基于改进非广延熵特征提取的双随机森林实时入侵检测方法

姚 东[1] 罗军勇[1] 陈武平[2] 尹美娟[3]
(解放军信息工程大学 郑州 450002)[1] (信息保障技术重点实验室 北京 100072)[2]
(数学工程与先进计算国家重点实验室 郑州 450002)[3]

摘 要 在网络骨干链路的高速、大数据量环境下,相对于正常数据,攻击及异常数据相对较少,进行实时入侵检测难度大。针对此问题,提出了一种基于改进非广延熵特征提取和双随机森林的实时入侵检测方法。利用非广延熵,提取出流量属性取值分布的多维特征,通过对非广延熵的改进来降低特征间的相关性。使用完整的特征样本集建立第一个随机森林检测模型,使用包含攻击数据的特征样本子集建立第二个随机森林检测模型,通过双随机森林检测算法实现对少量异常的有效检测。实验结果表明,该方法能够在有限流量信息的基础上获得较高的检测精确率和召回率,其时间和空间复杂度适当,适合于对骨干链路的实时入侵检测。

关键词 网络流量,入侵检测,非广延熵,随机森林
中图法分类号 TP393.08 **文献标识码** A

Online Double Random Forests Intrusion Detection Based on Non-extensive Entropy Features Extraction

YAO Dong[1] LUO Jun-yong[1] CHEN Wu-ping[2] YIN Mei-juan[3]
(PLA Information Engineering University, Zhengzhou 450002, China)[1]
(Science and Technology on Information Assurance Laboratory, Beijing 100072, China)[2]
(State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450002, China)[3]

Abstract This paper proposed an intrusion detection method that can be used in high speed network backbone. Based on non-extensive entropy with different parameters, the original distribution of the values of attributes was decomposed to high dimensional features. Using these detailed features, the detection model based on random forest was constructed. For the purpose of increasing detection accuracy and recall further, the second random forest detection model was constructed with the attack instances only. The experimental results suggest that proposed intrusion detection method can achieve competitive detection precision with a high recall.

Keywords Network traffic, Intrusion detection, Non-extensive entropy, Random forest

1 引言

随着网络规模的扩大、数据传输带宽的不断提高,面向高速骨干链路的实时入侵检测技术受到了越来越多的关注。高速骨干链路上的数据量大,实时处理难度大;攻击数据所占比例通常较小,检测难度大。本文旨在设计一种实时入侵检测方法,以有效地对骨干链路网络流量实施入侵检测。

面对骨干链路高速、大数据量的特点,实时检测应尽量减少对数据的处理和访问。目前大多数针对网络流量的检测算法都是基于流进行的,而流是在对网络数据包进行统计、合成基础上得到的一种网络数据形式,在骨干链路上直接对包实施检测可以减少由包到流、再由流提取特征的中间处理环节,因此本文选择基于包进行骨干链路的入侵检测。另外,实时入侵检测经常需要采用一些方法来减少所需处理的数据量。采样可以减少数据量,但采样是一个有损的信息处理过程,可能会丢失重要的信息,文献[1]通过实验证明采样后的数据用于异常检测将会导致异常检测算法误差增大,在异常数据较少时影响更大;降维方法,如常用的主成份分析法(PCA)[2],相对复杂度较高,可用于事后异常分析,不能满足骨干网的入侵检测要求。概要数据结构来源于数据流研究领域,适用于对大流量一次经过的数据进行快速查询,在应用于网络流量的检测方面出现了许多相关研究[3-5],结果显示概要数据结构具有最优的时间和空间约束。因此本文在数据预处理阶段采用它对流量数据的基本属性进行实时概要记录。

对于大量流量数据中少量异常流量的检测,首先要解决异常数据包特征的提取问题。根据骨干链路流量数据的特点和入侵检测的实时性要求,不适于通过对数据包的深度检测和多次访问得到详细的特征(如KDD cup99[6]的41维特征)来发现异常。有一些研究通过香农熵对流量属性取值的信息进行度量来发现异常,如文献[7-9]通过香农熵实现了对蠕虫和其它一些异常的检测。但是香农熵对分布变化的检测存在一些限制,即低维的流量熵值对少量异常具有线性不可分性,如文献[10]用香农熵对骨干网异常流量的检测,需要异常流量在总流量中的比例不低于4%。另外,文献[11,12]的研究

到稿日期:2013-02-07 返修日期:2013-06-03 本文受信息保障技术重点实验室开放基金(KJ-12-04)资助。
姚 东(1978—),男,硕士生,主要研究方向为计算机网络安全,E-mail:dojn@tom.com;罗军勇(1964—),男,教授,硕士生导师,主要研究方向为网络信息安全;尹美娟(1977—),女,博士,讲师,主要研究方向为网络信息安全。

