A Novel Approach for Univariate Outlier Detection

S.M.A.Khaleelur Rahman[1], M.Mohamed Sathik[2], and K.Senthamarai Kannan[3]

[1] Sadakathullah Appa College, Tirunelveli, Tamilnadu, India , email: smakrahman@gmail.com
[2] Sadakathullah Appa College, Tirunelveli, Tamilnadu, India , : mmdsadiq@gmail.com
[3] Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India , senkannan2002@gmail.com

Abstract -

In many applications outlier detection is an important task . In the process of Knowledge Discovery in Databases, isolation of outlying data is important. This isolation process improves the quality of data and reduces the impact of outlying data on the existing values. Numerous methods are available in the detection process of outliers in univariate data sets. Most of these methods handle one outlier at a time. In this paper, Grubb's statistics, sigma rule and fence rules deal more than one outliers at a time. In general, when multiple outliers are present, presence of such outliers prevents us from detecting other outliers. Hence, as soon as outliers are found, removing outlier is an important task. Multiple outliers are evaluated on different data sets and proved that results are effective. Separate procedures are used for detecting outliers in continuous and discrete data. Experimental results show that our method works well for different data.

Index Terms— Grubbs Test, Masking , Sigma rule, Univariate Outlier Detection

1 INTRODUCTION

.

OUTLIERS are the set of objects that are considerably dissimilar from the remainder of the data [1]. Outlier detection is an extremely important problem with a direct application in a wide variety of application domains, including fraud detection, identifying computer network intrusions and bottlenecks [5], criminal activities in e-commerce and detecting suspicious activities [5]. Many researches have been done for identifying outliers and many definitions were emerged. It is necessary to say that outliers are observations which are well deviated from the main data and not follow our assumed model. Outlier identification is challenging in multidimensional data. Many researchers argued that univariate outlier detection methods are useless but we favour this method because outlying data can be hidden in one or two dimensional view of the data. The determination of relevant variables is an important step in reducing the complexity of the model. During the deployment of the model, sometimes, the variable selection may improve the model accuracy. Different approaches have been proposed to detect outliers, and Grubbs statistics is used very often to evaluate measurements, coming from a normal distribution of data with size n, which are suspicious and far away from the main body of the data.

In this paper we detect the presence of outliers using Grubbs statistics' sigma rule, inner fence rule and outer fence rule. Removing outliers may improve the performance of data only when erroneous data mixed with the actual data. But in many situations, presence of outlier must be recorded to identify unusual behaviour.

2 RELATED WORK

Mathematical calculations are used to find out whether the outlier came from the same or different population . But we can not say that the result is correct. Many statistical methods have devised for detecting outliers by measuring how far the outlier is away from the other values. This can be the difference between the outlier and the mean of all points or the difference between the outlier and the mean of the remaining values, or the difference between the outlier and the next closest value. Next, standardize this value by dividing some measure of scatter, such as the SD of all values, the SD of the remaining values, or the range of the data. Finally, compute a P value answering this question: If all the values were really sampled from a Gaussian population, what is the chance of randomly obtaining an outlier so far from the other values? If the P value is small, you conclude that the deviation of the outlier from the other values is statistically significant.

Different computer-based approaches have been proposed for detecting outlying data [13] and [12] but we cannot claim that this is the generic or acceptable method

universally. Therefore , these approaches were classified into four major categories based on the techniques used [10], which are: distribution-based, distance-based, density-based and deviation-based approaches. Distribution-based approaches [10] develop statistical models from the given data and then identify outliers with respect to the model using discordancy test. Data sets may follow normal or Poison distribution. Objects that have low probability to belong to the statistical model are declared as outliers. However, Distribution-based approaches cannot be applied on multidimensional data because they are univariate in nature. In addition, a prior knowledge of the data distribution is required, making the distribution-based approaches difficult to be used in practical applications. In the distance-based approach [12],[13] and [10], outliers are detected by measuring distance. Main limitations of statistical methods are countered by this approach. Rather than working on statistical tests , objects that do not have enough neighbours are defined based on the distance from the given object. Density-based approaches [17] compute the density of regions in the data and declare the objects in low dense regions as outliers. In [35], the authors assign an outlier score to any given data point, known as Local Outlier Factor (LOF), depending on its distance from its local neighborhood. A similar work is reported in [17]. Deviation-based approaches[3] and [7], do not use statistical tests or distance-based measures to identify outliers. The objects that deviate from the description are treated as outliers.

