An Emerge Approach in Inter Cluster Similarity for Quality Clusters

[1]H. Venkateswara Reddy [2]S. Viswanadha Raju [3]B. Suresh Kumar [4] C. Jayachandra
[1]Associate Professor in CSE,VCE, Hyderabad, India, venkat_nidhish@yahoo.co.in
[2]Professor in CSE,JNTUH, Hyderabad, India,viswanadha_raju2004@yahoo.co.in
[3]M.Tech (C.S.E),VCE, Hyderabad, India ,sureshkumargoud2006@gmail.com
[4]M.Tech (C.S.E),VCE, Hyderabad, India,chinnijayachandra@gmail.com

Abstract: **Relationship between the datasets is one most important issue in recent years. The recent methods are based mostly on the numerical data, but these methods are not suitable for real time data such as web pages, business transactions etc., which are known as Categorical data. It is difficult to find relationship in categorical data. In this paper, a new approach is proposed for finding the relationship between the categorical data, hence to find relationship between the clusters. The main aim is to identify the quality clusters based on the relationship between clusters. If there is no relationship between clusters then those clusters are treated as quality clusters.**

Keywords: Inter cluster similarity, Sliding windows, Outlier detection, Node Importance, Data labelling, co-relationship.

I. INTRODUCTION

Categorical data is consisting of categorical variables. The clustering of the categorical data has been deemed an important issue in data mining [1].The Goal of clustering is to partition the data points into several groups according to the predefined similarity measurements [11].

To improve the efficiency of clustering, sampling is used to scale down the size of the data base [2]. Sampling is used to speed up the clustering algorithm in [4] and [3]. In a typical approach, the sampling techniques are used on clustering to randomly choose a small set from the original database, and the clustering algorithm is implemented on the small sampled set. The clustering result can thus be obtained efficiently which will be similar to that obtained from the original database. The earlier works not fully studied on the problem of allocating the un-clustered data into appropriate clusters. The intention of clustering is to distribute each un-clustered data point into a suitable cluster without loss of generality. An incomplete clustering product obtained from the sampled database is generally not what the user actually needs. For example, when we perform clustering for "customers" segmentation with a sampling technique, a part of customers is sampled and grouped after clustering. However, the other customers which are not sampled will not obtain the cluster label and thus do not belong to any segment. In ROCK algorithm [5], a similar sampling approach is applied to speed up the whole clustering method, and the difficulty of allocating the un-clustered data is also discussed.

To improve the quality of the clusters it is required to know the relationship between clusters, in such case we have to follow a procedure. In the next Sections, the procedure involving different steps is discussed.

This paper discusses clustering the data base by using sampling method in section II, Outlier detection in section III, finding node importance in section IV, data labeling in section V, finding relationship between the clusters in section VI, and the paper concludes with experimental results.

II.DATA CLUSTERING

Clustering is alignment of the objects in to a group called cluster so that the objects of the same cluster are more similar to each other than objects from different clusters. Often, similarity is measured depending on distance evaluate. However, in categorical data it is difficult to find the distance between the categorical variables. Therefore, many algorithms perform clustering based on pattern reorganization or sampling window technique; with using of sampling technique algorithm in this paper we highlighted the mentioned data base shown Table 1.

Table 1:

Object	A_1	A_2	A_3
X_1	A	M	C
X_2	Y	E	P
X_3	X	E	P
X_4	Y	M	P
X_5	A	M	D
X_6	A	M	C
X_7	X	M	P
X_8	A	M	D
X_9	Y	M	P
X_{10}	A	M	C
X_{11}	B	E	G
X_{12}	X	M	P
X_{13}	B	E	D
X_{14}	Y	M	P
X_{15}	B	F	D
X_{16}	Y	M	P
X_{17}	X	M	P
X_{18}	Z	N	T
X_{19}	X	M	P
X_{20}	Y	M	P

From the above table we are dividing sample size as 5.so that

S_1={ X_1, X_2, X_3, X_4, X_5}, S_2={ X_6, X_7, X_8, X_9, X_{10}}, S_3={ X_{11}, X_{12}, X_{13}, X_{14}, X_{15}} and S_4={ X_{16}, X_{17}, X_{18}, X_{19}, X_{20}}

By Appling sampling technique on S_1 and S_2 we get the following clustered databases Clusters C_1 and C_2 Shown in tables below.

