Research Article

A Modified Hybrid Method Based on PSO, GA, and K-Means for Network Anomaly Detection

Yuan Yuan \(^1\) and Yuangang Li \(^2\)

\(^1\)School of Economics and Management, Shanghai University of Political Science and Law, Shanghai 201701, China
\(^2\)Faculty of Business Information, Shanghai Business School, Shanghai 200235, China

Correspondence should be addressed to Yuangang Li; liyuangang76@163.com

Received 6 June 2022; Revised 2 September 2022; Accepted 9 September 2022; Published 30 September 2022

Academic Editor: Chien Ming Chen

Copyright © 2022 Yuan Yuan and Yuangang Li. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Data anomaly detection plays a vital role in protecting network security and developing network technology. Aiming at the detection problems of large data volume, complex information, and difficult identification, this paper constructs a modified hybrid anomaly detection (MHAD) method based on the K-means clustering algorithm, particle swarm optimization, and genetic algorithm. First, by designing coding rules and fitness functions, the multiattribute data is effectively clustered, and the inheritance of good attributes is guaranteed. Second, by applying selection, crossover, and mutation operators to particle position and velocity updates, local optima problems are avoided and population diversity is ensured. Finally, the Fisher score expression for data attribute extraction is constructed, which reduces the required sample size and improves the detection efficiency. The experimental results show that the MHAD method has better performance than the K-means clustering algorithm, the support vector machine, decision trees, and other methods in the four indicators of recall, precision, prediction accuracy, and F-measure. The main advantages of the proposed method are that it achieves a balance between global and local search and ensures a high detection rate and a low false positive rate.

1. Introduction

With the explosive development of Internet technology in recent years, the network has penetrated into every aspect of people’s lives and work. While the network brings advantages such as convenient exchange of real-time information and easy access to open resources, its characteristics such as no supervisor and no defense also cause more and more network security problems [1]. At present, the challenges of network security mainly include the difficulty of detecting abnormal behaviors, inefficient processing of massive data, and the failure of achieving overall protection [2]. As one of the important means to deal with the challenges of network security, intrusion detection can monitor and respond to the unauthorized use and abuse of network resources and can protect the effectiveness and confidentiality of data [3]. Therefore, optimizing the network intrusion detection method to further improve its accuracy and efficiency is of great significance for maintaining network security and developing network technology.

The research directions of network intrusion detection usually include misuse intrusion detection and anomaly intrusion detection [4]. Misuse intrusion detection is applicable to the judgment of abnormal behaviors related to normal behaviors, and the accuracy of this method depends on the completeness of the detection knowledge base, so it is difficult to identify the changeable intrusion behaviors [5]. Anomaly intrusion detection describes various characteristics of the network under normal operation by constructing models and rules. When there is a big deviation from these characteristics, the network can be judged to be abnormal or attacked [6]. The failure rate of misuse intrusion detection has increased with the surge of network attack methods and attack objects. Therefore, anomaly intrusion detection has become the main research direction of network intrusion detection.
The research methods of anomaly intrusion detection mainly include the statistical method, the Bayesian network, the neural network, and data mining [7]. Statistical anomaly intrusion detection is the most traditional method, which is suitable for small-scale network detection. However, this method is slow and inefficient in detecting large-scale networks [8, 9]. The Bayesian network method and the neural network method have fast detection speeds, but they have high false alarm rates and poor adaptive abilities [10, 11]. Data mining is implemented through the analysis of audit records of the Internet to mine potentially useful information and use this information to detect anomalous intrusion [12]. Data mining has become the most important anomaly intrusion detection method because of its high efficiency and accuracy [13].

Most of the data mining methods are based on intelligent algorithms such as particle swarm optimization (PSO), genetic algorithm (GA), and clustering algorithm [14, 15]. However, these intelligent algorithms still have the defects of local optimum, slow convergence speed, and low calculation accuracy. In order to further improve the accuracy of anomaly detection, this paper proposes a modified hybrid anomaly detection (MHAD) method to realize the complementary advantages of the K-means clustering algorithm (K-means), the PSO, and the GA, that is, three intelligent algorithms with strong correlation.

In the process of anomaly detection, when the original network data package is decomposed by the protocol, the protocol field and the package payload could contain a large amount of data. Even if the optimized method is adopted to detect all the data, there will still be problems such as the increase in data dimension and the complexity of attribute relationships. Therefore, in order to improve the efficiency of anomaly detection, this paper defines the Fisher score expression for extracting data attributes based on the Fisher discrimination, so as to select the valuable attributes and eliminate the interference of redundant attributes on detection.

The rest of this paper is organized as follows: Section 2 introduces three intelligent algorithms related to the MHAD method and reviews literature on the improvement of these algorithms. Section 3 describes the implementation process and important steps of the MHAD method. Section 4 firstly introduces the data sources of experiments in this paper, then sorts and extracts the data attributes, and finally classifies the experimental data. Section 5 analyzes the function of attribute extraction and the effect of the MHAD method through two experiments. Section 6 summarizes the conclusions of this paper and introduces future research directions.

