DOI：10.7652/xjtuxb201508014

用于木马流量检测的集成分类模型

兰景宏[1]，刘胜利[1,2]，吴双[1]，王东霞[1,2]
（1.中国人民解放军信息工程大学数学工程与先进计算国家重点实验室，450001，郑州；
2.北京系统工程研究所信息保障技术重点实验室，100083，北京）

摘要：针对传统集成学习方法运用到木马流量检测中存在对训练样本要求较高、分类精度难以提升、泛化能力差等问题，提出了一种木马流量检测集成分类模型。对木马通信和正常通信反映在流量统计特征上的差别进行区分，提取行为统计特征构建训练集。通过引入均值化的方法对旋转森林算法中的主成分变换进行改进，并采用改进后的旋转森林算法对原始训练样本进行旋转处理，选取朴素贝叶斯、C4.5 决策树和支持向量机 3 种差异性较大的分类算法构建基分类器，采用基于实例动态选择的加权投票策略实现集成并产生木马流量检测规则。实验结果表明：该模型充分利用了不同训练集之间的差异性以及异构分类器之间的互补性，在误报率不超过 4.21%时检测率达到了 96.30%，提高了木马流量检测的准确度和泛化能力。

关键词：木马流量；集成学习；旋转森林；异构分类器；加权投票
中图分类号：TP393 **文献标志码**：A **文章编号**：0253-987X(2015)08-0084-06

Research on Ensemble Classification Model of Trojan Traffic Detection

LAN Jinghong[1], LIU Shengli[1,2], WU Shuang[1], WANG Dongxia[1,2]
(1. State Key Laboratory of Mathematical Engineering and Advanced Computing, The PLA Information Engineering University, Zhengzhou 450001, China; 2. Science and Technology on Information Assurance Laboratory, Beijing Institute of System and Engineering, Beijing 100083, China)

Abstract: An ensemble classification model for Trojan traffic detection is proposed to solve the problem that traditional ensemble learning methods overly depend on training samples, have low classification precision and poor generalization ability when they are applied to the detection of Trojan traffic. Traffic statistics features between Trojan communication and normal communication are distinguished and then extracted to build training sets. Equalization method is introduced to improve the principal component transform of rotation forest algorithm, and the updated rotation forest algorithm is used to rotate original training samples. Then, base classifiers are constructed by using three classification algorithms: Naive Bayes, C4.5 decision tree and Support Vector Machine. Integration is realized and the Trojan traffic detection rules are eventually established by using a weighted voting strategy based on the dynamic choice of instance. Experimental results show that the model makes full use of the diversity of different training sets and the complementarity of heterogeneous classifiers, and that a 96.30% detection rate is reached while the false positive rate is not higher than 4.21%, that is, both the accuracy and the generalization ability of Trojan traffic detection are improved.

Keywords: Trojan traffic; ensemble learning; rotation forest; heterogeneous classifier; weighted voting

收稿日期：2014-12-23。 作者简介：兰景宏(1990—)，男，硕士生；刘胜利(通信作者)，男，副教授。 基金项目：国家科技支撑计划资助项目(2012BAH47B01)；国家自然科学基金资助项目(61271252)；信息保障技术重点实验室开放基金资助项目(KJ-14-105)；上海市科研计划资助项目(13DZ1108800)。

随着互联网技术的快速发展，网络安全问题日趋严重。作为黑客进行远程控制和信息窃取的主要手段，木马攻击给网络安全造成了严重的威胁[1-2]。与其他恶意代码相比，木马的运行不会对主机造成明显的破坏，其主要的特点是隐蔽性。早期人们通常采用特征匹配的方法对木马通信流量进行检测，但是随着计算机技术的不断发展，木马新变种不断涌现。基于特征匹配的检测方法虽然具有很高的准确性，但是不能识别未知木马，需要不停地更新特征库，具有一定的滞后性[3]。基于通信行为分析的检测方法，利用通信流量的行为统计特征构建异常检测模型，无需特征库，能够识别变种木马，具有较好的通用性和应用前景。随着数据挖掘技术的迅速发展，数据分类技术从网络通信数据中自动提取行为特征和计算异常行为模式，极大地降低了开发成本。基于有监督学习的分类算法具有准确性高、模型可读性强的优点，在木马通信流量的检测中得到了广泛应用[4-6]。这类检测方法利用标记过的样本训练分类器并产生检测规则，其中常用的分类算法包括朴素贝叶斯(naive Bayes，NB)、决策树、神经网络、支持向量机(support vector machine，SVM)、隐马尔可夫模型等。

