A Long Short-Term Memory Network-Based Intrusion Detection Method for Power Grid Middle Platform

Zhengyang Ding¹*, Jijun Wang² and Yiqing Cheng²

¹State Grid Jiangsu Electric Power Co., LTD. Information & Techcommunication Branch, Nanjing, China
²Jiangsu Electric Power Information Technology Co., LTD., Nanjing, China

*Corresponding author email: cs_zyding@163.com

Abstract. The power grid middle platform data has the feature of time-correlation, high dimensionality, and susceptibility to external factors. Based on the inherent characteristics of the power data and the theory of Long Short-Term Memory networks (LSTM), this paper proposed a stacked-LSTM based intrusion detection algorithm for power grid middle platform data. Technically, the strategy of constructing a dataset with replacement is adopted firstly, thus K sets of data is constructed. Then, a three-layer LSTM network is employed to conduct the feature extraction of high-dimensional data. Meanwhile, a two-layer fully connected hidden layer is used to generate a neural network and achieved user feature data matching. The high probability events are used to classify the node that appears the most in the results of the K data sets. Finally, experiments on real dataset proved that compared with traditional algorithms, the proposed algorithm has improved detection accuracy while reduced false alarm rates.

1. Introduction

With the rapid development of social economy, power resources are playing an increasingly important role in modern daily life. However, in order to reduce the payment of electricity bills, some users resort to fraudulent means, causing serious economic losses to the power company. In order to solve this problem, the power company has taken many measures and achieved certain results. The rapid development of smart grid technology provides a new research method for the analysis of abnormal electricity consumption data. The user power consumption data collected by big data power equipment is more abundant [1]. Data volume has become richer, storage and network equipment have become more and more advanced.

Under this background, the research on power data intrusion detection based on big data and artificial intelligence technology has become a recent area of research focus in the field of power big data. Based on statistical learning theory, Sheng et al. [2] applied polynomial fitting technology to the detection of abnormal electricity consumption data, which effectively improved the efficiency of electricity consumption forecasting. Bianco et al. [3] analysed the inherent feature and correlations of users' electricity data, then converted the user's power data detection problem into a linear regression problem, and the accuracy of the user's electricity data detection has been further improved. Study [4] analysed the information entropy theory and autoregressive theory, proposed an autoregressive mobility model, which effectively reduces the negative impact of seasonal changes on the detection of abnormal electricity consumption. Combining mathematical modelling and expert system, Messinis et al. [5] equipped with the expert knowledge of power company's long-term operation and mathematically modelled the time correlation of electricity consumption data, and realized the
improvement of the detection accuracy of abnormal electricity consumption of users. Li et al. [6] modelled the power data as a clustering model based on K-means and support vector neural network. They proposed an improved line loss rate algorithm in the station area, which effectively improved the classification performance of the line loss rate of user power consumption data. From the perspective of economics, Shi et al. [7] modelled power user data as a game-theory model and studied the characteristics of power consumption data. Data analysis algorithms based on the artificial intelligence theories like as deep learning have earned wildly approve recently. With the assistance of recurrent neural network and LSTM network theory, [8] proposed deep learning-based consumer electricity data analysis, which effectively improved The accuracy of the data analysis results, experiments verified that their algorithm has high operating efficiency. However, with the rapid growth of power user data and power equipment, the dimensionality and data volume of power middle platform data have also increased rapidly, resulting in the performance of existing power data intrusion detection algorithms being insufficient to meet the detection requirements of such data.

To address the above challenge, this paper analyses the high time correlation and high dimensionality characteristics of user electricity consumption data, proposed a novel intrusion detection algorithm based on LSTM neural network for the power middle platform. Experiments verify that the proposed intrusion detection algorithm has more advantages than the competitors, and the performance of intrusion detection has been improved. The rest of this paper are organized as follows. Section 2 formulates the proposed method in detail. Then, experimental evaluations are presented in Section 3. The conclusions and future works are drawn in Section 4.

2. Proposed Method

2.1. Pre-processing

To improve the data quality and reduce the problem of low performance of intrusion detection algorithms caused by errors such as data duplication and missing data, the data needs to be pre-processed before the detection algorithm is executed. For duplicate data, a direct deletion strategy is adopted. For missing data and wrong data, establish the upper limit and lower limit of the data field firstly. Then, the random selection strategy is used to fill in the upper and lower limit interval. Finally, all data values are normalized to avoid data quality problems caused by different value ranges. Denote the original data series as \( X = \{x_1, x_2, \ldots, x_n\} \), the normalize process can be formulated as

\[
x'_i = \frac{x_i - \text{mean}(X)}{\max(X) - \min(X)}
\]

where \( \text{mean}(X) \) is the mean value of data series \( X \), \( \max(X) \) is the maximum of data series \( X \), and \( \min(X) \) is the minimum of data series \( X \). Thus, \( x'_i \) denotes the normalized value.

