An abnormal user behaviour detection method based on partially labelled data

You Lu*, Xuefeng Xi, Ze Hua, Hongjie Wu, Ni Zhang

School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou, P.R. China

Received 1 March 2014, www.tsi.lv

Abstract

Detecting abnormal user behaviour is of great significance for a secured network, the traditional detection method, which is based on machine learning, usually needs to accumulate a large amount of abnormal behaviour data for training from different times or even different network environments, so the data gathered is not in line with practical data and thus affects accuracy, and that increases overhead for data labelling. In light of these disadvantages, this paper proposes the detection method based on collaborate learning, it uses under-sampling method based on distance and distribution to generate training sample from imbalanced data, and semi-supervised learning method combined by ensemble classifying method to reduce demand for labelled data, it also uses differentiated member classifiers based on mixed perturbation method for collaborate training and selectively build ensemble classifier according accuracy to detect abnormal user behaviour. Experiments based on data from simulation and real network showed that this method can effectively detect abnormal behaviour and outperform traditional methods in several evaluating indicators.

Keywords: abnormal user behaviour detection, collaborative learning, support vector machine

1 Introduction

Abnormal user behaviour has become an increasingly serious threat to network security, behaviour such as worm, DDoS attack and botnet will burden network load, leading to dramatic drop of service quality, or even collapse of network. Therefore, accurate detection and in-time warning place an important role in network management [1, 2].

Abnormal user behaviour detection has always been a hot topic for network research. Thanks to the progress of machine learning, there are many different machine-learning methods been used in abnormal user behaviour detection. Among these methods, SVM (Support Vector Machine) [3-7] has gained more attention from researchers due to its high efficiency, stability and strong generalization ability; it can also overcome disadvantages as over-fitting, local extreme and curse of dimensionality in neural network and other methods. For example, Kim et al. proposed anomaly detection method based on SVM [4], and evaluated its performance via KDD99 data; Laskov et al. put forward one-class SVM method for intrusion detection [5], which performed well in respect of false alarm rate; Tsang et al. held up core vector machine CVM [6], which can finish fast training based on large data set; Khan et al. combined SVM and hierarchical clustering [7], which could improve the SVM method's efficiency and achieve high detection rate when dealing with large data set. Though most of current detection methods based on SVM have high efficiency, their performances are not perfect in real network environment. This is because, on one hand, the existing methods usually need to accumulate large amounts of abnormal behaviour data as training sample from different times or even different network environment, so the data gathered is not in line with practical condition and thus affects accuracy, but if practical data gathered in targeted environment and over continuous time is used as training sample, there is a new problem: abnormal user behaviour only accounts for a small part of traffic in real environment, which will cause imbalance of training data and lead to over-fit of SVM classifier, and again affects accuracy of classified detection. On the other hand, it is very expensive to obtain the label, with increasing and changing abnormal behaviour's model; the large overhead may lead to detection methods' late response to abnormal behaviour and consequently affects the effect of detection application.

There are many specific sampling methods such as under-sampling can construct the training data from imbalance traffic. However, traditional under-sampling method based on random sampling does not consider the selected subset's effect on accuracy of SVM classifier. For problem about overhead of labelling, the semi-supervised learning method can reduce the demand for labelled data by training the classifier by part-labelled sample data, but methods based on single classifier such as Self-Training [8] has low accuracy, so researchers combine the collaborative method with semi-supervised learning, such as Co-Training [9] based on two classifiers, Tri-Training [10] based on three classifiers, CoForest [11] based on n classifiers, and so on. But in iteration process of these methods, using 10-fold cross-validation to calculate the label's confidence can generate

* ***Corresponding author*** e-mail: luyou.china@gmail.com

large overhead. Moreover, all member classifiers are used in detection application, so some classifiers affected by noise accumulation should reduce the accuracy of detection application.

To solve the above two problems, this paper proposes an abnormal user behaviour detection method based on collaborative learning. First, in order to improve the traditional under-sampling methods, we calculate the sampling ratio based on distribution of majority class and distance between majority class's subsets and minority class in real data, thus balanced training sample is built on the premise that real data distribution is retained as much as possible, and classification accuracy is improved as well. Secondly, we combine collaborative learning method with semi-supervised learning method, trains member classifiers based on partially labelled data to reduce the need for labelled data. In the process of training, cross-validation is replaced by the integration of member classifiers' results in order to reduce the overhead. Finally, we use selective ensemble method to build the ensemble classifier according to the member classifiers' accuracy gradually calculated in the process of semi-supervised learning, and avoid the low accuracy member classifiers' affection to the effect of detection application. The experiment results based on simulation and practical network data showed that our method performs better in several evaluating indicators, compared with traditional methods.

The rest of the paper is organized as below: we present the basic concepts and abnormal user behaviour detection model in Section II. In section III, we introduce the methods of under-sampling, generation method of member classifiers, training and ensemble methods of member classifiers, and the process of detection. In section IV, we present the experiment, including the experiment environment and results analysis, and in section V, we make a conclusion and present some future works.

2 Model of abnormal user behaviour detection

2.1 RELATIVE CONCEPTS

Different user behaviour' network traffic has different statistical characteristics, which reflects the intrinsic characteristics of behaviour. The detection method based on machine learning is to train classifiers with labelled training samples, making it adapt to normal and abnormal behaviour's differences in terms of statistical characteristics, and then use them to classify real traffic in order to detect abnormal user behaviour. To better understand our detection method, we provide the following definitions:

Definition 1: behaviour characteristics. Factors of user's behaviour that could reflect differences between normal and abnormal behaviour and be used in statistics study, such as duration of flow, time between packet arrivals and so on. It can be represented by vector $C_{index}=\{C_1,C_{2,}\ldots,C_n\}$, in which C_i, $i \in [1,n]$ represents No.i recognition clues.

Current research literatures of traffic classification and network security provide many behaviour characteristics sets, Moore et al. even gives a list of 246 types of behaviour characteristics [12]. But in specific situations, these characteristics are usually redundant or irrelevant, and some of them need to be removed through feature selection. In this paper, in consideration of efficiency, principal component analysis is adopted as feature selection method. In light of the length of this paper, we are not going into details.