发现，香农熵适合度量符合高斯分布的信息，而对于骨干链路的流量，IP 和端口的观测值存在较强的重尾分布特征，而非广延熵适合对非高斯分布信息的度量，即通过不同的非广延参数将分布中不同区域的特征放大，将一维熵值变为关于该分布的多维熵值，增强了对少量异常的检测能力。

包含多种攻击的异常检测问题是一个具有高维特征空间的复杂多分类问题，检测所需的最佳特征信息组合对于不同种类的攻击各不相同，将这些特征组合在一起建立检测模型时，对于某类攻击，冗余的特征信息往往会增加学习算法搜索该类解空间的复杂度，降低学习效率；另外，过多的特征也易产生局部优化和过拟合等问题；此外，某类攻击或异常在学习样本中比例较小，或者属于它的有效特征较少、较弱时，也难以被检测出来。因此，需要一种学习算法来有效处理高维空间的多分类问题，并且该算法具备对多分类中“小”类的检测能力。随机森林算法是一种组合的多分类算法，在天文学、微阵列分析、DNA 检测和新药发现等领域有着广泛的应用，在处理高维数据和发现少量异常方面体现出较强的能力。

根据以上分析，本文提出一种在网络数据包基本属性概要记录上基于改进非广延熵特征提取的双随机森林实时入侵检测方法。针对使用非广延熵建立特征集的过程中存在的问题，采用非均匀参数和 Top-k 的方法进行了改进。通过分析随机森林检测结果中的误、漏报实例投票率，从加强少量异常检测能力的目的出发，构建了双随机森林检测模型。在 DARPA1999 入侵检测数据集上的实验结果表明，该入侵检测方法在使用数据包少量属性的基础上，可以同时获得较高的精确率和召回率，适用于骨干链路的实时入侵检测。

2 相关理论和技术

2.1 非广延熵

香农熵的定义为：

$$H\equiv-\sum_{i=1}^{n}p_i\log_2(p_i) \tag{1}$$

式中，$p_i=m_i/m$，m 为数据项出现的总数，m_i 为数据项 i 出现的次数，$m=\sum_{i=1}^{n}m_i$，香农熵反映了被考察对象取值的多样性。

非广延熵（Nonextensive Entropy）的概念来自于非广延统计力学，最常用到的是 Tsallis Entropy，它的定义为：

$$S_q(X)\equiv\frac{1}{q-1}(1-\sum_{i=1}^{n}(p_i)^q) \tag{2}$$

式中，q 是非广延参数，当 $q>1$ 时，熵值中概率较大的元素贡献较大，相当于把高概率区间的特征进行了放大；当 $q<-1$ 时，熵值中概率较小的元素贡献较大，相当于把低概率区间的特征进行了放大；特别地，当 $q\to1$ 时，非广延熵收敛于香农熵。

利用非广延熵的这个特点，可以通过不同的 q 参数，提取出流量属性观测值分布在不同概率区间的特征，这样无论攻击或异常所占比例多少，都会得到相应的特征。

2.2 随机森林分类算法

随机森林分类算法是数据挖掘领域的前沿新技术之一，由分类和回归树（CART）算法的提出者 Leo Breiman[13] 于 2001 年提出。随机森林算法具有众多的特点，在对分布非平衡数据进行分类时效果较好。随机森林是一个由多个树结构分类器组成的组合分类器，可以表示为 $\{h(x,\theta_k)\}$，其中 $\{\theta_k\}$ 是相互独立且服从相同分布的随机向量。对输入 x，每个树分类器 h 对其类属投出一票，最终由票数最多的结果决定 x 的类属。

整个随机森林算法包括两个阶段：树的生长阶段（1）—（3）和投票阶段（4）、（5）。

1）从训练集 D 中用 bootstrapped[14] 抽样，有放回地随机选取 $|D|$ 个数据样本作为建立决策树的样本。

2）用在标准决策树基础上改进的方法建立森林中的一棵决策树：

a）在树的每一个节点进行分裂时，随机从组成数据样本的 n 个属性中选择 k 个属性，一般遵循 $k=\log_2 n+1$ 的原则，且 k 在整个森林的建立过程中保持不变。

b）不对树进行剪枝。

3）根据设定的树的数量，重复步骤 1）和 2），建立起整个森林。

4）当对一个实例进行检测时，将之输入随机森林检测模型的每一棵树，让每一棵树都对它进行投票。

5）计数每一种投票结果（或计算出现概率），找出其中支持率最高的结果作为检测结果。

3 基于改进非广延熵特征提取和双随机森林的实时入侵检测方法

该方法共包括 3 个阶段：数据采集阶段、特征建立阶段和入侵检测阶段。

（1）数据采集阶段。数据采集阶段的工作是为了从高速、大数据量的骨干链路上快速对网络流数据中一个时间窗口内数据包的部分属性取值进行统计，以得到这个时间段网络流数据的基本统计特征。对由骨干链路顺序进入大规模网络的数据包按固定时间窗口进行切分，从每一个数据包 x_i 中提取出 5 个属性字段（sIP_i，dIP_i，$sPort_i$，$dPort_i$，$bytes_i$），运用概要数据结构进行记录，并根据新到来的数据包对记录实时更新，如图 1 所示。

