基于改进朴素贝叶斯算法的入侵检测系统

王 辉[1] 陈泓予[1] 刘淑芬[2]
(河南理工大学计算机科学与技术学院 焦作 454000)[1] (吉林大学计算机科学与技术学院 长春 130012)[2]

摘 要 随着互联网连通性的不断增强以及网络流量的日益增大,最近频繁发生的入侵事件再度凸显了入侵检测系统的重要性。针对朴素贝叶斯算法的缺陷,提出了一种改进后的朴素贝叶斯算法。该算法在原有的朴素贝叶斯模型基础上巧妙地引入属性加值算法,通过对分类参数的调控来实现简化分类数据复杂度的作用,并以计算出的最佳参数值来优化分类精确度。最后结合实验结果证明,在入侵检测框架中引入改进算法能够大幅度地降低入侵检测系统的误警率,从而提高系统的检测效率,减少网络攻击所带来的经济损失。

关键词 朴素贝叶斯,入侵检测系统,属性加值,调控参数,误警率
中图法分类号 TP311 **文献标识码** A

Intrusion Detection System Based on Improved Naive Bayesian Algorithm

WANG Hui[1] CHEN Hong-yu[1] LIU Shu-fen[2]
(School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454000, China)[1]
(School of Computer Science and Technology, Jilin University, Changchun 130012, China)[2]

Abstract With increasing Internet connectivity and traffic volume, recent intrusion incidents have reemphasized the importance of network intrusion detection system (IDS). According to the deficiency of the Naive Bayesian (NB) algorithm, this paper proposed an improved NB algorithm. This algorithm based on the original model is combined with a parameter of classification control. It can simplify the complexity of the classification of data and optimize the classification accuracy by computed parameter values. The experimental results prove that the algorithm used in the intrusion detection framework can drastically reduce the false alarm rate of IDS, thereby improve the detection efficiency and decrease economic damage brought by the cyber attack.

Keywords Naive Bayesian, IDS, Value attribute, Controlling parameter, False alarm rate

1 引言

目前,随着基于网络的服务(Web-based Services)逐渐延伸到社会的各项领域以及网络中存在的机密信息数量的大幅度增长,网络安全技术已经变得越来越重要。如何有效地检测并防范网络入侵以保障网络数据的安全性也开始受到各界人士的关注和重视。

近年来,网络攻击(Cyber Attack)的数量及严重程度也随着Internet用户及信息量的增加呈增长趋势。根据赛门铁克(Symantec)最新的互联网安全威胁调查报告[1]披露,从2004年初到2012年底,针对诸如信用卡号、密码以及其他金融信息的网络入侵记录由最初的900多万条增至约6600万条,增长幅度超过750%,由此造成的全球经济损失每年平均高达1140亿美元。因此,入侵检测技术(Intrusion Detection)的研究与发展就愈发显得重要,入侵检测系统(Intrusion Detection System, IDS)也已经成为大多数机构必不可少的基础设施。

入侵检测技术是一种要找出能够危害信息资源完整性、机密性和可用性的一组恶意行为的安全措施[2]。入侵检测处理的问题在于,如何在海量的未知网络事件数据中将正常事件(Normal)和异常事件(Anomaly)精确地分类,以达到过滤网络攻击事件和降低误警率的目的,同时也需兼顾对分类速率的优化[3]。根据文献[2,4]得知,基于数据挖掘(Data Mining)的入侵检测分为两大类:误用检测(Misuse Detection)和异常检测(Anomaly Detection)。误用检测是试图将网络流量中已知的攻击行为的样本与特征进行匹配,它具有误检率低、检测速率快的优点。缺点是误用检测无法检测出新型的入侵行为,但是可以依赖一种学习算法来弥补这些不足。在文献[5]中Guinde指出,这种学习算法是通过训练来得到每一个已标记为正常事件或入侵事件实例的数据集,即使该算法无法检测出未被包含在训练集(Training set)中的新的攻击类型,但是能够通过一个更新过的训练集来结合新型攻击实例再度自动重复训练。对于异常检测,Monowar[6]认为该技术首先建立基于正常网络事件的模型,并且在脱离这些模型的

到稿日期:2013-06-24 返修日期:2013-10-13 本文受国家自然科学基金项目(51174263),教育部博士点基金项目(20124116120004),河南省教育厅科学技术研究重点项目(13A510325)资助。

王 辉(1975—),男,副教授,主要研究方向为计算机网络及网络安全、无线传感器网络等,E-mail:wanghui_jsj@hpu.edu.cn;陈泓予(1989—),男,硕士生,主要研究方向为计算机网络与通信技术;刘淑芬(1950—),女,教授,博士生导师,主要研究方向为计算机协同工作、网络软件、仿真建模等。

情况下检测网络事件。异常检测模型能够检测出新型的攻击事件，因为该模型仅依赖于已知的正常事件。尽管异常检测方法具有优势，但自身会因有预先可能未观测到的正常事件而导致较高的误警率(False alarm rate)，因其较高的误检率和漏检率而使得该方法很难运用到实际中来。介于上述两种检测类型的特点，在入侵检测研究中，也曾提出过结合误用检测与异常检测两种方法的混合入侵检测模型，用以提升预测性能[7]。

