Postprint: Anomaly Detection Algorithm Based on Genetic Algorithm-Optimized OCSVM Dual-Contour Model

Authors: Yan Tengfei, Shang Wenli, Zhao Jianming, Qiao Feng, Zeng Peng

Date: 2018-08-13T00:00:00+00:00

Abstract

Addressing the particularities of the Modbus industrial bus protocol and the class imbalance inherent in industrial control data samples, One-Class Support Vector Machine (OCSVM) is employed to construct both a normal OCSVM model and an anomalous OCSVM model—constituting a dual-profile modality that simulates normal and anomalous modes of system communication, thereby enabling anomaly detection in industrial control systems. Concurrently, genetic algorithm-optimized dimensionality reduction of independent variables is applied within industrial control network intrusion detection contexts to compress and reduce the dimensionality of input variables, which prevents overfitting in the OCSVM model and addresses issues of low classification accuracy, enhances anomaly detection precision, reduces modeling time, and validates through simulation the effectiveness of the proposed algorithm for industrial control network anomaly detection.

Full Text

Abstract

Access control and network isolation [1], but manual rule setting is prone to errors. Meanwhile, network security middleware products affect system real-time operations. At present, firewall technology is insufficient for protecting large-scale networks and complex intrusion attacks [2]. How to shift industrial control security from "passive defense" to "active protection" has become a primary issue in the cybersecurity domain. Intrusion detection technology, as an active protection technology, can detect hidden intrusion behaviors in normal information flows across network boundaries [3], analyze potential threats, and conduct security audits, making it widely applicable to industrial control system network security.

Currently, anomaly-based intrusion detection systems are widely used in active protection of industrial control security [4]. Previous research on anomaly detection mostly employed machine learning methods, among which Support Vector Machine (SVM) possesses unique advantages in solving small-sample, nonlinear, and high-dimensional pattern recognition problems, thus being widely applied in industrial control anomaly detection [5,6]. Literature [7] utilized SVM classification methods to establish industrial control communication protocol data detection models, which can detect industrial control system anomalies to a certain extent. To improve detection accuracy, OCSVM parameters and kernel function parameters σ must be optimized. Common methods for optimizing one-class support vector machine parameters include random selection, grid search, particle swarm optimization, and genetic algorithms. Traditional random selection based on experience cannot guarantee detection accuracy, exhibits high randomness, and cannot ensure optimal parameter selection. Grid search primarily sets search ranges and step sizes for OCSVM parameters and kernel function parameters σ, establishing a two-dimensional grid for optimization with fixed steps. While it achieves good optimization effects with sufficiently small steps and large enough space, grid search suffers from slow optimization speed and tends to obtain local optimal solutions, severely impacting OCSVM detection effectiveness and real-time performance. Particle swarm optimization for OCSVM parameters is easy to implement with few parameters to adjust, offering fast optimization speed and short detection time, but it easily falls into local optima. Although genetic algorithm optimization of OCSVM parameters involves complex operations requiring selection, crossover, and mutation, it can obtain global optimal solutions and improve OCSVM model detection accuracy. This work aims to improve detection accuracy while reducing detection time and lowering false positive and false negative rates; therefore, genetic algorithms are selected to optimize OCSVM parameters and kernel function parameters σ.

Addressing the imbalance of Modbus industrial bus protocol and industrial control data samples, this paper utilizes the characteristic of One-Class Support Vector Machine (OCSVM) that requires only one class of samples to train anomaly detection models. We construct both normal OCSVM models and abnormal OCSVM models, i.e., dual-contour modes, to simulate normal and abnormal modes of system communication, achieving industrial control system anomaly detection. Meanwhile, research reveals that OCSVM models, due to numerous input variables that are not mutually independent, are prone to overfitting, resulting in low detection accuracy and long modeling time. Therefore, this paper employs genetic algorithms for feature reduction on collected industrial data to remove redundancy, selecting variables that best reflect the relationship between inputs and outputs to participate in industrial control system anomaly detection algorithm modeling, thereby shortening detection time and improving detection accuracy.