Detection method based on machine learning needs a certain amount of labelled user behaviour data as training samples, the basic procedure is as follows: capture user traffic according to behaviour characteristics, analyse behaviour data manually or in other methods, and label the data. Since SVM is a two-category classification method, the label could be set as $t \in \{1,-1,0\}$, in which 1 is positive tag and means normal behaviour, and -1 is negative tag and means abnormal behaviour, and 0 means unknown type. So the definition of training sample could be concluded.

Definition 2: training sample. Labelled user behaviour data that could be used to train classifiers, the training sample that consists of m entries of labelled data could be shown as follow:

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} & t_1 \\ x_{21} & x_{22} & \ldots & x_{2n} & t_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{m1} & x_{m2} & \ldots & x_{mn} & t_n \end{bmatrix}_{m \times (n'+1)} .$$

Every line of this sample is made up of measured value $x_{i,j}, i \in [1,m], j \in [1,n]$ on $C_{index}=\{C_1,C_{2,}\ldots,C_n\}$, with corresponding label $t_i, i \in [1,m], t_i \in \{1,-1,0\}$.

2.2 SUPPORT VECTOR MACHINE THEORY

Support vector machine is a machine learning method put forward by Vapnil et al. [3] in the 1990s. It minimizes structure risk on the basis of statistical theory and overcomes the barrier of empirical risk minimization in traditional methods, so it has good generalization ability even with small sample. Its core theory is to replace a nonlinear mapping with a kernel function that satisfies Mercer condition, which allows sample point imported to map a high dimensional feature space, and uses linearly separable plane to obtain approximate ideal classification result. If the linearly separable training samples: $S=\{(x_i,y_i), i=1,2,...,r\}, x_i \in R^d, y_i \in \{1,-1\}$.

Optimal separating hyper-plane in d-dimensional space is:

$$w \cdot x + b = 0 . \quad (1)$$

Then, seek optimal hyper-plane can be transformed into the problem of constrained optimization:

$$\begin{aligned} &\min \varphi(\omega) = \frac{1}{2}\|x\|^2 \\ &s.t. y_i[w \cdot x_i + b] - 1 \geq 0, i = 1, 2, ..., r \end{aligned} \quad . \tag{2}$$

The optimal classification function obtained at last is:

$$f(x) = sign(\sum_{i=1}^{r} a_i y_i (x_i, x) + b) . \tag{3}$$

In this function, if a_i does not equal to 0, then the sample is called support vector; b could be calculated when the support vector is selected. In linear inseparable samples, a slack variable ξ and a penalty parameter c could be added to the constraint condition in Equation (2), which turns it into:

$$\begin{aligned} &\min \varphi(\omega, \xi) = \frac{1}{2}\|x\|^2 + c\sum_{i=1}^{r} \xi_i \\ &s.t. y_i[w \cdot x_i + b] - 1 + \xi_i \geq 0, i = 1, 2, ..., r, \xi_i \geq 0 \end{aligned} \quad . \tag{4}$$

In this way, the minimum risk requirements for minimum misclassified samples and maximum class interval have been compromised, and optimal classification plane in broad sense is obtained. C>0 is a constant, which controls penalty for misclassification. According to functional theory, as long as kernel function $K(x, x')$ satisfies Mercer condition, it corresponds with some transformation space's inner product. Appropriate kernel function can transform the nonlinear separability problem in previous space into linear separability problem in feature space, therefore, appropriate $K(x_i, x')$ can transform nonlinear classification into linear classification without increasing computation complexity. After replacing inner product with kernel function, the classification decision-making function is:

$$f(x) = sign(\sum_{i=1}^{r} a_i y_i K(x_i, x) + b) . \tag{5}$$

2.3 ABNORMAL BEHAVIOUR DETECTION MODEL

Abnormal user behaviour detection procedure can be described as follows: first, constructing training data on the basis of real traffic. Since the imbalance of abnormal behaviour data can affect classifier's accuracy, a appropriate sampling method is needed to build a balanced training data on the premise that real data distribution is maintained as much as possible; then use semi-supervised learning technology to train classifiers because this method can reduce reliance on labelled data by using more unlabelled data, collaborate learning and selective ensemble are also incorporated in semi-supervised learning process to make up both sampling method and semi-supervised learning's adverse effects on classification accuracy and overhead. The last procedure is to design detection process. So the abnormal user behaviour detection model proposed in this paper, which includes sample processing, member classifiers building, semi-supervised learning, selective ensemble and abnormal behaviour detection, is shown in Figure 1.

