基于 AdaBoost 的入侵检测技术探索与分析①

阴国富

(渭南师范学院 数学与信息科学学院, 渭南 714000)

摘 要：阐明了入侵检测系统的监测过程，提出在入侵检测的分析方法中通过 AdaBoost 框架的循环迭代，在每次迭代中，由该算法产生一个带权值的分类器，迭代结束产生多个分类器，最后将这些分类器进行加权联合，得到一个具有较高识别率的分类器，进而克服采用单一分类算法产生的识别率难以满足系统要求的缺陷，从而达到系统对攻击识别率提高，误警率降低的目的，以 KDD99 作为实验样本数据源，仿真实验表明该方法检测预警准确率高。

关键词：样本特征; 分类器; 入侵检测; 权值; 分类决策

Intrusion Detection Technology Based on AdaBoost

YIN Guo-Fu

(Weinan Normal University, Mathematics and the Information Science Institute, Weinan 714000, China)

Abstract：This paper illuminates the intrusion detection system monitoring process and puts forward that in intrusion detection methods of analysis by the iterative AdaBoost framework, in each iteration, the algorithm produces a belt of the weight values classifier, the iterative end into multiple classifier. Finally, the classifier are weighted joint to get a higher rate of classifier, and hence overcome classification algorithm USES a single produced to meet the requirements of the recognition system defect, so as to improve the system to attack rate，reduce false alarm rate of purpose, in KDD99 were selected as the experimental data. The simulation experiments show that the method is accurate in early warning detection.

Key words：sample characteristic; classifier; intrusion detection; weight values; classification decisions

1 引言

入侵检测系统(Intrusion Detection System,IDS)的目的是检测网络上所有成功和未成功的攻击行为，其主要功能有：监测并分析用户和系统的轰动; 核查系统配置和漏洞; 评估系统关键资源和数据文件的完整性; 识别已知的攻击行为; 统计分析异常行为[1-3]。到目前为止，在入侵检测的分析方法中，单一分类算法的识别精度很难达到系统的要求。AadBoost 算法的理论告诉我们，可以通过操作样本集，产生多分类器的方法提升单一分类算法的识别率[4-5]。基于 AadBoost 的入侵分析方法一种是通过循环迭代产生若干个分类器，然后对这些分类器进行加权联合的多分类器入侵检测方法，该方法包含信息获取、特征的预处理、特征提取、及分类器设计、分类决策等步骤。入侵检测系统的检测过程如图 1 所示。

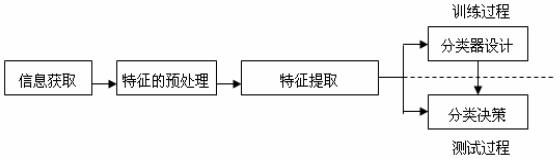

图 1 基于 AdaBoost 的入侵检测

其中，信息获取是通过各种手段获得原始数据; 特征的预处理是在将数据加工成一定表达方式的特征后，需要进行包括特征的数字化、标准化等一系列的

① 基金项目:陕西省自然科学基金(2011JM8020);渭南市自然科学基金(2011KYJ-1)
收稿时间:2011-11-23;收到修改稿时间:2012-03-05

预处理工作; 特征提取阶段是对样本作特征提取，通过特征提取提高系统的识别率与效率; 使用 AdaBoost 算法产生多个分类器，最终形成的分类器就可以用来检测代测试样本。

2 样本预处理

在用 AdaBoost 产生分类器之前，首先要对样本的特征进行一系列预处理工作[6-8]。

（1）样本属性特征的数值化：根据数据类型分为连续型数值、数据流量、离散型数值;

（2）类别的数值化：每个样本的最后一项属性表示它的类别，由于类别是非数值类型，而我们在构造基分类器时，所使用的类别标签必须是数字的，因此，必须对类别进行数值化;

（3）特征的归一化：为了防止特征参数因数量级差别较大而造成那些数量级小的特征难以发挥作用，因此需要对数据进行规一化处理。假设 $D=\{d_i \mid d_i \neq 0, i=1,2,\cdots,n\}$ 为一特征参数数据集合，$\alpha=\min\{d_i\}$，$\beta=\max\{d_i\}$，其中 $i=1,2,\cdots n$，归一化后的数据集合设为 $\overline{D}=\{\overline{d}_i \mid i=1,2,\cdots,n\}$。

