Comparison of Outlier Detection between Statistical Method and Undercomplete Autoencoder

Muhammad Ayaz Hussain, University of
Saif ur Rahman, University of
Christian Klaus, University of
Ioannis Iossifidis, University of

This paper is concerned about the analysis of power consumption data and the comparison of the performance between statistical method and undercomplete autoencoder in the determination of outliers in that given dataset. Firstly, the raw data was analyzed and a feature vector was constructed. Secondly, Tukey's test was implemented on this feature vector to determine the outliers. Finally, an undercomplete autoencoder was also used to detect outliers and the efficiency of both methods (statistical method and undercomplete autoencoder) were compared.

1. INTRODUCTION

Anomaly detection is a technique which is used to detect abnormal observations generated by a system. Anomaly detection is usually a domain specific problem. Therefore, it has been researched and implemented in many applications ranging diverse fields e.g. medical data [Chandola et al. 2009; Lauer 2001], anomaly power consumption [Araya et al. 2017], and cyber secrity [Kruegel and Vigna 2003; Hong et al. 2014] are among many. Anomaly detection in power consumption is vital because roughly 60% of the worlds electricity is consumed by residential buildings and by commercial sector (Araya, Grolinger, ELYamany, Capretz, & Bitsuamlak, 2017). The growth in the demand of power as well infrastructure of generation and distribution along with several constraints in order to meet the rising demands is one of the of the biggest problem, especially in developing countries. Anomaly detection in electricity consumption is of great significance because it can help reducing the wastage of electricity consumption generated by various sources including equipment or human related fault as unexpected variations can incur additional operating costs to their facilities. These fluctuations in electricity consumption can arise from various factors such as excessive use of heavy equipment like electric heaters in winters, room coolers during the summers or unusually high number of customers just before any public holiday etc.

One approach towards the detection of abnormality in energy consumption is to monitor energy consumption, manually. Once abnormality is found, it can be reported to concerned authorities in order to take due action to rectify it. However, this approach requires a lot of human effort. Therefore, this solution is less desirable. Another, possible solution is to first learn the machine learning or statistical model, and then later predict anomaly based on the leaned model [Araya et al. 2017; Martinelli et al. 2004] In this study, we have proposed a method that not only detect anomalies but also categorizes the given anomalous inputs into one of three categories, small, medium or big. Our method is divided into two steps: self labeling based on modified version of Tukey's test [Mcgill et al. 1978; Haynes 2013], and then the calculation of Probability density function (PDF) of gaussian distribution (Singh & Rajgopal, 1985). The parameters of gaussian distribution were estimated using maximum likelihood estimation (MLE) [Norden 1972]. We have compared the classification accuracy of our method with a semi-supervised sparse undercomplete autoencoder. The evaluation of models was performed on the data collected from 227 different superstores over the time period of 11 months owned by Tengleman warenhandelsgesellschaft KG, independently.

2. RELATED WORKS

There is a sizable literature regarding semi-supervised learning methods on big data, with some relating to power consumption such as, [Bhattacharjee et al. 2018] which also performed semi-supervised analysis to determine and counter data falsifications in power consumption which can be considered as anomalous power consumption. Since data falsification can range from individual customers tampering with the meter for electricity theft to certain groups orchestrating an attack against a rival company in a systematic manner compromising several meters, as their goals can be more complex than just electricity theft. For that, they proposed, a novel metric based on harmonic to the arithmetic mean ratios of daily power consumption. The ratios between harmonic and arithmetic mean remain highly stable over time as compared to just arithmetic mean. During various attacks, there is asymmetric growth(or decay) rates harmonic mean as compared to symmetric growth(or decay) rates of the arithmetic mean, which can help to infer the presence and type of falsification precisely. Their data consisted of real power consumption of 700 houses from Texas and 5000 houses from Ireland that belong to residential customers. One of the most important aspect of this work is that, it enables the quick identification of around or less than 10 days of monitoring, which is quite efficient as compared to pattern-based energy theft detector (CPBETD) which takes around 1 year, autoregressive moving average (ARMA) which took 1 month, or Entropy based method which also took 1 month.

