CAE-based classification method of electric power business

Ruiheng Ma*, Min Xiang, Rujie Lei and Chunmei Huang

Key Laboratory of Industrial Internet of Things and Networked Control, Chongqing University of Posts and Telecommunications, Chongqing, Chongqing, 400065, China
*Corresponding author’s e-mail: S180301014@stu.cqupt.edu.cn

Abstract. Aiming at the problem that abnormal data in the business classification of power data communication networks will reduce the business classification accuracy, a classification method based on convolutional autoencoder combined with lightGBM was proposed. First, the flow features of the electric power business were extracted through the convolutional autoencoder. Then, anomaly detection was realized according to the relationship between the reconstruction loss of the convolutional autoencoder and the set threshold. Finally, the electric power business was classified through the LightGBM classifier. The Moore data set used for simulation verification. The results show that the proposed method can effectively detect abnormalities, thereby improving the accuracy of the electric power business classification.

1. Introduction

With the rapid advancement of the country's "strong and smart grid" construction, the coupling between the power system and information technology has further improved [1], the number of power system nodes has increased, and network intrusions have become increasingly complicated. The electric power communication network puts forward higher requirements for security, real-time, and resource scheduling capabilities [2]. Through the identification and classification of the electric power business, the refined abnormal monitoring of the power communication network and differentiated QoS elastic guarantee are realized, thereby optimizing the carrying capacity of the network. Machine learning was used for feature extraction in traffic classification in [3-4]. CNN (Convolutional Neural Networks) was combined for traffic classification in [5-7]. The electric power business classification methods proposed by those researchers lack abnormal data analysis and cannot detect abnormal traffic.

An electric power business classification method based on Convolutional Auto-Encode (CAE) was proposed. This method detected abnormal data through reconstruction loss and then combined the LightGBM to achieve electric power business classification.

2. Anomaly detection and classification based on CAE and LightGBM

The particularity of the power industry required that the business classification method applied to the power information communication network must meet the accuracy of business identification greater than 92%, and the data content not to be interpreted [8]. The statistical flow characteristics of the electric power business were selected as the model input. A reasonable and adequate model combining the CAE with LightGBM was constructed to improve business classification accuracy. The overall architecture was shown in Figure 1.
The AE was a three-layer neural network structure, including an input layer, a hidden layer, and an output layer. The structure was shown in Figure 1, where \( n \) was the input data dimension, and \( m \) was the data dimension after feature extraction. The encoding process was expressed as equation (1), and the decoding process was expressed as equation (2).

\[
\begin{align*}
    z &= f(x) = s_f(\omega x + p) \\
    y &= g(z) = s_g(\omega_f z + q)
\end{align*}
\]

where \( \omega \) and \( p \) represented encoding weights and offsets, \( \omega_f \) and \( q \) represented decoding weights and offsets. The AE used the backpropagation to continuously adjust the model parameters until the output was equal to the input, thereby learning advanced abstract features. The similarity of the input and output was expressed as equation (3). The gradient descent algorithm was used to minimize the loss function of the AE, then the parameters could be solved.

\[
J_{AE}(\omega, p, q) = \sum (L(x, y)) = \sum \|x - y\|^2
\]

According to the classification experiment using CNN and its improved algorithm in [5-7], the results shown that CNN could effectively extract network traffic data characteristics. A 15-layer CAE combining CNN and autoencoder was designed, including the input layer, four convolutional layers and three pooling layers of the encoding network, three deconvolution layers and three de-pooling layers of the decoding network, and output layer. The CAE used a convolutional layer and a pooling layer to replace the fully connected layer. It can better mine the information between the feature data. The structure was shown in Figure 2 and Table 1.

![Figure 1. Overall architecture.](image)

![Figure 2. The 15-layer CAE structure](image)

| Layer | \( C_1 \) | \( S_1 \) | \( C_2 \) | \( S_2 \) | \( C_3 \) | \( S_3 \) | \( F \) |
|-------|----------|----------|----------|----------|----------|----------|------|
| Kernel size | \( 5*5*64 \) | \( 2*2 \) | \( 3*3*32 \) | \( 2*2 \) | \( 3*3*16 \) | \( 2*2 \) | \( 3*3*32 \) |
| Output | \( 16*16*64 \) | \( 8*8*64 \) | \( 8*8*32 \) | \( 4*4*32 \) | \( 4*4*16 \) | \( 2*2*16 \) | \( 1*1*32 \) |

| Layer | \( D_1 \) | \( U_1 \) | \( D_2 \) | \( U_2 \) | \( D_3 \) | \( U_3 \) | Output |
|-------|----------|----------|----------|----------|----------|----------|------|
| Kernel size | \( 2*2*16 \) | \( 2*2 \) | \( 3*3*32 \) | \( 2*2 \) | \( 5*5*64 \) | \( 2*2 \) | \( 5*5*1 \) |
| Output | \( 2*2*16 \) | \( 4*4*16 \) | \( 4*4*32 \) | \( 8*8*32 \) | \( 8*8*64 \) | \( 16*16*64 \) | \( 16*16*1 \) |

