基于集成分类器的恶意网络流量检测

汪洁，杨力立，杨珉
（中南大学信息科学与工程学院，湖南 长沙 410083）

摘 要：针对目前网络大数据环境攻击检测中因某些攻击步骤样本的缺失而导致攻击模型训练不够准确的问题，以及现有集成分类器在构建多级分类器时存在的不足，提出基于多层集成分类器的恶意网络流量检测方法。该方法首先采用无监督学习框架对数据进行预处理并将其聚成不同的簇，并对每一个簇进行噪音处理，然后构建一个多层集成分类器 MLDE 检测网络恶意流量。MLDE 集成框架在底层使用基分类器，非底层使用不同的集成元分类器。该框架构建简单，能并发处理大数据集，并能根据数据集的大小来调整集成分类器的规模。实验结果显示，当 MLDE 的基层使用随机森林、第 2 层使用 bagging 集成分类器、第 3 层使用 AdaBoost 集成分类器时，AUC 的值能达到 0.999。

关键词：恶意网络流量；攻击检测；攻击阶段；网络流量聚类；集成分类器
中图分类号：TP302
文献标识码：A
doi: 10.11959/j.issn.1000-436x.2018224

Multitier ensemble classifiers for malicious network traffic detection

WANG Jie, YANG Lili, YANG Min
School of Information Science and Engineering, Central South University, Changsha 410083, China

Abstract: A malicious network traffic detection method based on multi-level distributed ensemble classifier was proposed for the problem that the attack model was not trained accurately due to the lack of some samples of attack steps for detecting attack in the current network big data environment, as well as the deficiency of the existing ensemble classifier in the construction of multilevel classifier. The dataset was first preprocessed and aggregated into different clusters, then noise processing on each cluster was performed, and then a multi-level distributed ensemble classifier, MLDE, was built to detect network malicious traffic. In the MLDE ensemble framework the base classifier was used at the bottom, while the non-bottom different ensemble classifiers were used. The framework was simple to be built. In the framework, big data sets were concurrently processed, and the size of ensemble classifier was adjusted according to the size of data sets. The experimental results show that the AUC value can reach 0.999 when MLDE base users random forest was used in the first layer，bagging was used in the second layer and AdaBoost classifier was used in the third layer.

Key words: malicious network traffic, attack detection, attack phase, network flow clustering, ensemble classifier

1 引言

由于大量的网络用户和高速增长的云服务数量，网络安全大数据吸引了很多研究者[1-8]。网络恶意流量种类繁多且数量庞大，这种特性使得网络异常流量的检测必然属于大数据问题。除此以外，由于网络流量具有复杂性和实时性，导致网络流量的异常检测面临着一个巨大的挑战。网络恶意流量是互联网安全威胁之一，它的产生是由于网络黑客对网络中某个节点发起攻击。随着攻击技术的发展，

收稿日期：2017−10−12；**修回日期**：2018−07−19
基金项目：国家自然科学基金资助项目（No.61202495）
Foundation Item: The National Natural Science Foundation of China (No.61202495)

攻击场景越来越复杂，攻击的隐蔽性和多态性加深了攻击预防与检测的难度。

网络异常主要来自于网络/系统故障或者攻击行为。其中，系统/网络故障很容易发现并定位问题。但是，由于攻击行为具有隐蔽性、多样性和时间的非连续性，给网络异常检测带来困难。然而，如果尽早地发现攻击，做好防御措施就能够尽量地减少攻击带来的损害。许多研究者对网络异常检测展开了研究，提出了很多网络异常检测方法，包括基于单分类支持向量机和主动学习的方法[9]、基于主成分分析的方法[10]、基于时间序列分析的方法[11]和基于健壮多元概率校准模型的方法[12]等。

在网络异常流量检测领域中，近年来研究者们关注使用流量的统计特征在数据集中挖掘异常。目前，流量异常检测方法主要包括基于签名的异常检测、基于行为的异常检测、基于流量统计特征的异常检测。尤其许多研究者重点关注使用机器学习与数据挖掘的方法来分析异常流量统计特征，如最大报文大小、平均报文大小、TCP（transmission control protocol）数量等。然而，大多数的研究重点关注检测攻击类型，同目前的 IPS 一样，是针对单一攻击事件的检测。在真实情景下，攻击是有步骤可分的，攻击行为之间常常存在关联。在模型构建的过程中，如果缺失某个攻击步骤的样本，就会导致模型的训练不够准确。

目前，有许多研究者关注网络大数据的安全问题，并在大数据背景下提出了很多异常检测方法。例如，Zhang 等[5]提出基于离群点的异常检测方法（A-SPOT, adaptive stream projected outlier detection），使用的研究数据集是 KDD99。KDD99 是经典的入侵检测数据集。A-SPOT 在大量的高维度数据中首先构建子空间聚类模式，在子空间中检测异常。最近，在大数据集的处理问题上使用集成分类器得到了研究者的广泛关注。

分类器主要包含 2 种类型，基分类器和集成分类器[13]。常见的基分类器有 BayesNet、J48、SMO 等。集成分类器指的是结合简单基分类器构建的分类器模型，例如随机森林、AdaBoost。在人工智能和数据挖掘领域，研究者们已经提出许多创建集成分类器的方法，集成分类器同时也是一种多级分类器[14-15]。

传统的集成分类器使用单一的基分类器，并在集成分类器的生成阶段，使用所有的基分类器来处理数据，收集每个基分类器的输入并结合这些输入进行最后的决策。例如，随机森林就是一种传统的集成分类器。随机森林自动生成随机树集合，并联合所有随机树的结果共同决策。尽管研究者已经提出多级分类器，但这些方法在系统中使用的都是基分类器，没有把集成元分类器考虑到多层系统中[16-18]。

本文针对目前网络大数据环境攻击检测存在的问题，以及现有集成分类器存在的不足，设计了一个 N 层多级分布式自动化集成分类器（MLDE, multi-level distributed ensemble classifiers）来检测大数据环境下的网络恶意流量。MLDE 集成框架在非底层迭代使用集成元分类器，且每层使用的集成元分类器不同，而在最底层使用基分类器。整个 MLDE 集成框架自动将每层的分析结果合并起来传输给上一层进行分析，最后由顶层分类器实施最终的决策，实现框架的自动化。在整个 MLDE 集成框架构建过程中，不需要人为干预。在面临大数据处理时，这种特性使得 MLDE 框架的构建简单。另外，这种迭代分层的协作方法能够以一种并发的方式处理大数据集，同时能够根据数据集的大小来调整集成分类器的规模。

