Research on Network Data Security Based on RS-PS Support Vector Machine (SVM)

Su Li, Yu Linlin, Li Ting, Liu Xinxin
Dalian Institute of Science and Technology

Abstract. In the era of rapid development of digital information, resource sharing under the Internet environment, convenient access to information has brought us great convenience. However, as a new independent variable, the Internet itself has many unknown vulnerabilities, which provide an attack entrance for attackers and cause serious security risks. In this paper, intrusion detection research based on rough vector set and particle swarm optimization mainly involves particle swarm optimization (PSO), learning of rough function, etc. Based on some shortcomings of the algorithm optimization in the field of intrusion detection, a rough set(RS)-particle swarm(PS) SVM network data security detection method is proposed to provide security guarantee for information services.

1. Introduction
Intrusion detection technology (ID) is a supplement to traditional network security technology. It analyzes and judges by collecting audit logs in computer systems or data streams in the network, and discovers whether there are behaviors and violations of security policies in the network or system. The signs of attacks help the system deal with network attacks and further improve the network security environment [1]. Traditional intrusion detection system (IDS) is unable to deal with network attacks intelligently and actively, and security administrators are often required to manually track hacker attacks. The poor mobility performance, and the defects of network protocols cause great troubles to IDS. Therefore, the traditional intrusion detection system can no longer meet the needs of people, and people have developed active intelligent intrusion prevention system (IPS) on the basis of IDS [2]. Compared with IDS, the intrusion prevention system (IPS) improves the detection of new intrusion methods, and at the same time, when detecting intrusions, it can actively respond to attacks, protect the host from intruders, and increase the security of system administrators, increase the safety management capacity of the system administrator (include security audit, surveillance, attack recognition and response), improve the integrity of the information security infrastructure, and improve the adaptive ability of the intrusion detection system. However, the current intrusion prevention system (IPS) still has many problems, such as the slow speed and low detection rate of the system in the massive network data, For another example, some new attack data are difficult to be detected by the detection system, and the system has high false alarm rate and missing alarm rate [3].

Therefore, in order to continuously cope with the endless levels of attack methods, it has become an urgent need to study new network security technologies or make breakthroughs in original technologies.

2. Network data security monitoring

2.1. Network data security features
The basic idea of PCA is to reduce the dimension of a data set which is composed of many interrelated variables, whether heavy or light, while retaining the variables existing in the data set until the maximum extent. The same operation is done by converting variables into a new set of variables known as primary components [4,5]. The main components are orthogonal and ordered, so that the variables existing in the original variables are less retained when moving down in order. In this way, the maximum information content of the original data can be retained. Such a set is called the principal component. The goal of linear discriminant analysis (LDA) is to find the linear mapping to the optimal low-dimensional subspace through the feature space of data, and the maximum separation between classes in all possible low-dimensional subspaces [6,7]. Since PCA can only remove related linear structures but not nonlinear related structures, a semi-supervised LDA method is proposed to overcome this limitation. Specific methods and steps are as follows:

1. Suppose the number of input original data characteristic particle swarm is \(N\), and each characteristic particle \(x_i\) group has \(m\) characteristic attributes, then the matrix of input data can be expressed as

\[
x_i = (x_{i1}, x_{i2}, \cdots, x_{im}), i = 1, 2, \cdots, n
\]

(1)

The mean particle swarm \(\bar{x}\) and the covariance matrix \(\Sigma\) of sample \(x\) are calculated as follows:

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, i = 1, 2, \cdots, n
\]

(2)

\[
\Sigma = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'}, i = 1, 2, \cdots, n
\]

(3)

In Formula 2 and 3: \(T\) represents the unit of time.

2. \(\phi_1 \geq \phi_2 \geq \cdots \geq \phi_m\) can be obtained after the eigenvalue decomposition of the covariance matrix;

3. Calculation formula to obtain the sample information content \(\theta_i\) of the \(i\)th principal component,

\[
\theta_i = \frac{\phi_i}{\sum_{i=1}^{m} \phi_i}, i = 1, 2, \cdots, m
\]

(4)

4. According to information content \(\theta_i\) to calculate formula (5), \(\sigma\) (cumulative information content) is obtained to determine the number \(K\) of principal component samples, and the first \(K\) feature vectors are selected to constitute the projection matrix \(\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_K]\) representing the maximum information content, and \(\alpha_i\) is \(m\)-dimensional particle swarm [8].

\[
\sigma = \sqrt{\sum_{i=1}^{K} \theta_i}
\]

(5)

