Research on Intelligent Analysis Technology of Network Security Risk Based on Big Data

Yibing Zhao¹, Juling Zhang²*, Siyu Xiang², Yong Tang²

¹State Grid SiChuan Electric Power Company, ChenDu, SiChuan, 610000, China
²Electric Power Research Institute of State Grid Sichuan Electric Power Company, ChenDu, SiChuan, 610000, China

*Corresponding author’s e-mail: Zhangjl076x@sc.sgcc.com.cn

Abstract. Aiming at the severity and distribution of abnormal behaviour, attack behaviour, compliance, and vulnerability of equipment in the power information system, this article applies big data technology to analyse and evaluate abnormal behaviour in the network, and uses machine learning to improve the network, the ability to identify security risks and the accuracy of assessments provide strong guarantees and technical support for improving the security and reliability of power information systems.

1. Introduction
With the advent of the information age, power information systems and network security are developing rapidly. Massive amounts of information are appearing in power systems and networks. Now we are in the era of big data, with the improvement of computing power and the further development of machine learning algorithms, we can use big data technology to process massive amounts of information and data, which also promotes the development of various industries. At the same time, with the flow of massive data in the power system, it has brought more unstable factors to the network security in the power information system. The existence of a large amount of data will provide certain convenience for hackers to invade, and it may also cause network security operators to operate errors and so on.

Machine learning algorithms are widely used in speech, computer vision, natural language processing, knowledge graphs, finance and other directions, empowering all walks of life. Machine learning algorithms include linear regression, logistic regression, support vector machines, decision trees. Therefore, machine learning algorithms are used to extract the characteristics of massive data and networks, and then use appropriate algorithms to further process these features. Finally, we can use the model to predict and evaluate the abnormal behavior and attack behavior of equipment in network security, and can detect bad behaviour in the network in time and deal with it immediately. It can realize the intelligent analysis of network security risks, accelerate the pace of network security construction in our country's power information system, and build a networked, secure and intelligent power information system.

2. Related research
With the development of network security, more and more methods are applied in the process of network security analysis. Domestic and foreign scholars have analysed network security and put forward some different views. Foreign scholars use restricted Boltzmann machines to extract features
of the original network traffic, and use deep belief networks to classify features, which have also achieved certain results.

Xie Peng, Li Jianhua and others [5] used the Bayesian network analysis method to analyse network security. This method has achieved good results, but it lacks the calculation of the posterior probability of abnormal events.

Chen Xiaojun et al. [7] derived the probability of occurrence of the event through an abnormal event. This method solves the problem that Xie Peng et al. did not calculate the posterior probability of an abnormal event, but this method ignores the posterior probability of the node associated with the event. Impact.

Ma Chunguang et al. [9] used Bayesian methods to identify and evaluate the network, but as the number of nodes in the network increases, the computational complexity increases exponentially, and timely identification and evaluation may not be possible.

Based on the above analysis, this paper uses the support vector machine algorithm in machine learning to conduct intelligent analysis and research on network security risks.

3. SVM model based on machine learning

3.1. Comparison of algorithms

KNN can be said to be one of the simplest classification algorithms. At the same time, it is also one of the most commonly used classification algorithms. The full name of KNN is K Nearest Neighbours. KNN performs classification by measuring the distance between different feature values. The idea of the KNN algorithm is very simple: for any n-dimensional input vector, it corresponds to a point in the feature space, and the output is the category label or predicted value corresponding to the feature vector.

The KNN algorithm is a very special machine learning algorithm because it has no learning process in the general sense. Its working principle is to use training data to divide the feature vector space and use the result of the division as the final algorithm model. There is a sample data set, also called a training sample set, and each data in the sample set has a label, that is, we know the correspondence between each data in the sample set and its classification. After inputting unlabelled data, compare each feature of the unlabelled data with the corresponding feature of the data in the sample set, and then extract the classification label of the data with the closest feature in the sample (nearest neighbour). Generally speaking, we only select the first k most similar data in the sample data set, which is the origin of K in the KNN algorithm, and finally select the category with the most occurrences of the k most similar data as the classification of the new data.

But the disadvantage of the KNN algorithm is that it is not sensitive enough to special points, and in the power system, every abnormal point will cause a huge threat to the system, so the KNN algorithm is not suitable in the power information system.

The decision tree model is a tree structure. The decision tree is composed of nodes and directed edges. There are two types of nodes: internal nodes and leaf nodes. Internal nodes represent a feature or attribute, and leaf nodes represent a class. Decision tree (Decision Tree), also known as decision tree, is a predictive analysis model expressed in the form of a tree structure (including binary trees and polytrees). The decision tree algorithm is to classify each situation from the root node to the leaf node. The leaf node is the classification of each situation,

The shortcoming of the decision tree algorithm is that it is easy to cause over-fitting. If there is over-fitting in the power grid information, it may lead to inaccurate judgments of abnormal conditions in the power grid information system, leading to system failure.