1 Modbus TCP Communication Protocol Analysis

Industrial control systems refer to control systems oriented toward industrial fields. There are significant differences between industrial control system information security and traditional IT network information security, with data exchange protocols being a major distinction. Traditional IT information systems use the TCP/IP protocol stack, whereas industrial control systems use dedicated communication protocols or specifications, mostly adopting the Modbus TCP communication protocol [10].

Modbus is an industrial fieldbus protocol standard. The Modbus protocol is an application layer message transmission protocol that includes ASCII, RTU, and TCP message types. The protocol itself does not define the physical layer but defines the message structure that controllers can recognize and use [10]. The Modbus protocol standard is divided into two categories: Modbus on serial links and Modbus on TCP/IP. Through the Modbus/TCP message transmission protocol, abnormal detection systems can be built. Literature proposed using OCSVM to detect intrusion origins and time nodes, creating a cluster-based intrusion detection model that improved detection accuracy but had poor real-time performance. Literature proposed combining k-means clustering with OCSVM for intrusion detection, which could detect malicious network traffic and achieve intrusion detection. Although detection time improved, it was prone to overfitting when facing high-dimensional nonlinear data samples. Literature combined methods with OCSVM to reduce fault detection rates, avoiding the impact of initial values on the separation matrix, thereby shortening delay time.

2 OCSVM Anomaly Detection Model

To improve detection accuracy, OCSVM parameters and kernel function parameters σ must be optimized. This paper uses genetic algorithms to optimize the dimensionality reduction of independent variables in industrial control networks. This method improves the accuracy of anomaly detection and reduces modeling time. Simulation results show that the proposed algorithm is effective for anomaly detection in industrial networks.

2.1 OCSVM Principle

For the imbalance of industrial control anomaly data, how to achieve good detection performance on unbalanced categorical data is a concern. One-Class Support Vector Machine solves situations where only one class of samples is available for training classifiers. This paper adopts OCSVM to establish dual-contour models under normal and abnormal modes. Through collaborative discrimination mechanisms, the decision function (i.e., optimal hyperplane) is obtained, resulting in an industrial intrusion detection model based on OCSVM.

OCSVM first maps the input space to a high-dimensional space through a kernel function, separating them from the origin as much as possible in the high-dimensional space. That is, it seeks to construct an optimal hyperplane in the

feature space, assuming the coordinate origin as abnormal samples with class label -1 and normal samples with class label +1. The goal is to determine the boundary of normal samples, i.e., the optimal hyperplane. Data outside the boundary is classified as abnormal, achieving maximum margin between normal industrial data and abnormal data (coordinate origin). For training data x, f(x) indicates its position in the high-dimensional space relative to the hyperplane; f(x) being positive is considered normal class, while negative is considered abnormal class.

When industrial control anomaly monitoring data is linearly inseparable, kernel functions are introduced. This paper uses the Gaussian kernel function to transform data from nonlinear in the input space to linearly separable in the feature space, enabling accurate classification of anomaly detection data, as shown in Figure 3. In two-dimensional space, the optimal hyperplane is a straight line.

OCSVM seeks to minimize the objective function:

where: x_i are sample data, l is the number of training samples, is a trade-off parameter, Φ is the mapping from original space to feature space, ω and are the normal vector and offset of the hyperplane in feature space, respectively.

The Lagrangian function is introduced:

where: α_i and β_i are Lagrange multipliers. Introducing the Gaussian kernel function to map the sample space to the feature space yields the dual problem:

The Gaussian kernel function in feature space is:

From the above, we can obtain :

The decision function is:

2.2 Genetic Algorithm Optimization

The genetic algorithm optimization for variable dimensionality reduction modeling uses industrial intrusion detection data, including anomaly detection labels (normal samples +1 and abnormal samples -1), directly extracted features (Modbus/TCP length, address code, port number, etc.), and constructed features (number of abnormal codes in data addresses per unit time, number of connected device identifiers, number of read function codes, etc.), totaling 25 feature data points. Clearly, these 25 input variables have certain relationships and are not mutually independent. To shorten modeling time and improve modeling accuracy, variables with major influencing factors are selected from the 24 input variables (excluding detection labels) to participate in final modeling [13].

