Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

A novel approach to intrusion detection using SVM ensemble with feature augmentation

Jie Gu[a,c], Lihong Wang[d], Huiwen Wang[a,b], Shanshan Wang[a,c,*]

[a] School of Economics and Management, Beihang University, Beijing 100191, China
[b] Beijing Advanced Innovation Center for Big Data and Brain Computing (BDBC), Beijing 100191, China
[c] Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations, Beijing 100191, China
[d] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China

ARTICLE INFO

Article history:
Received 18 January 2019
Revised 26 April 2019
Accepted 29 May 2019
Available online 2 June 2019

Keywords:
Ensemble learning
Intrusion detection
Marginal density ratios transformation
Network security
Support vector machine

ABSTRACT

Network security has been a very important problem. Intrusion detection systems have been widely used to protect network security. Various machine learning techniques have been applied to improve the performance of intrusion detection systems, among which ensemble learning has received a growing interest and is considered as an effective method. Besides, the quality of training data is also an essential determinant that can greatly enhance the detection capability. Knowing that the marginal density ratios are the most powerful univariate classifiers. In this paper, we propose an effective intrusion detection framework based on SVM ensemble with feature augmentation. Specifically, the logarithm marginal density ratios transformation is implemented on the original features with the goal of obtaining new and better-quality transformed training data; SVM ensemble was then used to build the intrusion detection model. Experiment results show that our proposed method can achieve a good and robust performance, which possesses huge competitive advantages when compared to other existing methods in terms of accuracy, detection rate, false alarm rate and training speed.

1. Introduction

Nowadays, computer networks and the Internet have become an essential component of modern society (Tsai et al., 2009). Our daily activities are becoming more and more dependent on the networks where considerable information related to organizations and people is also stored at the same time. When been compromised, it may cause huge losses. Network security is becoming more and more important and has attracted increasing attention (Gan et al., 2013). Many measures are taken to protect networks from intrusions and attacks, including firewall, data encryption, intrusion detection system and other techniques.

Among these, intrusion detection system (IDS) which aims to classify user's activity into normal or intrusion-related behavior based on rules or models has received considerable attentions (Luo and Xia, 2014; Tjhai et al., 2010). Although promising improvements in IDSs detection approaches have been achieved, the intrusion detection is an ongoing research area where many problems need to be further solved, such as high dimensionality, huge volume, constantly changes in environments and real-time detection (Bamakan et al., 2016). In this paper, our main goal is to propose an effective

[*] Corresponding author at: School of Economics and Management, Beihang University, Beijing 100191, China.
E-mail address: sswang@buaa.edu.cn (S. Wang).

https://doi.org/10.1016/j.cose.2019.05.022

intrusion detection framework with an accurate and robust performance, as well as a fast training speed.

The review of the related works in intrusion detection domain indicates that learning algorithm and data quality are two main determinants of the performance of intrusion detection method (Aburomman and Reaz, 2017; Wang et al., 2017). In this paper, we propose a new and effective intrusion detection framework (DT-EnSVM) combining both the ensemble learning and data transformation technique. Specifically, we first apply ratio transformations to obtain a new and high-qualified training data, then SVM was chosen to train the base learners, since it has been proven to be an effective method,and finally a nonlinear combination method was used to aggregate these SVM classifiers to construct an ensemble-based intrusion detection model. The empirical results on NSL-KDD dataset and the other two datasets in intrusion detection domain demonstrate that the proposed method possesses competitive advantages compared with the existing methods in terms of accuracy, detection rate, false alarm rate and training speed.

Compared with the existing literatures on IDS, the main contributions of this paper are four folded. First, combination of data quality-improvement technique and ensemble learning which can obtain an accurate and robust performance. Second, comparison between classifier ensemble and base classifier with respect to detection accuracy. Third, Fuzzy c-means was used to enhance the effectiveness of an ensemble by maximizing the differences among training datasets for base learners as much as possible. Fourth, the proposed methodology is more flexible, and can be applied to a variety of practical applications where the classification of some groups of objects such as spam detection, and credit scoring are of interested.

The remainder of this paper is organized as follows. Related works are presented in Section 2. In Section 3, we briefly review the ensemble learning and the logarithm marginal density ratios transformation. Section 4 describes the details of the proposed intrusion detection framework (i.e., DT-EnSVM). Section 5 presents the experiment settings, the results, and a discussion of comparisons of our proposed method with others. Finally, Section 6 comes to conclusion remarks and future works.

2. Related works

Various researches on IDS have been investigated (Liao et al., 2013; Tsai et al., 2009). Many researchers have long considered intrusion detection as a classification problem (Amini et al., 2016; Bamakan et al., 2016; Gan et al., 2013; Luo and Xia, 2014), where they aim to classify the incoming sample data into normal or abnormal category. Various intrusion detection methodologies have been proposed, including support vector machine (SVM) (Bamakan et al., 2016; Feng et al., 2014; Horng et al., 2011; Li et al., 2012; Raman et al., 2017), Decision Tree (Eesa et al., 2015; Kim et al., 2014), K-nearest neighbor (Liao and Vemuri, 2002; Lin et al., 2015), Naive Bayes (Koc et al., 2012; Mukherjee and Sharma, 2012), Artificial Neural Network (Manzoor et al., 2017; Wang et al., 2010), Self-organizing map (Ippoliti and Zhou, 2012; Lee et al., 2011), and so on.

Among the aforementioned approaches, researches have shown that classifier ensemble or multiple classifier systems can obtain better performance in comparison with single classifier (Tama, 2017). Recently, ensemble learning receives a growing interest in improving the quality of a single machine learning technique (Farahani et al., 2018; Wang et al., 2011). In contrast to ordinary machine learning approaches that try to learn only one hypothesis from training data, ensemble learning attempts to construct a set of hypotheses and then aggregates them to make a final decision (Kittler et al., 1998; Tao and Zeng-lin, 2012). An ensemble is usually composed of multiple learners which are called base learners. Most importantly, the generalization ability of an ensemble is usually much stronger than a single learning method, which makes ensemble learning more attractive. Moreover, ensemble methods have shown to achieve desirable performances compared with the single classifier systems for a wide range of applications (Amini et al., 2016; Nanni and Lumini, 2009). Accordingly, ensemble learning can be considered as an effective approach to intrusion detection.