As far as the use of autoencoder for classification is concerned, those have been used in previous works for classification tasks such as [Fan et al. 2018] which used autoencoder based unsupervised anomaly detection method on a building energy data on a single facility which was in detail as 113 variables were considered which were generally classified into 4 distinct classes which are, (1) time variables (i.e., Month, Day, Hour, Minute and Day type); (2) outdoor variables (e.g., outdoor dry-bulb temperature and relative humidity); (3) operating parameters of the chiller plant (e.g., the temperatures and flow-rates of chilled water and condenser water); (4) energy variables (e.g., the total building cooling load and electricity consumption of the chiller plant).The highest classification accuracy is achieved using the high-level features generated by the 1D convolutional autoencoder trained without conditional information and using a 10% masking noise level. In general, convolutional autoencoders have slightly better performance than the feed-forward fully connected autoencoders. Similar to other machine learning problems, the accuracy of classification depends upon the quality of dataset. Since they used a vast number of variables in their feature vector, therefore, unsupervised learning was possible.

Semi-supervised learning using Softmax classifier with sparse autoencoder for classification of water quality is done by [Yuan and Jia 2016]. Their dataset consisted of 60 unbalanced classification records, 100 training samples are generated for each water quality assesment grade. Their feature vector consisted of 14 representative features of water quality which were chemical oxygen demand (COD), dissolved oxygen (DO), total phosphorus (TP), 5-day biochemical oxygen demand (BOD5), NH3-N, NO3- N, oil, chlorophyll, PH, electrical conductivity (EC), turbidity, CL, total coliform (TColi) and temperature (T). They used 500 labeled training samples and 7,932 unlabeled records in all. Their proposed method method showed 98.8% accuracy which was greater as compared to other supervised algorithms such as Softmax classifiers, BP-NN, RBF-NN and SVM.

3. METHODOLOGY

3.1. Data Preparation and Analysis

Initially, the raw data was provided by the company, which consisted of several entries. Those entries were obtained from the energy consumption data of 227 stores on 15 min basis for around 11 months as shown in the Figure 1. Analysis of Figure 1 shows that there is an obvious increase in power consumption during winters possibly due to the usage of electric heaters and an major drop in the power consumption around Christmas and New Year (just above January).

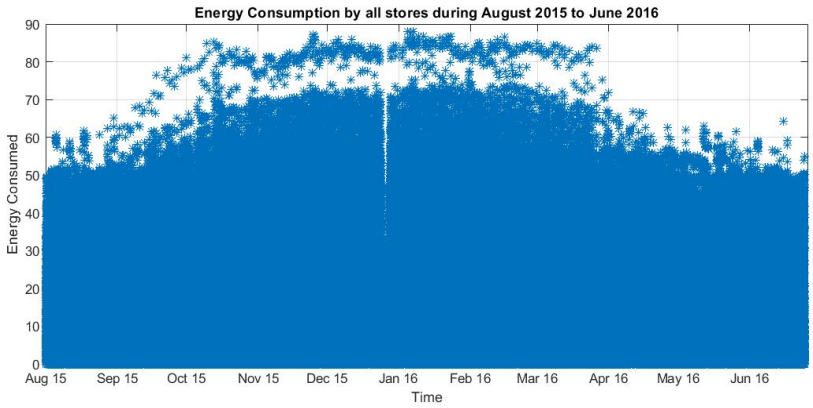


Fig. 1. Energy Consumption by all stores between August 2015 and June 2016

Afterwards, a random month is selected from the whole time period (in this case November 2015) for visualization and energy consumption of all stores is plotted as shown in Figure 2.