因此，研发高效的 IDS 的关键在于如何降低系统的误警率，提升分类的精准度。目前，许多较为有效的入侵检测分类模型被提出。文献[8]提出了结合网络协议分析技术和决策树(Decision Tree，DT)挖掘技术的一种新型模型，该模型分析数据包的协议类型，并根据协议类型来确定最佳的 DT 算法进行入侵检测，该方法在高速网络环境中对入侵的检测具有良好的效果。也有其他学习算法运用在入侵检测领域当中，如支持向量机 (SVM)、遗传算法、人工神经网络等[9-11]。

基于对前者的研究，本文提出了一种改进的朴素贝叶斯算法(Naive Bayes，NB)，该算法是在传统的朴素贝叶斯模型(NBM)基础上引入调控参数，通过对参数值的调整来控制分类的精确度以计算出最佳的分类效率，最后结合实验测试与上述方法的分类性能进行比较，实验数据结果显示出本文方法在入侵检测中在事件分类的精确度和误分率方面都具有较强的优势，尤其是在与改进前的 NB 算法进行比较时，体现出本文方法在准确率方面比传统 NB 算法有较为明显的提升。

本文第 2 节介绍了 Naive Bayes 的相关研究；第 3 节阐述了传统朴素贝叶斯算法，并详细地描述了本文提出的改进后的朴素贝叶斯算法，构建了朴素贝叶斯分类模型，并将该模型运用在入侵检测系统当中进行网络事件的分类流程；第 4 节采用经典的 KDD'99 入侵检测数据集的数据资源对基于本文算法的入侵检测技术进行模拟实验，并给出实验结果，同时对结果数据进行比较，做出分析；最后总结本文。

2 朴素贝叶斯相关研究

朴素贝叶斯算法是基于一个简单假设所建立的一种贝叶斯(Bayesian)方法，NB 假定样本的不同特征属性对样本的归类影响是相互独立的。朴素贝叶斯分类器(Naive Bayesian Classifier，NBC)是以贝叶斯定理为基础的一种应用，目前被广泛使用在信息领域，如垃圾电子邮件过滤、模式识别、入侵检测等[12-14]。选择朴素贝叶斯分类器的原因在于，较其他分类器而言，朴素贝叶斯分类器只需要根据少量的训练数据估计出必要的参数(变量的均值和方差)，对缺失数据不太敏感。由于变量独立假设，只需要估计各个变量的方法，而不需要确定整个协方差 $COV(X_1, X_2)$ 矩阵。

Panda 等人[13]已将朴素贝叶斯模型有效地运用在 IDS 当中，并建立了初级的入侵检测流程框架。在文献[14]中，Nouria 提出了一种朴素贝叶斯算法的扩展算法，该算法利用特征简化与决策树相结合的技术快速地获得最大后验概率 MAP)，在分类效率和准确度上有了明显的提升，但该算法在提取特征和校验合法数据的过程中会条件性地产生大量的伪数据(Pseudo-data)，因此会对异常事件分类造成许多干扰。文献[15]介绍了半监督式的朴素贝叶斯算法，该改进算法是基于并行系统下的计算方法，因此在并发的情况下处理海量网络数据具有较大优势，但不足之处在于针对小型和中型局域网络的异常数据检测效果不佳。局部加权朴素贝叶斯(LWNB)[16]的设计是对 NB 的一种较好的改进，通过与判别频率估计(DFE)的结合引出了逐渐缩小空间(GCS)算法，首先逐步缩小空间并划分出子空间，然后再使用朴素贝叶斯分类器对缩小的范围进行分类。这样的流程设计大大提高了分类器的分类速度和泛化正确率，但是此方法的缺陷在于针对预处理的过程过于复杂，同时增加了时间和空间复杂度 $O(n)$。文献[17]在朴素贝叶斯原有的基础上提出了新的权重(Weight)计算方法，该方法对网络数据提取的特征值进行估计，经过对大量数据特征属性值的选取后，能够大幅度剥离干扰数据，简化和精炼出纯度较高的有效数据，这样既能有效地提高分类器的准确率和召回率(Recall rate)，又能够降低误分和漏分事件的发生率。但该算法在预先提取特征属性值时还存在模糊性，对分类器性能的提升还不够稳定。

本文在借鉴上述成功经验的基础上，针对以上方法的不足，提出了一种改进的 NBM。该模型在原有入侵检测框架上进行了扩展，引入特征选择和离散化的预处理思想，一定程度上简化和降低了海量网络数据的复杂度和伪劣度，同时大幅度提升了事件分类的速度和精确度。与传统的 NB 入侵检测模型不同的是，本文算法在原有 NBM 的基础上引入了属性加值算法，通过对调控参数 θ 值的适当选取来有效地调控网络事件分类的准确度和误分率，并结合机器学习和数据集不断训练的方法，直到得出最佳的分类情况。实验结果表明，该模型具有较高的灵活性和可扩展性，不受应用环境领域的约束，能适用于各种概率性事件的分类情况。该算法具备以下优点：

