基于加权特征筛选的入侵检测系统

王鹏英 黄 海 黄晓平
(浙江理工大学计算机技术教研部 杭州 310018)

摘 要 网络攻击隐蔽性高,手段多样。传统检测系统特征提取不全,数据包易丢失,漏报、错报率高。为提高检测率,提出一种基于加权特征筛选的入侵检测算法。首先对网络数据包进行特征提取;然后采用支持向量机交叉验证对全部特征进行筛选,并计算各特征的权值;最后以加权保留特征构建入侵检测模型。仿真实例结果表明,该检测算法提高了入侵检测率,是一种有效的网络入侵检测方法。

关键词 入侵检测,特征筛选,特征加权,支持向量机
中图法分类号 TP391 **文献标识码** B

Intrusion Detection System Based on Choosing Characters and Weighting Characters

WANG Peng-ying HUANG Hai HUANG Xiao-ping
(Instructional Division for Computer Technology of Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract The network intrusion means are diversification. Traditional detection system can not extract feature very well. Packet is easy lost, and omission and misstatement rates are high. In order to improve the detection rate, this paper proposed an intrusion detection algorithm based on weighted feature selection. Firstly, features were extracted from network packets, then support vector machine (SVM) was used to select feature based on cross validation and calculate the feature values, lastly, intrusion detection mode was set up based on the weighted reserves features. The results of simulation experiment show that the proposed algorithm improves the intrusion detection rate. It is an effective network intrusion detection method.

Keywords Intrusion detection, Choosing characters, Weighting characters, Support vector machine (SVM)

1 引言

计算机网络技术迅速发展,其安全性问题日益突出。传统的单纯防御措施已无法满足安全性要求。计算机入侵检测系统能对入侵程序进行主动跟踪,并及时对入侵行为发出警报,已经成为当前网络安全的研究重点与热点问题[1]。传统的入侵检测系统主要基于主机的系统日志、各种程序日志等审计跟踪数据,通过模式匹配的方式进行入侵行为检测,存在要求主机系统本身具有较好的安全配置、易被入侵行为逃避审计、实时性差等缺点;也有部分检测系统基于网络数据进行入侵检测,但也存在无法检测不同网段的数据包、处理加密会话过程难等缺陷[2]。此外,传统入侵检测系统在其识别率不够时,一般采取增加大量信息特征的手段进行弥补。特征过多,一方面会导致计算复杂度增加,使得检测系统检测速度减慢,尤其是在当前高速网络情况下,检测系统会丢弃部分处理不及时的数据包,而被丢弃的包若含有入侵行为,则可导致检测系统漏报率大大增加;另一方面,增加的大量特征可能存在大量的冗余信息,甚至包含了无用特征,该类特征的存在会严重削弱入侵检测模型的识别能力[3,4]。

针对传统入侵检测系统检测数据来源局限、高速网络下漏报、误报增加的问题,本文基于加权特征筛选融合主机审计数据与网络数据包提出了一种新的入侵检测算法。不管是基于主机跟踪审计数据或是基于网络数据包的检测系统都存在检测数据不全,导致漏报误报高的问题。本文通过融合两种检测数据对入侵行为进行检测,增了检测范围,可以有效提高入侵检测正确率;另外,因数据的融合以及网络速度的提高,检测系统需处理的特征数量会急剧增加,其中可能包含了大量冗余或无用特征,不仅加大了检测系统的运算时间,对检测模型性能也造成较大的负面影响。本文基于支持向量机(Support Vector Machine, SVM)交叉验证对全部特征进行检测,并剔除所有对提高模型精度有不利影响的特征;最后基于强制筛选,确定各保留特征的权值。在保留加权特征的前提下构建了入侵检测模型并进行了仿真实验,结果表明所构模型在低速网络与高速网络情况下均保证了较高的检测精度。

2 入侵检测原理

入侵检测主要是指对当前收集的计算机数据进行分析,以确定计算机是否被非法行为入侵。当系统被非法行为入侵

到稿日期:2011-02-10 返修日期:2011-05-15 本文受2011年浙江省自然科学基金(Y1110157)资助。
王鹏英(1977—),女,硕士,主要研究方向为计算机应用,E-mail:waterwpy@zstu.edu.cn;黄 海(1975—),男,博士,讲师,主要研究方向为计算机网络安全;黄晓平(1974—),女,博士,讲师,主要研究方向为计算机应用。

