DOI: 10.3969/j.issn.1006-4729.2013.06.017

一种基于DBN的网络入侵检测算法

徐东辉[a], 王 勇[b], 樊汝森[a]

(上海电力学院 a.电子与信息工程学院,b.计算机科学与技术学院, 上海 200090)

摘 要:针对传统BP神经网络在检测速度、精度、复杂度等方面的缺陷,提出了一种基于深度信念网(deep belief nets,DBN)的网络入侵检测算法,将数据通过双层RBM结构降维,再用BP神经网络反向微调结构参数,从而简化了数据复杂度,减少了BP神经网络的计算量.通过对KDD99数据集仿真实验表明,该算法对于大数据拟合快,检测精度较高.

关键词:深度学习; 神经网络; 入侵检测; 网络安全

中图分类号: TP393.07;TP393.08 **文献标志码**:A **文章编号**: 1006-4729(2013)06-0589-04

A Network Intrusion Detection Algorithm Based on DBN

XU Donghui[a], WANG Yong[b], FAN Rusen[a]

(*a. School of Electronic and Information Engineering, b. School of Computer Science and Technology, Shanghai University of Electric Power, Shanghai* 200090, *China*)

Abstract: DBN(deep belief nets)-based intrusion detection algorithm is proposed for improving the traditional BP neural network in detection speed, accuracy, complexity, etc. The BP neural network structure will do a fine turning on parameters after the double RBMs structure reducing the dimension of data. In this way, the complexity of data and BP neural network is simplified. The experiments with KDD99 dataset show that the new algorithm is in excellent performance in large data fitting and intrusion detection.

Key words: deep learning; neural networks; intrusion detection; network security

随着网络技术的快速发展,网络安全面临的问题也日益纷繁复杂.入侵检测系统(Intrusion Detection System,IDS)是一种主动的安全防护措施,它从系统内部和各种网络资源中主动采集信息,分析可能的入侵攻击行为,能有效地保证网络的安全.它是网络安全系统的重要组成部分.[1]

近年来,入侵检测在基于神经网络的模型研究方面取得了较大进展:文献[2]提出了基于云模型和BP神经网络的入侵检测模型,但只对特定的攻击名称具有较好的识别率,攻击类型的拟合较差;文献[3]介绍了基于模糊积分的多神经网络融合在入侵检测中的应用,但其收敛时间长,检测精度不理想;文献[4]提出了一种基于KPCA和BP神经网络的多核入侵检测分类系统,该模

收稿日期: 2013-09-30

通讯作者简介:徐东辉(1989-)男,在读硕士,浙江海宁人.主要研究方向为电力信息安全.E-mail:xdh_1989@126.com.

基金项目:信息安全国家重点实验室(中国科学院软件研究所)开放式基金(04-02-1);上海市教育委员会创新基金(11YZ192);上海市"科技创新行动计划"重点项目(11511504400).

型需要耗费大量的收敛步数.

由此可见,神经网络迭代次数多,收敛速度慢,易陷入局部最优的缺点难以用模型结合的方式弥补,但是它对于大数据的拟合能力却又难以舍弃.因此,在高速的网络环境下,如何建立一个能够适应海量数据的快速收敛模型成为亟待解决的问题.

1 DBN 模型

1.1 DBN 简介

深度信念网(deep belief nets,DBN)是由若干受限玻尔兹曼机(restricted boltzmann machine,RBM)和一个 BP 神经网络构成的深度结构.其训练过程主要包括两个方面:一是利用 RBM 结构训练,筛选数据特征信息;二是将各层 RBM 连结,在最后一层设置 BP 神经网络,将 RBM 输出作为 BP 神经网络的输入,并利用数据进行监督训练,构成整个深度结构.[5]

DBN 将原始输入进行逐层的特征提取,从具体到抽象,使得神经网络得到的输入成为一个更加易于分类的特征向量.同时,多层 RBM 组成的深层结构使得在特征提取过程中的错误或者冗余信息被逐层弱化,并最终在 BP 神经网络的反向调整中使模型达到整体最优.与传统神经网络相比,DBN 深度结构的优势在于克服了传统神经网络在深度结构增加时训练时间长、易陷入局部最优、大数据处理慢的缺点.

1.2 RBM 算法的实现

RBM 由多伦多大学教授 HINTON G E 首先提出.[6]他从热力学能量函数的角度,基于热力平衡的原理,将玻尔兹曼机进行参数修剪,使之能够运用于有监督和无监督学习,通过大量实例的学习,提取数据的特征.RBM 结构如图 1 所示.