Table 2: C_1

Object	X_1	X_5	X_6	X_8	X_{10}
A_1	A	A	A	A	A
A_2	M	M	M	M	M
A_3	C	D	C	D	C

Table 3: C_2

Object	X_2	X_3	X_4	X_7	X_9
A_1	Y	X	Y	X	Y
A_2	E	E	M	M	M
A_3	P	P	P	P	P

When carrying on doing sampling on S_1, S_2 we also get some outliers which do not belong to any other cluster (If any).

III. Outlier Detection

Outlier detection is one of the majority significant subjects in recent duration. Outlier detection is the process of detecting errors in data. The recent methods are mostly based on the Numerical data, but these methods are not suitable in real time data such as web pages, business transactions etc., which are known as Categorical data. It is difficult to find outliers in categorical data. Outlier detection is the process of detecting the data object which is exceptional from the large amount of data. This process is used in telecommunications, financial fraud detections and data cleaning, to improve the quality of the services. An Outlier is defined as “The data objects that do not comply with the general behaviour or model of the data. Such data objects, which are grossly different from or inconsistent with remaining sets of data are called Outliers”[14]. But according to Hawkins’ definition, “An outlier is an observation that deviates so much from other observations so as to arouse suspicion that it was generated by a different mechanism “[10]. Although other definitions are also specified by the researchers and they faced many problems while applying for real time data.

For detecting outliers in large databases, the standard error deviation function method is proposed.

Standard deviation (S) = $\sqrt{\frac{1}{n-1}\sum (x_i - \mu)^2}$ ....(1)

Mean of objects $\mu = \frac{1}{n}\sum x_i$ ....(2)

Standard deviation error (S_{ed})= $\frac{x_i - \mu}{S}$ ....(3)

Symbols used shown in table

Symbol	Meaning
N	Number of objects in database in DB
x_i	Value of attribute in database DB

By applying these functions on cluster C_1, and cluster C_2 we can get the outliers from those clusters.
S_{ed} for Cluster C_1 : The cluster C_1 is shown below

Object	X_1	X_5	X_6	X_8	X_{10}
A_1	A	A	A	A	A
A_2	M	M	M	M	M
A_3	C	D	C	D	C

Applying the S_{ed} on cluster C_1
x_A=5,x_m=5, x_c=3, x_n=2 and n=4 {A,M,C,D}
The mean of the cluster C_1

$$\mu = \frac{1}{4}(x_A + x_M + x_C + x_D)$$

$$\mu = \frac{1}{4}(5 + 5 + 3 + 2)$$

$$\mu = \frac{1}{4}(15)$$

μ= 3.75
Standard error deviation is as follows.

$$S = \sqrt{\frac{1}{(4-1)}\left[(5-3.75)^2 + (5-3.75)^2 + (3-3.75)^2 + (2-3.75)^2\right]}$$

S=1.5
Error detection in Cluster C_1

$$S_{ed(A)} = \frac{5-3.75}{1.5} = 0.83$$

$$S_{ed(M)} = \frac{5-3.75}{1.5} = 0.83$$

$$S_{ed(C)} = \frac{3-3.75}{1.5} = -0.5$$

$$S_{ed(D)} = \frac{2-3.75}{1.5} = -1.16$$

The S_{ed} of Every attribute is shown in the table below.

Table 4:

Attribute	S_{ed}
A	0.83
M	0.83
C	-0.5
D	-1.16

Standard Error deviation Cluster C_1

Table 5:

Object	X_1	X_5	X_6	X_8	X_{10}
A_1	A =0.83	A =0.83	A =0.83	A =0.83	A =0.83
A_2	M=0.83	M=0.83	M=0.83	M=0.83	M=0.83
A_3	C=-1.16	D=-0.5	C=-1.16	D=-0.5	C=-1.16
Total S_{ed}	S_{ed}=0.5	S_{ed}=1.16	S_{ed}=0.5	S_{ed}=1.16	S_{ed}=0.5

The possible solutions for S_{ed}

i) Total S_{ed}= +ve then object belongs to cluster.

i i) Total S_{ed}= -ve then object is outlier.

From the results shown in Table5 it is concluded that there are no outliers in the Cluster C_1. Similarly by applying the same procedure on Cluster C_2, there will not be any outliers.

