基于径向基函数神经网络的网络流量识别模型

刘 晓

(暨南大学 信息科学与技术学院,广东 广州 510000)

摘 要:提出了一种基于径向基函数神经网络的网络流量识别方法。根据实际网络中的流量数据,建立了一个基于 RBF 神经网络的流量识别模型。先介绍了 RBF 神经网络的结构设计及学习算法,针对 RBF 神经网络在隐节点过多的情况下算法过于复杂的缺点,采用了优化的算法计算隐含层节点。仿真实验证明,该模型具有较好的准确率、低复杂度、高识别效果和良好的自适应性。

关键词:RBF 神经网络;流量识别;流量分类

中图分类号:TP183 **文献标识码**:A **文章编号**:1674-7720(2012)02-0077-03

Modeling network traffic based on radial basis function nerual network

Liu Xiao

(Computer Science Department, Jinan University, Guangzhou 510000, China)

Abstract: This paper presents a method of network traffic identification based on RBF (Radial Basis Function) neural network. With a large amount of real traffic data collected from the actual network, a nonlinear network traffic model based on radial basis function neural network theory was constructed to identify the network traffic. Firstly present the structure design and leaning algorithm of RBF neural network and then in order to reduce the artificial complexity of the RBF when too many hide layer units, present an optimize algorithm to calculate the numbers of hide layer units. Finally prove this identification method in the application of network traffic has the characteristics of high accuracy, low complexity and high recognition efficiency, and the practical feasibility in real-time traffic identification.

Key words: RBF neural network; traffic identification; traffic classification

随着互联网业务量的急剧增长,网络性能和服务质量方面的问题日益突出。在网络资源有限的情况下,建立网络流量模型,识别网络流量,及时作出控制或者调整,将会极大提高网络性能和服务质量。尤其是随着近年来互联网技术的发展,网络主要流量已经由传统的 FTP、TELNET 和 HTTP 向 P2P 和 IM 服务转变。传统的网络流量识别方法已经不能满足当前网络发展的需求。

神经网络对非线性函数关系具有良好的逼近能力,所以本文提出了一种基于 RBF 函数神经网络的网络流量模型。RBF 神经网络为局部神经网络模型,计算速度快、实时性好,相对于传统的线性流量模型具有更高的逼近能力和良好的自适应性,并可克服基于 BP 神经网络的流量模型训练时间长及计算复杂度高的不足。

1 RBF 神经网络结构及学习算法

1.1 RBF 神经网络结构

RBF 神经网络是 20 世纪 80 年代由 MOODY J 和 DARKEN C 提出的一种神经网络模型,是具有单隐层的前馈网络,属于局部逼近网络,已证明能以任意精度逼近任一连续函数。RBF 神经网络的结构如图 1 所示。

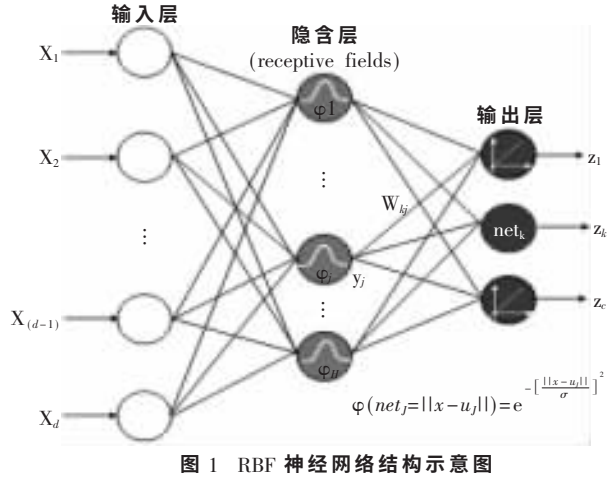

图 1 RBF 神经网络结构示意图

网络由输入层、径向基函数隐含层、输出层三层构成。低维空间非线性可分的问题总可以映射到一个高维空间，使其在此高维空间中为线性可分[1]。RBF 的输出单元部分构成一个单层感知机，只要合理选择隐单元数（高维空间的维数）和作用函数，就可以把原来的问题映射为一个线性可分问题[2]。RBF 网络中输入到隐含层的映射是非线性的，而隐含层到输出的映射是线性的。隐含层的节点数与实际问题的要求有直接的关联，过多的节点数会导致学习时间过长和低容错率，所以必须优化隐含层的节点数。隐含层的节点数可以采用式(1)计算：

$$h=\sqrt{(n+m)}+a \tag{1}$$

其中 n 是输入层的节点数，m 是输出层的节点数，a 是 1~10 的常数[3]。

隐含层基函数采用高斯函数：

$$\Phi_i(x)=\exp(-\frac{||x-c_i||^2}{2\sigma^2}) \tag{2}$$

隐节点的输出加权后进入输出层，输出层是其隐含层的线性组合[4-5]，即：

$$Y(x)=w_0+\sum_{i=1}^{n} w_i\Phi_i(||x-c_i||\sigma_i)=w_0+\sum_{i=1}^{n} w_i\exp(-\frac{||x-c_i||^2}{2\sigma^2}) \tag{3}$$

其中 $x\in R_n$ 为输入向量，$\Phi(\cdot)$是高斯核函数，$\|\cdot\|$ 是欧几里德范数，$c_i\in R_n$ 为第 i 个隐节点的场中心，$\sigma_i\in R$ 为第 i 个隐节点的场域宽度，n 是隐含层节点数，w_i 为第 i 个隐节点的基函数与输出节点的连接权值，w_0 为调整输出的偏移量。