2. Theories and Literature of Algorithms

2.1. K-Means Clustering Algorithm. K-means is a kind of iterative clustering algorithm. The steps of this algorithm are as follows: first, randomly select K individuals from N data individuals as the initial clustering center. Then, calculate the distance between each individual and each clustering center. Finally, assign each individual to the nearest clustering center [16]. A clustering center and all individuals assigned to it represent a cluster. When each new individual is assigned to a cluster, the clustering center is recalculated based on the existing individuals in the cluster. This iterative process is repeated until one of the following three termination conditions is satisfied: no individual is reallocated; the clustering center is not changed; or the quadratic sum of error is locally minimal.

Let the individual sample set be \( X = \{X_1, X_2, \ldots, X_i, \ldots, X_n\} \). The goal of the K-means is to find a data set \( Z = \{Z_1, Z_2, \ldots, Z_m, \ldots, Z_k\} \), \( 2 \leq k \leq n \), so as to minimize the sum of intraclass dispersion as shown in the following equation:

\[
\min J_c = \sum_{m=1}^{k} \sum_{i=1}^{n} d(X_i, Z_m). \tag{1}
\]

In (1), \( X_i \) is a D-dimensional real vector, \( Z_m \) is the m-th clustering center, and \( d(X_i, Z_m) \) is the distance between an individual and the corresponding clustering center. If (2) is satisfied, then \( X_i \) has been allocated to the most suitable clustering center \( Z_m \).

\[
J_i = d(X_i, Z_m) = X_i - Z_m = \min_{m=1,2,\ldots,k} X_i - Z_m. \tag{2}
\]

2.2. Particle Swarm Optimization. Particle swarm optimization (PSO) is an evolutionary algorithm based on swarm intelligence. Firstly, the PSO determines the dimension of the solution space according to the number of variables in the problem to be optimized. Secondly, the particle swarm is randomly initialized in a given solution space so that each particle has its own initial position and velocity. Finally, the optimal solution is found through iteration [17].

In each iteration, each particle updates its position and velocity in the solution space by tracking two extremums [18]. One extremum is the optimal solution found by the individual particle itself, which is called the individual extremum. Another extremum is the optimal solution reached by all particles in the particle swarm, which is called the global extremum. After finding these two extremums, the particle can update its velocity and position according to (3) and (4):

\[
V_{i}^{\text{iter}+1} = \omega \cdot V_{i}^{\text{iter}} + c_1 \xi (p_{i}^{\text{iter}} - X_{i}^{\text{iter}}) + c_2 \eta (p_{g}^{\text{iter}} - X_{i}^{\text{iter}}), \tag{3}
\]

\[
X_{i}^{\text{iter}+1} = X_{i}^{\text{iter}} + \gamma \cdot V_{i}^{\text{iter}+1}. \tag{4}
\]

In (3) and (4), \( V_{i}^{\text{iter}} \) is the velocity of particle \( i \) in the iter-th iteration. \( X_{i}^{\text{iter}} \) is the current position of particle \( i \) in the iter-th iteration. \( i \) is the population size, \( i = 1, 2, \ldots, N \). \( c_1 \) is the weight coefficient of the optimal value of the particle’s own history, which represents the cognition of the particle itself. \( c_2 \) is the weight coefficient of the optimal value of all particle tracking, which represents the cognition of the particle to the whole population. \( p_{i}^{\text{iter}} \) is the position of the individual extremum of particle \( i \). \( p_{g}^{\text{iter}} \) is the position of the global extremum of all particles. \( \xi \) and \( \eta \) are random
numbers evenly distributed in $[0, 1]$. $\gamma$ is the speed coefficient of the position update.

$\omega$ is the inertia weight, indicating the effect of the particle’s previous velocity on the current velocity. The larger $\omega$ can enhance the global search ability, and the smaller $\omega$ can enhance the local search ability, so $\omega$ plays a balancing role between the global search and the local search. The calculation formula of inertial weight is shown in the following equation:

$$\omega = \omega_{\text{max}} - \frac{(\omega_{\text{max}} - \omega_{\text{min}})}{\text{iter}_{\text{max}}} \times \text{iter}. \quad (5)$$

In (5), $\omega_{\text{max}}$ and $\omega_{\text{min}}$ represent the maximum value and the minimum value of the weight, respectively. $\text{iter}$ represents the current iteration number, and $\text{iter}_{\text{max}}$ represents the maximum iteration number. Usually, $\omega$ decreases linearly from the initial maximum value $\omega_{\text{max}} = 0.9$ to the minimum value $\omega_{\text{min}} = 0.4$ with the increase of iterations.

