Data Mining: Finding Outliers from Different Types of Data using Dissimilarity Data Structure

L. Sunitha[1], M. BalRaju[2], J. Sasikiran[3]

[1]Department of computer Science and Engineering
Vidya Vikas Institute of Technology, Hyderabad, India
sunitha_lingam@yahoo.com

[2]Department of computer Science and Engineering
Vidya Vikas Institute of Technology, Hyderabad, India
jb_raju_cse@yahoo.co.in

[3]Department of computer Science and Engineering
Vidya Vikas Institute of Technology, Hyderabad, India
jsasikiranj@yahoo.co.in

Abstract: ***An Outlier is an extreme value in a data set. Using clustering techniques we can detect outliers. Outlier means values that are far away from any cluster. In this paper we tried to find out outliers from Inter-Scaled Variables, Binary Variables, Categorical and Ordinal Variables by using Dissimilarity Data Structure. All similar objects are grouped and objects which are not belonging into any cluster are considered as outliers.***

Keywords: Outlier, Dissimilarity, Cluster, Inter-Scaled, Binary, Categorical

1. Introduction

Outlier detection [1][2] is one of the interesting tasks in data mining. Outlier is not a noisy, so it is not removed from data sets and it should be analyzed. Outlier mining has wide applications [3][4] involves in fraud detection, detecting of unusual usage of credit cards or telecommunication services. The identification of outliers has also received much attention from the computing community .However; there appear to be much less work on how to decide whether outliers should be retained or rejected. In statistical community, a commonly-adopted strategy when analyzing data is to carry out the analysis both including and excluding the suspicious values.

1.1 Outlier definition and Example

An outlier is an observation that lies an abnormal distance from other values. The definition leaves it up to the analyst to decide what will be considered abnormal. For finding abnormal observations it is necessary to characterize normal observations. Outliers should be investigated carefully. Example a student data set containing roll no and marks consider the following table.

Rollno	1	2	3	4	5	6
Marks	45	56	43	58	90	58

Here rollno 1,3 are grouped into one cluster (c1) and 2,4,6 are grouped into cluster(c2), rollno 5 the marks 90 are far from remaining objects. So here it is considered as outlier.

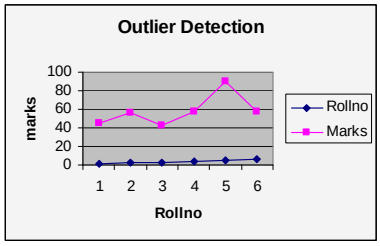

Figure 1: Line Chart for outlier

Dissimilarity Matrix: Dissimilarity Matrix [5] stores a collection of proximities that are available for all pairs of n objects. It is represented by an n-by-n table. where d(i ,j) is the dissimilarity between i and j.

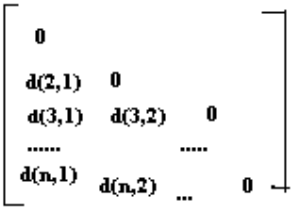

Figure 2: Dissimilarity Matrix

In general d (i, j) is a nonnegative number that is close to 0 when objects i and j are highly similar, and becomes larger the more they differ. Since d(i ,j)=d(j ,i)and d(i ,i)=0.

2. Finding Outlier from Inter-Scaled Variables

Inter-Scaled variables are continuous values that can be measured on a linear scale examples height, weight, temperature etc. The dissimilarity between the objects described by Interval Scaled Variables is measured buy

popular distance measures like Euclidean distance and Manhattan distance [6]

(1) Euclidean distance

$$d(i,j)= \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \text{---- (1)}$$

(2) Manhattan distance

$$d(i,j)= \sum_{i=1}^{n}(x_i - y_i)^2 \quad \text{---- (2)}$$

Example:

SID	1	2	3	4	5	6
Name	John	Rama	Shiva	Venkat	Smith	Raja
Weight	60	73	64	75	100	65

Manhattan distance

d(1,2)=|60-73|=13, d(1,3)=|60-64|=4, d(1,4)=|60-75|=15,d(1,5)=|60-100|=40,
d(1,6)=|60-65|=5,d(1,2)=|60-65|=5, d(2,3)=|73-64|=9,d(2,4)=|73-75|=2, d(2,5)=|73-100|=27,d(2,6)=|73-65|=12,
d(3,4)= |64-75|=11, d(3,5)= |64-100|=36, d(3,6)= |64-65|=1,d(4,5)= |75-100|=25, d(4,6)= |75-65|=10, d(5,6)= |100-65|=35

|0 |
|13 0 |
| 4 9 0 |
|15 2 11 0 |
| 40 27 36 25 0 |
| 5 12 1 10 35 |
--- ---

Hence sid 1, 3 and 6 are one cluster and sid 2 and 4 are one cluster, but sid 5 is consider as outlier dissimilarity is larger.

