Intra cluster similarity of clique by positive and negative objects

Raghunath kar , Minakshi Sahu , Tarini Charan Tripathy
Department of Comp.Ssience & Engg., Roland Institute of Technology, Berhampur, Odisha.
Department of Comp.Ssience & Engg., Kalam Institute of Technology, Berhampur, Odisha.
Department of Electronics & Communication Engg.Kalam Institute of Technology, Berhampur, Odisha.
karrajblocan@yahoo.com saiminakshi@gmail.com tarinitripathy@yahoo.com

***Abstract*-Clustering of high dimensional data sets is a challenge in now a days. Finding of good clusters is a big problem for the researchers. The outlier detection of clusters and finalize about the noisy data also an important thing in highdimensional data sets. In this paper it is researched about intracluster similarity of clusters with respect to occurrence of positive and negative objects through RandIndex. There are many methods are present to find good clusters from the high dimensional datasets like BIRCH, CLARA, CLARANS, DBScan, PAM, CLIQUE etc.The CLIQUE(Clustering in Quest) is a dimension-Growth subspace clustering method is used to find clusters. Here the process starts at single dimensional subspaces grows upward to higher dimensional ones. After finding the clusters from CLIQUE, with a user defined threshold value the occurrence of positive and negative objects are also found and it compares with the entropy of positive and negative objects with the value of RandIndex.**

Keywords: CLIQUE, positive and negative objects, RandIndex etc.

I. INTRODUCTION

The CLIQUE is a clustering method for dimensional growth data sets. Here each dimension partitioned in to a grid structure. Each grid structure is consisting with number of data objects. As per the consisting of data objects, the subsequent dense units are found. If the grid contains the noisy data it is a separate thing but if the units(cells) are dense, then the clusters are found on the basis of user defined threshold value. The sparse and crowded area with respect to objects in non overlapping rectangular units is found. Clusters are basically found from the maximal set of connected dense units with minimal cover. After finding the clusters the positive and negative objects through RandIndex() method, then the positive and negative objects and their ratios are calculated from the individual clusters. Finally the entropy of clusters with respect to positive and negative objects are calculated and compared with the value of RandIndex.

II. PROBLEM STATEMENT

If the set of objects are *n* then the set of objects are represented as like

$$M = \{o_{1,\ldots\ldots,} o_n\},$$

Suppose $P = \{p_1, p_2, \ldots\ldots, p_k\}$ and

$Q = \{q_1, q_2, \ldots\ldots, q_l\}$, represent two different partitions of the objects M such that $\bigcup_{i=1}^{k} p_i = S = \bigcup_{j=1}^{l} q_j$ $and\ p_i \cap p_i' = \emptyset\ and\ q_i \cap q_j' = \emptyset\ \ for\ 1 \leq i \neq i' \leq k\ \ and\ \ 1 \leq j \neq j' \leq l$

Suppose P related to the class and Q related to the clusters of those classes. The objects are classified as per the following criteria.

Let the parameters a, b, c, d represents the status of the objects with respect to presence their in classes and clusters. These are calculated as below.

'a' represents the assignment of number of pairs of objects that are placed in same class and same cluster. It is treated as True Positive (TP).

'b' represents the assignment of number of pairs of objects that are placed in same class but different cluster. It is treated as False Negative(FN).

'c' represents the assignment of number of pairs of objects that are placed in different classes and same clusters. This is known as False Positive (FP).

'd' represents the assignment of number of pairs of objects that are placed in different classes and different clusters. This is known as True Negative (TN).

After analyzing the objects the RandIndex() is found on the basis of agreements and disagreements of the objects. Here the quantities TP and TN are treated as agreements and FP and FN are treated as disagreements of the objects over the clusters. The RandIndex() is always with in the range *0 and 1*.It is represented by the following formula.

$$\mathrm{RI} = \frac{a+d}{a+b+c+d} \quad or\ \mathrm{RI} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$$

Here RI is the RandIndex() and a,b,c,d or TP,TN,FP,FN are the parameters described above.

III. RELATED WORK

In the CLIQUE method if the k dense units is dense, then the projections in $(k-1)^{th}$ dimensional space. Finally for each cluster it determines the objects with minimal coverage of maximum connected dense units. Let us assume the following example as given in the figure 1.