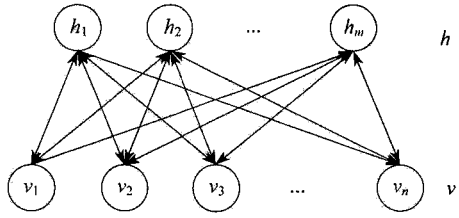


图 1 RBM 结构

RBM 由两个部分组成:可见层 v(visible)和隐藏层 h(hidden),一般将可见层作为输入层,隐藏层作为输出层(又称特征提取层).层间节点两两相连,每条连接都有相应的权值.

RBM 结构的主要特点是可见层和隐藏层可以相互表示,是一个特征提取结构.而多个 RBM 的叠加则是数据特征由具体到抽象的过程.本文所采用的深度结构 DBN 即是采用两个 RBM 逐层提取数据特征,降低了数据维数,减少了 BP 层反向误差修正的计算量,弥补了多层 BP 收敛慢、计算复杂的缺点.

若用 v_i 表示可见层节点值,h_j 表示隐藏层节点值,a_i 和 b_j 表示它们的偏置,w_{ij} 表示节点之间的权值,则:

$$P(h_j = 1) = \frac{1}{1 + \exp(-a_j - \sum_i v_i w_{ij})} \tag{1}$$

由于隐藏层和可见层可以相互表示,则有:

$$P(v_i = 1) = \frac{1}{1 + \exp(-b_i - \sum_j h_j w_{ji})} \tag{2}$$

随后 HINTON G E 提出了训练 RBM 的 CD(Contrastive Divergence)准则,[6]提出权值向量 $\boldsymbol{\theta}(a,b,w)$ 的更新公式为:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + u \frac{\partial \log_p(v,h)}{\partial \boldsymbol{\theta}} \boldsymbol{\theta}_i \tag{3}$$

参照 CD 准则,本文中的 RBM 采用基于 CD 准则的批训练算法如下:

(1)利用随机函数,初始化 a,b,w,设定迭代次数 a_{alph},批次大小(batchsize),动量 u;

(2)以 RBM 输入为向量 $\boldsymbol{v}_0$,代入式(1)和式(2)计算向量 $\boldsymbol{h}_0,\boldsymbol{v}_1,\boldsymbol{h}_1$;

(3)按批次求得 $\mathrm{sum}\left(\frac{\partial \log_p(v,h)}{\partial \boldsymbol{\theta}}\right)$,并除以批次大小,按式(3)更新权值向量 $\boldsymbol{\theta}$;

(4)判定迭代次数,若迭代次数 $x < a_{alph}$,则转至(2),若 $x = a_{alph}$ 则算法停止.

2 基于 DBN 的网络入侵检测模型

2.1 系统架构

基于 DBN 的网络入侵检测系统,首先利用 Excel 将 KDD99 数据集进行数据编码和归一化处理;然后将数据输入到一个[40,20]RBM 结构,数据由 40 维降至 20 维,再通过[20,20]的

RBM 结构,深度提取数据特征;最后将第二个 RBM 结构的输出作为 BP 层的输入,再经过监督学习,调整整个 DBN 结构的参数. 由于两层 RBM 都对数据进行了降维和特征提取,使得 BP 层能快速收敛到最优结果.

2.2 数据来源

1998 年,MITLincol 实验室进行了一项入侵检测评估项目,收集了 9 周时间的 TCPdump 网络连接和系统审计数据. 项目仿真各种用户类型、各种不同的网络流量和攻击手段,使之类似于一个真实的网络环境. 每个网络连接被标记为正常(normal)或异常(attack),异常类型被细分为 4 大类(DOS,R2L,U2R,PROBING)共 39 种攻击类型,其中 22 种攻击类型出现在训练集中,另有 17 种未知攻击类型出现在测试集中.

之后,来自哥伦比亚大学的 STOLFO Sal 教授和来自北卡罗莱纳州立大学的 LEE Wenke 教授采用数据挖掘等技术对以上的数据集进行特征分析和数据预处理,形成了一个新的数据集,如表 1 所示. 1999 年 KDD CUP 采用了该数据集,命名为 KDD99 数据集. 虽然年代较久远,但 KDD99 数据集仍然是网络入侵检测领域的事实基准,为网络入侵检测研究奠定了基础.

表 1 KDD99 数据类别统计

标 签	类 别	训练集	测试集
0	NORMAL	97 278	60 593
1	PROBE	4 107	4 166
2	DOS	391 458	229 853
3	U2R	52	228
4	R2L	1 126	16 189

2.3 算法实现

本实验采用 KDD99 数据集中 10% KDD 数据集作为训练集,corrected 数据集作为测试集. 其中,测试集中包含 17 种未知入侵类型,共18 729 行连接记录,占总测试量的 6%. 因此,corrected 数据集不仅要求测试模型具有强大的数据拟合能力,更需要有一定的泛化能力.

(1)数据处理 由于数据中有字符编码,也有数据编码,故应首先对数据进行编码和归一化整理.