3. Binary Variables

Binary variable has only 2 states: 0 and 1 or T and F or Y and N. To compute dissimilarity between the objects described by binary variables we consider a contingency table [7] represented 2 by 2 matrix

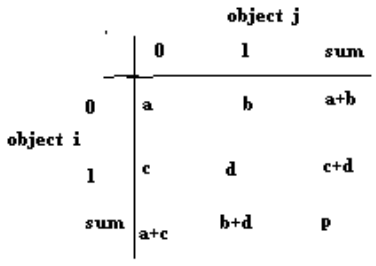

Figure 3: Contingency table

If both the states are equally valuable and carry same weight then it is a Symmetric binary variable that is, there is no preference on which outcome should be coded as 0 or 1, Example: Gender (states MALE and FEMALE). Asymmetric if the states are not equally important ie do not carry same weight. By convention, we shall code the most important outcome, which is usually the rarest one, by 1 (e.g., HIV positive), and the other by 0 (e.g., HIV negative).

Dissimilarity between objects i and j for Symmetric Binary Variable

d(i ,j)= (b + c)/(a + b+ c +d)

Dissimilarity between objects i and j for Asymmetric Binary Variable

d(i ,j)= (b+ c)/(a + b +c)

Example

Name	*gender*	*fever*	*cough*	*Test-1*	*Test-2*	*Test-3*	*Test-4*
Jack	M	y	N	P	N	N	N
mary	F	y	N	P	N	P	N
Jim	M	y	P	N	N	N	N

Gender is a symmetric attribute and the remaining attributes are asymmetric binary Let the values Y and P be set to 1, and the value N be set to 0

Name	gender	fever	cough	Test 1	Test 2	Test 3	Test 4
Jack	M	1	0	1	0	0	0
mary	F	1	0	1	0	1	0
Jim	M	1	1	0	0	0	0

d (i, j) = (b+ c)/(a + b +c)
d (jack, Mary) = (0+1)/ (2+0+1)=0.33
d (jack, Jim) = (1+1)/ (1+1+1)=0.67
d (Mary, Jim) =(2+1)/(1+1+2)=0.75

Dissimilarity between jack and Mary is minimum. Hence they can be put in to one cluster and mary, jim and jack, jim dissimilarity is maximum so Jim is an outlier i.e. it is not similar to other 2 objects.

4. Nominal / Categorical Variables

A generalization of the binary variable that can take more than 2 states, e.g: Variable Colour can take values red, yellow, and blue, green

d(i ,j) =(p –m)/p

Where p: - total no: of categorical variables in the dataset

m:- no: of matches i.e. i and j are in same state

Example:

Empid	*Age*	*Work Category*
E1	Young	Permanent
E2	Middle	contract
E3	young	permanent
E4	Middle	contract
E5	senior	contract

No: of Var = p = 2

d(E1,E2)=(p-m)/p=(2-0)/2=1,E1,E2 are not similar
d(E1,E3)=(2-2)/2=0(means dissimilarity is zero),so these are similar.d(E1,E4)=(2-0)/2=1(means dissimilarity is non zero),so these are not similar.
d(E1,E5)=(2-0)/2=1

Dissimilarity is non zero, so these are not similar.
d(E2,E3)=(2-0)/2=1
Dissimilarity is non zero, so these are not similar.
d(E2,E4)=(2-0/2=0
Dissimilarity is zero, so these are similar.
d(E2,E5)=(2-0)/2=1
Dissimilarity is non zero, so these are not similar.
d(E3,E4)=(20)/2=1
Dissimilarity is non zero, so these are not similar
d(E3,E5)=(2-0)/2=1
Dissimilarity is non zero, so these are not similar.
d(E4,E5)=(2-0)/2=1 Dissimilarity is non zero, so these are not similar

Therefore the cluster one c1(E1,E3), cluster two c2(E2,E4) and E5 not belonging to any cluster ,hence it is considered as outlier.