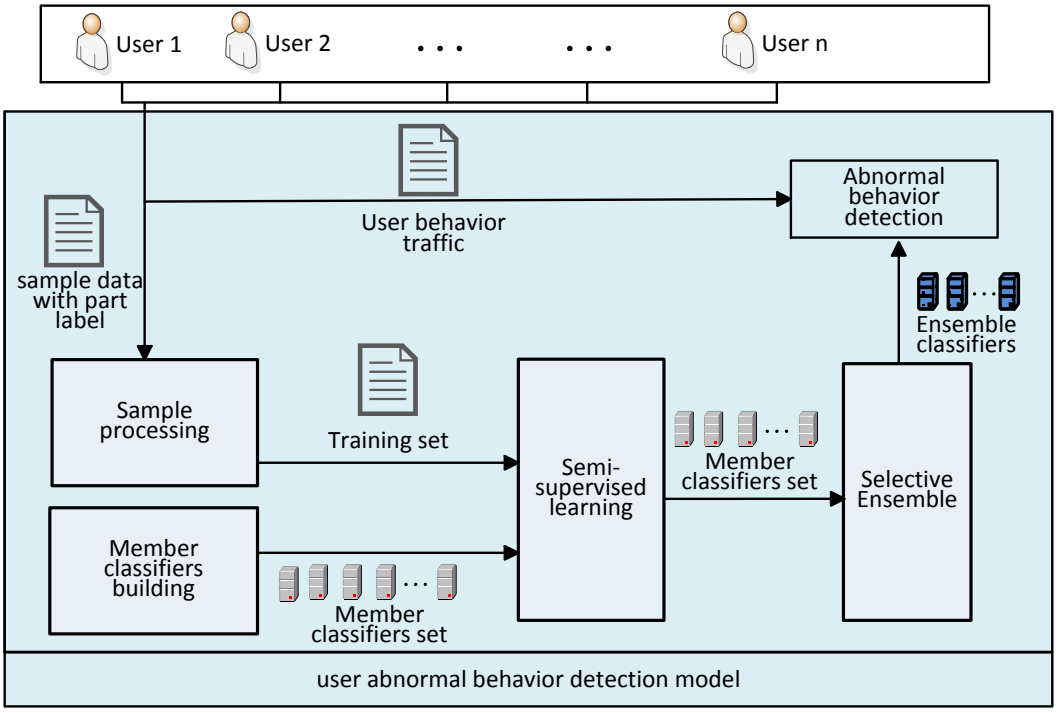


FIGURE 1 Abnormal Behaviour Detection Model

Sample processing module: using under-sampling method based on distance and distribution to process user behaviour traffic and construct training sample.

Member classifiers building module: generate a certain number of member classifiers by mixed perturbation method on the basis of feature and parameter for followed semi-supervised learning.

Semi-supervised learning module: using member classifiers' collaboration to conduct semi-supervised learning.

Selective ensemble module: select member classifiers according accuracy and integrate them into ensemble classifier.

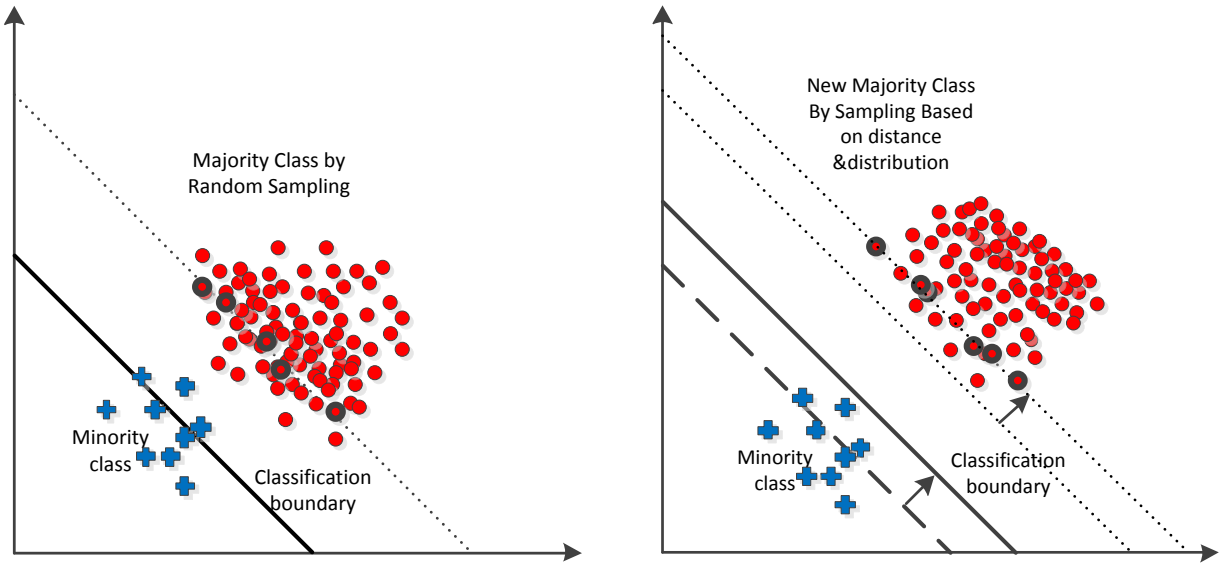
Abnormal behaviour detection module: classify user behaviour traffic by ensemble classifier and detect abnormal behaviour.

3 Abnormal behaviour detection method

3.1 CONSTRUCTION OF TRAINING SAMPLE

The quality of training sample is of great importance to the accuracy of detection. The traditional machine learning method usually accumulates abnormal behaviour data as sample, since these data is gathered over a long period of time or even from different network environments, they may be not in line with practical condition and affect sample quality. In order to obtain high quality sample, it is better to sample and label data from the targeted network and over a continuous period of time. However, abnormal behaviour data only accounts for a small part of real traffic and training sample may be imbalanced if using uniform sampling method, which may seriously affect detection ability of balanced-data based SVM classifier. The traditional sampling methods for imbalanced data are oversampling and under-sampling, the former is not appropriate because it requires that minority class must be a convex set, but the nature of abnormal data is unknown. The latter obtains balanced sample number by reducing the number of majority class sample, it has no requirement for the distribution properties of majority class, so our method uses under-sampling to construct training sample.

However, for the SVM based abnormal behaviour detection method proposed in this paper, traditional under-sampling methods [13], such as random under-sampling, DROP and CNN algorithm, still have deficiencies, because these methods only randomly select a subset of the majority class, and do not consider the selected subset's effect on the accuracy of SVM classifier. In fact, inappropriate subsets of majority class may lead to disappointed classification result, as shown in Figure 2(a), its subsets are too close to minority class, which causes SVM classification boundary moving to minority class and consequently reduces classification accuracy. If the distance between subsets of majority class and minority class is taken into consideration while sampling, and reduce subsets close to minority class and increase those far away, as shown in Figure 2(b), then the classification boundary can return to correct position. Besides, distance cannot be the only deciding condition of sampling ratio, distribution of majority class data also affects classification accuracy and should be considered as well, that's to say, if cluster majority class data, then majority subset data should account for a higher proportion in the sample, and minority subset should account for a lower proportion. Training sample constructed by this way can retain distribution of majority class to the largest extent, and ensure accuracy of classification.