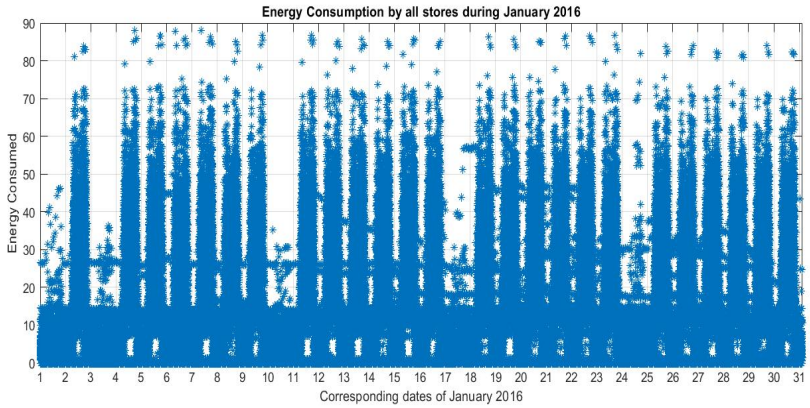


Fig. 2. Energy Consumption by all stores during January 2016

Still, it was difficult to pinpoint or visualize any single store from that plot. Therefore, a random store was chosen (in this case Store 105) for in depth analysis as shown in Figure 3.

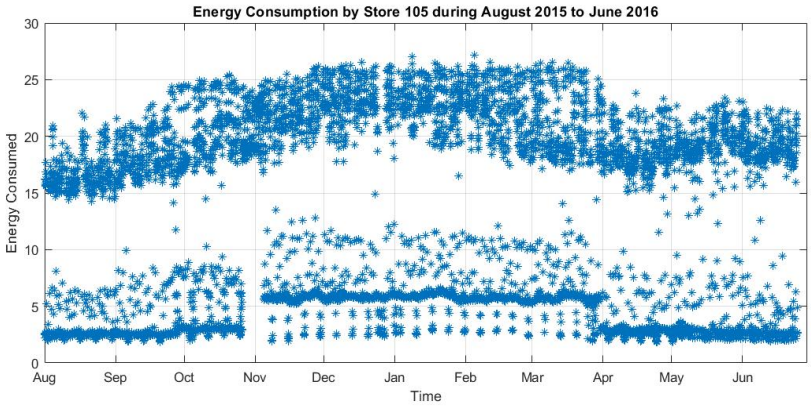


Fig. 3. Energy Consumption by Store 105 between August 2015 and June 2016

We can again see the noticeable increase in energy consumption during winter months from October till April as well as lower consumption during Christmas and New year holidays. Again, we used November which was randomly picked earlier to visualize the energy consumption by Store 105 as shown in Figure **??**.

In order to select meaningful features to be implemented in Machine Learning or Statistical Modeling algorithms, it was necessary to analyze the given data throughly. After a comprehensive analysis, it was decided that the amount of energy consumption depends upon multiple factors, which are the store itself as all stores differs from one another as each having its particular size, different number of energy consuming equipments and even have different geographical location which affect the amount of energy consumption. Aside from that, time of the day is another factor as during working hours energy consumption is higher. There is also a reduction in power consumption during the day even when the store is closed.

The data consisted of the energy readings from 227 different stores with 238 sensor values (as certain stores had multiple sensors) taken every 15 minute which spanned over a period of 11 months. The provided data was then throughly examined and it was subdivided into 2 main categories which were working days and non-working days as energy consumption showed different trends in both categories. Therefore, a Feature Vector was developed which consisted of respective month, hour of the day, store number, energy consumed, a tag for working day or non-working day. As the data was gathered from the sensors in the real world therefore, it was highly unbalanced outliers estimated to be around 2-5% of the total dataset.

3.2. Statistical Method for Anomaly Detection

In order to detect the outliers, a feature vector was developed, which consisted of the hour of the day, store number, energy consumed, a tag for working and non-working day and on that data a modified version of Tukey's test test [Mcgill et al. 1978] was performed, which involves the calculation of median of training data which is referred to $Q2$ and then again median of values lesser and greater than that of $Q2$ is calculated. Median of values lower than $Q2$ is refereed as $Q1$ (lower quartile) and those having greater than $Q2$ is referred as $Q3$ (upper quartile) as shown in Figure 8.

