

基于样条权函数神经网络 P2P 流量识别方法

侯善江¹, 张代远^{1,2,3}

- (1. 南京邮电大学 计算机学院, 江苏 南京 210003;
2. 江苏省无线传感网高技术研究重点实验室, 江苏 南京 210003;
3. 南京邮电大学 计算机技术研究所, 江苏 南京 210003)

摘要:样条权函数神经网络是一种新兴的神经网络,克服了很多传统神经网络(如BP、RBF)的缺点:比如局部极小、收敛速度慢等。它具有拓扑结构简单、精确记忆训练过的样本,反映样本的信息特征,求得全局最小值等优点。基于这些优点,文中提出了一种基于样条权函数神经网络 P2P 流量识别方法。通过提取 P2P 流量特征,运用样条权函数神经网络结构对 P2P 流识别。Matlab 仿真和模拟实验结果表明了这种方案的可行性,与传统神经网络相比,样条权函数神经网络在时间效率上具有明显优势。

关键词:样条权函数;神经网络;P2P;流量识别;插值

中图分类号:TP301

文献标识码:A

文章编号:1673-629X(2014)07-0021-04

doi:10.3969/j.issn.1673-629X.2014.07.006

P2P Traffic Identification Based on Spline Weight Function Neural Network

HOU Shan-jiang¹, ZHANG Dai-yuan^{1,2,3}

- (1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003, China;
3. Institute of Computer Technology, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)

Abstract: Spline weight function neural network is a new kind of neural network. It overcomes many defects of traditional neural networks (like BP, RBF), such as local minima, slow convergence, at the same time has many advantages, such as simple structure, remembering trained samples, reflecting the characteristics of the sample information, finding global minima directly and so on. A method of P2P traffic identification based on spline weight function neural network is presented in this paper based on advantages of this neural network. The structure of spline weight function neural network can identify P2P traffic by extracting characteristics of P2P traffic training. Matlab simulation and experimental results show the feasibility of the scheme. Compared with the traditional neural network, spline weight function neural network has obvious advantages in time efficiency.

Key words: spline weight function; neural network; P2P; traffic identification; interpolation

0 引言

随着互联网技术和应用深入发展, P2P 技术改变了传统的互联网通信模式。据统计, 目前 P2P 流量占用了 60% 以上的互联网络资源^[1], 网络带宽被大量消耗、网络资源监管困难以及安全隐患等诸多问题随之出现, 如何有效地对 P2P 流量进行识别与控制已经成为当前研究的热点问题。传统技术对 P2P 流量的识

别方法较单一, 一般只采用针对固定端口的监视, 这就可能导致对网络流量的分析失效。为此, 研究 P2P 流量的识别对网络管理和流量监控具有非常重要的现实意义^[2]。

样条权函数神经网络是文献[3]中提出的一种全新的神经网络。它克服了传统神经网络存在局部极小、收敛对初值敏感及收敛速度慢等缺点^[4], 并且该算

收稿日期: 2013-09-02

修回日期: 2013-12-06

网络出版时间: 2014-02-24

基金项目: 江苏高校优势学科建设工程资助项目(yx002001)

作者简介: 侯善江(1989-), 男, 硕士研究生, CCF 会员, 研究方向为人工智能; 张代远, 教授, 博士, 硕士生导师, 研究方向为人工智能、计算机体系结构、计算机应用等。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20140224.0930.058.html>

法的神经网络拓扑结构非常简单,只有两层,其中输入层的权与神经元全互连,输出层没有权。只要知道输入样本向量和输出样本向量的维数,就可以确定神经网络的拓扑结构。基于样条权函数神经网络的诸多优点,文中将样条权函数神经网络应用到 P2P 流量识别和检测的过程中,可有效解决传统流量识别方法存在的问题。

1 样条权函数神经网络基本原理

样条权函数神经网络这种全新的人工神经网络,包括第一类样条权函数神经网络和第二类样条权函数神经网络^[3]。以多输入单输出样条权函数神经网络为例,对网络拓扑结构做介绍,图 1(a)为多输入单输出的第一类权函数(和函数)神经网络拓扑结构,图 1(b)为多输入单输出的第二类权函数(积函数)神经网络拓扑结构。

