



Supervised and Unsupervised Neural Networks: Experimental Study for Anomaly Detection in Electrical Consumption

Joel García^{1(✉)}, Erik Zamora^{1(✉)}, and Humberto Sossa^{1,2(✉)}

¹ Instituto Politécnico Nacional - Centro de Investigación en Computación,
Av. Juan de Dios Batiz S/N, GAM, 07738 Mexico City, Mexico
joel.gaid.we@gmail.com, {ezamora,hsossa}@cic.ipn.mx

² Tecnológico de Monterrey, Campus Guadalajara, Av. Gral. Ramón Corona 2514,
45138 Zapopan, Jalisco, Mexico

Abstract. Households are responsible for more than 40% of the global electricity consumption [7]. The analysis of this consumption to find unexpected behaviours could have a great impact on saving electricity. This research presents an experimental study of supervised and unsupervised neural networks for anomaly detection in electrical consumption. Multilayer perceptrons and autoencoders are used for each approach, respectively. In order to select the most suitable neural model in each case, there is a comparison of various architectures. The proposed methods are evaluated using real-world data from an individual home electric power usage dataset. The performance is compared with a traditional statistical procedure. Experiments show that the supervised approach has a significant improvement in anomaly detection rate. We evaluate different possible feature sets. The results demonstrate that temporal data and measures of consumption patterns such as mean, standard deviation and percentiles are necessary to achieve higher accuracy.

Keywords: Anomaly detection · Neural networks · Supervised Unsupervised · Statistic

1 Introduction

Energy production has had a considerable impact on the environment. Moreover, different international agencies, such as OECD [11] and UNEP [7], estimated that the global energy consumption will increase by 53% in the future. Saving energy and reducing losses caused by fraud are today's challenges. In China, for instance, numerous companies have adopted strategies and technologies to be able to detect anomalous behaviours of a user's power consumption [14]. With the development of internet of things and artificial intelligence, it is possible to take measures of household consumption and analyze them for anomalous behaviour patterns [3].

Anomaly detection is a problem that has been widely studied with many techniques. To be able to apply the correct one, it is necessary to identify the sort of anomaly. In [5] Chandola, Banerjee and Kumar presented a research about anomaly types found in real applications and the most suitable method of attacking them. Anomalies are value groups which do not match with the whole dataset. They can be classified within three categories: point, contextual and collective anomalies. Point anomalies refer to individual data values that can be considered as anomalous with respect to the remainder. Meanwhile, contextual anomalies take advantage of complementary information related to the background of the problem in order to discover irregularities. Collective anomalies are similar to the contextual type. Collective refers to the fact that a collection of instances can be considered anomalous with respect to the rest of the elements. Our research considers the contextual anomalies in electrical consumption, because they need extra knowledge besides the usage measurements, to be detected. According to Chandola *et al.* [5], for anomaly detection in power consumption the most suitable methods are: Parametric Statistic Modelling, Neural Networks, Bayesian Networks, Rule-based Systems and Nearest Neighbor based techniques. The focus of this paper is on two neural network approaches: supervised and unsupervised learning. Furthermore, we present a feature selection study. The results are compared with a traditional statistical method based on the two-sigma rule of normal distributions.

The remaining sections of the paper are organized as follows: Sect. 2 presents a brief description of related work to the detection of electrical consumption anomalies. Section 3 describes the neural network methods proposed and the statistical procedure. Section 4 outlines the experiments and Sect. 5 gives the results obtained. Finally, Sect. 6 states the conclusion and future work.

2 Related Work

Diverse techniques have been used for identifying anomalies; statistics procedures and machine learning are some of them. In [2] and [12], two research works of 2016 and 2017 respectively, an unsupervised learning technique is presented. Both cases use autoencoders as the pattern learner engine. The difference is the composition of the feature sets. In [12] an 8-feature set is used while in [2] there is an increase in the number of temporal and generated data that results in a 15-feature set. So, in this research, we analyze feature sets by evaluating three possible groups, two based on the ones used in [12] and [2], and a third new feature set with 10 variables.

