IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 2 (May. - Jun. 2013), PP 101-106
www.iosrjournals.org

Hyper Graph Deviation Comparison & Cluster Relation In Categorical Data

Anvesh K[1], V Bhavya[2], B Suresh Kumar[3]
[1]*Assistant Professor, Department of IT, Vardhaman College of Engineering, Hyderabad*
[2]*Assistant Professor, Department of CSE, Vardhaman College of Engineering, Hyderabad*
[3]*Assistant Professor, Department of CSE, Vardhaman College of Engineering, Hyderabad*

Abstract: *Outlier detection is one of the most important issues in recent years. Outlier detection is the process of detecting errors in data. The recent methods are mostly based on 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 outliers in categorical data. In this paper, we propose an approach to find outliers those are Comparison of Deviations. In Comparison of deviation method, we use hyper graph to calculate the deviations of each object in the database, and we measure the similarity between attributes in database.*
Key words: *Categorical data,Hyper graph, Deviation, Outliers, Hot algorithm, Similarity objects.*

I 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.

II 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

NOTIONS AND SYMBOLS

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	De 1
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.

III 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	De 2
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.

IV 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.

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

$$f(I_r^n) = 1 - \left(\frac{-1}{\log k} \right) * \sum_{y=1}^{k} P(I_{yr}^n) \log(P(I_{yr}^n))$$

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

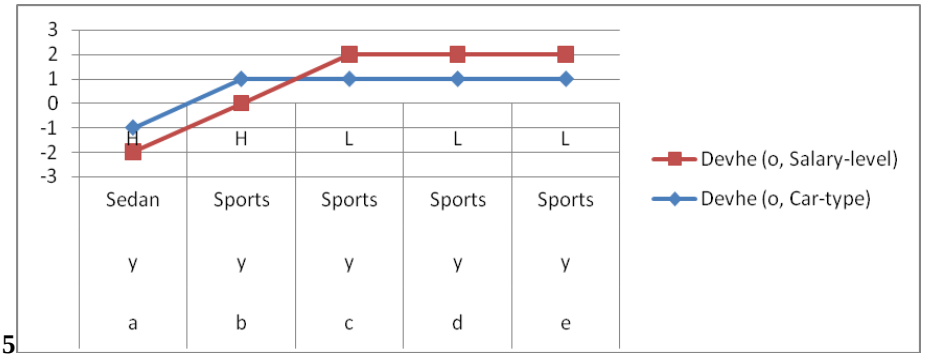
After filtering of outliers it is important that finding the importance for every object in data record for that we are using Chen node importance method.

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

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

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

Graph generated after simulation of deviation values of object 3 and



5

Logical Notations

$W(C_i, I_{ir})$	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:

(a) 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.

(b) 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.

(c) 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.

(d) 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.

(e) 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.

V Relation Between The Clusters And Objects

In this Section we are going to find the relationship between the modes (13). Depending on the node importance value, we can measure the relationship between the clusters and the object in the same cluster.

To find the relationship between the clusters, select a node from the clusters which has high importance in the particular cluster and apply similarity measure to that mode. The dissimilarity measures are as follows.

Definition:Let IS = (U, A, V, f) be a categorical information system,and P $\subseteq$ A. For any a $\in$ P and x, y $\in$ U, a dissimilarity measure betweenobjects x and y with respect to a is defined as

$$dis_a(x,y) = \mu_{\{y\}}^{\{a\}}(x) = \frac{\left|[x]_{\{a\}} \cap \{y\}\right|}{\left|[x]_{\{a\}}\right|} \text{.....[4]}$$

In definition, the domain of the rough membership function is an object y of U, not the universe U. The degree of relative overlap between the object x and the object y means the dissimilarity between the object x and the object y. The dissimilarity $dis_a(x, y)$ can be also described as:

$$dis_a(x,y) = \frac{f(x,a) \equiv f(y,a)}{\sum_{z \in U} f(x,a) \equiv f(y,a)} \text{.....[5]}$$

Where

$f(x,a) \equiv f(y,a) = 1$ If(x, a)=f(y, a)..... [6]

$f(x,a) \equiv f(y,a) = 0$ otherwise..... [7]

Following is a definition of similarity between two objects over several attributes defined in terms of dissimilarity between objects..