随着新型互联网应用的不断涌现，网络流量日趋复杂。流量的多样性和动态变化对木马流量的检测提出了严峻的挑战。由于过于依赖某种既定的数学模型且对待分类数据的空间分布要求较高，单一的分类算法已经不能满足木马检测的实际需求，而集成学习综合考虑多个分类器对同一检测对象的分类结果，能得到更加准确和全面的判断，已经成为目前研究的热点。

基分类器的精度以及基分类器之间的差异性是影响集成效果的两个重要因素，一个良好的集成分类器必须由精度较高的基分类器组成，同时基分类器之间差异性较大[7]。文献[8]进一步指出，在保证不降低基分类器精度的前提下，提高基分类器之间的差异性能有效提高集成的精度。一些传统的集成学习方法忽略了基分类器差异度和精度之间的关系，集成精度的提高缺乏理论和方法上的指导[9-11]，导致运用到木马流量的检测中存在对训练样本要求较高、集成精度难以提升、泛化能力差等问题。针对木马通信行为检测中的流量分类问题，本文提出了一种基于改进旋转森林算法的木马流量集成分类模型。

1　木马流量检测的集成分类模型

网络上不同的通信行为一般具有不同的流量统计特征，这些统计特征反映了通信的本质特性。基于改进旋转森林算法的木马流量集成分类模型，从本质上说是利用机器学习的理论和方法对木马通信和正常通信反映在流量统计特征上的差别进行区分。为了保证检测模型的可靠性，需要提取有效区分木马通信和正常通信的行为统计特征。由于本文的研究重点是分类模型的改进而不是通信行为特征的提取，因此借鉴了文献[5-6]的研究成果并采用文献[5]中给出的方法对网络流量进行预处理，提取一次完整的通信过程中产生的多条数据流并利用基于时间戳的聚类算法生成数据流簇，在数据流簇的基础上提取木马在请求连接阶段和交互操作阶段的行为统计特征。请求连接阶段的检测特征为平均数据包长度、平均数据包包头长度、源端口有序度、数据流簇中数据流最小差异度。交互操作阶段的检测特征为平均流长度、下载小包数量、上传大包数量、数据流持续时间方差、数据流簇持续时间/最长数据流持续时间、源端口有序度、下行小包数量/小包总数量、最长数据流持续时间、上行大包数量/大包总数量、主连接数据包时间间隔方差、数据量上传下载比、数据包数量上传下载比。需要说明的是：本文把小于 100 B 的数据包称为小包，否则为大包；主连接是数据流簇中持续时间最长的数据流，主要用于传输木马控制命令和一些心跳信息。

为了实现对木马通信流量快速准确地检测，首先需要对采集的网络流量进行处理，提取行为统计特征构建训练集，然后训练分类器并产生检测规则。在木马通信流量的检测中，所采用的分类算法是影响检测精度的重要因素。本文提出的木马流量检测集成分类模型的总体框架如图 1 所示。

异构分类器集成学习包括 3 个模块：①训练流量预处理模块从网络出口处捕获进出局域网的通信流量，进行白名单过滤后对通信流量进行行为特征提取和计算，标记后生成训练集；②基分类器构建模块采用改进的旋转森林算法对训练样本进行处理，采用 C4.5、SVM 和 NB 3 种分类算法生成多个基分类器；③动态加权投票集成模块采用基于实例动态选择的加权投票策略实现集成并形成检测规则。

