c_1 and c_2 as given below

$c=(c_1 \curlyvee c_2) = ((2 \leq x \leq 4) \wedge (2 \leq y \leq 4)) \vee ((3 \leq x \leq 5) \wedge (1 \leq y \leq 3))$

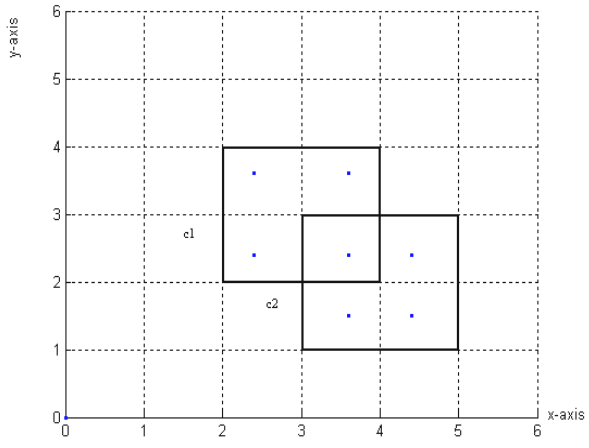

Figure 3

VI EXPERIMENAL RESULT

[A] IMPLEMENTATION OF MATRIX DECOMPOSITION METHOD

In this paper we study the matrix factorization method to find the maximal regions. So from figure 3 the two regions are c_1 and c_2. Let the c_1 and c_2 consisting the data points are as shown in the figure 3. From the association rule mining we observed that the cluster from the dense regions with refers to figure 3 are shown is as figure 4. Where the regions are with minimal description is outlier and it is a cluster. This is also done by matrix decomposition method to identify a cluster from the two regions named as $c_1$and c_2.

$$A = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$$

(Data points from the regions c_1 and c_2.)

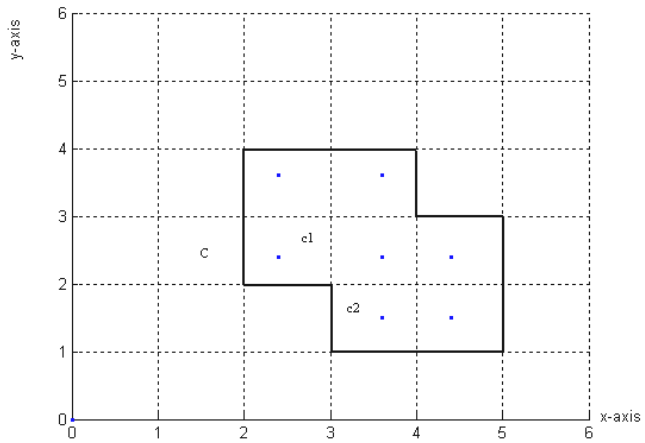

Figure 4

From the figure 4 it is clear that region c_1 contains the data points x_1,x_2,x_3,x_4 and the region c_2 consisting with the data points x_4,x_5,x_6,x_7. This is represented as $c_1=\{x_1,x_2,x_3,x_4\}$ and $c2=\{x_4,x_5,x_6,x_7\}$. The matrix and tabular form is denoted as

$$A = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$$

Here the data point x_4 is the repeated elements in the two subspaces. So to avoid the overlapping of the subspaces we have to decide whether the element should present in which subspaces or regions? In the figure 5 let us assume that x_4 is a common point and it is present both the regions c_1 and c_2. The DNF for the data point x_4 is $((3 \leq x \leq 4) \wedge (2 \leq y \leq 3))$.