$\overline{d}_i$ 采用指数归一化方法确定。

$$\overline{d}_i=\frac{2}{1+\exp(-\gamma d_i)}-1=\frac{1-\exp(-\gamma d_i)}{1+\exp(-\gamma d_i)} \tag{1}$$

其中，γ 为满足式 $\beta\gamma \le k$，k 为常数，一般取 $k=20$。指数中的常数是为保证对大数据区分性，如果不加该参数，对所有的 d_i 大于某一个固定的值时，d_i 都接近于 1，显然对于大数据不易区分。

因为 IDS 样本数据的属性中包含很多零元素，所以适合采用指数归一化的方法，最后的实验结果也验证了采用指数归一化比较适合于 IDS 数据的有效性，该方法能够提高系统的识别率。

3 样本特征提取

文章通过使用主成分分析方法（Principal Component Analysis，PCA）对样本进行特征提取[9-10]。

主成分分析(PCA)是一种把多个原有变量转化为为数不多的若干个线形无关的综合变量的统计方法。它通过寻找一个线形变换 A，使得 Ax 的截断在均方差意义下为最优，定义如下：

设 p 维总体 x 的 p 个随即变量为 $x_1, x_2 \cdots, x_p$，且总体 x 遵从正态分布，即有 $x \approx N(\mu, \sum)$，其中 μ 为总体 x 的均向量，$\sum$ 为总体 x 的协方差矩阵。它们的线性组合最多可以构成 p 个综合变量为：

$$\begin{cases} f_1=a_{11}x_1+a_{12}x_2+\cdots+a_{1p}x_p=a_1{}^{T}x \\ f_2=a_{21}x_1+a_{22}x_2+\cdots+a_{2p}x_p=a_2{}^{T}x \\ \cdots \\ f_p=a_{p1}x_1+a_{p2}x_2+\cdots+a_{pp}x_p=a_p{}^{T}x \end{cases} \tag{2}$$

对每个 $i(i=1,2,\cdots p)$，应满足如下的规范化条件：

$$a_i^{T}a_i=a_{i1}^2+a_{i2}^2+\cdots+a_{ip}^2=1 \tag{3}$$

和互不相关条件，即不同综合变量的协方差为零：

$$\mathrm{cov}(f_j,f_i)=0 \quad j<i, j=1,2,\cdots,p-1 \tag{4}$$

并在满足这两个条件的前提下，方差 $\mathrm{var}(f_i)$ 达到最大，则称综合变量 f_i 为总体 x 的第 i 个主成分。线形变换 A 为：

$$A=\begin{bmatrix} a_1 & a_2 & \cdots & a_p \end{bmatrix}^{T} \tag{5}$$

从以上的主成分定义可以看出：

1) 新的综合变量 f_i 为原有自变量的线形组合，原有自变量的信息是用综合变量的方差来表达，综合变量的方差越大表示包含原有自变量的信息越多;

2) f_1 是 $x_1, x_2, \cdots, x_p$ 的一切线形组合中方差最大的，称为第一主成分;

3) f_2 是 f_1 与不相关的 $x_1, x_2, \cdots, x_p$ 的一切线形组合中方差最大的，称为第二主成分;

4) 主成分分析可以实现数据的压缩，并达到揭示变量之间内在关系和进行统计解释的目的。

4 基分类算法BP神经网络设计

设计一个 BP 神经网络，作为入侵检测系统的基分类算法。首先把 BP 网络视作独立的识别系统，使用 KDD99 作为数据样本构造出训练集、测试集; 然后，用训练集训练 BP 神经网络，再用测试集进行测试，逐步调整、优化网络参数，确定网络各种指标; 最后，将这个合理的 BP 网集成到基于 AadBoost 的入侵分析检测系统中。考虑到训练效果实现的简易性，选择建立三层 BP 神经网结构，如图 2 所示。