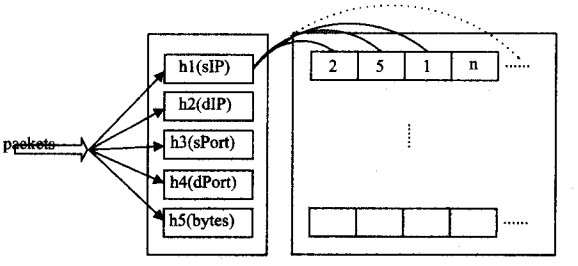

图 1 概要数据结构

其中 $h_n(n\in\{1,\cdots,5\})$ 表示不同的哈希函数。概要数据结构用一个 n 行 m 列的二维数组来存储每个属性字段不同取值的统计量，对应 $n\times m$ 个计数器。数组的每一行对应一个哈希函数，一行中的 m 个计数器表示在一个时间窗口该属性取值的 m 种情况，每个计数器记录一种取值情况在该时间窗口出现的次数。

（2）特征建立阶段。异常数据包在骨干链路上相对数量较少，对一个时间窗口网络流数据基本统计特征的影响很难被察觉，因此特征建立阶段的工作就是把基本统计特征放大、分解，使少量异常在高维空间中体现出较为明显的特征。这个工作由非广延熵完成，并通过对非广延熵计算的改进减小了多维特征之间的相关性，联合各属性的分布特征和时间窗

口内流量的简单统计特征，得到代表一个时间窗口整体流量特征的特征向量。特征提取的具体方法见 3.1 节。

(3)入侵检测阶段。如图 2 所示，首先利用随机森林从训练集全体中学习到的检测模型 I，对流量特征向量进行检测，满足投票阈值的结果作为最终检测结果，检测结束；不满足投票阈值的结果，利用随机森林从训练子集(由训练集中包含各类攻击的数据组成的子集)中学习得到的检测模型 II，对流量特征向量进行检测，根据检测阈值得到最终检测结果。具体方法见 3.2 节。

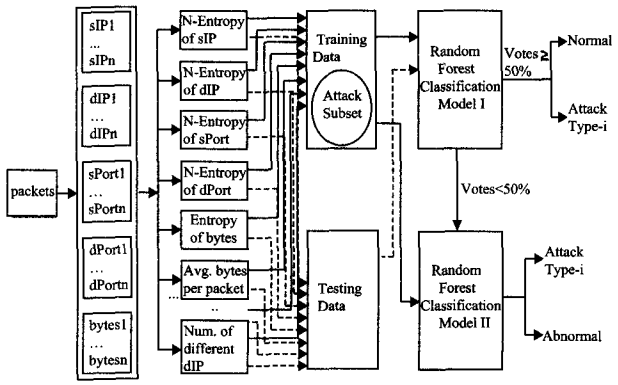

图 2　训练和检测过程

3.1　基于改进非广延熵的特征提取

3.1.1　基于非广延熵的特征提取存在的问题

在采用多分类技术的入侵检测模型中，流量特征是决定检测精确率的重要因素，使用非广延熵可以在更细的粒度上体现流量属性取值分布的特征。

文献[15]使用 Netflow[16]数据，根据式(2)，在[−2,2]区间以 0.25 为间隔，均匀选取 17 个参数 q，计算 IP、端口等属性在一个时间窗口的非广延熵值，用不同属性的非广延熵值组合代表该窗口内流量的总体特征。在实验中发现，使用该方法直接对流量数据包建立特征进行入侵检测，精确率较低，误报和漏报较多。通过分析发现，特征集的建立过程中存在两个问题：1)对参数 q 的均匀选取没有很好地利用重尾分布的特点，同时使用较多的参数对概率区间过细的划分会导致特征间相关性的加强，概率取值的动态范围体现出的特征受到较大限制；2)计算公式中的累加过程是全局的，因此对局部特征的放大会引入噪声。

另外，组合分类器比其中的单个分类器更精确需满足一个条件：单个分类器是“精确”的，任意两个分类器之间是“相异”的。随机森林算法中单个分类器的算法是在原有单个分类器算法上的改进，因此其在“精确”程度上是优于原单个分类算法的；要保证“相异”就是要保证单个分类器之间的错误是独立的，不会对相同的样本犯同样的错误。由于特征决定了分类的结果，同时也决定了分类器的错误，因此考虑通过改进特征，即消除特征的相关性来达到分类错误之间的独立。但是在建立特征时是对原有特征的扩展(放大和截取)，不可能完全避免特征之间的相关性，尽管如此，在轻微相关的情况下，组合还是可以提高分类的准确率。原因如下，根据组合分类器错误率公式：