木马流量检测包括 2 个模块：①检测流量预处理模块为对待检测的通信流量进行行特征提取和计算，根据基于实例动态选择的加权投票策略统计每

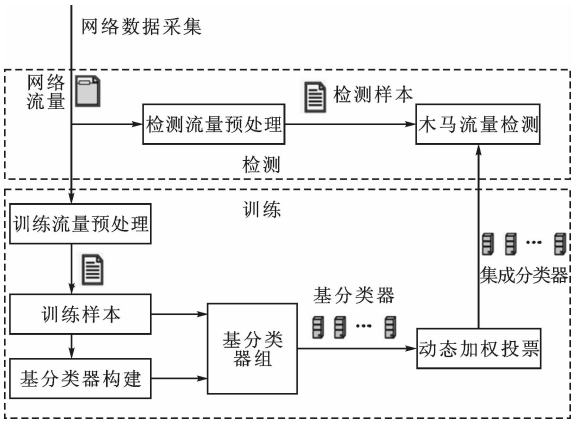

图 1 木马流量检测的集成分类模型

个基分类器投票的权重；②木马流量检测模块利用生成的检测规则对待检测的通信流量进行分类检测，分类结果包括木马流量和正常流量两种，木马流量的类别为 1，正常流量的类别为 0，根据检测结果生成详细的报警信息。

2 基于改进旋转森林的集成学习

旋转森林算法[12]最初由 Rodriguez 等提出，其核心思想是利用主成分分析法(principal component analysis，PCA)对原始样本数据进行旋转处理，从而生成有差异的训练集。这种旋转处理的好处在于保证基分类器分类精度的同时增加了基分类器之间的差异性，因而提高了集成的分类精度。

2.1 旋转森林算法的改进

传统的旋转森林算法使用 PCA 变换对原始样本数据进行变换处理，PCA 变换一般以样本的协方差矩阵为基础进行分析，适用于原始特征之间呈线性关系的情形，其计算过程容易受特征的量纲和数量级的影响。为此，本文引入均值化的方法对旋转森林算法中的 PCA 变换进行改进[13]。

假设有 n 个训练样本，每个样本有 m 个特征属性，原始样本数据可表示为矩阵 $\boldsymbol{X}=(x_{ij})_{n\times m}$，均值化是把矩阵 $\boldsymbol{X}$ 中每个数据点除以对应列的平均值，均值化后的训练集表示为矩阵 $\boldsymbol{Y}=(y_{ij})_{n\times m}$，即有

$$y_{ij}=x_{ij}/\bar{x}_j(i=1,2,\cdots,n;j=1,2,\cdots,m) \quad (1)$$

式中：$\bar{x}_j=(\sum_{i=1}^{n}x_{ij})/n$，为第 j 列的平均值。

下面计算矩阵 $\boldsymbol{Y}$ 的协方差矩阵 $\boldsymbol{U}=(u_{ij})_{m\times m}$。因为 $\boldsymbol{Y}$ 的每一列的平均值都是 1，故而有

$$u_{ij}=(\sum_{k=1}^{n}(y_{ki}-\bar{y}_i)(y_{kj}-\bar{y}_j))/(n-1)=$$

$$(\sum_{k=1}^{n}(y_{ki}-1)(y_{kj}-1))/(n-1)=$$

$$(\sum_{k=1}^{n}((x_{ki}/\bar{x}_i)-1)((x_{kj}/\bar{x}_j)-1))/(n-1)=$$

$$(\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j))/(\bar{x}_i\bar{x}_j(n-1))=$$

$$s_{ij}/(\bar{x}_i\bar{x}_j) \quad (2)$$

式中：$s_{ij}(i,j=1,2,\cdots,m)$ 是矩阵 $\boldsymbol{X}$ 的协方差矩阵中的元素；r_{ij} 为矩阵 $\boldsymbol{X}$ 的相关系数。下面求均值化后的矩阵 $\boldsymbol{Y}$ 中各个特征(即矩阵的列)的相关系数

$$r'_{ij}=u_{ij}/(u_{ii}u_{jj})^{1/2}=$$

$$s_{ij}/(\bar{x}_i\bar{x}_j(s_{ii}/\bar{x}_i\bar{x}_i)^{1/2}(s_{jj}/\bar{x}_j\bar{x}_j)^{1/2})=$$

$$s_{ij}/(s_{ii}s_{jj})^{1/2}=r_{ij} \quad (3)$$

均值化后特征之间的相关系数保持不变，表明均值化后保留了原始矩阵数据里的信息，不会影响训练样本的质量和分类器的精度。使用旋转森林对原始样本数据进行处理时，用均值化后的矩阵 $\boldsymbol{Y}$ 代替矩阵 $\boldsymbol{X}$ 进行 PCA 变换。

