International Journal of Computer Science and Informatics

Volume 1 | Issue 2 Article 6

October 2011

A Study On High Dimensional Clustering By Using Clique

Raghunath Kar
Roland Institute of Technology, Berhampur, Orissa, karrajblocan@yahoo.com

Susanta Kumar Das
Dept. of Computer Science Berhampur University, dr.dassusanta@yahoo.co.in

Recommended Citation

Kar, Raghunath and Kumar Das, Susanta (2011) "A Study On High Dimensional Clustering By Using Clique," *International Journal of Computer Science and Informatics*: Vol. 1 : Iss. 2 , Article 6.
DOI: 10.47893/IJCSI.2011.1019
Available at: https://www.interscience.in/ijcsi/vol1/iss2/6

Available online at www.interscience.in

A Study On High Dimensional Clustering By Using Clique

[1]Raghunath Kar & [2]Susant Kumar Dash
[1]Roland Institute of Technology, Berhampur, Orissa,
[2]Department of Computer Science, Berhampur University, Orissa
Email: karrajblocan@yahoo.com, 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.

The grid is a cell where the data points are present. We check the dense units from the structure by applying different algorithms. Finally the clusters are formed from the high dimensional data sets.

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 as 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.

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. CLIQUE ALGORITHM

This algorithm first partitions the data space into a grid. This is done by partitioning each dimension into equal interval known as units. Then it identifies the dense units (which is greater than an input threshold parameter *ø*). Finally the clusters are formed on the basis of the following algorithm.

CLIQUE(Data D)

<u>Part-1</u>

1. for each dimension *p* in D
2. partition *d* in to equal intervals
3. fix a minimum input parameter *ø* (by applying APRIORI property).
4. find the dense units.
5. *m*=2
6. while(true)
7. for each combination of *m* dimensions $p_1, p_2, \ldots\ldots, p_m$.

8. for each intersection *j* of dense units along *m* dimensions.
9. if *j* is dense
10. identify *j* as a dense unit.
11. if no more units are identified as dense, break the loop.

Part-2(Clustering)

12. The clusters are formed with maximal sets of dense regions from rectangular connected regions.

IV. PROBLEM STATEMENT

Let $S = \{S_1, S_2, \ldots, Sn\}$ is a set of bounded, totally ordered domains and $N = \{A_1, A_2, \ldots, An\}$ is a *n*-dimensional space. The input vector consists of a set of *n*-dimensional points$V = \{v_1, v_{2,\ldots\ldots}, v_n\}$ where $v_i = \{v_{i_1}, v_{i_2}, ., v_{i_n}\}$. The k^{th} component of v_i is drawn from domain S_k. We partition the data space *N* in to non-rectangular units. A unit is a cell from the subspaces. The subspace consisting with number of cells. It is also consisting with intersection of intervals. We call a unit *u* is dense if *u* is greater than ø (which is an input threshold parameter). A cluster is a union of dense (connected cells) units in *n*-dimensions. If the two *n*-dimensional units u_1 and u_2 are connected if they have a common face. If a third dimension u_3 exists then u_1 is connected to u_3 and u_2 is connected to u_3 where the input parameters of $u_1 = \{u_{p_1}, u_{p_2}, \ldots\ldots\ldots, u_{p_n}\}$ and $u_2 = \{u'_{p_1}, u'_{p_2}, u'_{p_3}, \ldots\ldots, u'_{p_n}\}$ and there are *n*-dimensions, let the dimensions are $S_{p_{k\ldots\ldots}} S_{p_{k-1}}$ such that $u_{p_1} = u'_{p_1}$. 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 ∩ 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 data base 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 we can take to form a cluster. If no units or cell contains the minimum threshold value or points then it follows the following rules.

(i) Let *ø*=3, project along Y-axis and count number of points from the cells. Here the number of points in the individual cells is only one. It is not satisfied with ø, and then the total number of points along the axis is six. It is greater then ø. So the connected regions are a cluster.

(ii) Let *ø* is 4 and project along X-axis and count number of points from the cells. The number of points from the cells is five. It satisfies the minimum threshold input parameter *ø*, so the region defined as a rectangular box in the figure-2 is a cluster.