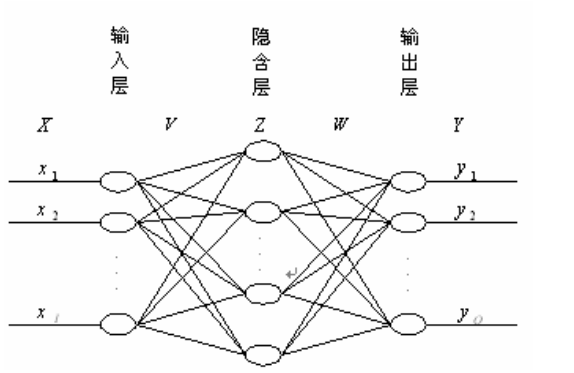

图 2 BP 网络结构图

设计的网络结构由输入层、输出层，和中间隐含层组成。其中，输入层的数目集 X 内的元素个数必须与样本的属性个数相等，输出层数目集 Y 内的元素个数须与样本的类别数相等，该 BP 网络所用的数据集的样本属性为 51 维，经特征提取后降维为 22 维，类别是 10 类，因此这个神经网络的输入层有 22 个处理单元，输出层有 10 个处理单元。

BP 网络通过训练集训练进行学习，用网络的均方根误差 RMS 来定量的反映 BP 网络学习的性能。E_{RMS} 定义为：

$$E_{RMS}=\sqrt{\frac{\sum_{p=1}^{m}\sum_{j=1}^{n}\left(d_{pj}-y_{pj}\right)^2}{m\quad n}} \tag{6}$$

式中 m 是训练集中模式的个数；n 是网络输出层单元的个数。

在 BP 网络学习过程中，按照梯度最速下降算法，均方根误差应是逐渐减小的。一般选择 0.1，这个指标的上限可以根据具体应用灵活设定。本文的 BP 网络选择精度为 0.05，经过测试，可以达到要求。初始权值矩阵是在一定范围内按均匀分布随即产生的。在初始权值下，对于给定输入模式，如果输出层单元的总输入与阀值相差甚远，需要对权值有较大的修正，随着网络的不断学习，不断地修正权值，误差 E_{RMS} 将不断减小，达到要求，停止训练。

5 基于AdaBoost的入侵检测方法

由于 AdaBoost 算法是框架算法，因此它要将其它的各个部分成功的组织起来，并与预处理、测试有机的结合，才能形成本检测系统，并达到高识别率的系统效果。

5.1 系统实现的具体步骤

基于 AdaBoost 的入侵检测系统操作流程如图 3 所示。

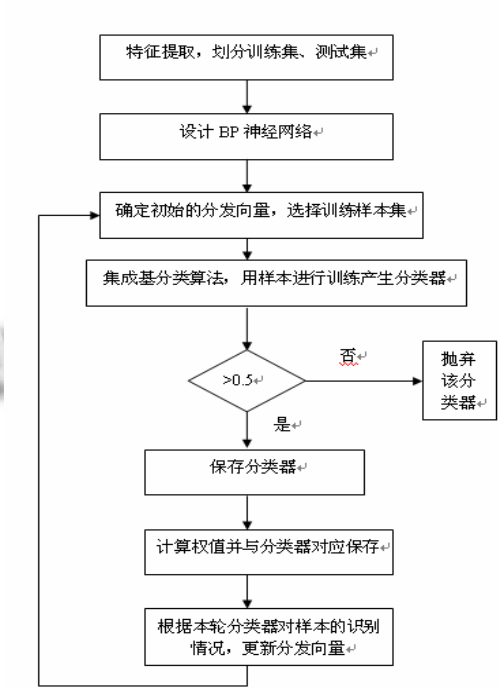

图 3 基于 AdaBoost 的入侵检测系统操作流程

5.2 训练轮数的确定

为了使 4.1 的步骤执行过程能够有效的将各个部分组织起来进行训练，还需要确定一个重要的参数即训练轮数，即进行多少次循环产生多少个分类器就能达到联合错误率最低。通过对 AdaBoost 的理论分析知：1）最终的联合分类器的精度并不随着产生的分类器个数的增加而提高；2）有时甚至会出现过配现象，即若训练轮数过多，随着次数的增加，最终的联合分类器的精度反而会降低，或者分类精度也只是趋于平缓，不会再有大的下降；3）随着训练轮数的增加，训练及测试的时间也都会随之增加，这不符合入侵检测系统的基本要求。因此，选择恰当的训练轮数，对最终的分类效果影响很大。

6 仿真实验

实验操作系统为：windowXP 环境、使用仿真分析软件有：Matlab、VC++和 SqlServer2000。为便于实验，从 KDD99 数据集中选取出 8 个类别约 1400 条样本记录进行训练和测试。

6.1 训练轮数的确定

训练轮数作为一个非常重要的影响因素，选取非常关键， 文章通过试验法确定该参数，借助图形查找到最佳的训练轮数，即联合错误率最低的而训练轮数最少。图 4 是总的错误率随训练轮数的变化曲线图，可以看到在训练轮数为 18 时，错误率最小为 0.0042。

