Improved Data Preprocessing Approach to Short-Term Load Forecasting

Athanasios Ioannis Arvanitidis
Dept. of Electrical and Computer Engineering
University of Thessaly
Volos, Greece
atarvanitidis@uth.gr

Dimitrios Bargiotas
Dept. of Electrical and Computer Engineering
University of Thessaly
Volos, Greece
bargiotas@uth.gr

***Abstract*—One of the most critical aspects for the smooth operation of power systems is short-term load forecasting. Forecast accuracy has a significant impact on an electricity utility's economic viability and reliability. Thus, robust deep learning methods, such as artificial neural networks, are implemented in order to achieve higher accuracy load forecasting results. In this paper, a new preprocessing method of the input data of a neural network, which emphasizes on the importance of specific input data, that show a higher Pearson's correlation coefficient with the output result, is proposed. This work implements the proposed preprocessing technique and compares the results with those derived from the classical min-max scaling methods. Numerical results of next hour's load forecasting, based on a multi-layer perceptron with the implementation of the proposed data scaling approach, show higher precision than the typical scaling method, demonstrating the importance of our work.**

***Index Terms*—short-term load forecasting, data preprocessing, scaling techniques, multi-layer perceptron**

I. Introduction

One of the most critical parts of effective power system management is the ability to forecast electrical load consumption. The accuracy of predictions has a direct impact on the economic feasibility and dependability of electricity systems. Short-Term Load Forecasting (STLF) covers a time span of one hour to one week and it is utilized for day-to-day power system operations, such as economic dispatch, demand response, energy transaction scheduling, power flow analysis, and power system reliability and stability research [1]. Short-term load forecasting has traditionally been performed using approaches such as time series models, regression-based algorithms, and Kalman filtering [2]. Recently, methods based on artificial intelligence and deep learning algorithms have been widely employed for power system optimization, since they outperform conventional approaches in terms of generalization and prediction [3]. Their primary applications include optimum power system operation and management, load forecasting and energy price forecasting.

In recent years, approaches based on Artificial Neural Networks (ANNs), as well as other computational intelligence methodologies, have emerged as potentially robust methods for short-term load forecasting. The increased availability of data due mainly to the expanded installation of new power meters and the breakthrough in the computational capability of current computers have contributed significantly to the recent success of neural networks. STLF is mostly dependent on historical load data, such as load data from prior days, weeks or yeas, as well as temperature and humidity data. The availability of load data per minute utilized by different types of neural networks achieves impressive performance as it brings even greater accuracy in the forecast results [4]. However, the data entered into the neural networks are not used in raw format, but they undergo into various types of preprocessing, such as eliminating outliers, handling missing values and feature scaling, so that they can be used properly and increase the efficiency of the forecasting model. Min-Max, z-score, standard and max absolute normalization are the most reputable techniques for scaling input data. Despite their extensive use, these strategies have certain drawbacks [5], which provides an opportunity to develop novel scaling techniques that increase the predictive abilities of ANNs.

This paper presents a unique data preprocessing technique that differs from earlier work in that it highlights the significance of specific input data by using neural networks to forecast next hour's load. The proposed technique focuses on the importance of certain neural network's input variables in relation to output variables, resulting in improved prediction outputs than usual preprocessing methods. This approach is applied to data from the Greek Power System and is utilized by a Multi-Layer Perceptron (MLP) for short-term load forecasting.

Our paper is developed as follows. In Section 2, the prevalent and most widely used preprocessing techniques of neural network's input data are presented. Section 3 presents precisely the analysis of the enhanced scaling method we propose as well as the improvement in accuracy that results in the short-term load forecasting under consideration, while Section 4 concludes the paper.

II. Preprocessing Techniques of Neural Network's Input data

Data preprocessing aims at making the raw data at hand more amenable to neural networks. This includes vectorization, normalization, handling missing values, and feature extraction [6].

