Advances in Computer Science Research, volume 88

International Conference on Computer, Network, Communication and Information Systems (CNCI 2019)

An Industrial Control Intrusion Detection Method Combining Semi-Supervised LDA and PSO-SVM

Liu Zhan[a], Jie Ling[b] and Peng Lin[C]

Faculty of Computer Guangdong University of Technology Guangzhou, China

[a]13286832426@163.com, [b]jling@gdut.edu.cn

Abstract. In order to improve the accuracy of industrial control intrusion detection, this paper proposes a fusion semi-supervised LDA and PSO-SVM method, using the cumulative contribution rate ω to determine the principal component analysis (PCA) accounted for the proportion of semi-supervised LDA algorithm. The experimental results show that compared with a single PCA or LDA and PSO-SVM combination, this method of combining semi-supervised LDA and PSO-SVM has advantages, high accuracy of anomaly detection, and low false positive rate.

Keywords: PCA, LDA, industrial control system, intrusion detection, SVM.

1. Introduction

With the continuous development and deepening of various innovative applications such as industrial Internet, intelligent manufacturing, and Internet of Things, the trend of industrial control systems (ICS) interconnection has become more apparent. A large number of networked control devices and data exchange facilities and diversified data access methods make industrial control devices more vulnerable to external attacks[1]. The functional design of the industrial control terminal has made it impossible to consider the need for safety protection during design. The limited computing storage resources also restrict the access of security protection measures, resulting in the occurrence of industrial network security incidents represented by the "seismic network" virus in recent years[2]. The use of 0-day vulnerabilities and the diversity of virus variants have exposed the traditional vulnerabilities-based security protection measures[3], and the security protection methods designed to be accurate and efficient and able to cope with unknown attacks have become the focus of current research.

Intrusion Detection System (IDS) is the second line of defense for network security. It monitors network data transmission and analyzes data to find abnormal behavior. IDS introduces machine learning methods such as rough set theory, decision trees, random forests, artificial neural networks, and support vector machines (SVM) to better cope with unknown attacks. SVM is an edge-based classifier, which is based on small sample learning with good generalization ability. It is often used in real world classification applications. At the same time, there is no over-fitting problem that can't be solved by ANN, but it is still affected by the size of the data set, resulting in system failure due to insufficient memory or long training. In real life, data is usually large, unrelated or outdated. SVM lacks the ability to immediately possess the importance of features[4], which ultimately increases the significant computational difficulty and may lead to poor prediction accuracy. Therefore, this paper proposes a semi-supervised dimension reduction method combined with PCA and LDA and an industrial control anomaly detection method combined with PSO-SVM, which improves the accuracy of anomaly detection. The simulation experiment is carried out on the industrial control data set provided by Mississippi State University Key Infrastructure Protection Center, which has been processed numerically. The results show that the proposed method has high detection accuracy.

2. Related Research

Literature[5] proposes the application of improved K-means algorithm in Intrusion detection, and uses the average method to solve the selection of initial clustering centers by traditional K-means algorithm. The improved support vector machine (SVM) algorithm proposed in

literature[6][7] optimizes the parameters of SVM and improves the detection rate of each attack type. Literature[8] proposed a Snot-based protocol analysis and detection system, which has a good detection effect on specific protocols and known illegal data packets. Reference[9] proposes an intrusion detection method based on grid optimization of SVM parameters, and finds that the algorithm has high detection accuracy. The literature[10-11] preprocesses the data set based on Principal Component Analysis (PCA), which accelerates the convergence speed and improves the detection efficiency.

In the above research, the literature[5-9] adopts the traditional intrusion detection method. Due to the high dimensionality of the industrial control data and the diversity of the attack types, the intrusion detection method takes too long or the accuracy is low. Literature[10-11] only uses feature extraction based on PCA to reduce data dimensionality, but because PCA method can only remove information about linear structure rather than nonlinear structure, it can‘t solve the information dimensionality reduction of nonlinear structure.

3. The Method Principle and Procedure

3.1 Data Dimension Reduction Based on Semi-supervised LDA.

By analyzing the basic idea of the PCA and LDA algorithms, the PCA algorithm can be used for all samples, and the intrinsic information of the data is retained to the utmost after the dimensionality reduction, but the category information of the data is ignored, and there are such cases as the matrix size is too large and the solution complicated question. Therefore, in order to realize the complementary advantages of the two, a semi-supervised LDA algorithm is proposed for feature extraction: firstly, the PCA algorithm is used to reduce the sample features, and then the combination of PCA and LDA is used for the second dimension reduction.