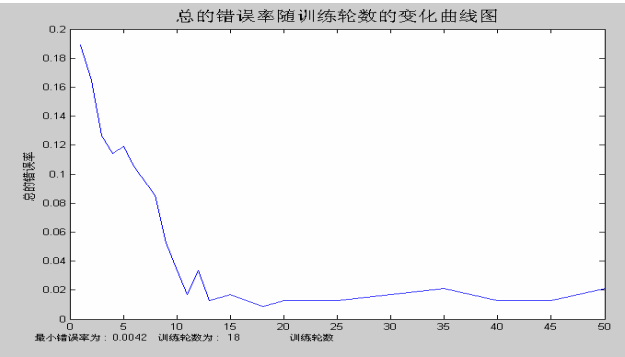

图 4　训练轮数与测试错误率关系图

6.2 训练结果对比与分析

该 BP 网所用的数据集与整个系统所用的数据集相同，首先对数据集作预处理与特征提取，确定了训练论数，将数据集的 80%用于训练，20%用于测试。

1）经过特征提取的样本集的训练结果：

表 1　经过特征提取的样本集的训练参数及结果

内容	值
样本总数：	958
对训练集的测试错误率：	0.0042
分类器个数：	18
训练时间：	4563.6780 秒

表 2　各分类器的错误率与权值对应表

错误率	权值	错误率	权值	错误率	权值
0.0498	2.9521	0.2917	0.8925	0.3593	0.5782
0.0779	2.4648	0.3306	0.7052	0.3694	0.5348
0.1549	1.6967	0.3329	0.7049	0.3963	0.4211
0.1742	1.5560	0.3475	0.6317	0.4004	0.4036
0.1864	1.4736	0.3530	0.6070	0.4188	0.3278
0.2358	1.1756	0.3568	0.5896	0.4383	0.2482

2）　没有经过特征提取的样本集的训练结果：

表 3　未经过特征提取的样本集的训练参数及结果

内容	值
样本总数：	958
对训练集的测试错误率：	0.0112
分类器个数：	18
训练时间：	10258.1310 秒

表 4　各分类器的错误率与权值对应表

错误率	权值	错误率	权值	错误率	权值
0.1354	1.8536	0.2567	1.0634	0.2955	0.8688
0.1410	1.8071	0.2581	1.0560	0.2967	0.8630
0.1677	1.6020	0.2635	1.0278	0.3111	0.7951
0.1813	1.5079	0.2768	0.9603	0.3202	0.7529
0.2080	1.3371	0.2931	0.8802	0.3655	0.5516
0.2501	1.0981	0.2950	0.8712	0.3695	0.5345

从上面的两个结果可以看出，在作了特征提取后，对训练样本集识别错误率有所减小，从 0.0112 减小到 0.0042，训练时间明显缩短。由此可见，通过对样本进行特征提能够有效提高识别率。比较发现 AdaBoost 算法生成的分类器的错误率越小其权值越大。

7　结论

入侵检测已经成为信息安全策略中非常重要的部分，而对于入侵检测系统的分析方法的研究更是该领域的热点所在，本文对 AdaBoost 算法进行了深入探讨，在此基础上，设计并实现了一种新的基于 AdaBoost 算法的多分类器入侵分析方法，该方法对攻击有很强的识别能力，识别率也较高。但依然存在很多问题需要进一步研究，如从理论上研究确定训练轮数与数据集样本的个数、样本的属性维数等因数的关系。

参考文献

1 陈项,安常青,车学农.分布式入侵检测系统及其认知能力.软件学报,2001,12:2.

2 郭红刚,方敏.AdaBoost 方法在入侵检测中的应用.计算机应用,2005,1.

3 马奎斯德萨.模式识别——原理、方法及应用.北京:清华大学出版社,2002.21–46.

4 边肇祺,张学工,等.模式识别.北京:清华大学出版社,2000.250–257.

5 肖立中,刘云翔.适合于入侵检测的分步特征选择算法.计算机工程与应用,2010,(11).

6 易哲,李伟生.基于粗糙集和遗传约简算法的入侵检测方法.计算机工程与应用,2010,(21).

