第 34 卷 第 5 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.5 Computer Engineering March 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)05—0225—03 文献标识码：A 中图分类号：TP309.2

基于 Boosting 模糊分类的入侵检测

唐晓衡，夏利民
(中南大学信息科学与工程学院，长沙 410075)

摘 要：提出一种基于 Boosting 模糊分类的入侵检测方法。采用遗传算法来获取入侵检测的模糊规则，利用 Boosting 算法不断改变训练样本的分布，使每次遗传算法产生的模糊分类规则重点考虑误分类和无法分类的样本。以 kddcup'99 为数据源进行了仿真实验，结果表明该方法具有良好的分类识别性能。

关键词：模糊分类；遗传算法；Boosting 算法；入侵检测

Intrusion Detection Based on Boosting Fuzzy Classification

TANG Xiao-heng, XIA Li-min
(Department of Information Technology, Central South University, Changsha 410075)

【Abstract】This paper proposes a method for intrusion detection based on Boosting fuzzy classification. Fuzzy rules involved in intrusion detection are obtained by genetic algorithm, and Boosting algorithm is employed to change the distribution of training instances during each round of training, so that new fuzzy classification rules extracted by genetic algorithm will put more emphasis upon the instances misclassified or uncovered. Simulation experiments with the data set kddcup'99 show that the method has good recognition performance.

【Key words】fuzzy classification; genetic algorithm; Boosting algorithm; intrusion detection

1 概述

计算机系统的互联提高了信息资源共享空间的广度和深度，但计算机信息有共享和易于扩散等特性，它在处理、存储、传输和使用上有着严重的脆弱性，因此，安全已经成为计算机网络系统的关键问题之一。

入侵检测系统作为一种可以放置在受保护网络内部的原始过滤器，能在不影响性能的情况下对网络进行监测，提供对内、外部攻击和误操作的实时保护[1]。

入侵检测由Anderson于 1980 年首次提出，他将入侵行为划分为外部闯入、内部授权用户的越权使用和滥用 3 种类型，并提出用审计方式追踪监视入侵产生的威胁[2]。但是入侵行为的判决因素往往比较复杂，并且涉及到系统运行的多个方面。比如在检测DoS攻击时，判决入侵行为所依赖的因素包括CPU利用率、内存利用率和网络流量等；即使是判决一个网络访问是否为攻击，也要考虑各种因素以进行综合评判。但以往许多入侵系统(如Snort, IDES)往往只针对少数几个特征进行匹配过滤，这直接导致了误报率过高[3]。

此外，通过数据挖掘和机器学习等方法建立的检测模型也存在误报率较高的缺点，因为它们强行将检测对象行为分为正常和入侵两类，或者将所有的异常行为都视作入侵行为[4]。异常实际上是一个模糊概念，主要反映在入侵特征模糊、特征匹配模糊以及检测标准模糊。

采用模糊理论能很好地解决上述问题。将待检测因素集的各因子分别进行模糊化处理，取得各自相应的隶属度函数，然后运用模糊规则集来对网络行为进行综合判决，从而更有效、更客观地判断一个网络行为是否为入侵。

针对以上入侵检测中存在的一些问题，并为了进一步提高模糊分类器的分类能力，本文首先利用遗传算法的空间随机寻优能力从初始种群中找出适应度最好的模糊规则；然后利用 Boosting 算法依据当前模糊规则的判定能力改变训练样本的权值，再次调用遗传算法，以期望下一轮产生的模糊规则能判定那些当前误判的训练样本。以上过程反复迭代运行，直到产生出预先设置的模糊规则数量为止。最后使用加权投票方式来对所有模糊规则进行集成。整个方法的体系结构如图 1 所示。

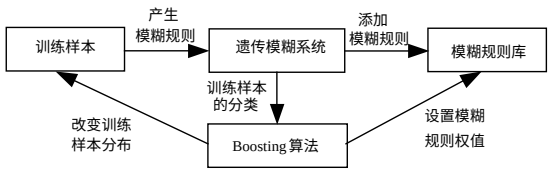

图 1 本文系统的体系结构

2 模糊分类方法

在入侵检测中，网络上各种攻击相互交叉、混合，其界限无法清晰划分，因此，基于模糊规则的分类系统比一般系统更适合解决入侵检测这种难以明确界定各种攻击行为的问题。设 $x=\{x_1,x_2,\cdots,x_n\}$ 为入侵检测中的一个样本，x_n为该样本的第n维特征，并以其作为模糊分类器的输入变量，则模糊规则的一般形式为

$$R_j: \text{if } x_1 \text{ is } A_{1j} \text{ and } x_2 \text{ is } A_{2j} \text{ and}\cdots\text{and } x_n \text{ is } A_{nj} \text{ then } y=c_j[w_j] \quad (1)$$