(1)分类速度快，算法复杂度低；

(2)误报率、漏检率低，误差小；

(3)结构严谨，适应性好，灵活度高；

(4)稳定性好，扩展性强。

3 Naive Bayes 入侵检测模型

3.1 基于改进的 NB 分类模型

贝叶斯决策理论是主观贝叶斯派归纳理论的重要组成部分。贝叶斯决策就是在不完全情报下，对部分未知的状态用主观概率估计，然后用贝叶斯公式对发生概率进行修正，最后再利用期望值和修正概率做出最优决策。其核心思想是：利用先验概率估计后验概率。

朴素贝叶斯分类模型以贝叶斯决策理论为基础，是简化后的贝叶斯概率模型。该分类模型具有实现简单、分类速度快、准确度高的特点，是目前使用最为广泛的分类模型之一。其核心算法如下：

样本 $A=(a_1, a_2, \cdots, a_n)$ 为 n 维布尔向量，用来表示网络事件 A 中特征 $\forall a_i$ 是否出现。网络事件类别 $\exists C \in (C_1, C_2, \cdots, C_m, f)$ 为 m 个类的分类问题，映射函数 $f: A_i \rightarrow C_j$ 表示任意未确定事件实例 A_i 被归类为 m 个类别中的一个 $\exists Cj$ 标签。现有网络事件训练样本 $X_1, X_2, \cdots, X_N$，其中 $X=(x_1, x_2, \cdots, x_t)$ 为 t 维布尔向量，训练样本类别 $c_1, c_2, \cdots, c_k$ 为 k 个事件类别标签。现考虑待分类网络事件样本 $\exists Y_i=(y_1, y_2, \cdots, y_n) \in (Y_1, Y_2, \cdots, Y_N, C_j)$，属于每个类别 C_j ($\forall j=1, 2, \cdots, n$)的概率情况。计算步骤如下：

1. 估计训练样本中类别 c_j 的概率 $P(c_j)$：

$$P(C=c_j)=\frac{\sum_{i=1}^{C}N(c_j)}{\sum_{s=1}^{\infty}T_s} \tag{1}$$

其中，$\sum_{j=1}^{C}N(c_j)$为类别为 c_j 的训练样本数，$\sum_{s=1}^{\infty}T_s$ 为总训练样本数。

2. 估计训练样本中特征 $\exists a_i$ 在事件类别 $\forall c_j$ 中出现的相对概率值 $P(a_i|c_j)$：

$$P(A_i=a_i|c_j)=\frac{\sum_{i=1,j=1}^{\infty}S(\exists a_i\in\forall c_j)}{\sum_{j=1}^{C}N(c_j)} \tag{2}$$

其中，$\sum_{i=1,j=1}^{\infty}S(\exists a_i\in\forall c_j)$表示存在类别为集合 $c_j(1\leqslant j\leqslant m)$ 并包含特征 $\forall a_i$ 的训练样本数。

3. 根据式(1)、式(2)可计算出训练样本中特征 $\forall a_i$ 的概率值 $P(a_i)$：

$$\begin{aligned}P(a_i)&=\sum_{i=1,j=1}^{A\cup C}P(a_i|c_j)P(c_j)\\&=\sum_{i=1,j=1}^{A\cup C}\frac{\sum_{i=1,j=1}^{\infty}S(\exists a_i\in\forall c_j)}{\sum_{j=1}^{C}N(c_j)}P(c_j)\end{aligned} \tag{3}$$

4. 由式(3)，根据贝叶斯公式计算出待分类样本中出现特征 $\exists a_i$ 时样本属于 c_j 的相对概率 $P(c_j|a_i)$：

$$P(c_j|a_i)=\frac{P(c_j)\prod_{j=1}^{\infty}P(a_i|c_j)}{P(a_i)} \tag{4}$$

5. 根据朴素贝叶斯的独立假设得出样本 $\forall Y_i$ 属于类别 c_j 的概率 $P(k)$：

$$P(k)=\prod_{i=1,j=1}^{U}P(c_j|a_i)P(Y=Y_i),(\forall i,j,1\leqslant k<\infty) \tag{5}$$

6. 使用同样的方法求出样本 y_i 属于其他各个类别 $c_j(1\leqslant j\leqslant k)$ 的概率：$\exists P=(P_1,P_2,\cdots,P_k,f)$，映射关系条件 f：$(y_1,y_2,\cdots,y_n)\rightarrow(c_1,c_2,\cdots,c_m),(\forall m\neq n)$。再对这 $\forall k$ 个概率值进行归一化，排序得到待分类样本 y_i 属于各个类别的相似度，计算出最大后验概率 MAP：

$$\begin{aligned}MAP&=\arg\max_{\exists m\in M}P(c_j|y_i)\\&=\arg\max_{\exists m\in M}\frac{\prod_{i=1}^{n}P(y_i|c_j)P(c_j)}{P(y_i)}\\&=\arg\max_{\exists m\in M}\prod_{i=1}^{n}P(y_i|c_j)P(c_j)\end{aligned} \tag{6}$$

7. 根据上述条件，推出相应的朴素贝叶斯分类器 NBC 定义公式：

$$Classifier(y_1,y_2,\cdots,y_n)=\arg\max_{\exists c_j\in C(1\leqslant j\leqslant m)}P(c_j)\prod_{i=1,j=1}^{n}P(y_i|c_j) \tag{7}$$