2.2 基分类器构建

构建分类器需要选择合适的分类算法，在机器学习领域分类算法种类繁多且有不同的适用范围，但是大多数都是内存驻留算法。本文选取以下几个评估指标作为基分类器构建算法的选择标准[14]：①预测的准确率为模型准确地预测未知的或未出现过的数据类别的能力；②分类速度为构建和使用分类模型的时间开销；③鲁棒性为分类模型的抗噪声能力和对残缺数据的预测能力；④可伸缩性为在较大数据量的情形下有效生成模型的能力；⑤可解释性为分类模型产生的规则提供的洞悉和理解的层次。

文献[15]指出，没有哪一种单一的分类算法能在所有的网络环境中都能取得比其他分类算法更好的分类效果。为了充分利用不同分类算法之间的差异性和互补性，本文选取了 C4.5 决策树、NB 和 SVM 3 种具有代表性和较大差异性的分类算法用于基分类器的构建。朴素贝叶斯算法设计简单、运行速度快、抗噪性能好，被广泛应用于文本分类领域，虽然朴素贝叶斯算法对属性之间的独立性要求较高，但是训练样本经过 PCA 变换后属性之间的关联性减弱。旋转森林算法最初是针对决策树算法设计的，决策树算法在构建分类器时采用信息增益率选择向下的分支属性，在分类精度和效率方面具有较好的折中，但是在实际运用中容易产生过拟合现象且对连续属性处理陡峭。SVM 算法[16]建立在统计学习理论和结构风险最小原理的基础上，兼顾训练误差和泛化能力，能够有效处理非线性数据，在样

本有限的情况下仍能取得较好的泛化性能，具有良好的稳定性和可伸缩性。

为方便性起见，人为设定每种分类算法生成的基分类器个数相同。基于改进旋转森林算法的基分类器构建过程如下。

输入　原始训练集 $\boldsymbol{X}=(x_{ij})_{n\times m}$，行为统计特征 $\boldsymbol{F}=(f_1,f_2,\cdots,f_m)$，基分类器个数 l，特征集合划分数 K。

输出　基分类器集合 $G=\{g_1,g_2,\cdots,g_l\}$。

构建步骤为：

(1)按照上一小节给出的均值化方法对矩阵 $\boldsymbol{X}$ 进行处理，生成数据矩阵 $\boldsymbol{Y}$；

(2)for $i=1$ to l；

(3)把 $\boldsymbol{F}$ 随机划分为 K 个子集(每次划分均不同)，每个子集中包含的特征数为 $M=m/K$，如果不能整除则把余下几个特征加入最后一个子集；

(4)for $k=1$ to K；

(5)对于第 k 个子集 F_{ik}，从 $\boldsymbol{Y}$ 中取出与 F_{ik} 对应的样本数据 $\boldsymbol{Y}_{ik}$(即列)，随机选取 $\boldsymbol{Y}_{ik}$ 中 75% 的数据，记为 $\boldsymbol{Y}'_{ik}$；

(6)对 $\boldsymbol{Y}'_{ik}$ 进行 PCA 变换，得到 M 个特征值，计算不为 0 的 $N(N\leqslant M)$ 个特征值对应的特征向量 $\boldsymbol{a}_{k1},\boldsymbol{a}_{k2},\cdots,\boldsymbol{a}_{kN}$ 和矩阵 $\boldsymbol{A}_{ik}=(\boldsymbol{a}_{k1},\boldsymbol{a}_{k2},\cdots,\boldsymbol{a}_{kN})$；

(7)end；

(8)合并 $\boldsymbol{A}_{i1},\boldsymbol{A}_{i2},\cdots,\boldsymbol{A}_{iK}$ 生成第 i 个基分类器对应的旋转矩阵

$$\boldsymbol{A}_i=\begin{bmatrix}\boldsymbol{A}_{i1} & & & \\ & \boldsymbol{A}_{i2} & & \\ & & \ddots & \\ & & & \boldsymbol{A}_{iK}\end{bmatrix}$$