其中，R_j表示第j条模糊规则；A_{nj}指与x_n相联系的模糊集；y为判决属性；$c_j \in \{c_1,c_2,\cdots,c_m\}$ 代表判决属性所属类别；w_j代

基金项目：湖南省自然科学基金资助项目(05JJ30121)
作者简介：唐晓衡(1974－)，男，硕士研究生，主研方向：模式识别和机器学习；夏利民，教授
收稿日期：2007-03-15 **E-mail：**xiaohengtang@163.com

表该条规则的置信度。

入侵检测中的一个模糊规则示例如下：

If dst_bytes is Slightly Low and duration is Middle then intrusion_type is DoS [0.6]

它表示当被检样本的 dst_bytes 属性值隶属 Slightly Low 模糊集，并且 duration 属性值隶属 Middle 模糊集时，有 0.6 的可信度证明此样本属于一个 DoS 攻击。

对于任意给定的一个输入样本 x，模糊规则 R_j，$j\in\{1,2,\cdots,m\}$触发后，该规则匹配样本 x 的程度为

$$\mu_{R_j}(x)=\mu_{R_j}(\{x_1,x_2,\cdots,x_n\})=\min_{i=1}^{n}\mu_{A_{ij}}(x_i) \tag{2}$$

其中，$\mu_{A_{ij}}(x_i)$为样本x的第i维特征相对模糊子集A_{ij}的隶属度。根据规则自身的置信度，该规则相对样本x的真值为

$$R_j(x)=w_j\mu_{R_j}(x) \tag{3}$$

这样，对于输入的一个样本，通过式(3)的计算会得到每个规则的相应真值和由该规则所指定的样本分类结果。把具有分类结果 $c_j=c_k, k\in\{1,2,\cdots,m\}$ 的所有规则真值进行相加求和计算，取和最大的分类结果作为该样本的所属类别输出，计算式为

$$C_{\max}(x^k)=\operatorname{argmax}_{C_k}\sum_{R_j/c_j=c_k}R_j(x) \tag{4}$$

3 模糊规则的获取

遗传算法是基于自然选择和种群基因的一种并行随机搜索算法。这种方法不要求对问题结构有深入的了解，只需要知道目标函数即可；由于是多点搜索，因此能减小陷入局部最优值的可能性；提高了系统的并行性。本文通过遗传算法来训练模糊规则，目的在于能够最大限度地归纳模糊规则，使其正确判定尽可能多的样本，同时错误判定的样本尽可能的少。

3.1 模糊规则的编码

对应于式(1)的一条模糊规则，规则后件和权值是固定的，可变部分主要反映在规则前件中模糊子集的定义上。本文中，每条规则均使用自身定义的模糊子集，并且取每个模糊子集的隶属度函数为高斯函数，即

$$\mu_{A_{ij}}=\exp\left[-\frac{1}{2}\left(\frac{x_i-\bar{x}_i^j}{\sigma_i^j}\right)^2\right] \tag{5}$$

根据式(5)，模糊子集A_{ij}可以由 $\bar{x}_i^j$ 和 σ_i^j 确定。将第j条规则中每个模糊子集的这 2 项数值连接起来，形成如$(\bar{x}_1^j,\sigma_1^j),(\bar{x}_2^j,\sigma_2^j),\cdots,(\bar{x}_n^j,\sigma_n^j)$的一个实数向量，构成第$j$条模糊规则的编码，由此形成遗传算法中的一个个体。

3.2 适应度函数的确定

根据文献[5]，适应度计算应该建立在规则能正确判定的样本数量上。除此之外，第 4 节描述的 Boosting 算法按照样本分类难度给样本赋予了权值，在计算规则适应度时应同时考虑样本的权值，在计算中还要体现该规则误判的样本数量。由此可得适应度函数计算式为

$$f_{j1}=\frac{\sum_{k|c_k=c_j}w^k\mu_{R_j}(x^k)}{\sum_{k|c_k=c_j}w^k} \tag{6}$$

$$f_{j2}=\frac{\sum_{k|c_k\neq c_j}w^k\mu_{R_j}(x^k)}{\sum_{k}w^k\mu_{R_j}(x^k)} \tag{7}$$