Outlier detection methods can be divided as univariate methods, and multivariate methods. Currently many researches are performed on multivariate methods. Another fundamental taxonomy of outlier detection methods is between parametric (statistical) methods and nonparametric methods that are model-free [41]. Statistical parametric methods working on the assumption that underlying observations are known [30] or, at least, they are based on statistical estimates of unknown distribution parameters [6]. These methods flag as outliers those observations that deviate from the model assumptions. They are often unsuitable for high-dimensional data sets and for arbitrary data sets without prior knowledge of the underlying data distribution [25].

Outliers are subject to masking and swamping. The definitions from[4] and [3]explain these masking effects. One outlier masks a second outlier. In the presence of the first outlier the second is not considered as an outlier but can be considered as an outlier only by itself. Thus, after the deletion of the first outlier the second is appeared as an outlier. In masking the resulting distance of the outlying point from the mean is small.

2.1 Masking Effect

Second observation can be considered as an outlier only by itself not in the presence of first outlying value. If first skewed value is deleted then only the second observation becomes a outlying observation. Masking is an effect which occurs when a group of outlying observations skews the mean and the covariance estimates toward it and makes the outlying value from the mean is small.

2.2 Swamping effect

Second observation can be considered as an outlier only under the presence of the first value. After the deletion of the first outlier the second observation becomes a non-outlying observation. Swamping occurs when a group of outlying instances skews the mean and the covariance estimates toward it and away from other non-outlying instances, and the resulting distance from these instances to the mean is large, making them look like outliers[5].

3 Proposed Work

In this paper, we use a new approach for outliers detection to detect the values that are in an abnormal distance from other values. It uses Grubbs test and three various criteria such as three sigma rule , inner fence criterion and outer fence criterion. Abnormal examples are displayed. Grubbs Test for Detecting Outliers is working with the assumption that the data follows a normal distribution[16]. Each time a potential outlier is identified, that value is temporarily removed and the test is run again and again until no other potential outliers are identified. The default value of alpha is .05, but there is an option to change this value.

Outliers are removed by using filtering rules inner fence and outer fence[16]. Values are clustered around some central values. We use the procedure to tell how other values are far away from central values. These far away values are labeled as outliers . We call them outliers because they lie well outside our expected range. In Grubbs' test, also known as the ESD method , the first step is to quantify how far the outlier is from the others? The ratio Z is calculated which is the difference between the outlier and the mean divided by the SD. If Z is large, the value is far from the others. Note that we calculate the mean and SD from all values, including the outlier. Normally 5% of the values in a Gaussian population are more than 1.96 standard deviations from the mean, our main assumption to conclude that the outlier comes from a different population if Z is greater than 1.96[37]. The Z value

for the query is compared with 5% significance level. No user parameters are used and all parameters are derived directly from data. Population mean and SD from other data are calculated first. Some historical data may also be used for calculating mean and SD.

In fraudulent transactional activities, detecting outlying data is not to remove specific record(s) but to isolate them separately. Hence removing outlying data may defeat our purpose. But in the process of noise removal the outlying data must be removed to enhance data quality.

If calculated value of Z is greater than the critical value , then the P value will be less than 0.05. This is because , if all the data were really sampled from a single Gaussian distribution less than a 5% chance that outlier are encountered from the remaining data in either direction. In this Grubbs test 95% confidence level is applied . First suspicious observations are removed from the data set. After the removal of deviated values , the next observations are confirmed by Grubbs test. Here we proved that run the Grubbs test after removing deviated data may not find new outliers but the calculation were performed and found that continuous statistics such as Average , SD and Z scores are changed. We admit that this method only works for testing the most extreme value in the sample. Here we remove the outlier and run Grubbs' test again to see if there is a second outlier in the data. If we do this , we cannot use the same table.

3σ test is used as a special case of Grubbs test with a single value of 3 ie constant critical value of 3 independent sample size. Masking effect affects the result of identifying outliers in Grubbs test so that here we run Grubbs test for removing outliers. Two procedures are normally used for outlier detection, Single-step and Sequential Procedures[3]. Single-step procedures identify all outliers at once as opposed to successive elimination or addition of datum. We use sequential procedures. At each step, one observation is tested for being an outlier. Largest measurement is tested first, if it is declared as an outlier, it is deleted from the dataset and the procedure is repeated. If it is declared as a non-outlying observation, the procedure is stopped. As found in [2], in outward testing procedures, the sample of observations is first reduced to a smaller sample. The statistics are calculated on the basis of the reduced sample and then the removed observations in the reservoir are tested in reverse order to indicate whether they are outliers. The outward testing procedure is stopped when no more left in the tray for observation [29].