$$error=\sum_{i=1}^{n}C_n^i p^i(1-p)^{n-i} \tag{3}$$

式中，n 为基分类器的数量，p 为每个基分类器的错误率，只要每个分类器的错误率不大于 50%，最终组合分类器的错误率一定不会大于单个分类器。

3.1.2　特征提取过程的改进方法

根据以上分析，对非广延熵的特征提取过程进行了改进。减少参数 q 的个数，在取值空间[−2,2]中，将均匀取值改为与重尾分布相对应的非均匀取值，具体取值为 −2、−1.1、0.01、0.5、1.1、1.3、1.5、1.7、2 这 9 个值，在概率动态变化较大的重尾分布头部，用较多的 q 参数对较多的区间进行特征放大考察，提高对变化的捕捉能力，在相对取值较为单一的尾部用较少的 q 参数，避免特征的重复和相关。

设在窗口 T 内，属性 M 共有 n 种取值，通过 $p_i=a_i/a$ 计算出每一种取值出现的概率，其中 a_i 表示 M 的第 i 种取值出现的次数，$a=\sum_{i=1}^{n}a_i$ 表示属性 M 在窗口内出现的总次数，经排序后得到关于 M 的有序概率集合 $P_M=\{p_1,p_2,\cdots,p_n\}$，$\sum_{i=1}^{n}p_i=1$，其中排序是为了方便后续 Top-k 的计算。

对非广延熵的计算过程，用 Top-k 的思想进行改进，将公式中累加的过程集中在对熵值贡献较大的取值上，降低局部特征中的噪声。具体算法是：对于一个给定的参数 q，从有序概率集合 P_M 中找出对熵值贡献最大的 k 个 p_i，k 是满足 $\sum_{i=1}^{k}(p_i)^q\geqslant\alpha\cdot\sum_{i=1}^{n}(p_i)^q$ 的最小索引值，其中 α 是百分比参数，表示 Top-k 中前 k 项值所占的百分比，在本文中取经验值 80%。由 k 得到 $a'=\sum_{i=1}^{k}a_i$，以 a' 为标准重新计算 $p_i'=a_i/a'$，$i=1,\cdots,k$。用 $p_i'(i=1,\cdots,k)$ 计算得到改进的非广延熵值 S_q'。

用改进的非广延熵对概要记录中的 4 个属性(源、目的 IP，源、目的端口)在窗口 T 内的分布进行特征提取。包字节数的分布特征用香农熵度量，因为与其它 4 个具有重尾分布的属性相比，包字节数的取值在大多数情况下都较为单一，由异常引起的变化也较为明显。再结合计算过程中得到的 5 个属性在窗口内不同取值的个数和平均包数等简单的统计特征，共同构成了该窗口流量数据的一个特征向量 $L_{T_1}=(l_{T_{11}},l_{T_{12}},\cdots,l_{T_{1n}})$，从训练集中得到的所有这样的特征向量将作为输入用于训练分类检测模型。

3.2　基于双随机森林的入侵检测方法

第一个随机森林的训练样本是从所有窗口得到的特征向量集 $L_T=\{L_{T_1},L_{T_2},\cdots,L_{T_n}\}$，按随机森林算法的树生成过程，得到随机森林的分类模型$\{h(x,\theta_k)\}$，参数 θ 代表了随机森林两个方面的随机要素：通过 Bagging[17]从 L_T 中等权重随机获得训练样本；从 L_{T_i} 中随机选择特征进行树的节点分裂。θ_k 表示 θ 的第 k 次实现，与随机森林中第 k 棵树相对应。

研究单随机森林在检测时的投票结果，发现当分类发生错误时，支持该结果的最大投票率通常都较低。通过对稀少类的加权来提高分类的精确率，权重较难选择，也较难在提高精确率的同时获得较高的召回率。文献[18,19]采用了多个检测模型组合的方式来提高检测的精确率，同时保持较低的误报率。根据模型组合思想对检测算法进行改进，建立第二个随机森林检测模型，而且第二个检测模型的构建是基于一个相对平衡的训练集，通过两个模型的协同工作来提高检测的精确率和召回率。

第二个随机森林检测模型的建立方法是从训练集 L_T 中选择出包含攻击的特征向量 L_{T_i}'，组成新的训练集 $L_T'=\{L_{T_1}',L_{T_2}',\cdots,L_{T_n}'\}$。在 L_T' 的基础上得到第二个随机森林

分类模型$\{h'(x,\theta_k')\}$。

在检测阶段，对于一个输入的待测特征样本x，进行投票决策：$\arg\max_j \frac{1}{t}\sum_{k=1}^{t} I(h(x,\theta_k)=j)$，基中$I(\cdot)$是示性函数，$t$是随机森林中树的总数。当$j\geqslant t/2$时，信任该投票结果，并将其作为最终的检测结果；当$j<t/2$时，将$x$送入第二个检测模型。根据第二个检测模型的投票：$\arg\max_{j'} \frac{1}{t'}\sum_{k=1}^{t'} I(h'(x,\theta_k')=j')$，设定阈值为$t'/3$，当$j'\geqslant t'/3$时，将投票结果作为最终的检测结果；当$j'<t'/3$时，将$x$记为异常用于后续分析，以确定其为误报或新的攻击类型。