Using genetic algorithms for optimization requires mapping the solution space to the coding space, where each code corresponds to a solution (i.e., chromosome or individual). The coding length is designed as 24, with each bit of the chromosome corresponding to an input variable. If a bit of the chromosome is "1," the corresponding input variable participates in final modeling; conversely, if "0,"

the corresponding input variable is not used for final modeling. This paper uses the inverse of the mean square error of test set data as the fitness function for genetic algorithms to optimize variable dimensionality reduction for Modbus/TCP industrial data input. Through continuous iteration, the most representative input variables are selected to participate in OCSVM intrusion detection model building. The genetic algorithm optimization variable dimensionality reduction flowchart is shown in Figure 4.

A. BP Model Construction. To compare prediction effects before and after genetic algorithm optimization, a BP model is first established using all 24 input variables.

B. Initial Population Generation. Generate n random initial data with string structures, where each string-structured data serves as an individual, and a population consists of n individuals.

C. Fitness Function Calculation. For OCSVM, the time complexity for obtaining a satisfactory solution can be worst-case, where Q is the number of support vectors. The complexity of the OCSVM method is related to the probability of individual behavior changes in the genetic algorithm [14], while the complexity of SVM, although not having a fixed proportion, is related to the number of support vectors and the size of the training set. Therefore, the OCSVM algorithm complexity is easily affected by the selected training set. The process is conducted offline and will not affect the online real-time application of OCSVM. Increasing offline optimization time to obtain optimal OCSVM classification performance is worthwhile. Due to the introduction of the dual-contour model, the number of OCSVMs to be trained will increase, so the corresponding training time will be extended, but the optimization of structure and training process is beneficial.

Fitness Function Calculation. The inverse of mean square error is selected as the fitness function:

where: f(x_i) is the predicted value of the i-th sample in the test set; t_i is the true value of the i-th sample in the test set; n is the number of test set samples.

Selection Operation. This step uses a proportional selection operator. The sum of all individual fitness values in the population is calculated. The relative fitness of population individuals is calculated as the probability of individuals being selected and inherited to the next generation. The roulette wheel selection method is adopted to randomly generate numbers between 0 and 1 to determine the number of times each individual is selected. The larger the fitness, the greater the chance of the individual being selected. If an individual can be selected multiple times, its genetic genes will expand in the population, and the probability of the individual being selected for the optimized independent variable set will increase.

Crossover Operation. Single-point crossover operator is used to achieve compression and dimensionality reduction of independent variables, and arithmetic

crossover operator is used to optimize BP neural network weights and thresholds.

Mutation Operation. Single-point mutation operator is used to achieve dimensionality reduction of input independent variables, and non-uniform mutation operator is used to optimize BP neural network weights and thresholds.

Optimization Result Output. After multiple iterations, the variable combination that best represents the input-output relationship is obtained.

Selection of Training Set and Test Set. After dimensionality reduction of independent variables by genetic algorithm, the corresponding training set and test set are obtained for anomaly detection by the OCSVM model.

This paper uses genetic algorithms to optimize OCSVM parameters and kernel function parameters σ. Research shows that the main factors affecting the quality of OCSVM models are the values of parameters and kernel function parameters σ. Applying genetic algorithms to optimize and σ can obtain optimal solutions and improve the detection accuracy of the model. Therefore, this paper uses genetic algorithms to optimize the parameters of the OCSVM model.

Using OCSVM to construct abnormal OCSVM models and normal OCSVM models for functional control behavior, simulating abnormal and normal modes of system communication, anomalies in the network can be discovered in real-time through collaborative discrimination. The real-time collaborative discrimination process of the dual-contour model is shown in Figure 5. The real-time collaborative discrimination engine captures and extracts functional control behavior information in real-time, using it as input data for the model to perform anomaly determination. The collaborative discrimination mechanism of the dual-contour model is briefly described as follows:

D) If the abnormal OCSVM model determines "normal" and the normal OCSVM model also determines "normal," then the final result is "normal."

E) If the abnormal OCSVM model determines "abnormal" and the normal OCSVM model determines "normal," then the final result is "abnormal."