Research on traffic classification based on flow statistical characteristics [9], Moore et al. used 248 characteristics to describe the traffic in detail. Therefore, a 16*16 matrix was constructed as the input of...
the model. Since the feature dimension was less than the number of matrix elements, the zero padding was performed at the end of the matrix. Besides, to avoid the loss of the edge information of the input layer too fast and to keep the matrix size before and after the convolution consistent, the boundary of the current layer matrix was filled with 0 to save the boundary information, and then pooling was used to reduce the data dimension. Considering the high requirements for real-time and accuracy of power business classification, too few convolution times will lead to insufficient feature extraction, and too many convolution times will cause the gradient to disappear. Therefore, a 15-layer CAE model was selected and the effectiveness of the model was verified through experiments. Also, when the number of network layers was fixed, the coding dimension greatly influenced the effect of model feature extraction and anomaly detection. When choosing coding dimensions, the dimensional features of the input data and the validity of the extracted features should be considered. Considering comprehensively, 32 was selected as the coding dimension of the network.

The anomaly detection method based on CAE only used normal data to train the CAE. Then draw box plots based on the reconstruction loss of normal and abnormal data on the test data set. Generate a sequence in the interval between the 75th quantile of normal data reconstruction loss and the 25th quantile of abnormal data reconstruction loss with a step length of 0.05, and test the anomaly detection accuracy rate on the test set. Take $\delta$ when the detection accuracy was the highest as the threshold. Finally, input the verification set data into the network and use equation (3) to calculate its reconstruction loss $L_v$. If $L_v > \delta$ was abnormal data, otherwise, it was normal. After anomaly detection, those data that are judged to be normal were sent to the LightGBM classifier for classification.

3. Simulation results and analysis
The computer used in this experiment has 16GB of running memory, 3.4GHZ main frequency, GTX1070TI graphics card, and Tensorflow1.14 deep learning framework under the Windows operating system.

Since the electric power communication network can be regarded as a hybrid of the electric power system and communication network, some communication scenarios in the current computer network can be applied to the power communication network [10]. Therefore, the Moore data set based on the flow statistical feature analysis method was used to verify the proposed method.

It is necessary to eliminate the dimensional relationship between the data to make the data comparable. The Z-score normalization method was used to normalize each feature in the data set. For the results, three evaluation metrics were used: accuracy (A), trust (T), overall accuracy (OA), which are defined as:

$$A = \frac{TP \times (TP + FN)^{-1}}{100\%}$$  \hspace{1cm} (4)

$$T = \frac{TP \times (TP + FP)^{-1}}{100\%}$$  \hspace{1cm} (5)

$$OA = \frac{(TP + TN) \times (TP + FN + FP + TN)^{-1}}{100\%}$$  \hspace{1cm} (6)

TP was the number of samples correctly classified as type X, FN was the number of samples misjudged as Not-X, FP was the number of samples incorrectly classified as type X, TN was the number of samples correctly classified as Not-X.

To verify the effectiveness of the 15-layer CAE, the anomaly detection effects of three different structures were compared. The results were shown in Table 2.

| Number of layers | OA (%) | Positive (%) | Negative (%) | Parameters |
|------------------|--------|--------------|--------------|------------|
| 11               | 94.19  | 93.18        | 97.97        | 96,113     |
| 15               | 95.47  | 94.98        | 97.27        | 86,401     |
| 19               | 93.21  | 92.02        | 97.66        | 85,769     |

Table 2 shown that the 11-layer CAE had a large number of parameters due to fewer convolution times and too fast dimensionality reduction. The 19-layer CAE increased the number of convolutions, which reduced the number of parameters, but the structure became more complicated. The 15-layer CAE
had the highest accuracy, a moderate amount of parameters, and a relatively simple structure, which can meet the electric power business classification requirements. Table 3 shown the accuracy of anomaly detection using different coding dimensions under the 15-layer CAE.

Table 3. The anomaly detection rate of different coding dimensions.

| Dimension | OA (%) | Positive (%) | Negative (%) | Parameters |
|-----------|--------|--------------|--------------|------------|
| 8         | 87.58  | 84.52        | 99.01        | 83,305     |
| 16        | 94.03  | 93.34        | 96.62        | 84,337     |
| 32        | 95.43  | 94.90        | 97.42        | 86,401     |
| 64        | 93.43  | 92.04        | 98.64        | 90,529     |
| 128       | 94.43  | 93.67        | 97.29        | 98,785     |

Table 3 shown that as the coding dimension increases, the overall accuracy reaches the highest when the coding dimension is 32 dimensions, and the amount of network parameters is in the middle. To verify the anomaly detection effect of this method, the BP auto-encoding anomaly detection method, the anomaly detection method based on random forest and information entropy, the KNN clustering anomaly detection method, and the One-class-SVM anomaly detection method were compared. The results were shown in Table 4.

