PressAcademia Procedia - (PAP), ISSN: 2459-0762

PressAcademia Procedia

2nd World Conference on Technology, Innovation and Entrepreneurship
May 12- 14, 2017, Istanbul, Turkey. Edited by Sefer Şener

OUTLIER DETECTION METHOD BY USING DEEP NEURAL NETWORKS

DOI: 10.17261/Pressacademia.2017.577
PAP-WCTIE-V.5-2017(15)-p.96-101

Olgun Aydin[1], **Semra Erpolat Tasabat**[2]
[1]Mimar Sinan Fine Art University. olgun.aydin@ogr.msgsu.edu.tr
[2]Mimar Sinan Fine Art University. semra.erpolat@msgsu.edu.tr

ABSTRACT

Detecting outliers in the data set is quite important for building effective predictive models. Consistent prediction can not be made through models created with data sets containing outliers, or robust models can not be created. In such cases, it may be possible to exclude observations that are determined to be outlier from the data set, or to assign less weight to these points of observation than to other points of observation. Lower and upper boundaries can be created to exclude outliers from the dataset, and models can be created using the data between those boundaries. In this study, it was aimed to propose a different perspective on outlier detection methods by creating upper bounds with the aid of deep neural networks using skewness, kurtosis and standard deviation values obtained from the dataset with trained models.

Keywords: Deep neural network, outliers, outlier detection, modelling, predictive modelling

1. INTRODUCTION

It is inevitable that there are deficiencies or some problems in the data obtained in real life. When analyzing data sets containing a large number of variables, researchers are faced with unusual observations that may have potentially harmful effects on the results of the problem and are called outliers. Since outliers increase the value of the error variance, they also have an impact on the power of statistical tests. While it is important to clear the data from the outliers, it is very important to determine the outliers and analyze them specially for some study areas. It is of utmost importance to examine outliers separately in situations such as weather forecasting, fraud detection, detection of unexpectedly responding patients to different types of drugs (http://www.cse.yorku.ca/~jarek/courses/6412/lectures/Outliers.ppt) Outliers can appear in almost any data set in any application domain. Due to measurement errors or momentary extreme conditions, the data set may contain outliers. For this reason, outlier detection is a research field that researchers place importance on.(Hawkins, 1980). In the scope of the study, a different method is proposed for the density based method which is a statistical based method used for outlier detection. With this motivation, data sets without normal outliers were created and outliers were added to these data sets in a controlled manner. It is aimed to determine outlier by using deep neural networks using skewness, kurtosis, standard deviation with generated data sets.

2. LITERATURE REVIEW

In this section, outlier detection methods used in the literature and summary information are given. In addition, summary information on deep neural networks is included.

2.1. Outlier Detection Methods

The method of outlier detection based on statistical distribution is based on the standard deviation method and the calculation of Interquartile Range (IQR) calculated before the boxplot is drawn. In the standard deviation method, values that fall outside the range of [Mean -2*Standard deviation, Mean +2*Standard deviation], or [Mean -3*Standard deviation,Mean+3*Standard deviation] are considered as outliers, if the data has normal distribution. In another method IQR, first quartile (Q1) followed by third quartile (Q3) is calculated. IQR is calculated from Q3-Q1. The data outside the

range [Q1 - 1.5 * IQR, Q1 + 1.5 * IQR] is marked as outlier.

One of the methods based on statistical tests is the Dixon method. In this method, the data set is sorted and the different test statistic is calculated according to the state of the smallest and largest value. The calculated test statistics is compared with the corresponding critical value to determine outlier values. Another method performed by statistical testing is Rosner. In this method, values are assigned in the range 0-9 according to the distance from the average starting from the farthest distance. 9 data is the most distant from the average. 0 is the closest to the average. The data considered to be outliers are removed from the individual data set and the average and standard deviation of each term is calculated. The test statistic is calculated with these calculated mean and standard deviations. These test statistics are compared with the critical values and it is decided whether the observation is outlier or not (Aggarwal 2013).