With the development of storage equipment and storage technology, the amount of user electricity data stored by power companies has exploded, providing abundant data for abnormal electricity usage detection. When performing an algorithm for detecting anomalies in electricity usage, over-fitting problems often occur. To decrease the impact of overfitting on the detection results, in the process of generating the training set and the test set, the strategy of constructing data sets with replacement is adopted to construct multiple data sets to improve the randomness of the data sets. Next, select K data sets, use 90% of them as the training set for training, and use the remaining data as the test set for testing.

Meanwhile, intrusion detection algorithms need to have the ability to process high-dimensional data if we want to analyse data within a long period. To solve this problem, based on the LSTM theory, we proposed an anomaly data detection algorithm that can solve high-dimensional data, and realize the full mining of the key features hidden in high-dimensional data. The more key features that are mined, the better the detection performance of the detection algorithm, and the more favourable it is to propose a superior detection algorithm.
2.2. Stacked-LSTM

In traditional neural networks, such as Recurrent Neural Network (RNN), the hidden layer can only be set to one state. Therefore, this type of network is very sensitive to short-term input, has the problem of disappearing gradients, and is not suitable for processing long series data, especially time-series data. The LSTM network adds an extra state in the hidden layer to store the long-term state, avoiding the problem of vanishing gradient caused by RNN. Therefore, LSTM is particularly suitable for processing time series data and is well-known for solving the time-series data prediction and anomaly detection. The schematic figure of the logical architecture of the LSTM unit is shown in Figure 1. When the data at time $t$ is input to the LSTM, it will participate in the calculation along with the long-term state and output at time $t-1$ to obtain the output of the LSTM at time $t$. After dimensional transformation, it becomes the predicted data at time $t$.

![Figure 1. An instance of the LSTM neural logical architecture.](image)

The key to LSTM is to control the long-term state $c_t$. Make each output of the network the result of the long-term state participating in the operation to avoid the problem of vanishing gradient. To this end, LSTM sets up 3 gates: forget gate, input gate, and output gate. First is the forget gate, which is used to control whether the LSTM forgets the state of the previous hidden cell with a certain probability. The second is the input gate, which is used to process the input of the current sequence position. The unit state at the current moment is updated at this stage. The calculation formula is shown in formula (2)

$$c_t = \text{sigm}(f_t) \times c_{t-1} + \text{sigm}(i_t) \times \tanh(c_t)$$

After this stage, the combination of LSTM current memory and long-term memory completes the update of the united state. The last is the output gate, which is used to calculate the output of the current sequence position. At this stage, the calculation formula for calculating the output at the current moment is as follows

$$h_t = \text{sigm}(o_t) \times \tanh(c_t)$$

The output gate and the unit state $c_t$ together determine the output of the LSTM. Figure 1 is only the logical architecture diagram of the LSTM unit. In fact, each operation involves a neuron ($\lambda$ is the number of neurons in the hidden layer). The output of the LSTM is also a $\lambda$-dimensional vector. Therefore, the LSTM model needs to set up a fully connected layer at the last, and change the dimensionality of the output of the LSTM layer to make the result present the desired dimension.

Like other neural networks, stacking hidden layers can make the model deeper and get accurate output. The LSTM model composed of multiple LSTM layers is called a stacked LSTM model, and its model structure is shown in Figure 2.
3. Experimental Evaluation

3.1. Basic Configurations

In our experiment, the well-known KDD99 dataset is used for evaluation. The experiments are running under Python3.5 and TensorFlow1.14 platform on a server with I7-9700K CPU, GeForce 2080Ti GPU, and 24G RAM. For comparison, ALL-AGL, MHCVF, and HAST-II are employed to evaluate our proposed stacked-LSTM method.

In our experiments, the pros and cons of intrusion detection systems are determined by three indicators, i.e., detection rate (DR), false positive rate (FPR), and accuracy (ACC). These evaluation indicators can be obtained from the confusion matrix of the model. The confusion matrix refers to the table obtained by filling in different positions in the table according to the actual classification and prediction classification of different items in the classification model. The final statistics are obtained.