After determining $Q1$ and $Q3$, then Interquartile Range (IQR) is calculated by subtracting the value at $Q3$ by $Q1$.

After determining $Q1$ and $Q3$, then Interquartile Range (IQR) is calculated by subtracting the value at $Q3$ by $Q1$.

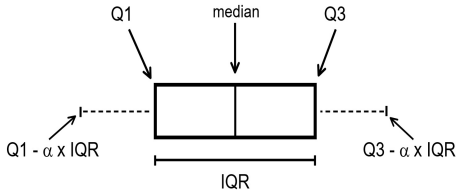


Fig. 4. Depiction of quartiles as used in the detection and removal of Outliers.

$$IQR = Q3 - Q1 \tag{1}$$

After calculating the IQR, we use it to calculate the Tukey's Fences or Inner Fences by the following equation,

$$InnerFenceLowerLimit = Q1 - \alpha(IQR) \tag{2}$$

$$InnerFenceUpperLimit = Q1 + \alpha(IQR) \tag{3}$$

where,
α is the factor which varies the inner fence limits.

Any value lying beyond inner fence limits (upper and lower) can be considered as an outlier, and therefore, it is removed. The results of designed outlier detection algorithm are shown in Figure 6, 7(a) and 8(a).

After the outliers have been removed, the test input is given to the Anomaly detector. The test input consists of an additional learning parameter β, which is used to give an equivalent amount of standard deviation to the value of energy consumption. For example, if $\beta = 2$, then we can consider energy data which is 2σ or 2 standard deviations away from the mean as not an anomaly. The lower the value of β the higher the chance of getting the test input as an anomaly and vice versa.

$$y - \beta\sigma \leq x \leq y + \beta\sigma \tag{4}$$

where,
y is output
x is test input
σ is standard deviation
β is the learning parameter

As the dataset used in this project was highly varied and there were several energy values that stayed the same for the whole period of time. In order to calculate the quartiles and medians, there must be at least 4 unique values in a dataset, therefore for those dataset having less unique values, some product of their respective average was used to determine Inner and Outer Fences.

3.3. Autoencoder

An autoencoder always consists of two parts, the encoder and the decoder, which can be defined as transitions ϕ and ψ such that:

$$\phi : \mathcal{X} \to \mathcal{F} \tag{5}$$

$$\psi : \mathcal{F} \to \mathcal{X} \tag{6}$$

$$\phi, \psi = \underset{\phi,\psi}{\arg\min} \ \|X - (\phi \circ \psi)X\|^2 \tag{7}$$

In a simple case, if there is one hidden layer, the encoder stage of an autoencoder takes the input $\mathbf{x} \in \mathbb{R}^d = \mathcal{X}$ and maps it to $\mathbf{z} \in \mathbb{R}^p = \mathcal{F}$.

$$z = \sigma(Wx + b) \tag{8}$$

This image $\mathbf{z}$ is usually referred to as *code*, *latent variables* or *latent representation*. Here, σ is an element-wise activation function such as sigmoid function, hyperbolic tangent or a rectified linear unit. W is weight matrix and b is a bias vector. After that, the decoder stage of the autoencoder maps $\mathbf{z}$ to the reconstruction $\mathbf{x}'$ of the same shape as $\mathbf{x}$.

$$x' = \sigma'(W'z + b') \tag{9}$$

where, σ', W' and b' for the decoder may differ in general from the corresponding σ, W and b for the encoder, depending on the design of the autoencoder.

Autoencoders are also trained to minimize reconstruction errors (such as squared errors):

$$\mathcal{L}(x, x') = \|x - \sigma'(W'(\sigma(Wx + b)) + b')\|^2 \tag{10}$$

where, x is usually averaged over some input training set.

