计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：基于随机森林的网络入侵检测方法
作者：苟继军，李均华，陈晨，陈一鸣，吕奕达
网络首发日期：2019-07-18
引用格式：苟继军，李均华，陈晨，陈一鸣，吕奕达．基于随机森林的网络入侵检测方法．计算机工程与应用.
http://kns.cnki.net/kcms/detail/11.2127.TP.20190718.1428.012.html

网络首发时间：2019-07-18 17:47:54
网络首发地址：http://kns.cnki.net/kcms/detail/11.2127.TP.20190718.1428.012.html

基于随机森林的网络入侵检测方法

苟继军[1]，李均华[1]，陈　晨[2]，陈一鸣[1]，吕奕达[2]

1.国网四川省电力公司经济技术研究院，成都 610041
2.西安电子科技大学,西安 710071

摘　要：为了提高网络安全水平，及时对网络攻击进行主动检测，本文提出了一种基于随机森林的网络入侵检测模型。该模型能够对大流量攻击进行分布式检测，且检测算法在引入了两个随机性后，即可降低网络流量内不同属性特征字段的噪声，并消除关联性，以便更为便捷、迅速的对攻击进行主动检测。最后，将经典的 Adaboost 组合多分类器方法与本文提出的算法在检测率，正确率，精确率三个方面进行对比，体现了本文算法的优越性，为大数据时代下网络安全提供了更好的保护。

关键词：网络安全；机器学习；　随机森林方法；攻击检测
文献标识码：A　　**中图分类号：**TN915.08　　doi：10.3778/j.issn.1002-8331.1903-0139

苟继军，李均华，陈晨，等.基于随机森林的网络入侵检测方法.计算机工程与应用
GOU Jijun, LI Junhua, CHEN Chen, et al. Network intrusion detection method based on random forest. Computer Engineering and Applications

Network Intrusion Detection Method Based on Random Forest

GOU Jijun[1], LI Junhua[1], CHEN Chen[2], CHEN Yiming[1], LÜ Yida[2]
1. State Grid Sichuan Economic Research Institute, Chengdu 610041, China
2. Xidian University，Xi'an 710071, China

Abstract： In order to improve the level of network security and detect network attack actively in time, this paper proposes a network intrusion detection model based on random forest.The model can perform distributed detection on large traffic attacks, and after introducing two randomness, the detection algorithm can reduce the noise of different attribute feature fields in network traffic and eliminate the correlation, so that it is more convenient and rapid to actively detect attack. Finally, the classical Adaboost combinatorial multi-classifier method is compared with the algorithm proposed in this paper in three aspects: detection rate, accuracy rate and precision rate, which reflects the advantages of this algorithm and provides better protection for network security in the era of big data.

基金项目：基于大数据的电力无线通信网络态势感知技术研究合同 国家电网公司科技项目资助(No.SGSCJY00GHJS1800016)。
作者简介：苟继军（1977—），男，高级工程师，硕士，主要研究方向为信息与网络安全；李均华（1971—），男，高级工程师，本科，主要研究方向为通信系统与安全；陈晨（1978—）男，博士生导师，副教授，主要研究方向为无线通信，物联网，E-mail：cc2000@mail.xidian.edu.cn；陈一鸣（1990—），男，博士，主要研究方向为通信系统与安全；吕奕达（1995—），男，研究生，主要研究方向为网络攻击检测，机器学习。

Key words： network security; machine learning; random forest method; attack detect

1 引言

随着信息社会的不断发展，以及大数据、云计算等新兴技术的应用，网络安全对于个人、企业、国家都有重要的意义。网络在提供给人们生活便利的同时，网络安全问题日益凸显，一旦出现安全事件，将造成极大的经济损失、社会影响。所以事前主动检测、防御，对于网络稳定、可靠运行具有重要的意义。近年来，由于网络攻击多发生于广域网环境，导致攻击检测工作面临挑战，因此要求研究人员必须结合大数据技术，研究新的攻击检测方法。一些传统的网络攻击流量检测方法和检测技术[1](如流量分析、特征提取、模式匹配和专家系统)在高速和大规模的互联网环境中很难高效准确地检测攻击，难以适应如今的大数据时代，必须对其进行改进。

针对这一问题，我们利用在各个领域都广泛应用、能进行自我学习、主动检测防御的机器学习算法，研究了基于随机森林的入侵检测模型，并且与近期流行的分布式并行计算框架 MapReduce 相结合，可有效的主动针对网络攻击进行检测，通过引入两个随机性，降低网络流量内不同属性特征字段的噪声，并消除了彼此的关联性，提高了攻击检测的检测率、正确率、精确率，在大数据时代背景下，为网络安全提供了更好的防御。