The basic idea of principal component analysis (PCA) is to calculate a set of new features arranged in order from importance to size from a set of features. The new feature is a linear combination of original features. Specific steps are as follows:

- Let the number of original input data samples be n, and each sample has p feature attributes, that is, the input data matrix is

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T, \quad i = 1, 2, \dots, n \tag{1}$$

This formula can explain the covariance matrix of xi

$$S = \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T, i = 1, 2, \dots, n \tag{2}$$

Where

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i, i = 1, 2, \dots, n \tag{3}$$

- Eigenvalues decomposition of covariance matrix S is performed to obtain eigenvalues sequence $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

- The sample contribution rate of the principal component is expressed as

$$\rho_i = \lambda_i \left(\sum_{j=1}^{p} \lambda_j \right)^{-1}, \quad i = 1, 2, \dots, p \tag{4}$$

- The number of principal component samples k is determined by the cumulative contribution rateω, and the eigenvectors corresponding to the first k eigenvalues are selected to form the projection matrix $A = [a_1, a_2, \dots, a_k]$, and B is the P-dimensional vector.

$$\omega = \sum_{i=1}^{k} \rho_i \tag{5}$$

The basic idea of LDA is to project lower dimensional space with labeled data by projection method. In this low-dimensional space, similar samples are as close as possible and heterogeneous samples are as far away as possible. The specific steps are as follows:

- The input data is $A = [a_1, a_2, \dots, a_k]$, ai is a p-dimensional vector, p is the number of features after PCA dimensionality reduction. Use LDA for dimensionality reduction:

  The total number of training samples$N = \sum_{m=1}^{C} N_m$, sample number C. The number of samples of a certain type is N_m, and the average value of this class is $\overline{a'_m} = \frac{1}{N_m}\sum_{l=1}^{N_k} a'_{ml}$. The average value of all training samples is $\overline{a'} = \frac{1}{N}\sum_{i=1}^{N} a'_i$.

- The intra-class and inter-class dispersion matrices S_w' and S_b' are:

$$S_w' = \sum_{m=1}^{C}\sum_{l=1}^{N_m}(\overline{a'_m} - a'_{ml})(\overline{a'_m} - a'_{ml})^T \tag{6}$$

$$S_b' = \sum_{m=1}^{C} N_m(\overline{a'_m} - \overline{a'})(\overline{a'_m} - \overline{a'})^T \tag{7}$$

- Then the semi-supervised matrix can be defined as:

$$S_b = (1-\omega)S_b' + \omega I \tag{8}$$

$$S_w = (1-\omega)S_w' + \omega S \tag{9}$$

Where I is the unit matrix. $\omega \in [0,1]$ can adjust the role of PCA and LDA in semi-supervised dimension reduction. When $\omega = 0$, semi-supervised dimension reduction completely changes to LDA algorithm; when $\omega = 1$, the algorithm completely transforms into PCA algorithm.

- According to Fisher discriminant analysis expression:

$$J(\sigma) = \frac{\sigma^T S_b \sigma}{\sigma^T S_w \sigma} \tag{10}$$

σ is any column vector. Fisher's linear discriminant analysis is to find σ that maximizes $J(\sigma)$. Taking this vector as the projection direction, we can find a projection matrix W composed of a set of optimal discriminant vectors. The projection matrix is solved by the following generalized eigenvalues:

$$S_b\sigma = \beta S_w\sigma \tag{11}$$

That is to say, the eigenvectors corresponding to the maximum eigenvalues of the first d $(d < C-1)$ of the matrix $S_w^{-1}S_b$ are obtained to form the projection matrix $W = [\sigma_1, \sigma_2, \dots, \sigma_d]$.

3.2 Particle Swarm Optimization for Parameters of Support Vector Machines.

As a traditional supervised learning model, the core of support vector machine (SVM) is to find a hyperplane that can correctly partition the feature space of binary classification data, while support vector machine (SVM) is the nearest sample point to the hyperplane. Constructing and solving the

optimal hyperplane can be transformed into solving a quadratic regression problem in the original space:

$$\begin{cases} \min \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{n}\xi_i \\ s.t.\ \ y_i[(\omega \cdot x_i) + b] \geq 1 - \xi_i \\ (\xi_i \geq 0; i = 1, 2, \ldots, n) \end{cases} \tag{12}$$

Where: ω is the weight vector; b is the classification threshold; ξ_i is the relaxation factor; C is the penalty factor.