在以上数学模型的基础上，构建出 NBM 结构，该分类模型拥有一个类别节点 $\exists C\in(C_1,C_2,\cdots,C_m)$ 和 n 个待分类节点 A_i，其中每个 A_i 由 n 个相互独立的特征值 a_i 组合描述，节点间的映射关系为：$f:\{C_1:\langle A_1,(a_1,a_2,\cdots,a_m)\rangle,C_2:\langle A_2,(a_1,a_2,\cdots,a_m)\rangle,\cdots,C_k:\langle A_n,(a_1,a_2,\cdots,a_m)\rangle\},(\forall m\neq n)$。所有待分类节点 A_i 都隶属于共有类别节点 C，且每个待分类节点间的关系也是相互独立的。NBC 结构示意图如图 1 所示。

由图 1 可知，关系模型中各节点间的联系及其结构意义同时也描绘出了贝叶斯决策定理的由先验概率求出后验概率的核心思想，再通过后验概率更新样本集合节点的训练学习过程，最后计算出最大可能性的分类情况。因此，尽管实际上独立假设常常是不准确的，但结合实验得出，NBC 的若干特性依然能够让其在实践过程中取得令人惊奇的分类效果。

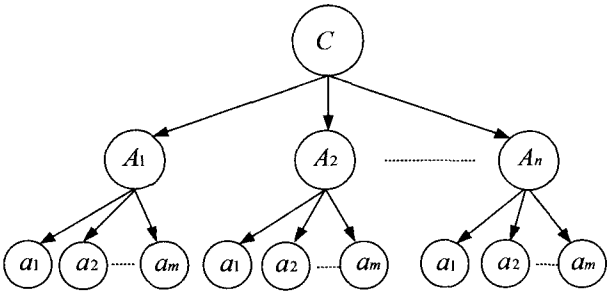

图 1　NB 分类模型结构示意图

3.2　引入属性加值算法的 NBC

从理论上讲，朴素贝叶斯分类算法是通过计算出待分类节点所属类别的最大后验概率 MAP 来确定最终分类节点 c_j $(\forall j,1\leqslant j\leqslant n)$。但在实际当中，不同因素的影响，例如特征属性样本节点选取的不同和训练样本集合的不同，会造成 NBM 的分类精确度在不同程度上的衰减。因此，针对上述朴素贝叶斯算法的缺陷，本文将在其模型的基础上引入属性加值算法，以提升 NBC 的分类精确度并大幅度地降低其误分率。

首先针对入侵检测技术可将待分类网络事件 $A_i(\forall i,1\leqslant i\leqslant n)$ 归类为正常事件(Normal) $C_N(N$ 为常量)或者异常事件(Anomaly)$C_j(\forall j,1\leqslant j\leqslant m)$，其中有全集合 $U_C=\{C_N,C_j|\exists P(C_N)+P(C_j)=1\}$。通常再根据朴素贝叶斯公式分别计算出后验概率 $P(C_N|A_i)$ 和 $P(C_j|A_i),(\forall i,j,1\leqslant i\leqslant j\leqslant n)$。因此，仅根据计算得出的最大后验概率 MAP 来模糊地判断 $P(C_j|A_i)>P(C_N|A_i)$，最后将其归类于某一种事件类型，是不够严谨的。这样不仅会造成较高的误检率，同时还会制造出错误的样本集合数据，从而在一定程度上降低了入侵检测技术的预测能力和鲁棒性。

因此，本文在上述算法的基础上引入属性加值法，通过调控参数值 θ 来进一步控制分类的精度。通过适当地选取 θ 值使得 NBC 的分类效率达到最佳的优化效果。由上式进一步得出：$\frac{P(C_j|A_i)}{P(C_N|A_i)}>1$，引入控制因子 θ 转换而得出 $\ln\frac{P(C_j|A_i)}{P(C_N|A_i)}>\theta$，最后推算出由 θ 值影响的改进判别式：

$$\frac{P(C_j|A_i)}{P(C_N|A_i)}>e^{\theta}\Rightarrow\frac{P(C_j|A_i)}{1-P(C_j|A_i)}>e^{\theta}\Rightarrow P(C_j|A_i)>\frac{e^{\theta}}{1+e^{\theta}} \tag{8}$$

由式(8)的推算可以得知，经过对调控参数 θ 值的不同选取就可以通过映射关系函数推算式 $f:\{U_{MAX}[e^{\theta}/(1+e^{\theta})]\rightarrow\psi\}$ 计算出与之相对应的判别值 ψ。根据夹逼定理可知，给定极限 $\lim P(C_j|A_i)=\psi$，根据任意正数 α，当 $\exists n>N$ 时，$|P(C_j|A_i)-\alpha|<\psi$，即 $\alpha-\psi\leqslant P(C_j|A_i)\leqslant\alpha+\psi$ 成立，其中 $\alpha-\psi>0$ 且 $\alpha+\psi<1$。最后根据计算出的极限 $f(X_P)$ 就可以估算出最合适的 ψ 值，从而可以更加准确地判定出事件 A_i 属于的某一类别 C_j。其中，选取的最佳参数 θ 值是根据实验测试计算出的对比数据所产生的最终结果。