Grubbs' test is also called the ESD method (extreme studentized deviate). Various steps are

- Calculate the ratio Z as the difference between the outlier and the mean divided by the SD

$$Z = \frac{|\text{mean} - \text{value}|}{SD}$$

- If Z is large, the value is far from the others. Calculate the mean and SD from all values, including the outlier
- Since 5% of the values in a Gaussian population are more than 1.96 standard deviations from the mean, then we conclude that the outlier comes from a different population if Z is greater than 1
- Calculate the SD from the data. The presence of an outlier increases the calculated SD
- Z does not get very large, Z can not get larger than $(N-1)/\sqrt{N}$, where N is the number of values
- The critical value increases with sample size
- If calculated value of Z is greater than the critical value in the table, then the P value is less than 0.05 on the assumption that there is less than a 5% chance that encountered an outlier
- Once identified an outlier, exclude that value from your analyses
- Remove the outlier, run Grubbs' test again to see there is a second outlier in the data

4 Results and Discussion

Now, we investigate our proposed method by using an artificial data set with two dimensions and values are entered to demonstrate our outlier detection approach. This data set has 3 attributes with 50 examples. All 3 attributes are continuous attributes.

Tables I, II and III display the detailed result for each variable. It displays individual variables, Grubbs statistic's cut value for different attributes. Filtering parameters are sigma rule, inner fence and outer fence. For each criterion, limit values Lower Bound and Upper Bound values are displayed and outliers are noticed. The Grubbs p-value is .05 and multiple of sigma is 3.

Following observations are noticed

The attribute Cholestral only contains outliers for the Grubbs test at the significance level of 5% for all three criteria

As for Cholestral there are four outlying values for the 3-sigmas criterion , five outlying values for the Inner fence criterion and only one value for the Outer Fence criterion.

By using the Inner Fence criterion, only one outlier for the attribute Thalac, four outlier for the attribute Oldpeak

and five outliers for the attribute Cholestral. As per the Outer Fence criterion, there is only one outlier in cholesterol.

Table IV shows abnormal values appeared in different records and their respective attributes are displayed. Domain values of each field was tested so that number of outliers detected are displayed. 8th,48th, 168th 175th and 252nd values are abnormal for the variable Cholestral, 81st, 147th, 163rd and 216th values are abnormal for the variable Oldpeak and only one 258th value is abnormal for the only variable Thalac.

In Table V and VI, we calculate statistical values such as mean with and without the outlying values. We compare the present result with the previous values in Table VII.

We visualize the results in a scatter plot. Figures 1, 2 and 3 display deviated values for the variable Cholesteral , Thalak and Oldpeak. The examples 8,48,168,175 and 252 seem abnormal for the variable Cholesteral. Examples 81,147,163 and 216 for Oldpeak and the example 256 appears abnormal for Thalac.

TABLE I

UNIVARIATE OUTLIERS DETECTION

DETAILED RESULTS FOR EACH VARIABLE
SIGMA RULE

Variable	Grubbs Stat.	Sigma rule		
	Cut:3.693	L.B	U.B	Detected
cholesteral	6.0817	94.6005	404.7180	4
thalac	3.3963	80.1806	219.1749	1
oldpeak	4.4970	-23.8563	44.8563	2

TABLE II

INNER FENCE RULE

Variable	Grubbs Stat.	Inner Fence		
	Cut:3.6936	L.B	U.B	Detected
cholesteral	6.0817	111.000	383.000	5
thalac	3.3963	83.500	215.500	1
oldpeak	4.4970	-24.000	40.000	4

TABLE III
OUTER FENCE RULE

Variable	Grubbs Stat.	Outer Fence		
	Cut : 3.6936	L.B	U.B	Detected
cholesteral	6.0817	9.00	485.00	1
thalac	3.3963	34.00	265.00	0
oldpeak	4.4970	-48.00	64.00	0

TABLE IV
DETECTED OUTLIERS

No.of example	No.of detection	Variable(s)
8	1	cholesteral
48	1	cholesteral
81	1	oldpeak
147	1	oldpeak
163	1	oldpeak
168	1	cholesteral
175	1	cholesteral
216	1	oldpeak
252	1	cholesteral
258	1	thalac

TABLE V

Attribute	Min	Max	Std-dev/avg
cholesteral	126	564	0.2070
thalac	71	202	0.1548
oldpeak	0	62	1.0907

TABLE VI

Attribute	Min	Max	Std-dev/avg
cholesteral	126	360	0.1796
thalac	88	202	0.1524
oldpeak	0	40	1.0585