在训练 MLDE 之前，对数据进行一系列的预处理。首先，从数据流中提取统计特征，然后对特征进行预处理，例如对其进行离散化等。由于网络恶意流量是复杂的，常常包含多个攻击阶段，例如，在整个 DDoS（distributed denial of service）的攻击活动过程中，常常包含多个攻击步骤，如扫描、安装特洛伊木马、发起 DDoS 等。不同的攻击阶段包含不同的网络流量统计特征。如果将所有的攻击阶段数据流放在一起分析，很难训练出精确的攻击检测模型。因此，在训练 MLDE 之前，采取了无监督学习模式来预处理数据集，将数据集划分到不同的簇，使得每个簇的数据集尽可能属于一个攻击阶段。另外，由于网络恶意流量中常常会包含噪音，因此在预处理阶段采用种子扩充算法去除这些噪音流。

本文的主要目标是为大数据环境下的网络恶意流量开发 MLDE 分类器，总的来说本文的贡献包括以下几个方面。

1) 提出了基于攻击阶段无监督学习的异常检测框架。

2) 提出 N 层 MLDE 集成分类器对网络恶意流量进行检测。MLDE 分类器是针对大数据环境设计的，当数据集较小时，MLDE 分类器能够转化为仅仅使用基分类器或者仅使用整个分类器的一部分

对数据集进行预处理。

3) 使用种子扩充算法移除噪音，提高分类器的正确性。

2　基于攻击阶段的恶意网络流量检测框架

2.1　攻击阶段相关定义

本文研究内容包括基于异常流量的数据预处理技术、攻击阶段搜索方法和基于攻击阶段的异常流量集成检测技术 3 部分。在进行这些研究之前，需要对攻击阶段的相关概念进行定义。

目前经典的对网络攻击阶段的定义来自著名的攻击链模型（cyber kill chain model）[19-22]。深入理解这个模型能够帮助安全机构在遇到安全威胁时选择正确的安全防御措施。攻击链模型将攻击分为 7 个攻击阶段。

1) 踩点（reconnaisance）：在这个阶段，攻击者通过对公开的可得到的信息进行研究，对攻击目标进行相关决策。

2) 武装（ weaponization）：为了攻击目标，攻击者收集攻击需要用到的设备和工具。收集到的信息越多，则越有利于攻击的进行。攻击的形式包括利用 Web 应用程序漏洞（Web application exploitation）、现有的或自己开发的恶意软件（ off-the-shelf or custom malware）、复合文档漏洞（compound document vulnerabilities ）、水坑攻击（ watering hole attacks）等。

3) 投送（delivery）：攻击者通过邮件或者其他的方式发送恶意载荷（payload）到目标。

4) 攻击（exploitation）：执行攻击载荷。

5) 安装（installation）：该阶段的操作可能需要较久的时间，安装对象可能是一个恶意软件或可执行代码。安装的恶意软件或可执行代码能够打开一个通道使得外部机器或网络设备能够访问目标。而且，这个通道需要在一定的时间内保持活跃。

6) 命令与控制（command and control）：攻击者创建一个命令与控制通道，这个通道能够持续的控制和操作已经能够访问的目标机器。

7) 收割（action on targets）：在攻击者能够访问目标机器后，可能进行一次或多次的漏洞利用，直到攻击者的目标已经实现。

除了经典的攻击链模型对攻击阶段的定义外，很多研究者在自己的研究工作中也对攻击阶段进行了归纳。例如，绿盟将攻击分为 5 个阶段，分别是侦查阶段、定向攻击阶段、“攻陷+入侵”阶段、安装工具阶段和恶意活动攻陷阶段。侦查阶段即是对网络、系统、端口、漏洞等进行探测扫描，了解目标信息。定向攻击阶段利用堆栈漏洞、Web 漏洞、逻辑配置漏洞、内存漏洞等对主机实施渗透攻击。当攻击进入“攻陷+入侵”阶段，表示主机已经被成功攻陷，攻击者可以做进一步的系统权限提升或者进一步攻击目标系统中的其他服务，如 ftp 登陆、telnet 密码破解等。安装工具阶段即是在目标机器上安装恶意软件，比如安装木马程序，通过这些恶意软件实现对目标的持续控制。在恶意活动阶段，攻击者实现此次攻击的最终目的，收取最终利益。

使用聚类的方法分析异常流量，即是尝试从网络流量相似性的角度来划分攻击阶段。虽然属于同一攻击阶段的流量具有相似性，但是并不是所有属于同一攻击阶段的流量都绝对是相似的。例如，扫描流量属于攻击阶段流量，针对大部分普通端口的扫描流量是相似的，但是这些普通端口的扫描流量与扫描 http（hypertext transfer protocol）服务和 ftp 服务的流量是不相似的。所以，本文依据攻击粗细粒度的不同对攻击阶段等概念作如下定义。

1) 攻击场景（attack scenario）：无论攻击是否成功，一次完整的攻击活动称为攻击场景，它由多个攻击阶段组成。

2) 攻击阶段（attack stage）：不同的攻击阶段在攻击场景中完成不同的任务。不同的攻击阶段之间具有逻辑关系和时间先后顺序。每个攻击阶段由多个元攻击阶段组成。参考大量文献后，本文将攻击场景划分为 5 个攻击阶段，并对一次成功或失败的攻击场景所依赖的攻击阶段组合进行总结，如表 1 所示。在真实情景下，收集到的某一攻击类型的攻击阶段组合是其中的一种。

在这里攻击的 5 个阶段分别是：侦查阶段、扫描阶段、获取目标权限阶段、控制目标阶段和发起攻击阶段。侦查阶段即是攻击者通过各种渠道尽可能多地了解目标。侦查阶段所采用的手段包括：社会工程、互联网搜索、域名管理/搜索服务、搜寻垃圾数据、非侵入性的网络扫描等。为了寻找网络环境中的漏洞，攻击者针对目标以及目标周边的网络环境实施带有侵入性地扫描。扫描内容包括：开放的端口、开发的应用服务、操作系统/应用程序漏洞、保护性差的数据传输链路、局域网/广域网设备的品牌和型号等。获取目标权限阶段意味着攻击已经开

表 1 攻击阶段组合

序号	侦查阶段（踩点阶段）	扫描阶段（武装阶段）	获取目标权限阶段（投送、攻击阶段）	控制目标阶段（安装，命令与控制阶段）	发起攻击阶段（收割阶段）	攻击结果
1	√					失败
2		√				失败
3			√			成功
4				√		成功
5					√	成功
6	√	√				失败
7		√	√			成功
8			√	√		成功
9				√	√	成功
10	√	√	√			成功
11		√	√	√		成功
12			√	√	√	成功
13	√	√	√	√		成功
14		√	√	√	√	成功
15	√	√	√	√	√	成功

始，可以通过电子邮件、即时通信软件、社交网络或是应用程序/系统漏洞等获取目标权限。控制目标阶段即是保持攻击者与目标主机或者肉机连接，常见的方式如安装木马程序等。发起攻击阶段即是开展一系列目标主机的非正常活动。

