Research Article

Towards Effective Network Intrusion Detection: A Hybrid Model Integrating Gini Index and GBDT with PSO

Longjie Li^{1,2}, Yang Yu,¹ Shenshen Bai,^{1,3} Jianjun Cheng,¹ and Xiaoyun Chen¹

¹School of Information Science & Engineering, Lanzhou University, Lanzhou 730000, China

²China Information Technology Security Evaluation Center, Beijing 100085, China

³Department of Electronic and Information Engineering, Lanzhou Vocational Technical College, Lanzhou 730070, China

Correspondence should be addressed to Longjie Li; ljli@lzu.edu.cn

Received 17 August 2017; Revised 24 January 2018; Accepted 11 February 2018; Published 26 March 2018

Academic Editor: Eduard Llobet

Copyright © 2018 Longjie Li et al. 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.

In order to protect computing systems from malicious attacks, network intrusion detection systems have become an important part in the security infrastructure. Recently, hybrid models that integrating several machine learning techniques have captured more attention of researchers. In this paper, a novel hybrid model was proposed with the purpose of detecting network intrusion effectively. In the proposed model, Gini index is used to select the optimal subset of features, the gradient boosted decision tree (GBDT) algorithm is adopted to detect network attacks, and the particle swarm optimization (PSO) algorithm is utilized to optimize the parameters of GBDT. The performance of the proposed model is experimentally evaluated in terms of accuracy, detection rate, precision, F1-score, and false alarm rate using the NSL-KDD dataset. Experimental results show that the proposed model is superior to the compared methods.

1. Introduction

With the prompt development of the Internet, a variety of network security problems has continued to occur. At the same time, many security mechanisms have been implemented to maintain the security of computer systems. Among these mechanisms, intrusion detection systems (IDSs) play a vital role in protecting computing infrastructures from attackers and intruders.

The primary task of an IDS is to discover and prohibit abnormal connections from network traffics. In general, intrusion detection approaches can be divided into two categories: *misuse-based detection* and *anomaly-based detection* [1–3]. A misuse-based detection method stores in advance signatures of already known intrusive behaviors in a database and determines whether a network connection is an attack by matching its characteristic with the signatures. If matching one signature, this connection is an attack. These methods are effective in identifying the well-known network attacks with low false alarm rate. However, they will fail when meeting unknown intrusions whose properties do not match any

signature in the database [4]. To address this problem, a misuse-based method needs to regularly update its database. Whereas, obtaining the signatures of new intrusive behaviors is usually expensive. Conversely, an anomaly-based detection method trains models for normal behavior and detects network intrusions by identifying traffics that significantly deviate from normal profile [5]. The basic hypothesis in anomaly-based detection methods is the characteristic of an abnormal connection is far from those of normal connections. In an anomaly-based detection system, we no longer need to store signatures of attack patterns. The anomaly-based detection methods can recognize not only unknown intrusive behaviors but also unknown future attacks. That is an advantage over misuse-based methods. Nevertheless, they may misclassify some normal traffics that are located in the boundaries between normal and abnormal behavior. Because new attacks keep emerging, anomaly-based methods have captured more attention from researchers.

To achieve reliable detection results, great efforts have been made by researchers. In the early stages, rule-based expert systems [6] and statistical methods [7] were applied

in intrusion detection systems. But the worse performance of these approaches when dealing with large-scale network traffics relegates their application to small datasets. In nature, anomaly-based intrusion detection is a classification problem. Hence, researchers have sought to develop solutions for IDS by utilizing various machine learning techniques [8], such as decision trees, support vector machines (SVMs), naive Bayes, and artificial neural networks (ANNs). When developing an IDS, the primary goal is to achieve the best possible accuracy. To this end, hybrid methods have been largely proposed to further enhance the accuracy of intrusion detection when compared to using individual machine learning approaches [4, 9, 10]. The basic idea behind a hybrid model is to significantly improve the detection performance by means of combining several machine learning techniques. In general, the styles of existing hybrid models mainly include ensemble classifier [9, 11, 12], clustering plus classification [13–15], and feature selection plus classification [3, 4, 16–21].

As mentioned above, hybrid models show a new way for network intrusion detection. In this paper, a novel hybrid model, namely, *GINI-GBDT-PSO*, is proposed to put forward a solution for IDS with high accuracy. The proposed method integrates the *Gini index*, the *GBDT* (gradient boosted decision trees) algorithm (GBDT has some other names, e.g., *GBRT* (gradient boosted regression tree), *MART* (multiple additive regression tree), and *tree net.*) [22], and the *PSO* (particle swarm optimization) algorithm [23] together. Feature selection, which extracts an optimal subset of features to represent the whole dataset, is a key step in intrusion detection because training a classifier based on the optimal subset can not only reduce the learning time but also improve its accuracy [19]. In the proposed method, we adopt the *Gini index*, which has been used for dimensionality reduction in the study of text mining [24–26], to extract significant features from original datasets. After feature selection, the proposed method uses the *GBDT* algorithm to train a prediction model on the optimal feature space. The *GBDT* algorithm is a powerful supervised learning method, which integrates the gradient boosting framework and decision tree technique into one ensemble model [22]. Due to its successful applications in many research fields, such as disease modeling [27, 28], web-search ranking [29, 30], and travel time prediction [31], we argue that the *GBDT* algorithm will be quite competent to the work of intrusion detection. The parameters of *GBDT*, for instance, learning rate ν , are optimized by the *PSO* algorithm. Finally, we experimentally study the performance of the *GINI-GBDT-PSO* method on the *NSL-KDD* dataset [32], compared with several baselines including single classification methods and hybrid models. Experimental results show that the proposed hybrid method improves the detection performance in comparison with baselines.