(9)按照原始训练集 $\boldsymbol{Y}$ 中各列所在的原始位置对 $\boldsymbol{A}_i$ 的列重新进行排序，记为 $\boldsymbol{A}'_i$，由此生成新的训练集 $\boldsymbol{Y}_i=\boldsymbol{Y}\boldsymbol{A}'_i$；

(10)在 C4.5、SVM 和 NB 算法中选择一种分类算法在训练集 $\boldsymbol{Y}_i$ 上生成一个基分类器；

(11)end。

2.3　动态加权投票集成

基分类器生成后，多数投票法是一种简单高效的集成方式，投票判决时每个分类器赋予的权值正比于它在训练集上的分类正确率。由于网络流量的动态变化，待检测流量的统计特征的分布是变化的，因此每个基分类器对网络流量的检测能力也是变化的，一种更为合理的加权方法应该考虑当前待检测流量的统计特征的分布与训练集之间的关系。

本文提出一种基于实例动态选择的加权投票策略实现集成。首先对训练样本集进行处理，使用 k 均值算法把训练样本聚类成 L 个簇，计算每个簇的中心；然后统计每一个基分类器对 L 个簇内所有样本分类的结果，计算分类正确率。

待分类实例 $\boldsymbol{E}$ 的分类过程如下。

步骤 1　计算待分类实例 $\boldsymbol{E}$ 与 L 个簇中心的距离，选择距离最近的 L_1 个簇；

步骤 2　分别统计每个基分类器对 L_1 个簇内所有样本的分类正确率 $\beta_k(k=1,2,\cdots,l)$，归一化后即为投票权重

$$w_k=\beta_k/\sum_{i=1}^{l}\beta_i(k=1,2,\cdots,l)$$

步骤 3　每个基分类器独立地给出 $\boldsymbol{E}$ 的分类结果，结合所赋予的投票权重实现集成分类。分类阈值可以根据实际的网络环境设定，使得检测的误报率和漏报率达到较好的折中。一般把分类结果大于 0.5 的数据流簇判定为木马通信流量。

本文采用异构欧氏距离来计算待分类实例与聚类簇中心的距离。待分类实例 $\boldsymbol{E}=(e_1,e_2,\cdots,e_m)$ 与某个聚类簇的中心 $\boldsymbol{C}=(c_1,c_2,\cdots,c_m)$ 之间距离为

$$d(\boldsymbol{E},\boldsymbol{C})=\left(\sum_{i=1}^{m}h_i^2(\boldsymbol{E},\boldsymbol{C})\right)^{1/2}$$

式中

$$h_i(\boldsymbol{E},\boldsymbol{C})=\begin{cases}0, & e_i=c_i\\ 1, & e_i\neq c_i\end{cases}(\text{第 } i \text{ 个特征是离散型})$$

$$\frac{|e_i-c_i|}{e_{i,\max}-e_{i,\min}}(\text{第 } i \text{ 个特征是连续型})$$

3　实验测试与分析

3.1　实验环境与数据

为了验证本文提出的检测模型的有效性，基于本模型构建了木马流量检测系统，在某实验室局域网出口处进行实际测试，检测系统环境部署如图 2 所示。通过交换机的镜像端口捕获进出局域网的网络流量进行检测，采用 Alexa 网站提供的访问量排名前 1 000 的域名对应的 IP 地址生成白名单对网络数据包进行过滤，减轻后台检测程序处理的压力。

实验测试主机 60 台，平均带宽约为 5 Mb/s，其中 10 台主机植入木马(控制端在局域网外)。为保证实验结果的说服力和方便性，本文使用的木马样本全部来自互联网，包括灰鸽子、pcshare、Bifrost、Gh0st 等国内外常见木马，随机进行各种操作生成木马流量，正常流量由实验室互联网出口流量生成，