2.2 框架细节

本文提出基于攻击阶段的异常流量自动检测框架如图 1 所示，主要包括 6 个步骤。

1) 训练离散标准：首先人工标记训练集为 5 个攻击阶段，并利用标记后的流量训练有监督的离散标准。

2) 离散化：利用训练好的离散标准离散化正常流量数据集和异常流量数据集，然后对离散后的数据集进行二元化。

3) k-means 聚类：利用 k-means 对“混合训练集”进行聚类，为扩展实验 k-means 算法的输入参数从 1 扩展到 8。其中，混合训练集指数据集中包含正常流量和异常流量。

4) 去噪音：使用种子扩充算法移除 k-means 聚类结果中的噪音[23]。

5) 训练攻击阶段模型：利用去噪音后的 k-means 聚类结果, 训练多层集成分类器 MLDE 模型。

6) 检测：验证模型精度。

本文所有实验采用的数据集来自麻省理工学院林肯实验室网络安全信息科学组[24]。基于攻击阶段的异常流量自动检测框架的第一步是训练离散标准。在训练离散标准之前，根据数据集的说明文档将其人为地标记为 5 个阶段，即搜索阶段、扫描阶段、获取目标权限阶段、控制目标阶段和发起攻击阶段。

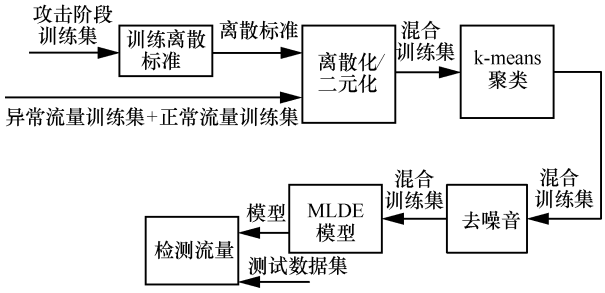

图 1 基于攻击阶段的异常流量检测框架

3 特征提取以及数据预处理

3.1 特征选择

本文使用异常流量的统计特征来研究异常流量的分析与检测。Wang 等[25]将 IP 数据流定义为互联网中 2 个节点之间的所有 IP 分组。IP 数据流是异常流量分析与检测的基本单元。本文使用五元组定义 IP

数据流：源 IP、源端口、目的 IP、目的端口和协议。

IP 数据流由一个 N 维特征向量描述，即 $X=\{x_1,x_2,\cdots,x_N\}$，其中，N 是特征属性的数目，x_i 指从单一 IP 数据流提取出的某个特征属性。在整个异常流量的分析中，并不考虑数据流的方向。Andrew 等[26]提出了 248 个可用在流量分类与聚类的复杂统计特征。他们的工作对流量分类的贡献巨大。然而，在流量分析中，选用大量的特征会造成巨大的计算开销。已有研究展示了某些少量的特征组合在流量分类中拥有足够的区分度[27-28]。类似的，本文对这些常见的被选用的特征进行区分度探讨。

特征选择和转换在机器学习中扮演了一个重要的角色。Comar 等[29]使用提取自 Narus 语义流量分析器（STA, semantic traffic analyzer）的少于 248 个特征的 108 个属性研究恶意软件的检测。其研究对象包括蠕虫、木马软件等流量的检测。他们提取的部分流量统计属性包括：数据分组的数目、有效载荷的字节大小等。使用这些流量统计属性进行分类面临的挑战包括：1) 同时处理连续值和离散值；2) 特征的关联特性；3) 缺失值的处理。选择越多的统计特征并不意味着对分类越好，有许多关于特征选择的文章对这点进行了论证。但是，Comar 等的工作证明了这 108 个统计属性在异常流量分析中具有区分度。本文的目标是在这些属性中发现最具有区分度的流量统计属性。

Lim 等[30]讨论了流量特征及其组合在网络流量分类中的区分度。他们将流量特征进行分组，然后使用机器学习算法对这些特征组合及特征进行评估，以此找出最具有区分度的属性度量，并对这些属性的区分度进行排名。研究结果显示分组大小和端口信息在所有的算法中有着最高的精度。除此而外，还进一步分析了不同数目的数据分组的大小的区分度，例如，最大和平均分组大小总体上比其他单一的独立分组大小的精度更高。

根据上述研究者的工作，可以得出结论：统计信息 pkt size，pkts，bytes，duration，tcp flag 在流量分类中拥有区分度。结合 Comar 等人的研究，本文选择的统计特征如表 2 所示。

3.2　数据预处理

异常流量的统计属性通常是连续值。为了使得挖掘更为有效，常在预处理阶段中使用数据变换策略[31]，这些策略包括：光滑策略、属性构造策略、聚集策略、离散化策略、二元化策略等。由于网络异常流量的属性值差别很大，直接使用连续的属性值进行分类聚类会导致分类精度不高。因此首先需要对其进行预处理。已有研究表明离散化对于分类有很大的影响，它能够进一步提升分类的准确度[30]。受此启发，本节主要研究离散化对于聚类的影响。

表 2　流特征属性

特征	描述
pkts	报文总数
pkt_noPayload	无负载报文总数
bytes	传送的字节总数
pay_bytes	所有负载的字节总数
duration	流持续时间
maxsz	最大的报文尺寸
minsz	最小报文尺寸
avfsz	平均报文尺寸
stdsz	报文大小的标准偏差
maxpy	最大的负载尺寸
minpy	最小的负载尺寸
avgpy	平均负载尺寸
stdpy	负载尺寸的标准偏差
synflag	SYN 的数目
rstfalg	RST 的数目
pushflag	PSH 的数目
finflag	FIN 的数目
ackflag	ACK 数目
syn_ackflag	SYN_ACK 的数目

离散化指的是将连续值划分为多个区间，并给划分区间打上区间标签或者概念标签。在这个过程中，连续的数值型属性离散成标称型属性。离散化方法包括有监督的离散方法和无监督的离散方法，典型的有监督的离散方法是 Ent-MDL。这个方法基于最短描述法则，同时也是基于熵的离散化方法。在离散化划分过程中，Ent-MDL 使用基于最短描述法则的方法自上而下停止基于熵的划分。典型的无监督离散化方法是分箱。分箱在离散化过程不需要使用类信息，所以是无监督的离散化方法。这是一种自顶向下的分裂方法，包括等宽分箱和等频分箱。等宽分箱指的是对连续值进行划分，每个划分区间的长度相同。等频分箱指的是每个分箱的数据值个数相同。分箱后使用箱均值、箱中位数、箱边界替换箱中的每一个值，可将属性离散化。