4 实验与分析

4.1 实验数据构建

为了验证本文提出的实时入侵检测方法，需要选择合适的入侵检测数据集。在该领域，公开提供的数据集并不多，常用的有DARPA1999[20]和KDDCUP99入侵检测数据集。本文选择DARPA1999入侵检测数据集作为实验数据集，1)因为该数据集的数据是tcpdump格式的原始流量数据，2)其中很多的攻击事件所占比例较小。从对流量属性的采集记录开始，直到特征提取和入侵检测，该数据集都能完整支持本文算法的验证。KDD cup99是在DARPA1999基础之上，由哥伦比亚大学IDS实验室整理形成的入侵检测数据集，该数据集已经不再包含原始的流量数据包，而是文本格式的特征集，而且针对DARPA1999中攻击数据比例较少而难于检测的问题，人为增加了攻击数据的数量，因此它不适用于本文算法的验证。

DARPA1999入侵检测数据集共包含5周的数据，前3周用于学习训练，第4和第5周用于测试。本文选择训练集中含有攻击实例的第2周数据对检测模型进行学习训练，在第4周的数据上进行测试。

DARPA的实验环境模拟了内网和外网，在采集实验数据的时候分别从边界路由器的内、外两处进行了采集，因此每一天的数据都包含inside和outside两个部分。每天的采集开始时间是早上8点，结束时间为第二天早上6点，包含22小时的流量数据。用长度为60(单位：秒)的窗口对流量数据进行切分，使用概要数据结构对数据包的源IP、目的IP、源端口、目的端口和字节数在该时间窗口的出现次数进行记录和更新。从数据包中提取信息的工作由开源的SiLK[21]工具完成。用改进的非广延熵对流量特征进行提取，与其它所选特征一起组成特征集。

特征集需加上最终的类属标记来形成训练集，根据DARPA提供的truthlist(攻击的信息列表)对特征集进行标记，含有攻击数据的时间窗口按DARPA中的4种攻击类别(U2R、R2L、PROBE、DOS)分别进行标记，不含有攻击数据的窗口则标记为NORMAL。由于DARPA采集环境的时钟使用EST(美国东部时间)，而truthlist中的时间使用UTC(世界标准时间)，因此在标记时需要消除5个小时的时间差异(EST落后UTC 5小时)，构建准确的训练集。

4.2 实验设计

为了全面验证本文提出的方法，通过3个实验的对比和分析依次说明：相对于香农熵，非广延熵对流量分布特征的提取更加全面、细致；改进的非广延熵在减小特征相关性方面的作用；随机森林在对分布非平衡数据进行分类检测时的优势，以及双随机森林算法对入侵检测的有效性。

实验使用Waikato大学的开源数据挖掘工具Weka[22]进行了对比和分析，选择了另外3种常用的多分类算法：朴素贝叶斯(Naïve Bayes)、贝叶斯网络(Bayes Network)和C4.5共同进行实验。朴素贝叶斯和贝叶斯网络同是基于贝叶斯理论的技术，使用时要求特征间条件独立或基本独立，当独立条件不满足时，朴素贝叶斯的分类会受到限制，相对来说，贝叶斯网络只要求某个特征条件独立于其非直接前驱节点，因此通过这两类算法可以验证改进后的非广延熵在减小特征相关性方面的作用。C4.5与随机森林都是基于树的分类算法，通过与C4.5的对比可以体现出基于树的组合分类器——随机森林在分类能力上的提高。

实验结果通过两个指标：精确率(记作Pre.)和召回率(记作Rec.)来反映。

4.3 实验结果分析

在香农熵特征集上的检测结果如表1所列，由于属性的分布特征只通过一维的熵值来反映，几乎所有算法的综合检测效果都不理想。相对来说在对DOS类攻击检测时的精确率和召回率明显高于其它类攻击，这是因为DOS类攻击的时间较为集中，特征较为明显。对Probe类攻击本身来说，它的特征也是比较明显的，但由于该测试集上的数据较少，攻击间隔时间较长，导致在一个时间窗口中与正常数据包相比特征不够突出，因此检测效果也不佳。随机森林检测出了所有攻击类型，体现出了它对非平衡数据的检测能力，但召回率偏低。说明在攻击数据包较少的情况下，用一个熵值难以反映出攻击的特征。

表1 香农熵特征集上的精确率和召回率

	Bayes Network		C4.5		Naïve Bayes		Random Forest	
	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.
normal	0.92	0.995	0.924	0.998	0.919	0.96	0.927	0.995
u2r	0	0	0	0	0.02	0.017	1	0.052
r2l	0	0	0	0	0.091	0.014	0.313	0.035
probe	0.115	0.068	0.1	0.023	0	0	0.75	0.068
dos	0.727	0.323	0.853	0.468	0.295	0.306	0.913	0.508