5. Fusion LDA is used for dimension reduction again: let the number of the \(P\)-th sample type be \(n_p\), then the total number of training samples is \(n = \sum_{p=1}^{\varphi} n_p\), \(\varphi\) is the number of sample types, the mean value of the \(P\)-th sample type is \(\bar{\alpha}_p = \frac{1}{n_p} \sum_{i=1}^{n_p} \alpha_{p,i}\), and the mean value of all training samples is

\[
\bar{\alpha}' = \frac{1}{n} \sum_{i=1}^{n} \alpha_{i}'
\]
The calculation formula of the between-class dispersion matrix $\delta_m'$ is

$$\delta_m' = \sqrt{\sum_{k=1}^{v} \sum_{l=1}^{n_k} (\overline{\alpha}_k' - \alpha_{kl}')(\overline{\alpha}_k' - \alpha_{kl}')} (6)$$

The calculation formula of the between-class dispersion matrix $\delta_\beta'$ is

$$\delta_\beta' = \sqrt{\sum_{k=1}^{v} n_k (\overline{\alpha}_k' - \alpha')'} (7)$$

Calculate the total divergence matrix $\delta'$ of PCA through the covariance matrix of formula (3):

$$\delta' = (n-1) \times k \quad (8)$$

Combining the cumulative information content $\sigma$ in formula (5), the scatter matrix of semi-supervised LDA can be defined as:

$$\delta_\beta = \delta_\beta' + \sigma (\chi - \delta_\beta') \quad (9)$$

$$\delta_\epsilon = \delta_\epsilon' + \sigma (\delta' - \delta_\epsilon') \quad (10)$$

Where, $\chi$ is the identity matrix and $\delta'$ is the total divergence matrix in PCA. $\sigma \in [0, 1]$ can adjust the dimension reduction effect of PCA algorithm in semi-supervised LDA. When $\sigma = 0$, the semi-supervised dimension reduction is completely changed to LDA algorithm. When $\sigma = 1$, the algorithm completely transforms to PCA algorithm [9].

2.2. Abnormal monitoring response

At the end of the dimension reduction of semi-supervised LDA data, the projection matrix is used as the characteristic particle swarm after the dimension reduction of data, so as to be used for the training and detection of anomaly detection methods. Response injection attacks will change the server-to-client response, thereby providing incorrect system status information. Injection attack will indicate that the system is still below the high alarm (HH) or above the low alarm (LL) [10]. Command injection attack refers to sending forged control or configuration commands to cause system behavior to change. Potential effects of malicious command injection include process loss of control of the device, interruption of device communication, and unauthorized modification of device configuration information and parameter configuration information. Denial of service attacks will try to exhaust the resources of communication links and system programs, leading to insufficient system resources and causing system crashes.

3. Experimental testing and analysis

3.1. Experimental data

In this open data set, each network data has 27 data dimensions, of which the last one indicates whether the network data is normal business data or attack data (1-7 in turn indicates the above attack types). In order to make the research brief, this paper only discusses the normal and abnormal situations of the data, and does not consider the classification of multiple SVM attacks. Before model training and testing, data preprocessing such as standardization and normalization should be carried out to ensure that the results will not be affected by different value ranges of data attributes.

3.2. Experimental methods

Since normalization is not as good as standardization when principal component analysis is used for data dimension reduction, z-SCORE standardized data pretreatment method is selected in this paper to eliminate the influence of these adverse factors, which is specifically described as follows:
\[ \lambda = \sqrt{\frac{\tau - \eta^2}{\phi}} \quad (11) \]

In Formula 11, \( \tau \) is the original input numeric data of an attribute, \( \eta \) is the average value of the attribute, \( \phi \) is the standard deviation of the attribute, and \( \lambda \) is the standardized result of the numeric data.

After standardized processing of the data set, semi-supervised LDA algorithm is used to reduce the data dimension, adjust the \( \sigma \) in formula (9) (10), namely accumulative information content in PCA algorithm, and adjust the bias degree of semi-supervised dimension reduction method. After verification, this paper chooses \( \omega = 0.9 \) to reduce the dimensionality of the data set to obtain better detection results.

In order to test the effect of the rough set-particle swarm optimization algorithm proposed in this paper on intrusion detection data processing, two identical intrusion data sets are selected at the same time, one of the data is preprocessed by the algorithm, and then the data is passed through the improved particle swarm optimization for training, the other data is not processed by the optimization algorithm, but simply processed through feature extraction and normalization, and then the data is trained through PKELM. The change curves of the optimal fitness value and the number of iterations of the two different algorithms in the evolution process are shown in Fig. 1 and Figure 2.

![Figure 1](image.png)

Figure 1 The optimal fitness value of different algorithms
In the process of iteration, the fitness value obtained from the data optimized by the algorithm is generally better than that obtained from the data not processed by the particle swarm optimization algorithm, and the convergence speed is relatively fast.