The rest of the paper is organized as follows. Section 2 gives the criteria used to evaluate the performance of intrusion detection models. The detail of the proposed hybrid method is described in Section 3. In Section 4, the performance of the proposed method is experimentally evaluated compared with several baselines. Section 5 briefly reviews

TABLE 1: Confusion matrix.

		Predicted	
		Abnormal	Normal
Actual	Abnormal	TP	FN
	Normal	FP	TN

some hybrid intrusion detection models. Finally, Section 6 concludes this work.

2. Evaluation Criteria

Table 1 shows a confusion matrix that is used to represent the information related to the actual and predicted classifications performed by a detection model. In Table 1, TP is the number of abnormal connections that are correctly detected; TN is the number of normal connections that are correctly identified; FP is the number of normal connections that are incorrectly classified as abnormal; FN is the number of abnormal connections that are incorrectly judged as normal.

In this paper, the performance of intrusion detection models is evaluated by five widely used measures: accuracy, precision, detection rate (DR), F1-score, and false alarm rate (FAR) [14, 33, 34]. The calculations of these evaluation measures are defined as follows:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (1)$$

$$DR = \frac{TP}{TP + FN}, \quad (2)$$

$$\text{precision} = \frac{TP}{TP + FP}, \quad (3)$$

$$F1\text{-score} = \frac{2 \times \text{precision} \times DR}{\text{precision} + DR}, \quad (4)$$

$$FAR = \frac{FP}{FP + TN}. \quad (5)$$

The accuracy and detection rates evaluate the capability of an intrusion detection model to correctly predict connections and detect abnormal events, respectively. The precision reflects the proportion of real abnormal events in all returned abnormal connections. The F1-score is the harmonic mean of precision and DR, which balances the precision and DR of a detection model. The false alarm rate measures the ratio of normal connections incorrectly classified as abnormal to total normal connections. Therefore, higher values of accuracy, detection rate, precision, and F1-score and lower values of false alarm rate show better detection performance for network intrusion detection models.

3. The Proposed Method

As mentioned in the Introduction, this work aims to improve the performance of network intrusion detection by constructing a hybrid model. To this end, we propose the *GINI-GBDT-PSO* method that combines the *Gini index*, the *GBDT* algorithm, and the *PSO* method into one framework.

In the proposed model, we use Gini index to select important features from original datasets and then adopt GBDT as a classification approach to detect abnormal events on the feature filtered datasets. Because manually configuring the parameters of GBDT is not an easy work, we use PSO as an optimizer to find the high-quality parameters for GBDT.

For the sake of clarity, the stages of the system framework are listed as follows:

- (1) Dataset preprocessing
- (2) Featuring a selection using the Gini index
- (3) Training a prediction model with GBDT optimized by PSO
- (4) Identifying network intrusions by the prediction model

3.1. NSL-KDD Dataset. In this study, the NSL-KDD dataset [32] is used as benchmark. NSL-KDD is an improved version of the popular KDDCup99 dataset (<http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>), which solves some inherent problems existing in KDDCup99 dataset. Each instance in NSL-KDD dataset is a TCP/IP connection record depicted by 41 different features and classified as one of the following classes: normal event, denial of service (DoS) attack, probe attack, user to root (U2R) attack, and remote to local (R2L) attack. The detail description about the NSL-KDD dataset can be found in [35–37]. Table 2 lists the number of instances in the training and testing sets. As seen in Table 2, the number of instances in the training and testing sets of NSL-KDD is in the reasonable range, which makes it affordable to conduct experiments on whole dataset. However, researchers have usually run experiments on randomly selected small portion of the KDDCup99 dataset, which may cause inconsistent evaluation results. In NSL-KDD dataset, it is free of redundant records in training set and duplicate records in testing set, so the classifiers will not be biased towards more frequent records.

There are three symbolic, two binary, and 36 continuous features among the 41 features in NSL-KDD dataset. To use the proposed method, symbolic features should be converted into numeric features. In this paper, the simple scheme used in [12] is adopted to handle symbolic features. The scheme maps symbolic values to integer values with range from 1 to M , where M is the number of distinct symbols for a feature. For class labels, normal is mapped to 1, DoS to 2, probe to 3, U2R to 4, and R2L to 5. Furthermore, we normalize the values of each feature into [0, 1].

3.2. Gini Index. Usually, the dataset for network intrusion detection contains many features. However, not every feature contributes to the task of detecting intrusion. Feature selection, which can remove redundant or irrelevant features, is a crucial step for intrusion detection. Based on the optimal feature space, we can not only enhance the speed of training a classifier for network intrusion detection but also improve its detection performance.