(下转第 31 页)

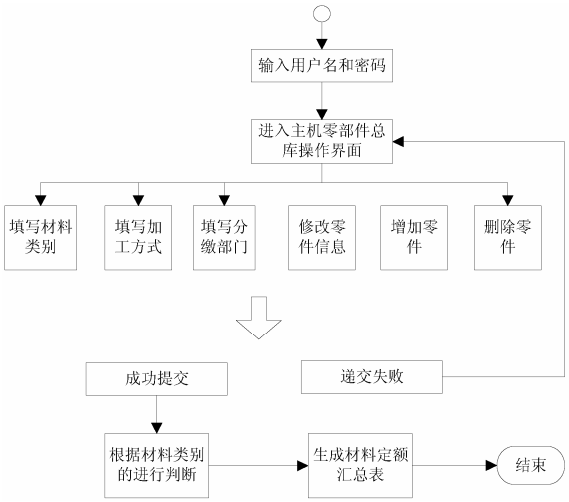

图 9 主机零部件总库的执行控制流程

综合查询模块为所有用户所共享，根据用户权限的不同，系统将自动确定所能执行的操作，用户只需通过选择或输入多种查询条件提交后便可得到所要查找的数据并能通过超链接提供数据相关的操作，如图 10 产品结构 BOM 的检索结果。

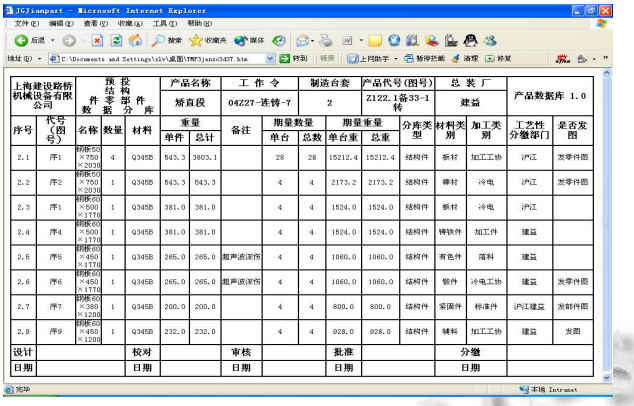

图 10 产品结构数据的检索结果

4 结论

本项目针对某离心机制造企业对产品结构和物料的管理需求，建立了计算机辅助管理的物料管理系统，该系统以计算机网络、数据库系统为技术基础，现代管理理论和方法为指导，覆盖设计产品结构及物料的信息流通的全过程，最大限度地利用企业的人、财、物、设备、技术和信息资源，加快产品数据流通的效率，提高了数据流通的准确程度，同时降低工作人员的劳动强度，最终以提高企业的经济效益和市场竞争能力为目标。

参考文献

1 Henrik J, Peter A, Kjell O. A system for information management in simulation of manufacturing processes. Advances in Engineering Software,2004,35:10–11.

2 陈伟忠.珠江电信 PDM 系统需求分析及实施.北京邮电大学,2008.

3 于晓,仲梁维等,基于产品的 BOM 自动生成方法.精密制造与自动化,2006,(3).

4 贾颖莲.基于.NET 平台的产品数据管理技术的研究与实现,2005.

5 Yeh RT. Software and Database Engineering:Towards a Common Design Methodology.Issue in Data Base Management,1979.

6 张婧.基于产品结构与配置管理的 PDM 系统研究与开发,2006.

7 王保健.ASP.NET 网站建设专家.北京:清华大学出版社,2005.

8 章立民.ASP.NET 开发实践范例宝典(使用 C#).北京:科学出版社,2010.

9 解本巨,李宗颜,宫生文.LINQ 从基础到项目实践,2010.2.

10 祖晓东.基于 ASP.NET2.0 实现 WEB 打印方法的探讨.

11 董培征,杨学良.ActiveX 技术在 Web 应用中实现本地端打印.微计算机应用,2001,22(4):202–204.

(上接第 72 页)

7 Schwenk H, Bengio Y. Boosting Neural Neural Networks. Neural Computation,2000,12(8):1869–1887.

8 陈军,徐蕾.用一种改进的蚁群聚类算法进行网络入侵检测.沈阳航空工业学院学报,2010,(1).

9 Giacinto G, Roli F. Intrusion Detection in Computer Networks by Multiple Classifier Systems. Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02),1051-4651/02 IEEE.2002.

10 沈学利,张纪锁.基于 BP 网络与改进的 PSO 算法的入侵检测研究.计算机工程与科学,2010,(6).