后，通常会引起系统日志文件、系统进程、系统目录、网络协议分析、网络流量等的变化，通过收集与分析该类数据即可检测到入侵行为，并发出警报[5]。传统的入侵检测系统可分为4个主要功能组件：事件产生器(Event generators)、事件分析器(Event analyzers)、响应单元(Response units)与事件数据库(Event databases)。事件产生器主要功能是在计算机各个检测位点收集源数据，并按事件分析器的要求进行数据预处理与特征提取；事件分析器是入侵检测系统最为核心的部分，其基于已经产生的特征进行分析处理，以确定入侵行为是否发生；响应单元的主要功能是在系统检测到入侵行为后，采取相应的措施来保护系统的安全性或完整性，如终止相关进程、修复文件、修复连接等；事件数据库的主要功能是储存入侵行为数据，以供需要时使用。其流程如图1所示。

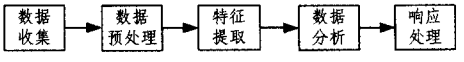

图1 传统入侵检测流程图

在入侵检测系统中，如何在高速网络情况下提高系统的检测速度，减少入侵信息的丢失并收集足够的有效入侵行为数据，是其技术难点。传统模型因考虑运算时间问题，要么仅收集基于主机日志的数据，要么仅收集基于网络数据包的数据，导致收集的数据过于单一，提供的信息有限，漏报、误报率非常高。本文通过融合主机日志数据与网络数据包数据，最大程度地提高数据覆盖面，增加数据信息的充分性、事实性与可靠性。两类特征的融合使得检测特征数目较大，其中不乏信息冗余或无用的特征。一方面，过多的特征使得检测系统运算速度减慢，导致在高速网络情况下会丢失大量信息数据包；另一方面，多余的特征可严重削弱检测模型的性能，这些多余特征理应被剔除。本文基于SVM交叉测试逐一检验提取的全部特征，并剔除相应的多余特征，以简化模型。另外，保留的特征对检测模型的贡献率也应该因其能提供的信息量不同而有所差异，所以本文进一步基于强制汰选，为各保留特征赋予相应的权值，最后基于加权保留特征构建入侵检测模型。改进的入侵检测模型流程见图2。

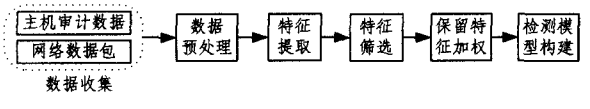

图2 改进的入侵检测模型流程图

3 加权筛选模型构建过程

3.1 特征转换

本文入侵检测系统基于SVM，所以提取的特征必须维数相同且都为数值向量形式。直接收集的入侵行为特征包含非数值向量，例如基于网络数据包的41维常用特征中包含了3维非数值特征，其具体数据格式如下：

1,tcp,discard,RSTO,0,0,2,0,0,0,2,0,1,0,0,1,0,0,0,1,0,0,236,9,0.00,0.00,1.00,1.00,0.04,0.07,0.00,255,9,0.04,0.07,0.00,0.00,0.00,0.00,1.00,1.00,Neptune。

1,icmp,ecr_i,SF,520,0,1,0,5,1,0,3,0,0,0,0,0,0,1,0,1,0,473,473,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,smurf。

1,tcp,finger,SF,520,0,1,0,0,2,0,1,0,8,0,0,0,0,1,0,0,0,78,2,1.00,1.00,0.00,0.00,0.01,0.08,1.00,1,6,1.00,0.00,1.00,0.50,1.00,0.17,0.00,0.00,land。

显然，其中第2－4维特征为非数值格式，需转换为SVM数值格式。其转换过程如下。

(1)统计某列非数值特征的类型数目 n。例如上例的第3列特征包含3种不同类型，n 记为3；

(2)对该列基于 $n-1$ 位数进行二进制编码。上例第3列特征2位二进制编码可表示为00,01,10，分别对应 discard,ecr_i,finger；

(3)循环所有特征，直到所有非数值格式特征全部编码完毕。

3.2 SVM交叉特征筛选

基于主机审计数据与网络数据包可收集大量的特征，但不是每个特征都对提高入侵模型检测精度有贡献。多余的特征不仅增加入侵模型的运算时间，还会削弱模型的分类精度。本文基于SVM交叉测试对每个特征进行验证。

