第31卷 第1期 石家庄铁道大学学报(自然科学版) Vol. 31 No.1
2018年3月 Journal of Shijiazhuang Tiedao University(Natural Science Edition) Mar. 2018

基于主成分分析和概率神经网络的入侵检测方法

赵广振[1]， 张翠肖[1]， 武辉林[2]， 高 婧[1]， 李 旋[1]

(1.石家庄铁道大学 信息科学与技术学院,河北 石家庄 050043;
2.河北省科学院 应用数学研究所,河北 石家庄 050081)

摘要:针对神经网络在入侵检测的应用中存在入侵数据冗余信息多,数据量大,训练时间长,易陷入局部最优等问题,提出了一种基于主成分分析(PCA)和概率神经网络(PNN)的入侵检测方法。首先使用PCA对数据进行特征降维,解决了入侵数据冗余信息多的问题;然后使用PNN建立入侵检测模型;其次,使用粒子群算法(PSO)解决概率神经网络参数的优化问题;最后使用KDD99数据集对该模型进行测试。实验结果表明:该方法能够有效提高检测的效果,而且检测速度明显提高。

关键词:入侵检测;主成分分析;粒子群算法;概率神经网络;网络安全

中图分类号:TP183 **文献标志码**:A **文章编号**:2095-0373(2018)01-0091-05

0 引言

截至2016年6月,我国网民规模达7.10亿,互联网普及率达到51.7%[1]。网络的快速发展给社会带来巨大的进步,但是其安全问题也造成了巨大的损失,如何保证网络的安全一直是人们关注的焦点。网络安全是由防护、检测、反应和恢复4个层次构成的一种综合防御体系[2]。入侵检测(Intrusion detection)作为防御体系的第二道防线,它是防火墙的合理补充,帮助系统应对网络攻击,是网络安全防护的重要技术之一。现有的入侵检测方法主要有以下几种:贝叶斯网络、马尔科夫模型、人工免疫原理、人工神经网络、支持向量机、极限学习机以及级联入侵检测等[3]。其中人工神经网络因其具有独特的结构,非线性模拟能力、强大的学习能力和自适应能力,在入侵检测系统中得到广泛应用。

虽然神经网络功能强大,在入侵检测的应用中仍然存在一些缺点。比如文献[4]采用改进的BP神经网络用于入侵检测,虽然利用可变学习速率算法、动量和批处理技术来提高网络的性能,但是数据的维数是复杂多变的,如果数据维数过高的话会导致整个网络的计算复杂度大幅度增加。又比如文献[5]利用遗传算法来解决神经网络收敛速度慢和陷入局部最优的问题,但是遗传算法搜索问题的解空间的大小跟网络参数的个数有关,这无疑会增加网络的训练时间,另外遗传算法处理高纬度的数据具有局限性。深度学习的出现使得解决复杂攻击数据有了新的解决思路,其通过组合低层数据特征形成更加抽象的高层特征来表示属性或者类别[6],不同于浅层神经网络,深度学习有多达5～6层,甚至10多层的神经网络结构。比如文献[7]采用了深度学习中的自编码网络模型实现对网络特征的提取,通过softmax分类器对特征数据进行分类。虽然保证了识别率,但是由于深度学习独特的网络结构,其训练速度比较慢。

针对以上问题,提出一种基于主成分分析和概率神经网络入侵检测方法。首先使用PCA对数据进行特征降维,消除冗余信息;然后使用概率神经网络(PNN)建立入侵检测模型;其次,使用粒子群算法(PSO)优化概率神经网络的参数;最后使用KDD99数据集对该模型进行测试。

1 主成分分析

主成分分析(PCA)是统计学中对数据分析的有力工具,将高维数据集变换到低维空间,保留最多的

收稿日期:2016-10-27 责任编辑:车轩玉 DOI:10.13319/j.cnki.sjztddxxbzrb.2018.01.16
基金项目:河北省科学院两院合作项目(161306);石家庄铁道大学研究生学院校企合作项目
作者简介:赵广振(1992—),男,硕士研究生,研究方向信息安全,E-mail:992924293@qq.com
赵广振,张翠肖,武辉林,等.基于主成分分析和概率神经网络的入侵检测方法[J].石家庄铁道大学学报:自然科学版,2018,31(1):91-95.