### 2.3. Genetic Algorithm

The genetic algorithm (GA) is a random-parallel search algorithm based on natural selection and gene genetics [19]. This algorithm considers the problem’s solution set as a population, then makes the solution better and better through genetic operations such as selection, crossover, and mutation, and finally obtains the global optimal solution. Selection, crossover, and mutation are the three basic steps of the GA [20].

The selection operation is the process of selecting individuals with high adaptive value from the parent population to produce a new population. It is also the process of individuals replicating themselves according to their viability. This operation reflects the biological "survival of the fittest" law.

The crossover operation is implemented in two steps. Firstly, the newly replicated individuals are paired in the matching pool obtained by the selection operation. Then, the crossover points are selected at random, and the paired individuals are cross-bred to generate new individuals.

The mutation operation is to change the gene values of individuals in a population. The mutation operation has two functions. One is that the local random search ability of mutation operations can accelerate the convergence to the optimal solution when the neighborhood of the optimal solution has been approached. The other is that the mutation operation can maintain the population diversity to prevent the occurrence of local convergence.

### 2.4. Literature Review

Anomaly detection is the main mechanism to reduce possible network intrusions, and algorithm-based anomaly detection can effectively distinguish between "normal" and "abnormal" behaviors in the network [21]. At present, K-means, PSO, and GA have been widely used in network anomaly detection, and these three algorithms have their own advantages and disadvantages in practice. As an unsupervised partition clustering algorithm, K-means has the advantages of being simple to operate and fast convergence. However, the clustering results of this algorithm are affected by the selection of the initial center. The sensitive reflection of outliers and the weak global search ability also cause the K-means to fall into a local optimum [22]. The PSO has the advantages of high search diversity and few adjustment parameters. However, in the face of solving complex optimization problems, the PSO is easy to fall into the local minimum value, with slow convergence speed and low accuracy [23]. The GA can effectively reduce the search space to have a fast convergence rate, and each individual in the population has made a good contribution to obtaining possible solutions to the problem [24]. However, there are conversion errors in the coding and decoding process of the GA, which tend to omit individuals close to the extreme value, resulting in a decrease in algorithm efficiency with the increase of variables. Many scholars have carried out research on improvement research for the defects of the K-means, PSO, and GA.

For the optimization of the K-means, Basha et al. [25] proposed an improved K-means based on the entropy distance between data point attributes, which could effectively remove outliers in the data set and greatly improve the clustering accuracy. However, the algorithm had a poorer measurement effect for attributes with larger distance values, so it was not suitable for the case of many attribute categories. Zhang et al. [26] proposed a new K-means based on density canopy to solve the problem of determining the most appropriate cluster number and best initial population. The density canopy was taken as the preprocessing process of the K-means, and its result was taken as the clustering number and initial clustering center of the K-means. This proposed algorithm had good antinoise performance, but the parallel performance needed to be improved.

For the optimization of the PSO, Huang et al. [27] proposed a generalized Pareto model based on the PSO for anomaly detection. Because the generalized Pareto model was multidimensional, the search ability of the particle swarm was improved by introducing a comprehensive learning strategy, and the possibility of the particle swarm falling into a local optimum was reduced by using dynamic neighborhoods. However, because this proposed model required some prior knowledge for parameter estimation, it increased the complexity of detection. Vijayakumar et al. [28] pointed out that using rules to select detection data could filter out the basic features that had a direct impact on network anomalies, thereby effectively reducing the computational complexity of anomaly detection in time and space. Ganapathy et al. [29] proposed a new PSO-based rule extractor whose main advantage was the comprehensibility of the extraction rules. They used an extension of the PSO fitness function under time constraints, thus achieving the goals of maximizing classification accuracy and minimizing the number of input features.

For the optimization of the GA, Chiba et al. [30] designed a new intrusion detection model by combining the Back Propagation Neural Network (BPNN) and the Improved Genetic Algorithm (IGA). Because the learning rate and the momentum term were two of the most relevant parameters affecting the performance of the BPNN classifier, they used IGA to find the optimal or near-optimal values of
these two parameters to ensure a high detection rate, high accuracy, and low false alarm rate. Zhang et al. [31] proposed an intrusion detection model based on the improved GA and deep trust networks. Through multiple iterations of the improved GA, the optimal number of hidden layers and the number of neurons in each layer were adaptively generated, so that the network could obtain a high detection rate in the face of different types of attacks. However, due to the high classification accuracy and good generalization ability of the proposed model, its training time was long.

For the optimization of combining multiple algorithms, Alguliyev et al. [32] proposed a multicriteria improvement method based on the weighted PSO for the problems of predefined cluster centers and the existence of multiple local minimum points in the K-means. This proposed method took the intercluster distance as the optimization criterion to minimize the intracluster distance and maximize the intercluster distance. This method outperformed K-means in terms of robustness and anomaly detection accuracy. In order to improve the performance of network intrusion detection system, Almoman [33] proposed a feature selection model based on PSO, GA, gray wolf optimizer, and firefly optimization. The proposed model was a rule-based pattern recognition method that could achieve high detection accuracy using few features, but the effectiveness of this model under a deep learning architecture remained to be studied. Moukhafi et al. [34] proposed a new intrusion detection system based on hybrid GA and support vector machine (SVM) combined with PSO feature selection, in which PSO was used to select the most influential features to learn the classification model, so as to improve the efficiency of detecting known and unknown attacks. The main advantage of this proposed system was that it greatly reduced the size of the original training data set and simplified the optimization of SVM parameters by the GA.