(a) Classification boundary of random sampling (b) Classification boundary of sampling based on distance

FIGURE 2 The effect of sampling methods to classification boundary

In light of above analysis, this paper proposed under-sampling method based on distance and distribution, the main idea is to cluster majority class (normal user behaviour data) in data to be sampled and obtain its distribution information, then calculate the distance between different subsets of majority class with minority class (abnormal user behaviour data), at last set sampling ratio based on the size of subset and its distance. The principle is that the more items the subset has, the higher sampling ratio; and the farther the subset is from minority class, the higher the ratio, thus enabling training sample to reach a compromise between retaining as many data distribution information as possible and making sampled data of majority class being as far away from minority class as possible.

Another noteworthy problem is that our method uses semi-supervised learning method (it will be introduced later) which uses partially labelled sample for training, therefore, not all data to be sampled is labelled. This makes it even more difficult to determine majority class and minority class. This paper uses semi-supervised clustering technology to deal with it, which means clustering all data to be sampled (both labelled and unlabelled data) into two categories, then study the number of data labelled as -1 (abnormal behaviour) in both subsets, the subset with more data labelled as -1 is minority class, otherwise it is majority class.

In conclusion, sampling procedure used in this paper is as follows:

Step 1. Sample practical traffic according to uniform proportion or equal proportion, form data set to be sampled, label part of the data manually or in other ways (for training effect, data labelled as -1 needs to be accumulated to certain threshold before stops labelling). Assume there are s entries of data to be sampled, which is shown as follow: $Source = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} & t_1 \\ x_{21} & x_{22} & \dots & x_{2n} & t_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_{s1} & x_{s2} & \dots & x_{sn} & t_n \end{bmatrix}$,

$t_i \in \{1, -1, 0\}$ are labels and $x_{i,j}, i \in [1, s], j \in [1, n]$ are user behaviour data based on $C_{index}=\{C_1, C_2, \dots, C_n\}$.

Step 2. Cluster *Source* into two subsets (we use Spherical K-Means algorithm), and study the number of data labelled as -1 (abnormal behaviour) in both subsets, set the subset with more data labelled -1 as majority class Mayor, the other one is Minor. Assume there are s_1 entries of data in Mayor and s_2 entries in Minor, and $s_1 + s_2 = s$, so: $Major = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n'} & t_1 \\ x_{21} & x_{22} & \dots & x_{2n'} & t_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_{s_1 1} & x_{s_1 2} & \dots & x_{s_1 n'} & t_{n'} \end{bmatrix}$,

$Minor = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n'} & t_1 \\ x_{21} & x_{22} & \dots & x_{2n'} & t_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_{s_2 1} & x_{s_2 2} & \dots & x_{s_2 n'} & t_{n'} \end{bmatrix}$.

Calculate the central value of minority class: $\bar{x}_i = \frac{\sum_{j=1}^{s_2} x_{ij}}{s_2}$.

$\overline{Minor} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_{n'})$, in which $\bar{x}_i = \frac{\sum_{j=1}^{s_2} x_{ij}}{s_2}$.

Step 3. Cluster majority class Mayor (we use Clique algorithm) into K subsets $A_1, A_2, \dots, A_k$, assume there are $Count(A_i)$ entries of data in subset A_i, calculate central value of every category $\bar{A} = (\bar{a}_1, \bar{a}_2, \dots, \bar{a}_{n'})$, in which $\bar{a}_i = \frac{\sum_{j=1}^{Count(A_j)} x_{ij}}{Count(A_j)}$, calculate the distance between A_i and central value of minority class $\overline{Minor}$: $Dist_{(\overline{Minor}, A_i)} = \sqrt{(\bar{x}_1 - \bar{a}_1)^2 + (\bar{x}_2 - \bar{a}_2)^2 + \dots + (\bar{x}_{n'} - \bar{a}_{n'})^2}$.

Step 4. Calculate the sampling ratio of subset A_i in Mayor:

$$Ratio_{A_i} = \frac{Dist_{(\overline{Minor}, A_i)}}{\sum_{j=1}^{k} Dist_{(\overline{Minor}, A_i)}} \cdot \frac{Count(A_j)}{\sum_{j=1}^{k} Count(A_j)}. \quad (6)$$

According to ratio, number of sample A_i can be calculated:

$$Size(A_i) = s_0 \cdot \frac{Ratio_{A_i}}{\sum_{j=1}^{k} Ratio_{A_i}} + count(A_i). \quad (7)$$

In which $count(A_i)$ is the number of labelled data in subset A_i, s_0 is the pre-set number of data item after under-sampling of majority class, and $s_0 \approx s_2$.

Step 5. Randomly sample unlabelled data in subset A_i according to the number $Size(A_i)$, and add all labelled data, after processing data in all subsets, combine majority class's processing result with minority class's data, thus forming the training sample Y: $Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n'} & t_1 \\ x_{21} & x_{22} & \dots & x_{2n'} & t_2 \\ \dots & \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mn'} & t_m \end{bmatrix}_{m \times (n'+1)}$, in which $m = s_0 + s_1$, $t_i \in \{1, -1, 0\}$.

3.2 Member classifier training and ensemble

In order to reduce demand for labelled data in training process, semi-supervised learning method is adopted. It is a reasonable choice to use partially labelled data to train SVM classifier. The main idea of semi-supervised learning is to train classifier with labelled data in the sample and classify unlabelled data, then add classification result with high confidence to labelled data for future iterative learning, thus using the "knowledge" obtained from unlabelled data to further strengthen classifiers. Traditional collaboration based semi-supervised learning (such as two classifiers based Co-Training [9] and three classifiers based Tri-Training [10]) still face problems like noise accumulation and computational overhead. Therefore, some researchers combined ensemble classification with collaborative learning, such as n classifiers based Co-Forest [11] method. It uses member classifiers' ensemble classification result as confidence to reduce overhead, but in process of the iteration, for every member classifier $F_i(i\ i \in [1,n])$, all the other classifiers' $F_j(i \in [1,n]$ and $j \neq i)$ ensemble classification results need to be calculated and determined whether the result satisfies their own condition of convergence, which consequently generates a large overhead. In view of this, this paper further improves this method by calculating confidence on the basis of all member classifiers' ensemble classification results in every iteration process, then updates all member classifiers' labelled data and calculates overall condition of convergence to reduce overhead. However, this method cannot ensure optimization of every member classifier, so selective ensemble method is introduced, we increase the number of member classifiers (n) in semi-supervised learning. When constructing ensemble classifier at last, select member classifiers according accuracy, and exclude member classifiers that fail in fully optimized. Since research shows that when member classifiers reach optimal performance, there is an upper limit [14] for the number of member classifiers needed (20-30), so ensemble classifier based on accuracy can still assure accuracy.

