

基于自适应模糊聚类的 LM-BPNN 网络故障诊断

蒋静芝, 孟相如, 张 立

(空军工程大学 电讯工程学院, 陕西 西安 710077)

摘 要: 提出一种综合运用模糊聚类 and 神经网络改进算法的网络故障诊断方法. 根据网络故障特征数据量大、且存在冗余和冲突的特点, 基于模糊聚类思想, 提出了以聚类中心为核、自适应半径来优选样本的数据预处理方法. 在进行故障特征的学习训练时, 针对 BP 神经网络用于网络故障诊断时训练次数多、收敛慢和易振荡的局限性, 使用结合了 Levenberg-Marquardt 的改进算法. 理论分析和实验结果表明, 文中提出的网络故障诊断方法能达到诊断更快速、更准确的效果.

关键词: 网络故障诊断; 模糊聚类; 自适应核区; LM-BPNN

中图分类号: TP306; TP393

文献标识码: A

文章编号: 1000-7180(2010)07-0153-04

LM-BPNN Based on Adaptive Fuzzy Clustering Algorithm to Network Fault Diagnosis

JIANG Jing-zhi, MENG Xiang-ru, ZHANG Li

(Telecommunication Engineering Institute, Air Force Engineering University, Xi'an 710077, China)

Abstract: Owing to the fact that the characteristic data of networks are huge and have redundancy and conflict each other, a method of data pretreatment based on fuzzy clustering is proposed. The center of each kind is used to be a core, with adapting the radius to sample from training samples. During learning of the characteristic data, Levenberg-Marquardt algorithm is imported in order to resolve the weaknesses of BP, such as large training times, slow convergence and incidental surge. Theoretical analysis and experimental results show that this method is efficient for network fault diagnosis.

Key words: network fault diagnosis; fuzzy clustering; adaptive hard core; LM-BPNN

1 引言

随着网络规模日益扩大, 维护与管理更加复杂, 现有的网络故障诊断技术难以适应网络的高速发展制约了网络效能的发挥. 目前, 基于无模型知识方法的神经网络 (Neural Network, NN) 在非线性、自学习和并行处理方面的优势, 使其成为网络故障诊断方法研究的热点之一.

网络动态信号携带的状态信息存在大量冲突和冗余, 直接输入神经网络训练将导致网络结构复杂、故障诊断精确性低. 为此, 文中提出在模糊聚类的基础上, 利用自适应优选核区对训练样本预处理, 保证

训练精度的同时实现数据的有选择压缩, 减少 NN 学习时间.

网络故障诊断中应用最多的是误差反向传播神经网络模型 (Backpropagation NN, BPNN). 研究表明^[1], 基于 BPNN 的网络故障诊断有很强的非线性映射、泛化和容错能力, 同时也存在训练次数多、收敛慢和易振荡的缺陷. 因此, 文中使用 Hagan 等^[2]提出的结合 Levenberg-Marquardt 优化算法的 BPNN 进行故障识别.

综合以上两种改进, 文中提出了基于模糊聚类自适应核区与 LM-BPNN 的网络故障诊断方法, 并通过实验仿真验证了所提方法的性能优势.

收稿日期: 2009-06-16; 修回日期: 2009-09-28

基金项目: 陕西省自然科学基金项目 (SJ08F14)

2 模糊聚类及自适应优选核区的设计

预处理训练样本能避免因网络状态信息存在冲突和冗余使神经网络的训练过程复杂、诊断精度不高,故成为故障诊断的重要步骤之一.文中的预处理方法设计如下:针对训练集中不同类型的故障样本,经模糊聚类获取聚类中心,以各中心为核、自适应距离为半径形成核区,其内样本构成新训练集.

2.1 模糊聚类

传统聚类方法的分类是硬划分.而网络故障的模糊性及故障与征兆的不确定性,适合进行软划分,因此引入模糊逻辑来处理聚类问题.

在模糊聚类分类样本的同时确定了各类的聚类中心,该中心代表了原始数据的基本趋势和特征,已有研究^[3]将聚类中心直接送入 BPNN 学习.聚类中心替代所有样本,减少了数据量、缩短了训练时间,但神经网络的学习需要一定数量的训练样本,只用中心数据并不能保证网络学习的准确性.故文中提出以各聚类中心为核、利用自适应半径选取核区,将核区的样本作为神经网络的输入.