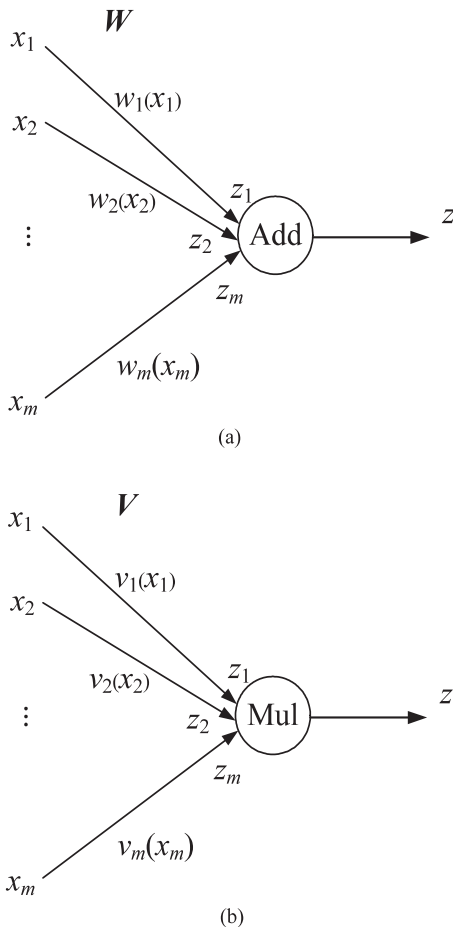


图 1 多输入单输出的权函数神经网络拓扑结构

可以看出这种神经网络拓扑结构简单,只有两层。只有输入层的权与神经元全互连,输出层没有权^[3]。而传统方法神经网络至少三层,通常各层之间的权全互连,这种神经网络结构比传统方法要简单得多。下面以多输入单输出的第一类权函数(和函数)神经网络为例介绍算法基本原理。

图 1 中 $x_i (i = 1, 2, \dots, m)$ 是输入样本,代表这个网络输入样本维数为 m , $w_i(x_i)$ 是输入样本 $x_i (i = 1, 2, \dots, m)$ 对应的权函数, $x_i (i = 1, 2, \dots, m)$ 经过权函数变换之后的输出为 $z_i (i = 1, 2, \dots, m)$, 圆圈 Add 表示加法器, Mul 表示乘法器, z 表示网络的输出^[3]。

$$z = \sum_{i=1}^m z_i \quad (1)$$

$$z_i = w_i(x_i) \quad (2)$$

z_i 与 z 的关系可以通过如下的加权系数联系起来,即

$$z_i = \eta_i z \quad (3)$$

式(3)中的加权系数 $\eta_i (i = 1, 2, \dots, m)$ 满足^[3]

$$\sum_{i=1}^m \eta_i = 1, 0 \leq \eta_i \leq 1$$

由式(2)和式(3)可得^[3]

$$z = (1/\eta_i) w_i(x_i) \quad (4)$$

假设每一个输入是由 m 维向量构成,输出样本由一维向量构成,共有 $N + 2$ 个样本需要训练, $w_i(x_i)$ 表示神经元与第 $i (i = 1, 2, \dots, m)$ 个输入节点相连的理论权函数, x_i 表示 m 维输入向量的第 i 个分量。由于有 $N + 2$ 个需要训练的样本,节点 x_i 有 $N + 2$ 个输入量,组成一个 $N + 2$ 维向量记为

$$\mathbf{x}_i = (x_{i0}, x_{i1}, \dots, x_{i(N+1)}) \quad (5)$$

对应的目标向量记为

$$\mathbf{z} = (z_0, z_1, \dots, z_{(N+1)}) \quad (6)$$

为了求得权函数,除了知道 x_i 的取值之外,还要知道函数 z_i 的值,根据式(3)和式(5)得到^[3]

$$\mathbf{z} = (w_i(x_{i0}), w_i(x_{i1}), \dots, w_i(x_{i(N+1)})) = (\eta_i z_0, \eta_i z_1, \dots, \eta_i z_{N+1}) \quad (7)$$

权函数 $w_i(x_i)$ 的形式将由输入向量 \mathbf{x}_i 和输出向量 \mathbf{z} 决定,根据插值理论得到对应的插值点为^[3]:

$$\mathbf{Ip}_i = \{\mathbf{Ip}_{i0}, \mathbf{Ip}_{i1}, \dots, \mathbf{Ip}_{i(N+1)}\} = \{(x_{i0}, \eta_i z_0), (x_{i1}, \eta_i z_1), \dots, (x_{i(N+1)}, \eta_i z_{N+1})\} \quad (8)$$