原始数据信息[8]。PCA 的基本原理如下:假设空间中存在 p 个具有相关性的变量,通过线性组合等方法消除其相关性,使其变成一组线性无关的变量,新变量要尽可能准确地反映原始变量的信息而且新变量的个数要少于原始变量。在这里用方差来度量信息,通常线性组合的方法有很多种,为了准确反映原始变量的信息,将方差最大的那组线性组合作为第一主成分,以此类推分别建立第二、三等主成分。需要注意的是,各个主成分之间是线性无关的。具体步骤如下:

第一步:将样本数据组成的矩阵进行标准化,目的是消除量纲对数据的干扰

$$x'_{ij}=(x_{ij}-\mu_j)/\sigma_j \tag{1}$$

式中,μ_j 为第 j 列数据的均值;σ_j 为第 j 列数据的方差。标准化后的数据 x'_{ij} 组成的矩阵为 $\boldsymbol{X}'$。

第二步:计算协方差矩阵 $\boldsymbol{C}_X$

$$\boldsymbol{C}_X=\frac{1}{p-1}\sum_{i=1}^{p}(\boldsymbol{X}'_i,\boldsymbol{X}_i'^{\mathrm{T}}) \tag{2}$$

第三步:采用雅克比方法计算协方差矩阵的特征值$(\lambda_1,\lambda_2,\cdots,\lambda_p)$与相应的特征向量 $\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_p$;

第四步:选择重要的主成分,并计算主成分的贡献率 η 和累计贡献率 $\sum\eta$

$$\eta=\frac{\lambda_i}{\sum_{k=1}^{p}\lambda_i},(i=1,2,\cdots,p) \tag{3}$$

$$\sum\eta=\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{p}\lambda_i},(i=1,2,\cdots,p) \tag{4}$$

第五步:根据累计贡献率,一般累计贡献率大于 85%以上就可以满足需求,这时取前 k 个特征值对应的特征向量$(\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_k)$组成变换矩阵 $\boldsymbol{Q}(p\times k)$。

第六步:计算降维后的新样本矩阵 $\boldsymbol{Y}$。

$$\boldsymbol{Y}_{m\times k}=\boldsymbol{X}'_{m\times p}\boldsymbol{Q}_{p\times k} \tag{5}$$

2 基于 PCA 的概率神经网络入侵检测模型

2.1 概率神经网络

概率神经网络是由径向基神经网络发展而来的一种前馈型神经网络,其理论依据是贝叶斯最小风险规则(贝叶斯决策理论)[9]。不同于传统神经网络的 Sigmoid 函数,它通常采用高斯函数作为网络的激活函数。其层次模型,由输入层、模式层、求和层、输出层共 4 层组成。网络的基本结构如图 1 所示。

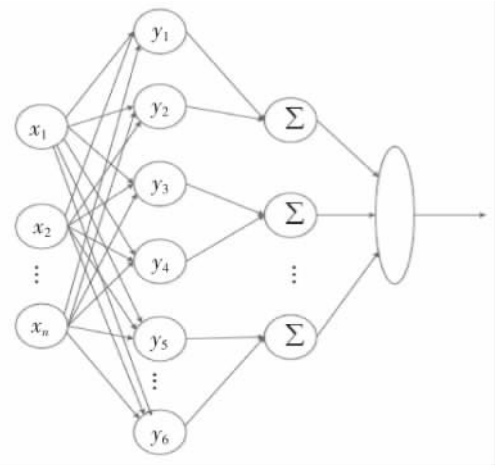

图 1 概率神经网络结构

输入层将样本数据输入到网络中,其神经元的个数同样本数据的维数相等。模式层神经元个数等于样本数据的总数目。假设样本数据为 $\boldsymbol{X}(x_1,x_2,\cdots,x_n)$,模式层的第 i 个神经元与输入层的权值矩阵为 $\boldsymbol{W}_i$ 则该神经元的输出 s_i 为

$$s_i=\exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{W}_i)^{\mathrm{T}}(\boldsymbol{X}-\boldsymbol{W}_i)}{2\delta^2}\right] \tag{6}$$