TABLE 2: Number of instances in NSL-KDD.

	Normal	DoS	Probe	U2R	R2L	Total
Training set	67,343	45,927	11,656	52	995	125,973
Testing set	9711	7458	2421	200	2754	22,544

The goal of feature selection is to get a group of significant features from the whole dataset, such that these selected features are very important for training a classification model.

To do that, we employ in this paper the *Gini index* to undertake the mission of feature selection. The Gini index, which was developed by Corrado Gini, an Italian statistician and sociologist, in 1912, was originally used to measure the statistical dispersion of income distribution across different population sectors. Nowadays, in the research of data mining, Gini index has been widely used in text mining [24–26].

Given a dataset D , which includes instances from l classes $C_1, \dots, C_l$, let f be a feature that has k distinct values in D . According to the values of f , we divide D into k disjoint subsets $D_1, \dots, D_k$. The score of feature f measured by Gini index is define as follows:

$$G(f) = \sum_{i=1}^k \frac{|D_i|}{|D|} \left(1 - \sum_{j=1}^l p_{ij}^2 \right), \quad (6)$$

where p_{ij} is the probability that an instance in D_i belongs to class C_j .

The range of $G(f)$ is [0, 1]. The smaller $G(f)$ is, the more important feature f is. For any D_i , if all instances in it belong to the same class, then $\sum_{j=1}^l p_{ij}^2 = 1$, and $G(f) = 0$. At this time, feature f has the strongest discrimination.

3.3. GBDT Algorithm. The *gradient boosting* is a machine learning technique for classification and regression problems, which produces a strong ensemble model by combining a series of weak prediction models in an iterative fashion. The *GBDT* (gradient boosted decision trees) algorithm is a powerful ensemble learning algorithm, which extends and enhances the classification and regression tree model according to gradient boosting [22]. The GBDT algorithm iteratively constructs decision trees. In each iteration, a decision tree is trained from the residuals of the previous tree. Then, the accumulation of predicted results of all trees is the final result.

Given the training samples $\{(x_i, y_i)\}_{i=1}^n$, where x_i is a sample and y_i denotes the label of sample x_i . Let $F(x)$ be a prediction model and $L(y, F(x))$ be a loss function. For any sample x_i , $F(x_i)$ is the prediction of x_i , and $L(y_i, F(x_i))$ is the lossness between $F(x_i)$ and y_i . The goal of GBDT is to learn an optimal model $F^*(x)$ such that $\sum_{i=1}^n L(y_i, F(x_i))$ is minimized for a specified loss function $L(y, F(x))$.

To do that, the GBDT algorithm first builds an initial decision tree $F_0(x)$, then iteratively constructs m new trees. In each iteration, a new tree $h(x)$ is added to reduce the residuals, which are obtained by the given loss function

Require:Training set: $\{(x_i, y_i)\}_{i=1}^n$; number of iterations: m Loss function: $L(y, F(x))$; learning rate: ν **Ensure:** The predict model: $F(x)$

1. $F_0(x) = 0$;
2. for $t = 1 \rightarrow m$ do
3. $y'_i = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right] F(x) = F_{t-1}(x), \quad i = 1, 2, \dots, n$;
4. Train a new tree $h_t(x)$ according to $\{(x_i, y'_i)\}_{i=1}^n$
5. $\rho_t = \arg \min_{\rho} \sum_{i=1}^n L(y_i, F_{t-1}(x) + \rho * h_t(x))$;
6. $F_t(x) = F_{t-1}(x) + \nu * \rho_t * h_t(x)$;
7. end for
8. return $F_m(x)$;

ALGORITHM 1: GBDT algorithm.

$L(y, F(x))$. Therefore, the optimal model $F^*(x)$ of GBDT can be defined as follows:

$$F^*(x) = F_0(x) + \nu * \sum_{t=1}^m \rho_t * h_t(x), \quad (7)$$

where m is the number of iterations, and $\nu (0 < \nu < 1)$ is the shrinkage parameter that controls the learning rate of GBDT. $h_t(x)$ is the tree trained in the t th iteration and ρ_t is the weight of $h_t(x)$.

The procedure of GBDT algorithm is shown in Algorithm 1. In line 1, tree $F_0(x)$ is initialized. In this paper, we simply set $F_0(x) = 0$. Line 3 calculates the pseudoresiduals, which represent the difference between real value and predicted value. In GBDT, the loss function is applied to obtain residuals at each iteration. Next, a new tree is trained according to the pseudoresiduals in line 4 and the weight of this tree is determined in line 5. Then, the prediction model is updated in line 6.