IV. NODE IMPORTANCE

The problem of evaluating node importance [13] in clustering is one of the active research topics in recent days and many methods have been proposed. For finding the node importance, the Our-NIR[9] method is used which is proposed by H.V.Reddy and S.ViswanadhaRaju et.al.

According to Our-NIR Method, the primary thing of Our-NIR is based on the notion of representing the clusters by the importance of the attribute values. This representation is more efficient than using the representative points. After thorough scrutiny of the literature, it is clear that clustering categorical data is untouched mainly due to the complexity involved in it.

A time-evolving categorical data is to be clustered within the due course. Therefore clustering data can be understood as follows: let there is a series of categorical data points D, where each data point is a vector of q attribute values, i.e., $P_j = (\mathrm{P}_j^1, P_j^2, ...P_j^q)$. And A = {$A_1$,$A_2$,..., A_q}, where A_a is the a^{th} categorical attribute, $1 \leq a \leq q$. The window size N is to be given so that the data set D is divided into several continuous subsets S_t, where the number of data points in each S_t is N. The superscript number t is the identification number of the sliding window and t is also called time stamp. Here, we consider the first N data points of data set D.

This makes the first data slide or the first sliding window S_0. Our motive is to cluster every data slide and relate the clusters of every data slide with previous clusters formed by the previous data slides. Several notations and representations are used in our work to simplify the process of presentation.

Here we consider a symbolic representation for the r^{th} node in cluster i is N [i, r], The number of data points in cluster C_i is m_i, and k is number of clusters.

This Method contains three rules as follows

A. Rule1 (Probability of node N [i, r])

The probability of node (P_i) in the cluster can be calculated as follows:

$$P_i = \frac{\left| N_{[i,r]} \right|}{m_i} \quad .... (4)$$

B. Rule 2 (Frequency of node N [i, r])

The distribution of the node in the clusters is calculated as follows.

$$d(N[r]) = \frac{\left| \sum_{y=1}^{k} P(N[y,r])^2 \right|}{2} \quad .... (5)$$

where

$$P(N[y,r]) = \frac{|N[y,r]|}{\sum_{z=1}^{k} |N[z,r\}]|}$$

C. Rule 3 (Weighted Function)

The importance of the node N [i, r] is calculated by the product of Rule 1 and Rule 2:

$$W(c_i, N_{[i,r]}) = \mathrm{P}_i * d(N_{[i,r]}) \quad \text{.. (6)}$$

The weighting function is designed to measure the distribution of the node between clusters based on the information theorem [12]. For easy understanding of the Node importance apply sampling method on S_1 and the resultant clusters are C_1, C_2. By applying Our-NIR Method for cluster C_1 the importance of each node as shown in the table 6.

Table 6

Node	Importance
A_1=A	0.5
A_2=M	0.3125
A_3=C	0.5
A_4=D	0.5

If any objects are moved as outliers, with using of node importance method that objects can be labeled. The main aim of this method is to reduce the outliers, and also by using this method on other sampling windows to label the objects into related clusters.

V. DATA LABELING

In this data labeling work the unlabeled data can be moved in to their related clusters. Applying any clustering algorithm on S_1, resultant Clusters are C_1 and C_2 shown in table 8 and table 9.

Table 7:

Object	A_1	A_2	A_3
X_1	A	M	C
X_2	Y	E	P
X_3	X	E	P
X_4	Y	M	P
X_5	A	M	D

Obtained clusters are :

Table 8:

C_1			
Object	A_1	A_2	A_3
X_1	A	M	C
X_5	A	M	D

Table 9:

C_2			
Object	A_1	A_2	A_3
X_2	Y	E	P
X_3	X	E	P
X_4	Y	M	P

S_2 is shown Table 10. Now we perform data labeling on S_2 by considering S_2 data points as unlabeled data points to move into respective clusters.

Table 10:

Object	A_1	A_2	A_3
X_6	A	M	C
X_7	X	M	P
X_8	A	M	D
X_9	Y	M	P
X_{10}	A	M	C

Object X_6 contain the attribute values as{A,M,C} so by taking three node importance of X_6 it is moved into the cluster C_1 like that X_7, X_9 moved into Cluster C_2 and X_6, X_8, X_{10} can moved into C_1 cluster.