根据插值理论,能够构造出近似权函数。

这种全新的神经网络训练后的权值是输入值的样条函数,而 BP 等算法训练得到的权值是常数,无法反映与输入值的内在联系,样条权函数学习算法训练后的权函数能映射输入值的信息特征,经理论分析可知,其样本代价函数为 0,即能够得到全局最优解,解决了 BP 等算法局部极小的问题;样条权函数学习算法的主要计算工作量是求解样条函数,可以归结为求解线性方程组问题,其收敛速度相比于其他算法也是很快的;随着输入样本数量的增加,训练后的样条权函数更能加强网络的泛化能力^[3]。

2 基于样条权函数神经网络 P2P 流量识别模型

P2P,即 Peer to Peer,意思是“点对点”。P2P 网络是建立在现有网络应用层之上的一层覆盖网。称 P2P 网络中的参与者为节点,节点之间的通信方式不同于传统 C/S 模式下客户端与服务器之间的通信,客户端单纯地从服务器获取资源^[5-6]。而 P2P 网络中每个节点都共享一部分资源,其他节点可以同时访问这些资源。每个节点既是资源的提供者,又是资源的获取者^[7-8]。

传统的 P2P 流量识别技术大致可分为端口识别技术、深层数据包检测和基于流特征的识别技术三大类^[9]。这些技术中,基于端口的识别技术已经逐渐被淘汰,目前工程应用中大多数流量识别产品都采用深层数据包检测的方法,而基于流特征的技术是建立在数理统计基础上的,需要与其他方法相结合。在实际应用中,并非使用一种技术,通常情况下是以一种识别技术为主,配合其他方法使用。

P2P 作为一项 IP 服务,一般来说它的包具有消息联系和数据传输两个状态,而且都以很高的速率进行变换^[10]。几种常见的 IP 服务流特征见表 1,通过比较 P2P 能够很容易地与其他服务类型分开^[11-12]。区分 P2P 和 FTP 这两种服务类型的依据是消息联系,P2P 有消息联系,FTP 没有用户交互,这样就可以很简单地二者区分开来。

表 1 常见的 IP 服务特征

项目	平均传输速率	持续时间	字节数
HTTP	高	短	由低到高
VPN	低	长	高
Games	低	长	高
Streaming	中	长	高
Telnet	低	长	中
FTP / P2P	由中到高	长	高

当流的包有大小变化并且变化的跨度比较大时,每个流的包大小的均方差值就很大。P2P 流的包大小变换频率高于其他服务类型。因此,选择每个流的包大小均方差值、每个流的包大小变换频率、每个流的包大小平均值、每个流的包的数目、每个流的总共的字节数这 5 个值作为输入特征值^[13],来判别是 P2P 流还是非 P2P 流。识别模型如图 2 所示。

- 模型的工作流程为:
- (1)对网络流量数据集进行 P2P 流量特征提取形成 P2P 流量特征子集;
 - (2)对提取的特征子集构建基于样条权函数神经网络 P2P 流量识别分类器;

(3)将测试数据集输入到神经网络分类器模型进行 P2P 流量识别。

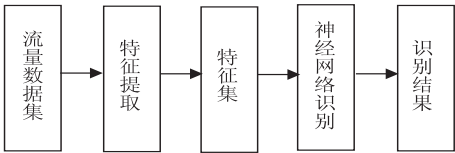


图 2 基于样条权函数神经网络 P2P 流量识别模型

由输入特征值决定了需要建立的样条权函数神经网络模型有 5 个输入神经元,即 $m = 5$;对于识别结果,可选取一个输出神经元,根据输出值大小,便可判断是否为 P2P 流。

3 实验仿真

运用 Matlab 工具箱对采集的流量数据进行分析,获得了 20 组流量状态样本数据,有相同个数的 P2P 流和非 P2P 流。选取 5 组测试样本,其中前 3 组为 P2P 流而后 2 组为非 P2P 流。在实验中选择第一类样条权函数神经网络作为识别模型的工具,构造权函数时选取三次样条插值函数,加权系数均选为 0.2。如果是 P2P 流,对应的输出值记为 1,如果是非 P2P 流,对应的输出值记为 0。对于训练好的神经网络输出结果选取 0.5 作为界限,若 $z \geq 0.5$,则为 P2P 流,若 $z < 0.5$,则为非 P2P 流。经过训练,对测试样本的测试结果如图 3 所示。