两者混合生成训练样本，其中正常数据流簇与木马数据流簇的比例为 8∶1。

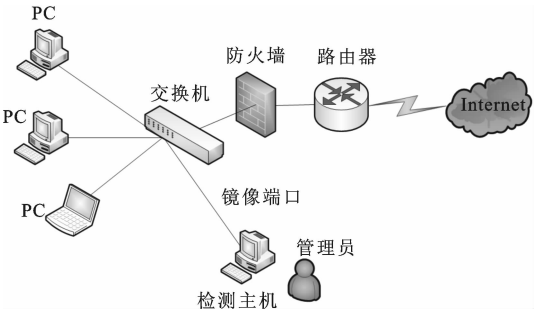

图 2 木马流量检测系统环境部署

3.2 实验评估指标和对比模型

本文采用入侵检测领域常用的检测率和误报率作为检测能力的评估指标，各项检测指标为：

数据流簇检测率(detection rate)设为

$$D = t_t / n_t$$

数据流簇误报率(false positive rate)设为

$$F = f_t / n_n$$

式中：n_t 表示测试中所有的木马通信数据流簇的个数；n_n 表示所有的正常通信数据流簇的个数；t_t 表示被正确识别的木马通信数据流簇的个数；f_t 表示正常通信数据流簇中被识别为木马数据流簇的个数。

用于实验对比测试的集成方法包括 Bagging[10]、Adaboost[11]、Rotboost[12]。3 种集成方法选用相应文献中给出的最优参数，即迭代次数均为 3，Rotboost 和本文方法进行旋转变换时特征子集的大小固定为 4。

3.3 实验结果分析

进行实验测试时，采集全网流量进行检测。为了测试检测模型的检测精度尤其是泛化能力，测试集中采集的部分木马流量未参与训练。本文采用 ROC 曲线(receiver operating characteristic curve，ROC)来评估集成学习方法的性能。ROC 曲线在入侵检测领域应用广泛，它体现了检测率和误报率之间的平衡，低于 ROC 曲线的区域(the area under ROC curve，AUC)的面积是评价集成学习方法的重要指标，它的取值范围是[0，1]，AUC 的面积越大，表示该集成学习方法的性能越好。

获取训练集后，分别采用 Bagging、Adaboost、Rotboost 和本文提出的集成学习方法构建检测分类模型，采用 Naive Bayes、C4.5 和 SVM 异构集成的方式，每种不同类型的基分类器均为 6 个。测试时通过调整检测的分类阈值获得每种集成学习方法的 ROC 曲线，如图 3 所示。

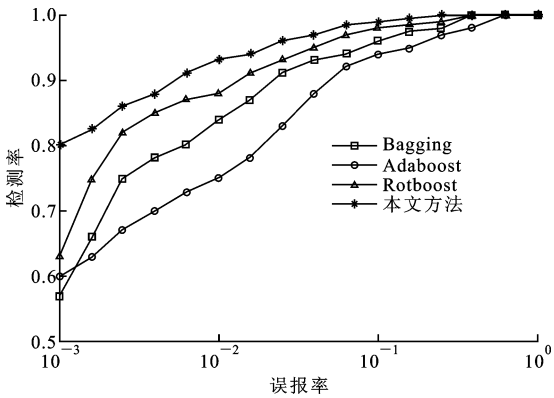

图 3 4 种集成学习方法 ROC 曲线对比

由图 3 可知，本文提出的木马流量检测模型与传统的几种集成学习方法相比具有更好的检测精度和泛化性能，当误报率为 1%左右时，本文模型已经达到 93%左右的检测率，体现了较好的分类性能。对异构集成和同构集成的检测效果进行了对比测试，表 1 给出了分类阈值为 0.5 时 4 种集成学习方法分别采用单一分类算法集成和多种分类算法集成的方式构建的检测模型的检测率和误报率的对比。由实验结果可知，对于每一种集成学习方法，异构集成的方式均能取得比同构集成更好的检测效果。

4 结 论

针对传统的集成学习方法运用到木马流量的检测中存在对训练样本要求过高、集成精度难以提升、泛化能力差的问题，提出一种基于改进旋转森林算法的木马流量集成分类模型。首先利用改进的旋转森林算法对原始样本数据进行处理，生成有差异性

表 1 4 种集成学习方法检测率和误报率的对比

分类方法	检测率/%				误报率/%			
	Bagging	Adaboost	Rotboost	本文方法	Bagging	Adaboost	Rotboost	本文方法
naive Bayes	91.81	83.90	91.38	93.65	6.08	6.44	5.46	5.19
C4.5	92.43	84.26	92.35	94.70	5.85	5.14	6.02	4.56
SVM	90.37	85.73	90.36	95.08	5.40	5.75	5.25	4.81
3 种方法集成	92.62	86.42	93.51	96.30	5.25	4.88	4.66	4.21