In [13] Lyu *et al.* propose a hyperellipsoidal clustering algorithm for anomaly detection focused on fog computing architecture. Jui-Sheng Chou, Abdi Suryadinata Telaga implemented a neural network called ARIMA (Auto-Regressive Integrated Moving Average) to identify contextual anomalies [6]. The network learns the normal patterns of energy usage, then it uses the two-sigma rule to detect anomalies. An interesting fact is that, the ARIMA architecture has only one hidden layer. Thus, it is convenient to explore more complex models. In this

study, we test various deep neural architectures and compare their accuracy. In [18] Yijia and Hang developed a waveform feature extraction model for anomaly detection. It uses line loss and power analysis. Ouyang, Sun and Yue proposed a time series feature extraction algorithm and a classification model to detect anomalous consumption [15]. They treat the problem as with point anomalies and exclude other information besides the consumer’s usage. On the contrary, we consider temporal information and generated data such as mean, standard deviation or percentiles since a point value could be abnormal in one context but not in another.

3 Methodology

In this section, we introduce the components and techniques used in the supervised and unsupervised learning method, and the statistical procedure. Autoencoder and ROC curve are used by the unsupervised approach, while multilayer perceptrons correspond to the supervised approach.

3.1 Autoencoder

The unsupervised method for anomaly detection is based on the neural model known as autoencoder or autoassociator [4]. The structure of an autoencoder has two parts; an encoder and a decoder. The goal of this network is to encode an input vector x so as to obtain a representation $c(x)$. Then it should be possible to reconstruct the input data patterns from its representation. The cost function for this model is the Mean Squared Error MSE given in (1):

$$MSE = \frac{1}{D} \sum_{i=1}^D (X_i - \hat{X}_i)^2 \quad (1)$$

where \hat{X}_i , predicted value; X_i , observed value; and D , sample size.

We expect the autoencoder to capture the statistical structure of the input signals. The Reconstruction Error RE, defined in (2), is the difference between the current input and the corresponding output of the network. It is a metric that shows how similar those elements are.

$$RE = \sqrt{\sum_{j=1}^D (X_j - \hat{X}_j)^2} \quad (2)$$

where X , input vector of D different variables; and \hat{X} , vector of D different variables of the constructed output by the autoencoder.

Then, it is necessary to establish a threshold for the Reconstruction Error. A small threshold might result in large False Positives (FP), which are normal patterns identified as anomalies. On the other hand, big thresholds achieve higher False Negative (FN) or anomalous values unidentified.

3.2 Receiver Operating Characteristics

The Receiver Operating Characteristics (ROC) curve has been widely used since 1950 in fields such as electric signal detection theory and medical diagnostic [10]. In a ROC graph, x-axis represents the False Positive Rate FPR (or 1-Specificity) and the y-axis the True Positive Rate TPR (or sensitivity). All possible thresholds are then evaluated by plotting the FPR and TPR values obtained with them. Hence, we can select a threshold that maximizes the TPR and at the same time minimizes the FPR. This policy would correspond to point (0, 1) in the ROC curve, as shown in Fig. 1. In this work, we focus on that policy.

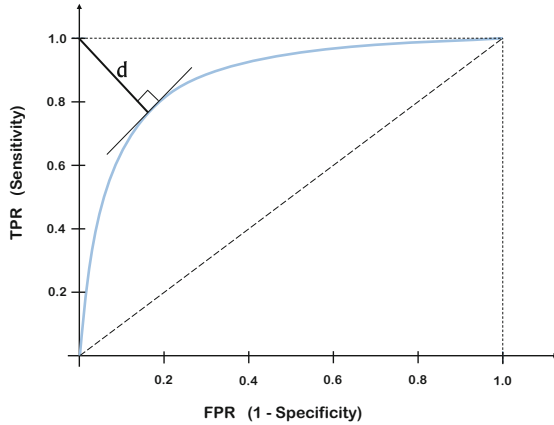


Fig. 1. ROC curve analysis for determination of optimal threshold.

Once the desired point in the plot has been established, the most suitable threshold can be determined by calculating the shortest distance, from that point to the curve, with the Pythagoras theorem given in (3):

$$d^2 = (1 - \text{sensitivity})^2 + (1 - \text{specificity})^2 \quad (3)$$

where d is the shortest distance from a desired point to the ROC curve.

