第 37 卷 第 21 期
Vol.37 No.21

计 算 机 工 程
Computer Engineering

2011 年 11 月
November 2011

• 安全技术 •

文章编号：1000—3428(2011)21—0108—03 文献标识码：A 中图分类号：TP183

基于 PCA 的 3 种改进 BP 算法性能研究

李志清[1]，傅秀芬[2]

(1. 广州行政学院信息网络中心，广州 510070；2. 广东工业大学计算机学院，广州 510075)

摘 要：现有入侵检测系统的效率和准确率较低。为此，提出一种基于主成分分析的特征提取方法。对数据源进行特征降维，将获得的主成分作为 BP 神经网络的输入数据进行识别。分析原始 BP 算法存在的问题，研究 RPBP、CGBP、LMBP 3 种改进 BP 算法，并进行仿真实验，结果表明，与原始 BP 算法相比，改进算法收敛速度快，漏报率和误报率低，能有效改善入侵检测的识别效果。

关键词：入侵检测；主成分分析；神经网络；BP 算法；误报率；漏报率

Performance Study on Three Kinds of Improved BP Algorithm Based on Principal Component Analysis

LI Zhi-qing[1], FU Xiu-fen[2]

(1. Information Network Center, Guangzhou Administration Institute, Guangzhou 510070, China;
2. School of Computer, Guangdong University of Technology, Guangzhou 510075, China)

【Abstract】A feature extraction method using Principal Component Analysis(PCA) is proposed to improve the efficiency and accuracy of intrusion detection. This method reduces data dimensions and views some principal components as the inputs of BP neural network to finish data recognition. In order to overcome the problems in standard BP algorithm, three kinds of improved BP algorithm are studies and simulated, experimental results show that compared with standard BP algorithm, RPBP, CGBP and LMBP algorithm have well convergent speed and low false positive rate and false negative rate, they can improve recognition effect of three kinds of improved BP algorithm.

【Key words】intrusion detection; Principal Component Analysis(PCA); neural network; BP algorithm; false positive rate; false negative rate

DOI: 10.3969/j.issn.1000-3428.2011.21.037

1 概述

随着网络技术的迅速发展，网络入侵技术日益表现出不确定性、复杂性和多样性等特点，而入侵检测技术的发展趋势就是在能有效检测出已知入侵行为的同时，对未知攻击也有检测能力，从而能够应付多变的入侵手段[1]。采用人工智能的方法提高检测的正确率，是入侵检测的研究方向之一。目前最常用的方法是将入侵检测与神经网络算法相结合，利用神经网络的自适应和自学习能力来提高入侵检测系统的性能[2-3]。本文把 3 种改进的 BP 神经网络算法应用到经主成分分析降维的网络数据，检验 3 种改进算法对攻击数据的识别效果。

2 主成分分析法

主成分分析法(Principal Component Analysis, PCA)，又称为主元分析，是一种通过降维来简化数据结构的方法[4]，主要用于多变量问题中。提取出来的每个主成分是原来多个变量的线性组合。用数学语言可简述如下：

设 $\boldsymbol{X}^{\mathrm{T}}=\left(X_1,X_2,\cdots,X_p\right)$ 为 P 维随机向量，每个随机变量 $X_i\left(i=1,2,\cdots,p\right)$ 有 n 个样本，它的主成分为：

$$\begin{aligned}&F_1=e_1^{\mathrm{T}}X=e_{11}X_1+e_{21}X_2+\cdots+e_{p1}X_p\\&\vdots\\&F_p=e_p^{\mathrm{T}}X=e_{1p}X_1+e_{2p}X_2+\cdots+e_{pp}X_p\end{aligned}\tag{1}$$

主成分的方差决定它反映总体信息的多少，第一主成分 F_1 在所有线性组合中方差最大，后面依次是 $F_2,F_3,\cdots,F_p$。值得注意的是，各主成分之间的相关系数为 0。