VI. CO-RELATION BETWEEN THE CLUSTERS

In this section we are proposing the relationship between the clusters based on correlation method .The relationship between the clusters shows the inter cluster similarity. The proposed Co-Relation method for finding the relation between the clusters is as follows.

$$r = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\left[n\sum X^2 - \left(\sum Y\right)^2\right]}} \quad \text{...} (7)$$

X= Node importance of every attribute in cluster C_i .
Y = Node importance of every attribute in cluster C_{i+1}.
n= Number of attributes in the relation belongs to a single cluster.
With using of this method we find the relationship between the clusters. The scale of the relationship between two variables is measured by the sign and absolute value of co-relation.

- The scale of co-relation is between -1 to +1.
- If the value is negative then there is no relation between the clusters.
- If the value is positive then category of relation show in table 11.

Table 11:

Value of r	Relation
0.90 to 1.00	Very high relation
0.70 to 0.89	High relation
0.50 to 0.69	Moderate relation
0.30 to 0.49	Low co-relation
0.00 to 0.20	Little if any co-relation

Here it is noticed that if any attribute has more than one value then choose the node corresponding to that attribute which is having Maximum node importance. If we form the clusters by taking S_1 we get the two clusters C_1 and C_2.
Then apply the method on cluster C_1 and C_2 the values are shown in table12.

Table 12:

X	Y	XY	X^2	Y^2
A=0.5	Y=0.33	0.165	0.25	0.1089
M=0.277	E=0.33	0.0914	0.0767	0.1089
C=0.25	P=0.5	0.125	0.0625	0.25
$\sum(x)$=1.027	$\sum(Y)$=1.16	$\sum(XY)$=0.3814	$\sum(x^2)$=0.3892	$\sum(Y^2)$=0.4678

and n=3 {A,M,C } or {Y,E,P}
then the relation between the cluster is

$$r=\frac{3(0.3814)-(1.027)(1.16)}{\sqrt{[3(0.3892)-(1.027)^2][3(0.4678)-(1.76)^2}}$$

r=-0.2
The value of r is negative so we conclude that there is no relationship between the clusters. Which means that the inter cluster between clusters very low so these C_1 and C_2 clusters are quality clusters.

Algorithm for cluster Relationship:

INPUT : Database (DB)
OUTPUT : Outliers (O), Relationship (C_i, Ci_{+1})
<u>Method</u>
Step 1:Apply any sampling algorithm(Ex. ROCK) on DB and form the initial clusters.
Step 2: for each node i in Cluster C_n in every level Use (3) for outlier detection in clusters.
Step 3 : end for.

Step 4: Use the Node importance (6) For node i in every cluster C_i.
Step 5:for each i in unlabeled S_i move the attributes in C_i by using node importance method // Label the unlabeled data points .
Step 6: end for .
Step 7: For each attribute value $A_i = I_r$
If ($I_r = I_{r+1}$)
{
max(C_i,$A_i = I_r$)
}
Step 8: End if
Step 9: End for.
Step 10: Apply the Co-Relationship method (7)
If (r=-ve)
{
print no relation
}
else
{
there is relation between C_i and C_{i+1}
}

Step 11:End .

VII. EXPERIMENTAL RESULTS

In this section, we demonstrate the performance of the proposed work on clustering categorical data by a thorough experimental study on the real dataset. In Section VII.I, the test environment and the dataset used are described. Next section, the evolving processes of clustering results is visualized on the real dataset.

VII.I Test Environment and Dataset:

All of our experiments are conducted on a PC with an Intel Corei3 processor with 1 GB memory and the Windows7 professional operating system. In all experiments, the k-modes [6] clustering algorithm is chosen to do the initial clustering and reclustering on the datasets. As the k-modes algorithm is dependent on the selection of initial cluster centers, we utilize an initialization method, which was proposed in [7], to obtain initial cluster centers before executing the k-modes. For to developing this paper we use the Java language and backend as My Sql .

The KDD-CUP'99 network-intrusion-detection stream dataset [8], which has been used earlier to evaluate several stream-clustering algorithms and DCDAs, is used in our study. The network-intrusion-detection dataset consists of a series of transmission control protocol (TCP) connection records from two weeks of LAN traffic managed by the Lincoln Laboratories at the Massachusetts Institute of Technology. Every record can also communicate to a normal connection or an intrusion. As a result, the data contain a total of 23 classes including the class for "normal connection."