If the feature space $\mathcal{F}$ has lower dimensionality than the input space $\mathcal{X}$, then the feature vector $\phi(x)$ can be regarded as a compressed representation of the input x. If the hidden layers are larger than the input layer, an autoencoder can potentially learn the identity function and become useless. However, experimental results have shown that autoencoders might still learn useful features in these cases.

In case of Undercomplete autoencoders, $\mathcal{F}$ has lower dimensionality than the input space $\mathcal{X}$, then the feature vector $\phi(x)$ can be regarded as a compressed representation of the input x.

In this paper, an undercomplete autoencoder was trained in a semi-supervised fashion on the values of normal energy consumption data. Afterwards, the trained model was used and evaluated on a pre-trained dataset.

4. EXPERIMENTS AND RESULTS

Outlier detection algorithm was developed and it performed adequately well in removing the outliers. Figure 5 shows the energy consumption data of each hour for all working days in January 2016, where as, figure 6 shows the same data as in figure 5 after the outlier removal algorithm implemented on it.

In Fig 7, we are comparing the accuracy of prediction of both algorithms with manually classified energy data of two stores using confusion matrices.

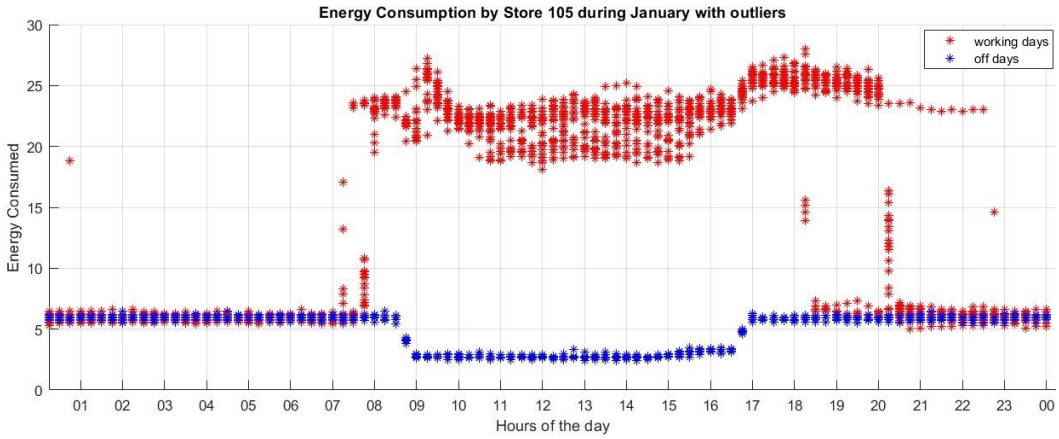


Fig. 5. Energy Consumption by Store 105 during working days of January 2016 with outliers