1.2 RBF 神经网络学习算法

(1)对训练数据进行聚类，把基函数分别分配给每一个聚类。选择一组初始的中心值$\{\tilde{\mu}_1,\tilde{\mu}_2,\cdots,\tilde{\mu}_K\}$，用 K-均值聚类算法计算出中心值$\tilde{\mu}_k(1\leqslant k\leqslant K)$和宽度$\tilde{\sigma}_j$：

$$\tilde{\mu}_k=\frac{1}{p_k}\sum_{X\in C_k}X,\ \tilde{\sigma}_k=\frac{1}{p_k}\sum_{X\in C_k}||X-\tilde{\mu}_k||^2 \tag{4}$$

(2)计算隐含层的输出。

(3)实际输出与期望输出进行比较，应用梯度下降法训练权重，使得均方最小更新权重。

权重的改变值：

$$E(w)=\frac{1}{2}\sum_{k=1}^{K}(T_k^p-O_k^p)^2=\frac{1}{2}\sum_{k=1}^{K}\{T_k^p-f(\sum_{j=1}^{J}w_{jk}\Phi_j^p)\}^2 \tag{5}$$

其中，$\delta_j^p=(T_j^p-O_j^p)f(w_{jk}\Phi_j^p)$ (6)

如果是线性的则为：

$$\delta_j^p=[T_j^p-O_j^p] \tag{7}$$

权重更新：

$$w_{jk}^{new}=w_{jk}^{old}+\Delta w_{jk}=w_{jk}^{old}+\eta\delta_j^p\Phi_j^p \tag{8}$$

(4)对输入的 N 组数据重复步骤(2)~步骤(3)N 次。

(5)重复步骤(2)~步骤(4)，直至误差小到可接受的程度。

2 识别过程

流量识别过程分为四个部分：数据获取过程、数据预处理过程、数据训练过程和测试数据分类过程。重点在于建立一个 RBF 神经网络模型对网络流量进行分类。

(1)数据获取过程是通过数据获取模块提取网络连接记录和分析特征，以选择合适的网络特征属性作为原始的输入值。选择一组最合适的特征子集作为 RBF 神经网络的原始输入值。

(2)数据预处理过程是将特征子集映射到[-1,1]的范围[4]。

(3)数据训练过程是将经过预处理后的网络流量特征子集作为 RBF 神经网络模型的训练集。

(4)根据 RBF 神经网络的输出对网络流量进行分类。

3 试验与分析

本文选用流量文库 http://newsfeed.ntcu.net/中给出的两组实际数据进行实验，两组数据分别如表 1、表 2 所示。

表 1 实际数据一

类别	流量数	百分比/%	应用
attack	94	0.494	病毒攻击
database	183	0.962	Sqlnet,oracle
ftp-control	22	0.116	ftp 控制
ftp-data	72	0.378	ftp 数据
ftp-pasv	177	0.930	ftp-pasv
interactive	8	0.042	ssh,klogin,rlogin,telet
mail	1 278	6.716	imap,pop2/3,smtp
p2p	116	0.610	BitTorrent
services	77	0.405	DNS,ident,ldap,ntp
www	16 612	87.303	www
other	389	0.204	其他
总计	19 028	100	

表 2 实际数据二

类别	流量数	百分比/%	应用
database	329	1.382	Sqlnet , oracle
FTP	1 701	7.147	ftp
interactive	2	0.084	ssh , klogin , rlogin , telet
mail	2 726	11.453	Imap , pop3 , smtp
services	220	0.924	Dns , ident , ldap , ntp
www	18 559	77.976	www
other	170	0.714	QQ 游戏，在线视频
总计	23 801	100	

RBF 网络在数据一中采用 248 个输入层节点、262 个隐含层节点和 11 个输出层节点的结构；在数据二中采用 248 个输入节点、260 个隐含层节点和 8 个输出层节点的结构。实验结果如表 3 所示。

本文提出了一种基于 RBF 神经网络的网络流量识别方法。通过测试两组开发的网络流量数据集，证

表 3 实验结果

类别	识别率/%	
	数据一	数据二
ATTACK	86.3	
DATABASE	98.8	99.6
FTP-CONTROL	96.2	
FTP-DATA	99.7	99.6
FTP-PASV	91	
INTERACTIVE	65.5	66
MAIL	98.1	98.7
P2P	93.4	89.6
SERVICES	99.7	99.2
WWW	98	97.9
OTHER	98.1	98.1
Total	97.044 8	97.445 4

明该方法具有较高的准确度、低复杂性和良好的自适应性。

参考文献

[1] Shi Zhongzhi.Neural Network[M].Beijing:Higher Education Press,2009.

[2] COVER T M.Geometrical and statistical properties of system of linear inequalities with applications in pattern recognition[J].IEEE Transactions on Electronic Computer,1965(14):326-334.

[3] Fei Sike Technology R&D Center.Matlab Application[M].Beijing:Electronic Industry Press,2005.

[4] MOORE A W,ZUEV D.Discriminators for use in flow-based classification[A].Intel Research,Cambridge,2005.

[5] 王俊松.基于 Elman 神经网络的网络流量建模及预测[J].计算机工程,2009(9):190-191.

(收稿日期:2011-09-16)

作者简介:

刘晓,男,1986 年生,硕士,研究生,主要研究方向:人工智能。