Clustering-based methods view small-sized clusters, including an observational dimension, as clustering outliers. Some examples for such methods include segmentation around medoids (PAM) and cluster large applications (CLARA) A modified version of these for spatial extensions, called CLARANS And a fractal size based method. Since the main objectives are clustering, these methods are not always optimized for outliers. In most cases, uncertain detection measures are implicit and can not be easily understood from clustering procedures (Ben-gal, 2005). Spatial methods are closely related to clustering methods. Lu et al. (Lu et al., 2003) define non-spatial values as spatially cited objects that differ significantly from local values (Ben-gal, 2005).

2.2. Deep Neural Networks

Deep Learning is a new field in Machine Learning research. Deep Learning; In general, studies that help to understand the meaning of images, sounds and texts. (http://deeplearning.net/tutorial/). Prior to 2006, attempts to train deep architects failed: a deeply supervised feedforward neural network training gives worse results (compared to one or two layers of hidden layer) (both training and test failure). In 2006, three studies led by Hinton's revolutionary work destroyed many memorabilia

- Hinton, G. E., Osindero, S. and Teh, Y., A fast learning algorithm for deep belief nets Neural Computation 18:1527-1554, 2006
- Yoshua Bengio, Pascal Lamblin, Dan Popovici and Hugo Larochelle, Greedy Layer-Wise Training of Deep Networks, in J. Platt et al. (Eds), Advances in Neural Information Processing Systems 19 (NIPS 2006), pp. 153-160, MIT Press, 2007
- Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra and Yann LeCun Efficient Learning of Sparse Representations with an Energy-Based Model, in J. Platt et al. (Eds), Advances in Neural Information Processing Systems (NIPS 2006), MIT Press, 2007

The following three principles have been included in three studies.

- The learning of impressions as unchecked is used to train each layer.
- One layer at a time is being trained on the previously educated, unchecked.
- Use audited training to fine tune all layers. (http://www.iro.umontreal.ca/~pift6266/H10/notes/deepintro.html)

3. DATA AND METHODOLOGY

Controlled deviations were added to the data set generated from the normal distribution. In this process, we first derive 500 random numbers from the normal distribution with a mean of 10 and a standard deviation of 2. Random numbers were then derived from the normal distributions with averages of 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65 and standard deviation 2. Then all these values are multiplied by 10 to obtain outliers. These operations were repeated 10,000 times. The mean, standard deviation, skewness, kurtosis values of data sets with outliers and without outliers were calculated and recorded at each step. All calculations were performed in R Studio environment by following script.

```
lbf <- NULL

ubf<- NULL

mxf <- NULL

dxf <- NULL

mxoutf <- NULL

dxoutf <- NULL
```

```
ubcf <- NULL
skwf <- NULL
krtsf <- NULL
for (i  in 1:10000)
{
 for(j in 1:10)
 {
        x <- rnorm(500,mean = 10,sd = 2)
        out_up<- abs(rnorm(50, mean = 10+5*j, sd = 2))*10
        x_out <- c(x,out_up)
        mxout <- mean(x_out)
        sdxout <- sd(x_out)
        mx<- mean(x)
        sdx<- sd(x)
        lb <- mx - 3*sdx
        ub <- mx + 3*sdx
        skw <- skewness(x_out)
        krts <- kurtosis(x_out)
        ubc <- (floor(min(out_up)) - mxout) / sdxout
        ubcf <- rbind(ubcf,ubc)
        lbf <- rbind(lbf, lb)
        ubf <- rbind(ubf, ub)
        mxf <- rbind(mxf, mx)
        dxf <- rbind(dxf, sdx)
        mxoutf <- rbind(mxoutf, mxout)
        dxoutf <- rbind(dxoutf, sdxout)
        krtsf <- rbind(krtsf, krts)
        skwf <- rbind(skwf, skw)
 }
}
```

Figure 1 shows the histogram of the 500 units data without outliers obtained from the normal distribution with a mean of 10 standard deviations of 2 in the above step. As you can see, the dataset is symmetrical.