2 相关研究

针对攻击检测这一热点问题，国内外学者对网络流量异常做了各种研究，检测算法主体基于特征、统计、数据挖掘与机器学习等方式进行检测。

基于特征的检测算法核心是通过提前人为设定网络异常流量的一般特征，检测算法再进行特征匹配。论文[1]在分析网络流量的基础之上，针对时、空方面的流量特征进行深度探究，使用大数据挖掘技术，对通信网络流量的行为特征参数进行提取，然后对网络流量异常行为进行分布式检测，能够实时对异常信息进行反馈。论文[2]通过对比单窗口聚类异常检测算法的不足，对其进行改进，在每个单窗口中用优化的 k-means 算法对数据进行初步聚类检测，最终综合多个窗口的结果得出异常流量，增加了检测正确率，提高了检测的效率。

基于统计分析的检测算法核心是对网络流量数据进行分层次的统计分析，利用网络攻击与流量数

据统计量的对应关系来检测。论文[3]提出以统计为基础的异常流量检测方法，可在常见网络环境、且误检率偏低时，更为迅速的对攻击进行检查。但此时往往需对阈值予以确定，且阈值无法自动伴随流量模式的改变而调整。论文[4]无需对阈值加以确定，而直接在网络流量审计数据内，对检测规则进行总结。

基于数据挖掘攻击检测不需要人工设置异常阈值，能从网络流量中自我总结出检测规则，一般有支持向量机、聚类、决策树等算法。论文[5]表明，在持续更新数据的同时，数据挖掘可对检测模型进行微调，以便与全新异常行为相适应，提高了检测异常流量时的准确度。论文[6]表明，对比以统计分析为基础的异常流量检测方法，数据挖掘算法复杂度更高、且效率偏低。假如要在海量网络流量环境下适用，仍需对此进行优化。论文[7-8]均通过聚类方式来分析异常流量。前者选定流量特征直方图作为图像，以便进行聚类。后者则基于 Chameleon 算法，对全新聚类算法进行总结。论文[9-10]则对 SVM，即支持向量机算法加以选用，以便检测 BGP 异常流量。针对 BGP 数据样本较少、多变、且维度较高等特性，多通过数据预处理、调整参数等方式，以便与之相适应。论文[11]以及论文[12]通过贝叶斯、决策树算法来检测异常流量。论文[13]将卷积神经网络引入网络攻击检测，将日志信息特征提取到灰度图，通过大数据平台 spark 处理日志信息，不断迭代生成最新的特征，通过卷积对数据进行降噪。论文[14]是一篇关于机器学习不同方法在攻击检测上的应用，对典型的理论模型进行了各个优化参数的对比，对于选取对攻击检测来说更具优势的机器学习方法具有重要意义。论文[15]使用朴素贝叶斯分类器实现异常检测系统，结果表明没有产生误报，正常和异常识别率达到 98%和 89%。论文[16-17]利用无监督聚类算法，Blowers 等人使用基于密度的聚类算法 DBSCAN 对正常与异常流量进行分组。

随着网络数据量越来越大，特征越来越复杂，基于特征和统计的检测算法已不适应实际应用场景。分布式计算平台的出现，计算能力的增强，使得基于机器学习的检测算法越来越流行，这是未来主要发展和研究趋势。

目前，一些经典的机器学习算法，例如基于概

率的贝叶斯[18]，决策树，SVM 被证明在面对高维度、海量数据时，攻击检测中的准确率不高且误报率高，而随机森林能够处理高维度数据，并且不用做特征选择，模型泛化能力强。所以本文在各种机器学习算法的基础上，重点将经典的 Adaboost 组合多分类器方法与本文提出的随机森林算法在检测率，正确率，精确率三个方面进行对比，来体现本文算法的优越性。

3 MapReduce 原理概述

MapReduce[19]是一个非常流行的分布式并行计算框架。MapReduce 是 Hadoop 的两个核心之一，Hadoop 主要包括两部分，一是分布式文件系统 hdfs，二是分布式计算框 MapReduce。两者的联系是通过 MapReduce 在 Hadoop 平台上进行分布式的计算编程。MapReduce 核心思想是简单的多业务逻辑被复杂的任务执行过程分解。

我们可以将 MapReduce 分开来理解：映射是对集合中的每个目标应用相同的操作，也就是说，如果要将一个窗体中的每个单元格乘以二，那么将该函数单独应用于每个单元格的操作属于映射(它体现了移动计算而不是移动数据)；化简指的是为了返回综合结果遍历集合中的元素。也就是说，输出表单中的任务数量和任务属于化简。