Furthermore, the quality of input data is another important determinant. Intrusion detection systems generally confront with data at large scale and high dimensions. They typically suffer from high training complexity. Therefore, it is of great importance to perform data transformation or data reconstruction on the original data to obtain a high-qualified training data for improving detection accuracy and increasing training speed. Wang et al. (2017) employed ratio transformations to improve the data quality. Experiment results on NSL-KDD dataset show that ratio transformations can greatly boost the detection accuracy. Therefore, we are encouraged to adopt ratio transformations to enhance detection performance.

As far as we know, only Wang et al. (2017) made an attempt to combine ratio transformations with a single SVM model. Nevertheless, our currently proposed method distinguishes from it in several aspects. First, the scheme to build the intrusion detection model is quite different. There is only one single detection model in Wang's method, whereas our proposed method here is an ensemble-based model that several detection models are constructed. In this sense, Wang's method can be seen as a special case of our current proposed framework. Second, our proposed method has a good capability to deal with sophisticated scenarios. Generally, heterogeneities among data samples are usually an essential characteristic of data, especially in data streams, such as the concept drift problem. However, in Wang's method, the data was used as a whole to train detection model, which does not take this heterogeneity into consideration. In contrast, our proposed method considers this important characteristic and use it to boost detection accuracy. Thirdly, our proposed method is more suitable to confront with large-scale data. The method proposed in Wang et al. (2017) may suffer from low efficiency when deals with large-scale data. In this paper, however, our proposed method is designed to increase the efficiency for large-scale data, and then can well solve this problem.

3. Preliminary

To better illustrate the proposed intrusion detection framework, we first briefly review the main principles of ensemble

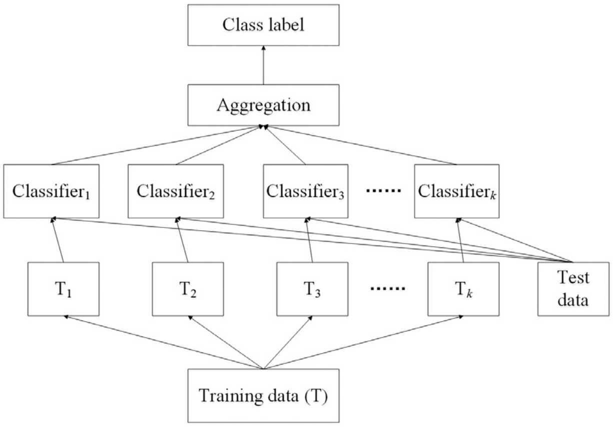


Fig. 1 – The detailed process of classifier ensemble.

learning in Section 3.1 and as well as that of ratio transformations in Section 3.2.

3.1. *Ensemble learning*

Ensemble learning is a machine learning paradigm where multiple base classifiers are trained and then aggregated to construct the final classifier (Nanni and Lumini, 2009; Wang et al., 2011). The generalization ability of an ensemble is usually much stronger than a single base classifier (Ghodselahi, 2011; Kim et al., 2003). Furthermore, Hansen and Salamon (1990) had proven that an ensemble can achieve a better performance than base classifiers when there are significant differences between base classifiers.

In practice, to achieve a good ensemble, we should maximize the differences among base classifiers as much as possible. This requirement can be met by using different training data for each base classifier.

The detailed process of a classifier ensemble is illustrated in Fig. 1. There are three main steps in constructing an ensemble:

- **Step 1**. Generate heterogeneous training data. The key factor in an ensemble is to maximize the differences among base classifiers, which can be achieved by generating different training data for different individuals.
- **Step 2**. Aggregate base classifiers. After training multiple base classifiers, we need to aggregate them to construct the final classifier under a given strategy.
- **Step 3**. Test the ensemble. For a testing sample, introduce it to these base classifiers and aggregate the outputs using the strategy selected in Step 2 to obtain the final classification result.

Note that SVM has been proven to be an effective method for classification problem, since it obeys the principle of structure of risk minimization to maximize the generalization ability. In addition, solving an SVM is essentially a convex optimization problem, the obtained locally optimal solution must be a globally one. Therefore, SVM is chosen here to train the base classifiers.

3.2. *Ratio transformations: data quality-improvement technique*

Knowing that, for univariate classification problems, the marginal density ratio is the best classifier, and the decision boundary is,

$$\{x : f(x)/g(x) = 1\} = \{x : \log f(x) - \log g(x) = 0\}, \tag{1}$$

where g and f denote the class conditional densities for class 0 and class 1, respectively.

Therefore, if we use the marginal density ratio as a new feature to replace the original one, then solving the original classification problem turns out to find a threshold value, which is more accurate and efficient.

According to the basic independence assumption of the Naive Bayes model, we can generalize the univariate case to multivariate situation, that is, we first calculate the marginal density ratios for each feature and then use them as new features in multivariate classifier.

Specifically, following Fan et al. (2016), suppose we have n pairs of observations $S = (\mathbf{X}_i, Y_i), i = 1, 2, \ldots, n$, where $\mathbf{X}_i \in \mathbb{R}^p$ denotes the features and $Y_i \in \{0, 1\}$ denotes the corresponding binary response. Denote by $g_1(x_1), \ldots, g_p(x_p)$ and $f_1(x_1), \ldots, f_p(x_p)$ the class conditional densities for class 0 and class 1, respectively, with x_j being the jth original feature space for $j = 1, 2, \ldots, p$, that is, $(X_j \mid Y = 0) \sim g_j(x_j)$ and $(X_j \mid Y = 1) \sim f_j(x_j)$. The **term** $\log \frac{f_j(x_j)}{g_j(x_j)}$ has been called the logarithm marginal density ratio transformation (ratio transformation for short) for the jth feature for $j = 1, 2, \ldots, p$.

Note the marginal densities $f_j(x_j)$ and $g_j(x_j)$ for jth feature are unknown, we need to estimate them first. The process of data quality-improved technique are shown as follows:

- Step 1. Data split. Randomly split S into two mutually exclusive parts, denoted by $S_1 = (\mathbf{X}^{(1)}, Y^{(1)})$, $S_2 = (\mathbf{X}^{(2)}, Y^{(2)})$, and let N_1 and N_2 be the number of samples in S_1 and S_2, respectively, which satisfy $S_1 \cap S_2 = \varnothing$, $S_1 \cup S_2 = S$ and $N_1 + N_2 = N$.
- Step 2. Kernel estimation of class conditional densities. Apply kernel density estimation to S_1 for class conditional densities and denote the estimates by $\hat{g}_j(x_j)$ and $\hat{f}_j(x_j)$ for $j = 1, 2, \ldots, p$. Specifically, let $\mathbf{X}^{1+}$ denote the set of corresponding $\mathbf{X}_i^{(1)}$ of $Y_i^{(1)} = 1$ for $i = 1, 2, \ldots, N_1$ in S_1, that is, $\mathbf{X}^{1+} = \{\mathbf{X}_i^{(1)} \mid Y_i^{(1)} = 1, i = 1, 2, \ldots, N_1\}$, and let $\mathbf{X}^{1-}$ denote the set of the corresponding $\mathbf{X}_i^{(1)}$ of $Y_i^{(1)} = 0$ for $i = 1, 2, \ldots, N_1$ in S_1, namely, $\mathbf{X}^{1-} = \{\mathbf{X}_i^{(1)} \mid Y_i^{(1)} = 0, i = 1, 2, \ldots, N_1\}$ that satisfy $\mathbf{X}^{1+} \cap \mathbf{X}^{1-} = \varnothing$, $\mathbf{X}^{1+} \cup \mathbf{X}^{1-} = \mathbf{X}^{(1)}$. The density estimates $\hat{g}_j(\cdot)$ and $\hat{f}_j(\cdot)$ are based on samples $\{\mathbf{X}_1^{1+}, \mathbf{X}_2^{1+}, \ldots, \mathbf{X}_{N_1^+}^{1+}\}$ and $\{\mathbf{X}_1^{1-}, \mathbf{X}_2^{1-}, \ldots, \mathbf{X}_{N_1^-}^{1-}\}$, respectively, where N_1^+ and N_1^- are the number of samples in $\mathbf{X}^{1+}$ and $\mathbf{X}^{1-}$, respectively, and satisfy $N_1^+ + N_1^- = N_1$. These concepts can be notated as follows:

$$\hat{g}_j(x) = \frac{1}{N_1^- h} \sum_{i=1}^{N_1^-} K\left(\frac{X_{ij}^{1-} - x}{h}\right), \tag{2}$$

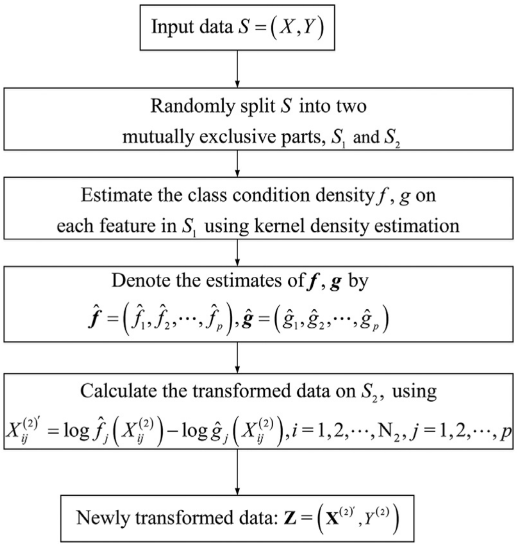

Input data $S=(X,Y)$
Randomly split S into two mutually exclusive parts, S_1 and S_2
Estimate the class condition density f, g on each feature in S_1 using kernel density estimation
Denote the estimates of $\boldsymbol{f}$, $\boldsymbol{g}$ by $\hat{\boldsymbol{f}}=(\hat{f}_1,\hat{f}_2,\cdots,\hat{f}_p)$, $\hat{\boldsymbol{g}}=(\hat{g}_1,\hat{g}_2,\cdots,\hat{g}_p)$
Calculate the transformed data on S_2, using $X_{ij}^{(2)'}=\log\hat{f}_j(X_{ij}^{(2)})-\log\hat{g}_j(X_{ij}^{(2)}), i=1,2,\cdots,N_2, j=1,2,\cdots,p$
Newly transformed data: $\mathbf{Z}=(\mathbf{X}^{(2)'},Y^{(2)})$

Fig. 2 – Detailed procedures of density ratio transformation.

and

$$\hat{f}_j(x) = \frac{1}{N_1^+ h}\sum_{i=1}^{N_1^+} K\left(\frac{X_{ij}^{1+} - x}{h}\right), \tag{3}$$

for $j = 1, 2, \ldots, p$, where $K(\cdot)$ is a kernel function and h is the bandwidth.

- Step 3. Data transformation. Calculate the transformed observations on S_2,

$$\mathbf{X}_i^{(2)'} = \log(\hat{\mathbf{f}}(\mathbf{X}_i^{(2)})) - \log(\hat{\mathbf{g}}(\mathbf{X}_i^{(2)})), \tag{4}$$

where $X_{ij}^{(2)'} = \log \hat{f}_j(X_{ij}^{(2)}) - \log \hat{g}_j(X_{ij}^{(2)})$, $i \in S_2$, $j = 1, 2, \ldots, p$. Finally, the newly transformed data is obtained and denoted as $\mathbf{Z} = (\mathbf{X}^{(2)'}, Y^{(2)})$

To better clarify the ratio transformations process, we display the specific steps in Fig. 2. Ratio transformations make an attempt to combine these powerful marginal density ratios and use them as new features, which can fully and sufficiently exploits the classification information contained in the original data. As a result, the newly transformed data are more concise and the quality of original data can be greatly improved.

4. Proposed framework for intrusion detection

In this section, we will list the detailed procedures in the DT-EnSVM. By integrating the powerful quality-improved transformation with an SVM ensemble, we can build a powerful intrusion detection model with a high accuracy, a low training complexity and a robust performance. More specifically, the quality-improvement technique is used to reconstruct the original features to provide high-qualified and concise training data. Then, an SVM ensemble classifier is trained using the newly transformed data, and finally, the intrusion detection model is built.

The proposed intrusion detection framework mainly consists of three parts, that is, data split, new data formation, ensemble-based intrusion detection model. Each part of our proposed framework is briefly described in the following subsections.

4.1. *Data split*

As informed previously, the key factor in an ensemble is to generate different datasets to train different base classifiers. Fuzzy-c means (FCM) clustering is a modified clustering algorithm that allows a piece of data to belong to two or more clusters, making it more suitable to practical scenario and having been frequently used in pattern recognition (Ghodselahi, 2011). Here, FCM clustering is chosen here to generate heterogeneous training data on the original dataset.

Data split is an important step in our proposed intrusion detection framework, maximizing the differences among base classifiers; and then determining the performance of the ensemble.

4.2. *New data formation*

After generating different training data, we first conduct the data transformation using ratio transformations on the original data to obtain high-qualified transformed data before building the intrusion detection model.

Forming new data using quality-improvement technique is the most important part in our proposed detection framework, since it can not only improve the accuracy of the detection model, but also can reduce the training time, which is of great importance in real-time intrusion detection systems.