接下来通过实验对比有监督的离散化方法和

无监督的离散化方法对聚类的影响。实验所采用的数据集是攻击阶段流量，采用的聚类算法包括 k-means、EM、farthestfirst、canopy，采用的有监督离散化方法是 Ent-MDL，采用的无监督离散化方法是分箱。如图 2~图 5 所示。横坐标表示聚类的数目，纵坐标表示未正确聚类实例的百分比。这个百分比越小表示聚类准确度越高，聚类效果越好。

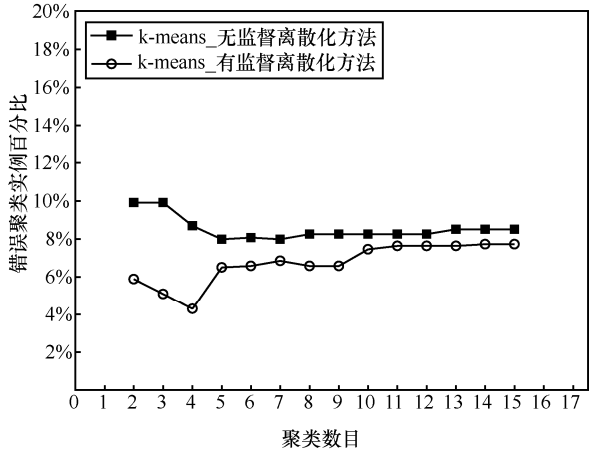

图 2　不同离散化方法对 k-means 影响

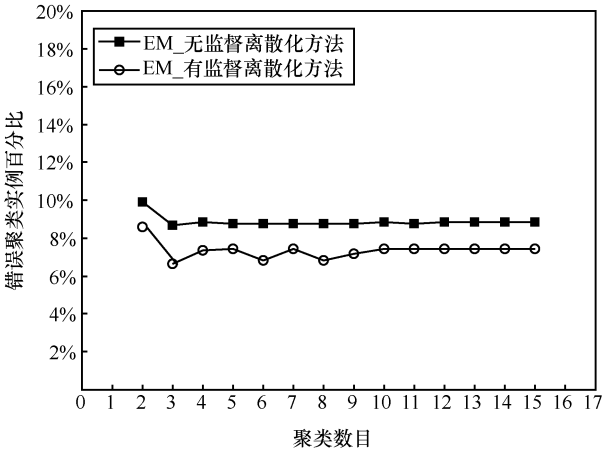

图 3　不同离散化方法对 EM 影响

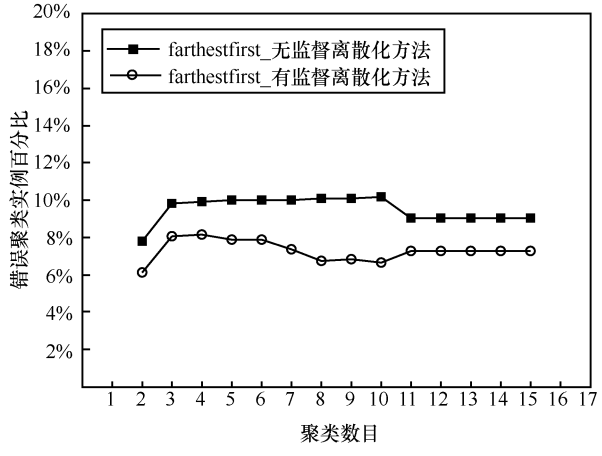

图 4　不同离散化方法对 farthestfirst 影响

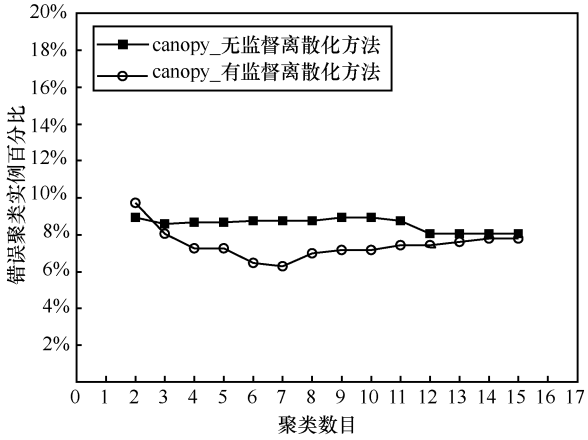

图 5　不同离散化方法对 canopy 影响

从图 2~图 5 中可以看出，使用无监督离散化方法离散的数据集对聚类算法进行评估，其未正确聚类实例的百分比高于使用有监督的离散化方法离散的数据集，即使用有监督的离散化方法处理数据集能够取得更好的聚类结果。基于此，本文在处理异常流量时采用有监督的离散化方法来离散数据集。

3.3　数据聚类

正如前面所提，网络恶意流量包含多个阶段，因此本文首先应用 simple k-means 将流量聚成 5 类。simple k-means 是经典 k-means 算法，Weka 实现了该算法，在文献[32]进行了详细地描述。由于网络恶意流量中经常包含了一些噪音流量，而 simple k-means 能很好地聚成 5 类，但是并不能去除其中的噪音流量，因此本文进一步对每一类应用种子扩充算法[23]去除其中的噪音流量。种子扩充算法的目标是发现那些与簇中其他节点距离均比较大的节点，并把它们作为噪音流。具体对于种子扩充算法的描述请见文献[23]，本文不做细述。

4　集成分类器结构

本文提出的 MLDE 框架理论上包含 *N* 层，是一种通用的多级分类器方法。但实际上由于计算机计算能力的不同，真实情景下，普通个人的计算机能实现的层次仅为 3 层或 4 层。考虑到面向大数据时框架的通用性，本文叙述 *N* 层 MLDE 框架的构建方法。

MLDE 框架如图 6 所示。MLDE 在不同的层次使用不同的分类器。第 1 层是最底层，最底层使用基分类器，如 BayesNet、J48、SMO（Sequential Minimal Optimization）等。而在第 2 层至第 *N* 层框架使用集成元分类器（ensemble meta classifier），包括 AdaBoost、bagging、dagging 等。在这里，第 *N*–1

层分类器是第 N 层中的一部分。第 N–2 层是第 N–1 层中不可分割的一部分。同上，底层是第 2 层中不可分割的一部分。图 6 中箭头的方向指示了数据流的方向。底层分类器使用基分类器来分析原始数据实例，并将分析结果上交给第 2 层分类器。从第 2 层分类器以后，每一层的分类器都使用集成元分类器。第 2 层的集成分类器收集来自第 1 层的基分类器的输出，并将输出的结果进行合并分析，然后将本层的输出结果传递给当前层次的上一层分类器，即第 3 层分类器。类似地，当第 N–1 层分类器收集到来自第 N–2 层分类器的输出结果，同时合并分析这些结果，将本层的结果传递给第 N 层/顶层分类器。最终，由顶层分类器做出最后决策。