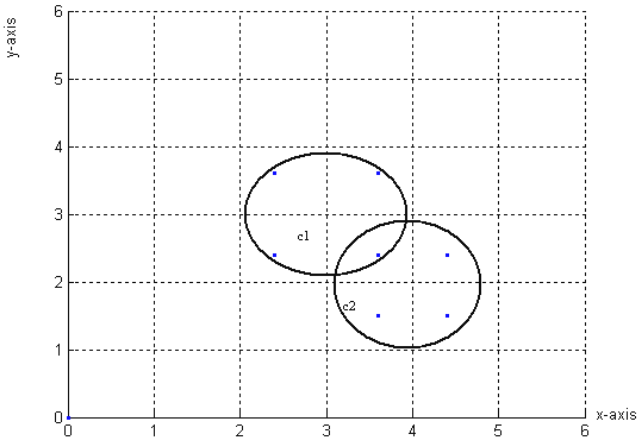

Figure 5

To solve this problem we take the help of the matrix decomposition method. Where A is a matrix and the elements $x_1 \ldots x_7$ are belonging to regions c_1 and c_2.

$$A = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$$

And we follow the Gram-Schmidt QR Decomposition as

$$x_1 = v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$v_2 = x_2 - \frac{x_2 \cdot v_1}{v_1 \cdot v_1} \cdot v_1$$

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \frac{1}{4} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -1/4 \\ -1/4 \\ -1/4 \\ 3/4 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$v_2 = \begin{pmatrix} -1/4 \\ -1/4 \\ -1/4 \\ 3/4 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$u_1 = \frac{1}{\|v_1\|} v_1 = \frac{1}{2} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$u_2 = \frac{1}{\|v_2\|} . v_2 = \frac{2}{\sqrt{15}} \begin{pmatrix} -1/4 \\ -1/4 \\ -1/4 \\ -3/4 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1/2\sqrt{15} \\ -1/2\sqrt{15} \\ -1/2\sqrt{15} \\ \sqrt{3}/2\sqrt{5} \\ 2/\sqrt{15} \\ 2/\sqrt{15} \\ 2/\sqrt{15} \end{pmatrix}$$

$$Q = (u_1, u_2) = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{cc} c1 & c2 \end{array} \\ \begin{pmatrix} 1/2 & -1/2\sqrt{15} \\ 1/2 & -1/2\sqrt{15} \\ 1/2 & -1/2\sqrt{15} \\ 1/2 & -\sqrt{3}/2\sqrt{5} \\ 0 & 2/\sqrt{15} \\ 0 & 2/\sqrt{15} \\ 0 & 2/\sqrt{15} \end{pmatrix} \end{array}$$

$$R = Q^T A = \begin{pmatrix} 1/2 & 1/2 & 1/2 & 1/2 & 0 & 0 & 0 \\ -1/2\sqrt{15} & -1/2\sqrt{15} & -1/2\sqrt{15} & \sqrt{3}/2\sqrt{5} & 2/\sqrt{15} & 2/\sqrt{15} & 2/\sqrt{15} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 2 & 1/2 \\ 0 & \sqrt{15}/2 \end{pmatrix}$$

So A=QR hence proved.

The matrix A is decomposed in to two parts as Q and R where

$$R = \begin{pmatrix} 2 & 1/2 \\ 0 & \sqrt{15}/2 \end{pmatrix}$$

$$Q = (u_1, u_2) = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{cc} c1 & c2 \end{array} \\ \begin{pmatrix} 1/2 & -1/2\sqrt{15} \\ 1/2 & -1/2\sqrt{15} \\ 1/2 & -1/2\sqrt{15} \\ 1/2 & -\sqrt{3}/2\sqrt{5} \\ 0 & 2/\sqrt{15} \\ 0 & 2/\sqrt{15} \\ 0 & 2/\sqrt{15} \end{pmatrix} \end{array}$$

Here the region c_1 belongs to the data points $c_1=\{x_1,x_2,x_3,x_4\}$ and the region c_2 belongs to the data points after decomposition as $c_2 = \{x_1,x_2,x_3,x_4, x_5,x_6,x_7\}$. But as referenced to figure 4 c_2 is formed a new region and it is sufficient to show that it is the minimal description of the region. So it is proved that if the data point x_4 lies in the c_1 then it will not a cluster. But if the data point will lie in the c_2 then it is a union of c_1 and c_2 and finally it is a minimal description of the region and it is a cluster.