3.2.1 SVM及交叉测试

SVM是一种基于统计学习理论的机器学习方法[6]，其基于结构风险最小，通过核函数映射到高维空间进行函数决策。本文以径向基核函数(RBF核)作为最优核函数。

使用RBF核时，需要设置两个参数：C 和 γ，其参数的选择是否合适直接决定了其分类精度的高低。对于特定问题，并不知道哪一对参数组合是最优的，需要进行最优参数搜索，以确定一对效果较好的 C、γ 组合。基于训练集寻找最优参数组合的这一过程即交叉验证过程：对于一个 v 组的交叉验证，首先将训练集分成大小一致的 K 个子集，确定第一份子集为验证集，其后以剩余的 $K-1$ 个子集训练，并预测验证集。直到每份子集全部被预测完毕，最后所得精度即可看作是对模型决策性能的评估。

3.2.2 特征筛选

针对某一特定的数据集，其 n 次交叉测试精度可用于模型的决策性能评估。假定数据包含 m 个特征，其交叉测试筛选过程如下：首先对包含全部特征的数据基于SVM进行10次交叉测试，得交叉测试精度 $rate_0$；随后剔除第 $i(i=1,\cdots,m)$ 个特征，计算其交叉测试精度 $rate_i$。比较 $rate_i$ 与 $rate_0$ 大小，$rate_i$ 大于 $rate_0$，则表示第 i 个特征的剔除能有效提高模型的分类精度，所以应该剔除所有 $rate_i$ 大于 $rate_0$ 的第 i 个特征。

3.3 特征加权

特征汰选后，假设还剩余 j 个特征，称其为保留特征。每个保留特征代表的信息各不相同，信息量也各有差异，其对入侵检测模型的贡献理应不同。同样，基于保留特征交叉测试可获得各保留特征对模型的贡献率，即权重。首先基于全部保留特征实施交叉测试，得其交叉测试精度 $rate_{save}$；然后强制逐个地剔除每个保留特征，计算其交叉测试精度 $rate_i(i=1,\cdots,j)$。各保留特征权重计算公式如下：

$$w_i=\frac{rate_i-rate_{save}}{\sum_{i=1}^{j}(rate_i-rate_{save})}$$

各保留特征相应乘上各自的权重，产生加权保留特征，最后基于加权保留特征构建入侵检测模型。

4 仿真实验

4.1 数据来源

本研究实验数据通过模拟麻省理工学院林肯实验室提供的 KDD CUP99 数据集产生，总共包含 7000000 条连接信息，被标记为正常或被攻击。每条连接信息包括流量特征集、主机流量特征集、基本特征集和内容特征集等 41 维网络数据特征。另外，包含基于主机审计数据选择重要端口扫描次数、探测操作系统版本次数、IP 扫描次数、用户在磁盘上所建目录数量、敏感数据访问次数、用户的源、非法连接数、特殊文件访问数量、用户的目的地址、超级用户权限访问次数、用户程序所占线程数、在当前时期口令错误的次数、用户连接建立成功率等 13 个特征。随机抽取 2000000 条信息作为独立测试样本，剩余为训练样本。

4.2 模型及识别结果

基于上述新方法对训练样本构建入侵检测模型(screen-weight)，并对独立测试进行判别；同时另外构建了 6 个参比模型，用于结果对比。这 6 个参比模型是基于遗传算法的入侵检测模型(GA)；基于神经网络的模型(NN)[7]；基于 SVM 的模型，既不进行特征筛选也不加权；仅基于主机审计数据的模型(HIDS)，除数据特征外，其它过程与新模型 screen-weight 一致；仅基于网络数据包的模型(NIDS)；screen 模型，除不进行加权外，其它与新模型一致。另外，为检测不同网络流量下各模型的入侵检测情况，7 个检测模型分别在 900Mb/S 与 1300Mb/s 两种情况下进行了仿真实验。各模型检测精度见表 1。

表 1 不同模型检测精度

模型	网络流量为 900Mb/s 时检测精度	网络流量为 1300Mb/s 时检测精度
GA	83%	66%
NN	85%	71%
SVM	90%	75%
HIDS	89%	89%
NIDS	91%	87%
screen	96%	90%
screen-weight	98%	95%