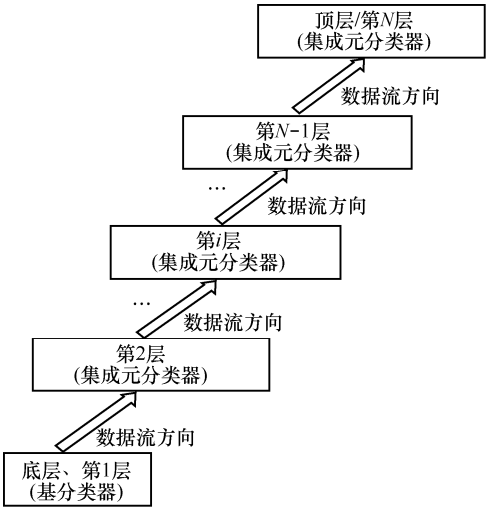

图 6　N 层 MLDE 框架

在构建 MLDE 时，框架的初始化方向即是从顶层到底层的分类器的调用方向，图 7~图 9 动态地展示整个初始化/调用过程。图中箭头指示的方向是调用方向/初始化方向，反映了分类器之间的关系。正如图 7 所示，每一层的分类器可以包含一个或多个分类器。第 N–1 层分类器由第 N 层分类器生成，即是 N–1 层分类器作为第 N 层分类器的输入参数进行初始化。第 N–1 层分类器由第 N–2 层分类器生成，此时第 N–2 层分类器作为第 N–1 层的输入参数进行初始化，如图 4~图 5 所示。类似地，依次从第 N 层、第 N–1 层……，直第 2 层使用集成元分类器进行初始化。当初始化最底层/第 1 层时，第 2 层将第 1 层分类器作为输入参数，且第 1 层的分类器使用基分类器。图 9 展示了初始化后的整个 MLDE 框架。

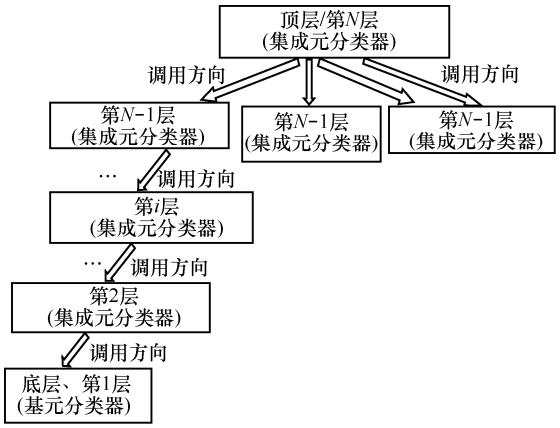

图 7　顶层调用关系

使用 MLDE 的优势是它能自动的构建整个系统，由于在每一层使用不同的集成元分类器，因此下一层的集成分类器作为上一层的集成分类器的参数，即下一层分类器作为上一层的集成分类器的一部分。这种构建方式使得模型的建立简单有效。在不同层次使用不同的集成元分类器，增强了整个框架的分类能力和并行处理能力。层次越多的 MLDE 框架在构建模型时，消耗的计算机内存越多。当数据集较小时，MLDE 可以由多层分类器退化为一级分类器。这样使得框架能满足不同规模大小的数据集的需求。

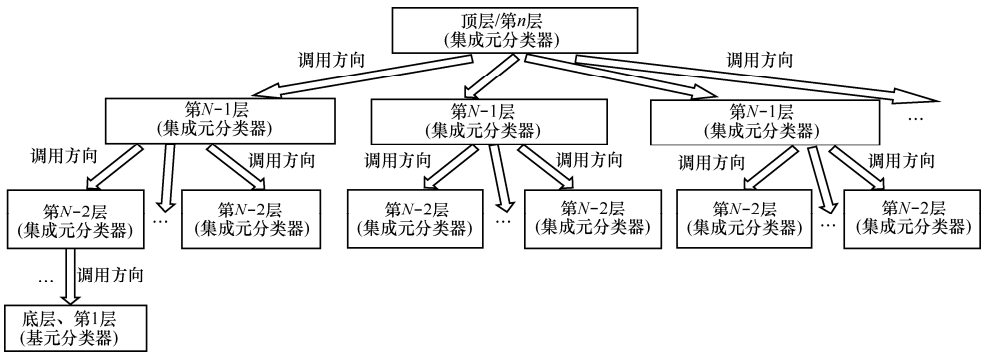

图 8　第 N 层到 N–2 层调用关系

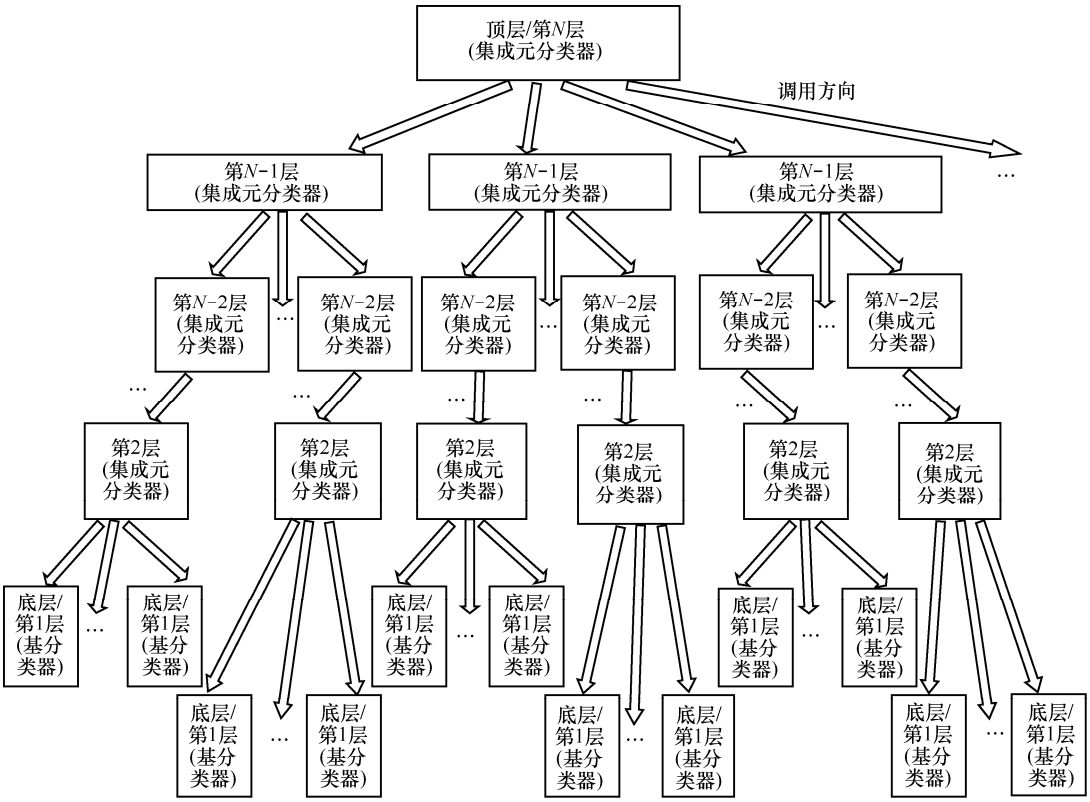

图 9　MLDE 调用关系

4.1　基分类器

经典的基分类器包括：SMO、FURIA、SPegasos、random forest、MLP、DTNB（decision table Naïve Bayes）、BayesNet、J48。本节对这些基分类器作一个简单的介绍。