Let IS = (U, A, V, f) be a categorical information system, and P $\subseteq$ A. For any x, y $\in$ U, the similarity measure between x and y with respect to P is defined as:

$$d_P(x,y) = \sum_{a \in P} d_a(x,y) \text{.....[8]}$$

where

$$d_a(x,y) = 1 - dis_a(x,y) \text{.....[9]}$$

If we apply these methods in our clustersC_1,C_2 as follows

Cluster C_1 **Table 6**

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

Cluster C_2 Table 7

Rid	Name	Age-range	Car-type	Salary-level
1	Mike	Middle	Sedan	Low
2	Jack	Middle	Sedan	High
4	Alice	Middle	Sedan	Low
7	Bob	Middle	Sedan	High
9	Helen	Middle	Sedan	High

From the database we can rename the names for easy analysis, that is
{Mike, Jack, Mary, Alice, Frank, Linda, Bob, Sam,Helen,Gary}={x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9,x_{10}}

So from the database we have two clusters C_1,C_2. In Cluster C_1 {x_3,x_4,x_5} are having the maximum importance and in Cluster C_2 {x_7,x_9,x_{10}} are having maximum importance .

So from Cluster C_1select $x_4$and from C_2select x_7 as mode of the cluster. From the definition of Cluster Similarity $d_p(x_4, x_7) = 3$, which means that we can't find any similarity between these two clusters.

VI Relationship Between The Nodes

If we apply the process for every node in the different clusters, it will be as shown in the table below.

Table 8

(x_3,x_7)	**3**
(x_3,x_9)	**3**
(x_3,x_{10})	**3**
(x_4,x_7)	**3**
(x_4,x_9)	**3**
(x_4,x_{10})	**3**
(x_5,x_7)	**3**
(x_5,x_9)	**3**
(x_5,x_{10})	**3**

We can conclude that there is no relationship between these two clusters. We can also apply this same cluster also. So we can find the relation or similarity between the nodes in the same cluster.

Algorithm

Hypergraph deviation method using node similarity

Input : Dataset D, min_sup,θ
Output : Outlier (o,C,A),Similarity(x_a,x_b).
// C is set of common attributes
// A is the outlying attribute

Method

Step 1: Apply any clustering algorithm (Forexample, EM clustering) on D and find clusters c_1, c2 ...cn.
Step 2: Build Hypergraph and construct the hierarchy.
Step 3: For each node i in Cluster c_n in every level Use [1] for outlier detection in a cluster.
Step 4: End for.
Step 5: ApplyNode importance [2] for node i in every cluster.
Step 6: Then select the node for which node importance value is high.
Step 7: Select a data point (or) node as mode of the cluster in every attribute with maximum nodeimportance.
Step 8: Repeat step 7 for every cluster generated in step 2.
Step 9: The relationship between two clusters is same as similarity between modes of the cluster then find this relationship between cluster using [9].
Step 10: End.

VII Conclusion

In this paper we proposed a method for finding the outliers in Clusters by comparison of deviation method. This helps in detecting outliers in large type of categorical data, and we apply similarity method between two clusters to determine the relationship or similarity between clusters.

References

[1]. Aoying Zhou,LiWei,Weining Qian,Wenjin .HOT:Hypergraph-baesd Outlier for Categorical Data.
[2]. Hung-LengChen,Kung-TaChunag,Member IEEEandMing-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. Aggarwal 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.
[10]. S. Ramaswamy, R. Rastogi, and K. Shim. Efficient algorithms for mining outliers from large data sets. In Proc. of SIGMOD'2000, pages 427–438, 2000.
[11]. 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.
[12]. Ruts and P. Rousseeuw. Computing depth contours of bivariate point clouds. Journal of Computational Statistics and data Analysis, 23:153–168, 1996.
[13]. Fuyuan Cao , Jiye Liang , Deyu Li , Liang Bai , Chuangyin Dang A dissimilarity measure for the k-Modes clustering algorithm in Knowledge-Based Systems 26 (2012) 120–127