式中,δ 是一个很重要的参数,称为扩散常数或者散布常数。这个参数需要在实验中进行调整,过大或者过小都会影响网络的性能。

求和层单元只和属于某个类别的神经元相连,并将其概率累加并传到输出层。输出层的每个神经元对应一种类别,它根据求和层传来的各个类别的概率密度函数,将概率密度函数最大的那个神经元输出

为 1,其它神经元输出为 0,输出为 1 的神经元对应的类别即为网络识别出的类。

2.2　粒子群算法

粒子群优化算法(PSO)是一种进化计算技术(Evolutionary Computation),1995 年由 Eberhart 博士和 Kennedy 博士提出[10]。该算法是一种群智能算法,它通过研究鸟群搜索食物过程中的行为,从中发现个体与群体之间存在的协作和信息共享的规律,并利用这种规律对问题的解空间进行搜索,以求获得最优解。

粒子群算法是一种并行算法,具有容易实现、精度高、收敛快等优点。针对研究问题的解空间,粒子群算法首先在该空间内初始化一些随机解,这些解就是一些粒子,粒子的状态包括粒子的速度和位置两个方面。然后使用适应度函数评价粒子所处位置是否最优,使用两个全局变量 $pbest$ 和 $gbest$ 来记录粒子个体和群体所寻找过的最优位置。对于每个粒子,计算其适应值,若优于 $pbest$ 就将其作为 $pbest$,若优于 $gbest$ 就将其作为 $gbest$,接着更新粒子的速度和位置。粒子速度和位置的更新规则如下

$$V_i = wV_i + c_1 \times c_1 \times rand() \times (pbest_i - x_i) + c_2 \times rand() \times (gbest_i - x_i) \tag{7}$$

$$x_i = x_i + V_i \tag{8}$$

式中,V_i 是粒子的速度;$rand()$是介于 0 和 1 之间的随机数;x_i 是粒子的当前位置。c_1 和 c_2 是学习因子 。

如果粒子的速度或者位置超出搜索的范围就将其设定为最大速度或者边界位置。粒子更新完毕后如果没有找到满足要求的最优解就继续迭代搜索。搜索最优解的停止条件一般选为最大迭代次数或者粒子群搜索到的最优位置满足预定最小适应阈值。

2.3　基于 PCA 的概率神经网络入侵检测模型

基于以上所做的工作,提出了一种基于 PCA 的概率神经网络入侵检测模型(PCA-PNN)。如图 2 所示。

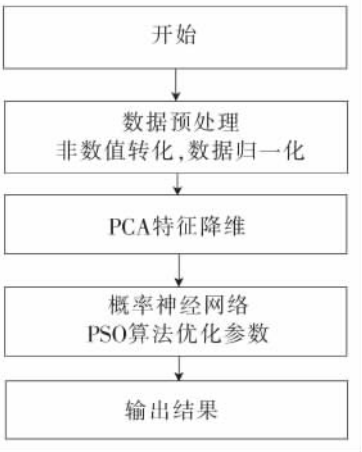

图 2　基于 PCA 的概率神经网络入侵检测模型

该模型首先利用 PCA 对训练数据和测试数据进行降维,在降维的时候先对训练数据进行降维,然后用降维产生的降维矩阵乘以测试数据就可以对测试数据进行降维,以保证新的训练数据和测试数据处在同一个基中;接着将数据输入到 PNN 网络中进行训练,采用粒子群算法对 PNN 网络的散布常数 Spread 进行优化,最后测试 PNN 网络的性能。

3　实验及结果分析

采用的实验工具是 Matlab R2010a,为了衡量算法的优劣,将未经过粒子群算法优化和特征降维的 PNN 网络作为对比。为评价上文提出的入侵检测模型,选用 KDD99 数据集。该数据集是由麻省理工学院 Lincoln 实验室仿真美国空军局域网环境建立的网络测试数据集,包含约 500 万条网络连接记录,除了正常的数据之外,还有 Dos、R2L、U2R 和 Probe 4 个大类的攻击数据。在研究中通常使用该数据集中的 10%训练数据集和 10%测试数据集[11]。定义了几个检测指标,分别是:

检测时间,完成预测所需的时间;