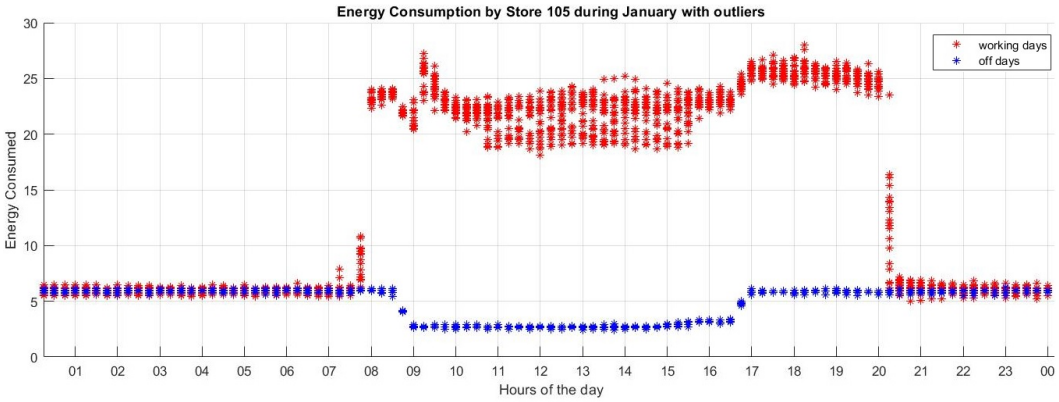
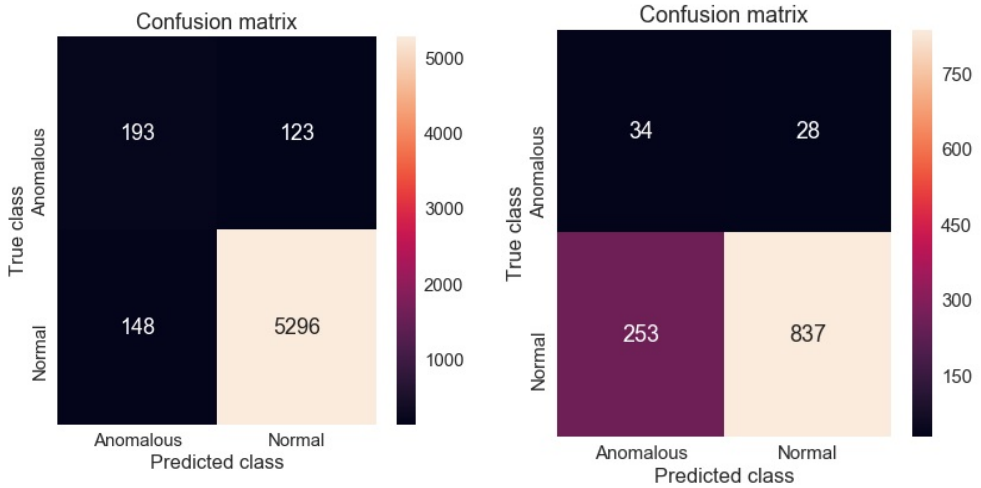


Fig. 6. Energy Consumption by Store 105 during working days of January 2016 without outliers

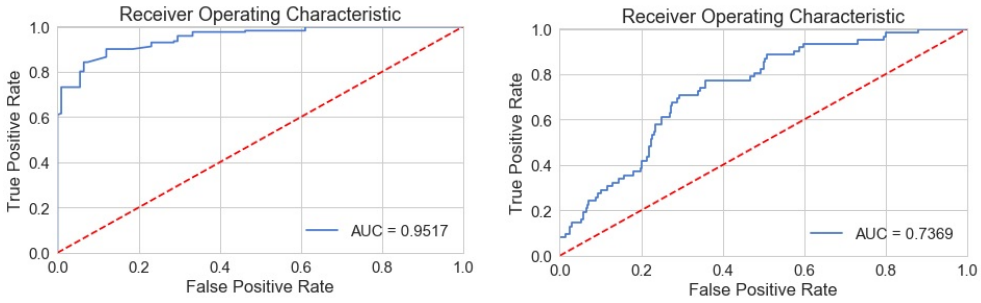


(a) Confusion Matrix of Outlier Detection Algorithm

(b) Confusion Matrix of Autoencoder.

Fig. 7. Confusion Matrices

The same dataset which was used to make confusion matrices in Fig 7 is used to plot Receiver Operating Characteristics curves in Fig 8.



(a) ROC curve of Outlier Detection Algorithm.

(b) ROC curve of Autoencoder.

Fig. 8. ROC curves

The same dataset which was used to make confusion matrices in Fig 7 is used to plot Receiver Operating Characteristics curves in Fig 8.

As we can see from the comparison of Fig 7 and Fig 8, we can interpret that the performance of Statistical method which is based on Tukey's test performed remarkably better with around 95% correlation with manual classification but the shortcoming of this method is that the dataset should be arranged in a proper manner as a single missing values in the dataset can make the classification ineffective. Whereas, classification using Autoencoder performed adequately with around 70-75% correlation with manually classified data but it required pre-classified data containing no outliers.