In GBDT algorithm, the least absolute deviation function, the least square function, and the Huber function are commonly used as the loss function. In this paper, we adopt the least square function that is defined as follows:

$$L(y, F(x)) = \frac{1}{2} (y - F(x))^2. \quad (8)$$

3.4. PSO Algorithm. Inspired by the population behavior of bird flocking and fish schooling, Kennedy and Eberhart proposed the PSO (particle swarm optimization) algorithm in 1995 [38]. This algorithm is a derivative-free, zero-order method and thus can be applied to solve a variety of optimization problems by simulating social behavior [12]. Compared to other optimization algorithms, the PSO technique is easy to implement, robust, and scalable and can quickly find approximately optimal solutions [34, 39]. In the research of IDSs, PSO is a popular optimization technique [12, 34, 40, 41].

At the outset, a group of particles, which represent a population of candidate solutions, is generated with random positions and velocities in the problem space. Then, each particle flies through the problem space by following the particle that is on the best-known position. To define how best a position is, a fitness value is assigned to a particle according to its

position. Thus, the optimization problem becomes to find the best position that has the best fitness value. In this work, we define the detection accuracy (see (1)) as fitness function. To obtain the optimization goal, the PSO algorithm iteratively updates the position and velocity of each particle in the problem space guided by the best position identified so far by itself as well as the best coordinate tracked by entire particle swarm.

Formally, in the n -dimensional problem space, the position and velocity of the i th particle at the t th iteration is defined as $X_i(t) = (x_{i1}(t), x_{i2}(t), \dots, x_{in}(t))$ and $V_i(t) = (v_{i1}(t), v_{i2}(t), \dots, v_{in}(t))$, respectively. Then, the particle changes its position and velocity according to its position and velocity at current stage, the best-known position identified by itself so far (denoted as P_{best}^i), and the global best-known position identified by entire particle swarm so far (denoted as G_{best}). Equations (9) and (10) show the calculation of the velocity and position, respectively.

$$V_i(t) = \omega V_i(t-1) + c_1 r_1 (P_{best}^i - X_i(t-1)) + c_2 r_2 (G_{best} - X_i(t-1)), \quad (9)$$

$$X_i(t) = X_i(t-1) + V_i(t), \quad (10)$$

where ω represents the inertia weight. c_1 is the particle acceleration factor, and c_2 is the population acceleration factor. r_1 and r_2 are two independently positive random numbers between 0 and 1.

The detailed process of the PSO algorithm is described in Algorithm 2. In Algorithm 2, N is the total number of particles in the population, and T is the total number of iterations. $fitness(\cdot)$ is the function that gives the fitness value of a position.

4. Experiments

This section experimentally studies the detection performance of the proposed hybrid model (i.e., GINI-GBDT-PSO), compared with six baselines. The six baselines contain three individual classification algorithms, that is, SVM [42], random forests (RF) [43], and C4.5 [44], and three hybrid models, that is, FC-ANN [14], CFA-DT [19], and IGCR-ANN [3]. The NSL-KDD dataset [32] is the benchmark. All experiments are implemented in the MATLAB 7.0 environment.

4.1. Setting of GINI-GBDT-PSO. For the proposed method, before training an intrusion detection model, the first task is to extract the optimal subset of features. The Gini index was originally used to determine the degree of inequality of income distribution across different population sectors. The international community has usually used 0.4 as the warning line for the gap between rich and poor [45]. In our experiments, we consider a feature whose score of Gini index is less than or equal to 0.4 as important feature. As a result, 18 important features are selected from the NSL-KDD dataset, which are listed in Table 3.


```

1. for each  $i \in [1, N]$  do
2.   Randomly initialize the position  $X_i$  and velocity  $V_i$  of current particle.
3.   Mark the position of current particle as its best-known position,  $P_{\text{best}}^i$ .
4. end for
5. Calculate the global best known position  $G_{\text{best}}$ .
6. for each  $t \in [1, T]$  do
7.   for each  $i \in [1, N]$  do
8.     Update the position and velocity of the  $i$ th particle based on (9) and (10).
9.     if  $\text{fitness}(X_i(t)) > \text{fitness}[P_{\text{best}}^i]$  then
10.       $P_{\text{best}}^i = X_i(t)$ ;
11.     end if
12.     if  $\text{fitness}(P_{\text{best}}^i) > \text{fitness}(G_{\text{best}})$  then
13.       $G_{\text{best}} = P_{\text{best}}^i$ ;
14.     end if
15.   end for
16. end for

```

ALGORITHM 2: PSO algorithm.

TABLE 3: Important features selected from NSL-KDD.

Id	Data type	Feature
f1	Continuous	Duration
f2	Symbolic	protocol_type
f3	Symbolic	Service
f4	Symbolic	Flag
f5	Continuous	src_bytes
f6	Continuous	dst_bytes
f7	Binary	Land
f8	Continuous	wrong_fragment
f12	Binary	logged_in
f23	Continuous	Count
f25	Continuous	serror_rate
f26	Continuous	srv_error_rate
f29	Continuous	same_srv_rate
f32	Continuous	dst_host_count
f33	Continuous	dst_host_srv_count
f34	Continuous	dst_host_same_srv_rate
f38	Continuous	dst_host_serror_rate
f39	Continuous	dst_host_srv_serror_rate

In order to train the best possible prediction model, parameters of the GBDT algorithm are optimized by PSO. However, the PSO algorithm also has several parameters (see (9)). Some papers discussed the values of ω , c_1 , and c_2 [34, 46, 47]. In this paper, we set the values of ω , c_1 , and c_2 as suggested in [47]. The details of parameter setting for PSO are outlined in Table 4. According to the parameter setting of PSO, the optimized parameters for GBDT are obtained, which are presented in Table 5. In Table 5, the last parameter (i.e., minimum number of leaf) is set for the decision trees used in GBDT.