 ISBN: 978-1-61208-967-6

A. Feature Selection

Feature selection is one of the initial steps in studying and understanding the dataset in order to construct a robust prediction model. The selection of suitable variables as input data for the neural network, in order to boost the accuracy of the prediction outcomes, is referred to as feature selection. Pearson's r correlation coefficient between each pair of variables is used to select features as input data. The Pearson correlation coefficient is a metric for determining the strength of a linear relationship between two variables of the dataset taking numbers between -1 and 1. When it is near to one, it indicates a significant positive association. When the coefficient is near to -1, it indicates a significant negative association. Finally, coefficients close to 0 indicate that no linear association exists. The Pearson correlation coefficient ignores whether a variable is considered dependent or independent, evaluating all variables identically. Pearson's correlation coefficient is given by (1):

$$r_{\mathrm{X,Y}} = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} \tag{1}$$

where $cov(X,Y)$ is the covariance, σ_X is the standard deviation of X and σ_Y is the standard deviation of Y.

Therefore, feature extraction and identification are one of the most important steps in the field of energy forecasting, including short term load forecasting [7]. The variables in the dataset that exhibit the strongest linear correlation with the load should be utilized in order to achieve higher accuracy and reduce the complexity of load forecasting. It is necessary to identify which characteristics selected from dataset are containing the most relevant information helping to provide accurate predictions. This crucial step is also applicable in the field of Energy where artificial intelligence algorithms are widely used [8]. In our work, the features with the highest Pearson's correlation with the output variable of the proposed MLP are selected as neural network's inputs.

B. Data Scaling

Data scaling is one of the most critical operations that should performed on the input data. Machine Learning (ML) methods, with a few exceptions, do not perform well when the input numerical characteristics have extremely varied scales [9]. In general, neural networks do not accept relatively big values or input data that are heterogeneous, i.e., there are substantial differences in the order of magnitude. As a result, to boost the neural network's performance, input data should contain values inside a closed interval.

Differences in the magnitude of scaling across input variables may increase the difficulty of the problem being approached. A model with large scale values is frequently unstable, which means it may perform poorly during learning and be sensitive to input values, resulting in larger generalization error. A basic linear rescaling of the input variables is one of the most prevalent types of preprocessing. In [10], the authors highlight that input data normalization can enhance neural networks' overperformance by reducing effectively the estimation errors and the computational time needed.

In forecasting approaches based on time series data, the most common normalizing methods are min-max, decimal scaling, z-score, median, and sigmoid normalization. A comparative study of these standard normalization techniques on the time series forecasting is presented in [11]. The authors use deep recurrent neural networks to predict the Bombay and New York stock exchanges by normalizing the input data using the methods described above and analysing the outcomes to determine which methodology is preferred. Meanwhile, Ogasawara et al. [12] propose an adaptive normalization technique for normalizing non-stationary time series. This innovative approach is used along a feed-forward neural network in order to predict numerous economic factors, producing greater results than the traditional normalization methods. Furthermore, in [13], the authors study the effectiveness of batch normalization technique in different types of convolutional neural networks concluding that the implementation of a normalization approach to the input data is inevitable.

The issue of data scaling has also influenced researchers' efforts for STLF, as it applies directly to the various types of ANNs used in the literature. Specifically, Che et al. [14] examine various machine learning algorithms in the STLF issue. In their work, the authors propose a fusion load forecasting model based on Support Vector Machines (SVM), Random Forests (RF), Long Short-Term Memory (LSTM) neural networks along with the Ensemble Empirical Mode Decomposition algorithm for dealing with the abnormal data. Their approach was tested on 15-min interval data yielding Mean Absolute Percentage Error (MAPE) lower than 3%. Furthermore, Yi et al. [15] propose a Multi-Temporal-spatial-scale Convolutional Network (MTCN) in order to reduce the data noise error, improve the time series features and enhance the prediction accuracy. The input data used in this model have been normalized via the standard min-max normalization method. The model has been tested using load data from Chinese power system producing better results in compare to the traditional ANN models used in the STLF issue. In [16], Kwon et al. study the impact of minimum-maximum, z-score and decimal normalization approaches to the input data of a MLP for the prediction of the load for 24 hours on weekdays. Using load and temperature data of the past two days of the Korean power system, came to the conclusion that the conventional min-max scaling outperforms the other two methods as it produces MAPE of 1,97%.