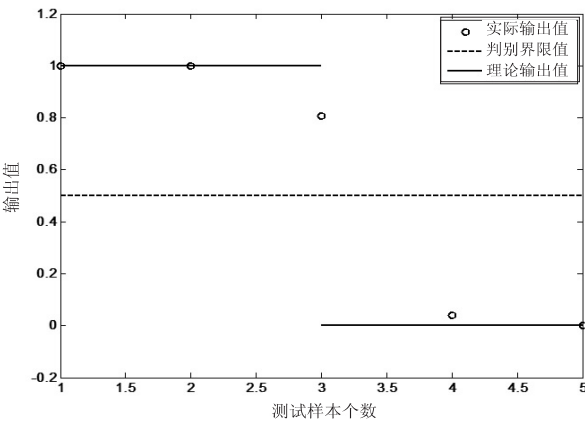


图 3 测试样本测试结果

由测试结果可以看出前三组输出值 $z \geq 0.5$,可判定为 P2P 流,后两组输出值 $z < 0.5$,可判定为非 P2P 流,识别结果是完全正确的。所以,基于样条权函数神经网络的 P2P 流量识别方法是可行的。

基于 BP 神经网络 P2P 流量识别方法已有不少学者进行了研究,但在网络训练时间上,样条权函数神经网络具有明显优势。选取上述的 20 组样本利用 BP 神经网络进行训练,网络模型结构选为 5-4-1,误差选为 0.05,所需训练时间为 2.773 s,而基于样条权函数神经网络的识别方法仅为 0.021 s。样条权函数神经网络

络能够精确记忆训练过的样本,而 BP 神经网络对于训练过的样本,实际输出值与目标值之间存在着误差,由此可见基于样条权函数神经网络的 P2P 流量识别方法在多方面是优于基于 BP 神经网络的 P2P 流量识别方法的。

4 结束语

由于样条权函数神经网络算法克服了传统神经网络存在局部极小、收敛对初值敏感及收敛速度慢等缺点,所以,文中提出的基于样条权函数神经网络的 P2P 流量识别方法具有独特的优势,实现了对 P2P 流量的准确快速识别。由于该算法是基于流量特征的,所以可以检测加密甚至是未知的 P2P 流量。但是,该算法识别结果的实时性不够,不能实现流量包的瞬时捕获识别。

参考文献:

- [1] Sen S, Wang J. Analying peer-to-peer traffic across large networks[J]. IEEE/ACM Transactions on Networking, 2004, 12(2): 219-232.
- [2] Huang Shucheng, Qu Yahui. Survey on data stream classification technologies[J]. Application Research on Computers, 2009, 26(10): 3604-3609.

(上接第 20 页)

参考文献:

- [1] 谢明明, 沈湘衡, 贺庚贤, 等. 空间相机仿真测试数据自动判读系统的设计[J]. 计算机测量与控制, 2010, 18(6): 1277-1279.
- [2] 吴伟, 张威, 潘顺良, 等. 自动判读系统在载人航天器电测中的应用[J]. 航天器环境工程, 2011, 28(6): 628-631.
- [3] Robinson R, McNab D, McNab A. A rule-based expert system for automatic error interpretation within ultrasonic flaw model simulators[J]. IEEE Trans on Systems, Man and Cybernetics, Part A: Systems and Humans, 2006, 36(6): 1202-1210.
- [4] 贺宇峰, 赵光恒, 吕从民, 等. 基于 CLIPS 专家系统的自动数据判读方法[J]. 中国科学院研究生院学报, 2011, 28(4): 505-513.
- [5] Du Zhiwei, Huang Yi. Design of automatic interpretation expert system for weapons ballistic testing data[C]//Proc of 3rd international symposium on systems and control in aeronautics and astronautics. Harbin: IEEE, 2010: 986-989.
- [6] 杨玉梅, 刁永锋. 基于 UML 顺序图的 Petri 网建模[J]. 计算机技术与发展, 2007, 17(10): 130-133.
- [7] Shaw O, Steggle J, Wipat A. Automatic parameterisation of