2.2 自适应优选核区的设计过程

在以聚类中心为核的基础上形成核区,有选择地扩大训练样本的数量,既保证神经网络训练的准确性,又防止增加样本数据间的冗余和矛盾.但由于每类样本分布不同,若选取等长的核区半径,将导致选取的新训练样本集合中各种故障样本数量不等,数量少的故障类型得不到充分学习,数量大的故障类型的样本间可能引入了冗余和矛盾.为克服这个问题,文中设计了以各核区内样本个数相等作为准则自适应确定核区半径的方法,并用某一故障类型样本集的处理为例加以说明.

定义 1 核区.原样本集构成的 m 维空间中,以聚类中心 O_i 为核、问题所需距离为界构成的区域称为核区 Q_i .

核区内的样本个数记为 q_i ,包括核 O_i ,其中, $i = 1, 2, \dots, c$, c 为聚类数.

定义 2 核区半径.核区边界到核的距离,称为核区半径,记作 r_i .

分析可得,要使新训练样本集合中各种故障样本数量相等,就是实现 $q_i = t (i = 1, 2, \dots, c)$, t 称为核区目标样本个数.

自适应核区选取的具体步骤如下:

在包括 N 个样本的某类故障总体 $X = \{X_1, X_2, \dots, X_n\}$ 中,每个样本有 m 个特征: $X_i = \{X_{i1},$

$X_{i2}, \dots, X_{ij}, \dots, X_{im}\} (i = 1, 2, \dots, n)$, 其中, X_{ij} 表示第 i 个样本的第 j 个特征值.

(1) 计算各指标的均值 $\bar{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij} (j = 1, 2, \dots, m)$, 标准差 $S_j = \left[\frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 \right]^{\frac{1}{2}}$, 转换标准差 $X_{ij} = \frac{X_{ij} - \bar{X}_j}{S_j}$;

(2) 利用标准化数据,求得样本间的相似系数,建立模糊相似矩阵 R ;

(3) 计算传递闭包,得到模糊等价矩阵 R^* ;

(4) 计算 γ 截矩阵,实现样本分类、获取各类中心.

(5) 根据样本数量及训练性能要求,确定目标个数 S .

(6) 任取一个样本计算与核的距离 d .并令 $r_i = d$,求得该 Q_i 对应的 q_i .

(7) ① 若 $q_i \gg t$ 或 $q_i \ll t$,则重复(6);

② 若 $q_i \leq t$,则 r_i 在 d 基础上适当增大;

③ 若 $q_i \geq t$,则 r_i 在 d 基础上适当减小.

(8) 重复(7)直到 $q_i = t$.

通过上述方法自适应选取 c 个核区半径 r ,从原样本集中选取了 $c \times t$ 个样本数据的新训练集合.

用经自适应核区选取形成的训练样本集合代替全体样本有 3 个优点:①解决了原始数据的冲突问题;②减少了数据的冗余度;③压缩了数据的计算量.用核区代替聚类中心点的优点是:解决了训练数据不够导致的神经网络精度不高的问题,使网络具有良好的兼容性和扩展性.

3 基于 LM-BPNN 的网络故障诊断系统

3.1 LM-BPNN

最优化理论中,LM 算法是一种非常有效的优化设计方法,从收敛速度和收敛性来看,它是牛顿法和最速下降法的一种折中.基于 LM 最优化方法的 BP 算法,简称 LM-BP 算法.已有研究^[4]证明,LM-BP 算法性能稳定,收敛速度非常快,可应用于实时性要求较高的在线检测.另有研究^[5]显示:如果要求精度比较高,则该算法的优点尤其突出.由此可见,采用 LM-BP 神经网络不仅能够克服 BP 神经网络的缺陷,保证训练准确性,并且更能适合网络故障诊断的时效性要求.

3.2 网络故障诊断系统

文中提出的基于自适应模糊聚类的 LM-BPNN

网络故障诊断系统结构如图 1 所示.其中, I 为训练学习过程:将原训练样本经自适应优选核区后形成的新样本集输入 LM-BPNN 进行学习,并将学习结果放入故障知识库; II 指的是网络的实时检测过程:将从网络实时采集的信息输入知识库进行比对,若为已知故障直接得出结论,为未知则再输入 LM-BPNN 学习得出新的故障类型同时存入知识库;并按故障类型的判断结果,通知网络管理员修理故障,保证网络效能的良好.

