www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 5 May, 2013 Page No. 1688-1691

OUTLIER TEST FOR CATEGORICAL DATA BY USING HYPERGRAPH DEVIATION COMPARISON

[1]H. Venkateswara Reddy [2]S. Viswanadha Raju [3]B. Suresh Kumar [4]C. Jayachandra

[1]Associate Professor in CSE,VCE, Hyderabad, India,
Email: venkat_nidhish@yahoo.co.in

[2]Professor in CSE,JNTUH, Hyderabad, India,
Email: viswanadha_raju2004@yahoo.co.in

[3]M.Tech (C.S.E),VCE, Hyderabad, India ,
Email: sureshkumargoud2006@gmail.com

4M.Tech (C.S.E),VCE, Hyderabad, India,
Email: chinnijayachandra@gmail.com

ABSTRACT

Outlier detection is one most important issue in recent years. 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. In this paper, we propose an approach to find outliers that is Comparison of Deviations. In Comparison of deviation method, hyper graph is used to calculate the deviations of each object in the database. After comparing all deviations which are having most negative term that objects are treated as outliers.

Index terms: Categorical data, Hyper graph, Deviation, Outliers, Hot algorithm**.**

1. INTRODUCTION:

Outlier detection is one of the essential technologies in data mining to detect the outliers. Outlier detection is the process of detecting the data object which is exceptional from the large amount of data. This process is used for Telecommunications, financial fraud detections, data cleaning and improves the quality of the services. The definition of outlier is “The data objects that don’t comply with the general behaviour or model of the data. Such data objects, which are grossly different from or in consistent with remaining set of data are called Outliers”. But according to the Hawkins the definition for the outlier is “An outlier is an observation that deviates so much from other observation as to arouse suspicious that it was generated by a different mechanism “(11). Although some different definitions are specified by the researchers and they faced many problems when they applied for the real time data. Depending Hawkins definition the respected authors Wenjinand, AoyingZhou, and Liwei Weining Qien proposed a method and they used the algorithm Hypergraph-based Outlier Test (HOT) (1) for finding the outliers.

2. OUTLIER MINING METHOD

A novel outlier mining (1) method and Hot algorithm which is based on the hyper graph model for Categorical data. According to HOT algorithm, the process for finding outlier is shown in below steps:

Step 1: Building the hierarchy of the hyper edges.

Step 2: Construct multidimensional array.

Step 3: finding Outlier in the array.

By using HOT algorithm the Outliers can be detected and deviation of data object "o" on attribute "A" is defined as $Dev^{he}(o, A) = \frac{S_A^{he}(x_0) - \mu_{S_A^{he}}}{\sigma_{S_A}^{he}}$ where

$\mu_{S_A^{he}} = \frac{1}{\| A^{he} \|} * \sum_{x \in A} S_A^{he}(x)$ is the average value of $S_A^{he}(x)$ for all $x \in A^{he}$

And $\sigma_{S_A^{he}} = \sqrt{\frac{1}{\| A^{he} \|} * \sum (S_A^{he}(x) - \mu_{S_A^{he}(x)})^2}$ .... (1)

is the standard deviation of $S_A^{he}(x)$ for all $x \in A^{he}$.

Here hyper edge *he* and data object "*o*" in it is defined as an outlier with common attribute C and outlying Attribute A, In which C is the set of attribute that have value appear in the frequent item set corresponding to *he*, if Dev^{he}(o, A) $< \theta$ ....(2)

The threshold of deviation θ determines how abnormal the outlier will be usually θ is set to a negative value (1).

The HOT algorithm find the deviation on the following data.

Table 1

Rid	Name	Age-range	Car-type	Salary-level
1	Mike	Middle	Sedan	Low
2	Jack	Middle	Sedan	High
3	Mary	Young	Sedan	High
4	Alice	Middle	Sedan	Low
5	Frank	Young	Sports	High
6	Linda	Young	Sports	Low
7	Bob	Middle	Sedan	High
8	Sam	Young	Sports	Low
9	Helen	Middle	Sedan	High
10	Gary	Young	Sports	Low

Construction of Hype graph for Table 1 is shown in Table 2.

Table 2: Hyper graph modelling

HyperedgeID	Frequent itemsets	Vertices
1	('Middle',*,*)	1,2,4,7,9
2	('Young',*,*)	3,5,6,8,10
3	(*,'Sedan',*)	1,2,3,4,7,9
4	(*,*,'Low')	1,4,6,8,10
5	(*,*,'High')	2,3,5,7,9
6	('Middle', 'Sedan',*)	1,2,4,7,9

Symbols Used:

Notion	**Meaning**
N	The number of objects in database DB
||DS||	The number of elements in set DS
A,A$_i$	Each denotes an attribute.
B,C	Each denotes a set of attributes.
V_o^i	The value of attribute A$_i$ in object o.
A^{DS}	The set of values of A appearing in dataset DS. Then A^{he} and A^{DB} denotes the A's values appear in hyper edge he and whole database respectively
S$_A^{DS}$(x)	Given x ∈ A and dataset DS, it is the number of objects in DS havingValues x in A. Similar to A^{DS}, S$^{he}_A$ and S$^{DB}_A$ are defined respectively.

By applying deviation on this data objects 3,5,6,8 and 10 on attribute Car-type. The result will be shown in the below table.