The Lagrange multiplier method is used to solve the problem (12) and the optimal classification function is

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N}\alpha_i y_i (x_i \cdot x) + b\right) \tag{13}$$

Where N is the number of support vectors; α_i ($i = 1, 2, \ldots, N$) is the Lagrange multiplier. If the sample set is non-linear separable, the optimal classification surface is constructed in the high-dimensional feature space by mapping it to the linear separable high-dimensional feature space. There is a corresponding relation $K(x_i, x_j) = \left(\Phi(x_i) \cdot \phi(x_j)\right)$ for the kernel function $K(x_i, x_j)$ satisfying Mercer condition. Therefore, the optimal classification function in high-dimensional feature space is

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N}\alpha_i y_i K(x_i, x) + b\right) \tag{14}$$

Particle swarm optimization (PSO) is an iteration-based optimization algorithm. Particles search iteratively in space following the optimal particles. Each particle will confirm the fitness value according to the fitness function, which is the standard to measure the quality of particles. The updated formulas for velocity and position are

$$v_i^{k+1} = \omega v_i^k + c_1 r_1 (p_i^k - x_i^k) + c_2 r_2 (p_g^k - x_i^k) \tag{15}$$

$$x_i^{k+1} = x_i^k + v_i^{k+1} \tag{16}$$

Where k is the current number of iterations; v_i^k is the current particle velocity; x_i^k is the current particle position; p_i^k, p_g^k is the current individual optimal solution and global optimal solution; ω is the inertia weight; c1, c2 are the acceleration factors; r1, r2 are the random numbers in [0, 1].

In this paper, PSO is selected to optimize the parameters of SVM. The specific steps are as follows:

- Randomly generate the initial particle swarm, and set the current iteration number $k = 1$, given the upper limit of the number of iterations k_{max}.
- Calculating the fitness function to obtain the fitness value, the inertia weight ω, and updating the position information of the particle.
- Traversing the fitness value of each particle, if there is a better value than its p_i^k, update p_i^k, otherwise the original value remains unchanged. The updated p_i^k of all particles was compared with the global extreme value p_g^k, and the best value was saved in p_g^k.

- Determine whether the iteration termination condition is reached. If the particle fitness value meets the requirement or the upper limit k_{max} of the given iteration number has been reached, the algorithm ends, otherwise repeat step c to iterate

The number of particles in this experiment is 30, the upper limit of the iteration is k_{max}=300, the acceleration factor c_1=1.7, and c_2=1.5.

3.3 Intrusion Detection Based on Semi-supervised LDA and PSO-SVM.

The semi-supervised LDA and PSO-SVM based intrusion detection system uses the semi-supervised LDA algorithm to reduce the dimensionality of the industrial control system network data, and uses the PSO algorithm to find the optimal parameters of the SVM classification technology to detect abnormal behavior. The basic flow of intrusion detection is shown in Fig. 1.