4.2. Results and Discussion. At first, we analyze the performance of SVM, random forest, C4.5, and FC-ANN with

TABLE 4: Parameters for PSO.

Parameter	Value
Number of particles (N)	20
Number of iterations (T)	200
Inertia weight (ω)	0.6
Particle factor (c_1)	0.103
Population factor (c_2)	2.897

TABLE 5: Values of parameters for GBDT.

Parameter	Value
Learning rate (ν)	0.2
Number of iterations (m)	701
Minimum number of leaf	13

TABLE 6: Performance of four methods with different features.

Method	Accuracy	Precision	DR	F1-score	FAR
SVM ₄₁	71.40%	94.51%	55.64%	70.04%	4.27%
SVM ₁₈	73.41%	96.89%	66.28%	78.72%	2.81%
RF ₄₁	75.92%	96.85%	59.64%	73.82%	2.57%
RF ₁₈	76.57%	96.86%	60.83%	74.73%	2.61%
C4.5 ₄₁	80.27%	96.44%	67.84%	79.65%	3.31%
C4.5 ₁₈	80.58%	95.08%	69.49%	80.30%	4.75%
FC-ANN ₄₁	72.56%	97.96%	52.91%	68.71%	1.48%
FC-ANN ₁₈	75.80%	96.95%	59.36%	73.64%	2.47%

respect to different features. These four methods do not integrate feature selection; thus, we run them with all 41 features and the 18 selected features in Table 3, respectively. The results are listed in Table 6. For a method M , M_{41} and M_{18} in Table 6 denote M with the 41 and 18 features, respectively. It is evident from Table 6 that the performance of each method with the 18 features is better than that with all

TABLE 7: Performance comparison of various methods in detection of abnormal connections.

Method	Accuracy	Precision	DR	F1-score	FAR
SVM	73.41%	96.89%	66.28%	78.72%	2.81%
RF	76.57%	96.86%	60.83%	74.73%	2.61%
C4.5	80.58%	95.08%	69.49%	80.30%	4.75%
FC-ANN	75.80%	96.95%	59.36%	73.64%	2.47%
CFA-DT	72.24%	93.42%	73.58%	82.32%	3.45%
IGCR-ANN	77.83%	96.85%	63.10%	76.41%	2.71%
GINI-GBDT-PSO	86.10%	96.44%	78.48%	86.54%	3.83%

features. These results are easy to understand, since the 18 features have more discrimination than others. In the following of this part, the results of these four methods are all with the 18 features.

Next, the detection performance of the proposed method in comparison with the six competing methods are evaluated. Table 7 exhibits the overall performance of different intrusion detection approaches in terms of various criteria. The best result for each criterion is highlighted in boldface. Clearly, our model yields the highest accuracy, detection rate, and F1-score, which markedly outperforms the counterparts. It can be seen from Table 7 that all methods achieve very high precision. Although the precision of FC-ANN is the highest, its detection rate is the lowest. That makes its F1-score is only 73.64%, which is very low compared with others. For the false alarm rate, the proposed method does not perform well. In the future, we need to further reduce the false alarm rate of the proposed method.

After introducing the overall performance of seven intrusion detection models, then we present the detection performance of these models for four attack types in Figure 1. We can observe from Figure 1(a) that the proposed method outperforms all compared methods in terms of detection rate for all attack types. For the two low-frequency attack types, that is, R2L and U2R, the corresponding values of the proposed method are 34.44% and 14.66%, respectively. Whereas, all compared methods, especially SVM and FC-ANN, exhibit poor performance when meeting U2R and R2L attacks. The proportions of samples of these two attacks in training set are very low (see Table 2) which is a reason that gives rise to this scenario. However, the main reason is that the capability of these models to detect low frequency attack is insufficient. Figure 1(b) reports the precision of each detection model for four attack types. In Figure 1(b), the proposed method gets the highest precision for R2L attack; however, its precision in detection of DoS, probe, and U2R attacks is not good enough. The RF method yields the highest precision for DoS and probe attacks, but its precision for U2R and R2L is very low. To fairly evaluate the capabilities of all methods for individual attack type, Figure 1(c) depicts the F1-scores obtained by all methods. Figure 1(c) shows that the proposed model performs the best for DoS, U2R, and R2L attacks and gets the approximate best F1-score for probe attack. RF shows the best for probe attack and the second

best for Dos attack, but its F1-scores about U2R and R2L attack are very low.

In Table 8, we show the harmonic mean of each measure for different attack types. Since some values are 0 (see Figure 1), we add a small constant, 0.01, to each value when computing harmonic mean. It is clear from Table 8 that the proposed method ranks first in terms of both DR and F1-score and third with respect to precision. For DR and F1-score, our method outnumbers the compared methods. Although, the precision of our method is lower than that of C4.5 and CFA-DT, it is far better than that of the other four.

