Research on Application of Support Vector Machine in Intrusion Detection

Lihua Su¹, Wenhua Bai¹, Zhanghua Zhu², Xuan He¹
¹National University of Defense Technology, Xi’an, 710106, China;
²Army Engineering University, Chongqing, 400035, China

*Corresponding author e-mail: bwh@nudt.edu.cn

Abstract. There are some problems such as the high false negative rate and false positive rate in intrusion detection technology. The paper researches on the technology of intrusion detection based on support vector machine. The forming process of SVM classifier is analyzed. The choice of kernel function is discussed. The experimental results show that the SVM-based intrusion detection system with the appropriate kernel function has high detection-accuracy and it can effectively predict and classify the network attacks.

Keywords: Support Vector Machine, Intrusion Detection, Kernel Function

1. Introduction

Most of the traditional statistical learning and machine learning algorithms are based on the basic conditions of big data and large samples, but in practical applications, such harsh conditions are often impossible to meet, and more often are scenarios of small data and small samples. SVM based on these two theories can effectively solve the problems and disadvantages of traditional learning algorithms when dealing with small sample data. Support vector machine [1] (SVM) is a machine learning method. In recent years, SVM has been successfully applied to text classification, intrusion detection, handwritten character recognition, and image processing technology, etc. This paper mainly studied the application of SVM in intrusion detection. The SVM is suitable for small sample data, and is not sensitive to data dimension, so that SVM is suitable for high dimensional heterogeneous unbalanced data sets, and can be well applied in intrusion detection system.

2. Intrusion detection techniques based on machine learning

Login detection [2] is the process of monitoring computer networks and systems to detect security violations. It belongs to the category of pattern recognition and classification. The intrusion detection system works on the key nodes of the computer network system. It collects and analyzes the information of computer network or system in real time, checks for any security policy violations and signs of attacks, so as to achieve the aim of preventing attacks.

According to the plan division, the intrusion detection system can be divided into signature-based detection and anomaly-based detection. Anomaly-based detection is the normal identification and aggressive detection of unfamiliar environments. Machine learning is a new type of high-tech
intelligent learning technology that can discover and predict the state of unknown models, and use deep learning algorithms to mine and analyze the detected state data to find out the laws behind the data. Introducing machine learning into the intrusion detection system can greatly improve the accuracy and efficiency of detection, and effectively distinguish between normal users and abnormal users. Based on this, the application of machine learning in intrusion detection systems has always been a research hotspot.

Most of the traditional various machine learning algorithms is based on the assumption that the number of samples tends to infinity, and have higher requirements on the regularity of sample data distribution. The SVM is not sensitive to data dimension and due to the excellent the classification performance of SVM, the method has a very broad prospects in application of intrusion detection field.

3. Intrusion detection techniques based on SVM

3.1. The working process of intrusion detection system based on SVM

SVM is a kind of supervised learning, which can perform pattern recognition and classification of target objects, which belongs to a kind of linear classification. Its main idea is to transform a low-dimensional space into a high-latitude space through intelligent calculation, and then find the optimal classification method on this basis. The structure of the SVM-based intrusion detection system is shown in Figure 1.

![Image](image.png)

**Figure 1. Structure of Intrusion Detection System based on SVM**

Before starting the training process, the original sample data must be preprocessed, including data noise processing, data normalization and dimension reduction, etc. After the training sample is determined, the training algorithm is completed by using the sample data, and then the determined decision function is obtained, that is, the required SVM classifier is obtained. By using classifier, the state information of the object can be monitored as the input parameter for anomaly detection. The data source attributes used in the detection should be the same as the characteristics determined in the training, and similarly, the data of each attribute domain should be preprocessed, among which the main work is data normalization. After the preprocessing, the data from each detection source forms the vector to be detected in the input space and sends it to the established SVM classifier, which gives a classification result after calculation.

The key to the application of support vector machine in intrusion detection is to establish a SVM classifier. Similar to other intrusion detection methods, building a SVM classifier will generate the normal state model of the target system. The classification process of the classifier is to map some attribute state information of the target object to a classification result. SVM is to achieve such a classification goal by using a decision function, so the establishment of SVM classifier is to determine the decision function, the reference basis is pre-obtained sample data.

3.2. The choice of kernel function

During the process of the establishment of SVM classifier, one of the most important factors affecting decision function is the kernel function used in SVM. Among the classification algorithms of support vector machines, kernel function method is the most advanced and effective one. Kernel function can effectively solve the problem of "dimension disaster" in traditional classification methods. In general,
a kernel function satisfying Mercer's condition is selected, and the original input sample space is mapped to a high-dimensional feature space through a nonlinear mapping. In this way, the nonlinear non-separable problem can be solved by using the linearly separable method in support vector machines.

The commonly used kernel functions are:

1. Polynomial Function
   \[ K(x_i, x) = [(x_i \cdot x) + 1]^d \]

2. Gaussian Kernel Function (Radial Basis Function, RBF)
   \[ K(x_i, x) = \exp\left(-\frac{|x_i - x|^2}{\sigma^2}\right) \]

3. Sigmoid Function
   \[ K(x_i, x) = \tanh\left[y(x_i \cdot x) + c\right] \]

Among the above kernel functions, the polynomial kernel is a representative global kernel function. The global kernel function has an effect on the data in a large range, which is distributed more widely. Especially for those points far away from the test point, the global kernel function can still achieve a better classification effect. The global kernel function has the advantages of being able to obtain the global information of the data set and strong generalization ability.