(iii) Here the two dimensional space is partitioned into 6×6 grid. A unit is the intersection of intervals. If the dataset given in the following figure-3, and assuming that ø is 3 (minimum threshold value), none of these cells of two dimensional dataset are dense cells. If the points are projected along the X-dimension however, there are three 1-dimensional dense units. Two of these are connected and merged. Finally two clusters are formed.

(iv) Here also two dimensional spaces have been partitioned in to 6×6 grid. A unit is the intersection of intervals. If the dataset given in the following figure-4, and assuming that *ø* is 2 (minimum threshold value), none of these cells of two dimensional dataset are dense cells. If the points are projected along the Y-dimension however, there are three 1-dimensional dense units. Two of these are connected and merged. Finally two clusters are formed. It may increased upto more number of clusters depends on the data set of the example.

(v) 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.

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

Age(Y,"15-30") Λ occupation(Y,"student") $\Longrightarrow$ buys(Y,"laptop") Equation (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.

VI. EXPERIMENAL RESULT

So by applying the rule which is given above in section-(v). We will find the maximum and minimum region from grid structure of figure-5.

(i) The dense region has been shaded. A $\cup$ B is a cluster.

(ii) A or B independently are the maximal region contained in this cluster.

(iii) Where A $\cap$ B is not a maximal region.

(iv) The minimal description for this cluster in the DNF expression as in the figure - 5

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

For multidimensional clustering we propose two graphs as shown in figure-5 and figure-6. The graphs are partitioned in to 36 cells .The density regions are found by applying association rule The individual graph structures are given as a two dimensional grid structures as X-Y and X-Z. But for the high dimensional clustering the dimensions are merged. Here in the figure-5 the data points are present in X-Y dimension. In the figure-6 the data points present in X-Z dimensions. So the clusters are formed from X-Y-Z dimensions as shown in figure-7. The line represented in the three dimensional cube is the data points from the common regions of the X-Y and X-Z dimensions. This is a study of data points. It can be extended textual data on the basis of association rule mining of section-(v).

VII TIMECOMPLEXITY

The dens units are present in a given subspace can not be very large because each dense unit must have selectivity of at least an input parameter. From the number of dense units, if each one is visited then it checks its *2m*-neighbours (where m is the number of dimension). The CLIQUE algorithm finds the maximal connected regions or units, if the total number of dense units in a subspace is n then the total number of accessing time are *2mn*.

VIII FEATURE 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.

IX. 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.

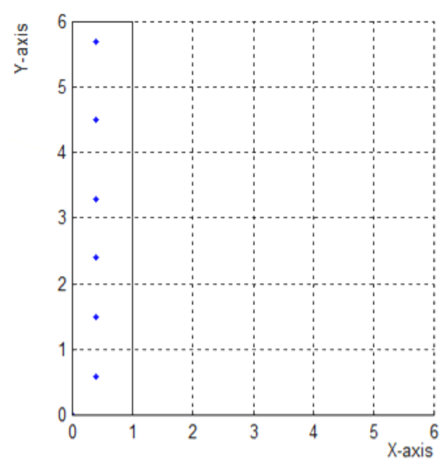

Fig 1 Number of points from all the cells are = 6

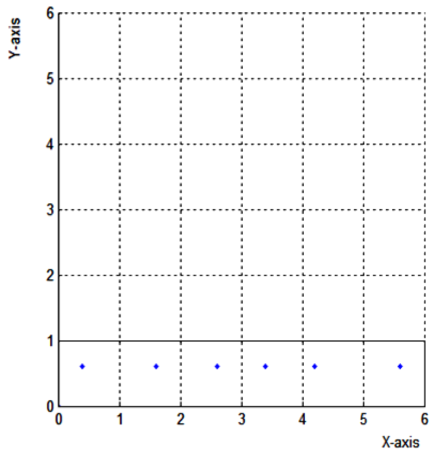

Fig 2 : Number of points from all the cells are =5

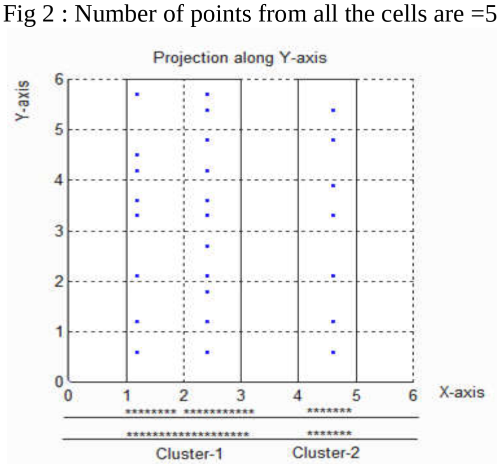

Fig 3 : Let the threshold value of ø is 3. No cell satisfies the minimum value. Here (X1-X3) merged as cluster-1 and (X4-X5) is a cluster-2.