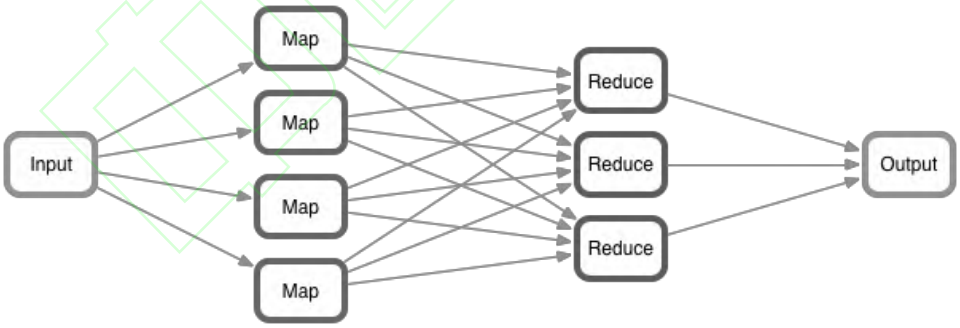

图 1 MapReduce 计算框架

MapReduce 计算框架如图 1 所示，简单理解，MapReduce 计算框架是把需要计算的数据放入到 MapReduce 中进行计算，然后返回一个我们期望的结果。所以首先我们需要一个源(计算的东西)，即输入(输入)，然后 MapReduce 通过定义一个好的计算模型来操作输入(输入)，并最终得到(期望的结果)

输出(输出)。在这里我们主要讨论的是 MapReduce 计算模型。MapReduce 计算模型如表 1 所示。

表 1　计算模型

函数	输入键值对	输出键值对
Map（）	<k1，v1>	<k2，v2>
Reduce（）	<k2，{v2}>	<k3，v3>

在运行一个 MapReduce 计算任务时，可划分任务进程为两部分，分别对应于 map 和 reduce。两输入输出均为键值对。总之，MapReduce 可以使程序并行执行，极大提高检测效率。

4　组合分类器分类方法

目前，大多攻击检测算法都应用了集成方法，该方法具有优越的泛化性能，很适合实际的攻击检测场景。集成学习是通过构造和组合多个学习器完成的，因此称为多分类器系统。集成学习比单一学习器更优越的泛化能力，它使通过将多个学习器进行结合而实现的。通过多分类器组合完成分类等，对比单一分类器而言，其往往更具优势，且泛化性更为突出，通过对单一分类器存在的局部行为的整合即可令整体分类性能变得更高。而就单一分类器学习算法而言，分类能力不足，在问题所处的分类空间里很难找到最优分类。 单分类器经典的算法有决策树，支持向量机等。多分类器算法有 Adaboost 和随机森林算法。通过以往的研究，基于决策树的基分类器存在过度拟合的问题，不能很好的解决噪声等问题，而 Adaboost 算法构成强分类器，能很好的解决过度拟合的问题，一般应用在二分类或者多分类场景，但缺点在于训练耗时，易受噪声数据影响。而基于随机森林的组合分类器既能有效解决过度拟合的问题，又对噪声数据有较强的健壮性。通过下文介绍，可以清晰的看出从单分类器到组合分类器，再到基于随机森林的组合分类器，检测算法逐渐优化的过程。

4.1　基于决策树的基分类器

作为分类方法中的一种，决策树[20]一般结合学习训练数据集，以便树得以形成，其可有效的分类测试数据集，若如下的训练数据集已经给定：

$$D=\{(x_1,y_1),(x_{2,}y_2),.....,(x_n,y_n)\} \quad （1）$$

那么，$x_i=(x_i^1,x_i^2,.....x_i^n)^T$ 为输入样本特征向量、n 以及 N 分别对应于特征数量、样本容量，y_i 即类标记。决策树主要包括有向边的层次结构、节点。决策树内的节点共包括三类：分别为内部、根以及叶节点。最初，当树得以生成时，根节点上将

聚集数据，此后结合递归方式来排序数据。决策树在选定分类属性时，一般选取最短属性。信息增益则是其中最为关键的选择准则。就决策树而言，需避免过拟合的问题。过拟合，即当噪声数据产生于训练样本数据集内时，而因训练样本、随机错误记录不多，往往会导致代表性样本无法生成。数学定义见如下，若此时有 H 这一空间存在，针对 $h \in H$ 这一假设，若有 $h` \in H$ 等其它假设存在，导致训练样本 h 对比 h'时，具备更低的错误率，但实际后者错误率明显小于前者，此时即为假设 h 过拟合训练样本数据。而使用组合多分类器是可以有效解决过拟合问题。