4.3. *Intrusion detection model: DT-EnSVM*

After data transformations, the new high-qualified data is used to train different base SVM classifiers. According to Kim et al. (2003) and Aburomman and Reaz (2017), in this paper, we chose a nonlinear combination method to aggregate these SVM classifiers. Specifically, the outputs of these base SVM classifiers are fed into another SVM to train the final detection model, resulting in a double-layer hierarchical SVM ensemble intrusion detection model. Denote by $f_{svm_i}(x)$, $i = 1, 2, \ldots, k$ the decision functions of base SVM classifiers in the lower layer, by $f_{svm_{final}}(x)$ that of the decision function of SVM classifier in the upper layer, and by $f_{final}(x)$ that of the decision function of SVM ensemble; thereby,

$$f_{final}(x) = f_{svm_{final}}(f_{svm_1}(x), f_{svm_2}(x), \ldots, f_{svm_k}(x)) \tag{5}$$

For clarity, the specific procedures of DT-EnSVM are illustrated in Fig. 3. As shown, it mainly consists of four parts: data split, data transformation, intrusion detection model building, intrusion detection.

- **Step 1**. *Data split*.
  Perform fuzzy c-means clustering on the original data to obtain k different subsets to train base classifiers.

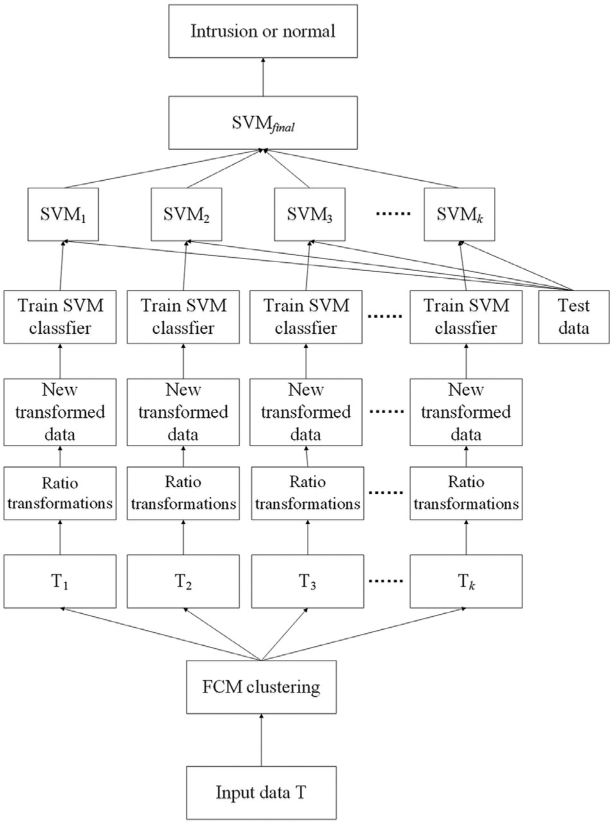


Fig. 3 – Framework of the proposed intrusion detection model.

- **Step 2**. *Data transformation.*
  Perform ratio transformations on the k subsets generated in Step 1 to obtain high-quality transformed data respectively.
- **Step 3**. *Intrusion detection model building.*
  Use the newly transformed data in Step 2 to train k different SVM classifiers in the lower layer and feed their outputs (a matrix with dimension $N \times k$, where N is the number of samples in original data, k is the number of base SVM classifiers) to the upper layer SVM and finally build the intrusion detection model.
- **Step 4**. *Intrusion detection.*
  Introduce a new testing sample to the detection model built in Step 3 to classify it as either an intrusion of a normal behavior.

Remark 1. It should be noted that during the process of ratio transformations shown in Section 3.2, if $f_j = g_j$, for some j, then the corresponding *jth* feature in the newly transformed data will have no contribution to classification because the *jth* feature is zero after transformation. Suggested by Fan et al. (2016), we used the transformed features and the original features jointly to build a new detection model named DT-EnSVM2. Compared to DT-EnSVM, the difference lies in Step 3 (Section 3.2), where DT-EnSVM2 does not merely use the transformed data, $(X^{(2)'}, Y^{(2)})$, but $(X^{(2)'}, X^{(2)}, Y^{(2)})$ to train the base SVM classifiers.

5. Experimental setting

5.1. *Dataset description*

Currently, there are only a few public datasets available in intrusion detection domain, where researchers can conveniently make performance comparisons among different detection models. NSL-KDD, which are mostly used in literatures, are used to evaluate our proposed method.

KDD'99 dataset has long been considered as benchmark dataset in intrusion detection domain. However, KDD'99 dataset suffers from some drawbacks. For example, there are redundant, duplicated records, causing the classifier trained with it biased toward the more frequently occurring records. In order to eliminate the undesirable influence, Tavallaee et al. (2009) proposed a much more effective dataset, namely, the NSL-KDD dataset, which is a modified version of KDD'99 dataset that eliminates all the redundant records and reconstructs the structure, making it more reasonable in both data size and data structure.

NSL-KDD dataset contains TCP connection records that consist of 41 information features plus one labeling feature. The 41 informational features are used to describe the details of each TCP connection in the dataset; the labeling feature helps to specify each connection as either normal or abnormal. This dataset can be available in https://github.com/defcom17/NSL_KDD or https://www.unb.ca/cic/datasets/nsl.html

5.2. *Experiment setup*

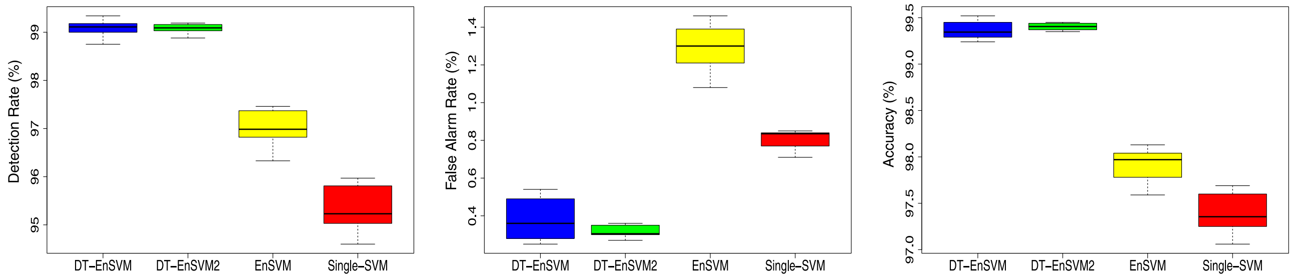
The empirical experiments in our study were all executed in a computer with an Intel Core i7-6700 CPU@ 3.40GHz with 16 GB RAM running Windows 7. The data transforation and model training were implemented using R and LIBSVM (Chang and Lin, 2007) (R version 3.3.2, R package: "e1071").