在进行主成分分析前，通常需要计算原始随机变量的相关矩阵。设样本数据矩阵为 $\boldsymbol{M}=\left(\alpha_1,\alpha_2,\cdots,\alpha_p\right)^{\mathrm{T}}=\left(\beta_1,\beta_2,\cdots,\beta_p\right)$，其中，$\beta_j\left(j=1,2,\cdots,n\right)$ 对应于每个随机向量 $\boldsymbol{X}$ 的样本向量；$\alpha_i\left(i=1,2,\cdots,p\right)$ 对应于每个随机变量的所有样本值构成的向量，则它的样本方差-协方差矩阵 $\boldsymbol{S}$ 和相关矩阵 $\boldsymbol{R}$ 分别为：

$$\boldsymbol{S}=\left(S_{ab}\right)=\left(1/n\right)\sum_{k=1}^{n}\left[\left(M_{ak}-\left(1/n\right)\sum_{c=1}^{n}M_{ac}\right)\left(M_{bk}-\left(1/n\right)\sum_{d=1}^{n}M_{bd}\right)\right]\tag{2}$$

$$\boldsymbol{R}=\left(R_{ab}\right)=S_{ab}/\left(\sqrt{S_{aa}}\sqrt{S_{bb}}\right)\tag{3}$$

为使主成分分析能均等地对待每一个原始随机变量，消除由于单位的不同而可能带来的不利影响，须将各原始随机变量作标准化处理，得到 $X_i^*\left(i=1,2,\cdots,p\right)$。$X^*$ 的样本方差-协方差矩阵 $\boldsymbol{S}$ 即为相关矩阵 $\boldsymbol{R}$。

由相关矩阵 $\boldsymbol{R}$ 求出它的特征值 $\lambda_t\left(i=1,2,\cdots,p\right)$，根据特征值的大小可确定主成分的次序并计算出各主成分的方差贡献率。此时，按顺序取前 $m\left(m\leqslant p\right)$ 个主成分，使其累计贡献率达到了一个较高的百分数(本文中为 90%)。接着逐一计算出选定主成分 $F_t\left(1\leqslant t\leqslant m\right)$ 和标准化随机变量 X_i^* 之间的相关系数 $\left(\rho Y_t,X_i^*\right)$(即因子负荷量)，并由 $\left(\rho Y_t,X_i^*\right)=\sqrt{\lambda_i}\cdot e_n$

基金项目：广东省自然科学基金资助项目(07001802)

作者简介：李志清(1981－)，男，讲师，主研方向：网络与信息安全，人工智能；傅秀芬，教授

收稿日期：2011-04-20 **E-mail：**zhiqinggz@gmail.com

计算出 e_n，从而得到 F_t 与原始随机向量 $\boldsymbol{X}$ 中 p 个变量的线性组合关系。事实上，主成分 F_t 的系数向量为第 t 个特征值 λ_t 所对应的正交化特征向量 $\boldsymbol{e}_t$。

3　BP 算法及其分析

BP 模型具有学习、联想和容错功能，并能进行大规模并行信息处理，对非线性系统具有很强的模拟能力，成为目前应用最为广泛的人工神经网络算法。BP 神经网络是采用误差逆向传播学习算法的多层前馈神经网络[5]，该网络模型有输入层、一个或多个隐层和输出层，层间多为全互连方式，同层单元之间不存在相互连接。

标准 BP 使用了优化中的梯度下降算法，对于复杂的非线性模型仿真从理论上来说其误差可以达到任意小的程度。但它仍然存在一些缺陷：

(1)BP 算法的收敛速度慢，常需要较多次数的迭代，而且随着训练样本维数的增加，网络性能会迅速下降。

(2)网络中隐层节点个数的选取尚无理论上的指导。

(3)从数学角度看，BP 算法是一种梯度最速下降法，这就可能出现局部极小问题。当出现局部极小时，表面看误差符合要求，但所得到的解并不一定是问题的最优解。因此，标准 BP 算法尚不完备。

4　3 种改进的 BP 算法

4.1　弹性 BP 算法(RPBP)