5. DISCUSSION/OUTLOOK/CONCLUSION/FUTURE WORK

In this paper, the performance of a semi supervised Autoencoder and a statistical model for outlier detection was compared. The experiment resulted in the varying results from each algorithms as Autoencoder did not require data to be arranged precisely and were robust against mixed up dataset whereas, Statistical based outlier detection model performed more accurately in detecting outliers but it required careful arrangement of the data.

In future, it is planned to use more variables in the feature vector of energy consumption data such size of the facility, number of energy consuming equipments, occupancy of the building, and other sensor data such as outside temperature and solar radiation. As more data is embedded in the feature vector, there will be more focus on using totally unsupervised learning using Autoencoder to discover and learn new pattern in the given dataset and do the classification more precisely.

REFERENCES

Daniel B. Araya, Katarina Grolinger, Hany F. ElYamany, Miriam A.M. Capretz, and Girma Bitsuamlak. 2017. An ensemble learning framework for anomaly detection in building energy consumption. *Energy and Buildings* 144 (2017), 191 – 206. DOI:http://dx.doi.org/https://doi.org/10.1016/j.enbuild.2017.02.058

Shameek Bhattacharjee, Aditya Thakur, and Sajal K. Das. 2018. Towards Fast and Semi-supervised Identification of Smart Meters Launching Data Falsification Attacks. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security (ASIACCS '18)*. ACM, New York, NY, USA, 173–185. DOI:http://dx.doi.org/10.1145/3196494.3196551

Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly Detection: A Survey. *ACM Comput. Surv.* 41, 3, Article 15 (July 2009), 58 pages. DOI:http://dx.doi.org/10.1145/1541880.1541882

Cheng Fan, Fu Xiao, Yang Zhao, and Jiayuan Wang. 2018. Analytical investigation of autoencoder-based methods for unsupervised anomaly detection in building energy data. *Applied Energy* 211 (2018), 1123 – 1135. DOI: http://dx.doi.org/https://doi.org/10.1016/j.apenergy.2017.12.005

Winston Haynes. 2013. *Tukey's Test*. Springer New York, New York, NY, 2303–2304. DOI: http://dx.doi.org/10.1007/978-1-4419-9863-7_1212

J. Hong, C. Liu, and M. Govindarasu. 2014. Integrated Anomaly Detection for Cyber Security of the Substations. *IEEE Transactions on Smart Grid* 5, 4 (July 2014), 1643–1653. DOI: http://dx.doi.org/10.1109/TSG.2013.2294473

Christopher Kruegel and Giovanni Vigna. 2003. Anomaly Detection of Web-based Attacks. In *Proceedings of the 10th ACM Conference on Computer and Communications Security (CCS '03)*. ACM, New York, NY, USA, 251–261. DOI: http://dx.doi.org/10.1145/948109.948144

Martin Lauer. 2001. A Mixture Approach to Novelty Detection Using Training Data with Outliers. In *Machine Learning: ECML 2001*, Luc De Raedt and Peter Flach (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 300–311.

Marco Martinelli, Enrico Tronci, Giovanni Dipoppa, and Claudio Balducelli. 2004. Electric Power System Anomaly Detection Using Neural Networks, Vol. 3213. 1242–1248. DOI: http://dx.doi.org/10.1007/978-3-540-30132-5_168

Robert Mcgill, John W. Tukey, and Wayne A. Larsen. 1978. Variations of Box Plots. *The American Statistician* 32, 1 (1978), 12–16. DOI: http://dx.doi.org/10.1080/00031305.1978.10479236

R. H. Norden. 1972. A Survey of Maximum Likelihood Estimation. *International Statistical Review / Revue Internationale de Statistique* 40, 3 (1972), 329–354. http://www.jstor.org/stable/1402471

Y Yuan and K.-B Jia. 2016. A semi-supervised approach for water quality detection based on IoT network. 7 (01 2016), 858–866.