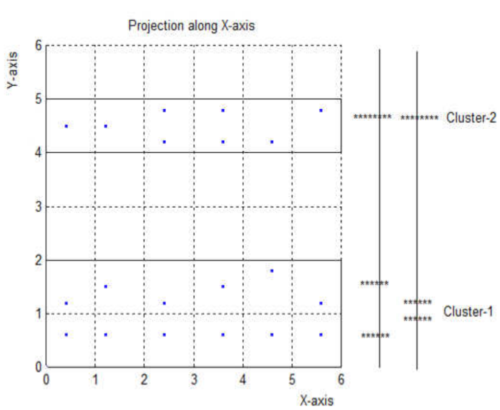

Fig 4 Let the threshold value of ø is 2. So no unit satisfies with minimum points. Here we merge (Y0-Y2) and it forms a cluster and (Y4-Y5) is a new cluster

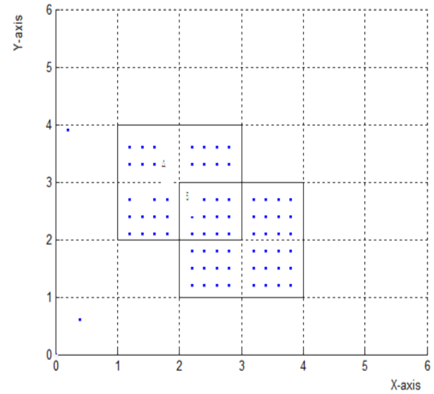

Fig 5 The DNF expression for minimal description for the dense region is as $((2\leq x \leq 4) \wedge (1 \leq y \leq 3)) \vee ((1 \leq x \leq 3) \wedge (2 \leq y \leq 4))$

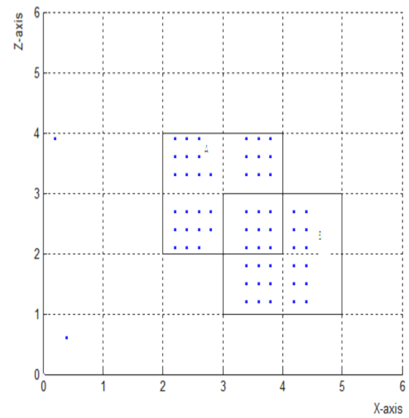

Fig 6 The DNF expression for minimal description for the dense region is as $((2\leq x \leq 4) \wedge (2\leq z \leq 4)) \vee ((3\leq x \leq 3) \wedge (1\leq z \leq 3))$

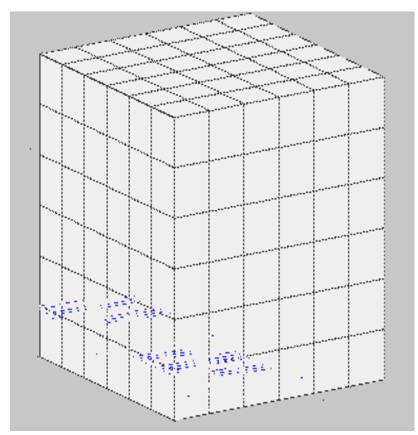

Fig 7 The DNF expression for minimal description for the dense region in fig-5is as $((2\leq x \leq 4) \wedge (1 \leq y \leq 3)) \vee ((1 \leq x \leq 3) \wedge (2 \leq y \leq 4))$ and the DNF expression for minimal description for the dense region in fig-6 is as $((2\leq x \leq 4) \wedge (2\leq z \leq 4)) \vee ((3\leq x \leq 3) \wedge (1\leq z \leq 3))$ .The data cube represented as a cluster from the x-y-z dimensions from fig-5 and fig-6.