In the following experiments, all 22 attack-types are seen as "attack." We utilize the class label which indicates that the record is a normal connection or an attack to identify the drifting concept. The majority of the associations in this dataset are normal, but occasionally, there might be a fracture of attacks. One of the objectives in the intrusion-detection system is to detect the changes of connections from normal to a burst of attacks or from the attacks back to normal, and those changes naturally correspond to a drifting concept. Therefore, this dataset is time-evolving data and is suitable for evaluating our algorithms. We utilize the 10% subset version, which is provided from the KDD-CUP'99 website for our experiments. In this dataset, there are 494 021 records, and each record contains 42 attributes (class label is included), such as the duration of the connection,

The number of data bytes transmitted from source to destination (and vice versa), the percentile of connections that have "SYN" errors, the number of "root" accesses, etc. Also, 34 attributes are continuous. We accept identical quantization on those numerical attributes where each attribute is quantized into five categorical values.

7.2 Evaluation on Accuracy:

In this evaluation, we perform clustering algorithms on the entire database to obtain the answer of the clustering result. Then, the work presented in this paper is adopted. The only difference in this evaluation is between the data labeling phase to label unlabeled data points.
After performing the sampling method on the database we can get clusters as C1 and C2, but sometimes that

cluster may contain outliers. So in order to Identify that outliers in the database we perform the outlier detection method that is Standard Deviation Error Method (S_{ed}) by following Equation= $\frac{x_i - \mu}{S}$

By applying this method on the datasets (DB) we can get the outliers. The Figure 1 shows the Insertion of Data into a table.

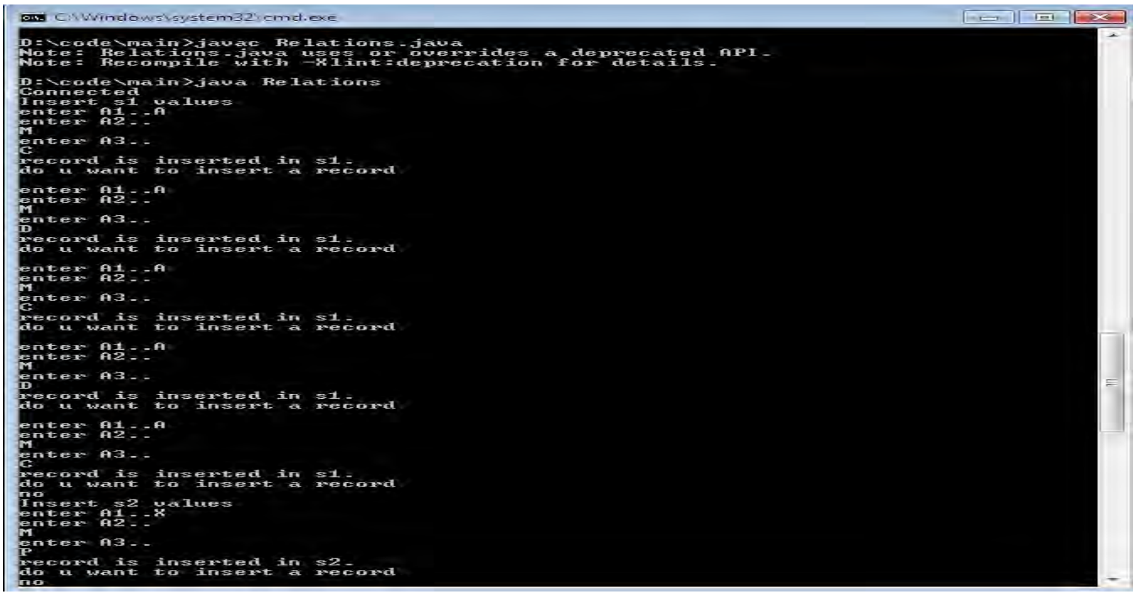

Fig 1: Data insertion into database

The Figure 1 can only contain the data which is stored in back end, but here our aim is to find S_{ed} for all the objects in clusters. The Figure 2 shows S_{ed} of all objects in cluster C_1 and C_2.