TABLE VII

Variable	Mean for 250 examples	Mean for 240 examples(exclude records 8,48,81,147 163,168,175,216,252 and 258)	Deviation (%)
Cholestral	249.6593	246.2731	+3.3862 %
Thalak	149.6778	150.1577	-0.4899 %
Oldpeak	10.5000	9.7231	+0.7769 %

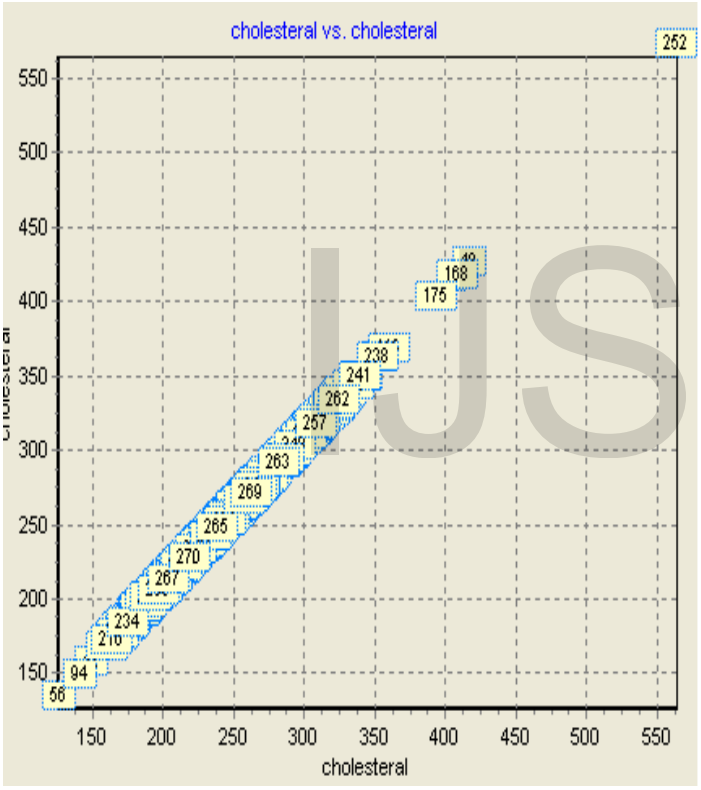

FIGURE 1 DISPLAYING OUTLYING VALUES

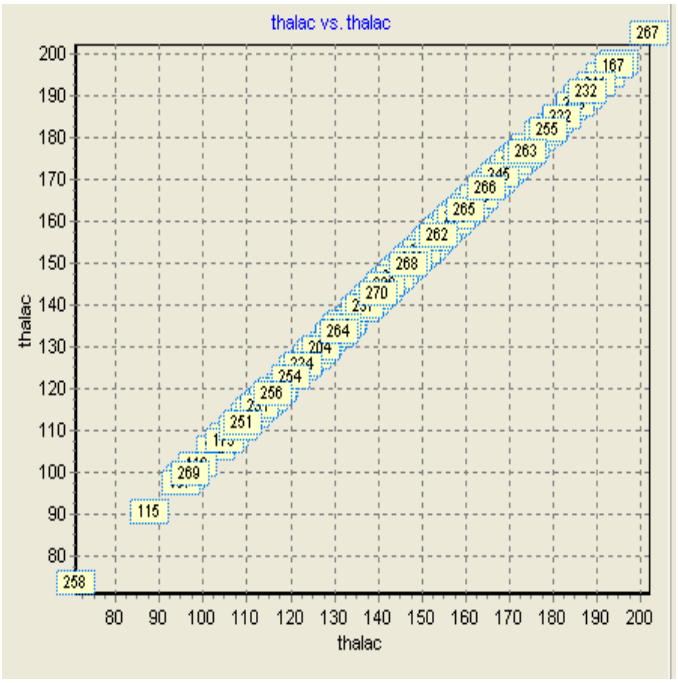

FIGURE 2

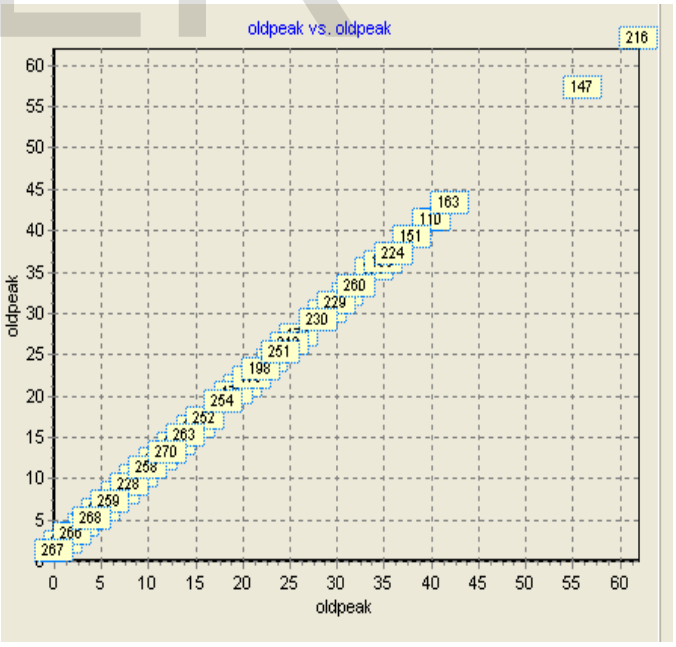

FIGURE 3

5 CONCLUSIONS

The method used does not require an apriori knowledge of the process model. It detects and replaces outliers on-line while preserving all other information in the data. We demonstrated that the proposed filter-cleaner is efficient in outlier detection and data cleaning for even non-stationary process data. Abnormal results are displayed for combination of variables. We delete outliers in the second run to demonstrate that there is a significant change in continuous statistic values after outlying values were removed. This is not the recommended strategy because abnormal examples contribute highly to explore data.

REFERENCES

[1] Hawkins, D. Identification of Outliers. Chapman and Hall, 1980.
[2] Hawkins, S., He H. X., Williams G. J., and Baxter R. A. (2002) "Outlier detection using replicator neural networks," In Proceedings of the Fifth International Conference and Data Warehousing and Knowledge Discovery (DaWaK02), Aix en Provence, France. ampel, F. R. (1971). A general qualitative definition of robustness. Annals of Mathematics Statistics, 42, 1887–1896.
[3] Davies L. and Gather U. (1993), "The identification of multiple outliers", Journal of the American Statistical Association, 88(423), 782-792..
[4] Iglewics , B., and Martinez, J. (1982), "Outlier Detection using robust measures of scale," Journal of Sattistical Computation and Simulation, 15, 285-293.
[5] Jiawei Han and Micheline Kamer, (2006) "Data Mining Concepts and Techniques" , Morgan Kaufmann Pub.