Table 4. The anomaly detection rate of different methods.

| Number | Method                        | OA (%) | Positive (%) | Negative (%) |
|--------|-------------------------------|--------|--------------|--------------|
| 1      | 15-layer CAE                  | 95.47  | 94.89        | 97.27        |
| 2      | BP AE                         | 92.52  | 92.22        | 93.37        |
| 3      | Random forest + entropy       | 88.68  | 98.34        | 52.83        |
| 4      | KNN                           | 91.60  | 90.11        | 97.54        |
| 5      | One-class-SVM                 | 89.90  | 90.66        | 87.25        |

Table 4 shown that the 15-layer CAE has the highest overall accuracy of anomaly detection. Although the positive class accuracy of Method 3 was higher than that of the 15-layer CAE, the negative class accuracy and the overall accuracy of Method 3 were lower. In summary, the anomaly detection method proposed was better than other methods. Considering the accuracy and real-time requirements of the electric power business classification, the classification accuracy of the five classification methods were compared. The results were shown in Table 5.

Table 5. The overall accuracy of different classification methods.

| Method                      | LightGBM | SVM | Random Forests | Logistic Regression | GBDT |
|-----------------------------|----------|-----|----------------|---------------------|------|
| OA (%)                      | 99.79    | 99.64 | 99.55          | 99.11               | 99.08|

Table 5 shown that the LightGBM method has the highest overall accuracy. And the calculation speed of LightGBM was nearly ten times faster than GBDT. To illustrate the effect of the method proposed on business classification when abnormal data appears, the other two methods based on logistic regression and information entropy and based on random forest and information entropy were compared. The results were shown in Table 6.

Table 6. OA of classification and OA of anomaly detection using different methods.

| Method                        | OA of classification (%) | OA of anomaly detection (%) |
|-------------------------------|--------------------------|-----------------------------|
| 15-layer CAE + LightGBM       | 95.27                    | 97.42                       |
| Logistic regression + entropy | 81.69                    | 19.90                       |
| Random forest + entropy       | 85.50                    | 37.41                       |

Table 6 shown that the overall accuracy of classification and abnormal detection of this method are higher than other methods. Table 7 shown the credibility and accuracy of each type.

Table 7. Class accuracy and class credibility using different methods.

| Type                          | 15-layer CAE + LightGBM | Logistic regression + entropy | Random forest + entropy |
|-------------------------------|--------------------------|-------------------------------|-------------------------|
| T (%)                         | 99.98                    | 99.98                         | 99.98                   |
| A (%)                         | 99.92                    | 88.89                         | 95.43                   |
Table 7 shown that the method proposed can maintain a high level of reliability and recognition accuracy of each type. Although the credibility of Type 6 was low, it still had significant advantages over the other two methods. The average test time was 0.26ms, which could meet the speed requirements of the electric power business classification. Also, Figure 3 shown that the data set used in the experiment had an uneven distribution of the sample number. However, this method could still accurately classify each type of sample and adapt to scenarios where business data was unevenly distributed in real network operation.

![Figure 3. Comparison of classification results of three methods.](image)

### 4. Summary

This article proposed an electric power business classification method based on CAE. The Moore dataset was used for simulation verification. The results show that this method can accurately detect abnormal data and effectively improve the existing business classification accuracy. Moreover, it had also achieved good classification results in categories with small sample sizes.

### References

[1] LIU T, TIAN J, WANG J.Z, et al. (2019) Integrated Security Threats and Defense of Cyber-physical Systems. Acta Automatica Sinica, 45 (1): 5-24.

[2] QI B, YE X, LI B, et al. (2020) Differential Protection Strategy of Link Fault for Distinguishing Protection Quality of Power Business. Automation of Electric Power Systems, 44(11):181-188.

[3] WEI Z K, XIA J B, et al. (2016) Research on Multi-feature Extraction and Classification Based on Random Forest. Transducer and Microsystem Technologies. 35(12):55-59.

[4] KE W L., Wang Y, Lei X, et al. (2017) Spark-Based Feature Selection Algorithm of Network Traffic Classification. In: International Conference on Computational Intelligence & Security. Hong Kong, pp. 140-144.

[5] WANG Y, ZHOU H.Y, et al. (2018) Network traffic classification method basing on CNN. Journal on Communications, 39(1): 14-23.

[6] Wang Y., An J, Huang W. (2018) Using CNN-based Representation Learning Method for Malicious Traffic Identification. In: 2018 IEEE/ACIS 17th International Conference on Computer and Information Science. Singapore. pp. 400-404.

[7] WANG W, ZHU Ming, et al. (2017) Malware traffic classification using convolutional neural network for representation learning. In: 2017 International Conference on Information Networking. Da Nang. pp. 712-717.
[8] Jiang Y, Hua J, Hu J, et al. (2015) Study on Traffic Identification Technologies of Electricity Grid Information and Communication Network, Informatization Research, 10-14, 18.

[9] A. W. Moore, D. Zuev, and M. Crogan. Discriminators for use in flow-based classification. Technical Report RR-05-13, Department of Computer Science, Queen Mary, University of London, September 2005.

[10] ZHANG Y, Wang L, Sun W, et al. (2011) Distributed Intrusion Detection System in a Multi-Layer Network Architecture of Smart Grids. In IEEE Transactions on Smart Grid. 796-808.