In order to eliminate the unit differences between features and to prevent the dominance of features with large value to those with small value ranges, we normalized the data into a range of [0,1] in advance.

A 10-fold cross validation method was used to train and test our proposed intrusion detection model, where the original dataset is randomly sampled into 10 mutually exclusive subset with equal size. In each run of our proposed model, nine subsets are selected to train the intrusion detection model and the remaining one is used to test the model. Thus, by running the model 10 times, each subset has an equal chance of being chosen to train and test the model. Finally, the performance of the proposed detection model is computed by averaging the results of each testing subset.

5.3. *Experimental results and discussion*

To evaluate the performance of proposed method in intrusion detection problem, the 10-fold cross-validation has been adopted and the algorithm were carried out ten times and then the obtained results were averaged. Accuracy, Detection Rate (DR), False Alarm Rate (FAR) were chosen to evaluate the performance of our proposed model as well as to conduct



(a) (b) (c)

Fig. 4 – Detection performances of DT-EnSVM, DT-EnSVM2, EnSVM, SVM. (a) shows the DR results, (b) shows the FAR results, and (c) shows the Accuracy results.

Table 1 – Confusion matrix.

		Predicted	
		Attack	Normal
Actual	Attack	TP	FN
	Normal	FP	TN

Table 2 – Performances of the four compared methods.

Metric	DT-EnSVM	DT-EnSVM2	EnSVM	SVM
Accuracy (%)	99.36 (0.10)	99.41 (0.06)	97.88(0.25)	97.39 (0.22)
DR (%)	99.07 (0.19)	99.09 (0.14)	96.93(0.50)	95.34 (0.46)
FAR (%)	0.38 (0.10)	0.31 (0.05)	1.29(0.13)	0.82 (0.07)

Note: Standard errors shown in parentheses are in percentage form.

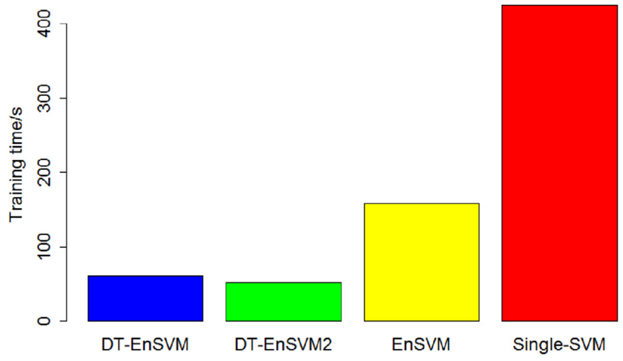


Fig. 5 – Training time required by the DT-EnSVM, DT-EnSVM2, EnSVM, and SVM methods.

expedient comparisons with other detection methods. Here we define the confusion matrix as shown in Table 1, where TP and TN mean true positive and true negative, respectively, indicating the attack and normal samples are correctly classified. A false negative (FN) refers to an attack sample that is wrongly classified as a normal one, and a false positive denotes normal sample that is falsely considered as an attack. Thus, these aforementioned performance metrics can be obtained using the following equations:

- Accuracy=$\frac{TP+TN}{TP+TN+FP+FN}$
- DR=$\frac{TP}{TP+FN}$
- FAR=$\frac{FP}{FP+TN}$

To verify the effectiveness of our proposed intrusion detection models, we first compare the performances of DT-EnSVM and DT-EnSVM2 with those of EnSVM (an SVM ensemble intrusion detection model without using ratio transformations) and SVM (a single SVM-based intrusion detection model using the original data). Fig. 4 gives the 10-fold cross validation results of DT-EnSVM, DT-EnSVM2, EnSVM and SVM with regard to Accuracy, DR and FAR.

From the comparison results in Fig. 4, we can find that EnSVM takes clear advantages over SVM in terms of detection rate and accuracy, indicating that an ensemble can boost the detection capability of a single SVM classifier. More specifically, a comparison of the detection rate for DT-EnSVM, DT-EnSVM2, EnSVM and SVM is shown in Fig. 4(a). This figure shows that the detection rates of our proposed intrusions detection models are far higher than those of EnSVM and SVM. The differences of false alarm rate between our proposed method and the other two detection models are depicted in Fig. 4(b) where both DT-EnSVM and DT-EnSVM2 have a clear advantage over EnSVM and SVM. Finally, Fig. 4(c) shows the detection accuracy of DT-EnSVM/DT-EnSVM2 is significantly higher than those of EnSVM and SVM. Therefore, it can be inferred that our proposed intrusion method is effective and consistently better than EnSVM and SVM.

To have a better understanding, the final comparison results are summarized in Table 2, resulting the same conclusion as shown in Fig. 4 that the performances of DT-EnSVM and DT-EnSVM2 are both better than those of EnSVM and SVM. As shown, the detection accuracy of our proposed method exceeds 99% (99.36% and 99.41%, respectively), while EnSVM and SVM only achieve 97.88%, 97.39% respectively. Besides, the detection rates of DT-EnSVM and DT-EnSVM2 are both above 99%, while those of EnSVM and SVM not yet reaches 97%. Furthermore, the false alarm rates of DT-EnSVM and DT-EnSVM2 are both below 0.4%, however, the FARs of EnSVM and SVM are both above 0.8%.

Moreover, to further exploit the advantages of our proposed method, a comparison on training time required by these four models is performed. As shown in Fig. 5, we can see that the

Table 3 – Performances of the four compared methods using random split scheme.

Metric	DT-EnSVM(RSplit)	DT-EnSVM2(RSplit)	EnSVM(RSplit)	SVM
Accuracy (%)	99.16 (0.09)	99.15 (0.11)	97.13(0.21)	97.39 (0.22)
DR (%)	98.97 (0.18)	99.06 (0.15)	94.99(0.46)	95.34 (0.46)
FAR (%)	0.67 (0.12)	0.77 (0.17)	1.02(0.09)	0.82 (0.07)

Note: Standard errors shown in parentheses are in percentage form.

Table 4 – Performance comparison between FCM scheme and random split scheme.