In this research, we evaluate 7 autoencoder models for anomaly detection. The threshold that optimises TPR and FPR is calculated in each case.

3.3 Multilayer Perceptron

A Multilayer Perceptron (MLP) is a class of artificial neural network formed by layers of perceptrons. MLPs usually have an input layer, some hidden layers and an output layer. Every perceptron is a linear classifier. It produces a unique output y by executing a linear combination of an input vector X , multiplied for a weights vector W^T , and adding a bias b , as shown in (4):

$$y = \varphi\left(\sum_{i=1}^n w_i x_i + b\right) = \varphi(W^T X + b) \quad (4)$$

Where φ is an activation function which is typically the hyperbolic tangent or the logistic sigmoid. In this work, we use the hyperbolic tangent function since it can output positive and negative values, which is relevant for the application.

MLPs frequently perform classification and regression tasks. This network learns the dependencies between inputs and outputs in supervised learning through the backpropagation training algorithm [1]. During the training process, weights and bias are modified in order to minimize a certain loss function. The loss or objective function can be defined in different ways; this implementation uses the Mean Squared Error shown in Eq. (1).

3.4 Statistical Procedure

The proposed models, based on neural networks, were compared with a statistical technique. This traditional procedure employs the properties of gaussian distributions. This probability distribution is extensively used when modelling nature, social and psychological problems [9]. The properties that the statistical method exploits are related to the standard deviation and coverage of the normal distribution.

The 3-sigma rule states three facts; the first establishes that, in the range $[\mu - \sigma, \mu + \sigma]$ is approximately 68.26% of the distribution. About 95.44% of the data is within two standard deviations of the median $[\mu - 2\sigma, \mu + \sigma]$. This characteristic is commonly named as two-sigma rule. Finally, 99.74% of the distribution data falls in the range $[\mu - 3\sigma, \mu + 3\sigma]$. Considering the interval which contains about 95% of the instances, the remaining 5% can be considered abnormal values as they are out of the expected range. In this context, an anomalous element is a value that does not match with the normal pattern of the tenant electricity consumption. We used the two-sigma rule to label data of the consumption dataset for the supervised neural network experiment (Fig. 2).

3.5 Metrics

The approaches seen in this paper utilize the F1-score as the metric to assess the model's performance. The F1-score, given by (5), relates the precision P (6) and recall R (7) of every model. Precision is also known as the positive predicted value while recall is also referred to as sensitivity:

$$F1_{score} = 2 \frac{PR}{P + R} \quad (5)$$

$$P = \frac{TP}{TP + FP} \quad (6)$$

$$R = \frac{TP}{TP + FN} \quad (7)$$

where P , precision; R , recall; TP , number of true positives; FP , number of false positives; and FN , number of false negatives.

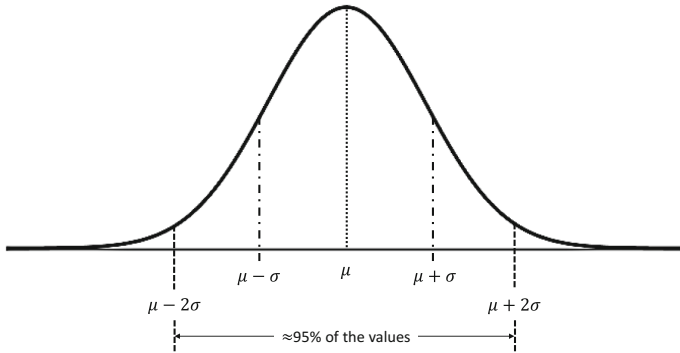


Fig. 2. Normal graph with distribution data properties.

4 Experiments

In this section, four experiments are presented: feature sets evaluation, supervised method, unsupervised method evaluation and the statistical procedure implementation.

4.1 Dataset

The information correspond to an individual household electric power consumption database generated by UC Irvine Machine Learning Repository [8]. It contains measurements of electric power consumption in a home with one-minute sampling rate over a period of almost 4 years. Data was split into two groups, measurements of 2007 are used for training and validation while values of 2008 are used for testing.

