Kalpa Publications in Engineering

Volume 2, 2018, Pages 1–6

Proceedings on International Conference on Emerging Trends in Expert Applications & Security (2018)

kalpa Engineering

Detection of Anomalous Value in Data Mining

Darshanaben Dipakkumar Pandya[1], Dr.Sanjay Gaur[2]

[1]Research Scholar, Department of Computer Science, Madhav University, Pindwara, Sirohi,
Rajasthan.

[2]Associate Professor, Department of Computer Science & Engineering, Jaipur Engineering College
And Research Center, Jaipur.

[1]pandya_darshana@rediffmail.com, [2]sanjay.since@gmail.com

Abstract

In the database of numeric values, outliers are the points which are different from other values or inconsistent with the rest of the data. They can be novel, abnormal, unusual or noisy information. Outliers are more attention-grabbing than the high proportion data. The challenges of outlier detection arise with the increasing complexity, mass and variety of datasets. The problem is how to manage outliers in a dataset, and how to evaluate the outliers. This paper describes an advancement of approach which uses outlier detection as a pre-processing step to detect the outlier and then applies rectangle fit algorithm, hence to analyze the effects of the outliers on the analysis of dataset.

Keywords: Data mining, Anomalous Values, Attribute, rectangle fit algorithm, Quartiles.

1 Introduction

An anomalous value in database is solitary of the principle problems featured in data analysis and in the prediction. The belongings of these anomalous values are highly reflected on the final results. Our chief goal is to achieve the final result without error in the consolidated form, which is use to take decisions. There are various forms of anomalous values in the database, among those; outlier values case is one of the harder cases to recover.

In this study, a method of outlier detection is introduced and discussed which provides an approach to treat anomalous values. This step treats the anomalous block of values from a real- world imbalanced database.

To illustrate, some real application case for outlier detection are portray underneath.

(1) Outliers are used for detection of computer hardware speed: Suppose that it usually takes about four seconds to download a gigabyte file from a main server, but one day the system becomes slower, instead, eight seconds are needed to perform the same task. While eight seconds may indicate a good performance, it is nonetheless, helpful to find the source of the delay in order to prevent more critical faults in the future. In this case, the download operation is the outlier while the delay is its witness.

(2) Mechanical failure in real world: Assume that someone's bike brakes are making a strange noise. Although they seem to be functioning properly, this is not a normal behavior and the bike is brought in for servicing. In this case, the bike brakes are the outlier and the noise is a witness for it.

(3) Integrity based on knowledge: If an abnormal property is discovered in a database, the source that reported this information would have to be checked from database. Detecting abnormal properties, that is, detecting outliers can also cause to an update of default rules in a knowledge base, i.e. suppose there is a rule that birds fly, and notice a bird that does not fly.

This occurrence of such an outlier in the theory would be reported to the programmer. The programmer investigates the case, finds out that the bird is actually a penguin; therefore he updates the knowledge base with the default "penguins do not fly."

2. Background on Anomalous Data

In this study, a statistical method is discussed which provides an approach to find out pattern to discover anomalous values from a real imbalanced database with massive anomalous values. Therefore, the objective of this method is to discover the best fitted value for the anomalous value and select records completely by removing outliers. The function of statistical methods has gained stuff in exploring estimation and prediction techniques. Zhang, S., Zhang, C., and Young [1] are the authors who have introduced data preparation for data mining and Applied Artificial Intelligence, HP Kriegel, P Kröger, A Zimek [2] are the persons who have discussed Outlier Detection Techniques. Yang Zhang, Nirvana Meratnia, Paul Havinga [3] are the persons who have introduced outlier detection techniques for Wireless Sensor Networks and the sensor nodes are integrated with sensing, processing and wireless communication capabilities. Charu C. Aggarwal, Philip S. Yu [4] are the authors who have introduced outlier detection for high dimensional data, Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad, Mustafa Mat Deris [5] are the pioneer Statisticians, who have created a Survey of Outlier Detection Methodologies and outlier detection . Zuriana Abu Bakar ,Rosmayati Mohemad , Akbar Ahmad , Mustafa Mat Deris [6] suggested a Comparative Study for outlier Detection Techniques in Data Mining and Experimental studies show that outlier detection technique using control chart is better than the technique modeled from linear regression , P.García-TeodoroaJ.Díaz-VerdejoaG.Maciá-FernándezaE.Vázquezb[7] are the scientists who invested Anomaly- based network intrusion detection Techniques , systems and challenges. Shohei Hido, Yuta Tsuboi,Hisashi Kashima, Masashi Sugiyama,Takafumi Kanamori[8] are the Statisticians, who have considered Statistical outlier detection using direct density ratio estimation of parameters and finding outliers in the test set based on the training set consisting only of inliers. The key idea is to use the ratio of training and test data densities as an outlier score. This approach is expected to have better performance even in high dimensional problems since methods for directly estimating the density ratio without going through density estimation are available, C. Aggarwal and P. Yu[9] are investigated outlier Detection for High Dimensional Data, Z. He, X. Xu and S. Deng [10] discussed various algorithms cluster based Local Outliers. The objective of proposed study is to determine the statistical technique which may be significant in the handling of anomalous attribute values.