It can be seen from the above analysis that algorithm-based network anomaly detection has achieved rich research results. At the same time, existing research has also confirmed the feasibility of combining multiple algorithms for detection optimization. However, there are still two main challenges to be solved. On the one hand, it is the imbalanced data volume. In the field of intrusion detection, the training set often contains a large number of normal samples, but few attack data. Unbalanced training samples will make the trained detection model biased; that is, the model will pay too much attention to normal data, resulting in a decrease in the recognition effect of abnormal data. On the other hand, it is the flexible change of data features. A good anomaly detection algorithm should have good temporal and spatial adaptability to deal with the problem that the distribution of data features may be different in different situations.

At present, there are few studies on these two challenges in the field of network intrusion, and few studies combine the three algorithms of PSO, GA, and K-means for anomaly detection. In order to effectively adapt to the complex network anomaly detection environment, this paper designs a network anomaly intrusion detection method based on drawing on the advantages and making up for the disadvantages of these three algorithms, so as to improve the accuracy of the results and ensure the efficiency of the method. At the same time, this paper constructs the Fisher score expression for data attribute extraction, so as to avoid the problem of biased detection results.

3. Modified Hybrid Anomaly Detection (MHAD) Method

The GA can realize a global effective search and increase the population diversity, but it does not take the historical and current state of the individual into account when performing mutation operations in a population. The PSO can perform iterative optimization according to the individual’s historical state and current state. However, when the individual optimal position and the global optimal position are close to each other, all individuals will evolve in this direction, resulting in premature convergence. The essence of anomaly detection is to cluster the normal data and the abnormal data, respectively, and the advantages of the GA and PSO can just make up for the disadvantages of each other. As a result, by using K-means clustering as the iterative judgment condition, the combination of the PSO and GA can be effectively realized. The MHAD method, based on the K-means, PSO, and GA, not only increases the diversity of individuals, but also ensures the good inheritance of individuals, so that the global optimal solution can be obtained efficiently and accurately. Figure 1 shows the implementation steps of the MHAD method, and each step of the method is explained in detail as follows:

Step 1: Assign values to parameters. The parameters that need to be assigned are the cluster number $K$, the
population size \( N \), the maximum allowable velocity \( V_{\text{max}} \), and the maximum number of iterations \( \text{itermax} \).

Step 2: Create the initial population. Within the range of each attribute value of the sample, the initial population is randomly generated according to the coding principle. Because the MHAD method is based on the K-means for anomaly detection, the cluster center can be selected as an individual in the population. Let the sample dimension be \( D \) and let the number of cluster centers be \( K \). Considering the position and speed of the particle, the real number coding is adopted. The encoded form of each particle is shown in the following equation:

\[
X_i = \begin{bmatrix}
x_{i1} & x_{i2} & \cdots & x_{iD} \\
x_{i1} & x_{i2} & \cdots & x_{i2D} \\
\vdots & \vdots & \cdots & \vdots \\
x_{ik1} & x_{ik2} & \cdots & x_{ikD} \\
\end{bmatrix}, \quad V_i = \begin{bmatrix}
v_{i1} & v_{i2} & \cdots & v_{i1D} \\
v_{i1} & v_{i2} & \cdots & v_{i2D} \\
\vdots & \vdots & \cdots & \vdots \\
v_{ik1} & v_{ik2} & \cdots & v_{ikD} \\
\end{bmatrix}
\] (6)

In equation (6), \( i = 1, 2, \ldots, N \), \( x_{ikD} \) is the D-dimensional position of individual \( i \) in the \( K \)-th cluster, and \( x_{i1} \) is the D-dimensional position of individual \( i \) in the \( K \)-th cluster.

Step 3: Calculate the adaptive value of particles in the population. According to (2), the distance \( J_i \) between \( X_i \) and the corresponding clustering center is obtained. Then, the adaptive value of \( X_i \) is calculated according to (7). The smaller the dispersion of the individual, the larger the adaptive value.

\[
\text{fit} (X_i) = \frac{1}{(J_i + 1)}
\] (7)

Step 4: Update the speed and position of the particles. This step determines the convergence rate and accuracy of the MHAD method. Select the top \( N \) particles in adaptive value, update the speed of the particle according to (3), and update the position of the particle according to (4).

Step 5: Cluster the particles in the population. According to the coding of the cluster center, the cluster is divided according to the nearest neighbor principle, and the new cluster center is calculated.