3.2.1 Member classifiers construction based on mixed perturbation

Since the nature of selective ensemble is still ensemble learning, which integrate the classification results of member classifiers to determine final classification, and obtain better performance than single classifier. Schapire et al. proved that the necessary and sufficient condition for ensemble classifier's higher accuracy than any other member classifier is that all member classifiers have higher accuracy and differences [14]. Since selecting member classifiers on the basis of accuracy only guarantees their accuracy, this paper, according to characteristics of SVM classifiers, designs a constructing method for member classifiers that ensures their differences.

It has already proved that SVM classifiers are characteristic-sensitive and parameter-sensitive [15]. Characteristic-sensitive means that different training sample according different subsets selected from feature space corresponds can generate different classifiers, and parameter-sensitive means that Gaussian kernel based SVM's classification ability is closely related to its parameter (ξ and penalty parameter C), and there is a "low discrepancy area" in parameter ξ and C's figure region, giving them the feature of low discrepancy in member classifiers within this area, the "low discrepancy area" is called Reg_{Low}. On the basis of above conclusion, this paper proposes the construction method of member classifiers based on mixed perturbation: first use feature perturbation technology to select different subsets (the number is u) from the user behaviour characteristics set $C_{index}=\{C_1,C_2,\ldots,C_n\}$; then with the help of parameter perturbation technology, randomly select v parameter ξ and w parameter C for Gaussian kernel within the Reg_{Low} region; at last combine them together, which can generate different member classifiers (the number of classifier is u*v*w), the detail of method is:

Algorithm 1: member classifier generation based on mixed perturbation

Input: training sample **Y**, behaviour characteristic set $\mathbf{C_{index}=\{C_1,C_2,\ldots,C_n\}}$, **u** (number of characteristic subspace), **v**(number of parameter ξ), **w** (number of parameter C)

Output: member classifier set $\mathbf{F_{all}=\{f_1,f_2,\ldots,f_{u*v*w}\}}$ and characteristic subspace set $\mathbf{C_{all}=\{C(f_1),C(f_2),\ldots,C(f_{u*v*w})\}}$

1: ***For*** **i = 1** ***to*** **u**

2: Randomly select m=n/2 characteristics entry from C_{index},
form characteristic subspace $C_{index}(i)=\{C'_1,C'_2,\ldots,C'_m\}$, then build new m-dimensional sample Y_i from sample Y according the characteristics in set $C_{index}(i)$.

3: Analyse sample Y_i, calculate its Reg_{Low} by the method in literature [20], select v parameters ξ and w parameters C

4: ***For*** **j = 1** ***to*** **v**

5: ***For*** **k = 1** ***to*** **w**

6: Use parameter ξ_i and C_k to generate member classifier $f_{(i-1)*u*v+(j-1)*v+k}$, add it to F_{all} and add $C_{index}(i)$ to C_{all} as $C(f_{(i-1)*u*v+(j-1)*v+k})$

7: ***Return*** $\mathbf{F_{all}}$ **and** $\mathbf{C_{all}}$.

3.2.2 Algorithm of collaboration-based semi-supervised training and selective ensemble

The basic procedure of the collaboration-based semi-supervised training and selective ensemble is:

i) after using labelled data in training data to train all member classifiers, use these member classifiers to classify unlabelled data in training data;

ii) integrate classification results, calculate the confidence of data's label, the value is the ratio of the number of member classifiers supporting this label to the total number of classifiers;

iii) select data with highest confidence (set the number as h) from those with confidence higher than threshold (0.5 in this case), and add them to training sample;

iv) iterate above steps until it reaches the maximum iteration number or can no longer update training data. Since confidence of classification results obtained in the iteration process can also reflect accuracy of different member classifiers, so classifiers' accuracy is also updated based on results integration during iteration. When training is completed, a certain number of member classifiers with highest accuracy can be directly selected to construct ensemble classifier, which will be used to detect abnormal behaviour. Algorithm detail is as follow:

Algorithm 2: collaboration-based semi-supervised training and selective ensemble

Input: $\mathbf{F_{all}=\{f_1,f_2,\ldots,f_{u*v*w}\}}$,$\mathbf{C_{all}=\{C(f_1),C(f_2),\ldots,C(f_{u*v*w})\}}$, **Y**, iteration number **Max**, ensemble classifier number **z**, number of renewed data **h**

Output: ensemble classifier $\mathbf{F_{resemble}=\{f_1,f_2,\ldots,f_z\}}$

1 ***For*** **every** member classifier $f_i \in F_{all}$, build new m-dimensional sample Y_i from sample Y according the characteristics in set $C(f_i)$, and set f_i's accuracy $Correct(f_i)=0$

2 Use labelled data in Y_i to train member classifier f_i

3 Use all member classifiers to classify unlabelled data in sample Y

4 Integrate classification results of unlabelled data by bagging method, and calculate its confidence with $Degree = \frac{Agree}{u*v*w}$ in which *Agree* is the number of member classifiers that support the label 1

5 Select classification results whose confidence exceed 0.5 and form renewed set Result, arrange items of Result in descending order based on confidence

6 ***If*** (Result =ϕ) or (iteration number > Max) *go to* 8

7 ***Else***
use top h items in Result to update all member classifiers' sample Y_i
check every item of updated data, if a member classifier labels this item correctly, add the classifier's accuracy 1
go to 2.