SVM（support vector machine）算法的目的是解决一个大的二次规划问题。训练一个支持向量机需要消耗大量内存和运行时间。为了简单化支持向量机算法，Platt 提出 SMO 算法（sequential minimal optimization）[33]，这个算法将一个大的二次规划问题分成一系列小的只包含 2 个变量的问题。FURIA (fuzzy unordered rule induction algorithm) 是一个基于模糊规则的算法[34]，它是对规则学习器 RIPPER（repeated incremental pruning to produce error reduction）算法的修正和扩展。FURIA 使用模糊规则代替常规规则，使用无序规则代替有序规则。基于在线的次梯度投影方法 SPegasos（primal estimated sub-gradient solver for SVM）[35]。这个方法在处理大规模数据集问题上已经取得了较好的实验结果。

random forest 是一种集成的机器学习方法[36]，这个方法利用随机重复抽样技术和节点随机划分技术去构造多个决策树。最后的分类结果通过投票获得。MLP(multilayer perception)是多层感知器技术[37]。单层感知器技术能够完成线性分离数据的分类问题，但是不能够解决非线性问题。通过增加一个新的隐藏层次，MLP 能够解决非线性问题。DTNB 是一个结合决策表和朴素贝叶斯分类算法的组合分类算法[38]。贝叶斯网络(BayesNet)描述的是一组变量所遵从的概率分布，它通过一组条件概率来指定一组条件独立性假定。J48 算法是 C4.5 算法在 weka 中的应用[32]，是决策树算法。

4.2　集成分类器

本文同时也研究了部分集成分类器的性能，它们包括 stacking、bagging、AdaBoost、multiboost、grading、decorate、dagging 等。stacking 算法是 Wolpert 为叠加泛化提出一个机器学习框架[39]。算法由 2 阶段构成：基分类器的输出结果被用来作为第 2 阶段集成分类器的输入。集成学习算法被用来揭示如何更好地集成每一个基分类器的输出结果。

bagging 算法的主要思想是给出一个弱学习算法和训练数据集，弱学习算法对这个训练样本进行多次分类[40]。采用多数投票的方法确定最终的分类结果。通过这种方式，最终结果的分类精度将会被提高。AdaBoost 是 boosting 算法的一种[41]。该算法

使用不同的弱分类器对同一个训练数据集进行分类，然后把所有弱分类器集合起来，构成一个强分类器。multiboost 结合 bagging 算法和 AdaBoost 算法[42]。合并这 2 个算法的原因为：1) bagging 算法主要是减少方差，然而 AdaBoost 不仅减少方差，同时减少偏差。2) 在减少方差时，bagging 比 AdaBoost 更有效。

grading 算法试图标志和修正基分类器中的不正确预测[43]。这是一个元分类技术，它使用基分类器的预测作为元级属性，使用“分等级的”预测作为元级类。decorate 是一个集成同构分类器算法[44]。基本思想是不断地增加人工样本到训练集，以确保在增加一个新的分类器后集成分类器的精度不会减少。dagging 算法是一个有效的集成分类方法，尤其当独立的分类器有着糟糕的时间复杂度时[45]。这个元分类器在数据之外构造一系列的节点和层级，而且对基分类器的副本提供数据通道。算法通过联合投票规则将弱学习算法的输出结果结合起来。

5　性能评估与分析

理论上 MLDE 是一个 N 层框架，N 大于或等于 1。但实际上，由于计算机硬件水平的限制，N 并不能无限大。所以，在对 MLDE 框架进行评估的时候，本文重点研究在不同的层次使用不同的集成元分类器是否能够增强分类效果。本文实验的计算机配置为处理器 intel(R) core(TM) i7-6700HQ CPU @ 2.60 GHz、内存 8.00 GB、64 位操作系统。

实验主要测试基分类器和集成分类器在构建 MLDE 时对恶意流量和正常流量的分类效果。本文使用 WEKA 中的 simpleCLI 来生成和执行分类器，并采用 10 倍交叉的验证方式来评估分类器的效率[46]，评估的结果用常见的分类器的性能评估指标（area under curve or ROC area）进行展示。

首先比较针对网络恶意流量的几种不同的基分类器的性能。实验过程中不对数据集进行数据预处理，得到的基分类器结果如图 10 所示。从图 10 可以看出，random forest 取得最好的实验结果，其次是 J48 算法。

基于图 10 的结果，在基于攻击阶段的异常检测框架中有一个步骤是“k-means 粗糙聚类”，实验过程中，在 k-means 聚类的过程中输入聚类数目 1~8 作为输入参数，从而生成 8 组 k-means 实验结果。然后，分别基于这 8 组实验结果使用 random forest 进行分类，分类结果如图 11 所示。随着 k-means 聚类数目的增加，AUC 也跟着增加。当聚类数目达到 5 的时候，AUC 开始保持稳定。所以，采用聚类数目是 5 的 k-means 实验结果进行下一步分析。观察了实验结果数据，发现 k-means 聚出的 5 类基本对应于之前标记的攻击 5 个阶段。这主要是因为在聚类之前，本文首先训练了离散标准，即人工标记训练集为 5 个攻击阶段，并利用标记后的流量训练有监督的离散标准。最后实验利用种子扩充算法来移除网络恶意流量的噪音，结果如表 3 所示，移除噪音能够进一步提升分类效率。在本数据集，去除的噪音比例大约在 10%。