[6] Hadi, A. S. (1992) "Identifying multiple outliers in multivariate data," Journal of the Royal Statistical Society. Series B, 54, 761-771.
[7] Acuna E. and Rodriguez C., (2004), A Meta analysis study of outlier detection methods in classification, Technical paper, Department of Mathematics, University of Puerto Rico at Mayaguez, available at academic.uprm.edu/~eacuna/paperout.pdf. In proceedings IPSI 2004, VeniceJin.
[8] Knorr, E., and Ng. R. (1998). Algorithms for mining distance-based outliers in large datasets. In Proc. 24th Int. Conf. Very Large Data Bases (VLDB), 392–403, 24–27.
[9] Knorr, E., Ng R., and Tucakov V. (2000). Distance-based outliers: Algorithms and applications. VLDB Journal: Very Large Data Bases, 8(3–4):237–253.
[10] Papadimitriou, S., Kitawaga, H., Gibbons, P.G., and Faloutsos, C., (2002) "LOCI: Fast Outlier Detection Using the Local Correlation Integral," Intel research Laboratory Technical report no. IRP-TR-02-09.
[11] Ramaswamy, S., Rastogi, R., and Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Dalas, TX.
[12] Rosner, B. (1975), "On the detection of many outliers," Technometrics, 17, 221-227.
[13] Rousseeuw, P., (1985), Multivariate estimation with high breakdown point. In: W.Grossmann et al., editors, Mathematical Statistics and Applications, Vol. B, 283-297,
[14] Akademiai Kiado: Budapest.Rousseeuw P., Leory A. (1987), Robust Regression and Outlier Detection. Wiley Series in Probability and Statistics.
[15] Breunig, M.M., Kriegel, H.P., Ng, R.T. and Sander, J. (2000) "Lof: Identifying density-based local outliers," In Proc. ACMSIGMOD Conf. 2000, 93–104
[16] Grubbs F. E., 1950, Sample Criteria for Testing Outlying Observations, Annals of Math. Statistics, vol. 21,
pp. 27-58.
[17] Grubbs F. E., 1969, Procedures for Detecting Outlying Observations in Samples, Technometrics, vol. 11, No. 1,
pp. 13-14.