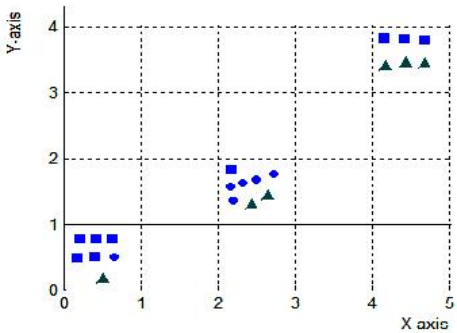

Fig 1 Data sets in the Grid

The grid is represented through two dimensions as X and Y. The dense units (cells) are of these dimensions are

$$\{x_0, \dots\dots\dots, x_5\}\ and\ \{y_0, \dots\dots, y_4\}.$$

The data objects are found here in the grid units as

$$\{(x_0, x_1), (y_0, y_1)\}, \{(x_2, x_3), (y_1, y_2)\} and\ \{(x_4, x_5), (y_3, y_4)\}.$$

These three units are the dense units .The units are not connected here. If we consider α=2 as a threshold value, the objects are present in the grids are greater than or equal to α. Here three clusters are formed with the dimensional subspaces. This is shown in figure 2.

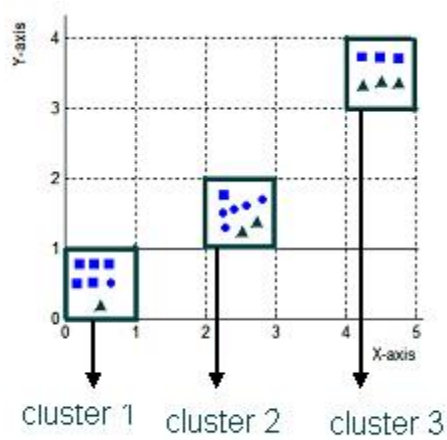

Figure 2 Clusters are formed from the Grid

The clusters are given with respect to dimensions.

$$q_1 = \{(x_0, x_1), (y_0, y_1)\}, q_2 = \{(x_2, x_3), (y_1, y_2)\} and\ q_3 = \{(x_4, x_5), (y_3, y_4)\}.$$

The idea about the positivity and negativity of the objects was not given previously. The idea of RandIndex() is introduced here to find out the positive and negative objects. This can be represented in the matrix form of classes and clusters This is defined in the figure 3.

class/cluster	q_1	q_2	------	q_l	sums
p_1	n_{11}	n_{12}	------	n_{1l}	n_1
p_2	n_{21}	n_{22}	------	n_{2l}	n_2
⋮	⋮	⋮	------	⋮	⋮
p_k	n_{k1}	n_{k2}	------	n_{kl}	n_k
sums	n_1	n_2	------	n_l	n_{kl}

Figure 3 Matrix form of partitioned dataset

Where TP can be defined by the formula as $\sum_{i,j}\binom{n_{ij}}{2}$. The FN can be represented by $\sum_i\binom{n_{i.}}{2} - \sum_{i,j}\binom{n_{ij}}{2}$. The FP can also be represented by $\sum_j\binom{n_{.j}}{2} - \sum_{i,j}\binom{n_{ij}}{2}$. Finally the TN can be represented by the formula as $\mathbf{TN} = \binom{n}{2} - (\mathbf{TP} + \mathbf{FN} + \mathbf{FP})$.

So by implementing this formals to our newly formed clusters (by CLIQUE) as located in the figure 2. At first we rearrange the objects in the matrix form. This is shown in figure 4.

class/cluster	q_1	q_2	q_3	sums
p_1	5	1	3	9
p_2	1	5	0	6
p_3	1	2	3	6
sums	7	8	6	n=21

Figure 4 Classification of Data in matrix form

Here the classes and clusters are given in the matrix form. The figures are the number of different data objects present in the clusters. P_1 represents the square, p_2 represents the circle. P_3 represents the triangle.