8 Select top z member classifiers according accuracy and form ensemble classifier $F_{resemble}=\{f_1,f_2,\ldots,f_z\}$.

9 ***Return*** $\mathbf{F_{resemble}}$

3.2.3 Abnormal behaviour detecting procedure

The procedure of using ensemble classifier to detect abnormal user behaviour is:

i) capture user traffic;

ii) classify the data with ensemble classifier;

iii) using the bagging method to vote for the classification results;

iv) determine whether the user behaviour is abnormal or not on the basis of simple majority rule (for convenience of judgment, set z, the number of member classifiers in ensemble classifier, as singular), detailed procedures can be described as:

Step1. Measurement: measure user behaviour traffic according to behavioural characteristics, and obtain data vector to be detected $D=\{d_1,d_2,\ldots,d_n\}$

Step2. Classification: input data vector D into member classifiers (z) to classify it.

Step3. Voting: vote to the data's label out coming from every member classifier.

Step4. Judgment: on the basis of simple majority rule, if output is labelled1, then it represents normal behaviour; if -1, then it represents abnormal behaviour.

Since user behaviour traffics constantly, the detection process is in iteration, as shown in Figure 3.

4 Experiment and analysis

4.1 EXPERIMENT INTRODUCTION

This paper uses data from simulation and real network environment to verify the detection method. Simulation experiment uses 10% subset of KDD99 data set, which is adopted by many researchers as the benchmark of abnormal user behaviour detection. Real network data come from the computer room network in Suzhou University of Science and Technology, the network topology is shown in Figure4.

a) Simulation experiment.

Since 10% subset of KDD99 data set includes 97278 entries of normal behaviour data and 396743 entries of abnormal behaviour data, which is not in line with them balance of abnormal behaviour data in real network, so our experiment sampled KDD99 training data set to form the imbalanced data to be sampled, and set some of the data's label as empty. At last, construct the test data set from KDD99 by the same way. The detailed condition of the simulation experiment is shown as Table 1.

TABLE 1 Condition of simulation experiment

data set	sum	abnormal data	labelled data
set 1	2000	1%	30%
set 2	2000	5%	30%
set 3	2000	10%	30%
set 4	2000	20%	30%
set 5	2000	30%	30%
set 6	2000	40%	30%
set 7	2000	20%	5%
set 8	2000	20%	10%
set 9	2000	20%	15%
set 10	2000	20%	20%
set 11	2000	20%	25%
set 12	2000	20%	30%

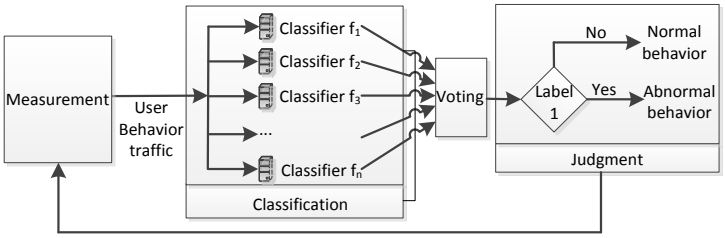


FIGURE 3 The process of detection

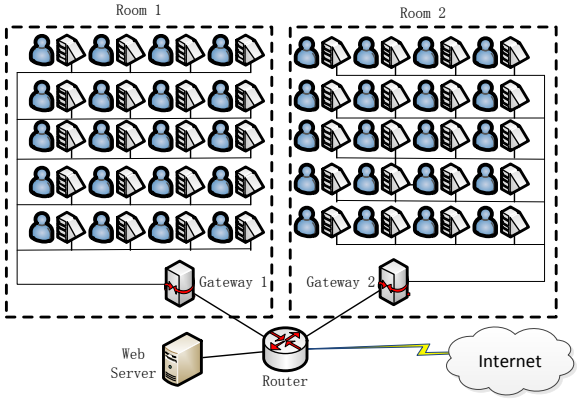


FIGURE 4 Topology of real-environment network

b) Experiment in real network environment

According to the University's course arrangement, we collect the network traffic when there were students doing network attack trials in network security course. There were 6 students conducting DDoS attack to the server in Classroom2, while 20 students in Classroom 1 doing normal activities (including Web search, VOD and P2P download). It took 45 minutes for traffic data collection (one class). Among traffic data collected, according to IP address, student behaviour in room 2 was defined as abnormal behaviour, while data collected from room 1 was defined as normal behaviour data. Data analysis found that abnormal behaviour data only accounted for 20% of total traffic, so the imbalance feature was satisfied. Behaviour characteristics used in collection is based on the KDD99's setting(use the characteristics that can be collected in our network environment), and construct over 3000 entries of behaviour data, select 1000 as data to be sampled and 2000 as test set. The detail is shown in Table 2.

TABLE 2 Condition of Real-environment experiment

data set	sum	labelled data
set 1	1000	5%
set 2	1000	10%
set 3	1000	15%
set 4	1000	20%
set 5	1000	25%
set 5	1000	30%

c) Contrast method.

Two contrast methods were adopted: the detection method based on single SVM classifier, and detection method based on Naive Bayesian classifier. The SVM classifier uses svm lib's tool to optimize parameters.

d) Evaluating indicator

Precision, Recall and F-Measure were used as evaluating indicators. Precision and Recall reflect detection method's ability to classify abnormal behaviour, and F-Measure was the harmonic mean of Recall and Precision, which could better evaluate detection method in a comprehensive way, therefore, these three indicators were widely used by researchers.

e) Hardware and software platform

The software is behaviour detection application developed by ourselves integrated with tools as svm lib, Weka, and so on, the database is SQL Server 2005, the hardware platform is Intel Core2 Quad 2.3GHz, 4GB memory, and the OS is Windows XP SP3.