检测精度,被正确分类的数据占总检测数据的比例;

检测率,被正确分类的入侵数据占总入侵数据的比例;

误报率,正常数据中被错误分类的数据占总正常数据的比例。

为了模拟真实的网络环境,分别在训练集和数据集的正常数据和 4 大类的攻击数据中随机取了 10 000条数据,其中正常数据占了 96%,具体数据的分布见表 1 所示。

实验预处理:每个样本有 42 个特征,其中有些是字符型的特征,因此必须要对样本进行预处理,将不同的字符型的数据替换为不同的数字。比如在网络服务类型这项中令 aol=1,auth=2,bgp=3,courier=4,以此类推。然后使用 Matlab 将样本数据的排列顺序打乱。接着将样本中表示类别的特征提取出来,作为标签数据,最后对样本数据进行 L2 范数归一化。

表 1 实验数据

数据类型	训练集样本数量	测试集样本数量	类别标签
Normal	9 600	9 600	1
Dos	200	200	2
R2L	76	76	3
U2R	39	39	4
Probe	85	85	5

数据降维:使用 PCA 算法对数据进行降维。在 PCA 中,原始数据矩阵被转换到新的矩阵中,这个新的矩阵的维度按照每个特征的贡献率从左到右排列。新矩阵中每个维度的贡献率见表 2 所示。

表 2 PCA 分析结果

维度	贡献率	累积贡献率
D1	0.476 0	0.476 0
D2	0.222 2	0.698 3
D3	0.148 7	0.847 0
D4	0.094 4	0.941 4
D5	0.030 6	0.972 1
D6	0.016 4	0.988 5
D7	0.007 4	0.995 9
D8	0.003 3	0.999 2
D9	0.000 7	1.000 0

一般来说达到 85%以上就能够很好地包含所有的信息,本文将累计贡献率设为 98%,取维度为 6。在降维的时候,要将训练集和测试集降到基相同的向量空间内。

训练和测试:将训练数据和标签数据输入到概率神经网络中进行学习,针对概率神经网络的散布参数 spread 的取值问题,使用粒子群算法对 spread 进行优化。粒子群算法的自适度函数为 $f=\frac{1}{n}\sum_{i=1}^{n}w_i$,其中 $n=5$,$w_i=\frac{\text{correct}_i}{\text{sum}_i}$,$\text{correct}_i$ 表示能被正确识别的第 i 类攻击的样本数量,sum_i 表示第 i 类样本的总数量。由于只针对 spread 一个参数,所以问题解的编码空间是一维的。实验中所有的参数见表 3 所示。

表 3 实验参数

参数名	参数值	备注
累积贡献率	0.98	满足需求,且维度为 6,大小适中
散布常数	0.023 2	粒子群算法产生的最优解
惯性权值	0.8	决定着局部收敛和全局收敛的速度
加速因子	2,2	建议 $c_1+c_2<=4$
迭代次数	50	超过 50 时,实验结果无明显变化
种群数量	50	同时参与搜索的粒子数量
粒子维度	1	由自适度函数的输入决定,本文针对散布常数

然后对网络模型进行测试,并计算出上述的检测指标。对比实验的具体结果见表 4 所示。

表 4 实验结果

实验方法	散布常数	时间/s	精确率/%	检测率/%	误报率/%
优化的 PAC-PNN	0.023 2	6.16	98.28	89	1.33
未优化的 PCA-PNN	0.01	6.22	83.68	87.75	16.49
	0.015	6.02	85.32	88.25	14.80
	0.02	6.10	97.83	89	1.8
	0.03	6.40	96.81	89	2.9
优化的 PNN	0.021 2	35.38	99.04	89.25	0.55
未优化的 PNN	0.05	34.68	98.78	83.25	0.6
	0.12	34.31	98.03	64.75	0.58
	0.153	34.06	98.01	56.5	0.26

在未优化的网络中,散布常数的上述取值是在众多的逐个尝试中取得的最好结果。从实验结果中可以看出,使用粒子群算法对 PCA-PNN 网络进行优化,能够找到散布常数的最优值,从而提高网络的性能,而且检测时间明显比未经过降维的网络短,因为降低数据的维度之后网络的运算负担就会减小很多,