$$Fitness_{R_j}=f_{j1}\cdot(1-f_{j2}) \tag{8}$$

其中，x^k为第k个样本；w^k为Boosting算法赋予x^k的权值；c_j为x^k所属类别。

3.3 模糊规则的获取

本文的遗传算法步骤如下：

(1)按 3.1 节编码形式随机产生 N 组个体构成初始种群。

(2)计算出每个个体的适应度值，并按适应度值从大到小进行排序。选择出 10%的优秀个体直接进入下一代，剩下部分按轮盘赌的比例选择法进行选择。

(3)从群体中随机选择 2 个个体，按式(9)、式(10)进行交叉产生新的个体：

$$x_k^i=\alpha x_{k-1}^i+(1-\alpha)x_{k-1}^j \tag{9}$$

$$x_k^j=(1-\alpha)x_{k-1}^i+\alpha x_{k-1}^j \tag{10}$$

其中，α 为一给定常数；x_{k-1}^i和 x_{k-1}^j为第 k-1 代选择出来进行交叉计算的父个体；x_k^i 和 x_k^j 为产生的新一代个体。

(4)从群体中随机选择一个个体，按式(11)进行变异产生新的个体：

$$x_k^i=x_{k-1}^i+\bar{\eta},\ \boldsymbol{\eta}\sim N(0,\boldsymbol{\sigma}) \tag{11}$$

其中，x_{k-1}^i为概率选出的父个体；x_k^i 为新一代个体；$\boldsymbol{\eta}$ 是正态分布随机向量。

(5)计算新个体的适应度，淘汰最差的 10%个体。

(6)转第(2)步重复进行，直到完成规定的进化次数。

4 Boosting 模糊分类

Boosting 模糊分类的基本思想是给定一个训练集$(x^1,c_1),(x^2,c_2),\cdots,(x^N,c_N)$，其中，$c_N\in\{c_1,c_2,\cdots,c_m\}$。首先给每一个训练样本赋以相等的权值，然后用遗传算法对训练集样本进行多轮训练，每轮训练得到一条模糊规则，并对训练失败的训练样本赋以较大的权值，以便在后续的学习中集中对这些比较难的训练样本进行学习，得到下一条新规则，计算每条规则对应的权值，分类效果好的规则对应的权值较大，反之较小。最终的分类采用加权投票方式对新样本进行分类判别[6]。

基于 Boosting 方法的模糊分类规则集产生方法如下：

(1)给定训练样本集

$$S=\{(x^1,c_1),(x^2,c_2),\cdots,(x^N,c_N)\},\ c_N\in\{c_1,c_2,\cdots,c_m\}$$

赋予每个样本相等的初始权值 $\omega^i(1)=1/N$。

(2)For t=1, 2,⋯, L Do　　//L 代表要产生的模糊规则数，//通常为经验值

1)利用遗传算法找出一条使适应度$Fitness$最大的模糊规则R_t；

2)在当前样本的分布下，计算模糊规则R_t的分类错误率E_t及规则对应的权值 α_t ：

$$E(R_t)=\frac{\sum_{i|C_i\neq C_t}\omega^i\mu_{R_t}(x^i)}{\sum_i\omega^i\mu_{R_t}(x^i)} \tag{12}$$

$$\alpha_t=\frac{1}{2}\ln\left(\frac{1-E(R_t)}{E(R_t)}\right) \tag{13}$$

3)根据错误率更新样本的权值：

$$\omega^i(t+1)=\frac{\omega^i(t)}{z_t}\begin{cases}e^{\alpha_t\mu_{R_t}(x^i)} & c_i\neq c_t\\ e^{-\alpha_t\mu_{R_t}(x^i)} & c_i=c_t\end{cases} \tag{14}$$

其中，z_t 是归一化因子。

(3)输出具有 L 条模糊规则的分类器。对于未知样本$x^k=\{x_1^k,x_2^k,\cdots,x_n^k\}$，由上述模糊分类器得到其类别：

$$C_{\max}(x^k)=\arg\max_{C_m}\sum_t\alpha_t\sum_{R_t/c_t=C_k}\mu_{R_t}(x^k) \tag{15}$$

由于采用了 Boosting 方法，每条模糊规则都着重对那些当前规则集不能正确分类的训练样本进行学习，使得到的新规则有很好的互补性、有利于分类。Boosting 方法采用了加权投票方式分类判别，体现了不同规则在分类中不同的作用，使其分类的正确率高于一般的模糊分类(如式(4))。

