Probabilistic Deep Autoencoder for Power System Measurement Outlier Detection and Reconstruction

You Lin, *Student Member, IEEE*, and Jianhui Wang, *Senior Member, IEEE*

Abstract—**A probabilistic deep autoencoder is proposed to reconstruct power system measurements in this paper, which can be utilized in outlier detection and reconstruction. A nonparametric distribution estimation method is employed to capture the uncertainty information of the measured data. The estimated confidence intervals of the measured data are extracted from the estimated distribution and used as input to the first layer of neural networks. Through multilayer encoding and decoding processes, the intervals of measurements are reconstructed, which are further applied to detect and replace outliers. Simulation results verify the effectiveness of the proposed method.**

Index Terms—**Power system measurements, outlier detection, data reconstruction, probabilistic deep autoencoder.**

I. INTRODUCTION

ACCURATE measurements are essential for improving the effectiveness of power system analysis techniques. Outliers and missing data are severe problems that impact the accuracy of such analyses. Methods have been proposed for detecting outliers and estimating the missing values [1]. In previous studies, the detection strategies are either based on the unconditionally estimated distributions or not effective for large-dimensional datasets [2]. Hence, these outlier detection models are not adaptive to the varying dependent conditions of the target to be detected. It is essential to find new detection strategies applicable to big data in power systems. Moreover, most of the data reconstruction methods are deterministic which ignore the important probabilistic information of the reconstructed data [3]. In comparison, many of the popular probabilistic machine learning models are mostly utilized for generating similar data that have the same characteristics as the original dataset instead of reconstructing the replacement of the bad data, such as variational autoencoder [4] and generative adversarial net [5]. These methods have a complex distribution inference in the latent layers, thereby resulting in the difficulty in solving the model parameters. Therefore, it is essential to propose new probabilistic machine learning methods for outlier detection and data reconstruction. In this paper, an easy-to-execute conditional estimation approach with a probabilistic deep autoencoder (PDAE) is proposed to construct the intervals utilized for measurement outlier detection and reconstruction. The proposed method has good accuracy and provides quantitative ranges for the data to be reconstructed.

Manuscript received February 25, 2019; revised May 11, 2019; accepted August 13, 2019. Date of publication August 22, 2019; date of current version February 19, 2020. The work of J. Wang was supported by the U.S. Natural Science Foundation under Grant 1800716. Paper no. PESL-00043-2019. *(Corresponding author: Jianhui Wang.)*

The authors are with the Department of Electrical and Computer Engineering, Southern Methodist University, Dallas, TX 75205 USA (e-mail: youl@smu.edu; jianhui@smu.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2019.2937043

II. METHODOLOGY

A. Framework

Firstly, we estimate the distributions of each measurement and corresponding confidence intervals with a coverage probability $\rho \in [0, 1]$ using a kernel density estimation (KDE) model. Secondly, constructing an activation function using the KDE estimated distribution and a coverage probability ρ corresponding to a given normalized measurement. The activation function is utilized in the first layer of the proposed PDAE model to provide the estimated lower and upper bounds which will be the input to the second layer of the PDAE model. Meanwhile, the multilayer structure is further constructed. Then, intervals with lower and upper bounds of the measurements are obtained. Measurements outside the intervals are regarded as outliers. The estimated intervals will be the most possible ranges of the reconstructed data points to replace the outliers in the dataset.

B. Pre-Estimated Confidence Intervals From Nonparametric Estimation

1) Data Normalization: Given the measured time series $x = \{x_t\}_{t=1}^{N}$, normalize the data into range [0, 1]. N is the number of measurements for each feature.