在非广延熵特征集上的检测结果如表2所列，基本上所有算法都检测出了4类攻击。基于贝叶斯技术的算法，对部分攻击的召回率提高较多，说明非广延熵的确提取出了少量异常的特征；但是精确率较低，而且对正常实例的召回率下降幅度较大，这说明非广延熵建立的特征之间存在较强的相关性，对于特征独立性要求较高的朴素贝叶斯算法受到的影响最大。C4.5算法是在全体特征集上通过信息增益来选择进行分裂的特征，由于出现了更多与异常相关的特征，因此它的综合检测效果在该特征集上好于其它算法。相对来说，同是基于树的随机森林在选择分裂特征时，随机选择少量的特征，因此在大量特征都有较强相关性的基础上，很难随机选择出最佳的分裂特征。

表2 非广延熵特征集上的精确率和召回率

	Bayes Network		C4.5		Naïve Bayes		Random Forest	
	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.
normal	0.994	0.509	0.973	0.994	0.991	0.018	0.925	0.994
u2r	0.022	0.308	0.415	0.332	0.009	0.923	0.448	0.134
r2l	0.044	0.75	0.642	0.348	0.046	0.116	0.895	0.152
probe	0.019	0.714	0.333	0.036	0.07	0.614	1	0.432
dos	0.056	0.375	0.78	0.542	0.13	0.042	0.969	0.431

在改进的非广延熵特征集上的检测结果如表3所列，所有算法对攻击检测的精确率和召回率均有大幅度的提升，尤其是本文提出的双随机森林检测算法。但是从基于贝叶斯技术的算法来看，特征之间的相关性还是对结果产生了一定的影响。这是由于减少 q 参数取值数量和计算过程中 Top-k 的处理都只是在一定程度上减少了特征间的相关性，并没有使特征间完全独立。在对相关性要求不高的检测算法中，这样的改进是可以接受的，并且能够改善检测效果。实验结果显示，非广延熵和双随机森林相结合进行入侵检测可以同时获得较高的精确率和召回率，证明了本文所提出的入侵检测方法的有效性。

表3 改进的非广延熵特征集上的精确率和召回率

	Bayes Network		C4.5		Naïve Bayes		Double RF	
	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.	Pre.	Rec.
normal	0.994	0.516	0.971	0.991	0.996	0.217	0.967	0.999
u2r	0.6	0.625	0.808	0.675	0.373	0.792	0.886	0.892
r2l	0.747	0.747	0.798	0.762	0.417	0.172	0.9	0.931
probe	0.918	0.738	0.895	0.836	0.803	0.803	0.934	0.934
dos	0.671	0.773	0.896	0.864	0.317	0.303	0.955	0.955

4.4 检测算法的时间复杂度

随机森林在构建时对于输入的变量集合进行随机分组，每组变量的个数是一个事先确定好的常量，这个量在整个森林的构建过程中是不变的。对于每组变量，利用 CART (Classification and Regression Trees)方法生成一棵树，生长过程中不进行剪枝操作。在生成树中的每个节点时，对输入该节点的变量，重复前面的随机分组，再重复 CART 方法，直到叶子结点为止。对于每组变量的个数，通常的选取原则是小于 $\log_2 M+1$ 的最大正整数，其中 M 是输入变量的个数。设在构建树时使用的样本数为 N，M 为输入变量的总个数，G 为每一组变量的个数，则在该样本集上构建随机森林的时间复杂度为 $G*\log_2(N)/M$。

在检测速度方面，对流量数据的实时概要记录和在此之上的特征生成都是在内存中完成，基于树的随机森林继承了树结构检测模型的优点，其时间复杂度为搜索单个 CART 树的常数倍。对于本文实验中所使用的数据集和 60 秒时间窗口而言，在实验环境(Pentium4 3GHz dual core 处理器，1G 内存)下，得到测试集第 4 周第 1 天 1289 个时间的窗口的处理时间的平均值为 12.515 秒。去除部分数据包较少的时间窗口对平均值的影响，处理一个时间窗口的时间低于 15 秒，大约为时间窗口长度的 1/4。对于数据量更大的环境可以对硬件进行进一步升级，因此其检测速度可适应实时应用的需求。

结束语 高速骨干链路的入侵检测是目前网络安全领域研究的热点，现有的检测技术多需要对流量数据进行存储、整合以及多次访问才能提取出有效的特征，不适用于对高速流数据的实时检测。本文提出了一种基于改进的非广延熵和双随机森林的实时检测方法：对流量数据，用概要数据结构实时记录相关信息，用改进的非广延熵进行特征提取，依据得到的特征，用双随机森林检测算法对攻击进行有效的检测。实验结果显示了使用该方法进行实时入侵检测的有效性和可行性。