According to existing literature on genetic algorithm computational complexity, assuming at generation W, the expected number of surviving bad patterns is given by the mean of a binomial distribution. When the genetic algorithm reaches a satisfactory solution after time, where λ and K are constants related to genetic algorithm pattern reliability [14].

D. Optimization Result Output. The inverse of mean square error is used as the fitness function:

where: f(x_i) is the predicted value of the i-th sample in the test set; t_i is the true value of the i-th sample in the test set; n is the number of test set samples.

2.3 Dual-Contour Decision Mechanism

The dual-contour decision mechanism works as follows:

1) If the normal OCSVM model determines "normal" and the abnormal OCSVM model determines "abnormal," the final result is "normal."
2) If the normal OCSVM model determines "abnormal" and the abnormal OCSVM model determines "normal," the final result is "abnormal."
3) If both OCSVM models determine "normal," the final result is "normal."
4) If both OCSVM models determine "abnormal," the final result is "abnormal."
5) If the abnormal OCSVM model and normal OCSVM model produce inconsistent determinations, further arbitration is required, considering false positive and false negative rates to construct weight values for different alarms to complete final anomaly determination.

The overall process is: Industrial control system traffic data collection → Genetic algorithm optimization for variable dimensionality reduction → One-class support vector machine model → Data capture and behavior analysis → Abnormal OCSVM model/Normal OCSVM model → Anomaly determination? → Alarm arbitration decision.

3 Experimental Results and Analysis

To verify the proposed OCSVM anomaly detection model, a simulation platform was built to extract industrial control traffic data using Wireshark software to capture data packets. Functions provided by the 7/% Neural Network Toolbox and Genetic Algorithm Toolbox were used to implement genetic algorithm optimization for variable dimensionality reduction in the 7/% environment. 1000 Modbus/TCP data packets from the industrial data network were extracted, normalized to [0,1], with 700 used as training data and the remaining 300 as test data for genetic algorithm optimization. After program execution, the population fitness function evolution curve is shown in Figure 7.

The results show that after dimensionality reduction, a set of variables was selected. The number of input variables participating in modeling after optimization is half of the total input variables. Comparing BP network test results before and after optimization screening, it was found that when using 11-dimensional input variables for modeling, prediction accuracy was improved and enhanced. Meanwhile, modeling time was reduced from 24.914s to 7.641s, indicating that using genetic algorithms for input variable compression and dimensionality reduction significantly shortens modeling time.

To verify the effectiveness of the proposed anomaly detection method, the genetic algorithm-optimized OCSVM dual-contour model was compared with Support Vector Machine algorithm, BP algorithm, and OCSVM algorithm. The same 700 data packets captured by Wireshark were used as training data, with another 300 data packets as test sets for network anomaly detection verification.

Through simulation experiments with OCSVM experimental results comparison, the results are shown in Table 1.

Table 1 Performance comparison of various algorithms

The results show that while none of the algorithms achieve 100% detection accuracy on the test set, the OCSVM detection accuracy can reach 98.00%, far higher than other algorithms, indicating that the GA-OCSVM dual-contour model proposed in this paper can effectively reduce the false negative rate and false positive rate of anomaly detection through real-time collaborative discrimination mechanism, thereby improving detection accuracy. Meanwhile, after genetic algorithm optimization, the dimensionality of input variables is reduced, shortening OCSVM model detection time and improving generalization ability and flexibility. The GA-OCSVM dual-contour anomaly detection model better addresses the problems of more normal data than abnormal data in industrial control networks and high input variable dimensionality affecting detection accuracy, making it more suitable for actual industrial control networks.

For verification of the dual-contour model anomaly detection algorithm built with OCSVM, 700 data packets from the captured 1000 data packets were used for model training, and the remaining 300 data packets were used for testing. When not using genetic algorithm optimization, the actual classification and predicted classification of the test set are shown in Figure 8, with classification accuracy being:

Stronger, more flexible. The GA-OCSVM dual-contour anomaly detection model better addresses the problems of more normal data than abnormal data in industrial control networks, high input variable dimensionality affecting detection accuracy, making it more suitable for actual industrial control networks. After genetic algorithm optimization, the actual classification and predicted classification of the test set are shown in Figure 9, with classification accuracy reaching 93.75% and 98.00% accuracy.