The true positive objective are calculated as

$$\binom{5}{2} + \binom{5}{2} + \binom{2}{2} + \binom{3}{2} + \binom{3}{2} = 10 + 10 + 1 + 3 + 3 = 27, \mathrm{TP} = 27$$

---------Equation (1)

The false negative objects are calculated as

$$\binom{9}{2} + \binom{6}{2} + \binom{6}{2} - 27 = 36 + 15 + 15 - 27 = 66 - 27 =$$

----------Equation (2)

The false positive objects are calculated as

$$\binom{6}{2} + \binom{6}{2} + \binom{5}{2} - 27 = 15 + 15 + 10 = 40 - 27 = 13,$$

---------Equation (3)

The true negative objects are calculated as per the following formula. Since we know that

$$\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP} = \binom{n}{2}.$$

------Equation (4)

If we put the values of equations (1) ,(2) ,(3) in the Equation(4) we get the result of TN

$$\mathrm{TN} = \binom{n}{2} - (\mathrm{TP} + \mathrm{FN} + \mathrm{FP}) = \binom{21}{2} - (27 + 39 + 13) =$$

----------Equation (5)

The Rand Index(RI) of this object are calculated as below

$$\mathrm{RI} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}} = \frac{27 + 131}{27 + 13 + 39 + 131} = 0.752$$

----------Equation (6)

If the clusters are formed with the objects of two partitions and they agree perfectly, the value of RandIndex is 1 otherwise 0.Now we analyze the Entropy value of positive and negative objects.

(a) Entropy

When we need to define a measure commonly used in information theory, called *Entropy*. If a given dataset(s), containing only positive and negative objects of some target concept, the *Entropy* of s, simply defined as

$$Entropy(s) = -p_p log_2 p_p - p_n log_2 p_n \ ,$$

---------Equation (7)

Where P_p is the proportion of positive objects and P_n is the proportion of negative objects. There are three basic properties of *Entropy* is given below.

(i)If all members of 'S' belong to the same class then *Entropy* is 0.
(ii)If the positive and negative data objects are equally present in the data sets then the *Entropy* is '1'.
(iii)If the unequal numbers of positive and negative data sets are present in a given set then the *Entropy* lies between 0 and 1.
Here the total positive objects (with true positive and false positive) is 40. The total negative objects (with true negative and false negative) is 170.The *Entropy* of the dataset is calculated as below.

$$Entropy(s) = -\frac{40}{210} log_2 \frac{40}{210} - \frac{170}{210} log_2 \frac{170}{210}.$$

--------Equation (8)

$$Entropy(s) = 0.137 + 0.075 = 0.212.$$

---------Equation (9)

IV. CONCLUSION

Clusters are analyzed with respect to positive and negative objects in CLIQUE. Here it proves from both the results that the unequal number of objects (positive and negative) is present. The results through RandIndex and through Entropy are calculated as it is given in the equation (6) and (9). We have seen that the disagreement of the objects through RandIndex is more. The limitation of both the methods are [0, 1].

V. FUTURE WORK

In future the performance of positive and negative objects within the individual cluster may be extended .In CLIQUE, Instead of taking threshold value it may be taken the *Entropy* measure for finding of good clusters. Implementation of Adjusted RandIndex for getting good clusters may be extended in future.

VI. REFERENCES

[1] Data Mining Concepts and Techniques Jiawei Han, Micheline Kamber

[2] An empirical study on principal component Analysis for clustering gene expression dada by Ka yee yeung,Watter L. Ruzzo.

[3] Insight into Datamining: Theory and Practice – K.P.Soman, Shyam Diwakar, V.Ajay.

[4] An Efficient Cell-based Clustering Method for Handling Large, High-Dimensional Data Jae-Woo Chang

[5] Automatic Subspace Clustering of high Dimensional Data for Data Mining Applications by Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, Prabhakar Raghavan

[6] Subspace Clustering for Uncertain Data by Stephan Gunnemann, Hardy Kremer, Thomas Seidl.

[7] Subspace Clustering for High Dimensional Data: A Review by Lance Parsons, Ehtesham Haque, Haun Liu.

[8] Mining Subspace Clusters: Enhanced Models, Efficient Algorithms and an Objective Evaluation Study by Emmanuel Muller.