为了使提出的方法更加完善，还需从以下 3 个方面开展进一步的研究：1)参数 q 的选取问题，研究如何根据每个窗口内不同属性取值的具体分布情况，自适应地对参数 q 的个数和取值做出决策；2)参数 α 的选取问题，研究如何根据具体分布特征进行自动选择，进一步降低特征之间的相关性；3)窗口策略，研究窗口大小对检测结果的影响，以及如何根据链路的实际情况自适应调整窗口的策略。

参 考 文 献

[1] Mai J, Chuah C N, Sridharan A, et al. Is sampled data sufficient for anomaly detection? [C]// Proceedings of the 6th ACM SIGCOMM conference on Internet measurement. ACM, 2006: 165-176

[2] Zargar G R, Baghaie T. Category-Based Intrusion Detection Using PCA[J]. Journal of Information Security, 2012, 3(4): 259-271

[3] Liu Yang, Zhang Lin-feng, Guan Yong. Sketch-based Streaming PCA Algorithm for Network-wide Traffic Anomaly Detection [C]// Distributed Computing Systems (ICDCS), 2010 IEEE 30th International Conference on. IEEE, 2010: 807-816

[4] Tang Jin, Cheng Yu, Zhou Chi. Sketch-based SIP flooding detection using Hellinger distance[C]// Global Telecommunications Conference, 2009, GLOBECOM 2009. IEEE, 2009: 1-6

[5] Li Ai-ping, et al. Detecting Hidden Anomalies Using Sketch for High-speed Network Data Stream Monitoring[J]. Appl. Math, 2012, 6(3): 759-765

[6] Hettich S, Bay S D. KDD cup 1999 data [EB/OL]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, 1999

[7] Wagner A, Plattner B. Entropy based worm and anomaly detection in fast IP networks[C]// Enabling Technologies: Infrastructure for Collaborative Enterprise, 2005, 14th IEEE International Workshops on. IEEE, 2005: 172-177

[8] Lakhina A, Crovella M, Diot C. Diagnosing network-wide traffic anomalies[J]. ACM SIGCOMM Computer Communication Review, ACM, 2004, 34(4): 219-230

[9] Li Xin, et al. Detection and identification of network anomalies using sketch subspaces[C]// Proceedings of the 6th ACM SIGCOMM conference on Internet measurement. ACM, 2006: 147-152

[10] 朱应武，杨家海，张金祥. 基于流量信息结构的异常检测[J]. 软件学报，2010，21(10)：2573-2583

[11] Ziviani A, Gomes A T A, Monsores M L, et al. Network anomaly detection using nonextensive entropy[J]. Communications Letters, IEEE, 2007, 11(12): 1034-1036

[12] Scherrer A, Larrieu N, Owezarski P, et al. Non-gaussian and long memory statistical characterizations for internet traffic with anomalies[J]. IEEE Transactions on Dependable and Secure Computing, 2007, 4(1): 56-70

[13] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32

[14] Mooney C Z, Duval R D. Bootstrapping: A nonparametric approach to statistical inference[M]. Sage Publications, Incorporated, 1993

[15] Tellenbach B, Burkhart M, Sornette D, et al. Beyond shannon: Characterizing internet traffic with generalized entropy metrics [J]. Passive and Active Network Measurement, 2009: 239-248

[16] Cisco Systems Inc. Netflow services solutions guide[OL]. http://www.cisco.com

(下转第 218 页)

假如有一个系统 UML 用例图中所包含的用例风险程度和价值的情况如表 3 所列。

表 3　某个 UML 用例图中各个用例的风险程度和价值

用例名	风险程度	价值
uc1	7.6	0.78
uc2	4.2	0.92
uc3	3.5	0.45
uc4	4.5	0.42
uc5	5.2	0.62
uc6	6.2	0.82

表 3 包含 6 个用例，按照价值可以将整个系统的用例排序为：*uc*2、*uc*6、*uc*1、*uc*5、*uc*3、*uc*4；按照用例的风险程度可以将整个系统的用例排序为：*uc*1、*uc*6、*uc*5、*uc*4、*uc*2、*uc*3。根据式(6)有：