4.2 Features Evaluation

Feature selection is the most important part in contextual anomaly detection. Choosing the right variables can lead to high anomaly detection rates. In the neural networks approach evaluation, we built three distinct feature sets: A, B and C. Feature sets A and C are based on the sets used in [12] and [2], respectively. The third set, B, was proposed with an average number of features. The three sets are shown in Tables 1, 2 and 3. Every feature set is built with data of three categories: **temporal information** such as season, day of the week, month and so on; a 12 hourly values **consumption window**; and **generated data** derived from the window consumption such as mean, standard deviation, subtraction between elements of the window, interquartile range and percentiles. Although all three sets have the 12 hourly consumption window, they contain a different number of temporal values and generated data.

The unit circle projection used for temporal features consists of a representation to encode periodic variables [17]. The mapping is calculated by a sine and cosine component, as shown in formula (8):

$$\hat{t}_{n,k} = \left\{ \cos\left(\frac{2\pi t_{n,k}}{p_k}\right), \sin\left(\frac{2\pi t_{n,k}}{p_k}\right) \right\} \quad (8)$$

where $k = \{month, day\ of\ week, week\ of\ year, \dots\}$; $\hat{t}_{n,k}$, periodic value; p_k , known period. For $k = month \rightarrow p_k = 12$, for $k = day\ of\ week \rightarrow p_k = 7$ and so on. In contrast to other encodings, such as one-hot [16], which increases the input dimensionality, the unit circle representation requires only two dimensions per periodic value. One-hot encoding maps categorical variables to integer values and then represents each one with a binary vector. In each vector, only the element whose index equals the integer value is marked as one, the remainders are marked as zero.

Table 1. Feature set A.

Feature	Description
Month of year	Unit circle projection with $p_k = 12$
Day of week	Unit circle projection with $p_k = 7$
Week of year	Unit circle projection with $p_k = 53$
Hour of day	Unit circle projection with $p_k = 24$
Electricity consumption window	Window with 12 samples of energy active in one hour $S_j, \{j = 1, \dots, 12\}$
First difference	Difference between last and first elements of a consumption window $S_{12} - S_1$

Table 4 shows the highest F1-scores derived from the supervised and unsupervised method. The biggest difference between the sets can be observed in the unsupervised approach. In this experiment, the third feature group outperforms the first two sets by having a 0.899 F1-score. However, with the supervised implementation the difference is smaller. Although the third set continues being the best, the remaining sets almost overlap in their performance.

4.3 Unsupervised Method

In this approach, seven autoencoder models were evaluated. As can be seen in Table 5, their structure varies from one layer for the encoder and the decoder part, up to three layers in each. The number of neurons in each layer is in terms of n . It depends on the characteristic vector length, which is different for every proposed set. That is, n is 21 for feature set A, 28 for set B and 33 for set C. In the model M4, the number of neurons is 33 for set A and B, and it is 40 for set C. In order to calculate the F1-score for anomaly detection, we have used the ROC curve analysis to obtain a threshold for the reconstruction error in each autoencoder model. Table 6 shows the scores achieved for the seven models with the three feature sets. For each feature set there is a model that accomplishes the best performance. For set A the advisable model is M7 while for set B it is M6 and M5 for set C.

Table 2. Feature set B.

Feature	Description
Day of year	Unit circle projection with $p_k = 365$
Season	1–4 Winter, spring, summer, autumn
Month of year	Unit circle projection with $p_k = 12$
Day of week	Unit circle projection with $p_k = 7$
Hour of day	Unit circle projection with $p_k = 24$
Electricity consumption window	Window with 12 samples of energy active in one hour $S\{j = 1, \dots, 12\}$
\bar{x}	Mean of sensor data values in each consumption window
s	Standard deviation of sensor data values in each consumption window
$S_{12} - S_1$	Difference between last and first elements of a consumption window
$\bar{x}_i - \bar{x}_{(i-1)}$	Difference between the means of i^{th} and $(i-1)^{st}$ consumption windows

4.4 Supervised Method

In this section, four models of MLP were designed following the pattern shown in Table 7. As in the unsupervised experiment, the numbers of neuron instances depend on the characteristic vector length (n is 21 for set A, 28 for set B and 33 for set C). The models were trained for classification with labeled data of 2007. We used the two-sigma rule of normal distribution for labeling. Then the models were tested with data of 2008. Table 8 reports F1-scores accomplished with the supervised approach. In this case, all MLPs have accuracy better than that obtained by the unsupervised model based on autoencoders. All models have a score above 0.93. Moreover, the difference between the highest values with each set is just one percentage point. That is, set A with M4 had 0.947 while set B with M1 obtained 0.957 and set C with M2 gave 0.967.