从而提高效率。从精确度、检测率和误报率这 3 个指标来看,虽然 PCA 方法尽可能地保留了原始数据的完整性,仍然还是有一些信息的损失,因此其性能比优化的传统 PNN 网络稍弱一些。

由于入侵数据在总的数据中只占了很小的比例,在未优化的 PNN 中,趋向于将数据分类到正常数据中,因此其检测入侵数据的检测率不高,但是误报率低。优化的 PCA-PNN 比起传统的 PNN 网络来说,无论是性能还是检测时间都有明显的优势。

4　结论

针对神经网络在入侵检测的应用中存在入侵数据冗余信息多,数据量大,训练时间长,易陷入局部最优等问题,提出了一种基于 PCA 的概率神经网络入侵检测方法。该方法通过 PCA 消除了数据的冗余,降低了数据的维度,缩短了网络训练的时间,使用粒子群算法提高了概率神经网络的性能。实验结果表明,将主成分分析的方法和粒子群优化算法与概率神经网络结合是有效的,具有一定的推广意义。但是本文是在公开数据集上所做的实验,实际网络环境中的数据比数据集更加真实,也更加复杂。因此,下一步的工作是将该方法运用到真实的网络中,通过在网络中得到的反馈来改进该方法。

参　考　文　献

[1]中国互联网络信息中心. 第 38 次中国互联网络发展状况统计报告[EB/OL].[2016-8-3].http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201608/t20160803_54392.htm.
[2]卿斯汉, 蒋建春, 马恒太, 等. 入侵检测技术研究综述[J]. 通信学报,2004,7:19-29.
[3]方向, 王丽娜, 贾颖. 智能化入侵检测算法研究综述[J]. 通信技术,2015,12:1321-1328.
[4]丁玲, 赵小刚. 改进 BP 神经网络用于入侵检测[J]. 微计算机信息,2012,3:131-132,84.
[5]沈夏炯, 王龙, 韩道军. 人工蜂群优化的 BP 神经网络在入侵检测中的应用[J]. 计算机工程,2016,2:190-194.
[6]尹宝才, 王文通, 王立春. 深度学习研究综述[J]. 北京工业大学学报,2015,1:48-59.
[7]李春林, 黄月江, 王宏, 等. 一种基于深度学习的网络入侵检测方法[J]. 信息安全与通信保密,2014,10:68-71.
[8]苑津莎, 尚海昆. 基于主成分分析和概率神经网络的变压器局部放电模式识别[J]. 电力自动化设备,2013,6:27-31.
[9]蒋芸,陈娜,明利特, 等. 基于 Bagging 的概率神经网络集成分类算法[J]. 计算机科学,2013,5:242-246.
[10] Kennedy J. Small worlds and mega-minds: effects of neighborhood topology on particle swarm performance[M]// Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on[C]. Washington USA: IEEE, 1999.1945-1950.
[11]张新有, 曾华燊, 贾磊. 入侵检测数据集 KDD CUP99 研究[J]. 计算机工程与设计,2010,22:4809-4812,4816.

Intrusion Detection Method Based on Principal Component Analysis and Probabilistic Neural Network

Zhao Guangzhen[1], **Zhang Cuixiao**[1], **Wu Huilin**[2], **Gao Jing**[1], **Li Xuan**[1]

(1.School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China;
2.Institute of Applied Mathematics, Hebei Academy of Sciences, Shijiazhuang 050081, China)

Abstract: In view of the redundant information, big volume of data and long training time of neural network application in intrusion detection, and easy to get trapped in local optimum, an intrusion detection method based on principal component analysis and probabilistic neural network is proposed in this paper. Firstly, the dimension of the original intrusion data is reduced by principal component analysis, which solves the problem of redundant information of intrusion data. Secondly, the probabilistic neural network is used to establish the intrusion detection model. Then, the problem that it is hard to optimize the parameters of the probabilistic neural network is solved by using particle swarm optimization algorithm. Finally, the KDD99 data set is used to test the performance of the model. The experimental result shows that both the precision rate and detection speed are improved effectively by using the proposed method.

Key words: intrusion detection; principal component analysis; particle swarm optimization; probabilistic neural network; network security