Figure 1: Histogram of Variables without Outliers Generated from Normal Distribution

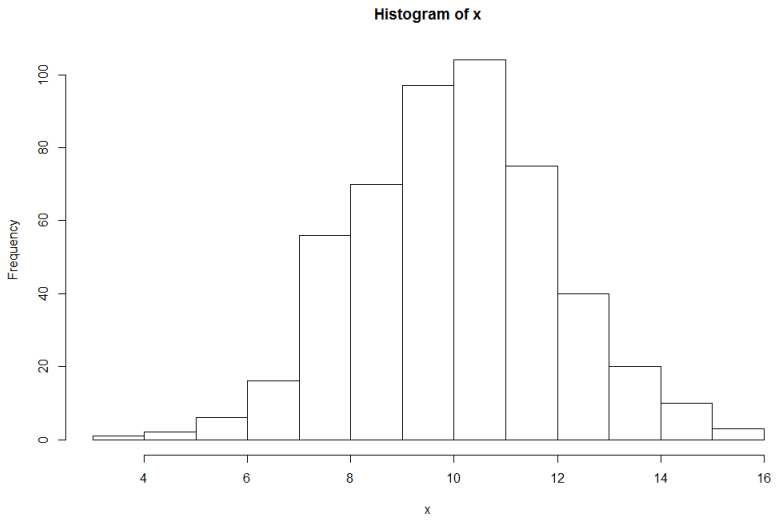
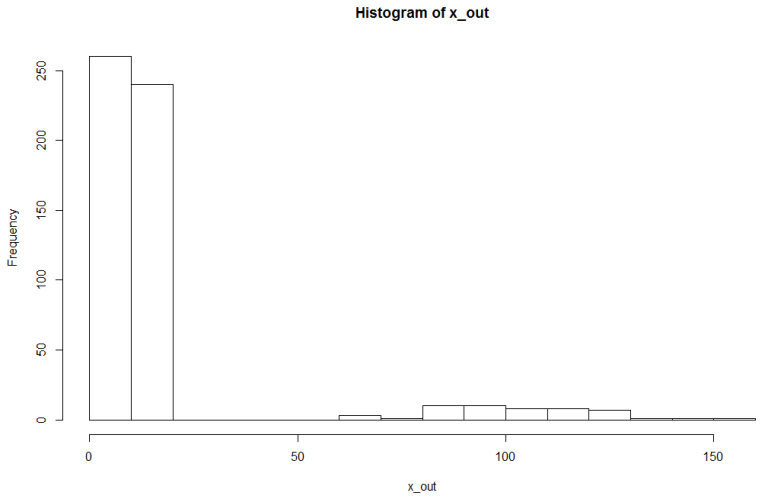


Figure 2 shows the histogram of 500 units of data added outlier to the data set obtained from the normal distribution with a mean of 10 standard deviations of 2 obtained in the above step. As you can see, the dataset is not symmetrical and is right skewed.

Figure 2: Histogram of Variables with Outliers Generated from Normal Distribution



As it is known that 99% of the data set with standard normal distribution is in the [Mean -3*Standard deviation,Mean+3*Standard deviation] With the help of this information, the outliers can be defined as those outside the range of [Mean -3*Standard deviation,Mean+3*Standard deviation]. +3. For this generalization, the upper limit coefficient can be defined as -3, and the lower limit coefficient can be defined as -3. If the data set is normal distribution, the data other than Mheat +/- 3 * standard deviation can be evaluated as outlier. However, the upper and lower limit coefficients are not defined as +3 or -3 when the data does not fit the normal distribution and the data is skewed to the right or to the left. These coefficients also vary according to the skewness, kurtosis and standard deviation of the data. If these coefficients can be estimated, values outside the range [Mheat -lower coefficient * standard deviation, Mheat + upper coefficient * standard deviation] can be defined as outlier. With this in mind, outliers were added to the data obtained from the normal distribution in a controlled manner and 100000 data sets were created to the right. Upper coefficients can also be calculated as it is known at which points the outliers are added. For this purpose, (floor (min (out_up)) - mxout) / sdxout formula is used. After all these calculations the final data set was created.