4.2 基于 Adaboost 的组合多分类器

Adaboost[21]是用于训练相同训练集的不同分类器的迭代算法。一般通过对数据分布的改变，最终来实现 Adaboost。本算法的一大亮点，即通过关联弱分类器、带权分类误差。需分离样本权重与数据集对照后的误差，由此可获得弱分类器样本权重分类误差，即带权分类误差。如下即为相关表示：

$$\varepsilon_i = \sum w_i(k) I[c_i(k) \neq y_k] \qquad (2)$$

此处，与 i 个弱分类器互为对应的带权分类误差，即针对 k 这一样本进行分类的结果，y_k 即与样本 k 对应的标签，$I[c_i(k) \neq y_k]$是一个指示函数，表示为：

$$I[c_i(k) \neq y_k] = \begin{cases} 0, c_i(k) = y_k \\ 1, c_i(k) \neq y_k \end{cases} \qquad (3)$$

结合上述公式，由此可得到，此种误差往往受到训练集内样本权重、弱分类器的样本误差等方面的影响。

弱分类器的权重与它自身的带权分类误差有关，表示为

$$a_i = \frac{1}{2} \ln(\frac{1-\varepsilon_i}{\varepsilon_i}) \qquad (4)$$

进而可认识到，以[0，1]作为该误差的范围，且与弱分类器权重之间为反比关系，也就说明，弱分类器权值伴随带权分类误差的减小而增大，反之，即减小。

更新训练样本权重的公式参见如下：

$$w_{i+1}(k) = \frac{w_i(k)}{sum(w_i)} \times \begin{cases} e^{-a_i}, c_i(k) = y_k \\ e^{a_i}, c_j(k) \neq y_k \end{cases} \qquad (5)$$

即当对 Ci 的权值进行获取后，再更改样本 k 的权重。当其能够准确的完成样本 k 分类时，即将样本权值降低。如果不这样做，则需提高权值。一般可划分算法为两部分，即训练以及分类阶段。

运用此种算法，可令过度拟合现象得以避免，但也存在欠缺之处。比如训练阶段时间较长,无法对随机模型的实现予以表示，因此，噪声数据对此算法产生影响较大。

4.3 基于随机森林的组合多分类器

就机器学习领域来看，随机森林[22]属于备受关注的一类集成学习技术，在入侵检测等领域已得到大量运用。首先，通过 Bootstrap 重采样方法的运用，自原始样本内对诸多子样本进行提取，再通过决策树来完成建模工作。此后，综合不同决策树预测。最终基于投票算法，获得最终的预测结果。它是一组树分类器，每个分类器都是一个分类回归树，并且不使用 CART 算法进行修剪。在针对不同数字构成的森林输出策略进行分类时，结合多数表决方法来完成，如下即为具体表示：

$$TC:\{C(x,a_i),i=1,2,.....,n\} \quad (6)$$

其中，x 是输入向量，a_i 是独立同分布的变量。

随机森林的特征：1.通过对每个节点的随机选择特征完成分类，即令决策分类树的关联性得以最小化，提高分类准确度，并有效地解决过拟合问题;2.可以预测属性特征在分类中的重要性，并且单个决策树增长得更快;3.针对噪声数据，鲁棒性较为突出。

随机森林属于组合决策树模型，首先通过 BACGUG 方法的运用，重建原始训练集，然后使用随机特征选择方法生成每个重组训练集的决策树。随机森林中个体分类器的多样性不仅取决于数据样本的干扰，而且还取决于属性特征的干扰。通过增加个体分类器之间的差异程度，使得最终分类器的泛化性提高。

在形成多个决策树之后，运用本算法，即可从单一向组合多分类器转换。若树木、森林具备更为突出的关联性，则森林分类相对更差。随机森林与单一决策树相比的优势在于避免过拟合，高分类精度和良好的稳定性。随机森林比其他组合多分类器集成方法对数据噪声更稳定。就组合决策树的全部树而言，它是对概率分布形成的随机变量进行运用。在对精度进行检测时，以弱相关、低精度作为关键条件，所以为实现低偏差，需确保树生长的深度达到最大。为了使弱相关得以实现，需确保应用随机化。随机变量的运用，可令决策树在分类方面变得更为准确，减少每个决策树之间的相关性，最终令森林整体的分类性能得以提升。结合如上分析，由此认识到，通过在单一分类器内引入两种随机性，以此使得过拟合现象得到解决，即为随机森林。以下是随机森林的几个重要数学概念，通过这几个随机森林的定理及分析，为我们之后如何更好