Most papers in the existing literature suggest that datasets should be subjected to a global normalization technique. In [17], Passalis et al. present some global normalization methods for the STLF issue. In contrast with the existing literature, this paper proposes that only some of the input variables should be normalized based on their impact on the predicted results. The proposed data scaling is done by multiplying certain input data with importance coefficient in order to obtain an order of magnitude that appropriately determines their influence in the result of the forecast.

 ISBN: 978-1-61208-967-6

III. Analysis of the Proposed Scaling Method

Following a thorough review of the literature, an innovative data processing technique is suggested and applied to certain specific data depending on the Pearson's correlation coefficient. An MLP neural network, which is used to predict the value of the next hour's load using historical temperature and load data from previous days and the previous hour, is presented in detail in this section. The data used, containing hourly load values, derives from the Greek national power system for the years 2013-2017, from which 80% is chosen as training set, while the remaining 20% consists the test set. Our proposed MLP neural network consist of three layers; an input layer, a hidden layer, and an output layer, as depicted in Fig. 1. Historical load data, meteorological data such as temperature, wind speed and direction, and data relating to the seasonality of the load, such as hour, day, month, etc., are included in the dataset. In order to reduce the complexity of the suggested forecasting model, only data with a high Pearson correlation coefficient related to the load variable are chosen as input variables. The input variables used for next hour's load forecasting are the following:

- Hour: The time of day for which the load forecast is made.
- Week Day: A characteristic coding to denote the day of the week.
- Holiday: Binary values are used to indicate whether a day is a holiday, which includes Greek state holidays, religious holidays and the weekends, or a normal working day
- Temperature: The hourly value (in Celsius) of the temperature of the day for which the load is forecast.
- D-7 Load: The value of the load at the corresponding time on the same day of the previous week.
- D-1 Load: The load value of the day preceding the one for which prediction is made, at the corresponding time.
- H-1 Load: The value of the previous hour's load on which the forecast is based.

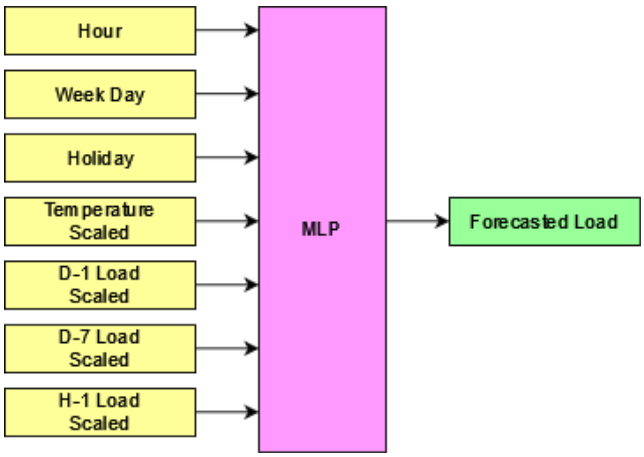

Fig. 1. Proposed MLP architecture for STLF.

Pearson's correlation coefficient of the input data is then calculated and compared to the neural network output, i.e., the load for the following hour. Data with a coefficient r near to +1 have a stronger impact on the outcome of the forecast and should thus be considered more important. The variables with higher r values compared to the output variable are $D\text{-}1Load$, $D\text{-}7Load$, $H\text{-}1Load$. As a result, in order to improve the predicting results, these variables with r approaching +1 in respect to the load variable are subjected to an improved scaling technique. The main benefit of this particular scaling for variables that have a strong correlation with the load variable is that they are given greater significance, allowing the neural network to use this knowledge and improve the forecasting accuracy. Fig. 2 summarizes the autocorrelation coefficient calculation results.