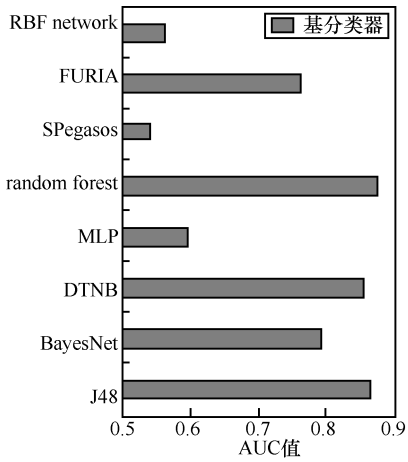

图 10　几种基分类器的评估结果

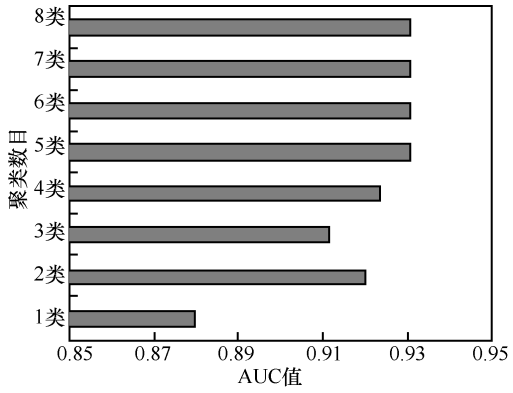

图 11　被聚类的数据集的 AUC 结果

实验还验证了多级集成分类器提高分类效果的能力。本文选择测试结果最好的 random forest 作为 MLDE 的第 1 层基分类器，然后再采用经典的集成分类器和使用较多的集成分类器放在 MLDE 的第 2 层与 random forest 组合在一起进行测试。测试组合包括：（AdaBoost，random forest）、（bagging，random forest）、（dagging，random forest）、（decorate，

random forest）、（grading，random forest）、（multiboost，random forest）、（stacking，random forest），括号中第 1 项表示 MLDE 第 2 层采用的集成分类器，第 2 项表示采用的基分类器。其中，组合（AdaBoost，random forest）在第 2 层分类器中表现最佳，如图 12 所示。从图 12 可以看出，组合（AdaBoost，random forest）、（bagging，random forest）、（dagging，random forest）、（decorate，random forest）取得了较好的实验结果。因此，在对第 3 层 MLDE 集成框架进行评估时，本文选用 AdaBoost、bagging、dagging 和 decorate 作为第 3 层的候选集成分类器，并对所有组合进行了测试。实验结果如图 13 所示。纵坐标表示第 2 层和第 3 层的集成分类器组合，其中左边的分类器是第 3 层分类器，右边的分类器是第 2 层分类器。在 3 层 MLDE 框架中，没有在不同层次中使用同一种集成元分类器。因为已有实验显示这种组合并不能提高分类效率，使得分类效果有所改善[16-17]。当 AdaBoost 在第 3 层中使用，bagging 在第 2 层中使用，random forest 作为第 1 层的基分类器时实验取得最好结果，AUC 达到了 0.99 以上。另外实验结果显示，由于 5 个阶段的流量有其各自的特点，因此识别难度是基本相同的。

表 3　　去噪音与未去噪音实验结果对比

	未去除噪音数据	种子扩充算法去除噪音之后的数据
AUC 值	0.9308	0.9432

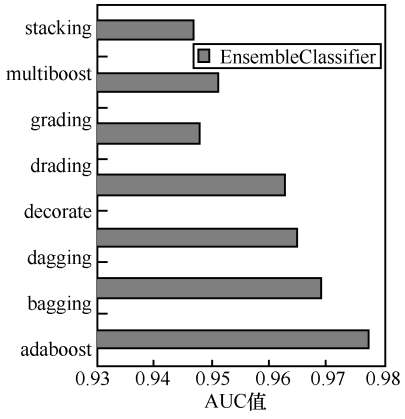

图 12　不同的集成分类器的 AUC 结果

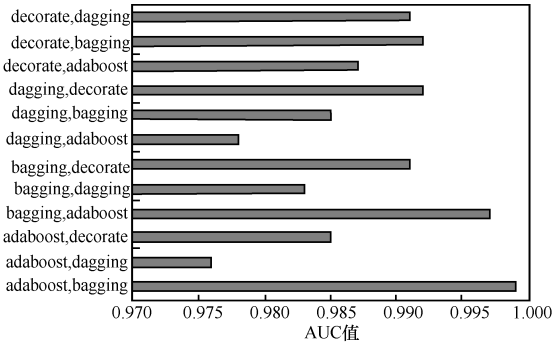

图 13　3 层 MLDE 实验结果

6　结束语

本文提出基于攻击阶段的异常流量集群检测技术（MLDE），并详细的解释了框架中的每个步骤，包括训练离散标准、离散化、k-means 粗糙聚类、去噪音、训练攻击阶段模型、检测等步骤。实验结果显示在 MLDE 框架中使用不同的集成元分类器组合能够进一步提升分类模型标识恶意网络流量的能力。其中，在所有的基分类器中，random forest 是最适合用来分类网络恶意流量数据集。比起其他的集成元分类器，AdaBoost 进一步提升分类效果的能力最强。在 3 层 MLDE 框架中，训练出的模型能够获得的最好 AUC 值达到 0.99 以上。

参考文献：

[1] MOKHTAR B,ELTOWEISSY M. Big data and semantics management system for computer networks[J]. Ad Hoc Networks, 2017, 57: 32-51.

[2] BROEDERS D, SCHRIJVERS E, SLOOT B VD, et al. Big data and security policies: towards a framework for regulating the phases of analytics and use of big data[J]. Computer Law & Security Review, 2017, 33(3):309-323.

[3] MANOGARAN G, THOTA C, KUMAR M V. MetaCloudDataStorage architecture for BIG DATA security in cloud computing[J]. Procedia Computer Science, 2016, 87: 128-133.

[4] XIA Y, CHEN J, LU X, et al. Big traffic data processing framework for intelligent monitoring and recording systems[J]. Neurocomputing, 2016, 181: 139-146.

[5] ZHANG J, LI H, GAO Q, et al. Detecting anomalies from big network traffic data using an adaptive detection approach[J]. Information Sciences, 2015, 318(C): 91-110.

[6] SARALADEVI B, PAZHANIRAJA N, PAUL P V, et al. Big data and hadoop-a study in security perspective[J]. Procedia computer science, 2015, 50: 596-601.

[7] WANG H, JIANG X, KAMBOURAKIS G. Special issue on Security, Privacy and Trust in network-based big data[J]. Information Sciences, 2015, 318(C): 48-50.

[8] SANCHEZ M I, ZEYDAN E, OLIVA A D L, et al. Mobility management: deployment and adaptability aspects through mobile data traffic analysis[J]. Computer Communications, 2016, 95: 3-14.

[9] 刘敬, 谷利泽, 钮心忻, 等. 基于单分类支持向量机和主动学习的网络异常检测研究[J]. 通信学报, 2012, 36(11): 136-146.
LIU J, GU L Z, NIU X X, et al. Research on network anomaly detection based on one-class SVM and active learning[J]. Journal on Communications, 2012, 36(11): 136-146.

[10] 钱叶魁, 陈鸣, 叶立新. 基于多尺度主成分分析的全网络异常检测方法[J]. 软件学报,2012,23(2):361-377.
QIAN Y K, CHEN M, YE L X. Network-wide anomaly detection method based on multiscale principal component analysis[J]. Journal of Software, 2012,23(2):361-377.