By identifying the overall accuracy, special effects, sensitivity, or Average, AUC, Kappa statistics, false alarm rate and false alarm rate, the experimental results obtained by particle swarm optimization algorithm are shown in Table 1 and 2.

**Table 1. Performance evaluation of different classifiers**

| Performance evaluation standard (%) | PKELM | Particle swarm optimization algorithm |
|------------------------------------|-------|-------------------------------------|
| ACC                                | 97.56 | 99.55                               |
| Special                            | 96.55 | 99.48                               |
| Sensitivity                        | 95.27 | 98.97                               |
| Average                            | 97.56 | 98.68                               |
| AUC                                | 95.54 | 98.48                               |
| kappa statistics                   | 93.58 | 98.34                               |
| FPR                                | 0.41  | 1.58                                |
| TNR                                | 2.47  | 6.57                                |

**Table 2. Performance evaluation of different algorithms**

|                                | PKELM | Particle swarm optimization algorithm |
|--------------------------------|-------|-------------------------------------|
| ACC                            | 0.95  | 0.99                                |
| Special                        | 0.96  | 0.99                                |
| Sensitivity                    | 0.96  | 0.98                                |
| Average                        | 0.97  | 0.99                                |
| AUC                            | 0.95  | 0.98                                |
| kappa statistics               | 0.96  | 0.99                                |
| FPR                            | 0.74  | 0.97                                |
| TNR                            | 0.69  | 0.95                                |
Table 2 records the values of performance indexes of different algorithms, among which the particle swarm optimization algorithm proposed in this paper has the best performance when detecting intrusion detection data, and has achieved certain effects both in improving the detection rate and reducing the false alarm rate. In the comparison algorithm, particle swarm optimization and other optimization methods are not used, and the detection effect is relatively low.

4. Conclusion
This paper conducts in-depth research on the characteristics of intrusion data, and analyzes the current network data with high dimensionality, complex structure, and large base.

(1) Aiming at the problem of high difficulty in identification of the intrusion data, rough set-particle swarm optimization algorithm is adopted to conduct cluster analysis on the intrusion data, so as to increase the aggregation degree of similar data and reduce the identification difficulty of the data; At the same time, the improved particle swarm optimization algorithm is used to improve the clustering performance.

(2) This paper makes appropriate improvements to the limitations of the particle swarm algorithm. When the particles fall into the local optimum, Gaussian disturbance is used to make the particles jump out of the local optimum. However, when the fitness function has multiple consecutive peaks, the particle may continuously fall into the local optimal.

References
[1] A.Wathiq.O.zulaiha,N.Mohd. Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Systems With Applications . 2017
[2] Qingchao Jiang,Xuefeng Yan. Parallel PCA–KPCA for nonlinear process monitoring[J] . Control Engineering Practice . 2018
[3] C.A. Carrillo-Santos,J.C. Seck-Tuoh-Mora,N. Hernández-Romero,L.E. Ramos-Velasco. Wavenet identification of dynamical systems by a modified PSO algorithm[J] . Engineering Applications of Artificial Intelligenc . 2018
[4] Dongshun Cui,Guang-Bin Huang,Tianchi Liu. ELM based smile detection using Distance Vector[J] . Pattern Recognition . 2018
[5] Syed Ali Raza Shah,Biju Issac. Performance comparison of intrusion detection systems and application of machine learning to Snort system[J] . Future Generation Computer Systems . 2018
[6] Hassan I. Ahmed,Nawal A. Elfeshawy,S. F. Elzoghdy,Hala S. El-sayed,Osama S. Faragallah. A Neural Network-Based Learning Algorithm for Intrusion Detection Systems[J] . Wireless Personal Communications . 2017 (2)
[7] Gang Song,Qun Dai. A novel double deep ELMs ensemble system for time series forecasting[J] . Knowledge-Based Systems . 2017
[8] Azzeddine Bakdi,Abdelmalek Koudri,Abderazak Bensmail. Fault detection and diagnosis in a cement rotary kiln using PCA with EWMA-based adaptive threshold monitoring scheme[J] . Control Engineering Practice . 2017
[9] Mirza M. Baig,Mian M. Awais,El-Sayed M. El-Alfy. A multiclass cascade of artificial neural network for network intrusion detection[J] . Journal of Intelligent & Fuzzy Sys . 2017 (4)
[10] Shu-Zhi Gao,Xiao-Feng Wu,Liang-Liang Luan. PSO Optimal Control of Model-free Adaptive Control for PVC Polymerization Process[J]. International Journal of Automation and Computing. 2018(04)