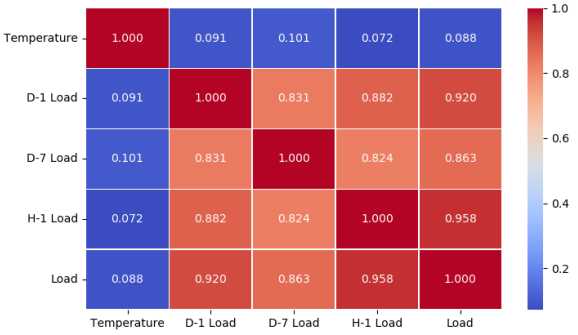

Fig. 2. Calculation of Pearson's correlation coefficient for input variables.

As previously stated, in order to produce more accurate prediction outcomes using neural networks, the input data should be scaled appropriately. Initially, the variables $Hour$, $WeekDay$ and $Holiday$ serve as labels for the day on which the prediction is created and are not susceptible to scaling. The $Temperature$ variable is subjected to the standard min-max scaling approach. Because of the large value of the coefficient r, the variables $D\text{-}1Load$, $D\text{-}7Load$ and $H\text{-}1Load$ are subject to both the standard min-max scaling approach and the modified min-max scaling method in (2). This paper proposes an enhanced min-max data preprocessing technique for STLF, that alters the order of magnitude of the variables $D\text{-}1Load$, $D\text{-}7Load$, $H\text{-}1Load$ giving them the appropriate weight, and compares the forecasing results with those obtained from the conventional implementation of the min-max method.

$$y = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \cdot ImpCoeff \tag{2}$$

where $ImpCoeff$ is an integer that appropriately identifies the significance of the data for the forecast result by allocating the input data within the closed interval [0, ImpCoeff].

A. Calculation of Importance Coefficient for the Enhanced Min-Max Scaling Method

The $ImpCoeff$ coefficient, which correctly attributes the significance of these variables, must be determined before

applying the suggested scaling method to the input data. $ImpCoeff$ is defined by the accuracy of the neural network prediction by calculating the resultant MAPE value through a trial-and-error procedure. At first, the coefficient accepts integer values in the range [1,100]. It is underlined that when the coefficient equals 1, the suggested method is associated with the traditional min-max scaling methodology. Fig. 3 depicts the MAPE values obtained by implementing the suggested MLP for the STLF at various $ImpCoeff$ values.

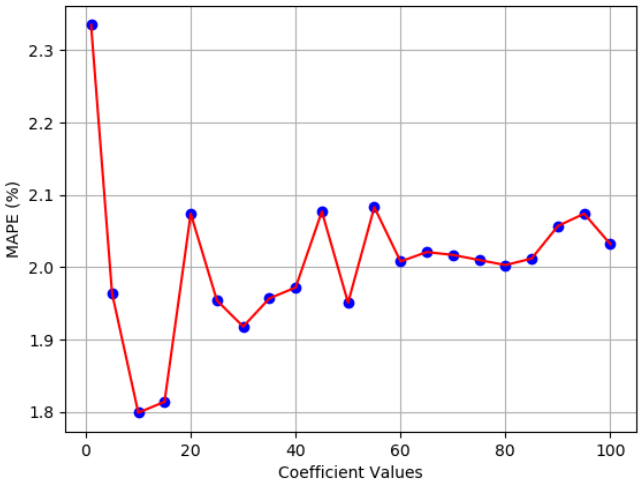

Fig. 3. MAPE calculation for the various ImpCoeff values.

Then it is discovered that for coefficient values in the interval [7,12], MAPE obtains the lowest values. Fig. 4 depicts the thorough computation of $ImpCoeff$ in this interval, stressing that when $ImpCoeff$ equals 10, MAPE yields the smallest feasible value. As a result, with $ImpCoeff$ equal to 10, our suggested scaling approach is obtained from (2).