Finally, we compared our method with the 18 selected features as well as all 41 features. The result in detection of abnormal connections is described in Figure 2. Clearly, the performance with the 18 features is better than that with 41 features. This scenario indicates that feature selection is an important technique in intrusion detection. Consequently, the combination of Gini index, GBDT algorithm and PSO technique are a competitive detection model.

In summary, the above experimental results demonstrate that our hybrid model can give overall better performance than single classification methods (i.e., SVM, C4.5, and RF) and some hybrid models (i.e., FC-ANN, CFA-DT, and IGCR-ANN) when detecting network intrusion. Specifically, for the low frequency attacks, our model substantially outnumbers the others. Consequently, the proposed model has a strong applicability for network intrusion detection.

5. A Brief Review of Hybrid Models

Recent years have witnessed a surge of research efforts about the application of hybrid techniques for intrusion detection. This section briefly lists some recent studies.

Abuomman and Reaz [12] developed an ensemble method based on the SVM and k -nearest neighbors (k -NN) classifiers, as well as the PSO algorithm to enhance the detection accuracy. Their approach trains six k -NN experts with different k s and six SVM experts with different parameters on the same dataset and then generates two new ensembles by combining the opinions of 12 experts using the PSO and metaoptimized PSO, respectively. The difference of both ensembles is in the setting of the PSO behavioral parameters. In the first ensemble, the behavioral parameters are manually selected, and in the second one, the behavioral parameters are optimized by using local unimodal sampling (LUS) [48]. Additionally, they also gave an ensemble method by weighing each expert using the weighted majority algorithm (WMA) [49], for the purpose of comparison. The experimental results on five randomly selected subsets from the KDDCup99 dataset show that the LUS-based ensemble achieves the best accuracy.

De la Hoz et al. [18] applied statistical techniques and the self-organizing map (SOM) approach to detect network anomalies. Two machine learning techniques, principal component analysis (PCA) and Fisher discriminant ratio (FDR), were considered in their work to remove irrelevant features and noises. Then, the probabilistic SOM (PSOM), a fuzzy version of the classical SOM, is used to model and classify the filtered feature space. Experiments run on the

FIGURE 1: Performance comparison of various methods for four attack types.

TABLE 8: Harmonic mean of each measure for four attack types.

Method	DR	Precision	F1-score
SVM	0.02%	0.02%	0.02%
RF	0.61%	0.64%	0.61%
C4.5	13.80%	46.94%	21.27%
FC-ANN	0.02%	3.01%	0.02%
CFA-DT	8.19%	32.42%	13.06%
IGCR-ANN	0.04%	0.04%	0.04%
GINI-GBDT-PSO	33.18%	22.18%	26.59%

FIGURE 2: Performance comparison of the proposed methods with different features.

NSL-KDD dataset show that the PCA + FDR + PSOM model has the best performance compared with the contrasted hybrid models.

De la Hoz et al. [17] also introduced another hybrid model into the network intrusion detection problem with the purpose of diminishing model complexity and promoting detection performance. In that paper, feature selection is

carried out by a multiobjective optimization, and both anomaly detection and attack classification are fulfilled by the growing hierarchical self-organizing maps (GHSOMs) [50]. The multiobjective approach is implemented by the NSGA-II algorithm [51] in which Jaccard coefficient is employed to measure the performance of a classifier for the corresponding class. Experimental results on the NSL-KDD dataset illustrate that their proposed model is better than other approaches in terms of detection accuracy and detection rate.

Wang et al. [14] proposed a hybrid model by uniting the fuzzy clustering and ANN. First, their model partitions the training set into several different groups using fuzzy clustering technique. Then, different ANN models are trained based on different groups. Finally, a fuzzy aggregation module is generated by aggregating all ANN models' results. Experimental results on the KDDCup99 dataset indicate that this model outperforms the decision trees, naive Bayes, and backward propagation neural network (BPNN).

Eesa et al. [19] presented a hybrid method, which combines the cuttlefish optimization algorithm (CFA) and the decision tree classifier, to detect network intrusions. In their model, the CFA is used to select significant attributes, and the decision tree algorithm is utilized to identify types of abnormal events. The performance of this hybrid model was evaluated on the KDDCup99 dataset. The results demonstrate when the number of attributes is less than 25, the detection accuracy and detection rate have obvious improvement.

Guo et al. [2] proposed a two-level hybrid approach for intrusion detection by exploiting the strengths of misuse-based and anomaly-based detection methods. This approach is composed of two anomaly detection components (ADCs) and one misuse detection component (MDC). In stage 1, the ADC 1 detects abnormal connections using the ADBCC method [2]. Then, the declared abnormal and normal connections are, respectively, sent to the ADC 2 and the MDC in parallel to further evaluated by k -NN. This hybrid approach was experimentally tested on the KDDCup99 and the Kyoto University Benchmark dataset. The results show that this approach is effective in detection abnormal connections with a low false alarm rate.