RBF kernel function is a kind of classical robust radial basis kernel function. Its parameters determine that it has good anti-interference ability to the noise in the data, and the performance of RBF kernel function is very sensitive to the parameters. Under certain conditions, the behavior of Sigmoid kernel is the same as that of RBF kernel, so in general, RBF kernel should be selected first.

When RBF kernel functions are used, the penalty factor C and \( \sigma^2 \) need to be selected. The \( \sigma^2 \) value affects the output response range; the smaller the \( \sigma^2 \) value is, the narrower the response range is, and the less the empirical risk of the resulting classification hyperplane is, but the structural risk is bigger; on the contrary, the bigger the \( \sigma^2 \) value is, the wider the response range is, and the less the structural risk of the resulting classification hyperplane is. The two parameters can be selected by cross-validation, that is, to divide all the data into equal parts and select one of them as the test set, the others as the training set. When one prediction is made for each training set, the accuracy of cross-validation is the proportion of data that are correctly classified. In general, the exponential growth of C and \( \sigma^2 \) is a good way to choose, the classification effect is also better.

4. Experiment parts

4.1. Data preprocessing

In the paper, the experiment adopted KDDCUP99 as the intrusion detection dataset. Each record in the data set corresponds to each TCP/IP connection. Each record is described by 41 characteristics (attributes). The original data set contains 22 attacks, it can be divided into four main types of attacks: DOS (denial of service attacks) and R2L (remote access of unauthorized), U2R (illegal access to the local super user) and Probing (scanning and detection). In the experiment, all types of attacks are classified as abnormal mode, and intrusion detection only needs to judge whether it is normal mode or abnormal mode.

According to the observation of Data set of KDDCUP99, the data set is a set of heterogeneous data. The 41 features (attributes) of each record have both textual and numerical values, some of which vary widely, while others have only two values, 0 and 1. For such a heterogeneous problem, data normalization is needed. The common treatment for text attribute values is to replace them with numeric values. Such as the last attribute value of each record is a text value, which is used to indicate
whether the record is normal or some kind of attack. If all records are divided into normal and abnormal cases only, the property value of the normal record can be set to +1, while the property value of the attack record can be set to -1. The data set should also be scaled properly to prevent some property values from being too large and others too small, so that large numbers drown out decimals. This also can reduce the computation time.

4.2. Experiment and result analysis
The experiment in this paper is completed on the software tool LIBSVM[6], the representative 10% data set in KDDCUP99, namely KddCup. data_10_percent, is selected as the experimental data. The data is divided into training set and test set. After pre-processing, SVM is trained with training data to obtain SVM classifier. Finally, the test data set is predicted. All attack types in the data set are classified into exception class. Detection only needs to determine whether it is normal mode or abnormal mode, and use detection Precision to evaluate detection performance, indicating the proportion of correctly classified data recorded in the total test data set.

![Figure 2. Detection performance of SVM on the test data set](image)

According to the previous analysis, RBF is selected as the kernel function in the experiment. From kddcup.data_10_percent, data of different sizes were selected as training data sets for detection, and relevant detection accuracy was obtained, as shown in the table 1. As can be seen from the experimental results, the average detection accuracy on the test data set is above 95%. It shows that the intrusion detection method based on SVM technology is very effective. Moreover, the more data in the training data set, the higher the detection accuracy of the SVM classifier. This is because the more training data, the more data representing the difference between normal and abnormal samples, so that the SVM detection accuracy will be improved with the increase of learning samples.

5. Conclusion
It can be seen from some experiments in literature [7] that intrusion detection based on support vector machine has the following characteristics: first of all, it does not need all the normal and abnormal information, and it can get a relatively ideal detection effect with less normal and abnormal execution; secondly, the training time and detection time required by this method are shorter than other methods, so this method can be upgraded at any time and carry out efficient real-time detection. By combining support vector machine with other methods, such as PCA analysis, a better intrusion detection system can be constructed, and multiple SVM classifiers can be used to detect different kinds of intrusion [8].

References
[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York, Springer, 1995.
[2] SANDEEP K. Classification and Detection of Computer Intrusions [D]. USA: Purdue University, 1995.
[3] Zhaoqi Bian, Xuegong Zhang. Pattern recognition. 2nd edition [M]. Beijing: Tsinghua University Press, 2000

[4] Zheng M M. Research on Mixed Kernel Function Based on Support Vector Machine [D]. Shanghai: East China Normal University, 2019.

[5] Information and Computer Science, University of California, Irvine. KDD Cup 1999 Data [EB/OL]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[6] LIN Chih-jen, LIBSVM: A library for Support Vector Machines [EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.htm

[7] Kuang F, Zhang S, Jin Z, et al. A novel SVM by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection [J]. Soft Computing-A Fusion of Foundations, Methodologies and Applications, 2015, 19(5):1187-1199.

[8] Thaseen I S, Kumar C A. Intrusion detection model using fusion of chi-square feature selection and multi class SVM [J]. Journal of King Saud University-Computer and Information Sciences, 2017, 29(4):462-472.