BP 网络的隐层采用 S 形传输函数。随着训练的进行，会出现梯度的幅度非常小，权值和阈值的修正量很小，导致训练的时间变得很长。

弹性 BP 算法的目的是消除梯度幅度的不利影响，所以在进行权值的修正时，仅用到偏导的符号，而其幅值却不影响权值的修正，权值大小的改变取决于与幅值无关的修正值。当连续 2 次迭代的梯度方向相同时，可将权值和阈值的修正值乘以一个增量因子，使其修正值增加；当连续 2 次迭代的梯度方向相反时，可将权值和阈值的修正值乘以一个减量因子，使其修正值减小；当梯度为 0 时，权值和阈值的修正值保持不变；当权值的修正发生振荡时，其修正值将会减小。如果权值在相同的梯度上连续被修正，则其幅度必将增加，从而克服了梯度幅度偏导的不利影响，即：

$$\Delta x(k+1)=\begin{cases}\Delta x(k)\cdot k_{inc}\cdot sign\big(g(k)\big)\\ \Delta x(k)\cdot k_{dec}\cdot sign\big(g(k)\big)\\ \Delta x(k)\end{cases} \tag{4}$$

其中，$g(k)$ 为第 k 次迭代的梯度；$\Delta x(k)$ 为权值或阈值第 k 次迭代的幅度修正值；k_{inc} 为增量因子；k_{dec} 为减量因子。

4.2　变梯度算法(CGBP)

最速下降 BP 算法是沿着梯度最陡下降方向修正权值的，虽然误差函数沿着梯度的最陡下降方向进行修正，误差减小的速度是最快的，但收敛的速度不一定是最快的。在变梯度算法中，沿着变化的方向进行搜索，使其收敛速度比最陡下降梯度方向的收敛速度更快。

变梯度算法的第 1 次迭代是沿着最陡梯度下降方向开始进行搜索的：

$$p(0)=-g(0) \tag{5}$$

然后，决定最佳距离的线性搜索沿着当前搜索方向进行：

$$\begin{cases}x(k+1)=x(k)+ap(k)\\ p(k)=-\boldsymbol{g}(k)+\beta(k)p(k-1)\end{cases} \tag{6}$$

其中，$p(k-1)$为第 $k+1$ 次迭代的搜索方向，从式(6)可以看出，它由第 k 次迭代的梯度和搜索方向共同决定。系数 $\beta(k)$ 按照 Fletcher-Reeves 修正算法定义为：

$$\beta(k)=\frac{\boldsymbol{g}^{\mathrm{T}}(k)\boldsymbol{g}(k)}{\boldsymbol{g}^{\mathrm{T}}(k-1)\boldsymbol{g}(k-1)} \tag{7}$$

4.3　LMBP 算法

LMBP 算法是为了在以近似二阶训练速率进行修正时避免计算 Hessian 矩阵而设计的[6]。当误差性能函数具有平方和误差的形式时，Hessian 矩阵可以近似表示为：

$$\boldsymbol{H}=\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J} \tag{8}$$

梯度的计算表达式为：

$$\boldsymbol{g}=\boldsymbol{J}^{\mathrm{T}}\boldsymbol{e} \tag{9}$$

其中，$\boldsymbol{H}$ 是包含网络误差函数对权值和阈值一阶导数的雅克比矩阵；$\boldsymbol{e}$ 是网络的误差向量。

LM 算法用上述近似 Hessian 矩阵按式(10)进行修正：

$$x(k+1)=x(k)-\left[\boldsymbol{J}^{\mathrm{T}}\boldsymbol{J}+\mu I\right]^{-1}\boldsymbol{J}^{\mathrm{T}}\boldsymbol{e} \tag{10}$$

当系数 μ 为 0 时，即为牛顿法；当系数 μ 的值很大时，式(10)变为步长较小的梯度下降法。牛顿法逼近最小误差的速度更快、更精确，因此应尽可能使算法接近于牛顿法，在每一步成功的迭代后(误差性能减小)，使 μ 减小；仅在进行尝试性迭代后的误差性能增加的情况下，才使 μ 增加。这样，该算法每一步迭代的误差性能总是减小的。

5　仿真实验与结果

5.1　实验数据集

实验中采用入侵检测领域比较权威的 KDD CUP 1999 数据集[7]。数据集中包含正常数据与多种异常数据，涉及 4 大类 22 小类入侵攻击行为：DoS(拒绝服务攻击)，R2L(远程非授权访问攻击)，U2R(非授权得到超级用户权限攻击)，PROBE(漏洞探测与扫描攻击)。每条数据包含 41 个特征属性(维数)，其中，34 个为连续变量，7 个为符号变量。