Table 3

Rid	**Name**	**Age-range**	**Car-type**	**Salary-level**	**De1**
3	Mary	Young	Sedan	High	-1
5	Frank	Young	Sports	High	1
6	Linda	Young	Sports	Low	1
8	Sam	Young	Sports	Low	1
10	Gary	Young	Sports	Low	1

Here De1= Dev^{he}(o, Car-type)

According to Hawkins Outlier definition object 3 is treat as an outlier, for the deviation value of object 3 is -1.

3. Problem Statement

The existing method select objects as car-type in data record attribute Age-range as young. By applying the same procedure considering object as Salary-level the resultant values are shown in Table 4.

Table 4

Rid	**Name**	**Age-range**	**Car-type**	**Salary-level**	**De2**
3	Mary	Young	Sedan	High	-1
5	Frank	Young	Sports	High	-1
6	Linda	Young	Sports	Low	1
8	Sam	Young	Sports	Low	1
10	Gary	Young	Sports	Low	1

De2= Dev^{he}(o, Salary-level)

According to HOT algorithm object 3 and 5 treated as the outliers. If the procedure continues on large databases then there is a chance of getting more outliers. The proposed Comparison of deviation method is used for outlier detection on large categorical databases effectively.

4. Comparison of Deviations

In this method we have to compare the deviation value of the data records.

Table 5 shows the deviation values of object 3 and 5. In the above table the deviation value θ is negative for object 3 in car-type and salary-level attributes and for object 5 the deviation value θ is negative in attribute salary-level.

According to the HOT algorithm object 3 and 5 consider as outliers. But, the proposed method comparison of deviation which gives the number of negative values that objects are consider as Outliers. The object 3 has more negative terms so that it is treated as outlier in the data record.

Table 5

Rid	**Name**	**Age-range**	**Car-type**	**Salary-level**	**Dev^{he} (o, Car-type)**	**Dev^{he} (o, Salary-level)**
3	Mary	Young	Sedan	High	-1	-1
5	Frank	Young	Sports	High	1	-1
6	Linda	Young	Sports	Low	1	1
8	Sam	Young	Sports	Low	1	1
10	Gary	Young	Sports	Low	1	1

After filtering of outliers it is important that finding the importance for every object in data record for that we are using Chen node importance method.

The importance value of the n-node set I^n_{ir} is calculated as follows (2):

$$w(c_i, I_{ir}) = \left(\frac{|I_{ir}|}{m_i} \right) * f(I^n_r)$$

$$f(I^n_r) = 1 - \left(\frac{-1}{\log k} \right) * \sum_{y=1}^{k} P(I^n_{yr}) \log(P(I^n_{yr}))$$

$$P(I^n_{yr}) = \frac{|I^n_{yr}|}{\sum_{z=1}^{k} |I^n_{zr}|}$$

With using of these equations we can find the importance of each data point in the data set or records.

Symbols Utilized in this Paper:

$W(C_i, I_{ir})$	The importance of I_{ir} in c_i
$\|I_{ir}\|$	The number of occurrence of I_{ir}
m_i	The number of data points in C_i

If the above equations are applied to the Table 3, the node importance of each object in data record is as following:

If the person age is middle and his Salary is Low then the importance for buying Sedan is

w (c_i, I^n_{ir}) =0.4.

If the person Age is middle and Salary is High then the importance for buying Sports car-type is w (c_i, I^n_{ir}) =0.6.

If the person Age is young and Salary is Low then the importance for buying Sports car-type is w (c_i, I^n_{ir}) =0.6.

If the person Age is young and Salary is High then the importance for buying Saden car-type is w (c_i, I^n_{ir}) =0.39.

If the person Age is young and Salary is High then the importance for buying Sports car-type is w (c_i, I^n_{ir}) =0.39.

By using node importance method all n-node sets the unlabeled data objects can be labelled.

6. Conclusion

Comparison of deviation method is used for finding the outliers in effective manner. This helps for detecting outlier in large type of categorical data. The procedure is to compare the all the deviations in data record by using hypergraph in that which gives the more negative deviation terms that will be treated as outliers. By using node importance method the unlabeled data objects can be labelled.

7. References:

[1] Aoying Zhou,LiWei,Weining Qian,Wenjin .HOT:Hypergraph-baesd Outlier for Categorical Data.

[2] Hung-Leng Chen ,Kung-Ta Chuang,Member,IEEE,and Ming-Syan Chen,Fellow ,IEEE,On Data Labeling for Clustering Categorical Data,November 2008

[3] C. Aggarwal and P. Yu. Outlier detection for high dimensional data. In Proc. Of SIGMOD’2001, pages 37–47, 2001.

[4] Quinlan, J. R., (1986). Induction of Decision Trees. Machine Learning 1: 81-106, Kluwer Academic Publishers

[5] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Proc. of VLDB’94, pages 487–499, 1994.

[6] V. Barnett and T. Lewis. Outliers In Statistical Data. John Wiley, Reading, New York, 1994.

[7] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. Optics-of: Identifying local outliers. In Proc. of PKDD’99, pages 262–270, 1999.

[8] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. Lof: Identifying density-based local outliers. In Proc. of SIGMOD’2000, pages 93–104, 2000.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proc. of KDD’96, pages 226–231, 1996.