的优化基于随机森林的检测模型指明了方向。

边缘函数表示为：

$$mg(X,Y)=av_k(I(h_k(X)=Y))-\max av_k(I(h_k(X)=j))\quad(7)$$

此处 I()即对应示性函数，Y 以及 j 分别对应于正确、非正确的分类向量，av_k 即对应于取平均值，如下即为泛化误差：

$$PE=P_{X,Y}(mg(X,Y)<0)\qquad(8)$$

其中，下标 X，Y 表示概率的定义空间。

森林中所有决策树的泛化误差都收敛于：

$$\lim_{n\to\infty}PE=P_{xy}(P_\theta(K(X,\theta)=Y)-\max P_\theta(K(X,\theta)=j)<0\quad(9)$$

其中，n 表示森林中决策树的个数。

该定理表明：在决策树数量不断增大的同时，泛化误差随之达到稳定上界。表明随机森林存在扩展性，且避免过拟合的作用较为突出。如下即为对应上界：

$$PE\le\frac{\overline{\rho}(1-s^{2)}}{s^2}\qquad(10)$$

在此之中，$\overline{\rho}$ 以及 s 分别对应各决策树相关度的均值、平均强度。

上述定理表明：随机森林的泛化性能取决于，各决策树之间的相关性越小，单棵分类树的分类能力强这两个因素。

4.4 算法分析

随机树模型能够对大流量攻击进行分布式检测，处理高维度的数据，训练速度快，容易做成并行化方法(训练时树与树之间是相互独立的)，并且不用做特征选择(因为特征子集是随机选择的) 。在训练完后，它能够给出哪些特征比较重要，模型泛化能力强，如果有很大一部分的特征遗失，仍可以维持准确度。它有两大随机性即可降低属性特征字段的噪声、且关联性也得以消除，以便确保检测的准确性。两个随机性具体指：

1.选定训练数据样本的随机性。随机森林训练数据样本的生成，一般结合随机采样返回来完成，且全新的数据集属于原始数据集的子集。因数据集存在差异性，导致新形成的森林也存在不同。所以，森林决策树即形成了随机的增长过程。

2.选定特征属性变量的随机性。随机森林算法，导致选定特征属性变量具备更强的随机性，在构建决策树、并进行生长时，无需开展剪枝工作。如此，特征属性变量即令分类精度得以明显提升。降低了森林中决策树之间的相关系数。

相比于 Adaboost 算法，其并不过分依赖于单一

分类器，噪声数据对此的干扰相对较小。

5　攻击检测模型

因决策树一般通过单个树来完成决策树分类，其在诸多单检测模型之中发挥关键作用。但却难以实现于分布式检测模型内。因此，此处即对以随机森林算法为基础的分布式检测模型进行检测方式、

5.1　分布式检测模型

此种模型可划分为四大块，如分布式分类检测、数据采集模块等，分布式检测模型具体如图 2 所示：

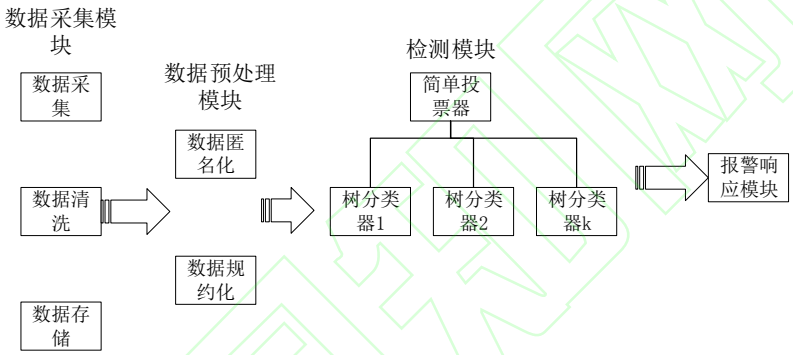

图 2　分布式检测模型

1.数据采集模块。通过对网络流量抓包技术的运用，以便对进行检测的数据流、数据包进行抓取。

2.预处理模块。需要对数据或数据流进行预处理的数据预处理方法，如数据匿名性和数据规范化操作等。

由于收集的网络数据是裸数据，这些数据通常无法通过细粒度统计分析和数据挖掘进行有效分析和攻击检测。

3.分布式分类检测模块。自节点上对部署攻击、正常流量的分类检测子模块进行总结，此后再向主节点汇总检测结果，并运用简单多数表决方式，以便得到分类结果。本模块均将 CART 算法运用于单一决策树分类器，并并行在分类进程中执行，所以具备相对独立性。CART 算法则对两点递归分割技术加以运用，此种技术划分样本集为两大样本子集，然后将它们分开以形成简单的二叉树。Gini 指数用于在分裂时测量 CART 算法的杂质。对于决策树的节点，Gini 索引计算如下 Gini 指数计算公式如下