5. Ordinal Variables

Ordinal variables are similar to categorical variable except that the states are ordered in a meaning full sequence.

Example1: medals of sport have states in sequence as gold, silver, bronze.

Htno	20	21	22	23	24	25
Grade	0.33	0	1	0.33	0.66	0

Example 2: designations in department are professor, associate professor, assistant professor.

5.1Procedure for finding outliers from ordinal variables

Step1: If v is ordinal variable having Mv ordered states representing a ranks riv =1,2,3,….., Mv .
Step2: If Xiv is value of v for the i^{th} object , then replace Xiv with its rank riv
Step 3: Normalize the ranks such that they fall in to a fixed range between [0.0 , 1.0] by using following formula

Ziv = riv - 1 / Mv – 1

Step 4: Dissimilarity is computed using any distance measure like Euclidean or Manhattan on Ziv value as used in interval scaled variables.

Example:

Htno	20	21	22	23	24	25
Grade	First	distinction	fail	first	second	distinction

Step 1: Here Result is an ordinal variable with 4 states i.e Mv = 4.

Let their ranks be Distinction =1, First Class = 2, Second Class = 3, Fail = 4

Step 2: Replace the ordinal variable values with its corresponding ranks. By this we have:

Htno	20	21	22	23	24	25
Grade	2	1	4	2	3	1

Step 3: Normalize the rank values by Ziv = riv - 1 / Mv – 1.

Z1v = (1-1)/(4-1) = 0/3 = 0
Z2v= (2-1)/(4-1) = 1/3 = 0.33
Z3v = (3-1)/(4-1) = 2/3 = 0.66
Z4v = (4-1)/(4-1) = 3/3 = 1

Now we replace the rank values with the calculated normalized values, by which we have

Htno	20	21	22	23	24	25
Grade	0.33	0	1	0.33	0.66	0

Step 4: now we use the manhatan distance measure to compute the dissimilarity matrix.

Manhattan distance

$$\mathbf{d(i,j)=} \sum_{i=1}^{n} (x_i - y_i)^2$$

d(20,21) = |0.33-0| = 0.33 ,d(20,22) = |0.33-1| = 0.67Similar calculations will give the matrix as:

	20	21	22	23	24	25
20	0					
21	0.33	0				
22	0.67	1	0			
23	0	0.33	0.67	0		
24	0.33	0.66	0.34	0.33	0	
25	0.33	0	1	0.33	0.66	0

21,25 20,23 24 22

20 and 23 dissimilarity is 0 , so these are belonging into one cluster, 21 and 25 dissimilarity is 0 , so these are belonging into one cluster and 24 , 22 are not similar to any other objects .Hence these considered as outliers.

6. Conclusion

The original outlier detection methods were arbitrary but now, principled and systematic techniques are used, drawn from the full gamut of Computer Science and Statistics. In this paper, we are discussed the finding outliers from various data types using dissimilarity structure matrix .The basic idea is dissimilarity of two objects is zero means those objects are similar. Dissimilarity is greater than zero means not similar. In this way similar objects are grouped into cluster and object not belonging to cluster or outliers.

References

[1] Outlier detection Irad Ben-Gal Department of Industrial Engineering Tel-Aviv University Ramat-Aviv, Tel-Aviv 69978, Israel.bengal@eng.tau.ac.il http://www.eng.tau.ac.il/~bengal/outlier.pdf
[2] Outlier Detection: A SurveyVARUN CHANDOLA University of Minnesota ARINDAM BANERJEE University of Minnesota and VIPIN KUMAR University of Minnesota, http://www.bradblock.com.s3-website-uswest1.amazonaws.com/Outlier_Detection_A_Survey.pdf
[3] Outlier Detection: Applications And Techniques, Karanjit Singh and Dr. Shuchita Upadhyaya ,HQ Base Workshop Group EME ,Meerut Cantt, UP, India, Kurukshetra University, Haryana, India, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January2012,http://ijcsi.org/papers/IJCSI-9-1-3-307-323.pdf
[4] Outlier Detection Techniques, Hans-Peter Kriegel, Peer Kröger, Arthur Zimek
[5] http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Clustering/Dissimilarity_Matrix_Calculation
[6] Manhattanhttp://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Manhattan_Distance_Metric.htm
[7] Text book Data Mining: Concepts and Techniques second Edition JiaweiHan University of lllionis at Urbana-Champaign Micheline Kamber.