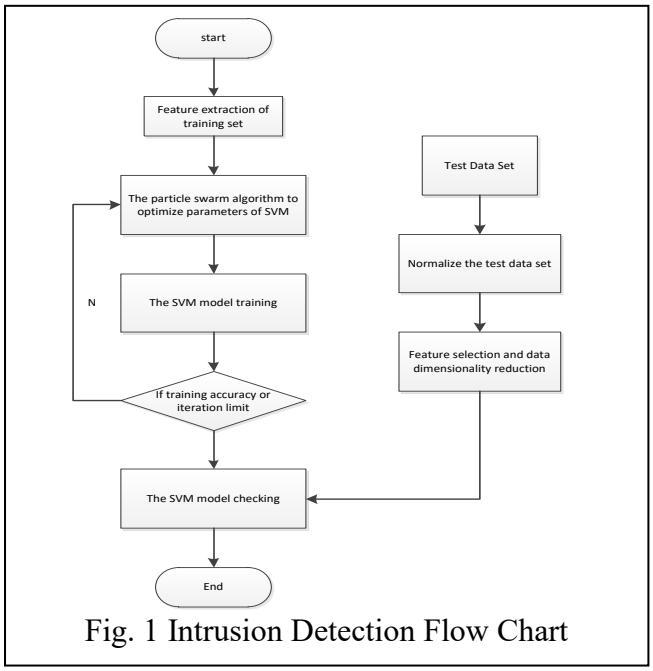

Fig. 1 Intrusion Detection Flow Chart

The whole process of intrusion detection system can be divided into four stages: data processing stage, support vector machine parameter optimization stage, model training stage and model testing stage. In the first step, we use semi-supervised dimensionality reduction method to integrate PCA algorithm with LDA algorithm. The minimal information loss of PCA and the maximum identifiability of LDA classes are complemented to achieve data dimensionality reduction, thus avoiding "dimensionality disaster". The second step, the use of constant iterative particle swarm algorithm to get the optimal support vector machine (SVM) parameters. The third step, after get eigenvector dimension after input to optimize parameters of support vector machine (SVM), to get classifier model. The fourth part, carried out on the test set input classifier model after data preprocessing, comparing test the model.

4. The Results of Simulation and Analysis

4.1 Data Sources and Data Preprocessing.

The Mississippi State University Critical Infrastructure Protection Center publishes data sets[12]. The public data set provides four data sets. This paper selects the transaction data set of 10% of the water storage tank system. The data set includes normal process control and process measurement as well as attacks against SCADA system. Attacks fall into four categories: reconnaissance, response injection, command injection and denial of service (Dos). In the data set, each network data including 27 marked characteristics, of which 26 for the connection characteristics of a tag, which is used to tag data category. Because the intrusion eigenvalues collected by industrial control system from the network have different measurement units and the characteristic attributes of samples are different, this paper carries out normalized pretreatment to eliminate the influence of these unfavorable factors. Detailed description as follows:

$$\tilde{T} = \frac{T - T_{min}}{T_{max} - T_{min}} \tag{17}$$

After normalizing the data set, PCA and LDA are combined to extract features. Adjust ω in formula (8) (9) for better experimental results. The experimental data set consists of 10000 in 10% Transaction Data Set of Water Tank System, and the experimental data is divided into two groups to form a training data set and a test data set. The test data set is further divided into three groups. Experiments, in order to get the optimal classification effect, need to select the proper kernel function, because of the radial basis function of the high-dimensional nonlinear data processing effect is better, so this paper selected the radial basis function as the kernel function. In order to evaluate the performance of classifier, this paper uses Precision, False Positive Rate (FPR) and False Negative Rate (FNR) as evaluation indicators.

4.2 The Experiment Results Analysis.

In order to reflect the influence of PSO parameter optimization on support vector machine, the traditional support vector machine and literature[9] are used as comparison models. Using the same data set, the test results are shown in Table 1.

Table 1 Experimental results of different optimized parameter methods

Parameter optimization method	Precision /%	FPR/%	FNR/%
Not used	78.92	2.56	15.53
GA optimization	84.73	2.01	8.53
Grid optimization	84.50	1.78	10.30
PSO optimization	85.90	1.54	9.64

The simulation results in Table 1 show that the detection rate of network intrusion detection using optimized parameters is much higher than that of traditional SVM, and the false alarm rate and false alarm rate are significantly reduced. Compared with other methods, the average precision of PSO is superior.

In order to verify the superiority of this algorithm, BP Neural Network (BPNN) is used as a comparison model. From Fig. 2 and 3, it can be seen that SVM overcomes the shortcomings of local optimum and slow convergence speed of neural network, and is more suitable for real-time and high detection accuracy requirements of network intrusion detection.

The experimental results in Table 2 show that the semi-supervised LDA method used in feature extraction can improve the detection results of PSO-SVM model.