B. Numerical Results

Before entering the proposed MLP neural network to estimate the following day's load, input data with the highest coefficient r value is exposed to both scaling strategies. Table I summarizes and compares the outcomes of both procedures. As predicted, the technique with the lowest MAPE is deemed to be more efficient.

TABLE I
MAPE CALCULATION FOR THE TWO SCALING TECHNIQUES OF INPUT DATA

Scaling Method	MAPE
Classic Min-Max Scaling	2.34%
Enhanced Min-Max Scaling	1.80%

It turns out that our enhanced Min-Max Scaling technique yields a lower MAPE value in the forecast. Despite its simplicity, this technique appropriately emphasizes the weight and importance of the input variables D-1$Load$, D-7$Load$ and

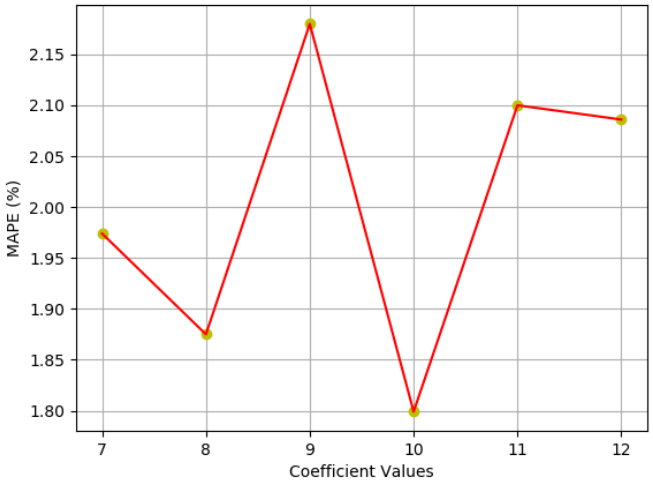

Fig. 4. Computation of optimum value for $ImpCoeff$.

H-1$Load$ in terms of MLP performance. When the proposed preprocessing technique is applied to the input data of a MLP neural network, the resulting value of MAPE decreases below to 2%, resulting in one of the lowest prediction value in the literature, based on data of the Greek interconnected power system. Fig. 5 and Fig. 6 provide a graphical comparison of prediction outcomes in 2017 using the proposed MLP for estimating next hour's load. In comparison to the usual scaling strategy, it is clear that the suggested method's outputs closely match the real load curve.

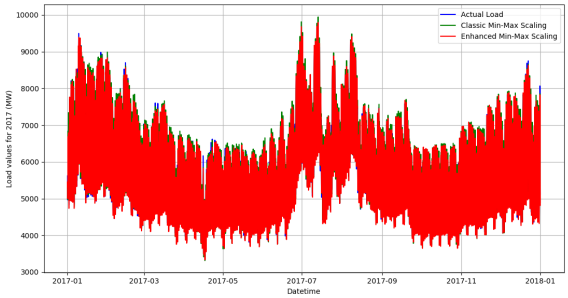

Fig. 5. Load curves for the year 2017.

IV. Conclusion

The increased use of neural networks in short-term load forecasting necessitates the development of novel data preparation approaches to increase the forecasting model's accuracy. In this paper, an enhanced preprocessing technique is presented that is applied to the input data of an MLP neural network to predict the value of the load in the following hour. This approach is based on the precise determination of a coefficient that assigns the proper importance to particular input data that demonstrate a high degree of correlation with

 ISBN: 978-1-61208-967-6

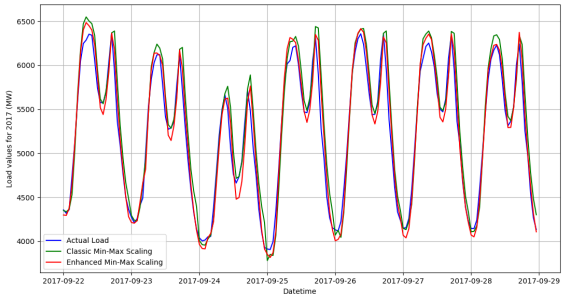

Fig. 6. Comparison of load prediction.

the proposed MLP's forecast output. Despite its simplicity, the findings of short-term load forecasting are more accurate when compared to other results in the literature that use data from the Greek interconnected system, as indicated by the low MAPE value, which is around 1.80%.