3. Outlier Analysis

An *outlier* is an observation that lies an abnormal distance from other values in a random sample from a population .In data mining outlier analysis can be done by various methods. The proposed method is based on replacing outlier attribute values discovered and then remove the data having outlier parentally from the data set. This method is very much useful for numerical attributes. In general, this method is search of rectangle fit value which is very close to the true mean of the attribute. Three main types of Outliers.

(1)Point Outliers: Observations anomalous with respect to the majority of observations a feature Like univariate outlier. **(2) Contextual Outliers**: Observations considered anomalous given a specific context. **(3)Collective Outliers**: A collection of observations anomalous but appear close to one another because they all have a similar anomalous value. Results of the outlier analysis are shown in below figure.

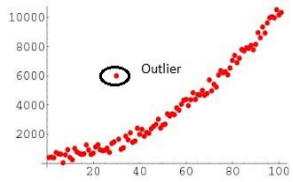


Fig. 1. Plot between different income and expenses in a company.

The plot shows the distribution of points between the two features of our data income length and expenses length, and the highlighted point in red circle shows the outlier point. The descriptive analysis applied on data. Following results, shown in table 1. Whereas that of analysis of data is shown in table 3.

Descriptive Statistics (with outlier)

Qty	N	Q1	Q3	Mean	Median	S.D
(list wise)	15725	31.5	83.25	524.1667	55.5	1367.557

Table 1: Descriptive analysis of the data based on table 3 (with outlier)

Descriptive Statistics (without outlier)

Qty	N	Q1	Q3	Mean	Median	S.D
(list wise)	1270	24	76.5	48.84	48	28.80

Table 2: Descriptive analysis of the data based on table 3 (without outlier)

There can easily observe here that the mean calculated in Table 1 is 524.1667 and the Median value is 55.5 i.e. much greater than the mean whereas in Table 2. It is nearer to the mean but statistically observe the z scores of the data length calculated by the analysis, these scores are different and if sum of without outlier is less than the sum of with outlier, then remove the data having outlier permanently from the data set. The below table 3 shows Rectangle fit approach of the dataset with outlier and outliers treatment by removing it from database.

