

文章编号:1001-9081(2009)03-0795-03

基于数据挖掘技术的板形缺陷模式识别

赵小燕, 张朝晖

(北京科技大学 信息工程学院, 北京 100083)

(sally_zxy@163.com)

摘要:介绍了建立带钢板形缺陷模式识别的数据挖掘过程。针对普通神经网络识别精度较低的缺陷,提出一种基于分层神经网络进行数据挖掘的新方法。该方法采用二叉树型结构,通过分层来细化预测范围并选用多个神经网络进行递推。实验结果证明了分层神经网络模型比普通神经网络模型的预测精度有较大提高,完全可以满足实际生产需要。

关键词:数据挖掘; 人工神经网络; 板形缺陷; 模式识别; 分层

中图分类号: TP311.13; TP391.4 **文献标志码:** A

Flatness defect pattern recognition with data mining technology

ZHAO Xiao-yan, ZHANG Zhao-hui

(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China)

Abstract: The flatness defect pattern recognition based on data mining technology was proposed. In order to solve low accuracy of normal BP (Back Propagation) network, a novel data mining algorithm based on hierarchical BP model was presented. The new model with binary tree structure reduced prediction range of each network and adopted several networks for degree elevation. Compared with the normal BP model, the new system precision was improved remarkably. The experimental results show this method can meet the requirements of the producing process.

Key words: data mining; artificial neural network; flatness defect; pattern recognition; hierarchy

数据挖掘(Data Mining, DM),也称为数据库知识发现(Knowledge Discovery in Databases, KDD),其目的是发现人们不易觉察的、隐含的模式,从而提高市场决策能力,检测异常模式,以及在过去的经验基础上预言未来趋势等^[1-2]。随着计算机与通信技术的迅猛发展及其在工业生产中的应用,生产现场大量生产过程数据被采集到上位计算机,数据挖掘将从这些数据中获取可用于信息管理、问题求解、判断决策、生产控制的有用知识。由于神经网络具有分布存储、非线性、自组织和学习性等特点,该方法在解决数据挖掘问题时具有一定的优势。通常普通神经网络识别精度较低,本文提出一种基于分层神经网络进行数据挖掘的方法,并将其应用在带钢板形缺陷预测中。实验结果证明了该方法比普通神经网络的预测精度有较大提高,完全满足生产需求。

1 建立带钢板形缺陷的数据挖掘过程

板形是板带产品的主要质量参数,实时监控板带的板形并快速预测出其板形缺陷模式对控制带钢质量非常重要。通常板形缺陷模式比较复杂,很难进行精确的数学描述。在生产中常定义几种简单的板形缺陷作为基本板形模式,如图1所示。图1中列出了带钢内部应力偏差相对应的板形缺陷及轧辊形状的关系,横坐标 x 表示板宽方向,纵坐标 σ 为残余应力。

板形检测装置通常是指装备有多个传感器的板形检测辊,带钢板形缺陷数据挖掘的主要任务和目的是将板形检测装置实测的一组板带横向应力信号 $\sigma_i^T (i = 1, 2, 3, \dots, m)$ 分解成适合板形调控机构控制的几种基本板形缺陷模式,从而为制定合理的板形控制策略提供依据。板形缺陷模式的数

据挖掘过程如图2所示。

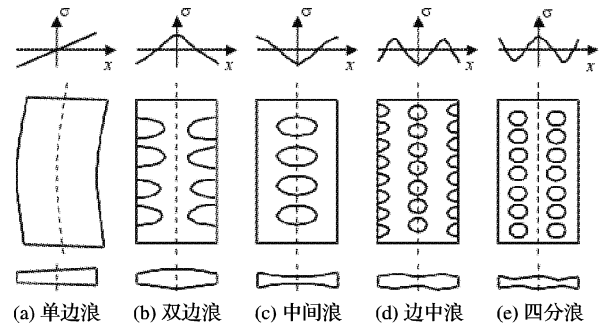


图1 板形基本模式及应力-板形对照图

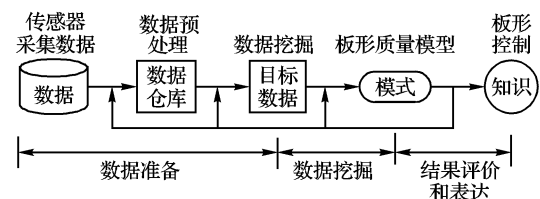


图2 带钢板形缺陷模式的数据挖掘过程

1.1 数据采集

多个传感器分布在板形检测辊的不同位置,检测辊通过与带钢接触将带钢内部的应力转换成传感器的输出,热轧过程检测辊转速通常为 15 r/s,由此大量的测量数据被采集到上位计算机,通过对测量数据的挖掘来预测分析板形缺陷。