5.2　数据预处理

KDD CUP 1999 数据集提供的 kddcup. data. corrected 文件中共包含 4 898 431 条原始数据，运用 SQL2000 数据库删除重复记录，得到 1 074 992 条有效数据，其分布为：Normal (812, 814)，DoS(247, 267)，R2L(999)，U2R(52)，PROBE(13, 860)。

由于每条数据的 2、3、4 维均为非数值形式，Matlab 无法识别，必须进行数值化。统计各维中出现的内容，按字母排序，并以序号代替原内容。运用 SQL2000 数据库，将数值化后的数据按照攻击方式分类，保存在不同文档中，并仅保留 41 位数值部分。

对某种攻击类型进行识别前，按 10:1 的比例从 Normal 和该类攻击中随机选取数据，并按 5:1 的比例分别保存在 PTrain.txt 和 PTest.txt 中，同时根据 PTrain.txt 中的数据条数构造 TTrain.txt(由若干 0 和 1 组成，0 对应正常数据，1 对应入侵数据)，供后续实验调用。

5.3　仿真阶段

根据实验要求，本文编写了 Matlab 程序，并保存于 PCATEST.m 文件中。由于对 4 个神经网络算法(包括原始 BP 算法和 3 个 BP 改进算法)进行仿真实验，本文创建 *newff* 函数，即每次调用时只需更改神经网络创建语句 *newff* 中的训练函数类型，即可实现不同的神经网络分类器功能。

$$net=newff\left(\min\max(P0),[B,1],\{'\tan sig','\log sig'\},'BP'\right)$$

神经网络误差性能目标值为 0.01，训练的最大步长为 2 000 步，学习率为 0.05，最小梯度值为 1×10^{-30}。神经网络输出结果以 0.5 为判定阈值，小于 0.5 为正常数据，大于 0.5 为攻击数据。

5.4 实验结果及分析

设定公式如下：

$$\text{误报率}=\frac{\text{被错误地判断为攻击的正常数据数}}{\text{测试样本中的正常数据总数}}\times100\%$$

$$\text{漏报率}=\frac{\text{被错误地判断为正常的攻击数据数}}{\text{测试样本中的攻击数据总数}}\times100\%$$

在仿真过程中，将训练精度设定为 0.001，检验 4 种神经网络算法对 4 种攻击(smurf、neptune、teardrop、nmap)的检测效果。限于篇幅，仅给出 smurf 攻击的检测结果和训练误差性能曲线。

对于 smurf 攻击，检测效果如表 1 所示，4 种神经网络算法对 smurf 攻击训练误差性能曲线，如图 1~图 4 所示，其中，训练精度设定为 0.001。

表 1 smurf 检测效果

算法	误报率/(%)	漏报率/(%)	收敛步数
BP	16	13.32	163
RPBP	6	11.76	5
CGBP	10	10.16	17
LMBP	8	10.76	7