5 Results and Discussion

The supervised approach outperforms the results of the unsupervised method and the statistical procedure. Besides, this method can be implemented on embedded systems due to the light structure of MLPs. In the supervised methods, we combine a statistical part and a neural network technique. This produces what can be named an hybrid model; at first, data rows are labeled with the two sigma rule and then, a multilayer perceptron is trained to execute classification. The feature set C performs better in both, supervised and unsupervised learning, as demonstrated by having the greatest F1-score values. This improvement is achieved by the increase of temporal and generated data. So we conclude that

Table 3. Feature set C.

Feature	Description
Day of year	Unit circle projection with $p_k = 365$
Season	1–4 Winter, spring, summer, autumn
Month of year	Unit circle projection with $p_k = 12$
Day of week	Unit circle projection with $p_k = 7$
Hour of day	Unit circle projection with $p_k = 24$
Electricity consumption window	Window with 12 samples of energy active in one hour $S\{j = 1, \dots, 12\}$
\bar{x}	Mean of sensor data values in each consumption window
s	Standard deviation of sensor data values in each consumption window
$S_{12} - S_1$	Difference between last and first elements of a consumption window
$\bar{x}_i - \bar{x}_{(i-1)}$	Difference between the means of i^{th} and $(i-1)^{st}$ consumption windows
$\bar{x}_{(i+1)} - \bar{x}_i$	Difference between the means of $(i+1)^{st}$ and i^{th} consumption windows
Q1	First quartile of the sensor data values in each window
Q2	Median of the sensor data values in each window
Q3	Third quartile of the sensor data values in each window
IQR	Interquartile range of the sensor data values in each window

Table 4. Performance of feature sets in neural networks approach.

Approach	F1-score		
	Feature set A	Feature set B	Feature set C
Unsupervised	0.540	0.791	0.899
Supervised	0.948	0.957	0.967

in order to have a high anomaly detection rate, additional information like the previously mentioned is needed. Contextual or temporal data represent relevant information concerning the individual consumption. This pattern depends on the moment it was measured, for example the season and the hour of the day, but also, generated data such as the mean and standard deviation of an hourly consumption reading can vary from one user to another. Analysing consumption in buildings to find unexpected behaviours is especially useful; it can lead users

Table 5. Structure of autoencoder models.

Model	Autoencoder					
	Encoder			Decoder		
	Number of neurons in layer					
	1	2	3	1	2	3
M1	$n/2$	-	-	n	-	-
M2	$2 * n/3$	-	-	n	-	-
M3	n	-	-	n	-	-
M4	33 or 40	-	-	33 or 40	-	-
M5	$n/2$	$n/3$	-	$n/2$	n	-
M6	$2 * n/3$	$n/2$	-	$2 * n/3$	n	-
M7	$2 * n/3$	$n/2$	$n/3$	$n/2$	$2 * n/3$	n

Table 6. F1-scores of unsupervised approach.

Model	F1-score		
	Feature set A	Feature set B	Feature set C
M1	0.457	0.762	0.807
M2	0.473	0.648	0.654
M3	0.518	0.588	0.863
M4	0.463	0.656	0.869
M5	0.530	0.789	0.899
M6	0.458	0.791	0.786
M7	0.540	0.782	0.793

Table 7. Structure of MLPs proposed.

Model	MLP			
	Number of neurons in layer			
	1	2	3	4
M1	$n/2$	2	-	-
M2	$n/2$	$n/3$	2	-
M3	$2 * n/3$	$n/2$	2	-
M4	$2 * n/3$	$n/2$	$n/3$	2

Table 8. F1-scores of supervised approach MLPs.