$$Gini(t) = 1 - \sum_k [p(c_k | t)]^2 \quad (11)$$

其中，Gini 指数，即 1 与类别 c_k 概率平方之和的差值，其对样本集合不确定程度予以体现。一般而言，该数值较大时，则样本集合具备更为突出的不确定性程度。

4.报警响应模块。就分布式分类检测模块而言，其中仅给出一个警报级别权重。并连续观测多个检测结果。若短时间内出现结果较为接近、或完全一致时，即对报警级别权重予以增加。最后，则基于投票策略机制，以便最终警报响应产生于主节点。并且主节点在末端通过轮询机，使得报警响应出现于系统。

5.2 分布式检测方法

为令分布式协同攻击检测特点得到满足，一般在已搭建形成的 Hadoop 集群系统内部署随机森林分布式检测模型，当前主要自多点来检测攻击流量。本方法的具体思想如下，即分别在系统的从节点位置部署所有随机森林内的决策树。所以，由此取得分类检测结果，并在主节点上创建集中分析处理模块。集中分析处理模块的主要功能是每个从节点上每个基本分类器获得的各种分类检测结果，并基于投票策略来针对记录进行投票。由此，即可取得分类检测的最终结果。如下即为分布式分类检测的具体步骤：

其一，预处理测试以及原始训练数据集，并不均分、或均分为同等于 Hadoop 集群内从节点数量的份数；其二，基于各个节点，运用选择训练数据样本的随机性，来对训练样本数据进行选定，并构建决策树，并在此基础上，再运用选定特征属性变量的随机性来进行属性特征的选定，此后完成训练；其三，分别在从节点内输入相关子测试数据集样本，以便完成分类检测，获得子结果，然后再向主节点发送结果；其四，通过集中分析处理模块进行投票，由此获得分类检测的最终结果。

5.3 实验指标

基于分类的检测模型，它的性能根据检测到异常流量的个数与正常流量的个数进行评价。TP(True Positive)表示攻击被正确预测的个数，TN(True Negative)表示将正常流量预测正确的个数，FN(False Negative)表示将攻击错误预测为正常的个数，由以上三个指标，可得出 DR,Accuracy,Precision

$$DR = \frac{TP}{TP + FN} \quad (12)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (13)$$

$$\Pr ecision = \frac{TP}{TP + FP} \quad (14)$$

DR 反映数据集中攻击样本被正确识别的比例，Accuracy 反映被正确分类的样本的比例，Precision 反映被检测为攻击的样本中，真正攻击样本的比例。

6 实验结果与分析

在攻击检测方向，任何一个检测算法最终要应用到不同的实际场景中，但不同的应用环境，数据量大小，网路环境都不相同，因此不存在唯一的标准去衡量检测算法的优越。为了在实验环境下采用一个统一的标准，大多论文都采用美国国防部高级计划研究局 1999 年用于攻击检测的公开数据集 KDD CUP99。为了证明本文方法在 DR，Accuracy，Precision 这些性能指标比 Adaboost 方法的优越性，在实验室环境下，采用 Linux 虚拟机，利用阿里云服务器，采用虚拟化的方法，搭建一个伪分布式 Hadoop 集群，其中有一个主节点，其余为从节点，受主节点管理。本文采用经典的 KDD CUP 1999 数据集进行机器学习训练，其中选取两类数据集，一类是异常数据集和正常数据集进行验证。并且，该数据集包含测试卷和训练集，有 41 项特征，其中 38 项是数值型，3 项是标称型数据。

本文选择 Adaboost 算法与本文算法进行性能比较，原因在于 Adaboost 算法作为机器学习中一种经典的强分类器算法，具有高精度的分类器，易实现，不需要做特征筛选，并且不用担心过度拟合的问题，很适合攻击检测这种多分类的应用场景，与本文算法进行比较，有较高的参考价值。

最终，本文对最后检测的结果进行三项指标的评估，分别是检测率，正确率，精确率，并且与 Adaboost 方法进行三个指标的对比，在仿真环境下证明了本文方法的优越性。

对于本文采用的随机森林方法，它有一个重要的特性，随机森林中决策树的数量即阈值的不同，对最终检测算法的有效性及检测效率有极大的影响。假如决策树过多，算法的效率会下降，若过小，对结果的分类精度就会下降，产生过拟合的现象。根据以往研究者的成果及真实的实验基础上，本文采用 8 个不同的阈值。