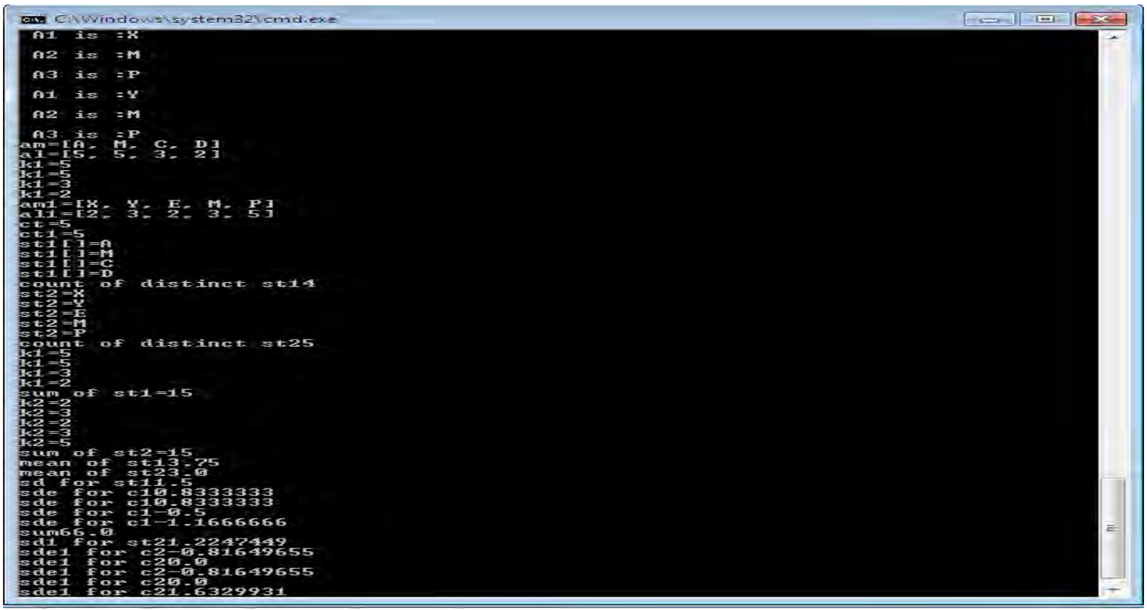

Fig 2: Standard Error deviation for each attribute in clusters.

By applying outlier method on C_1 and C_2 outlier can be retrieved. In Figure2 identifies the S_{ed} values for each object in C_1 and C_2. The method can't find any outliers in S_1 and S_2, but remaining Sampling windows can get the outliers that is shown in the Figure 3.

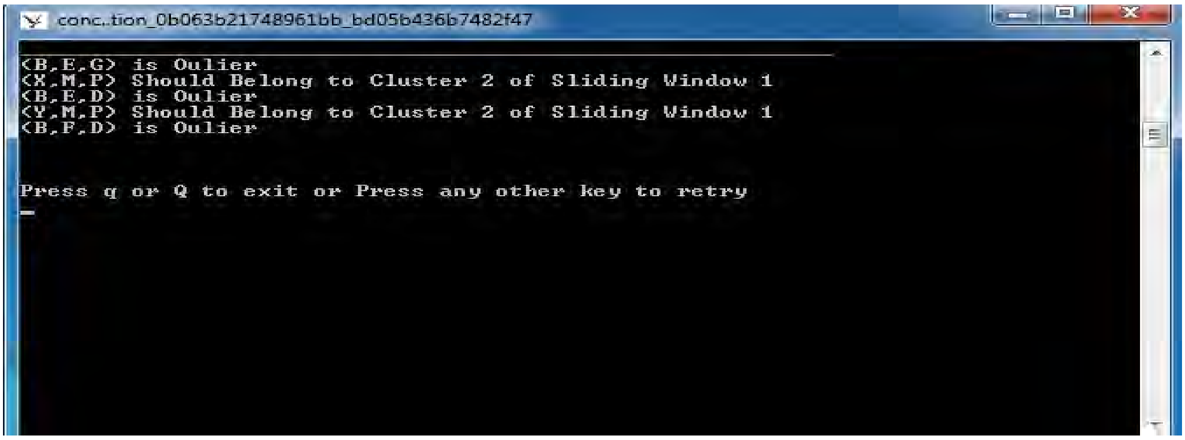

Fig 3: Outlier detection in Clusters by taking S_3 as unlabeled data point.

For data labelling here we use the node impotence method. The node importance for each node in cluster is shown in Figure 4.