3.3　基于 NBC 的入侵检测流程

实现入侵检测的方法从根本意义上讲就是设计一个事件分类器来将数据流中的正常与异常数据区分出来，从而实现

对攻击行为的报警功能。与此同时，IDS 对入侵事件的检测又是一个不确定行为，而朴素贝叶斯理论正适合于不确定的概率性事件，因此在针对 IDS 的研究设计时引入基于 NBC 的入侵检测技术是完全合理的。

针对上述原理的分析，本文将改进后的 NBC 运用在入侵检测模型的分类模块当中，通过一系列入侵检测流程的处理，最终得出合理的网络事件归类集合。该模型的入侵检测流程如图 2 所示。

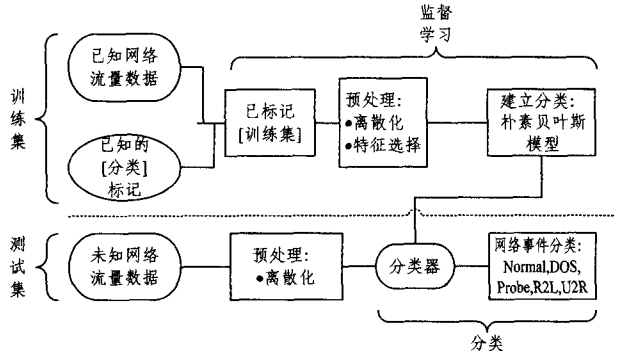

图 2　基于 NBC 模型的入侵检测流程

根据图 2 中的流程结构可知，第一阶段，首先将已知的网络流量数据 T_k 与样本中已知的类别 C_k 通过映射集合 $U_{TC}=(\exists T_k, C_k \mid \forall T_k \in C_k)(1 \leqslant k \leqslant n)$ 进行结合并标记到训练集(Training set，TrS)当中进行训练，将训练集中标记过的复杂数据进行预处理(离散化、特征选择)，接着将预处理中过滤出的有效而简单化的数据进行统计和计算，最后根据估计出的先验概率 $P(T_k \mid C_k)$，$(\exists k=1,2,\cdots,n)$ 建立出基于 NB 算法的事件分类器 Classifier。此阶段的整个检测流程是一个重复的监督学习过程，用以逐渐完善样本集中的归类数据，从而使得分类器具有较好的预测效果。第二阶段，首先将测试集(Test set，TeS)中的未知网络流量数据 T_u 进行分类前的预处理(离散化)，然后根据第一阶段预先建立好的 $NBC:\{(U_u=\langle T_u, C_k\rangle) \subseteq (U_k=\langle T_k, T_u, C_k\rangle)\}$ 进行分类测试，然后通过映射关系函数 $f_{NBC}: T_u \rightarrow C_k$，匹配未知事件与样本的标记事件，进而得出待分网络事件 $X=(X_1, X_2, \cdots, X_n)$ 的所属事件类别 $\exists C \in \{(C_j \neq \emptyset, 1 \leqslant j \leqslant n) \mid \text{Normal, DOS, Probe, R2L, U2R}\}$，同时再次更新样本集中的数据，以便能完备检测模型的检测性能。整个框架流程所采用的 TrS 和 TeS 中的数据资源均来自经典的 KDD Cup 1999(KDD'99)入侵检测数据集，该数据集的设计模式能够检验分类器模型的泛化能力和预测能力。针对海量而复杂的网络数据测试前的预处理过程，本文借鉴了诸如离散化(Discretization)和特征选择(Feature Selection)的思想，其数据处理的目的是：(1)使抽象的网络数据具体化；(2)去除冗余特征和不重要特征来简化数据；(3)降低时间和空间复杂度，提高分类器训练速度和检测精度。预处理网络数据在入侵检测流程中是必不可少的前提条件，是对分类器模型抗干扰性能的有效提升。

4　实验结果与分析

4.1　入侵检测数据集

本文实验数据采用 KDD Cup 1999(KDD'99)入侵检测数据集，此数据集中的数据包含两个部分：7 周时间的训练数据大概包含 5000000 多个网络连接记录，剩下的 2 周时间的测试数据大概包含 2000000 个网络连接记录。每个网络连接被标记为正常(Normal)或异常(Anomaly)，异常类型被细分为 4 大类共 39 种攻击类型，其中有 22 种攻击类型出现在训练集中，另有 17 种未知攻击类型出现在测试集中。4 种攻击类型及其描述和举例由表 1 所列，表 2 中的数据为网络事件 5 种类型的记录数目以及分别在 10% KDD'99 入侵检测数据集中的 TrS 和 TeS 的分布情况。

表 1　异常类型

4 种攻击类	类型描述	22 种攻击类型
DOS	拒绝服务攻击	Ping-of-death, syn flood, smurf etc.
R2L	来自近程主机的未授权访问	Guess password
U2R	未授权的本地超级用户特权访问	Buffer overflow Attacks etc.
Probe	端口监视或扫描	Port-scan Ping-sweep etc.