Step 6: Determine whether the maximum number of iterations is reached or whether the cluster center of each particle is no longer changed. If yes, go to Step 11. If not, go to Step 7.

Step 7: Filter the particles according to the adaptive value. Firstly, the distance between \( N \) particles and the new cluster center is obtained according to (2). Secondly, the adaptive value of each particle is calculated according to (7). Finally, the particles ranked in the top \( M \) in terms of the adaptive value are selected (\( M < N \)).

Step 8: Perform crossover operation on particles. The \( M \) particles are paired in unrepeatable pairings. Then, \( M \) new particles are generated by performing the crossover operation with probability \( p_c \) according to equation (8), \( p_r = \text{rand}[0, 1] \).

\[
\begin{align*}
X_{1\text{iter}} &= p_c \cdot X_{1\text{iter}} + (1 - p_c) \cdot X_{2\text{iter}} \\
X_{2\text{iter}} &= p_c \cdot X_{2\text{iter}} + (1 - p_c) \cdot X_{1\text{iter}} \\
V_{1\text{iter}} &= p_c \cdot V_{1\text{iter}} + (1 - p_c) \cdot V_{2\text{iter}} \\
V_{2\text{iter}} &= p_c \cdot V_{2\text{iter}} + (1 - p_c) \cdot V_{1\text{iter}}
\end{align*}
\] (8)

In (8), \( \text{iter} \) represents the current number of iterations. \( x_1, x_2 \) and \( \bar{x}_1, \bar{x}_2 \), respectively, represent the position of the two particles before and after this step. \( v_1, v_2 \) and \( \bar{v}_1, \bar{v}_2 \), respectively, represent the velocity of the two particles before and after this step.

Step 9: Perform mutation operation on particles. Perform the mutation operation shown in (9) and (10) for each particle \( X_{i\text{iter}} \) with random probability \( p_r \), \( p_r = \text{rand}[0, 1] \).

\[
\begin{align*}
X_{i\text{iter}} &= \begin{cases} 
X_{i\text{iter}} + c_i & \text{if } \text{fit} (X_{i\text{iter}} + c_i) > \text{fit} (X_{i\text{iter}}) \text{ and } p_r > 0.5 \\
X_{i\text{iter}} & \text{otherwise,}
\end{cases}
\end{align*}
\] (9)

\[
\bar{v}_i = \begin{cases} 
0.5 \cdot (\bar{v}_{i\text{iter}} + v_{i\text{iter}}) & \text{if } p_r < 0.5 \\
\bar{v}_{i\text{iter}} & \text{otherwise.}
\end{cases}
\] (10)

In (9), \( c_i \) is a random number uniformly distributed in \([x^L - \bar{x}_{i\text{iter}}, x^U - \bar{x}_{i\text{iter}}]\) interval, and \( x^L, x^U \) are the borders of the available interval. \( \text{fit} \) represents the adaptive value function as shown in (7).

Step 10: Update the number of iterations and the number of particles in the population. The \( M \) particles after the mutation operation are combined with the original \( N \) particles, so that the number of particles in the population becomes \( M + N \). After adding 1 to the number of iterations, go to Step 3.

Step 11: Output the clustering results.

4. Experimental Data

4.1. Data Attribute Extraction. Data acquisition and preprocessing are necessary parts of anomaly detection. The data selected for this paper is from the KDD Cup 1999 data set [35] (http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html), which is a network test data set established by Lincoln Laboratory in Massachusetts Institute of Technology to simulate the US air force LAN (Local Area Network) [36]. This data contains simulated intrusions in a
variety of network environments, including 22 attack types and 1 normal type, as shown in Table 1.

As can be seen from Table 1, identification types in the data set can be divided into Normal, DoS, Probe, R2L, and U2R. Each piece of data in the data set contains 42 attributes. The attribute format for a piece of data is shown as follows:

0, udp, ftp, SF, 105, 146, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 255, 254, 1.00, 0.01, 0.00, 0.00, 0.00, 0.00, portswipe.

In this piece of data, the first attribute represents the connection time. The second attribute indicates that the data is a normal connection or an intrusive connection. The third attribute is the service type, such as http, ftp, and smtp. The fourth attribute represents the connection tag, such as SF, REJ, and RSTR. The next 37 attributes are numeric attributes that represent record parameters at connection time. The last attribute is the class tag attribute, indicating whether the data is a normal connection or an intrusive connection.

In this paper, the data attributes of the classified data sets in KDD Cup 99 are extracted based on Fisher discrimination. Fisher discrimination is an attribute selection algorithm based on sample distance. When an attribute enables the maximum distance between samples of different clusters and the minimum distance between samples of the same clusters, this algorithm assigns the highest Fisher score to the attribute [37].