Akashdeep et al. [3] developed an intelligent system by means of feature selection and ANN classifier. The proposed system first ranks features according to information gain and correlation and then selects useful features by combining their ranks using a novel approach. Experimental results on the KDDCup99 dataset illustrate that the system is really encouraging.

6. Conclusion

Motivated by the successes of Gini index and GBDT algorithm in other fields, this paper proposed the GINI-GBDT-PSO method, a novel hybrid intrusion detection model to improve the performance of network intrusion detection systems. The proposed model first extracts the optimal subset of features from whole dataset by using the Gini index. Then, the GBDT algorithm, which is a gradient boosting approach, is adopted to detect abnormal connections. In addition, the PSO algorithm is employed to optimize parameters of the GBDT algorithm in the proposed model. This model can not only enhance the overall performance for network intrusion detection effectively but also improve the detection performance for each type of attack.

In order to validate the performance of our method, we performed experiments on the NSL-KDD dataset compared with six baselines. Five evaluation criteria are introduced to conduct fair comparisons, which are accuracy, detection rate, precision, F1-score, and false alarm rate. The experimental results demonstrated that the proposed model performs the best on the whole in comparison with baselines. These results indicate that the proposed model is an accurate and effective solution for network intrusion detection systems.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (no. 61602225) and the Fundamental Research Funds for the Central Universities (no. lzujbky-2017-192).

References

- [1] O. Depren, M. Topallar, E. Anarim, and M. K. Ciliz, "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks," *Expert Systems with Applications*, vol. 29, no. 4, pp. 713–722, 2005.
- [2] C. Guo, Y. Ping, N. Liu, and S.-S. Luo, "A two-level hybrid approach for intrusion detection," *Neurocomputing*, vol. 214, pp. 391–400, 2016.
- [3] Akashdeep, I. Manzoor, and N. Kumar, "A feature reduced intrusion detection system using ANN classifier," *Expert Systems with Applications*, vol. 88, pp. 249–257, 2017.
- [4] S.-W. Lin, K.-C. Ying, C.-Y. Lee, and Z.-J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," *Applied Soft Computing*, vol. 12, no. 10, pp. 3285–3290, 2012.
- [5] M. H. Bhuyan, D. Bhattacharyya, and J. Kalita, "Network anomaly detection: methods, systems and tools," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014.
- [6] K. Ilgun, R. A. Kemmerer, and P. A. Porras, "State transition analysis: a rule-based intrusion detection approach," *IEEE Transactions on Software Engineering*, vol. 21, no. 3, pp. 181–199, 1995.
- [7] D. E. Denning, "An intrusion-detection model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, 1987.
- [8] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: a review," *Expert Systems with Applications*, vol. 36, no. 10, pp. 11994–12000, 2009.
- [9] S. Peddabachigari, A. Abraham, C. Grosan, and J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems," *Journal of Network and Computer Applications*, vol. 30, no. 1, pp. 114–132, 2007.
- [10] F. Kuang, W. Xu, and S. Zhang, "A novel hybrid KPCA and SVM with GA model for intrusion detection," *Applied Soft Computing*, vol. 18, pp. 178–184, 2014.
- [11] G. Kim, S. Lee, and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Systems with Applications*, vol. 41, no. 4, Part 2, pp. 1690–1700, 2014.
- [12] A. A. Abuomman and M. B. I. Reaz, "A novel SVM-kNN-PSO ensemble method for intrusion detection system," *Applied Soft Computing*, vol. 38, pp. 360–372, 2016.
- [13] L. Khan, M. Awad, and B. Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering," *The VLDB Journal*, vol. 16, no. 4, pp. 507–521, 2007.
- [14] G. Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using artificial neural networks and fuzzy clustering," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6225–6232, 2010.
- [15] C. Xiang, P. C. Yong, and L. S. Meng, "Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees," *Pattern Recognition Letters*, vol. 29, no. 7, pp. 918–924, 2008.
- [16] W.-C. Lin, S.-W. Ke, and C.-F. Tsai, "CANN: an intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-Based Systems*, vol. 78, pp. 13–21, 2015.
- [17] E. De la Hoz, E. De la Hoz, A. Ortiz, J. Ortega, and A. Martínez-Alvarez, "Feature selection by multi-objective optimisation: application to network anomaly detection by hierarchical self-organising maps," *Knowledge-Based Systems*, vol. 71, pp. 322–338, 2014.
- [18] E. De la Hoz, E. De la Hoz, A. Ortiz, J. Ortega, and B. Prieto, "PCA filtering and probabilistic SOM for network intrusion detection," *Neurocomputing*, vol. 164, pp. 71–81, 2015.
- [19] A. S. Eesa, Z. Orman, and A. M. A. Brifcani, "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2670–2679, 2015.
- [20] Y. Chen, A. Abraham, and B. Yang, "Hybrid flexible neural-tree-based intrusion detection systems," *International Journal of Intelligent Systems*, vol. 22, no. 4, pp. 337–352, 2007.