Taking the two-class model as an example, the confusion matrix is as follows:

| Ground truth | Result of detection |
|--------------|---------------------|
| True         | TP                  | FN                  |
| False        | FP                  | TN                  |

In this way, the DR, FPR and ACC are calculated as

\[
DR = \frac{TP}{TP + FN} \tag{4}
\]

\[
FPR = \frac{FP}{FP + TN} \tag{5}
\]
Among the three evaluation indicators, ACC is used to evaluate the overall detection ability of the detection system, DR is employed to measure the model's ability to detect abnormal events and FPR is used to evaluate the detection system's ability to avoid misjudgment of normal conditions.

3.2. Performance Evaluation

To ensure the fairness of the evaluations, we used the same operating environment to test all algorithms and used the same training sample sampling rate. TABLE 2 shows the results of the ACC, DR, and FPR comparison of all methods under the KDD99 dataset. All experiments are rerun 10 times and the average value is obtained to list.

**Table 2. The Comparison of ACC, DR and FPR of Different Algorithms on the Dataset KDD99.**

| model    | ACC    | DR     | FPR   |
|----------|--------|--------|-------|
| ALL-AGL  | 91.76% | 85.31% | 1.08% |
| MHCVF    | 96.45% | 58.77% | 0.32% |
| HAST-II  | 95.59% | 93.26% | 0.17% |
| Stacked-LSTM | 97.10% | 96.73% | 0.08% |

Although the ACC and DR of ALL-AGL are both above 85%, the false alarm rate still reaches 1%. The FPR of MHCVF reaches a satisfactory level, but the detection rate is only 58.77%, which is difficult to meet the daily intrusion detection. HAST-II is currently an excellent algorithm for intrusion detection, and all indicators have reached satisfactory results. The Stacked-LSTM mentioned in this article employs the superiority of deep learning method to fully learn the hidden features in network traffic data, and reaches the highest level in each indicator.

At the same time, we also compared the time complexity of the four algorithms. As shown in Figure 3, our time complexity is less than the other three comparison algorithms.

![Figure 3. Comparison of test time required for different algorithms.](image)

4. Conclusions and Future Work

In this study, aiming to solve the problem of data intrusion detection in power middle platform systems, we proposed a stacked-LSTM model based on LSTM neural network. Firstly, the data feature is extracted through LSTM, and then the inner difference sequence is modeled for distribution, and a more appropriate anomaly score is dynamically assigned to each data to lift the accuracy of data intrusion detection. The KDD99 dataset is used to verify the performance of our proposed method.
Furthermore, if an intrusion attack against LSTM occurs, the proposed algorithm may fail. Therefore, in future work, data attacks based on LSTM will be the focus of our research.

Acknowledgements
This research was funded by the Science and Technology Project of State Grid Co., LTD (Research and Application of Service Design and Management Technology based on Data Middle Platform, 5700-202018181A-0-0-00).

References
[1] Tang X, Yao H, Sun Y, et al. Joint Modeling of Local and Global Temporal Dynamics for Multivariate Time Series Forecasting with Missing Values.//AAAI. 2020: 5956-5963.
[2] Sheng-Ju S. Implementation of cyclic redundancy check in data communication[C]//2015 International Conference on Computational Intelligence and Communication Networks (CICN). IEEE, 2015: 529-531.
[3] Bianco V, Manca O, Nardini S. Linear regression models to forecast electricity consumption in Italy[J]. Energy Sources, Part B: Economics, Planning, and Policy, 8(1): 86-93, 2013.
[4] Liu X, Nielsen P S. Regression-based online anomaly detection for smart grid data[J]. arXiv preprint arXiv:1606.05781, 2016.
[5] Messinis G, Dimeas A, Rogkakos V, et al. Utilizing smart meter data for electricity fraud detection[C]//SEERC Power Conf., Portoroz, Slovenia. 2016: 7-8.
[6] Zhou Q, Yu K, Chen X, et al. Calculation Method of the Line Loss Rate in Low-voltage Transformer District Based on PCA and K-Means Clustering and Support Vector Machine[C]//2018 International Conference on Power System Technology (POWERCON). IEEE, 2018: 4264-4271.
[7] Shi B, Zhang J, He Y, et al. Electricity consumption behavior analysis of single-phase power consumers in distribution network based on dynamic game theory[J]. Automation of Electric Power Systems, 41(14): 87-91, 2017.
[8] Adil M, Javaid N, Qasim U, et al. LSTM and Bat-Based RUSBoost Approach for Electricity Theft Detection[J]. Applied Sciences, 10(12): 4378, 2020.