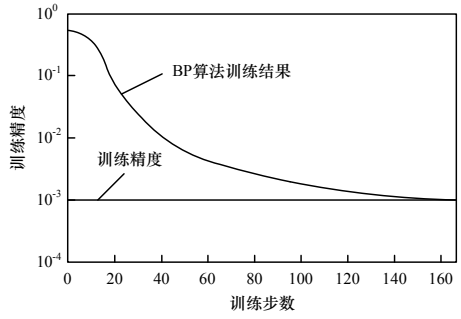

图 1 BP 算法训练误差性能曲线

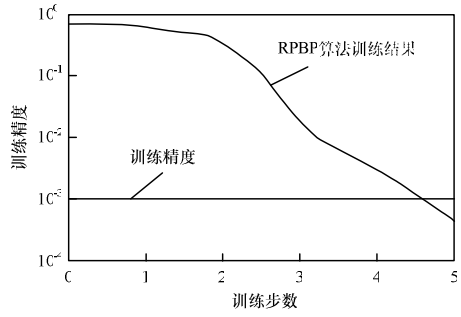

图 2 RPBP 算法训练误差性能曲线

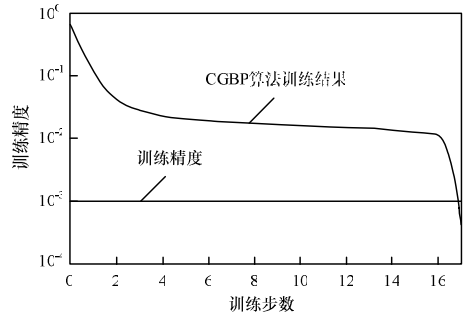

图 3 CGBP 算法训练误差性能曲线

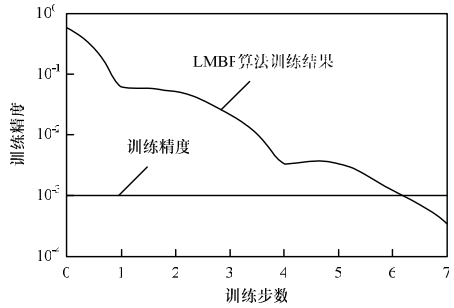

图 4 LMBP 算法训练误差性能曲线

对实验结果进行分析，能得到如下结论：

(1)实验数据显示，原始 BP 算法收敛相对较慢，发生了未收敛的情况。3 种改进算法的收敛速度与之相比都有明显提高，可见改进算法具有更好的实时性。

(2)原始 BP 算法的误报率较高，基本都在 15%以上，若不使用 PCA 技术一般都在 20%左右。经过改进后，误报率普遍降到 6%~10%，漏报率也有一定降低，说明改进后识别效果明显提高。

(3)在对 4 种攻击类型的检测中，攻击数量越多，检测效果越好。这充分体现了神经网络自身的特点，即输入数据越完备，网络权值跟阈值的调节就越精确，训练出的网络效果也就越好，识别率也就越高。

(4)对于具体的某种攻击，采用不同的改进算法，得出的检测效果是有明显区别的。因此，在设计入侵检测模型中，可以取长补短，把多种改进算法结合起来，运用“寻优”技术，在处理某种具体的攻击类型时，选择某种特定的“最优”算法，将大幅度提高整体识别效果。

6 结束语

单纯把神经网络用于入侵检测，面对海量高维数据，势必引起神经网络规模剧增。若采用传统的特征选择，删除部分维数降维，又会带来信息丢失的风险。把主成分分析和 BP 神经网络结合起来用于入侵检测，在充分利用原始信息的基础上，大幅度降低 BP 神经网络的输入矢量维数，简化结构，提高神经网络的精度和性能，改善入侵检测的识别效果。本文对 3 种改进 BP 算法进行仿真实验，结果表明 3 种改进算法可以使入侵检测系统具有更高效率和更好的检测效果。

参考文献

[1] 蒋建春, 马恒太, 任党恩, 等. 网络安全入侵检测: 研究综述[J]. 软件学报, 2000, 11(11): 1460-1466.

[2] Arulampalam G, Bouzerdoum A. A Generalized Feedforward Neural Network Architecture for Classification and Regression[J]. Neural Networks, 2003, 16(5/6): 561-568.

[3] 易晓梅, 陈 波, 蔡家楣. 入侵检测的进化神经网络研究[J]. 计算机工程, 2009, 35(2): 208-213.

[4] 艾玲梅, 李 营, 马 苗. 基于 EMD 和 PCA 的 P300 分类算法[J]. 计算机工程, 2010, 36(5): 182-183, 187.

[5] 葛君伟, 沙 静, 方义秋. 具有混沌学习率的 BP 算法[J]. 计算机工程, 2010, 36(23): 168-170.

[6] Wang Jingen, Lin Shang, Chen Shifu, et al. Application of Fuzzy Classification by Evolutionary Neural Network in Incipient Fault Detection of Power Transformer[C]//Proc. of the Int'l Joint Conf. on Neural Networks. New York, USA: IEEE Press, 2004.

[7] The University of California Irvine KDD Archive[EB/OL]. [2007-06-26]. http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.

编辑 索书志