Metric	DT-EnSVM (FCM)	DT-EnSVM2 (FCM)	EnSVM (FCM)
Accuracy (%)	99.36 (0.10)	99.41 (0.06)	97.88(0.25)
DR (%)	99.07 (0.19)	99.09 (0.14)	96.93(0.50)
FAR (%)	0.38 (0.10)	0.31 (0.05)	1.29(0.13)
Metric	DT-EnSVM (RSplit)	DT-EnSVM2 (RSplit)	EnSVM (RSplit)
Accuracy (%)	99.16 (0.09)	99.15 (0.11)	97.13(0.21)
DR (%)	98.97 (0.18)	99.06 (0.15)	94.99(0.46)
FAR (%)	0.67 (0.12)	0.77 (0.17)	1.02(0.09)

Note: Standard errors shown in parentheses are in percentage form.

Table 5 – Performance comparison of different intrusion methods.

	Method	Dataset	Accuracy (%)	DR (%)	FAR (%)
Bamakan et al. (2016)	TVCPSO-SVM	NSL-KDD	98.30	97.05	0.87
Bamakan et al. (2016)	TVCPSO-MCLP	NSL-KDD	97.44	97.26	2.41
Singh et al. (2015)	OS-ELM	NSL-KDD	98.66	98.26	0.99
Zhu et al. (2017)	I-NGSA-III+GHSOM-pr	NSL-KDD	99.24	N/A	N/A
Wang et al. (2017)	LMDRT-SVM	NSL-KDD	99.31	**99.20**[a]	0.60
Wang et al. (2017)	LMDRT-SVM2	NSL-KDD	99.28	99.16	0.61
Proposed Method	DT-EnSVM	NSL-KDD	99.36	99.07	0.38
	DT-EnSVM2	NSL-KDD	**99.41**[a]	99.09	**0.31**[a]

([a]Best values are presented in bold.)

training time cost of our proposed models is superior to those of EnSVM and SVM. EnSVM required more than three times' training time as much as DT-EnSVM2 does, and nearly 2.6 times as DT-EnSVM does. SVM needs much more time to complete the model training task. Accordingly, it can be inferred that our proposed models are both more concise and less complex than EnSVM and SVM, which can greatly increase the training speed and then reduce the training time.

The comparison results above shows that the proposed method in our study is superior to the EnSVM and SVM not only in intrusion detection performance, but also in training speed, verifying its effectiveness. Besides, both DT-EnSVM and DT-EnSVM2 have a small standard error, meaning the performance of proposed method is robust.

Therefore, our proposed intrusion detection framework can be characterized as a high accuracy, a high detection rate, a low false alarm rate, a fast training speed and a robust performance.

Remark 2. Inspired by one referee's suggestion, here we illustrate the benefits obtained from the fuzzy c-means clustering method (FCM). Actually, to construct a good ensemble, it should satisfy that the base classifiers are significantly different from each other. In this paper, we used FCM to meet this requirement. Here, we will demonstrate the effectiveness of FCM in boosting the performance of an ensemble.

We choose FCM and random split (RSplit) to obtain different subsets of original data in Step 1 in Section 4.3, and conduct a performance comparison between these two schemes. Experiment results are shown in Tables 3 and 4.

It can be found in Table 3 that performances of DT-EnSVM(RSplit) and DT-EnSVM2(RSplit) are better than that of EnSVM(RSplit) in terms of Accuracy, DR and FAR, indicating the effectiveness of ratio transformation technique in improving the detection performance. However, the performance of EnSVM(RSplit) is worse than that of SVM. Therefore, an ensemble using random split fails to achieve a good performance.

Furthermore, results in Table 4 show that the performance of DT-EnSVM, DT-EnSVM2 and EnSVM using FCM scheme are consistently better than the performance of corresponding model using random split scheme. Therefore, it can be inferred that FCM can boost the performance of the ensemble-based intrusion detection model.

To further evaluate our proposed intrusion detection framework, we compare DT-EnSVM, DT-EnSVM2 with other existing methods in intrusion detection domain using NSL-KDD dataset. The comparison results are summarized in Table 5.

To better interpret the advantages of our proposed intrusion detection framework, some of the recent researches are

Table 6 – Performance comparison of different detection methods on KDD'99 dataset.

	Method	Accuracy (%)	DR (%)	FAR (%)
Lin et al. (2015)	CANN	99.46	N/A	2.95
Lin et al. (2015)	KNN	99.89	N/A	0.32
Gan et al. (2013)	PLS+CVM	99.87	99.74	N/A
Wang et al. (2017)	LMDRT-SVM	99.93	99.94	0.10
Wang et al. (2017)	LMDRT-SVM2	99.92	99.93	0.14
	Single-SVM	99.39	99.29	0.19
	EnSVM	99.70	99.86	0.98
Proposed method	DT-EnSVM	99.92	99.93	0.11
	DT-EnSVM2	99.91	99.92	0.11

selected to conduct the performance comparison. As shown in Table 5, when compared to LMDRT-SVM/LMDRT-SVM2 in Wang et al. (2017), our proposed method has a better performance in detection accuracy, and especially in false alarm rate; however, a slightly lower in detection rate. Besides, Table 5 shows that our proposed models consistently outperform other detection methods in terms of accuracy, detection rate, false alarm rate. However, it should be noted that Table 5 just provide a snapshot of comparison between our proposed intrusion models and existing methods in intrusion detection problem. Hence, there might be some limitations in this comparison. Accordingly, we cannot claim that our proposed intrusion detection framework always performs better when compared to any of the other methods in the context of intrusion detection. However, according to the comparison results shown in Table 5, our proposed method still possesses a powerful competitive advantage in the intrusion detection domain.

5.4. *Additional comparisons*

Furthermore, two more datasets, KDD'99 (Gan et al., 2013) and Kyoto 2006+ (Song et al., 2011), are used to evaluate the effectiveness of our proposed intrusion detection framework. Different from NSL-KDD dataset, KDD'99 dataset contains redundant records. The Kyoto 2006+ dataset contains real traffic data collected from diverse types of honeypots deployed near Kyoto University.

Performance of our proposed models, EnSVM, SVM and those of other recent researches are summarized in Table 6 and Table 7, respectively.

As the results shown in Table 6, our proposed method still remain its competitive advantages when compared to other detection models. Our proposed method here is nearly the same as that of LMDRT-SVM/LMDRT-SVM2. Note that it is easy to obtain a relative high detection ability on KDD'99 dataset because of the redundant records, but not on NSL-KDD dataset. Considering the results in Tables 5 and 6, it can be inferred that our proposed method is better than LMDRT-SVM/LMDRT-SVM2 in a whole, especially in dealing with sophisticated problems. Besides, the performances our proposed models are consistently better than EnSVM and SVM in terms of accuracy, detection rate and false alarm rate, demonstrating the effectiveness of our proposed intrusion detection framework in boosting detection capability.