For the dichotomy problem, let the training data set be \( X = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\} \in \mathbb{R}^D \). Set the dimension of the original space as \( D \), the class tag as \( y_j \in \{+1, -1\} \), and the training sample number as \( N \). The definition of the Fisher score is shown in the following equation:

\[
F = \frac{S_b}{S_w} .
\]  

(11)

In (11), \( S_b \) is the dispersion between clusters, which describes the distance between two kinds of samples. \( S_w \) is the within-class dispersion, which describes the distance between samples of the same kind.

The calculation formula of \( S_b \) is shown in the following equation:

\[
S_b = \left( \frac{\Delta m_1 - \Delta m}{m_1} \right)^2 + \left( \frac{\Delta m_2 - \Delta m}{m_2} \right)^2 .
\]  

(12)

In (12), \( \Delta m_1, \Delta m_2 \) and \( \Delta m \) refer to the mean value of the normal samples, the mean value of the abnormal samples, and the mean value of all samples, respectively.

The calculation formula of \( S_w \) is shown in the following equation:

\[
S_w = \sigma_p^2 + \sigma_n^2 .
\]  

(13)

In (13), \( \sigma_p^2 \) and \( \sigma_n^2 \) are the variance of the normal samples and the variance of the abnormal samples.

Therefore, according to equations (11)–(13), this paper defines the Fisher score expression for the \( r \)-th attribute of a data is shown in the following equation:

\[
F_r = \frac{S_b}{S_w} = \frac{2}{k-1} \sum_{r=1}^{k} \left( \frac{\Delta m_{r} - \Delta m}{m_{r}} \right)^2 .
\]  

(14)

In (14), \( \Delta m_{r} \) is the mean value of the \( r \)-th attribute of the K-th cluster. \( m_{r} \) is the mean value of the \( r \)-th attribute of all samples. \( \sigma_{r}^2 \) is the variance of the \( r \)-th attribute of the K-th cluster.

4.2. Data Classification. There are about 5 million pieces of network connection data in the KDD Cup 1999 data set. Because it is difficult for the computer equipment used in experiments in this paper to calculate such a data amount, the subdata set of “kddcup.data_10.percent” in KDD Cup 1999 was selected. There are 493,751 records in this subdata set, including 97,278 abnormal records and 396,473 normal records. There are four types of abnormal data, that is, DoS, U2R, R2L, and Probe, and the specific classification identification of each type is listed in Table 1. In this paper, false alarm rate (FAR) and detection rate (DR) are used to evaluate the analysis results, and their definitions are shown as follows:

\[
\text{FAR} = \text{misjudgment number of normal records/total number of normal records}.
\]

\[
\text{DR} = \text{detected number of intrusion records/total number of intrusion records}.
\]

Considering the amount of data in “kddcup.data_10.percent,” this data set is divided into training set A1 and test set A2. The training set is mainly used to generate the detection model, including generating the clustering center vector. Because there are only about 50 abnormal data points of type U2R in the “kddcup.data_10.percent” data set, it is not necessary to analyze this type as a separate test set. In this paper, 1 training set and 4 test sets are classified through random sampling of “kddcup.data_10.percent,” as shown in Table 2.

---

**Table 1: Network data identification type.**

<table>
<thead>
<tr>
<th>Identification type</th>
<th>Meaning</th>
<th>Specific classification identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Normal record</td>
<td>Normal</td>
</tr>
<tr>
<td>DoS</td>
<td>Denial of service attack</td>
<td>Back, land, Neptune, pod, smurf, teardrop</td>
</tr>
<tr>
<td>Probe</td>
<td>Monitoring and other detection activities</td>
<td>Ipsweep, nmap, portsweep, Satan</td>
</tr>
<tr>
<td>R2L</td>
<td>Unauthorized access from remote machine</td>
<td>ftp_write, guess_passwd, imap, multihop, phf, spy, warezclient, warezmaster</td>
</tr>
<tr>
<td>U2R</td>
<td>Unauthorized access to root authority by ordinary users</td>
<td>buffer_overflow, loadmodule, perl, rootkit</td>
</tr>
</tbody>
</table>

---
The clustering algorithm can be applied to the network anomaly detection based on the following two basic assumptions:

The amount of normal data is much more than that of abnormal data.

The values of some attributes of the abnormal data deviate significantly from the normal value range.

The data in Table 2 satisfy the above two basic assumptions, so they can be used as experimental data for subsequent analysis.

There are great differences in the attributes of the experimental data, and the units of measurement are not uniform. Therefore, in order to eliminate the adverse effect on the accuracy of experimental results, the experimental data must be normalized. The normalization method for a matrix $D \times K$ is shown in the following equation:

$$\bar{x}_{\alpha \beta} = \frac{x_{\alpha \beta} - \bar{x}}{s}, \quad \alpha = 1, 2, \cdots, D; \quad \beta = 1, 2, \cdots, K. \quad (15)$$

In equation (15), $\bar{x} = 1/D \times K \sum_{\beta=1}^{K} \sum_{\alpha=1}^{D} (x_{\alpha \beta} - \bar{x})^2$. The normalized experimental data can be obtained through the calculation of (15).