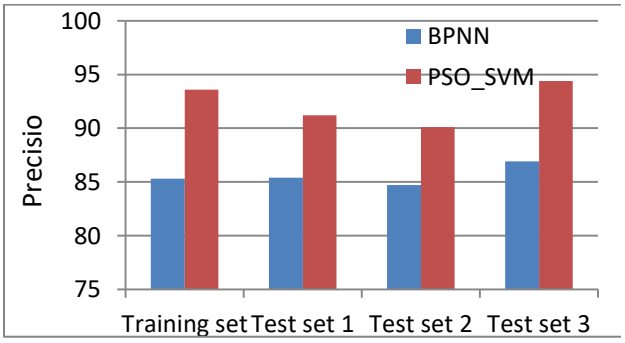

Fig. 2 Precision Ratio of BPNN and PSO_SVM

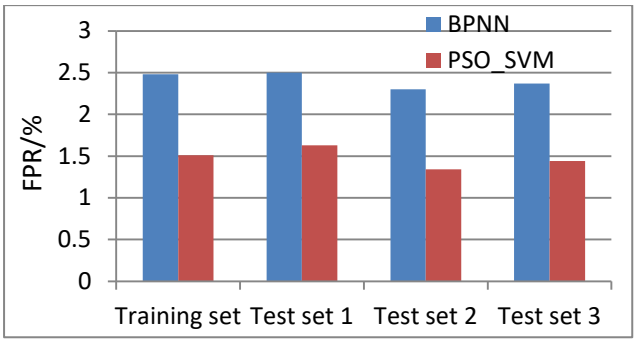

Fig. 3 FPR of BPNN and PSO_SVM

Table 2 Detection results of PSO_SVM by different dimension reduction methods

Dimension reduction method	Precision /%	FPR/%	FNR/%
Not used	85.90	1.54	9.64
PCA	93.23	0.67	11.66
LDA	94.31	1.96	7.41
SLDA	95.71	1.16	6.23

5. Conclusion

In view of the large amount of industrial control data and the characteristics of many dimensions, this paper proposes a fusion method of semi-supervised LDA and PSO-SVM. This method combines two feature extraction methods to reduce the dimension of original data to reduce feature vectors and improve the detection speed of data sets, and PSO algorithm is used to optimize parameters, construct classification model and carry out intrusion detection experiments. Normalized pretreatment of data sets and repeated experiments show that this intrusion detection method, which combines semi-supervised LDA and PSO-SVM, improves the precision of abnormal behavior and reduces the FNR.

Acknowledgments

This work is supported by the science and technology project of Guangdong Province (No.2017B090906003) and the project of Guangzhou Science and Technology (No.201802010043、201807010058).

References

[1]. Yi Hu, Dong Yu, Minglie Liu. Research status and development trend of industrial control network. Computer Science, 2010, 37(1): 23-27, 46.

[2]. An Yang, Limin Sun, Xiaoshan Wang, et al. Review of Intrusion Detection Technologies for Industrial Control Systems [J]. Computer Research and Development, 2016, 53 (9): 2039-2054.

[3]. Yingxu Lai, Zenghui Liu, Xiaotian Cai, et al. Review of Intrusion Detection in Industrial Control Systems [J]. Journal of Communications, 2017 (2): 143-156.

[4]. X.-s. Gan, J.-s. Duanmu, J.-f. Wang, W. Cong, "Anomaly intrusion detection based on pls feature extraction and core vector machine", Knowledge-Based Systems, vol. 40, pp. 1-6, 2013.

[5]. Yang Li. Application of K-means Clustering Algorith m in Intrusion Detection[J]. Computer Engineering, 2007, 33(14): 154-156.

[6]. LIU W J, QIN J T, QU H C, et al. Industrial Control Network Intrusion Detection Method Based on Improved Single-class Support Vector Machine[J]. Journal of Computer Application, 2017, 26(12): 1-5.

[7]. MasoodParvania, Georgia Koutsandria, VishakMuthukumar et al. Hybrid Control Network Intrusion Detection Systems for Automated Power Distribution Systems[C]// Ieee/ifip International Conference on Dependable Systems and Networks. California: IEEE Computer Society, 2014:774 -779.

[8]. Thomas Morris，Rayford Vaughn，Yoginder Dandass. A retrofit network intrusion detection system for modbus rtu and ascii industrial control systems [C]//System Science (HICSS)，2012 45th Hawaii International Conference on. IEEE：2012：2338-2345.

[9]. Gongrang Zhang, Fei Wan. Application of SVM based on grid search in intrusion detection [J]. Computer technology and development, 2016 (1): 97-100.

[10]. Mingyu Qi, Ming Liu, Yanming Fu, etc. Research on SVM Network Intrusion Detection Based on PCA [J]. Information Network Security, 2015 (2): 15-18.

[11]. Sen Zhao, Tingting Qiu. Intrusion Detection System for Wireless Sensor Networks Based on PCA [J]. Computer Engineering and Application, 2014, 50 (14): 88-91.

[12]. Morris T, Gao W. Industrial Control System Traffic Data Sets for Intrusion Detection Research[C]// International Conference on Critical Infrastructure Protection. Springer, Berlin, Heidelberg, 2014.