2) Probability Distribution: To estimate the probability density function $f(y)$ of the target variable y with its corresponding explanatory vector $\boldsymbol{x}_y$, the KDE model [6] can be formulated as,

$$f(y|\boldsymbol{x}_y)_{Y|X_Y} = \frac{1}{h_Y}\frac{\sum_{i=1}^{\mathrm{N}} K\left(\frac{\boldsymbol{x}_y-\boldsymbol{x}_i}{\boldsymbol{H}}\right)K\left(\frac{y-y_i}{h_Y}\right)}{\sum_{i=1}^{\mathrm{N}} K\left(\frac{\boldsymbol{x}_y-\boldsymbol{x}_i}{\boldsymbol{H}}\right)} \quad (1)$$

where bandwidth parameter vector $\boldsymbol{H} = [h_1, h_2, \ldots, h_D]^{\mathrm{T}}$; $h_1, h_2, \ldots, h_D$ and h_Y are respectively the bandwidth parameters of the D-dimensional explanatory vector $\boldsymbol{x}_y$ and 1-dimensional target y; D is the dimensionality of the explanatory vector $\boldsymbol{x}_y$; K is the kernel function and a Gaussian kernel function is utilized in this paper; and

$$K\left(\frac{\boldsymbol{x}_y-\boldsymbol{x}_i}{\boldsymbol{H}}\right) = \prod_{j=1}^{D} K\left(\frac{x_{y,j}-x_{i,j}}{h_j}\right) \quad (2)$$

For the measured time series $x = \{x_t\}_{t=1}^{N}$, the target to be estimated is the measured value x_t at time t, and its corresponding explanatory vector is the past D measurements. That is, $y = x_t$, and $\boldsymbol{x}_y = [x_{t-1}, \ldots, x_{t-D}]$. By estimating $h_1, h_2, \ldots, h_D$ and h_Y, the probability distribution $f(x_t)$ of x_t can be obtained from (1).

3) Confidence Intervals: From the above-estimated distribution $f(x_t)$, the $(\rho \times 100)$ percent confidence interval can be

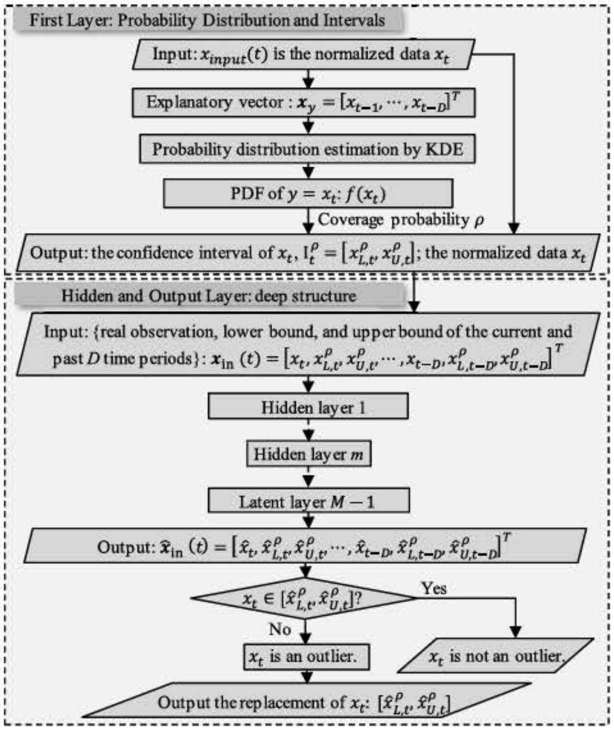

Fig. 1. Framework of the proposed PDAE model.

calculated from its α-quantile q_t^α and $(1-\alpha)$-quantile $q_t^{1-\alpha}$. Here, $\alpha \in [0, 0.5]$ and $\rho = 1 - 2\alpha$. Therefore, the estimated $(\rho \times 100)$ percent confidence interval I_t^ρ from the above distribution $f(x_t)$ can be represented by (3).

$$\mathrm{I}_t^\rho = \left[x_{L,t}^\rho, x_{U,t}^\rho\right] = \left[q_t^\alpha, q_t^{1-\alpha}\right] \tag{3}$$

where, $x_{L,t}^\rho$ and $x_{U,t}^\rho$ are the lower and upper bounds of the normalized measurement x_t at time t, respectively.

Therefore, given an input x_t, the $(\rho \times 100)$ percent confidence interval I_t^ρ can be calculated from the distribution function $f(x_t)$ and the pre-defined coverage probability ρ.