### 5. Experimental Results

The input parameters of experiments include the following: cluster number $K = 2$, population size $N = 50$, cross mutation particle number $M = 20$, maximum allowable velocity $V_{\text{max}} = 1$, maximum number of iterations $i_{\text{itermax}} = 500$, and $c_1$ and $c_2$ of PSO are 1. In this paper, the training set A1 is selected to obtain the clustering center and calculate the Fisher score of each data attribute. For symbolic attributes (such as protocol, service, and flag), enumeration can be used to map to numeric attributes. In the Fisher score sorting, all types of invasion can be classified as abnormal data. However, focusing only on recall improvement will reduce the probability of successful prediction of truly abnormal data. Therefore, in order to evaluate the detection results of each method, compared with normal data, it is necessary to pay more attention to the prediction of real abnormal data, so recall is introduced as a parameter. However, focusing only on recall improvement will reduce the probability of successful prediction of truly abnormal data.

#### 5.1. Experiment 1

In order to analyze the effect of the Fisher score sorting on attribute extraction, experiment 1 is designed in this paper. Firstly, according to the Fisher sorting results, the top 10, 20, and 30 attributes are extracted to form three groups. Secondly, 10, 20, and 30 attributes are randomly selected to form three groups. Finally, all 42 attributes are treated as one group. In experiment 1, data from Test set A2 are used to analyze these seven groups, respectively, and the MHAD method is used to calculate the FAR, DR, and running time, as shown in Table 3.

As can be seen from Table 3, after extracting data attributes, the accuracy of anomaly detection is improved, and the false alarm rate and the running time are reduced. The detection rate of the first group is higher than that of the sixth group, indicating that the Fisher score sorting could improve the accuracy and efficiency of results while reducing the amount of data required for detection. It can be seen from the results of the first six groups that the detection accuracy increases with the increase of the attribute number, but the detection rate is the lowest when all attributes are included, thus proving that the existence of redundant attributes can reduce the efficiency of anomaly detection.

#### 5.2. Experiment 2

On the basis of Experiment 1, Experiment 2 selects the top 15 attributes ranked by Fisher score and inputs the data from Test Set A2 to analyze the effect of the MHAD method. In Experiment 2, the K-means is adopted as the reference object because it represents the traditional anomaly detection method. The hybrid K-means-PSO method [38] is adopted as the reference object because it represents the improved anomaly detection method combining two algorithms. In addition, this paper uses the Support Vector Machine (SVM), Naive Bayes (NB), and Decision Trees (DT) as reference objects because these three methods are often used in network anomaly detection. The advantage of SVM is that the final decision function is determined by a small number of support vectors, which can avoid the adverse effect of the sample space dimension on the detection results. The advantage of NB is that the necessary parameters (mean and variance of variables) can be estimated from a small amount of training data, and it is less sensitive to missing data. The advantage of DT is that it has a strong ability to classify data, and there is no requirement for historical data when making judgments.

In Experiment 2, four evaluation parameters of precision, recall, prediction accuracy, and F-measure are used to evaluate the detection results of each method. The precision can judge the overall detection of the method. Compared with normal data, it is necessary to pay more attention to the prediction of real abnormal data, so recall is introduced as a parameter. However, focusing only on recall improvement will reduce the probability of successful prediction of truly abnormal data.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Normal data</th>
<th>Abnormal data</th>
<th>Normal proportion (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set A1</td>
<td>19502</td>
<td>498</td>
<td>97</td>
</tr>
<tr>
<td>Test set A2-DoS</td>
<td>24256</td>
<td>744</td>
<td>97</td>
</tr>
<tr>
<td>Test set A2-R2L</td>
<td>24256</td>
<td>744</td>
<td>97</td>
</tr>
<tr>
<td>Test set A2-Probe</td>
<td>24256</td>
<td>744</td>
<td>97</td>
</tr>
<tr>
<td>Test set A2-Mix</td>
<td>24256</td>
<td>744</td>
<td>97</td>
</tr>
</tbody>
</table>

### Table 2: Classification of data sets.

As can be seen from Table 3, after extracting data attributes, the accuracy of anomaly detection is improved, and the false alarm rate and the running time are reduced. The detection rate of the first group is higher than that of the sixth group, indicating that the Fisher score sorting could improve the accuracy and efficiency of results while reducing the amount of data required for detection. It can be seen from the results of the first six groups that the detection accuracy increases with the increase of the attribute number, but the detection rate is the lowest when all attributes are included, thus proving that the existence of redundant attributes can reduce the efficiency of anomaly detection.