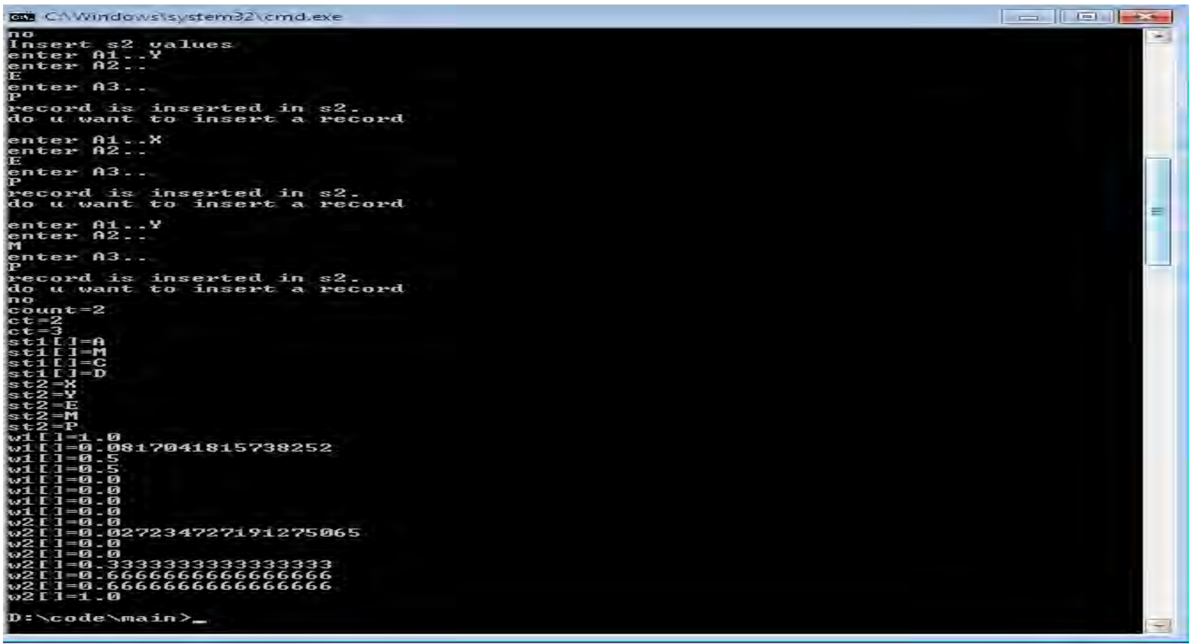

Figure 4: Node Importance of every attribute in cluster C1 and C2.

By finding the node importance of each node in the clusters, the next proposed method is to find relationship between the clusters. For finding the relationship between the clusters formula (7) is used.

The method explained in section 5 .The Figure5 shows the Relationship between the clusters.

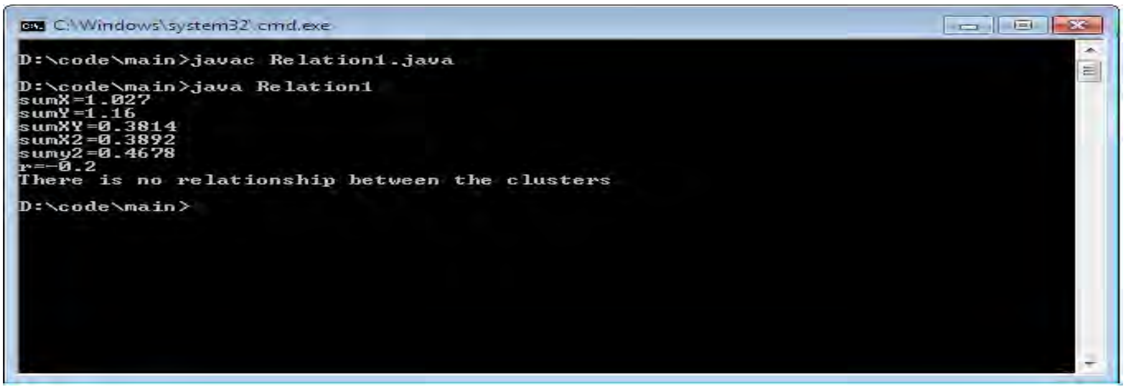

Fig 5: Cluster relationship.

The Figure5 contain the relationship value between the clusters is r=-0.2. According to definition if the r value is negative we can conclude that there is no relation between the clusters.