According to the results shown in Table 7, our proposed method can still achieve a good performance on Kyoto 2006+ dataset. Most importantly, the performance of our proposed method is still consistently better than EnSVM and Single-SVM, indicating the effectiveness of proposed intrusion detection framework one more time.

From the results above, it can be concluded that our proposed intrusion detection framework is effective. It not only can achieve a better and more robust performance, but also has a fast training speed, which is of essential importance in practice.

6. Conclusion

Many intelligence algorithms have been applied to improve the detection capability of an intrusion detection system. Among these methods, ensemble learning has received a increasing interest and shown to obtain better performance than single learning methods. Besides, the intrusion detection performance is also highly depend on the quality of training data. In this paper, we proposed an effective intrusion detection framework based on SVM ensemble with feature augmentation. Specifically, the quality-improved technique is used to provide concise, high-quality training data and SVM ensemble is applied to build intrusion detection model.

Table 7 – Performance comparison of different detection methods on Kyoto 2006+ dataset.

	Method	Accuracy (%)	DR (%)	FAR (%)
Ambusaidi et al. (2016)	CSV-ISVM	N/A	90.15	2.31
Ambusaidi et al. (2016)	LSSVM-FMIFS	N/A	97.80	0.43
Singh et al. (2015)	OS-ELM	96.37	94.24	2.05
Singh et al. (2015)	AdaBoost	92.15	90.40	6.55
Singh et al. (2015)	ELM	94.56	92.92	4.22
Wang et al. (2017)	LMDRT-SVM	98.33	99.85	3.25
Wang et al. (2017)	LMDRT-SVM2	98.47	99.85	2.96
	Single-SVM	97.53	99.63	4.66
	EnSVM	97.83	99.54	3.96
Proposed method	DT-EnSVM	98.34	99.82	3.21
	DT-EnSVM2	98.48	99.81	2.92

The empirical experiment results show that our proposed detection framework can achieve a robust performance with a high accuracy, a high detection rate, a low false alarm rate and a rapid training speed. When compared to other existing methods in intrusion detection problems, our proposed method shows strong superiority and is highly competitive. However, in this paper, we considered only the binary case of intrusion detection problems. Therefore, in future work, we plan to generalize our research to include different attack types.

Declaration of interest

None.

Acknowledgments

We are grateful to an Associate Editor and two anonymous referees for their very constructive comments and suggestions which helped improve the paper greatly. This research was financial supported by Natural Science Foundation of China (Grant Nos. 71420107025, 11701023, 61772151).

REFERENCES

Aburomman AA, Reaz MBI. A survey of intrusion detection systems based on ensemble and hybrid classifiers. Comput Secur 2017;65:135–52. doi:10.1016/j.cose.2016.11.004.

Ambusaidi MA, He X, Nanda P, Tan Z. Building an intrusion detection system using a filter-based feature selection algorithm. IEEE Trans Comput 2016;65(10):2986–98. doi:10.1109/TC.2016.2519914.

Amini M, Rezaeenour J, Hadavandi E. A neural network ensemble classifier for effective intrusion detection using fuzzy clustering and radial basis function networks. Int J Artif Intell Tools 2016;25(2). doi:10.1142/S0218213015500335.

Bamakan SMH, Wang H, Yingjie T, Shi Y. An effective intrusion detection framework based on MCLP/SVM optimized by time-varying chaos particle swarm optimization. Neurocomputing 2016;199:90–102. doi:10.1016/j.neucom.2016.03.031.

Chang CC, Lin CJ. Libsvm: A library for support vector machines. ACM Trans Intell Syst Technol 2007;2(3). doi:10.1145/1961189.1961199.

Eesa AS, Orman Z, Brifcani AMA. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Syst Appl 2015;42(5):2670–9. doi:10.1016/j.eswa.2014.11.009.

Fan J, Feng Y, Jiang J, Tong X. Feature augmentation via nonparametrics and selection (FANS) in high-dimensional classification. J Am Stat Assoc 2016;111(513):275–87. doi:10.1080/01621459.2015.1005212.

Farahani FV, Ahmadi A, Zarandi MHF. Hybrid intelligent approach for diagnosis of the lung nodule from ct images using spatial kernelized fuzzy c-means and ensemble learning. Math Comput Simul 2018;149:48–68. doi:10.1016/j.matcom.2018.02.001.

Feng W, Zhang Q, Hu G, Huang JX. Mining network data for intrusion detection through combining SVMS with ant colony networks. Future Gen Comput Syst 2014;37:127–40. doi:10.1016/j.future.2013.06.027.

Gan Xs, Duanmu Js, Wang Jf, Cong W. Anomaly intrusion detection based on PLS feature extraction and core vector machine. Knowl Based Syst 2013;40:1–6. doi:10.1016/j.knosys.2012.09.004.

Ghodselahi A. A hybrid support vector machine ensemble model for credit scoring. Int J Comput Appl 2011;17(5).

Hansen LK, Salamon P. Neural network ensembles. IEEE Trans Pattern Anal Mach Intell 1990;12(10):993–1001. doi:10.1109/34.58871.

Horng SJ, Su MY, Chen YH, Kao TW, Chen RJ, Lai JL, Perkasa CD. A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert Syst Appl 2011;38(1):306–13. doi:10.1016/j.eswa.2010.06.066.

Ippoliti D, Zhou X. A-ghsom: An adaptive growing hierarchical self organizing map for network anomaly detection. J Parall Distrib Comput 2012;72(12):1576–90. doi:10.1016/j.jpdc.2012.09.004.

Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Syst Appl 2014;41(4):1690–700. doi:10.1016/j.eswa.2013.08.066.

Kim HC, Pang S, Je HM, Kim D, Bang SY. Constructing support vector machine ensemble. Pattern Recogn 2003;36(12):2757–67. doi:10.1016/S0031-3203(03)00175-4.

Kittler J, Hatef M, Duin RP, Matas J. On combining classifiers. IEEE Trans Pattern Anal Mach Intell 1998;20(3):226–39. doi:10.1109/34.667881.

Koc L, Mazzuchi TA, Sarkani S. A network intrusion detection system based on a hidden naïve Bayes multiclass classifier. Expert Syst Appl 2012;39(18):13492–500. doi:10.1016/j.eswa.2012.07.009.