Ratio Based outlier detection							
OUTLIER CALCULATION EXAMPLE							
STANDARD DATA WITH OUTLIERS				OUTLIERS TREAMENT BY REMOVING IT FROM DATA BASE			
ID	ITEM	QTY	OUTLIERS FINDING	ID	ITEM	QTY	OUTLIERS REMOVED QTY
1	MOTHERBOARD	85	FALSE	1	MOTHERBOARD	85	85
2	RAM	81	FALSE	2	RAM	81	81
3	ROM	97	FALSE	3	ROM	97	97
4	CD DRIVES	66	FALSE	4	CD DRIVES	66	66
5	VIDEO CARD	65	FALSE	5	VIDEO CARD	65	65
6	MODEM	22	FALSE	6	MODEM	22	22
7	NETWORK DEVICES	78	FALSE	7	NETWORK DEVICES	78	78
8	PRINTERS	44	FALSE	8	PRINTERS	44	44
9	SCANNER	84	FALSE	9	SCANNER	84	84
10	UPS	59	FALSE	10	UPS	59	59
11	SPEAKERS	52	FALSE	11	SPEAKERS	52	52
12	VIDEO CARD	41	FALSE	12	VIDEO CARD	41	41
13	MOTHERBOARD	36	FALSE	13	MOTHERBOARD	36	36
14	ROM	4	FALSE	14	ROM	4	4
15	PRINTERS	14	FALSE	15	PRINTERS	14	14
16	SCANNER	2600	TRUE	16	SCANNER	2600	___
17	CD DRIVES	86	FALSE	17	CD DRIVES	86	86
18	VIDEO CARD	30	FALSE	18	VIDEO CARD	30	30
19	MODEM	72	FALSE	19	MODEM	72	72
20	NETWORK DEVICES	3	FALSE	20	NETWORK DEVICES	3	3
21	PRINTERS	41	FALSE	21	PRINTERS	41	41
22	SCANNER	52	FALSE	22	SCANNER	52	52
23	UPS	6355	TRUE	23	UPS	6355	__
24	VIDEO CARD	78	FALSE	24	VIDEO CARD	78	78
25	MOTHERBOARD	4	FALSE	25	MOTHERBOARD	4	4
26	ROM	20	FALSE	26	ROM	20	20
27	PRINTERS	36	FALSE	27	PRINTERS	36	36
28	SCANNER	20	FALSE	28	SCANNER	20	20
29	UPS	2500	TRUE	29	UPS	2500	__
30	SPEAKERS	3000	TRUE	30	SPEAKERS	3000	__
SUM		15725		SUM		1270	
MEAN		524.17		MEAN		524.16667	48.84615385
MEDIAN		55.5		MEDIAN		55.5	48
MODE		78		MODE		78	78
Q1		31.5		Q1		31.5	24
Q3		83.25		Q3		83.25	76.5
IQR		51.75		IQR		51.75	52.5
UPPER BOUND		160.88		UPPER BOUND		160.875	155.25
LOWER BOUND		46.13		LOWER BOUND		-46.125	-54.75
S.D		1367.6		S.D		1367.5587	28.80651636

Table 3: Rectangle fit approach of the dataset with outlier.

4. Proposed Approach

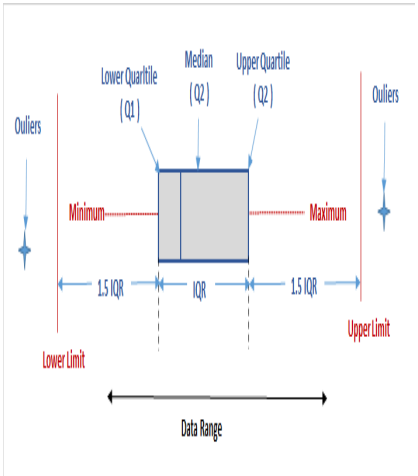
As there can be reviewed several different ways of detecting outliers here proposed a method which is a combination of different approaches, statistical and data mining. Firstly apply outlier detection and then Rectangle fit algorithm to group the data into parts for discovering outliers. Following are the steps of the algorithm in detail:

Step 1: Select a dataset on which outlier detection is to be performed from the database.

Step 2: Determine the first (Q1) and third quartiles of the data (Q3) given from dataset.
Step 3: Compute the inter quartile range (Q3-Q1) given from dataset.
Step 4: Determine the fences. Fences serve as cutoff points for determining outliers from the database. Lower fence A = Q1 – 1.5 (Compute Inter quartile Range)
Upper fence B = Q3 + 1.5 (Compute Inter quartile Range).
Step 5: If a data value is less than the lower fence A or greater than the upper fence B, it is considered an outlier in dataset.
Step 6: Then run the rectangle fit algorithm, for the dataset with and without the data Having outlier, using replicates in order to select proper centroids so as to overcome
The problem of local minima from the dataset.
Step 7: Compare the results for the sum of point to centroid distances from the Dataset.
Step 8: If **Sum (without outlier) < Sum (with outlier),** then remove the data entry Having outlier permanently from the dataset.