表 2　攻击类型在 10% KDD 数据集中的分布

攻击类型	训练样例	测试样例	10% KDD 训练集分布	10% KDD 训练集分布
Normal	97277	60592	19.69%	19.48%
DOS	391458	237594	79.24%	73.91%
R2L	1126	8606	0.23%	5.20%
U2R	52	70	0.01%	0.07%
Probe	4107	4166	0.83%	1.34%
Total	494020	311028	100%	100%

4.2　实验结果分析

实验过程中搭建的环境平台所使用的操作系统为 Windows XP，CPU 频率为 2.3GHz，内存为 4GB，500G 硬盘存储空间，编程工具为 MATLAB 8.0，入侵检测数据集则使用上述 10% KDD'99 入侵检测数据集的 TrS 和 TeS 进行模拟试验，不同的实验结果分别如表 3—表 5 所列。

表 3　改进前后 NB 算法检测率比较

Method	Normal	DOS	R2L	U2R	Probe
Improved NB Algorithm(TR%)	98.69	99.25	99.17	98.74	97.81
Improved NB Algorithm(FP%)	0.07	0.05	0.08	0.14	0.03
Traditional NB Algorithm(TR%)	96.28	95.87	92.12	93.85	94.74
Traditional NB Algorithm(FP%)	0.09	0.06	0.13	0.18	0.08

(1)由表 3 的实验结果可以得出，经过改进后的 NB 算法较传统的 NB 算法在对针网络入侵事件(DOS，R2L，U2R，Probe)分类的准确率(TR)和误检率(FP)方面都具有明显的改善。因此，本文通过将调控参数引入 NBM 并通过缩放参数极限值的方法来控制分类的精度是可行的。另外，在图 3 和图 4 中可以明显看出，改进后的 NB 算法在分类效果上的提升幅度还是比较可观的。并且，由于训练样本值的不断扩展，条件概率的估计误差也会随之逐渐地缩小，因此根据改进算法自身的连续学习性和可扩展性，会使得改进效果提升得更加理想。

表 4　θ 取值对检测率的影响

θ(取值)	0.0	0.5	1.0	1.5	2.0	2.5	3.0
TR(%)	95.26	96.37	97.63	98.21	97.74	95.62	94.45
FP(%)	9.42	7.16	6.55	6.28	7.83	8.24	9.71

(2)从表 4 中计算得出的调控参数 θ 的取值可以观察出，根据 θ 值的递增变化($\forall\theta\in[0,\infty)$)，TR 和 FP 在整个演化过程中会分别呈现出一个高峰值 $H(v)$ 和低谷值 $L(v)$，这两个值所对应的 θ 值是实验结果期望得到的最佳参数。由图 5 可知，首先，本文通过设置极限预警值(阈值)来划定参数值 θ 更加合理的取值范围($\forall\theta\in[0.5,2.5]$)，超越阈值之外的参数值部分为摒弃值 $V(ab)$，将不会出现在 θ 值的选取范围之内。然后在建立好的参数样本集合 $U_v=\{V_\theta=(0.5,0.6,\cdots,2.5)\mid f_u\}$ 中根据参数映射函数 $f_u:\{[V_{HL}=\langle H(v),L(v),\rangle]\rightarrow V_\theta\}$ 可计算出最佳的调控参数值 $\theta_{best}=1.5$。最后将得出的 θ_{best} 值注入本文算法，从而得出基于本文算法的入侵检测模型的最优检测效率。

表 5　多种事件分类方法的性能比较

Method	Accuracy(%)	Error Rate(%)
Improved Naive Bayes	98.21	6.28
决策树(DT)	97.83	7.72
支持向量机(SVM)	97.76	8.13
Naive Bayes	96.95	9.85
人工神经网络(ANN)	94.98	9.47
遗传算法(GA)	91.64	8.82

(3)将本文改进的 NB 算法与其它分类算法在入侵检测领域方面做了系统的比较，从表 5 的实验数据中可以看出，在精确度(Accuracy)和错误率(Error Rate)方面，改进方法较其他方法都体现出了更好的效果。本算法在规避 NB 算法缺陷的基础上改良了 NB 算法，并结合了预处理的成功经验加以优化，因此在分类精确度上明显超过了 Naive Bayes 分类方法。