表 2 表示本文模型与 Adaboost 在不同阈值下各参数 TP，FP，TN.FN 的实验结果。

表 2　本文模型与 Adaboost 在不同阈值各参数结果

阈值	本文随机森林方法				Adaboost 方法			
	TP	FP	TN	FN	TP	FP	TN	FN
25	222526	133	60460	14	222214	926	59667	485
50	222578	185	60442	6	222216	962	59648	219
75	222579	165	60428	11	222223	885	59711	288
100	222588	152	60453	9	222234	882	59722	196
125	222586	162	60482	8	222242	195	60333	228
150	222587	165	60465	7	222243	188	60405	265
175	222565	169	60489	6	222212	196	60401	256
200	222584	164	60475	7	222234	172	60412	249

经过由如下三个结果图可得，在 8 个阈值中下本文方法的 DR、Accuracy、Precision 几乎接近百分百，相比于 Adaboost 检测性能十分优越。

在 DR 方面，本文的检测模型在阈值 50 后基本保持在接近 100%，当阈值为 100 时最大，为 99.95%，而 adaboost 方法的 DR 值起伏较大，且均在 99.95% 以下，如图 3。在 Accuracy 方面，本文检测模型的阈值为 25 时，其值为 99.95%，之后有所下降，但基本都保持在 99.9 到 99.95 之间，而 adaboost 在阈值为 100 前非常低，基本不能使用，在 100 后 Accuracy 才大幅上升，与本文检测方法明显效果相差很多，如图 4 所示。在 Precision 方面，本文检测模型在所有的阈值上均优于 Adaboost 方法。其中，Adaboost 方法的阈值从 100 到 125 时有一个很大的上升，而上升幅度较大的原因正是因为 Adaboost 方法不存在两个“随机性”，所以只在阈值为 100 之后才与本文检测模型的 Precision 基本一致。如图 5 所示。

最终的结果表明：本研究所设计的攻击检测模型在 DR、Accuracy、Precision 三个方面都要比传统的 Adaboost 方法更具优越性，更具有优秀的检测性能和检测的稳定性。

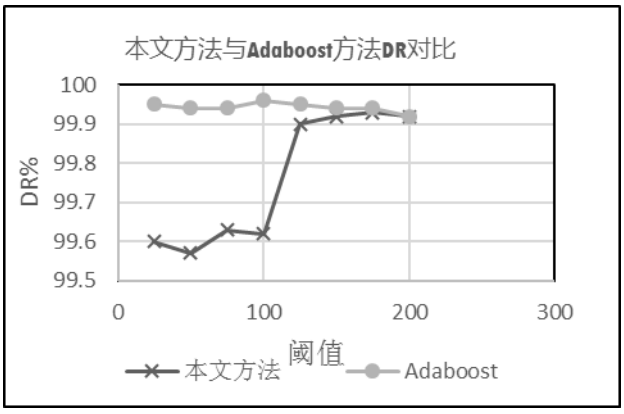

图 3　本文方法与 Adaboost 的 DR 对比

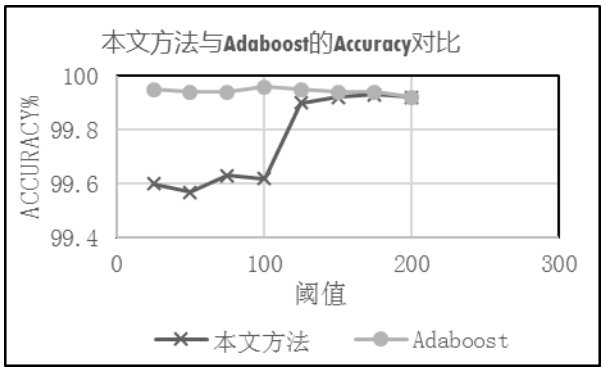

图 4　本文方法与 Adaboost 的 Accuracy 对比

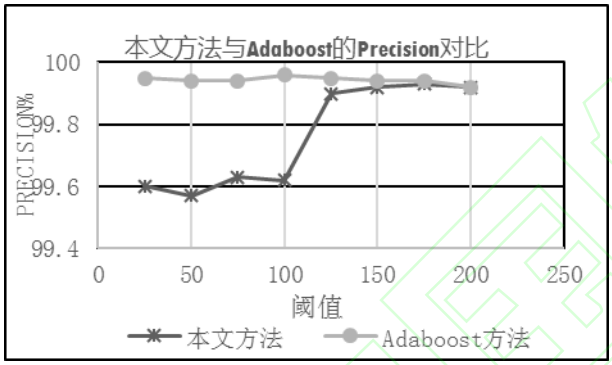

图 5　本文方法与 Adaboost 的 Precision 对比

7　结束语

本文为了解决在大数据环境下网络存在的安全问题，尤其针对攻击检测这一突出问题，基于机器学习与统计分析等方法，研究了一种基于随机森林的入侵检测方法，并与 MapReduce 分布式框架相结合，理论分析了检测模型的功能与步骤，且检测算法中在引入两大随机性后，即可降低网络流量内的属性特征字段的噪声，并使彼此关联性得以消除，提高了攻击检测的检测正确率，精确率，为网络安全提供了更好的保护。

参考文献：

[1] Buczak A L, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection[J]. IEEE Communications Surveys & Tutorials, 2015, 99: 1-26.