的训练集，在保证基分类器精度的同时增大了基分类器之间的差异性；然后选取具有代表性和较大差异性的 3 种分类算法构建基分类器；最后采用一种基于实例动态选择的加权投票策略方法实现集成。实验结果表明，与传统的一些集成学习方法相比，本文提出的分类模型具有更高的检测精度和泛化性能，提高了基于通信行为分析的木马检测方法的实用性。下一步将从木马通信行为特征提取和提高分类模型的自适应能力两方面开展研究，进一步提高木马流量检测的准确度和可信度。

参考文献：

[1] 国家互联网应急中心. 网络安全信息与动态周报[EB/OL]. (2014-11-06)[2014-11-18]. http://www.cert.org.cn/publish/main/uplood/File/2014CNCERT44.pdf.

[2] 国家互联网应急中心. 2013 年我国互联网网络安全态势综述[EB/OL]. (2014-03-28)[2014-11-20]. http://www.cert.org.cn/publish/main/upload/File/CNCERT%202013.pdf.

[3] ROESCH M. Snort-lightweight intrusion detection for networks[C]// Proceedings of the LISA 13th Systems Administration Conference. Berkeley, CA, USA: USENIX, 1999: 229-238.

[4] TEGELER F, FU X, VIGNA G, et al. Botfinder: finding bots in network traffic without deep packet inspection[C]// Proceedings of the 8th International Conference on Emerging Networking Experiments and Technologies. New York, USA: ACM, 2012: 349-360.

[5] 胥攀，刘胜利，兰景宏，等. 基于多数据流分析的木马检测方法[J]. 计算机应用研究，2015，32(3)：890-894.
XU Pan, LIU Shengli, LAN Jinghong, et al. Trojan detection method based on analysis of multiple data flow[J]. Application Research of Computers, 2015, 32(3): 890-894.

[6] 李世淙，云晓春，张永铮. 一种基于分层聚类方法的木马通信行为检测模型[J]. 计算机研究与发展，2012，49(S2)：9-16.
LI Shicong, YUN Xiaochun, ZHANG Yongzheng. A model of Trojan communication behavior detection based on hierarchical clustering technique[J]. Journal of Computer Research and Development, 2012, 49(S2): 9-16.

[7] KOTSIANTIS S. Combining bagging, boosting, rotation forest and random subspace methods[J]. Artificial Intelligence Review, 2011, 35(3): 223-240.

[8] KROGH A, VEDELSBY J. Neural network ensembles, cross validation, and active learning[C]// Proceedings of the Annual Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 1995: 231-238.

[9] BÜHLMANN P. Handbook of Computational Statistics[M]. Berlin, Germany: Springer, 2012: 985-1022.

[10] GARCIA N. Supervised projection approach for boosting classifiers[J]. Pattern Recognition, 2009, 42(9): 1742-1760.

[11] ZHANG Chunxia, ZHANG Jiangshe. Rotboost: a technique for combining rotation forest and adaBoost[J]. Pattern Recognition Letters, 2008, 29(10): 1524-1536.

[12] RODRIGUEZ J J, KUNCHEVA L I, ALONSO C J. Rotation forest: a new classifier ensemble method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2006, 28(10): 1619-1630.

[13] 高艳，于飞. 一种用于综合评价的主成分分析改进方法[J]. 西安文理学院学报：自然科学版，2011，14(1)：105-108.
GAO Yan, YU Fei. A modified principal component analysis algorithm for comprehensive evaluation[J]. Journal of Xi'an University of Arts and Science: Natural Science, 2011, 14(1): 105-108.

[14] HAN J, KAMBER M. 数据挖掘概念与技术[M]. 范明，孟小峰，译. 2 版. 北京：机械工业出版社，2007：185-188.

[15] CALLADO A, KELNER J, SADOK D, et al. Better network traffic identification through the independent combination of techniques[J]. Journal of Network and Computer Applications, 2010, 33(4): 433-446.

[16] VAPNIK V. The nature of statistical learning theory[M]. Berlin, Germany: Springer Verlag, 2000: 9-11.

（编辑　武红江）