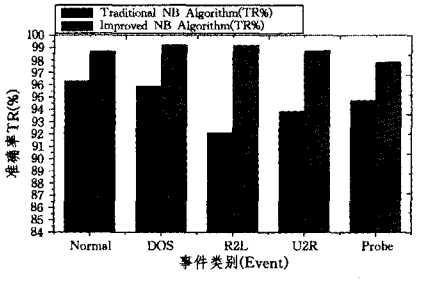

图 3　改进前后 NB 算法分类准确率(TR)的比较

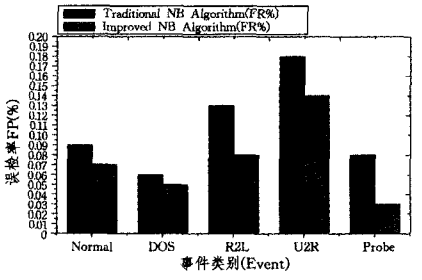

图 4　改进前后 NB 算法分类误检率(FP)的比较

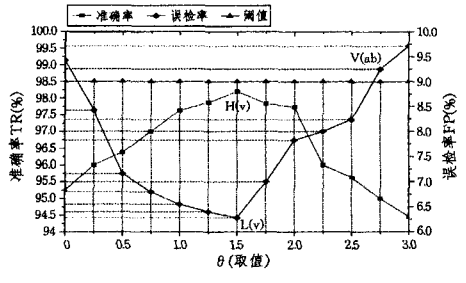

图 5　影响分类的 θ 取值的分布概率

结束语　针对互联网中海量而复杂的网络攻击事件，入侵检测行为其实是一系列不确定性行为过程的一个组合，鉴于朴素贝叶斯定理最适合解决概率性事件的原因，将朴素贝叶斯决策理论分类器(Classifier)运用到入侵检测技术当中是完全可行的。本文借鉴了前人的一些经验和思想，并针对其不足之处对 NB 算法做了适当的改进，提出了一种基于属性加值的 NB 改进算法。该算法在入侵检测流程中引入了预处理(Preprocessing)和数据挖掘(DM)等经典的数据处理方法，通过数据特征的进一步提取，降低了分类器分类的时间复杂度 $O(n)$。在此基础上，本算法通过实验测试计算出了最佳的分类调控参数 θ_{best}，并结合此参数获得了分类器最优的分类效果。最后经过实践的证明，基于本算法的 IDS 的误警率被大幅度地降低，并且在与其它分类算法的分类效果进行比较时，同样体现出了本算法较高的优势。但本算法各方面的性能仍然有很大的提升空间，在今后的工作当中，如何更加有效地细化调控参数并与其它高效率的分类方法进行巧妙的结合，从而更进一步提升分类器对复杂多变的网络数据的预测性能，是接下来的研究计划。

参 考 文 献

[1] 2013 Internet Security Threat Report (Symantec. com [OL]. http://www. symantec. com/security _ response/publications/threatreport. jsp

[2] Govindarajan M, Chandrasekaran R M. Intrusion detection usingneural based hybrid classification methods [J]. Computer Networks, 2011, 55(8): 1662-1671

[3] García-Teodoro P, Díaz-Verdejo J, Maciá-Fernández G. Anomaly-based network intrusion detection: Techniques, systems and challenges [J]. Computers & Security, 2009, 28(1/2): 18-28

[4] Mohammad M N, Sulaiman N, Muhsin O A. A novel intrusion detection system by using intelligent data mining in weka environment [J]. Procedia Computer Science, 2011, 3(1): 1237-1242

[5] Guinde N B, Ziavras S G. Efficient hardware support for pattern matching in network intrusion detection [J]. Computers &Security, 2010, 29(7): 756-769

[6] Bhuyan M H, Bhattacharyya D K, Kalita J K. Survey on Incremental Approaches for Network Anomaly Detection [J]. International Journal of Communication Networks and Information Security, 2011, 3(3): 226-239

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

[8] 杨杰，陈昕，万剑雄. 网络协议分析与决策树挖掘的入侵检测模型研究[J]. 计算机应用与软件，2010，27(2)：19-55

[9] 徐永华，李广水. 基于距离加权模板约简和属性信息熵的增量SVM入侵检测算法[J]. 计算机科学，2012，39(12)：76-86

[10] Li Lu, Zhang Guo-yin, Nie Jin-yuan. The Application of Genetic Algorithm to Intrusion Detection in MP2P Network [J]. Lecture Notes in Computer Science, 2012, 31(3): 390-397

[11] Wang Gang, Hao Jin-xing, Ma Jian. A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering [J]. Expert Systems with Applications, 2010, 37(9): 6225-6232

[12] 邓维斌，王国胤，洪智勇. 基于粗糙集的加权朴素贝叶斯邮件过滤方法[J]. 计算机科学，2011，38(2)：218-221

(下转第 119 页)

$(r^\beta)'$，于是标签 α 与标签 β 在下一轮(特指以两标签之一作为搜索目标的下一轮)必将再次发生响应碰撞。

②如果 $k'_{i\alpha}\neq k'_{i\beta}$ 且 $r'_{i\alpha}\neq r'_{i\beta}$，则有 $0<P[(r^\alpha)'=(r^\beta)']<1$，尽管这个概率很小，但两个标签在下一轮仍然有发生响应碰撞的可能。

③如果 $k'_{i\alpha}=k'_{i\beta}$ 但 $r'_{i\alpha}\neq r'_{i\beta}$，按照协议的设定，二维函数 f 对其中任意一维变量服从平均分布的属性，则必然有 $f(k'_{i\alpha}, r'_{i\alpha}\oplus r_b')\neq f(k'_{i\beta}, r'_{i\beta}\oplus r_b')$，即 $(r^\alpha)'\neq(r^\beta)'$，于是这两个标签不会在下一轮再次发生响应碰撞。

很显然，本文的设计正好符合了第 3 种情况，所以从根源上彻底避免了连环响应碰撞的可能性。

其次，由于更新 k_{ij} 与 r_{ij} 这两个秘密值的过程是在标签和 R_i 的硬件内部进行的，ID 值从来没有以明文或加密的方式进入无线环境的通信过程中；因此，标签的 ID 值没有暴露的危险，并且对秘密值更新方案的改进也没有破坏文献[6]中已经证明过的安全性。为了更加保险起见，可以赋予每个标签一个假名，记作 metaID，而 metaID=hash(ID)，更新秘密值时以 $r_4\oplus$ metaID 更新 r_{ij} 的值。