$B(uc1)=(6-3)+(6-1)=8$

$B(uc2)=(6-1)+(6-5)=6$

$B(uc3)=(6-5)+(6-6)=1$

$B(uc4)=(6-6)+(6-4)=2$

$B(uc5)=(6-4)+(6-4)=4$

$B(uc6)=(6-2)+(6-2)=8$

所以得出的敏捷迭代顺序为：*uc*1、*uc*6、*uc*2、*uc*5、*uc*4、*uc*3。

如果开发人员和用户更看重用例的价值，可以利用式(7)，对用例的价值对应的权重取更大一些，比如取 0.8，则风险程度的权重就取 0.2，于是：

$B(uc1)=0.8*(6-3)+0.2*(6-1)=2.8$

$B(uc2)=0.8*(6-1)+0.2*(6-5)=4.2$

$B(uc3)=0.8*(6-5)+0.2*(6-6)=0.8$

$B(uc4)=0.8*(6-6)+0.2*(6-4)=0.4$

$B(uc5)=0.8*(6-4)+0.2*(6-4)=2$

$B(uc6)=0.8*(6-2)+0.2*(6-2)=4$

所以得出的敏捷迭代顺序为：*uc*2、*uc*6、*uc*1、*uc*5、*uc*3、*uc*4。

如果开发过程中更看重用例的风险程度对产品和用户的影响，则可以设置风险程度的权值更大一些，其计算过程与上面类似。

结束语　本文通过研究得出如下的结论：

• 将敏捷开发迭代顺序从功能组为基础转向以 UML 用例图中的用例为基础。

• 改变了以往敏捷开发迭代顺序以单一指标为依据的缺陷，本文利用用例的风险程度和使用概率来综合考虑敏捷开发的迭代顺序。

• 以往敏捷开发迭代顺序的确定大多是以定性方法为主，主观随意性比较强。本文利用概率统计方法和模糊意见集中决策方法来定量确定敏捷开发的迭代顺序。

总之，本文提出的方法能够为软件开发人员在确定迭代顺序时提供一种定量的决策依据，而且可以根据用户和开发人员的关注点来调整项目的开发顺序，如果项目更重视可靠性，则风险因素成为重要考量的因素，如果项目更重视用户的使用情况，则价值成为主要关注的因素，即该方法能很好地反映出项目开发的迭代顺序随项目涉众人员关注点的不同而不同，使用起来比较灵活。

参 考 文 献

[1] 王晓华. 敏捷开发环境下软件可靠性分析及相关问题研究[D]. 贵阳：贵州大学，2008：49-54

[2] (美)柯恩. 敏捷估计与规划[M]. 宋锐，译. 北京：清华大学出版社，2007：95-200

[3] Cortellessa V. Harshinder Singh and Bojan Cukic. Early reliability assessment of UML based software models[C]// WOSP '02. Rome, Italy, July 2002：24-26

[4] Singh H, Cortellessa V, Cukic B, et al. A Bayesian approach to reliability prediction and assessment of component based systems[C]// Proc. Of 12th International Symposium on Software Reliability Engineering(ISSRE'01). 2001

[5] 胡文生，赵明，杨剑锋. 一种基于 UML 用例模型的软件可靠性分配方法[J]. 计算机科学，2012，39(6A)

[6] 胡文生，赵明，杨剑锋，等. 敏捷开发过程中的迭代策略分析[J]. 微电子学与计算机，2012，29(5)

[7] 侯福均，吴祈宗. 模糊偏好关系与决策[M]. 北京：北京理工大学出版社，2009(2)：72-77

[8] Johnstone C P, Lill A, Reina R D. Does habitat fragmentation cause stress in the agile antechinus? A haematological approach [J]. Journal of Comparative Physiology B: Biochemical, Systemic, and Environmental Physiology, 2011, 182(1): 139-155

[9] Mikulenas G, Kapocius K. An Approach for Prioritizing Agile Practices for Adaptation [M]. Information Systems Development, 2011, Part 7: 485-498

[10] 江瑜. 基于 UML 的敏捷建模方法研究[J]. 计算机工程与设计，2008，29(15)

[11] 段隆振，王凤斌，甘晟科，等. 基于敏捷化统一过程需求建模的研究及实践[J]. 计算机科学，2006，33(10)

[12] Limpens F, Team R, EPFL, et al. Towards agile competence development[C]// 2011 IEEE International Conference on Date of Conference. March 2012：667-672

[13] Swaminathan B, Jain K. Implementing the Lean concepts of Continuous Improvement and Flow on an Agile Software Development Project: An Industrial case Study[C]// AGILE India (AGILE INDIA). Feb. 2012：10-19

[14] Bustard, David. Beyond Mainstream Adoption: From Agile Software Development to Agile Organizational Change[C]// Engineering of Computer Based Systems (ECBS), 2012 IEEE 19th International Conference. April 2012：90-97

(上接第 196 页)

[17] Quinlan J R. Bagging, boosting and C4. 5[C]// Proceedings of the National Conference on Artificial Intelligence. 1996：725-730

[18] Siraj M M, Maarof M A, Hashim S Z M. A Hybrid Intelligent Approach for Automated Alert Clustering and Filtering in Intrusion Alert Analysis[J]. Journal of Computer Theory and Engineering, 2009, 1(5)：539-45

[19] Panda M, Abraham A, Patra M R. A Hybrid Intelligent Approach for Network Intrusion Detection[J]. Procedia Engineering, 2012, 30：1-9

[20] http://www.ll.mit.edu/mission/communications/ist/corpara/ideval/data

[21] http://tools.netsa.cert.org/SiLK

[22] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques (second ed.) [M]. Morgan Kaufmann Publishers, 2005