- [3] 张代远. 神经网络新理论与方法[M]. 北京: 清华大学出版社, 2006.
- [4] Yu Shiwei, Zhu Kejun, Diao Fengqin. A dynamic all parameters adaptive BP neural networks model and its application on oil reservoir prediction[J]. Applied Mathematics and Computation, 2008, 195(1): 66-75.
- [5] 邓河, 阳爱民, 刘永定. 一种基于 SVM 的 P2P 网络流量分类方法[J]. 计算机工程与应用, 2008, 44(14): 122-126.
- [6] 石萍, 陈贞翔, 荆山, 等. 基于对等特征的 P2P 流量识别方法[J]. 中国教育网络, 2007(2): 36-38.
- [7] 陈建华, 黄道颖, 张尧, 等. 计算机对等网络 P2P 技术[J]. 计算机工程与应用, 2003, 39(33): 162-164.
- [8] 陈浩. P2P 应用流量检测的研究与实现[D]. 北京: 北京邮电大学, 2007.
- [9] 张磊. 三种神经网络识别 P2P 流量的方法比较[D]. 重庆: 重庆大学, 2010.
- [10] 黄君毅, 吴静, 张晖. IP 流量分类算法中特征选择作用分析[J]. 计算机工程, 2010, 36(16): 68-70.
- [11] 张文, 侯立东. 基于传输层标志位的 P2P 流量识别技术[J]. 科技咨询, 2007(1): 163-164.
- [12] 陈宝钢, 张凌, 许勇, 等. 基于 P2P 应用的网络流量特征分析[J]. 计算机应用, 2007, 27(3): 531-533.
- [13] 沈富可, 常潘, 任肖丽. 基于 BP 神经网络的 P2P 流量识别研究[J]. 计算机应用, 2007, 27(S2): 44-45.

stochastic PetriNet models of biological networks[J]. Electronic Notes in Theoretical Computer Science, 2006, 115(3): 111-129.

- [8] 潘玲琳. 基于产生式规则的专家系统的研究实现[J]. 计算机技术与发展, 2007, 17(5): 66-68.
- [9] 吴晓, 赵光恒, 于英杰, 等. 仿真测试自动判读专家系统的设计与实现[J]. 系统仿真学报, 2005, 17(12): 3050-3052.
- [10] 周振红, 周洞汝, 杨国录. 基于 COM 的软件组件[J]. 计算机应用, 2001, 21(3): 6-8.
- [11] 钟金琴, 辜丽川. 一种面向对象的软件设计模式库的设计[J]. 计算机技术与发展, 2008, 18(9): 22-25.
- [12] Tilkov S, Vinoski S. Node.js: using JavaScript to build high-performance network programs[J]. IEEE Internet Computing, 2010, 14(6): 80-83.
- [13] 金晓鸥, 钟宝燕, 李翔. 基于 Rhino 的 JavaScript 动态页面解析研究与实现[J]. 计算机技术与发展, 2008, 18(2): 1-4.
- [14] 王怡苹, 李文海, 文天柱. 面向信号测试的路径搜索算法研究[J]. 仪器仪表学报, 2013, 34(7): 1650-1658.
- [15] 何国辉, 陈家琪. 游戏开发中智能路径搜索算法的研究[J]. 计算机工程与设计, 2006, 27(13): 2334-2337.

作者: 侯善江, 张代远, HOU Shan-jiang, ZHANG Dai-yuan
作者单位: 侯善江, HOU Shan-jiang(南京邮电大学 计算机学院, 江苏 南京, 210003), 张代远, ZHANG Dai-yuan(南京邮电大学 计算机学院, 江苏 南京 210003; 江苏省无线传感网高技术研究重点实验室, 江苏 南京 210003; 南京邮电大学 计算机技术研究所, 江苏 南京 210003)
刊名: 计算机技术与发展 
英文刊名: Computer Technology and Development
年, 卷(期): 2014(7)

参考文献(13条)

1. Sen S;Wang J Analying peer-to-peer traffic across large net-works 2004(02)
2. Huang Shucheng;Qu Yahui Survey on data stream classifica-tion technologies 2009(10)
3. 张代远 神经网络新理论与方法 2006
4. Yu Shiwei;Zhu Kejun;Diao Fengqin A dynamic all parame-ters adaptive BP neural networks model and its application on oil reservoir prediction 2008(01)
5. 邓河;阳爱民;刘永定 一种基于SVM的P2P网络流量分类方法 2008(14)
6. 石萍;陈贞翔;荆山 基于对等特征的P2P流量识别方法 2007(02)
7. 陈建华;黄道颖;张尧 计算机对等网络 P2P 技术 2003(33)
8. 陈浩 P2P应用流量检测的研究与实现 2007
9. 张磊 三种神经网络识别P2P流量的方法比较 2010
10. 黄君毅;吴静;张晖 IP流量分类算法中特征选择作用分析 2010(16)
11. 张文;侯立东 基于传输层标志位的P2P流量识别技术 2007(01)
12. 陈宝钢;张凌;许勇 基于P2P应用的网络流量特征分析 2007(03)
13. 沈富可;常潘;任肖丽 基于BP神经网络的P2P流量识别研究 2007(z2)

引用本文格式: 侯善江. 张代远. HOU Shan-jiang. ZHANG Dai-yuan 基于样条权函数神经网络P2P流量识别方法[期刊论文]-计算机技术与发展 2014(7)