C. Probabilistic Deep Autoencoder

The PDAE framework we propose is a multilayer neural network to extract the features and then reconstruct the input upper and lower bounds of the statistical data intervals. It consists of an input layer, multiple hidden layers, and an output layer. The framework of the proposed PDAE model is shown in Fig. 1. The M-layer PDAE model is formulated as follows.

First layer L_1: The first layer is a data pre-processing layer based on the pre-estimation model in Section II-B. The objective of the first layer is to refine the original normalized data using the KDE-estimated lower and upper bounds of the normalized measurement. The input $x_{input}(t)$ of this layer is the normalized original data. That is, $x_{input}(t) = x_t$. The activation function consists of the pre-defined coverage probability ρ and the KDE-estimated distribution function $f(x(t))$ obtained from (2) with respect to the input. The output of the first layer consists of the lower bound $x_{L,t}^\rho$ and upper bound $x_{U,t}^\rho$ of x_t, along with the original normalized value x_t.

To take advantage of as many features in dataset, the features of the past D time periods are also considered. Thus, the input of the first hidden layer is $\boldsymbol{x}_{\text{in}}(t) = [x_t, x_{L,t}^\rho, x_{U,t}^\rho, \ldots, x_{t-D}, x_{L,t-D}^\rho, x_{U,t-D}^\rho]^T$, as shown in Fig. 1.

Hidden layer L_m: With the input $\boldsymbol{x}_m(t,k)$, the output $\boldsymbol{y}_m(t,k)$ of the mth encoding layer in the kth iteration can be calculated from (4) and (5).

$$\boldsymbol{y}_m(t,k) = s_m(\mathbf{w}_m(k)\boldsymbol{x}_m(t,k) + \mathbf{b}_m(k)) \tag{4}$$

$$\boldsymbol{x}_m(t,k) = \boldsymbol{y}_{m-1}(t,k) \tag{5}$$

The corresponding output of the (M-m)th decoding layer is

$$\boldsymbol{y}_{M-m}(t,k) = g_m(\mathbf{w}_{M-m}(k)\boldsymbol{x}_{M-m}(t,k) + \mathbf{b}_{M-m}(k)) \tag{6}$$

$$\boldsymbol{x}_{M-m}(t,k) = \boldsymbol{y}_{M-m-1}(t,k) \tag{7}$$

where, $s(\cdot)$ and $g(\cdot)$ are the activation functions; $\mathbf{w}$ and $\mathbf{b}$ are the weights and bias of each layer, respectively.

Output layer L_M: Output the reconstructed values of $\boldsymbol{x}_{\text{in}}(t)$, represented by $\boldsymbol{y}_M(t,k) = \widehat{\boldsymbol{x}}_{\text{in}}(t,k) = [\hat{x}_t(k), \hat{x}_{L,t}^\rho(k), \hat{x}_{U,t}^\rho(k), \ldots, \hat{x}_{t-D}(k), \hat{x}_{L,t-D}^\rho(k), \hat{x}_{U,t-D}^\rho(k)]^T$.

The cost function $J(k)$ of the formulated model includes the mean square error, the L_2 regulation term Ω_{weight} [7], the sparsity term $\Omega_{sparsity}$ [7], and the dispersion regulation term $\Omega_{dispersion}$.

$$J(k) = \frac{1}{N}\sum_{t=1}^{N} \|\boldsymbol{x}_{\text{in}}(t) - \widehat{\boldsymbol{x}}_{\text{in}}(t,k)\|^2 + \lambda \cdot \Omega_{weight} + \beta \cdot \Omega_{sparsity} + \gamma \cdot \Omega_{dispersion} \tag{8}$$

where, $\Omega_{dispersion}$ is a penalty term for $\widehat{\boldsymbol{x}}_{\text{in}}(t,k)$ falling outside the estimated confidence intervals, which is calculated from (9). If $\widehat{\boldsymbol{x}}_{\text{in}}(t,k)$ falls outside the estimated confidence interval I_t^ρ, the dispersion term will be larger than $\widehat{\boldsymbol{x}}_{\text{in}}(t,k)$ falling into the interval.