5 仿真实验与结果

实验数据来自kddcup'99[7]入侵检测数据集。该数据集提供了约 490 万条数据，每条数据具有 42 个属性，其中，36 个属性为数值型；5 个为字符型；1 个为类别标签。其中包含 4 大类 22 种攻击，4 大类分别为：DoS攻击，如neptune, smurf；U2R攻击，如buff_overflow, perl；R2L攻击，如ftp_write, guess_passwd；PRB攻击，如portsweep, nmap。整个数据集很大，通常使用一个 10%的子集(包含 494 020 条记录，其中，normal占 19.69%；attack占 80.31%)来对分类器进行训练和测试。为了使攻击多样化，将出现频率低的攻击记录全部选取，而出现频率高的smurf, nepturn等攻击记录及正常记录各选取 10%，再将数据随机 10 等分，得到 10 个子集P1~P10，采取 10 重交叉验证方式进行测试，取均值作为最后的实验 结果。

对于离散形式的属性(如 logged_in 等)，其对应模糊子集视为经典集合，不参与遗传操作。而对于取值为连续形式的属性，通过归一化操作使其值在 0~1 之间。遗传算法的参数设定为：群体大小为 500；交叉率为 0.20；变异率为 0.01；进化代数为 40。每次进化最好的 50 个个体直接进入下一代，其他通过交叉和变异操作产生，适应度最差的 10%被淘汰。

实验 1 比较一般模糊分类器与本文Boosting模糊分类器的检测率和误报率，Boosting 训练次数取 25。结果如表 1 所示。

表 1 2 种模糊分类器的比较 (%)

	平均检测率	平均误报率
一般模糊分类器	95.82	10.32
Boosting 模糊分类器	97.67	6.86

在引入 Boosting 算法之后，平均检测率有一定提升，由 95.82%上升到 97.67%，提升率为 1.93%，误报率则有较大下降，由 10.32%下降到 6.86%，下降率为 33.53%。原因是采用了 Boosting 算法后，产生的模糊规则能更大程度地覆盖样本集，加权投票策略能更加正确地判定样本。

实验 2 Boosting 训练次数(即模糊规则数)对检测率和误报率的影响。结果如表 2 所示。

表 2 不同训练次数对分类结果的影响 (%)

Boosting 训练次数	平均检测率	平均误报率
15	89.62	10.82
20	94.43	8.75
25	97.67	6.86
30	98.22	6.23

由表 2 可见，Boosting 训练次数越多，分类器的性能越好，但当训练次数从 20 增加到 30，提升的比率逐渐减小，因此，在实际应用中需要在训练时间和性能之间作权衡。

6 结束语

入侵检测作为信息安全的重要组成部分，对防御网络攻击有着重要意义。本文提出一种新的基于 Boosting 模糊分类的入侵检测方法。实验表明，通过该方法构造的模糊分类器在入侵检测中具有更高的检测率和较低的误报率，具有较大的应用价值。

参考文献

[1] 陈爱斌, 夏利民. 基于Boosting算法的入侵检测[J]. 计算机工程, 2004, 30(11): 98-100.

[2] Anderson J P. Computer Security Threat Monitoring and Surveillance[R]. James P. Anderson Company, 1980-04.

[3] Han Sangjun, Cho Sungbae. Detecting Intrusion with Rule-based Integration of Multiple Models[J]. Computers & Security, 2003, 22(7): 613-623.

[4] Chen Wangbin, Wang Lisheng, Liao Genwei. Study on Intrusion Detection Technology Based on Sequence Pattern Mining[J]. Mini-Micro Systems, 2004, 25(5): 878-881.

[5] González A, Pérez R. Completeness and Consistency Conditions for Learning Fuzzy Rules[J]. Fuzzy Sets and Systems, 1998, 96(1): 37-51.

[6] 夏利民, 戴汝为. 基于Boosting模糊分类的滚动轴承故障诊断[J]. 模式识别与人工智能, 2003, 16(3): 323-327.

[7] KDD99. KDD99 Cup Dataset[Z]. [2006-10-10]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

（上接第 224 页）

物体灰度直方图跟人体相似，则也有可能被当作人体来跟踪统计，今后可以考虑结合图像中其他人体特征来进行进一步精确定位以解决这一问题，在遮挡十分严重的运动人体分割中可以利用多摄像头采集不同角度的图像来进行处理。

参考文献

[1] Haritaoglu I, Harwood D, Davis L. W4: Real-time Surveillance of People and Their Activities[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(8): 809-830.

[2] Haritaoglu I, Flickner M. Detection and Tracking of Shopping Groups in Stores[C]//Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, HI, USA: IEEE Computer Society, 2001.

[3] Lu Wenmiao, Tan Yap-peng. A Color Histogram Based People Tracking System[C]//Proc. of IEEE International Symposium on Circuits and Systems. Sydney, Australia: [s. n.], 2001.

[4] 傅永会, 张凤超, 张宪民. 一种改进的基于颜色直方图的实时目标跟踪算法[J]. 数据采集与处理, 2001, 16(3): 309-314.