另外，即使攻击者在下一轮幸运地破解了 r'_{ij} 的值，他除非在本轮同样幸运地得到了 r_4 的值，才能在同时掌握 r_4 与 r'_{ij} 两个值的情况下破解 metaID。而要得到 r_4 的值，则必须同时得到本轮的 k_{ij} 与 r_{ij} 的值，即意味着在本轮该标签就已经被攻破，这样就产生了一个悖论，因而 r_4 是安全的。所以，即使下一轮的 r'_{ij} 的值被破解，但由于本轮的 r_4 是安全的，metaID 必然是安全的，而标签的 ID 值更加是安全的。因此，更新后的 SSP$^+$ 协议既具有前向安全性，也具有后向安全性。

综上所述，经过改进后的 SSP$^+$ 协议，除继承了 SSP 协议的安全性能之外，还避免了连环响应碰撞的安全隐患，具有前向与后向安全性。所以，SSP$^+$ 协议是安全的。

结束语 随着 RFID 技术日益广泛地应用于人们的生产、生活，需要从大量的 RFID 标签中快速、安全地对目标标签进行搜索，因此设计安全、高效的 RFID 安全搜索协议具有重大的现实意义。本文根据这种现实需要，提出了搜索协议中容易被人们忽视的碰撞问题，并分析了碰撞问题发生的概率及可能引发的安全性问题。然后，在 SSP 协议的基础上，对其进行安全性扩展与改进，改进后的 SSP$^+$ 协议很好地弥补了这一缺陷，消除了遭受追踪攻击的隐患。

参 考 文 献

[1] 孙其博，刘杰，黎彝，等. 物联网：概念、架构与关键技术研究综述[J]. 北京邮电大学学报，2010，33(3)：1-9

[2] 卿斯汉. 安全协议 20 年研究进展[J]. 软件学报，2003，14(10)：1740-1752

[3] Tan C C, Sheng B, Li Q. Serverless search and authentication protocols for RFID[C]// 5th IEEE International Conference on Pervasive Computing and Communications. New York: IEEE Press, 2007: 24-29

[4] Ahamed S, Rahman F, Hoque E, et al. S^{3}PR: secure serverless search protocols for RFID[C]// International Conference on Information Security and Assurance. Hawaii: IEEE Press, 2008: 187-192

[5] Kulseng L, Yu Z, Wei Y. Lightweight secure search protocols for low-cost RFID systems[C]// 29th International Conference on Distributed Computing Systems. Washington: IEEE Press, 2009: 40-48

[6] 邓淼磊，衡晓鹏，鲁志波. 安全的 RFID 搜索协议[J]. 西安通信学院学报，2009，8(5)：78-81

[7] 曹峥，邓淼磊. 通用可组合的 RFID 搜索协议[J]. 华中科技大学学报：自然科学版，2011，39(4)：56-59

[8] van Le T, Burmester M, Medeiros B. Universally composable and forward secure RFID authentication and authenticated key exchange[C]// Proc. of the 2nd ACM Symposium on Information, Computer and Communications Security. New York: ACM, 2007: 242-252

[9] Burmester M, van Le T, Medeiros B. Provably secure ubiquitous systems: Universally composable RFID authentication protocols [C]// Proc. of the 2nd International Conference on Security and Privacy in Networks. Maryland: IEEE Press, 2006: 176-186

[10] 邓淼磊，马建峰，周利华，等. RFID 匿名认证协议的设计[J]. 通信学报，2009，30(7)：24-31

[11] Lee S M, Hwang Y J, Lee D H. Efficient authentication for low-cost RFID systems[C]// Proc. of the International Conference on Computational Science and Its Applications (ICCSA2005). Berlin: Springer-Verlag, 2005: 619-627

[12] Canetti R. Universally composable security: a new paradigm for cryptographic protocols[C]// 42th IEEE Annual Symposium on Foundations of Computer Science. Oakland: IEEE Press, 2001: 136-145

[13] Kurosawa K, Furukawa J. Universally composable undeniable signature[C]// Proc. of ICALP'08. Berlin: Springer-Verlag, 2008: 524-535

[14] Canetti R, Herzog J. Universally composable symbolic analysis of mutual authentication and key-exchange protocols[C]// Theory of Cryptography Conference. Berlin: Springer-Verlag, 2006: 380-403

(上接第 115 页)

[13] Panda M, Patra M R. Network Intrusion Det-ection Using Naive Bayes [J]. International Journal of Computer Science and Network Security, 2007, 7(12): 258-263

[14] Farid D M, Harbi N, Rahman M Z. Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection [J]. International Journal of Network Security & Its Applications, 2010, 2(2): 12-25

[15] 江凯，高阳. 并行化的半监督朴素贝叶斯分类算法[J]. 计算机科学与探索，2012，6(10)：912-918

[16] 欧阳泽华，郭华平，范明. 在逐渐缩小的空间上渐进学习朴素贝叶斯参数[J]. 计算机应用，2012，32(1)：223-227

[17] 周晓庆，肖顺文，肖建琼，等. 一种基于改进的权值调整技术数据源分类算法研究[J]. 计算机应用研究，2012，29(3)：916-918