3.2. Selection of SVM model

The detection and evaluation of abnormal behaviours in the power information system can actually be regarded as a classification process and a supervised learning process. Therefore, the SVM support vector machine algorithm in machine learning can be used to classify abnormal behaviours, and the
The relationship between abnormal behavior categories and attack behaviors can be found through classification. The modeling process using support vector machines is shown in Figure 1.

Support vector machine generalization and learning capabilities are stronger than other machine learning algorithms, and it is suitable for solving global optimal solutions. It is widely used in the process of network security risk identification and assessment. Different hyperplanes and classifications can be obtained according to different kernel functions. Function, and is suitable for small sample data sets.

The basic idea of using SVM learning is to solve the separation hyperplane that can correctly divide the training data set and have the largest geometric interval. As shown in Figure 1 below, it is the separating hyperplane (linearly separable).

Because the abnormal behavior in the power system is affected by many factors, we use the nonlinear support vector machine algorithm. The processing process of the nonlinear support vector machine algorithm is generally as follows:

Input: training data set

\[ T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \] (1)

among them

\[ x_i \in \mathbb{R}^p, y_i \in \{+1, -1\}, i = 1, 2, \ldots, N \] (2)

Output: separation hyperplane and classification decision function

(1) Select an appropriate kernel function \( K(x, z) \) and penalty parameter \( C > 0 \), construct and solve the convex quadratic programming problem

\[ \min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i \] (3)

\[ \text{s.t. } \sum_{i=1}^{N} \alpha_i y_i = 0 \] (4)

\[ 0 \leq \alpha_i \leq C, i = 1, 2, \ldots, N \] (5)

Get the optimal solution:
\[ \alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*)^T \quad (6) \]

(2) Calculation chooses a portion of \( a^i_*, a^j_* \) to meet the conditions \( 0 < a^j_* < C \),

\[ b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x_j) \quad (7) \]

(3) Classification decision function:

\[ f(x) = \text{sign} \left( \sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^* \right) \quad (8) \]

Commonly used kernel function, Gaussian kernel function:

\[ K(x, z) = \exp \left( -\frac{|x-z|^2}{2\sigma^2} \right) \quad (9) \]

The corresponding SVM is a Gaussian radial basis function classifier. In this case, the classification decision function is:

\[ f(x) = \text{sign} \left( \sum_{i=1}^{N} \alpha_i^* y_i \exp \left( -\frac{|x-x_i|^2}{2\sigma^2} \right) + b^* \right) \quad (10) \]

4. Model verification and analysis

4.1. Data set processing

The experimental data comes from the network information captured in the power system. There are a total of 5000 network data information. Take 3000 of them as a training set for training to build a model, and 2000 pieces of data as a test set to test and evaluate the model, and judge the model's evaluation of abnormal behaviour of the power network.

Since the data provided in the power system may have missing and abnormal situations, we use feature engineering methods to process the data to make the data set better adapt to the SVM algorithm and train a better model. Data pre-processing generally includes four processes, namely data cleaning, data integration, data transformation, and data specification.

Data cleaning is generally to process missing values, generally using the mean method, median, and mode to fill in. Here we use the mean method for processing, and for outliers we use deletion or mean value replacement for correction. Data integration is mainly to merge data. For example, the data is basically the same. If the data is not merged, it may cause data redundancy and increase the training time of the model. Data transformation can be based on specific data features such as square, open, and logarithmic. The data specification is mainly to normalize the data. For example, the data range provided in the power information system is too large. We can use normalization to reduce the data to a smaller range. On the one hand, the model training time can be reduced. Improve model generalization performance.

4.2. Evaluation index

The evaluation indicators used in the thesis are precision rate and recall rate. The accuracy rate is the ratio of the prediction accuracy of all samples to the total number of samples. Recall rate is the ratio of positive samples being predicted to find all positive samples correctly. Since the power information system data used in this paper is multi-category, the detection is actually a multi-category task. Therefore, when calculating the TP (True Positive Number), TN (True Negative Number), FP (False Positive Number) and FN (False Negative Number) of each category, the current category is regarded as a positive sample, and all other categories are regarded as negative sample. The formula of each evaluation index is shown below.

Precision rate: \[ TP = TP/(TP+FP) \]
Recall rate: \[ TP = TP/(TP+FN) \]

4.3. Experimental results

In the experiment, 3000 data sets are used for training, and k-person cross-validation is used during training to prevent over-fitting. The SVM model is set to iterate 10,000 times during training. The
The figure below is the accuracy curve on the data set. You can see when the number of iterations reached 10,000 times, the model achieved good results. The results are shown in Figure 3 below.