4.2 ANALYSIS OF EXPERIMENT RESULT

In simulation experiment shown in Figure 5, our method is much more stable and performs better than the contrast methods. Specifically, analyse the results of different proportion of abnormal data in training data, i.e. the different imbalance degree between abnormal behaviour data and total traffic.(meanwhile the proportion of labelled data is fixed at 30%), in Figure 5(a) (Precision), Figure 5(b) (Recall), Figure 5 (c) (F-Measure) shows the comparison results with abnormal data proportion at 1%(use data set 1 in Table 1), 5%(set 2 in Table 1), 10%(set 3 in Table 1), 15%(set 4 in Table 1), 20%(set 5 in Table 1), and 30%(set 6 in Table 1). As we can see from these results, contrast method 2 performs better than contrast method 1 if there is less abnormal data, because SVM is based on balanced data. As the proportion of abnormal data rises, contrast method 1's performance gradually gets close to contrast method 2 or even

outperforms it. But our method performs better than both of contrast methods in various situations, because of its advantages coming from the collaborative learning, ensemble classification, and special treatment to imbalance data as well. Then analyse the results of different proportion of labelled data in training data (meanwhile the proportion of abnormal data is fixed at 20%), in Figure 5(d) (Precision), Figure 5(e) (Recall), Figure5(f) (F-Measure) shows the comparison results with labelled data proportion at 5%(use data set 7 in Table 1), 10%(set 8 in Table 1), 15%(set 9 in Table 1), 20%(set 10 in Table 1), 25%(set 11 in Table 1), and 30%(set 12 in Table 1). These results showed that when there is less labelled data in the total data to be sampled, contrast method 1 is better than contrast method 2, because SVM has better generalization ability than Naive Bayesian method, but with labelled data increases, performance improvement of contrast method 2 is faster than that of contrast method 1, while our method performs better than both of the contrast methods and is more stable, because when there is less labelled data, our method can rely on collaboration-based semi-supervised learning technology, and when the number of labelled data increases, it can maintain stable due to the advantage brought by ensemble classification.

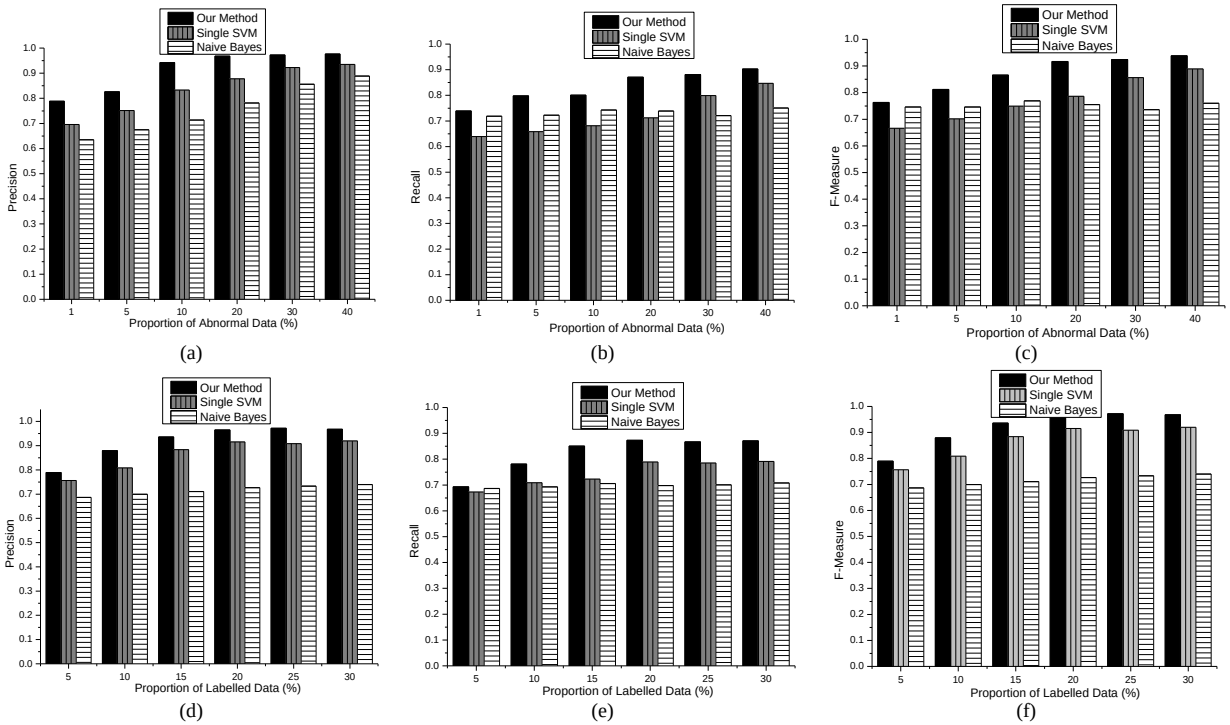


FIGURE 5 Results of Simulation Experiment

The result of real-environment experiment is shown in Figure 6. As showed in Figure 6, compared with simulation experiment result, our method maintains its advantages in all indicators. Since the proportion of abnormal behavioural data is fixed in real-time environment (about 20%), we analyse the results of different proportion of labelled data in training sample, in Figure 6(a) (Precision), Figure 6(b) (Recall), Figure 6(c) (F-Measure) showed the comparison results with labelled data proportion at 5% (use data set 1 in Table 2), 10% (set 2 in Table 2), 15% (set 3 in Table 2), 20% (set 4 in Table 2), 25% (set 5 in Table 2), and 30% (set 6 in Table 2). Results shows, thanks to the collaboration-based semi-supervised training and selective ensemble technology, our method is better than contrast methods when there is less labelled data.