[B] IMPLEMENTATION OF MATRIX PARTITION METHOD

So in our problem the elements from dense region are shown in the matrix format as.

$$A = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$$

Hence it is partitioned as

$$A_{11} = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \end{array}$$

$$A_{12} = \begin{array}{c} \\ x_4 \\ x_5 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$$

$$A_{13} = \begin{array}{c} \\ x_6 \\ x_7 \end{array} \begin{array}{c} \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \end{array}$$

From A_{11} $c_1=\{x_1, x_2, x_3\}$, $c_2= \{\}$
From A_{12} $c_1= \{x_4\}$, $c_2=\{x_4, x_5\}$
From A_{13} $c_1= \{\}$, $c_2=\{x_6, x_7\}$
$A_{11} \curlyvee A_{12} = \{x_1,x_2 ,x_3, x_4, x_5\}$
$A_{13} = \{x_6, x_7\}$
$c= \{A_{11} \curlyvee A_{12} \curlyvee A_{13}\} = \{x_1, x_2 , x_3, x_4, x_5, x_6, x_7 \}$
Here the union of all the regions are defined in the A_{11},A_{12},and A_{13} is $c=\{x_1, x_2 , x_3, x_4, x_5, x_6, x_7 \}$. This is equal to the result of the matrix decomposition method i.e. $c_2 = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7$.

VII CONCLUSION

For the CLIQUE clustering the both methods are helpful to find the clusters easily as compared to the clustering algorithm defined above. In the matrix partition method it is difficult to find the partition boundaries. After fixing the boundary the work is easy. But in the factorization method no such boundaries are fixed before the solution but the implementation is necessary.

VIII TIMECOMPLEXITY

If C-number of clusters
n – Highest Dimensionality
m- Number of input points
C-number of clusters
τ Number of Intervals
$\varnothing$ – Density Threshold, Then the time complexity of is CLIQUE Numerical Data $O(C^n+mn)$.But in the method of matrix factorization the same job may be done with the time complexity of $O(m^2n)$.

IX FUTURE WORK

In CLIQUE the clusters are formed with large overlap among the reported dense regions. It is difficult to find clusters of different density within different dimensional subspaces. We may use entropy as a measure of the quality of subspace clusters. The PROCLUS (projected clustering) is a typical dimension reduction subspace clustering method may implement to find clusters from high dimensional subspaces. The PROCLUS starts projections from high dimensional subspaces instead of single dimensional subspaces. The association

rule mining may also implement to find the clusters from high dimensional data sets. The factorization with Eigen value may implement the process in future.

X REFERENCES

[1] Data Mining Concept and Techniques- J. Han and M. Kamber
[2] Insight into Datamining: Theory and Practice – K.P.Soman, Shyam Diwakar,V.Ajay.
[3] An Efficient Cell-based Clustering Method for Handling Large, High-Dimensional Data Jae-Woo Chang
[4] Automatic Subspace Clustering of high Dimensional Data for Data Mining Applications by Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan.

[5] Subspace Clustering for Uncertain Data by Stephan Gunnemann, Hardy Kremer, Thomas Seidl.
[6] Subspace Clustering for High Dimensional Data: A Review by Lance Parsons, Ehtesham Haque, Haun Liu.
[7] Mining Subspace Clusters: Enhanced Models, Efficient Algorithms and an Objective Evaluation Study by Emmanuel Muller.
[8] Constraint-based Subspace Clustering by Elisa Fromont, Adrinna Prado, Celine Robardet
[9] Outlier Detection and Ranking Based on Subspace Clustering by Thomas Seidl, Emmanuel Muller, Ira Assent, Uwe Steinhausen.
[10] Analyzing Clique Overlap Martin G. Everett, Stephen P. Borgatti.
[11] Data Mining by Vikram Pudi, P.Radha Krishna.

[12] Matrix Decomposition Methods for Data Mining: Computational Complexity and Algorithms Pauli Miettinen
[13] A study on highdimensional clustering by using CLIQUE Raghunath Kar,Dr.susant Kumar Das