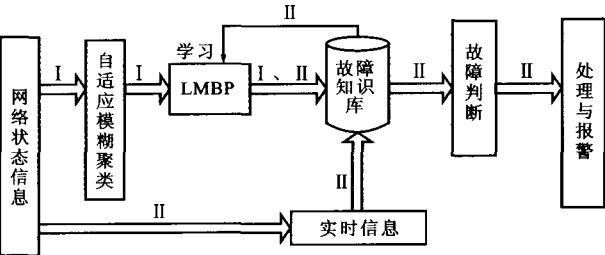


图 1 基于 LM-BPNN 的网络故障诊断系统

4 仿真验证

4.1 数据来源及样本结构

目前基于大型复杂网络的模拟和故障数据的获取相当困难,故文中采用 KDD'99 数据集^[6].该数据集取自模拟真实网络环境,数据集包含多种标出攻击类别的网络特征数据,可用于有监督的机器学习.

为确保数据的普适性,从 neptune、teardrop、portsweep、nmap 四种攻击类型的原始数据集中每隔选取训练和测试集,并对各种网络状态进行相应的编码.样本结构及对应状态编码如表 1 所示.

表 1 实验数据样本结构及对应状态编码

	训练集	测试集	对应编码
neptune	25	50	0001
teardrop	25	50	0010
portsweep	25	50	0100
nmap	25	50	1000
总计	100	200	—
样本属性	41	41	—

4.2 数据预处理

4.2.1 特征数值化及归一化处理

KDD'99 为异构数据集,有文字描述信息 http、telnet 等,也有变化范围较大的数值信息.将以字符表征的属性,按 KDD'99 数据集关于服务类型的说明数值化.

网络属性值具有不同的度量,这使得计算样本之间距离时,取值范围偏大的属性占了主导地位,弱化了取值范围偏小的属性的贡献,显然这样无法正确反映两个向量之间的差异,故本文采用最大值(或最小值)归一化方法加以处理.

4.2.2 模糊聚类及自适应核区形成

模糊聚类有很多种方式,本文用模糊 C 均值进行方法性能的验证.模糊 C 均值对训练样本进行聚类,将 neptune、teardrop、portsweep、nmap 四种网络状态的训练样本各自模糊聚类成 2 类,经综合考虑取核心区目标个数 t 为 5.通过 2.2 节的自适应核区选取方法得到典型样本组成新样本集.

4.3 实验及结果分析

在 MATLAB R2007 版本软件上进行仿真验证,有效利用神经网络和模糊逻辑工具箱的同时进行了自适应优选核心区的方法编辑和实现.

在单隐层神经网络能够实现任何精度收敛的研究基础上,采用三层网络拓扑结构形式.根据样本数据的特性,输入层节点数取 41,输出层节点取 4;根据隐层节点的选取原则以及进行实验对比收敛性能后,选取隐层节点数为 16.

为了验证所提方法的性能,将与 BP 神经网络、LM-BP 神经网络进行比较:

(1) 在 $\max_epoch = 5000$, $error_goal = 0.001$ 及 $lr = 0.05$ 的条件下,结果如图 2 所示.

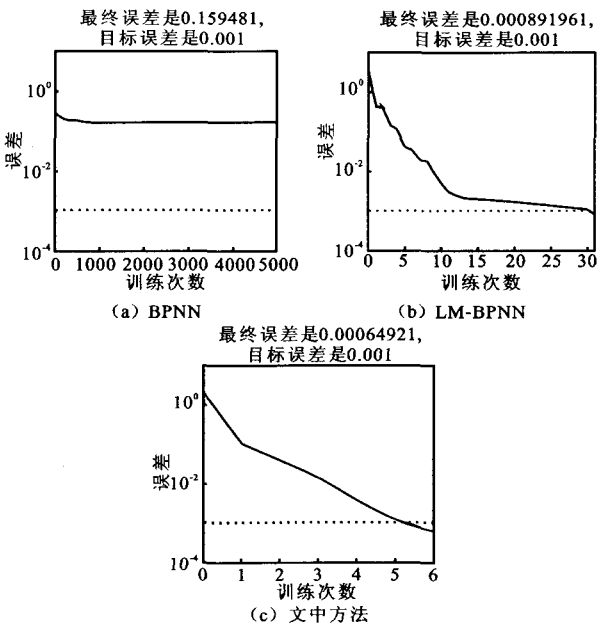


图 2 网络训练误差性能曲线

可以看出,图 2(a)的传统 BP 神经网络训练过程收敛非常缓慢,在最大训练次数为 5000 的条件下

下,无法实现目标误差 0.001 的收敛.图 2(b)中 LM-BP 神经网络的收敛速度相对较快,在 31 次训练后得到收敛;曲线抖动,说明存在收敛过度的情况.图 2(c)是文中方法的训练误差性能曲线,在相同的仿真环境中仅用 6 次就实现收敛,网络收敛速度最快;误差曲线平滑,基本上不存在收敛过度的现象.