[11] 郑黎明. 大规模通信网络流量异常检测与优化关键技术研究[D]. 长沙: 国防科技大学, 2012.
ZHENG L M. Key Technologies research on traffic anomaly detection and optimization for large-scale networks[D]. Changsha: National University of Defense Technology, 2012.
[12] 李宇翀, 罗兴国, 钱叶魁, 等. RMPCM: 一种基于健壮多元概率校准模型的全网络异常检测方法[J]. 通信学报, 2015,36(11):201-212.
LI Y C, LUO X G, QIAN Y K, et al. Network-wide anomaly detection method based on robust multivariate probabilistic calibration model[J]. Journal on Communications, 2015,36(11):201-212.
[13] ABAWAJY J H, KELAREV A, CHOWDHURY M. Large iterative multitier ensemble classifiers for security of big data[J]. IEEE Transactions on Emerging Topics in Computing, 2014, 2(3): 352-363.
[14] ABAWAJY J, CHOWDHURY M, KELAREV A. Hybrid consensus pruning of ensemble classifiers for big data malware detection[J]. IEEE Transactions on Cloud Computing, 2015,PP(99):1.
[15] ISLAM R, ABAWAJY J. A multi-tier phishing detection and filtering approach[J]. Journal of Network and Computer Applications, 2013, 36(1): 324-335.
[16] ISLAM M R, ABAWAJY J, WARREN M. Multi-tier phishing email classification with an impact of classifier rescheduling[C]//Pervasive Systems, Algorithms, and Networks (ISPAN).IEEE,2009: 789-793.
[17] ISLAM R, SINGH J, CHONKA A, et al. Multi-classifier classification of spam email on a ubiquitous multi-core architecture[C]//Network and Parallel Computing. IEEE, 2008: 210-217.
[18] ISLAM MR, ZHOU W, GUO M, et al. An innovative analyser for multi-classifier email classification based on grey list analysis[J]. Journal of network and computer applications, 2009, 32(2): 357-366.
[19] RUTHERFORD J R, WHITE G B. Using an improved cybersecurity kill chain to develop an improved honey community[C]//International Conference on System Sciences. 2016: 2624-2632.
[20] MIHAI I C, PRUNA S, BARBU I D. Cyber kill chain analysis[J]. Information Security and Cybercrime,2014, 3: 37.
[21] DALZIEL H. Securing social media in the enterprise[M]. Amsterdam: Syngress Publishing, 2015: 7-15.
[22] WINKLER I, GOMES A T. Advanced persistent security[M]. Amsterdam :Syngress Publishing, 2017: 179-184.
[23] 汪洁,何小贤. 基于种子——扩充的多态蠕虫特征自动提取方法[J]. 通信学报, 2014,35(9):12-19.
WANG J, HE X X. Automated polymorphic worm signature generation approach based on seed-extending[J]. Journal on Communications, 2014,35(9):12-19.
[24] LINCOLN LABORATORY. 2000 DARPA Intrusion Detection Scenario Specific Data Sets[EB].Lexington: Massachusetts Institute of Technology,2000.
[25] WANG Y, XIANG Y, ZHANG J, et al. Internet traffic classification using constrained clustering[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 25(11): 2932-2943.
[26] MOORE A, ZUEV D, CROGAN M. Discriminators for use in flow-based classification[M]. London: Queen Mary and Westfield College,2005.
[27] CASAS P, MAZEL J, OWEZARSKI P. Unsupervised network intrusion detection systems: Detecting the unknown without knowledge[J]. Computer Communications, 2012, 35(7): 772-783.
[28] WANG Y, XIANG Y, ZHANG J, et al. Internet traffic clustering with side information[J]. Journal of Computer and System Sciences, 2014, 80(5): 1021-1036.
[29] COMAR P M, LIU L, SAHA S, et al. Combining supervised and unsupervised learning for zero-day malware detection[C]// INFOCOM, 2013 Proceedings IEEE. IEEE, 2013: 2022-2030.
[30] LIM Y, KIM H, JEONG J, et al. Internet traffic classification demystified: on the sources of the discriminative power[C]//International Conference. ACM, 2010: 9.
[31] HAN J W, KAMBER M , PEI J. Data mining: concepts and techniques, Third Edition[M]. 3rd ed. San Francisco: Morgan Kaufmann Publishing, 2011: 211-321.
[32] QUINLAN J R. C4. 5: programs for machine learning[M]. Elsevier, 2014.
[33] PLATT J C. Fast training of support vector machines using sequential minimal optimization[M]. Advances in kernel methods.MIT Press,1999: 185-208.
[34] HÜHN J, HÜLLERMEIER E. FURIA: an algorithm for unordered fuzzy rule induction[J]. Data Mining and Knowledge Discovery, 2009, 19(3): 293-319.
[35] SHALEV-SHWARTZ S, SINGER Y, SREBRO N. Pegasos: Primal estimated sub-gradient solver for SVM[C]//Proceedings of the 24th international conference on Machine learning. ACM, 2007: 807-814.
[36] BREIMAN L. Random forests[J]. Machine learning, 2001, 45(1): 5-32.
[37] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning internal representations by error propagation[R]. California Univ San Diego La Jolla Inst for Cognitive Science, 1985.
[38] HALL M A, FRANK E. Combining naive bayes and decision tables[C]//FLAIRS Conference. 2008, 2118: 318-319.
[39] WOLPERT D H. Stacked generalization[J]. Neural networks, 1992, 5(2): 241-259.
[40] BREIMAN L. Bagging predictors[J]. Machine learning, 1996, 24(2): 123-140.
[41] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm[C]//ICML. 1996, 96: 148-156.
[42] WEBB G I. Multiboosting: A technique for combining boosting and wagging[J]. Machine learning, 2000, 40(2): 159-196.
[43] SEEWALD A K, FÜRNKRANZ J. An evaluation of grading classifiers[C]// International Symposium on Intelligent Data Analysis. Springer-Verlag, 2001: 115-124.
[44] MELVILLE P, MOONEY R J. Constructing diverse classifier ensembles using artificial training examples[C]// International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. 2003, 3: 505-510.
[45] KAI M T,WITTEN I H. Stacking bagged and dagged models[C]// Fourteenth international conference on machine learning. Morgan Kaufmann Publisher Inc. 1997:367-375.
[46] WITTEN I H, FRANK E. Data mining: practical machine learning tools and techniques[M]. Amsterdam: Elsevier/Morgan Kaufman, 2011.

[作者简介]

汪洁（1980−），女，湖南桃江人，博士，中南大学副教授，主要研究方向为网络与信息安全等。

杨力立（1992−），女，布依族，贵州安顺人，中南大学硕士生，主要研究方向为网络与信息安全等。

杨珉（1993−），男，江西南昌人，中南大学硕士生，主要研究方向为强化学习等。