Research on Electricity Information Acquisition System Based on Sample Data Mining Model

Dang zhongkui1, *, Fu Lei1
State Grid NingXia Information & Telecommunication Company, GuYuan 756000, China
*E-mail: dzk2005@foxmail.com

Abstract: In power user information collection and detection, power companies generally have a variety of different detection needs, or need to solve the problem while having additional requirements for certain aspects. Therefore, the SVM classification technology is used in the paper to carry out more detailed pattern recognition of power consumption characteristics for small-scale users or users with major suspicions. Moreover, given the imbalance of the abnormal electricity detection data set, a comprehensive processing model of unbalanced samples is constructed. Meanwhile, the differential evolution algorithm is applied to complete the SVM parameter optimization, which not only solves the problem that the SVM classification performance is significantly affected by the parameters, but also ensures the operating efficiency of the integrated classification model.

1. Introduction
The electricity development should be in line with the national economic development level and the principle of appropriate leadership. However, current power supply still cannot fully meet people’s needs, and some regions still have problems such as weak power grids and insufficient power supply. Therefore, improving the construction of power facilities and promoting the orderly development of the power industry are of great significance to the growth of the entire national economy[1-2].

Since the power user information data can better present the characteristics of users' power consumption behavior within the time period, the user's electricity behavior is effectively modeled through the power information data in the paper. What is more, the SVM is used to identify and deal with the user's electricity behavior pattern, so that the abnormal electricity consumption can be effectively identified.

2. Information Collection Based on Sample Data Sampling

2.1 Sample Feature Vector
The user's electricity load curve data is represented as a vector in the user's electricity feature space, and the feature vector $x^m$ is used to represent the user's electricity usage behavior pattern within the time period, based on which the identification of abnormal power users is completed. What is more, considering the time difference among users accessing the electricity information collection system, some users cannot obtain the long-term historical electricity data. Therefore, the user's historical electricity consumption data in a whole year is selected, and daily average load data in every two weeks is adopted as an element. A total of 26 element feature vectors are applied to represent the user's
electricity consumption behavior characteristics. The feature vector of the user's electricity consumption behavior pattern is $x_m = \left( x_n^m, n = 1, 2, \cdots, n \right)$, where H=26 represents the number of daily average electricity load elements among every two weeks in a whole year[3-4].

In the entire SVM- based abnormal electricity detection system, the normal sample acquisition method is relatively simple, while the abnormal sample acquisition cost is relatively high. Therefore, the different types of users have serious class imbalances[5-6].

2.2 Power Data Collection

Through the main station of the electricity information collection system, the historical electricity load information of all the users to be tested in a whole year is obtained, and a certain number of training samples including normal electricity samples and abnormal electricity samples are gained from the detection system library. In addition, the electricity sample is applied to construct the abnormal electricity detection model, as shown in Figure 1.

```
Input layer

Data preprocessing
User feature extraction
Anomaly data detection model
User characteristic data input
Abnormal behavior analysis

Analysis layer

Output layer

Suspect user verification
Suspected user abnormal behavior

```

The collection and acquisition of the user's electricity data to be detected are completed through the existing electricity information collection system, and then the original electricity data is pre-processed to eliminate the impact of incomplete data or attributes on the later abnormal detection. Meanwhile, the behavioral features of users' electricity consumption are firstly extracted from the processed data, which makes the detection more vivid. Secondly, by analyzing the user's electricity consumption behavior, a reasonable abnormality detection model is constructed for the user. Thirdly, the extracted electricity characteristic data is input to the detection model, which will be analyzed and judged by the detection system. Finally, the test results will be output and verified on site[7-8].

2.3 Data Pre-processing

The actual collected data is limited by the current operating status of the existing collection system. Generally, there will be problems such as missing data and wrong collection. Therefore, it is necessary to conduct reasonable detection on the collected data to correct inconsistent data and wrong data. Meanwhile, the missing values are filled, and the noise is removed so as to better smooth the user load curve and reduce the interference with the extraction of the later power mode. In addition, the user's electricity load curve data is used as a vector in the user's electricity feature space, and the feature vector $x_m^m$ is adopted to represent the user's electricity consumption behavior pattern.

The standardization process is to eliminate the differences among users' different electricity consumption scales, and the input feature vectors of the SVM detection model are normalized. Additionally, the standardization method is a kind of min-max normalization, and the interval after normalization is [0, 1]. The formula used here is as follows[9-10].

$$nx = \frac{x - \min x}{\max x - \min x}$$  \hspace{1cm} (1)
In Equation 1, X refers to the load data to be processed in the original input feature vector, and nx is the processed load data. Besides, $\max_x$ and $\min_x$ respectively represent the maximum and minimum load data in the current electricity feature vector.

3. Experimental Verification Analysis
The experimental design is as follows. A fixed number 520 of normal electricity samples are selected from the total electricity sample set, and abnormal electricity samples are selected according to different ratios of 1:1, 3:1, 5:1 and 10:1. All the selected samples are used to form a verification sample set. Moreover, since the training data and the verification data are not clearly divided, in order to reasonably verify the classification effect of several algorithms, a 50% cross-validation test method is used here, and the corresponding results obtained each time are recorded. Therefore, the arithmetic average is taken as the final algorithm performance evaluation. What is more, the detection performance of SVM on unbalanced samples after optimization by various algorithms is tested from the two aspects of minority F1 and G values, and the parameter optimization time is increased as the evaluation index. Additionally, in order to reduce the interference of accidental factors on the algorithm, the algorithm results are the average of 30 runs. Besides, the parameters of several group intelligent optimization algorithms are set as follows. The population size is N=40, and the maximum evolutionary generation is T=200. Additionally, the termination threshold is set to 0.0001. What is more, the SVM parameter is C, and the range of g is [2^{-10}, 2^{10}]. The specific test results are shown in Table 1-3.