- [21] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," *Information Sciences*, vol. 177, no. 18, pp. 3799–3821, 2007.
- [22] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
- [23] R. C. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Micro Machine and Human Science, 1995. MHS '95, Proceedings of the Sixth International Symposium on*, pp. 39–43, Nagoya, Japan, 1995.
- [24] N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," *Expert Systems with Applications*, vol. 39, no. 5, pp. 4760–4768, 2012.
- [25] C. C. Aggarwal, Y. Zhao, and P. S. Yu, "On the use of side information for mining text data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 6, pp. 1415–1429, 2014.
- [26] W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang, "A novel feature selection algorithm for text categorization," *Expert Systems with Applications*, vol. 33, no. 1, pp. 1–5, 2007.
- [27] K. B. Stevens and D. U. Pfeiffer, "Spatial modelling of disease using data- and knowledge-driven approaches," *Spatial and Spatio-temporal Epidemiology*, vol. 2, no. 3, pp. 125–133, 2011.
- [28] Y. L. Cheong, P. J. L. ao, and T. Lakes, "Assessment of land use factors associated with dengue cases in Malaysia using boosted regression trees," *Spatial and Spatio-temporal Epidemiology*, vol. 10, pp. 75–84, 2014.
- [29] A. Mohan, Z. Chen, and K. Weinberger, "Web-search ranking with initialized gradient boosted regression trees," *Journal of Machine Learning Research*, vol. 14, pp. 77–89, 2011.
- [30] S. Tyree, K. Q. Weinberger, K. Agrawal, and J. Paykin, "Parallel boosted regression trees for web search ranking," in *Proceedings of the 20th International Conference on World Wide Web*, pp. 387–396, New York, NY, USA, 2011.
- [31] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, Part B, pp. 308–324, 2015.
- [32] M. Tavallaei, E. Bagheri, W. Lu, and A.-A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Ottawa, ON, Canada, 2009.
- [33] R. Singh, H. Kumar, and R. K. Singla, "An intrusion detection system using network traffic profiling and online sequential extreme learning machine," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8609–8624, 2015.
- [34] S. M. H. Bamakan, H. Wang, T. Yingjie, and Y. Shi, "An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization," *Neurocomputing*, vol. 199, pp. 90–102, 2016.
- [35] L. Dhanabal and S. P. Shantharajah, "A study on NSL-KDD dataset for intrusion detection system based on classification algorithms," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 4, no. 6, pp. 446–452, 2015.
- [36] F. Iglesias and T. Zseby, "Analysis of network traffic features for anomaly detection," *Machine Learning*, vol. 101, no. 1–3, pp. 59–84, 2015.
- [37] R. A. R. Ashfaq, X.-Z. Wang, J. Z. Huang, H. Abbas, and Y.-L. He, "Fuzziness based semi-supervised learning approach for intrusion detection system," *Information Sciences*, vol. 378, pp. 484–497, 2017.
- [38] J. Kennedy and R. C. Eberhart, "Particle swarm optimization," in *Proceedings of IEEE International Conference on Neural Networks*, vol. 4, pp. 1942–1948, Perth, WA, Australia, 1995.
- [39] S. X. Wu and W. Banzhaf, "The use of computational intelligence in intrusion detection systems: a review," *Applied Soft Computing*, vol. 10, no. 1, pp. 1–35, 2010.
- [40] A. Karami and M. Guerrero-Zapata, "A hybrid multiobjective RBF-PSO method for mitigating DoS attacks in named data networking," *Neurocomputing*, vol. 151, Part 3, pp. 1262–1282, 2015.
- [41] A. Karami and M. Guerrero-Zapata, "A fuzzy anomaly detection system based on hybrid PSO-Kmeans algorithm in content-centric networks," *Neurocomputing*, vol. 149, Part C, pp. 1253–1269, 2015.
- [42] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.
- [43] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [44] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, CA, USA, 1993.
- [45] Y. Tao, X. Wu, and C. Li, "Rawls' fairness, income distribution and alarming level of Gini coefficient," 2014, <http://arxiv.org/abs/1409.3979>.
- [46] R. C. Eberhart and Y. Shi, "Comparing inertia weights and constriction factors in particle swarm optimization," in *Proceedings of the 2000 Congress on Evolutionary Computation*, vol. 2, pp. 84–88, La Jolla, CA, USA, 2000.
- [47] N. R. Samal, A. Konar, S. Das, and A. Abraham, "A closed loop stability analysis and parameter selection of the particle swarm optimization dynamics for faster convergence," in *2007 IEEE Congress on Evolutionary Computation*, pp. 1769–1776, Singapore, Singapore, 2007.
- [48] M. E. H. Pedersen and A. J. Chipperfield, "Local unimodal sampling," Tech. Rep. HL0801, Hvass Laboratories, 2008.
- [49] N. Littlestone and M. K. Warmuth, "The weighted majority algorithm," *Information and Computation*, vol. 108, no. 2, pp. 212–261, 1994.
- [50] A. Rauber, D. Merkl, and M. Dittenbach, "The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data," *IEEE Transactions on Neural Networks*, vol. 13, no. 6, pp. 1331–1341, 2002.
- [51] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.