[2] Klassen M, Yang N. Anomaly based intrusion detection in wireless networks using Bayesian classifier[C]//IEEE FifthInternational Conferenceon Advanced Computational Intelligence. [S.l.]: IEEE , 2012 : 257 - 264.

[3] Min.Network Traffic Abnormality Detection Algorithm Based on Selfadaptive Threshold[J]. Computer Engineering, 2009, 19.

[4] GonzalezH，Han J，Ouyang Y，et al. Multidimensional DaMi ningof Traffic Anomalies on LargeScale Road Netwo rks[J]. Transportation Research Record Journal of theTransportation Research Board，2011，2215(-1): 75-84.

[5] Muniyandi A P, Rajeswari R, Rajaram R. Network Anomal Detection by Cascading KMeans Clustering and C4.5 Decii onTree a lgorithm[J]. Procedia Engineering, 2012, 30: 174-2.

[6] Agarwal B, Mittal N. Hybrid Approach for Detection of AnomalyNetwork Traffic using Data Mining Techniques[J]. Procedia Technology, 2012, 6(4): 996-1003.

[7] Hu W, Gao J, Wang Y, et al. Online Adaboostbased Paraeterized Methods for Dynamic DistributedNetwork IntrusionDetection[J]. IEEE Transactionson Cybernetics, 2014, 44(1): 66-82.

[8] Zhang Zulkemine M, Haque A. Random-forests-based Network Intrusion Detection Systems[J]. IEEE Transactions on Systems, Man, and Cybernetics Part C (Applications and Revi ews), 2008, 38(5): 649-659.

[9] Singh K, Guntuku S C9 Thakur A. Big Data Analytics Frameworkfor Peer-to-peer Botnet Detection Using Random Forests [J]. Information Sciences, 2014, 278: 488-497.

[10] 孙红艳，张红玉.基于 SVM 的 BGP 异常流量检测[J]. 现代电子技术, 2010, 33(18): 118-120.

[11] 侯重远，江汉红，芮万智，等. 工业网络流量异常检测的概率主成分分析法[J]. 西安交通大学学报，2012, 46(2): 70-75.

[12] 冶晓隆，兰巨龙，郭通. 基于主成分分析禁忌搜索和决策树分类的异常流量检测方法[J]. 计算机应用，2013, 33(10): 2846-2850.

[13] 夏玉明,胡绍勇,朱少民,刘丽丽.基于卷积神经网络的网络攻击检测方法研究[J].信息网络安全,2017(11):32-36.

[14] 和湘,刘晟,姜吉国.基于机器学习的入侵检测方法对比研究[J].信息网络安全,2018(5):1-11.

[15] Amor N B, Benferhat S, Elouedi Z. Naive Bayes vs decision trees in intrusion detection systems[C]// Proceedings of the ACM Symposium on Applied Computing. Nicosia , Cyprus, 2004: 420-424.

[16] Blowers M, Williams J. Machine learning applied to cyber operations.Network Science and Cybersecu-rity[C]//Pino Red. Network Science and Cybersecurity. New York, USA: Springer, 2014: 155-175.

[17] Kanungo T, Mount D M, Netanyahu N S, et al. An efficient K-means clustering algorithm: Analysis and implementation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 24(7): 881-892.

[18] Hu W, Gao J, Wang Y, et al. Online Adaboost-based Parameterized Methods for Dynamic Distributed Network Intrusion Detection[J]. IEEE Transactions on Cybernetics, 2014, 44(1): 66-82.

[19] 何明亮，陈泽茂，左进. 基于多窗口机制的聚类异常检测算法[J]. 信息网络安全, 2016(11): 33-39.

[20] Tom W.著.Hadoop 权威指南[M].华东师范大学数学科学与工程学 院译.第 3 版.北京：清华大学出版社，2015：321-365.

[21] Tom Machine Mitchell 机器学习[M].曾华军，张银奎等译. 北京：机械工业出版社，2003: 49-50.

[22] 李航.统计学习方法[M].北京：清华大学出版社，2012：56-57.