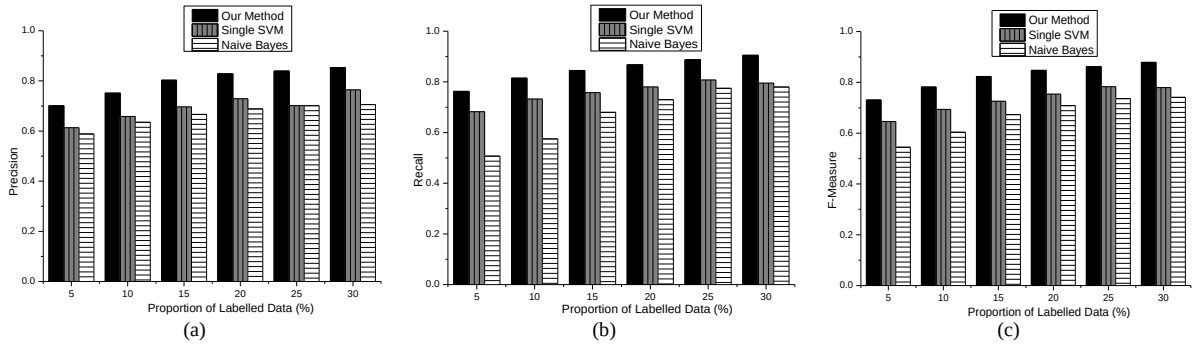


FIGURE 6 Results of Real-Environment Experiment

5 Conclusion

Traditional machine learning based abnormal user behaviour detection method need accumulating a large amount of abnormal behaviour data from different period of time or even different network environment, the data gathered is not in line with practical condition, and that increases overhead for data labelling, so they cannot detect abnormal user behaviour quickly or accurately. This paper proposes an abnormal user behaviour detection method based on collaborative learning to improve the traditional methods. It uses distance and distribution based under-sampling method to construct training sample from imbalance real data gathered in targeted environment over continuous time, trains different member classifiers by collaborative learning method on partially labelled data to reduce the overhead and labelled data, and constructs ensemble classifier based on accuracy to detect abnormal user behaviour accurately. The experiment results show that our method performs better in several evaluating indicators than traditional methods. Our next work includes optimize detection efficiency, and study user behaviour control mechanisms on the basis of abnormal behaviour detection result.

Acknowledgments

This work is supported by Jiangsu Provincial Natural Science Foundation of China under Grants No. BK20131154.

References

[1] Luo J, Han Z, Wang L 2009 Trustworthy and controllable network architecture and protocol framework *Chinese Journal of Computers* **3**(3) 391-404 *(In Chinese)*
[2] Lin C, Lei L 2007 Research on Next Generation Internet Architecture *Chinese Journal of Computers* **30**(5) 693-711
[3] Vapnik V 1998 Statistical Learning Theory *Wiley New York*
[4] Deepaa A J, Kavitha V 2012 A Comprehensive Survey on Approaches to Intrusion Detection System *Procedia Engineering* **38** 2063-9
[5] Davisa J J, Clark A J 2011 Data preprocessing for anomaly based network intrusion detection A review *Computers & Security* **30** 353–75
[6] Tsang W, Kwok J T, Cheung P M 2005 Core Vector Machine: fast SVM training on very large datasets *Journal of Machine Learning Research* **6** 363-92
[7] Khan L, Award M, Thuraisingham B 2007 A new intrusion detection system using support vector machines and hierarchical clustering *VLDB Journal* **16**(4) 507-21
[8] Hady M F A, Schwenker F 2013 *Semi-supervised Learning* Handbook on Neural Information Processing Springer Berlin Heidelberg 215-39
[9] Teichman A, Thrun S 2012 *The International Journal of Robotics Research* **31**(7) 804-18
[10] Zhou Z-H, Li M 2005 Tri-Training Exploiting Unlabelled Data Using Three Classifiers *IEEE Transactions on Knowledge and Data Engineering* **17**(11) 1529-41
[11] Li Ming, Zhou Z-H 2007 Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples *IEEE Transactions on System* **19**(11) 1479-93
[12] Moore A W, Zuev D 2004 Discriminators for use in Flow-based classification *Technical Report IRC-TR-04-028 Intel Research Cambridge*
[13] Bastista G E, Prati R C, Monard M C 2004 A study ofthe Behaviour of Several Methods for Balancing Machine Learning Training Data *ACM SIGKDD Exploration News letter* **6**(1) 20-9
[14] Schapire R E, Freund Y, Bartlet P 1998 Boosting the classification boundary a new explanation for the effectiveness of voting methods *The Annals of Statistics* **26**(5) 1651-86
[15] Valentini G, Dietterich T 2004 Bias variance analysis of support vector machines for the development of SVM based ensemble methods *Journal of Machine Learning Research* **5**(6) 725-75

Authors

You Lu, born in July, 1977, Suzhou, China

Current position, grades: Ph.D. candidate, lecturer.
University studies: School of Electronic and Information Engineering Suzhou University of Science and Technology, School of Computer Science and Engineering, Southeast University, Nanjing.
Scientific interest: next generation network architecture and user behaviour control.
Publications: 5 papers in Chinese top journals.

Xuefeng Xi, born in February, 1978, Suzhou, China

Current position, grades: Ph.D. candidate, Associate Professor.
University studies: School of Electronic and Information Engineering Suzhou University of Science and Technology, School of Computer Science and Engineering, Southeast University.
Scientific interest: network application, natural language processing, information extraction, parallel and distributed computing.
Publications: 3 papers in Chinese top journals.

Ze Hua, born in May, 1972, Suzhou, China

Current position, grades: Ph.D. candidate, Associate Professor
University studies: School of Electronic and Information Engineering Suzhou University of Science and Technology.
Scientific interest: network application, parallel and distributed computing.
Publications: 3 papers in Chinese top journals

Hongjie Wu, born in June, 1977, Suzhou, China

Current position, grades: PH. D. Associate Professor.
University studies: School of Electronic and Information Engineering Suzhou University of Science and Technology, School of Computer Science and Technology, Soochow University.
Scientific interest: network application, parallel and distributed computing.
Publications: 8 papers in International and Chinese top journals.

Ni Zhang, born in August, 1976, Shanxi, China

Current position, grades: Ph.D. candidate, lecturer.
University studies: School of Electronic and Information Engineering Suzhou University of Science and Technology, School of Computer Science and Engineering, Southeast University, Nanjing.
Scientific interest: network application, natural language processing, information extraction, parallel and distributed computing.
Publications: 2 papers in Chinese top journals.