In Experiment 2, four evaluation parameters of precision, recall, prediction accuracy, and F-measure are used to evaluate the detection results of each method. The precision can judge the overall detection of the method. Compared with normal data, it is necessary to pay more attention to the prediction of real abnormal data, so recall is introduced as a parameter. However, focusing only on recall improvement will reduce the probability of successful prediction of truly abnormal data.
normal data, so prediction accuracy is introduced as a parameter. Because recall and prediction accuracy are a pair of contradictory variables, the F-measure is introduced to balance these variables to evaluate each method more objectively. According to the confusion matrix analysis in Table 4, four evaluation parameters can be obtained as shown in equations (16)–(19).

\[
\text{Precision} = \frac{TA + TN}{TA + TN + FA + FN}, \tag{16}
\]

\[
\text{Prediction accuracy} (P) = \frac{TA}{TA + FA}, \tag{17}
\]

\[
\text{Recall} (R) = \frac{TA}{TA + FN}, \tag{18}
\]

\[
F - \text{measure} = \frac{2P \times R}{P + R}. \tag{19}
\]

Table 5 shows the detection results of the 6 methods in Test Set A2. In terms of recall, the MHAD method is 4% higher than the second-ranked hybrid PSO-K-means method and the SVM. In terms of accuracy, only the MHAD method exceeds 90%. In terms of prediction accuracy, the MHAD method improves by 15% over the worst-ranked K-means. In terms of F-measure, the MHAD method and the SVM are tied for first place, both at 71%. In the field of intrusion detection, the collected sample data is usually the egress and ingress traffic at the gateway or router of the entire network. These datasets have the characteristics of large volume, complex information, and difficult identification. Therefore, the number of samples available for anomaly detection is small, and traditional methods may easily lead to inaccurate model descriptions, resulting in a large error rate. In contrast, the MHAD method proposed in this paper has the advantages of small errors, high flexibility, and good stability.

Figure 2 shows the intraclass dispersion and change curves of the best individuals in each generation of the population in the iterative process of each method. When the maximum number of iterations is reached, the intraclass dispersion sum of the MHAD method is the smallest, indicating that the method has the highest convergence accuracy. At the same time, the MHAD method fluctuates slightly in the convergence process, which is due to the
introduction of GA to increase the diversity of particles. Although the K-means has the fastest convergence speed, it is easy to fall into a local optimum. The hybrid PSO-K-means method is difficult to ensure high classification speed and accuracy at the same time, and it is prone to misclassification. The SVM is difficult to solve the requirement of high-accuracy prediction under the condition of a small number of samples. Because each step of DT is accurate classification without the concept of ambiguity, a slight deviation in the selection of model nodes may directly affect the reliability of abnormal classification results. From Table 5 and Figure 2, it can be seen that the MHAD method has better performance than other methods, which verifies the effectiveness of the proposed method in network anomaly detection.

6. Conclusions

The rapid development and wide application of the Internet not only promote the prosperity of the social economy, but also bring unprecedented challenges to network security. Current network attacks are characterized by diversity, persistence, and concealment, which leads to a surge of abnormal network data and reduces the accuracy of original network anomaly detection methods. Therefore, it is extremely important to design an accurate and efficient abnormal data detection method to protect network security.

The MHAD method proposed in this paper extends the original anomaly detection framework and introduces the preprocessing idea of feature selection, which reduces the complexity and inferiority of massive network data and greatly improves the speed and accuracy of event classification. The MHAD method introduces the PSO on the basis of the K-means, which effectively controls the accuracy of network event classification by calculating the fitness value of the particles and optimizing the particles by combining the crossover mutation process of the GA until the best classification is obtained.

It can be seen from the experimental results that, compared with the other five anomaly detection methods, the average improvement rates of the MHAD method in terms of recall, precision, prediction accuracy, and F-measure are 5%, 14.8%, 7.8%, and 3.25%. In summary, the contributions of the MHAD method are that it is able to achieve a balance between global and local search and ensures a high detection rate and a low false positive rate. The MHAD method has high flexibility and scalability and is not constrained by the field of application environment, so this method is also suitable for other types of data detection problems.

At the same time, this paper defines the Fisher score expression for extracting data attributes based on the Fisher discrimination. By comparing the Fisher sorting data with randomly selected data and initial data, it is concluded that using the data with extracted valuable attributes for anomaly detection can improve the accuracy rate and reduce the required sample size. It is also proved that the redundant attributes of the data will have adverse effects on the anomaly detection. As a result, the Fisher score expression proposed in this paper can be applied to network anomaly detection under various data attribute conditions.

Due to the relatively long running time of the MHAD method, we will continue to optimize the method steps to shorten the detection time in the future. Our future research focus will also include the design of an abnormal weak correlation data detection method based on data attributes. In the large-scale network, there is a large amount of unusually weak correlation data generated from multiple data sources and multiple time periods. Although this data cannot directly interfere with network communication alone, it will pose a threat to network security when it accumulates to a certain amount. Therefore, it is of vital importance to improve the detection efficiency of abnormal weak correlation data.

Data Availability

We are using open source data, which can be found at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was Shanghai University of Political Science and Law.

References