Lee S, Kim G, Kim S. Self-adaptive and dynamic clustering for online anomaly detection. Expert Syst Appl 2011;38(12):14891–8. doi:10.1016/j.eswa.2011.05.058.

Li Y, Xia J, Zhang S, Yan J, Ai X, Dai K. An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert Syst Appl 2012;39(1):424–30. doi:10.1016/j.eswa.2011.07.032.

Liao HJ, Lin CHR, Lin YC, Tung KY. Intrusion detection system: a comprehensive review. J Netw Comput Appl 2013;36(1):16–24. doi:10.1016/j.jnca.2012.09.004.

Liao Y, Vemuri VR. Use of k-nearest neighbor classifier for intrusion detection1. Comput Secur 2002;21(5):439–48. doi:10.1016/S0167-4048(02)00514-X.

Lin WC, Ke SW, Tsai CF. Cann: An intrusion detection system based on combining cluster centers and nearest neighbors. Knowl Based Syst 2015;78:13–21. doi:10.1016/j.knosys.2015.01.009.

Luo B, Xia J. A novel intrusion detection system based on feature generation with visualization strategy. Expert Syst Appl 2014;41(9):4139–47. doi:10.1016/j.eswa.2013.12.048.

Manzoor I, Kumar N, et al. A feature reduced intrusion detection system using ann classifier. Expert Syst Appl 2017;88:249–57. doi:10.1016/j.eswa.2017.07.005.

Mukherjee S, Sharma N. Intrusion detection using Naive Bayes classifier with feature reduction. Procedia Technol 2012;4:119–28. doi:10.1016/j.protcy.2012.05.017.

Nanni L, Lumini A. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Syst Appl 2009;36(2):3028–33. doi:10.1016/j.eswa.2008.01.018.

Raman MG, Somu N, Kirthivasan K, Liscano R, Sriram VS. An efficient intrusion detection system based on hypergraph-genetic algorithm for parameter optimization and feature selection in support vector machine. Knowl Based Syst 2017;134:1–12. doi:10.1016/j.knosys.2017.07.005.

Singh R, Kumar H, Singla R. An intrusion detection system using network traffic profiling and online sequential extreme learning machine. Expert Syst Appl 2015;42(22):8609–24. doi:10.1016/j.eswa.2015.07.015.

Song J, Takakura H, Okabe Y, Eto M, Inoue D, Nakao K. Statistical analysis of honeypot data and building of kyoto 2006+ dataset for NIDS evaluation. In: Proceedings of the first workshop on building analysis datasets and gathering experience returns for security; 2011. p. 29–36. doi:10.1145/1978672.1978676.

Tama BA. Hfste: Hybrid feature selections and tree-based classifiers ensemble for intrusion detection system. IEICE Trans Inf Syst 2017;100(8):1729–37. doi:10.1587/transinf.2016ICP0018.

Tao C, Zeng-lin H. Svm ensemble based on boosting and kfcm. In: Zhang W, editor. In: Software engineering and knowledge engineering: Theory and practice. Springer Berlin Heidelberg; 2012. p. 593–9. doi:10.1007/978-3-642-29455-6_83.

Tavallaee M, Bagheri E, Lu W, Ghorbani AA. In: Proceedings of the second IEEE symposium on computational intelligence for security and defence applications. A detailed analysis of the kdd cup 99 data set; 2009. doi:10.1109/CISDA.2009.5356528.

Tjhai GC, Furnell SM, Papadaki M, Clarke NL. A preliminary two-stage alarm correlation and filtering system using SOM neural network and k-means algorithm. Comput Secur 2010;29(6):712–23. doi:10.1016/j.cose.2010.02.001.

Tsai CF, Hsu YF, Lin CY, Lin WY. Intrusion detection by machine learning: a review. Expert Syst Appl 2009;36(10):11994–2000. doi:10.1016/j.eswa.2009.05.029.

Wang G, Hao J, Ma J, Huang L. A new approach to intrusion detection using artificial neural networks and fuzzy clustering. Expert Syst Appl 2010;37(9):6225–32. doi:10.1016/j.eswa.2010.02.102.

Wang G, Hao J, Ma J, Jiang H. A comparative assessment of ensemble learning for credit scoring. Expert Syst Appl 2011;38(1):223–30. doi:10.1016/j.eswa.2010.06.048.

Wang H, Gu J, Wang S. An effective intrusion detection framework based on SVM with feature augmentation. Knowl Based Syst 2017;136:130–9. doi:10.1016/j.knosys.2017.09.014.

Zhu Y, Liang J, Chen J, Ming Z. An improved NSGA-III algorithm for feature selection used in intrusion detection. Knowl Based Syst 2017;116:74–85. doi:10.1016/j.knosys.2016.10.030.

Jli Gu is currently a Ph.D. student in School of Economics and Management at Beihang University, and researcher in Beijing Key Laboratory of Emergence Support Simulation Technologies for City Operations. He received the B.Eng. degree in Industrial Engineering from Nanchang University in 2014. His research interests are in the area of computer science and data mining, currently focus on information security.

Lihong Wang received the Ph.D. degree from Harbin Institute of Technology. She is currently a researcher in National Computer Network Emergency Response Technical Team/Coordination Center of China. Her current research interests are in the area of information processing and data mining.

Huiwen Wang received her B.Sc. degree from Beihang University (BHU), China, in 1982, DEA MASE, from Paris XI, France, in 1989, and Ph.D. degree in Engineering System from BHU in 1992. She is currently a professor in Management Science and engineering Department, School of Economics and Management (SEM), BHU. Also, she is director of SEM Academic Degrees Committee, and director of Research Center of Complex Data Analysis in BHU. Prof. Wang received National Science Fund for Distinguished Young Scholars http://www.nsfc.gov.cn/english/06gp/pdf/2011/041.pdf. Her general area of research is statistics and data analysis, with a recent focus on multivariate analysis for high-dimension complex data. She is an IASC member, a member of National Statistics Teaching Materials Review Committee, executive director of China Marketing Association, editorial member of Journal of Symbolic Data Analysis.

Shanshan Wang received her B.S. in Mathematics from Qingdao University in 2008, M.S. in Statistics and Probability from Beijing Normal University in 2011, and Ph.D. in Statistics from Beijing Normal University in 2014. She is currently assistant professor at Management Science and engineering Department, School of Economics and Management (SEM), Beihang University, Beijing China. Her main research interests are High-dimensional data analysis, Quantile Regression, non/semi-parametric modeling, and survival data analysis.