REFERENCES

[1] E. Kyriakides and M. Polycarpou, *Short Term Electric Load Forecasting: A Tutorial*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 391–418.

[2] M. Alamaniotis, D. Bargiotas, and L. Tsoukalas, "Towards smart energy systems: Application of kernel machine regression for medium term electricity load forecasting," *SpringerPlus*, vol. 5, 12 2016.

[3] A. I. Arvanitidis, D. Bargiotas, A. Daskalopulu, V. M. Laitsos, and L. H. Tsoukalas, "Enhanced short-term load forecasting using artificial neural networks," *Energies*, vol. 14, no. 22, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/22/7788

[4] D. Kontogiannis, D. Bargiotas, and A. Daskalopulu, "Minutely active power forecasting models using neural networks," *Sustainability*, vol. 12, no. 8, 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/8/3177

[5] P. Koprinkova and M. Petrova, "Data-scaling problems in neural-network training," *Engineering Applications of Artificial Intelligence*, vol. 12, no. 3, pp. 281–296, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197699000081

[6] D. Kontogiannis, D. Bargiotas, A. Daskalopulu, and L. H. Tsoukalas, "A meta-modeling power consumption forecasting approach combining client similarity and causality," *Energies*, vol. 14, no. 19, 2021. [Online]. Available: https://www.mdpi.com/1996-1073/14/19/6088

[7] I. Jebli, F.-Z. Belouadha, M. I. Kabbaj, and A. Tilioua, "Prediction of solar energy guided by pearson correlation using machine learning," *Energy*, vol. 224, p. 120109, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544221003583

[8] D. Rochman, E. Baugé, A. L. Vasiliev, H. Ferroukhi, S. Pelloni, A. J. Koning, and J.-C. Sublet, "Monte carlo nuclear data adjustment via integral information," *The European Physical Journal Plus*, vol. 133, pp. 1–23, 2018.

[9] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed. O'Reilly Media, Inc., 2017.

[10] J. Sola and J. Sevilla, "Importance of input data normalization for the application of neural networks to complex industrial problems," *IEEE Transactions on Nuclear Science*, vol. 44, no. 3, pp. 1464–1468, 1997.

[11] S. Bhanja and A. Das, "Impact of data normalization on deep neural network for time series forecasting," *ArXiv*, vol. abs/1812.05519, 2018.

[12] E. Ogasawara, L. C. Martinez, D. de Oliveira, G. Zimbrão, G. L. Pappa, and M. Mattoso, "Adaptive normalization: A novel data normalization approach for non-stationary time series," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, 2010, pp. 1–8.

[13] V. Thakkar, S. Tewary, and C. Chakraborty, "Batch normalization in convolutional neural networks — a comparative study with cifar-10 data," in *2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT)*, 2018, pp. 1–5.

[14] W. Guo, L. Che, M. Shahidehpour, and X. Wan, "Machine-learning based methods in short-term load forecasting," *The Electricity Journal*, vol. 34, p. 106884, 01 2021.

[15] L. Yin and J. Xie, "Multi-temporal-spatial-scale temporal convolution network for short-term load forecasting of power systems," *Applied Energy*, vol. 283, p. 116328, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261920317128

[16] B.-S. Kwon, R.-J. Park, S.-W. Jo, and K.-B. Song, "Analysis of short-term load forecasting using artificial neural network algorithm according to normalization and selection of input data on weekdays," in *2018 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, 2018, pp. 280–283.

[17] N. Passalis and A. Tefas, "Global adaptive input normalization for short-term electric load forecasting," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 1–8.

 ISBN: 978-1-61208-967-6