$$\Omega_{dispersion} = \sum_{t=1}^{N}\sum_{d=0}^{D}\left\{\left[2\hat{x}_{t-d}(k) - x_{L,t-d}^\rho - x_{U,t-d}^\rho\right]^2 + \left[2\hat{x}_{L,t-d}(k) - x_{L,t-d}^\rho - x_{U,t-d}^\rho\right]^2 + \left[2\hat{x}_{U,t-d}(k) - x_{L,t-d}^\rho - x_{U,t-d}^\rho\right]^2\right\} \tag{9}$$

Solving the parameters using a gradient method, we can obtain the reconstructed values of $\boldsymbol{x}_{\text{in}}(t)$, represented by $\widehat{\boldsymbol{x}}_{\text{in}}(t)$, as shown in Fig. 1. $\hat{x}_t$, $\hat{x}_{L,t}^\rho$ and $\hat{x}_{U,t}^\rho$ are the reconstructed values of x_t, $x_{L,t}^\rho$ and $x_{U,t}^\rho$, respectively.

If the normalized measurement x_t falls out of the reconstructed interval, it is an outlier. If x_t is an outlier or missing value, its estimated possible range is $[\hat{x}_{L,t}^\rho, \hat{x}_{U,t}^\rho]$.

Since D historical measurements are utilized, the proposed model is designed for scenarios with discrete and a limited number of consecutive outliers or missing values. However, the method can be applied indirectly in scenarios including more consecutive outliers or missing data points by preprocessing the raw data by disrupting the data sequence.

III. Numerical Study

Simulated power system measurements are utilized to verify the effectiveness of the proposed PDAE model. Outliers with a uniform distribution in the measurements are simulated.

Fig. 2 provides the reconstructed intervals with a coverage probability $\rho = 40\%$ which are marked with black solid lines. Obviously, the reconstructed intervals have good coverage of the replaced values of the outliers. In this figure, only one outlier escapes the detection. The outlier detection accuracy is 99.24%. We also analyze the outlier detection accuracy sensitivity of the proposed PDAE model with respect

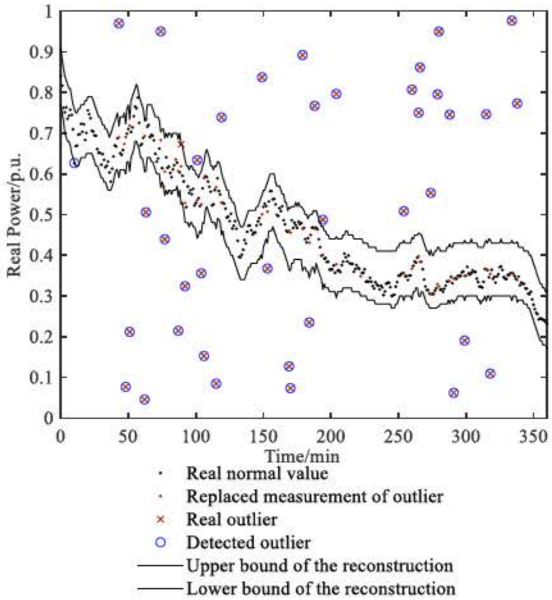

Fig. 2. Outlier detection and 40% coverage interval reconstruction with uniformly distributed outliers.

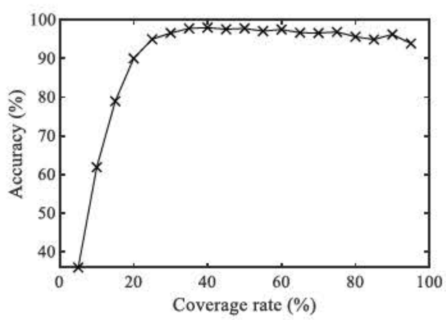

Fig. 3. Accuracy sensitivity of the proposed PDAE model with respect to different coverage probabilities.

to different coverage probabilities in Fig. 3. It demonstrates that the proposed PDAE model has good accuracy when the interval coverage probability $\rho \in [25\%, 80\%]$. Because if ρ approaches 0, the reconstructed interval may be too narrow, in which case more normal values may be detected as outliers. Vice versa, if ρ approaches 1, the reconstructed interval may be so wide that more outliers will be detected as normal values.