Table 1 Time spent on parameter optimization under different sample ratios (unit: s)

| Sample ratio | GA-SVM | PSO-SVM | DE-SVM |
|--------------|--------|---------|--------|
| 1:1          | 821.11 | 918.04  | 467.69 |
| 3:1          | 446.24 | 495.95  | 246.81 |
| 5:1          | 308.75 | 383.26  | 173.71 |
| 10:1         | 256.65 | 289.37  | 117.80 |

It can be seen from Table 1 that the time taken to optimize the parameters with the DE algorithm is significantly less than that of the PSO algorithm and the GA algorithm, which is only about 1/2 of the time spent by the latter two. Moreover, the differential evolution algorithm used in the paper can effectively improve the overall recognition accuracy of the abnormal power detection model based on SVM, and reduce the time spent on SVM parameter optimization. Meanwhile, the SMOTE+Bagging comprehensive processing model constructed for the problem of unbalanced samples can significantly improve the detection accuracy of a small number of abnormal users, which is of great value for improving the effect of abnormal power detection based on SVM.

Table 2 F values at different sample ratios

| Sample ratio | GA-SVM | PSO-SVM | DE-SVM |
|--------------|--------|---------|--------|
| 1:1          | 0.7566 | 0.7550  | 0.7603 |
| 3:1          | 0.7202 | 0.7003  | 0.7131 |
| 5:1          | 0.6782 | 0.6802  | 0.6817 |
| 10:1         | 0.4615 | 0.4706  | 0.4919 |

Table 3 G values at different sample ratios

| Sample ratio | GA-SVM | PSO-SVM | DE-SVM |
|--------------|--------|---------|--------|
| 1:1          | 0.8765 | 0.8742  | 0.8815 |
| 3:1          | 0.8448 | 0.8385  | 0.8482 |
| 5:1          | 0.7911 | 0.7983  | 0.8001 |
| 10:1         | 0.6087 | 0.5993  | 0.6144 |
Several algorithms have been used to optimize the SVM parameters, and good results have been achieved as well. In the case of balanced sample detection, the F value of the upgraded minority reaches a maximum of 0.76, which shows that seeking suitable optimization parameters in practical applications is of great significance for improving the performance of abnormal power detection based on SVM. Meanwhile, compared with genetic algorithm and particle swarm optimization, it is found that the comprehensive detection effect of optimizing SVM model based on differential evolution algorithm has been improved to a certain extent, which is not only reflected in the detection accuracy of abnormal users, but also in the efficiency of parameter optimization.

4. Conclusion
In order to realize more accurate electricity theft detection for small-scale users, a supervised abnormal electricity consumption detection scheme based on SVM is constructed by analyzing the characteristics of electricity consumption of users, which is verified with the measured data, so that the effectiveness of the method can be proved.

(1) Due to the serious sample imbalance problem in abnormal electricity detection, a sample combination processing model is constructed in the paper. What is more, through the classification performance test in the case of the different proportions of normal and abnormal electricity samples, it is proved that the processing model which is constructed and integrated can effectively improve the data imbalance problem of the training samples.

(2) As for the problem of SVM parameter selection, differential evolution algorithm is used to optimize SVM parameter selection. Moreover, by comparing with different parameter selection algorithms in different proportions of positive and negative samples, it is showed that the differential evolution algorithm can effectively complete the selection of high-quality parameters. What is more, compared with other parameter optimization algorithms, the differential evolution algorithm takes less time, which effectively guarantees the running efficiency of the Bagging integrated classification based on SVM.

References
[1] H.W. Hui, C.C. Zhou, S.G. Xu, F.H. Lin, A Novel Secure Data Transmission Scheme in Industrial Internet of Things, China Communications, vol. 17, no. 1, 2020, pp: 73-88.
[2] F.H. Lin, Y.T. Zhou, X.S. An, I.You, K.R. Choo, Fair Resource Allocation in an Intrusion-Detection System for Edge Computing: Ensuring the Security of Internet of Things Devices, in IEEE Consumer Electronics Magazine, vol. 7, no. 6, 2018, pp: 45-50. doi: 10.1109/MCE.2018.2851723.
[3] J.T. Su, F.H. Lin, X.W. Zhou, X. Lv, Steiner tree based optimal resource caching scheme in fog computing, China Communications, vol. 12, no.8, 2015, pp: 161-168.
[4] Xie Fei, Shao Shuyan, Peng Jie. Analysis of electric power wireless video monitoring system [J]. Shandong Electric Power Technology. 2017(04)
[5] Xu Xiaofei, Chen Liang. LZ77 power data compression algorithm using integer wavelet transform [J]. Journal of Xi’an University of Technology. 2018(03)
[6] Wang Jian. Discuss the application of data compression in remote communication of electricity information collection[J]. Communication World. 2017(13)
[7] Yan Haizhou, Xu Bugong, Li Haibin, et al. Application of wavelet denoising and compression in power plant equipment [J]. Electric Measurement and Instrumentation. 2017(11)
[8] Guo Zhihua, Wang Zuozheng, Li Na, et al. Application and Practice of Multiple Communication Methods in Qinghai Power Grid Electricity Information Collection System[J]. Qinghai Electric Power. 2016(04)
[9] Chen Jinmu, Zhang Ke, Zhou Chongcheng, et al. Research on the verification and promotion strategy of the relationship between households in Taiwan based on the collection system[J]. Instrumentation and Analysis Monitoring. 2019(01)
[10] Wu Chaofei, Ma Hao, Li Peng, et al. The influence of electromagnetic disturbance on communication performance in electricity information collection [J]. Hebei Electric Power Technology. 2019(01)