Model	F1-score		
	Feature set A	Feature set B	Feature set C
M1	0.942	0.957	0.966
M2	0.948	0.955	0.967
M3	0.939	0.932	0.951
M4	0.947	0.936	0.947

to save energy and for companies to avoid frauds. These are some ways to reduce the environmental impact of the electricity generation and usage.

6 Conclusion and Future Work

In this research, supervised and unsupervised approaches of neural networks for anomaly detection were researched. We compare them with a statistical procedure. Furthermore, feature selection has been studied. The results show that supervised learning outperforms the accuracy of unsupervised and statistical methods. Moreover, the feature set C performs best. The fact that the supervised approach takes advantage of statistical methods to label data before training is what leads to the improvement showed in the results regarding the anomaly detection rate. Because the usage behaviour can differ between buildings, the proposed methods for detecting anomalies in electricity consumption works individually for a certain home. In this sense, a topic for further research is the generation of models that can generalize the consumption in a group of houses. That way, we might decrease the resources occupied in training the neural network models. Future works could test the proposed neural network techniques with other real datasets. Additionally, other methods such as PCA or Support Vector Machines could be researched.

Acknowledgments. E. Zamora and H. Sossa would like to acknowledge CIC-IPN for the support to carry out this research. This work was economically supported by SIP-IPN (grant numbers 20180180 and 20180730) and CONACYT (grant number 65) as well as the Red Temática de CONACYT de Inteligencia Computacional Aplicada. J. García acknowledges CONACYT for the scholarship granted towards pursuing his MSc studies.

References

1. A beginner's guide to multilayer perceptrons. deeplearning4j.org/multilayer_perceptron. Accessed 25 May 2018
2. Araya, D.B., Grolinger, K., ElYamany, H.F., Capretz, M.A., Bitsuamlak, G.: Collective contextual anomaly detection framework for smart buildings, pp. 511–518. IEEE (2016)

3. Ashton, K.: That “Internet of Things” thing. *RFiD J.* **22**, 97–114 (2009)
4. Bengio, Y., et al.: Learning deep architectures for AI. *Found. Trends® Mach. Learn.* **2**(1), 1–127 (2009)
5. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. *ACM Comput. Surv. (CSUR)* **41**(3), 15 (2009)
6. Chou, J.-S., Telaga, A.S.: Real-time detection of anomalous power consumption. *Renew. Sustain. Energy Rev.* **33**, 400–411 (2014)
7. Costa, A., Keane, M.M., Raftery, P., O’Donnell, J.: Key factors methodology: a novel support to the decision making process of the building energy manager in defining optimal operation strategies. *Energy Build.* **49**, 158–163 (2012)
8. Dheeru, D., Taniskidou, E.K.: UCI machine learning repository (2017)
9. Gómez Chacón, I.M., et al.: Educación matemática y ciudadanía (2010)
10. Hajian-Tilaki, K.: Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Casp. J. Intern. Med.* **4**(2), 627 (2013)
11. IEA: World energy outlook 2011 executive summary (2011)
12. Tasfi, N.L., Higashino, W.A., Grolinger, K., Capretz, M.A.: Deep neural networks with confidence sampling for electrical anomaly detection, June 2017
13. Lyu, L., Jin, J., Rajasegarar, S., He, X., Palaniswami, M.: Fog-empowered anomaly detection in iot using hyperellipsoidal clustering. *IEEE Internet Things J.* **4**(5), 1174–1184 (2017)
14. Ouyang, Z., Sun, X., Chen, J., Yue, D., Zhang, T.: Multi-view stacking ensemble for power consumption anomaly detection in the context of industrial internet of things. *IEEE Access* **6**, 9623–9631 (2018)
15. Ouyang, Z., Sun, X., Yue, D.: Hierarchical time series feature extraction for power consumption anomaly detection. In: Li, K., Xue, Y., Cui, S., Niu, Q., Yang, Z., Luk, P. (eds.) *LSMS/ICSEE-2017. CCIS*, vol. 763, pp. 267–275. Springer, Singapore (2017). <https://doi.org/10.1007/978-981-10-6364-0-27>
16. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
17. Stover, C.: Unit circle
18. Yijia, T., Hang, G.: Anomaly detection of power consumption based on waveform feature recognition, pp. 587–591. *IEEE* (2016)