(2)字符编码 以数据第 2 列为例,共有 tcp,udp,icmp 3 种类型,必须将字符编码,编辑成 3 种,即 0,1,2,或者是 001,010,100 二进制编码类型. 考虑到第 3 项 service 有 70 种类型,故选择前一种编码方式.

(3)归一化处理 为了使之后的实验更加准确,必须将训练集和测试集进行归一化处理. 整理之后所有数据都在[0,1]区间.

(4)数据集整理 由于前 41 列中的第 20 列数据全为零,故将其删除,构成 40 维的数据,作为 DBN 的输入;将第 42 列作为决策项进行编码,如表 2 所示,作为 DBN 结构的输出.

表 2 决策项编码

序 号	类 别	编 码
0	NORMAL	1,0,0,0,0
1	PROBE	0,1,0,0,0
2	DOS	0,0,1,0,0
3	U2R	0,0,0,1,0
4	R2L	0,0,0,0,1

3 实验结果及分析

本实验环境为 Windows 7(64 位)操作系统,4 G 内存,MATLAB2012b.

首先,设定 RBM 和 DBN 迭代次数为 1,动量为 0.5,批次大小为 50;BP 的批次大小为 50,去迭代次数为 2,5,10,得到的误差率如表 3 所示.

表 3 Corrected 数据集测试结果

迭代次数/次			动 量		批次大小			误 差
RBM	DBN	BP	RBM	DBN	RBM	DBN	BP	/%
1	1	2	0.5	0.5	50	50	50	9.206 8
1	1	5	0.5	0.5	50	50	50	8.581 7
1	1	10	0.5	0.5	50	50	50	7.876 8

从表 3 可以看出,在 BP 迭代次数增加的过程中,误差逐渐减小,但迭代次数增加到一定程度,误差便趋于稳定,甚至大于 10 次会出现过拟合现象;DBN 在较低的迭代次数下仍有 90% 以上的测试精度,并且在较好的情况下能够达到 92.1%.

其次,剔除 18 729 行未知分类,进行测试,设定 RBM 和 DBN 的迭代次数为 1,动量为 0.2,批次大小为 25,BP 的批次大小为 50,得到的误差波

动如图 2 所示. 由图 2 可知,测试结果在[0.30,0.35]区间浮动,显示了 DBN 优秀的数据拟合能力.

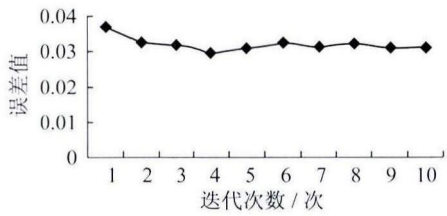


图 2 已知分类的测试结果

最后,将被剔除的 18 729 行未知分类作为测试集进行仿真实验,参数设置如下:RBM 和 DBN 的迭代次数为 1,动量为 0.2,批次大小为 25,BP 的批次大小为 50,得到的误差波动如图 3 所示.

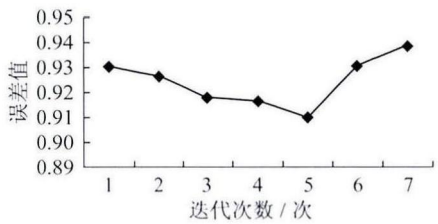


图 3 未知分类的测试结果

由图 3 可知,在迭代次数为 5 时,得到的最优误差为 0.91,表明 DBN 对于数据的泛化能力相对较差.

4 结 语

本文利用 RBM 和 BP 神经网络构成的 DBN 深度结构建立了电力网络入侵检测模型,并以 KDD99 数据集为训练和测试数据进行了仿真实验. 实验结果表明,DBN 在较低的迭代次数下能够得到较好的检测精度. 但通过对未知分类数据的测试表明,DBN 的泛化能力有所欠缺. 这需要我们在进行数据编码时,对于相似分类进行一定的数据预处理,以进一步提高其泛化能力.

参考文献:

[1] 肖敏,韩继军,肖德宝,等. 基于聚类的入侵检测研究综述[J]. 计算机应用,2008,28(6):34-38.

[2] 张铜. 基于云模型和 BP 神经网络的入侵检测[J]. 传感器与微系统,2011,30(1):116-118.

[3] 刘玉玲. 基于模糊积分的多神经网络融合在入侵检测中的应用[J]. 河北大学学报,2008,28(3):312-317.

[4] 刘继清. 一种基于 KPCA 和 BP 神经网络的多核入侵检测分类系统的研究[J]. 江苏科技大学学报,2011,25(6):602-605.

[5] HINTON G E, OSINDERO S, TEH Y W. A fastlearning-algorithmfordeepbeliefnets[J]. Neural Computation, 2006, 18(7):1 527-1 554.

[6] HINTON G E. A practical guide to trainingrestricted boltzmannmachines[D]. Toronto University of Tornto, 2010.

(编辑 胡小萍)