This paper use sampling method for dividing the data into clusters based on similarity. The attribute in S_1 and S_2 is having the similarity. The graph in Figure 7 shows the S_{ed} of C_1 and C_2 .

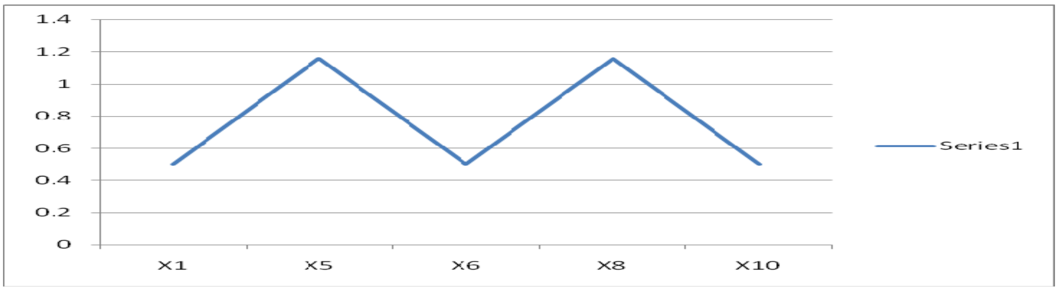

Fig 7: S_{ed} for each object in databases.

According the graph in Figure 7, the value is positive so there are no outliers in the cluster. Our work is to find the relationship between the clusters. For finding the relationship between clusters in this paper we proposed relationship method and results are shown in Figures.

VIII.CONCLUSION

In this paper we proposed a method for finding the relationship. The main reason behind the cluster relationship is finding the quality clusters. The measurement of quality cluster is Inter cluster similarity. Inter cluster similarity is nothing but the relationship between the cluster. When the Inter cluster similarity is low then that clusters are quality cluster. Our Proposed method shows the relationship between the clusters by using Node importance method.

REFERENCES

[1] M.-S. Chen, J. Han, and P.S. Yu, “Data Mining: An Overview from a Database Perspective,” IEEE Trans. Knowledge and Data Eng., 1996.

[2] N. Mishra, D. Oblinger, and L. Pitt, “Sublinear Time Approximate Clustering,” Proc. 12th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA), 2001.

[3] R.T. Ng and J. Han, “CLARANS: A Method for Clustering Objects for Spatial Data Mining,” IEEE Trans. Knowledge and Data Eng.,2002

[4] P.S. Bradley, U.M. Fayyad, and C. Reina, “Scaling Clustering Algorithms to Large Databases,” Proc. Fourth Int’l Conf. Knowledge Discovery and Data Mining (KDD), 1998.

[5] S. Guha, R. Rastogi, and K. Shim, “ROCK: A Robust ClusteringAlgorithm for Categorical Attributes,” Proc. 15th Int’l Conf. DataEng. (ICDE), 1999.

[6] Z.X.Huang,"Extensions to the k-means algorithm for clustering largedatasets with categorical values," Data Mining Knowl,Discov.,Vol2,no3,pp283-304,1998.

[7] F.Y.Cao,J.Y.Liang , and L.Bai,"A new initialization method for catagorical data clustering ," Expert Syst., Appl.,vol.36,no.7,pp.10223-10228,2009.

[8] UCI Machine learning Repository.(2010)[online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html

[9] S.Viswanadha Raju, H.Venkateswara Reddy, N.Sudhakar Reddy,G.Sreenivasulu, and Dr. KVN Sunitha," Our - NIR : Node Importance Representative for Clustering of Categorical Data" IJCST Vol. 2, Iss ue 2, June 2011.

[10] S. Shekhar, C.-T. Lu, and P. Zhang. Detecting graph-based spatial outliers: Algorithms and applications (a summary of results). In Proc. of KDD’2001, 2001.

[11] P. Berkhin, “Survey of Clustering Data Mining Techniques,” technical report, Accrue Software, 2002.

[12] Aggarwal, C.; Han, J.; Wang, J.; Yu, P. “A Framework for Clustering Evolving Data Streams,” Proc. 29th Int'l Conf. Very Large Data Bases (VLDB), 2003.

[13] Shannon, C.E., “A Mathematical Theory of Communication” Bell System Technical, 1948.

[14] http://www.slideshare.net/dataminingcontent/outlier-analysis