实验结果表明文中方法能克服 BP 神经网络的缺陷及网络故障特征数据量庞大所带来的问题,能够快速完成样本数据的学习训练,更好地适应了网络故障诊断的高时效性,具有很高的实时诊断价值.

(2) 零误差的要求下验证上述结论,在 $\max_epoch = 5000$, $error_goal = 0$ 及 $lr = 0.05$ 的条件下,结果如图 3 所示.

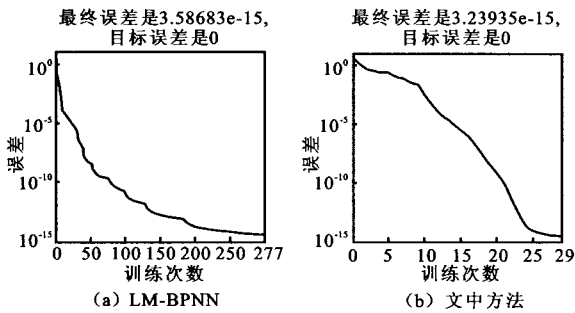


图 3 网络训练误差性能曲线

可以看出,在零误差的高要求下,LM-BP 神经网络和文中方法都能实现收敛,LM-BP 神经网络通过 277 次训练后才能实现收敛,而文中方法只需要 29 次训练就能实现收敛.与误差要求为 0.001 的实验比较后发现,LM-BP 神经网络的训练次数急剧增加,文中方法有增加但次数不多.另外,图 3 的曲线平滑度显示了文中方法发生收敛过度的次数较 LM-BP 神经网络明显减少.

(3) 应用测试集对文中方法的网络进行检测.如故障类型为 nmap 的特征值 $p = [0, 2, 14, 7, 0, 0, \dots, 0.95, 0, 0.95, 1, 0, 0]$ 输入网络,输出的结果为 $t = [0.9881 \ 0.0037 \ 0.0036 \ 0.0079]$,故障诊断准确.由于 BP 神经网络无法实现收敛,仅对用 LM-BP 网络和文中方法训练后收敛的网络进行故障检测正确率的统计如表 2.

可见,文中方法不仅在时效性方面优于 BP 神经网络、LM-BP 神经网络,同时在一定程度上提高

了检测的准确性.

表 2 故障检测的性能对比表

性能 \ 方法	LM-BP	文中方法
检测正确率	86.67%	91.67%

5 结束语

文中基于模糊聚类自适应核区对训练样本进行优选及 LM-BPNN 的网络故障诊断方法,有效解决了原始训练样本的冲突性、冗余性,压缩了数据量,同时克服了 BP 神经网络的自身训练次数多、收敛慢和容易出现振荡的缺陷,提高了算法的时效性,能够较好的满足网络故障的实时诊断要求,并在一定程度上提高了检测的准确性.

参考文献:

[1] 戚涌,李千目,刘凤玉.基于 BP 神经网络的网络智能诊断系统[J].微电子学与计算机,2004,21(10):10-13.

[2] Hagan M T, Menhaj M B. Training feedforward networks with the marquardt algorithm[J]. IEEE Trans on Neural Networks, 1994(5):989-993.

[3] 张宏烈,荆涛,刘彦忠.基于模糊聚类神经网络的化工过程预报模型研究[J].微计算机信息,2006,22(3):253-255.

[4] Deng Qingshan, Mei Guoping. Application of backpropagation neural network based on levenberg-marquardt algorithm in detection of fraudulent financial statements[C]// The 2th International Conference on Computer Science & Education. China, Wuhan, 2007:151-154.

[5] 侯亚丽,李铁.基于 LM 优化算法的 BP 神经网络目标识别方法[J].探测与控制学报,2008,30(1):53-57.

[6] University of California Irvine. UCI KDD archive[EB/OL].[2009-05-11]. <http://kdd.ics.uci.edu/>

作者简介:

蒋静芝 女,(1984-),硕士研究生.研究方向为网络故障诊断.

孟相如 男,(1963-),教授,博士生导师.研究方向为宽带通信网络技术、网络故障诊断.

张立 男,(1981-),博士研究生.研究方向为网络故障诊断、人工智能.