5. Experimental Results

There can be a hypothetical data which has been made by introducing some outlier values in the well known sales data. The above table 3 shows Rectangle fit approach of the dataset with outlier. Now must delete the outlier entry and save both the dataset i.e. with outlier entry and without outlier entry and run further the rectangle fit approach to do the analysis of the data and calculate the sum of points to the value in each case. After calculating both the values will compare them to check whether the presence of outlier increases the sum or not, if it increases then it must be removed. Value of the sum of all points in the if Sum (without outlier) < Sum (with outlier), then remove the data entry having outlier permanently from the dataset. Here in above example if sum (1270), Value of the sum of all points in the standard data without outlier is comes out to be, which is much lesser than the sum value of the dataset with outlier entries. Hence delete the outlier entry permanently from the dataset, because this entry is not at all useful and distorting our original dataset. The below fig. 2. And fig. 3. Shows the rectangle fit approach of the dataset with outlier.



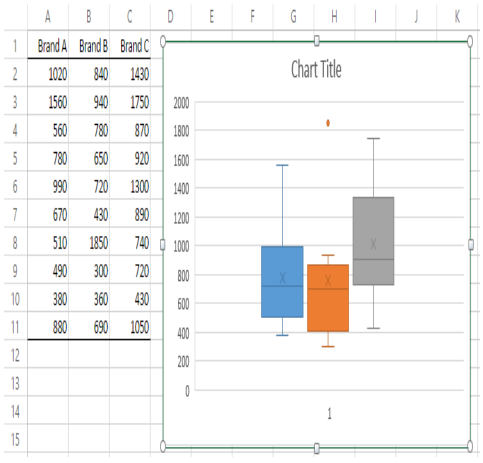


Fig. 2. Shows the rectangle fit approach with outlier. **Fig. 3.** Shows the rectangle fit Approach.

The above fig. 2. Shows Q1 (Lower Quartile), Q2 (Median), Q3 (Upper Quartile) and IQR (Inter Quartile range) using the rectangle fit approach of the dataset with outliers.

In the above fig. 3. Example data of a company is given and chart indicates data of Brand A, Brand B, Brand C shows the rectangle fit approach example of the dataset with outlier.

6. Conclusion

The conclusion lies in the fact that outliers are usually the unwanted entries which always affects the data in one or the other form and misreports the distribution of the data. Sometimes it becomes necessary to keep even the outlier entries because they play an important role in the data but in our case achieving and our main objective is to discovering outlier entries and i.e. to delete the outlier entries from database. Proposed approach provides proper consolidated report using data relative attributes of the database. It is obvious that values in the relative attribute or dependent attribute have certain correlations in the database. Furthermore the more work can be undertaken to identify the correlation between the attributes, which in turn shall help in discovery of anomalous values. One can also laid the emphasis on working upon the said research as a basis and evolve more types of patterns and distribution of values in the attribute for discovering anomalous values and its implications.

References

[1] Zhang, S., Zhang, C., and Young, "Data preparation for data mining", Applied Artificial Intelligence, Vol.- 17, Pp. 375-381, 2003.

[2] HP Kriegel, P Kröger, A Zimek, (2010), Outlier Detection Techniques (KDD).

[3] Yang Zhang, Nirvana Meratnia, Paul Havinga (2010), Outlier Detection Techniques for Wireless Sensor Networks: A Survey, Vol.-12, No-ssue- 2, Second Quarter 2010.

[4] Charu C. Aggarwal, Philip S. Yu (2001), Outlier detection for high dimensional data, Vol.-30, No. - 2, June 2001.

[5] Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad, Mustafa Mat Deris (2004), A Survey of Outlier Detection Methodologies, Vol.-22, No.-2, Pp. 85– 126.

[6] Zuriana Abu Bakar, Rosmayati Mohemad, Akbar Ahmad, Mustafa Mat Deris (2006) A Comparative Study for Outlier Detection Techniques in Data Mining.

[7] P.García-TeodoroaJ.Díaz-VerdejoaG.Maciá-FernándezaE.Vázquezb (2009), Anomaly-based network intrusion detection Techniques, systems and challenges.

[8] Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, Takafumi Kanamori (2009), Statistical outlier detection using direct density ratio estimation.

[9] C. Aggarwal and P. Yu, "Outlier Detection for High Dimensional Data". International Conference on Management of Data, Vol.-30, No.- 2, Pp-37– 46, May 2001.

[10] Z. He, X. Xu and S. Deng, "Discovering Cluster based Local Outliers". Pattern Recognition Letters, Vol. - 24, No. - 9-10, Pp. 1641 – 1650, June 2003.