4.3 结果分析

从以上结果可知，本文模型(screen-weight)在所有参比模型中检测精度最高，稳定性最好。

首先，从机器学习算法方面分析，通过遗传算法模型(GA)、神经网络模型(NN)、支持向量机模型(SVM)的比较可知，在网络流量为 900 Mb/s 时，SVM 精度最好；当网络速度增加时，其检测精度降低得最慢。显然，相比另外两种学习方法，基于结构风险最小的 SVM 在入侵检测系统中具有更好的适用性。

其次，从数据特征选择方面来看，在网络流量为 900Mb/s 时，基于网络数据模型(NIDS)的检测精度略高于基于主机审计数据的 HIDS 模型，但两个模型检测精度都明显低于本文模型。可见，融合两类数据能有效增加特征信息量，提高检测模型的检测精度。

从特征筛选方面来看，当网络流量为 900 Mb/s 时，经过特征筛选的 screen 模型其检测精度高于未经过特征筛选的 SVM 模型 6 个百分点。可见，未经过筛选的特征中包含了大量的冗余或无用特征，导致模型检测精度下降；当网络流量增加到 1300 Mb/s 时，screen 模型精度仅下降了 6 个百分点，而 SVM 模型精度下降了 15 个百分点。可以认为，过多的特征使得模型计算复杂度增加，而丢失了大量的信息数据，最终导致了模型精度急剧下降。

最后，在特征加权方面，screen-weight 模型检测精度在两种网络流速下都高于 screen 模型。显然，每个保留特征对模型贡献并非一致，应该根据其信息含量赋予相应的权值，以相互区分。

结束语 本文提出了一种新的入侵检测模型，通过融合主机审计数据与网络数据，提高数据的覆盖面，增加了特征信息量；并基于 SVM 交叉测试，删除其中冗余与无用特征，降低模型运算复杂度；最后基于强制筛选为各保留特征赋予相应的权值，以进一步提高模型的检测精度。新模型能在保证运算复杂度的情况下，有效利用信息量，具有较高的入侵检测精度，在入侵检测领域有较好的应用前景。

参 考 文 献

[1] 顾钧. 基于 KPCA 和 SVM 的网络入侵检测研究[J]. 计算机仿真，2010，27(7)：105-107

[2] 张艳芳，郭郁杰. 入侵检测关键技术研究[J]. 信息与电脑，2010，(8)：16-17

[3] 王树，杜启军，余桂贤，等. 网络入侵检测系统的最优特征选择方法[J]. 计算机工程，2010，36(15)：140-142，144

[4] Lee W K. Feature Selection of Intrusion Data Using a Hybrid-Genetic Algorithm Approach [J]. Wireless Networks，2007，13(6)：459-460

[5] 汪世义，陶亮，王华彬. 几种机器学习方法在 IDS 中的性能比较[J]. 计算机仿真，2010，27(8)：92-94，121

[6] Cortes C，Vapnik V. Support vector networks [J]. Machine Learning，1995，20：273-297

[7] 常卫东，王正华，鄢喜爱. 基于集成神经网络入侵检测系统的研究与实现[J]. 计算机仿真，2007，24(3)：134-137

[8] 肖敏，柴蓉，杨富平，等. 基于可拓集的入侵检测模型[J]. 重庆邮电大学学报：自然科学版，2010，22(3)：345-349

(上接第 88 页)

[11] Boccardi F，Huang H. Limited Downlink Network Coordination in Cellular Networks [C]// Personal，Indoor and Mobile Radio Communications (PIMRC). Athens，Greece，2007：1-5

[12] Papadogiannis A，Gesbert D，Hardouin E. A Dynamic Clustering Approach in Wireless Networks with Multi-cell Cooperative Processing [C]// IEEE International Conference on Communications (ICC). Beijing，China，2008：4033-4037

[13] Liu Jingxin，Wang Dongming. An improved dynamic clustering algorithm for multi-user distributed antenna system [C]// Wireless Communications & Signal Processing (WCSP). China，2009：1-5

[14] Zhou Sheng，Gong Jie，Niu Zhisheng，et al. A Decentralized Framework for Dynamic Downlink Base Station Cooperation [C]// Global Telecommunications Conference (GLOBECOM). China，2009：1-6