Table 1: Sample from Final Dataset (Independent Variables)

mxoutf	dxoutf	krtsf	skwf
22.52423	39.94923	6.767220	2.919109
26.88227	54.03466	6.659846	2.907539
32.05484	70.34434	6.342558	2.870150
36.33946	83.34157	6.283900	2.863794
40.65891	97.40719	6.206902	2.854884
45.31630	111.66793	6.167622	2.849995
49.81123	126.20862	6.145563	2.847488
54.20110	140.20838	6.147518	2.847924
59.38225	156.64846	6.129798	2.845660
63.70956	170.21744	6.116862	2.844120
23.05472	41.56753	6.844389	2.930907
27.27965	55.08199	6.617082	2.903642

A sample from the 100000x4 size final dataset is shown in Table 1. In the data set, mxoutf is about average of the data with outliers, dxoutf is about standard deviation of the data with outliers, krtsf is the kurtosis of the data with outlier, and skwf is the skewness of the data with outlier. Ubcf is about upper coefficient. In the study, ubcf is considered as a dependent variable, while dxoutf, krtsf, swkf are considered as independent variables. A deep neural network, LSTM, is used to detect long and short term dependencies. Before creating this neural network, 75% of the data is allocated as a train set and 25% as a test set. In train phase, epoch 50, batch size 30, mean squared error as loss function, and optimization algorithm as adam.

4. FINDINGS AND DISCUSSIONS

When the model training process is completed, it is seen that the final MSE is 0.0806. Figure 3 shows the variation of the loss function in each epoch in the train process. It was seen that the MSE was 0.0808 when the modeled train test set was applied.

Figure 3: Training Process

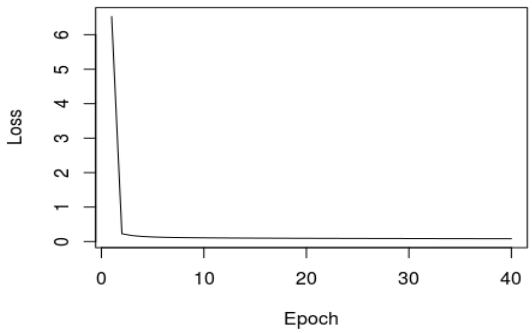


Table 2 also shows some estimation results by using . According to the results, it can be seen that the upper coefficients can be predicted consistently.

Table 2: Some Test Results

mxoutf	dxoutf	krtsf	skwf	ubcf (real)	ubcf (predicted)	Biass
-1,58549	-1,58345	3,20958	3,23297	2,99761	2,83683	0,16078
-1,21875	-1,21455	1,08475	1,05602	4,47566	4,62052	-0,14486
-0,85889	-0,86192	0,12233	0,13471	6,48482	6,65756	-0,17274
-0,51340	-0,50979	-0,21554	-0,20037	8,16181	8,43490	-0,27309
-0,17372	-0,17109	-0,27275	-0,28420	9,61548	9,69550	-0,08002
0,15581	0,16377	-0,55835	-0,55735	11,03645	11,63451	-0,59805
0,51716	0,51341	-0,54952	-0,55273	13,45952	13,39541	0,06412
0,87909	0,88744	-0,67357	-0,67475	15,16611	15,29549	-0,12938
1,21913	1,22310	-0,70991	-0,71117	16,88347	16,89834	-0,01487
1,54321	1,55068	-0,77332	-0,77348	18,91024	18,33228	0,57796

5. CONCLUSION

Before analyzing a dataset, it is very important to extract the dataset from the outliers. Models made with a data set containing outliers will not be able to achieve accurate information as a result of the inferences. In this study it is aimed to give a different perspective to outlier detection. As a result of the studies made, it is possible to detect the outliers and remove the dataset from the outliers through deep neural networks created using some descriptive statistics. In future studies, it is aimed to use other deep neural networks other than LSTM and to compare the results.

REFERENCES

Aggarwal, C.C. (2013), Outlier Analysis, Springer-Verlag New York

Hawkins, D. (1980), Identification of Outliers Chapman and Hall Hawkins, D. Identification of Outliers. Chapman and Hall.

http://www.cse.yorku.ca/~jarek/courses/6412/lectures/Outliers.ppt

http://deeplearning.net/tutorial/

http://www.iro.umontreal.ca/~pift6266/H10/notes/deepintro.html

Ben-Gal, Irad. "Outlier detection." Data mining and knowledge discovery handbook (2005): 131-146.

Osborne, Jason W., and Amy Overbay. "The power of outliers (and why researchers should always check for them)." Practical assessment, research & evaluation 9.6 (2004): 1-12.