Three benchmark outlier detection methods including z-score, DBSCAN, and iForest are utilized to verify the performance of the proposed PDAE method. The outlier detection results of the proposed PDAE method and the benchmark methods are shown in Fig. 4. The results show that DBSCAN and the proposed PDAE model outperform z-score and iForest models, while PDAE has a better performance in detecting outliers scattered on the margin of normal values. To verify the stability of all the benchmark methods, the average accuracy of four methods corresponding to 10 random tests are respectively presented in Table I. The results demonstrate that the proposed PDAE method is more accurate than the other benchmarks.

In this paper, we utilize 20000 samples in the training process to get a better balance in outlier detection accuracy and training time. The training time of the proposed model is

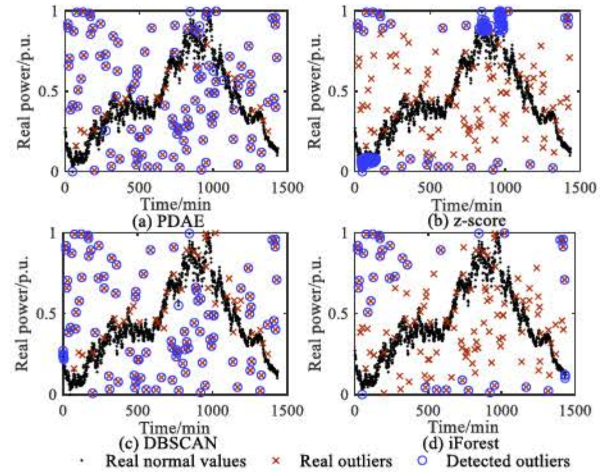

Fig. 4. Outlier detection results of different methods.

TABLE I
AVERAGE ACCURACY OF DIFFERENT OUTLIER DETECTION METHODS FOR 10 TESTS

	PDAE	z-score	DBSCAN	iForest
Average Accuracy (%)	98.19	85.37	96.86	93.10

less than 7 mins using MATLAB 2018b in a computer with 2.5 GHz, 8 GB of RAM and an Intel Core i5 CPU processor. Therefore, it can be utilized in online detection by updating the model every 7 mins or less time with more powerful computers.

IV. CONCLUSION

A probabilistic deep autoencoder (PDAE) is proposed in this paper, in which nonparametric estimated distributions are utilized to construct the uncertainty intervals of measurements in the first layer of neural networks. Deep autoencoder structures are formulated based on the nonparametric estimated uncertainty intervals of the measurements. Outliers can be accurately detected with the proposed PDAE model. The reconstructed intervals can provide good uncertainty estimation to substitute outliers in the measurements.

REFERENCES

[1] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, "A comparative evaluation of outlier detection algorithms: Experiments and analyses," *Pattern Recognit.*, vol. 74, pp. 406–421, Feb. 2018.

[2] G. B. Giannakis, Y. Shen, and G. V. Karanikolas, "Topology identification and learning over graphs: Accounting for nonlinearities and dynamics," *Proc. IEEE*, vol. 106, no. 5, pp. 787–807, May 2018.

[3] Y. Lin, J. Wang, and M. Cui, "Reconstruction of power system measurements based on enhanced denoising autoencoder," in *Proc. IEEE PESGM*, 2019, pp. 1–5.

[4] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014.

[5] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. NIPS*, 2014, pp. 2672–2680.

[6] R. J. Bessa, V. Miranda, A. Botterud, J. Wang, and E. M. Constantinescu, "Time adaptive conditional kernel density estimation for wind power forecasting," *IEEE Trans. Sustain. Energy*, vol. 3, no. 4, pp. 660–669, Oct. 2012.

[7] The Mathworks. *TrainAutoencoder*. Accessed: Oct. 2018. [Online]. Available: https://www.mathworks.com/help/deeplearning/ref/trainautoencoder.html