![Figure 3. Accuracy rate curve](image1)

![Figure 4. Loss curve](image2)

The loss function is used to evaluate the difference between the predicted value of the model and the true value. The better the loss function, the better the performance of the model. The loss function used by different models is generally different. Loss function is divided into empirical risk loss function and structural risk loss function. The empirical risk loss function refers to the difference between the predicted result and the actual result, and the structural risk loss function refers to the empirical risk loss function plus the regular term. And this article uses the empirical risk loss function.

The following figure is the loss value curve of different categories on the data set. It can be seen that when the number of iterations reaches 10,000 times, the model has achieved good results. The results are shown in Figure 4 below.

In the experiment, the traditional method based on keywords and the deep learning method based on convolutional network is used. Using these two methods to compare with the SVM method, the experimental results on the same test set are shown in Table 1 below. It can be seen that the SVM classification method has achieved good results with the same samples.

| Method                                           | Accuracy | Recall |
|--------------------------------------------------|----------|--------|
| Keyword-based traditional methods                | 73.1%    | 71.2%  |
| CNN-based deep learning method                    | 81.2%    | 81.7%  |
| Classification method based on SVM               | 83.7%    | 85.9%  |

5. Conclusions

With the complexity of the network environment, the network security in the power information system has attracted more and more attention. How to detect abnormal behaviours in the power system and their distribution is becoming more and more important. The detection of abnormal behaviours and attack behaviours in network security can generally be divided into four types, based on classification, statistics, clustering and information theory. According to the different processing methods of data characteristics, it can be divided into three different situations based on machine learning, deep learning and traditional methods using keywords. This article uses the SVM algorithm in machine learning to conduct intelligent analysis and research on network security risks. Compared with the keyword-based method, the operation steps are reduced and the accuracy is improved. Compared with the deep learning method, the complexity of the network security risk analysis model is reduced. From the experimental results, it can be seen that the SVM-based classifier detects abnormal behaviour in the power information system, and the accuracy and recall rate have been significantly improved compared with the other two methods. Therefore, good results have been
achieved. The focus of subsequent research will further improve the accuracy of the power safety assessment results and reduce the complexity of the power safety risk analysis model.

Acknowledgment
This work was supported by the State Grid SiChuan Electric Power Company of China (Project Name: Research and platform development of full flow advanced sustainable threat intelligent analysis technology based on deep learning).

References
[1] Zhenlong Yuan, Yongqiang Lu, Yibo Xue. Droid Detector: Android Malware Characterization and Detection Using Deep Learning[J]. Tsinghua Science and Technology. 2016(01)
[2] Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Octeau, Patrick McDaniel. FlowDroid[J]. ACM SIGPLAN Notices. 2014 (6)
[3] William Enck, Peter Gilbert, Seungyeop Han, Vasant Tendulkar, Byung-Gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel, Anmol N. Sheth. TaintDroid[J]. ACM Transactions on Computer Systems (TOCS). 2014 (2)
[4] Asaf Shabtai, Uri Kanonov, Yuval Elovici, Chanan Glezer, Yael Weiss. “Andromaly”: a behavioral malware detection framework for android devices[J]. Journal of Intelligent Information Systems. 2012 (1)
[5] Xie Peng, Li Jianhua, Ou X, etc. Using Bayesian Networks for Network Security Analysis [C] // 2010 IEEE / IFIP International Conference on Reliable Systems and Networks (DSN), Chicago, IEEE, 2010, 211-220.
[6] Poolsappasit N, Dewri R, Ray I. Dynamic security risk management based on Bayesian attack graphs [J]. IEEE Journal of Reliable and Secure Computing, 2012, 9(1): 61-74.
[7] Chen Xiaojun, Fang Binxing, Tan Qingfeng, etc. Malicious insider attack intention inference based on probabilistic attack graph model[J]. Chinese Journal of Computers, 2014, 37(1): 62-72.
[8] Gao Ni, Gao Ling, He Yiyue, etc. Dynamic security risk assessment model based on Bayesian attack graph[J]. Journal of Sichuan University (Natural Science Edition), 2016, 48(1): 111-118.
[9] Ma Chunguang, Wang Chenghong, Zhang Donghong, etc. Dynamic network risk assessment model based on attacker tendency[J]. Computer Research and Development, 2015, 52(9): 2056-2068.
[10] Liu Mingzhen. Particle swarm optimization support vector machine intrusion detection algorithm[J]. Computer Engineering and Applications, 2012, 48 (35): 71-74 + 105.
[11] Wang Xuesong, Liang Ximing. Improved ant colony algorithm to optimize support vector machine for network intrusion detection[J]. Computing Technology and Automation, 2015, 34 (02): 95-99.
[12] Xie Weizeng. Artificial bee colony algorithm optimization support vector machine network intrusion detection[J]. Microcomputer Application, 2017, 33 (01): 71-73.