4 Conclusion

This paper addresses the particularities of the Modbus industrial bus protocol in industrial control networks and the imbalance in industrial control data samples compared to traditional IT networks. By utilizing the OCSVM algorithm to construct both normal and abnormal OCSVM models (i.e., dual-contour modes) that simulate normal and abnormal modes of system communication, industrial control system anomaly detection is achieved. Simultaneously, this paper applies genetic algorithm optimization for variable dimensionality reduction to industrial control networks, achieving optimized processing of input variables, suppressing OCSVM model overfitting and low classification accuracy problems, improving anomaly detection precision, and reducing modeling time. Simulation verification demonstrates the effectiveness of the proposed algorithm for industrial control network anomaly detection, showing significant application value in industrial control anomaly detection algorithms.

References

[1] Rezai N, Wool Z. Anomaly detection in Modbus/TCP SCADA systems. International Journal of Critical Infrastructure Protection, 2015, 10: 1-12.

[2] Wang Ming, et al. Research on anomaly detection based on one-class support vector machine in industrial control systems. Smart Industry and Information Security, 2017, 36(23): 1-12.

[3] Chen Shanxue, et al. Research and development overview of intrusion detection technology in industrial control systems. IEEE Transactions on Information Forensics & Security, 2017, 3011-3023.

[4] Zhao Hui, et al. Industrial control system intrusion tolerance system based on failure detector algorithm. Journal of Jilin University (Information Science Edition), 2011, 29(4): 34-376.

[5] Li Chaolun, et al. 362-OCSVM-based anomaly detection research. Computer Applications Research, 2015, 32(6): 1778-1781.

[6] Chen Shuang, et al. Application of one-class support vector machine in networked control system anomaly detection. Electronics Letters, 2014, 50(25): 1935-1936.

[7] Wang Wanliang, et al. Industrial control network anomaly detection based on OCSVM. *Reference contains corrupted encoding and incomplete citation information.*

[8] Hu Yao, et al. Research on intrusion detection in SCADA systems. International Journal of Panfeng, 2015, 30(11): 1-12.

[9] Jian Fan, et al. Overview of intrusion detection technology research and development in industrial control systems. Smart Industry and Information Security, 2017, 36(23): 1-12.

[10] Knowledge W, Prince D, Hutchison D, et al. A survey of cyber security management in industrial control systems. *Reference contains corrupted encoding and incomplete citation information.*

[11] Liao Zhifeng, et al. Research on industrial control system security protection technology. Computer Applications Research, 2015, 32(6): 150-158.

[12] Hu Yao, et al. Research on anomaly detection based on one-class support vector machine in industrial control systems. Smart Industry and Information Security, 2017, 36(23): 1-12.

[13] Wang Wanliang, et al. Industrial control network anomaly detection based on OCSVM. ChinaXiv Cooperative Journal.

[14] Rezai N, Wool Z. Anomaly detection in Modbus/TCP SCADA systems. International Journal of Critical Infrastructure Protection, 2015, 10: 1-12.

[15] Wang Ming, et al. Research on anomaly detection based on one-class support vector machine in industrial control systems. Smart Industry and Information Security, 2017, 36(23): 1-12.

[16] Chen Shanxue, et al. Research and development overview of intrusion detection technology in industrial control systems. IEEE Transactions on Information Forensics & Security, 2017, 3011-3023.

[17] Zhao Hui, et al. Industrial control system intrusion tolerance system based on failure detector algorithm. Journal of Jilin University (Information Science Edition), 2011, 29(4): 34-376.

[18] Li Chaolun, et al. 362-OCSVM-based anomaly detection research. Computer Applications Research, 2015, 32(6): 1778-1781.

[19] Chen Shuang, et al. Application of one-class support vector machine in networked control system anomaly detection. Electronics Letters, 2014, 50(25): 1935-1936.

Note: Figure translations are in progress. See original paper for figures.