1.2 数据处理

数据质量好坏会直接影响所建模型的质量,本文选取实际生产中 400 组中具有不同板形模式的测量数据,其中 300 组作为神经网络训练数据集,另取 100 组数据作为测试数据

收稿日期:2008-09-09;修回日期:2008-11-14。

作者简介:赵小燕(1974-),女,河北邢台人,博士研究生,主要研究方向:检测技术、数据挖掘; 张朝晖(1965-),男,河北保定人,教授,主要研究方向:先进检测技术与信号处理、嵌入式测量仪器。

集,以检测模型的识别精度。

首先将数据进行处理,根据设定的板形目标 σ_i^A 和一组实测的离散应力信号 $\sigma_i^T (i = 1, 2, 3, \dots, m)$ 计算板形应力偏差,并将所有的数据过去噪、剔除错误数据、归一化等预处理。

数据挖掘模型的输入变量、输出变量分别为板形应力偏差 $\Delta \sigma_i^Z$ 和板形模式系数 a_j 。根据 9 段板形检测系统建立的板形应力偏差参数集 P 和板形模式系数集 O 的内容如下:

$$P = [\Delta \sigma_1^Z, \Delta \sigma_2^Z, \Delta \sigma_3^Z, \Delta \sigma_4^Z, \Delta \sigma_5^Z, \Delta \sigma_6^Z, \Delta \sigma_7^Z, \Delta \sigma_8^Z, \Delta \sigma_9^Z]$$

$$O = [a_1, a_2, a_4]$$

上式中: $\Delta \sigma_1^Z \sim \Delta \sigma_9^Z$ 为 9 个测量段所对应板带横向的归一化板形应力偏差信号; $a_1 \sim a_4$ 为不同板形缺陷模式的系数。

1.3 数据挖掘

数据挖掘融合了多个不同学科领域的技术与成果,使得目前的数据挖掘方法表现出多种多样的形式。知识发现类数据挖掘技术主要方法有人工神经网络、支持向量机、决策树、遗传算法、粗糙集、规则发现和关联顺序等^[3-7]。本课题针对带钢板形缺陷预测建模主要采用了人工神经网络方法。

图 1 中的板形基本模式可以采用勒让德正交多项式各项分量表示,如式(1)所示:

$$\begin{cases} P_1(x) = x \\ P_2(x) = \frac{1}{2}(3x^2 - 1) \\ P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3) \end{cases} \quad (1)$$

板形曲线可表示为:

$$\Delta \sigma_i^Z = a_1 P_1(z) + a_2 P_2(z) + a_4 P_4(z) \quad (2)$$

系数 a_1 、 a_2 、 a_4 的正负符号对应单边浪、中间浪、双边浪、四分浪及边中浪等不同的板形缺陷模式,数据挖掘的过程就是从 $\Delta \sigma_i^Z$ 数值中预测出不同的缺陷模式 $P_1(z)$ 、 $P_2(z)$ 、 $P_4(z)$ 及系数 a_1 、 a_2 、 a_4 的大小,从而采取相应的板形控制手段进行调节。

1.4 知识应用

通过数据挖掘所得到的不同板形缺陷模式为板形控制回路提供反馈信号,使板形控制形成闭环。通常板形控制手段有斜辊控制、弯辊控制、冷却控制、基本冷却控制等,以消除不同板形缺陷对带钢质量的影响。

2 基于分层 BP 的板形缺陷模式识别

2.1 普通神经网络建模

目前多种不同结构的神经网络模型用于板形的缺陷模式预测中^[8-9]存在的主要问题有:网络的隐层节点个数和网络的初始权阈值是随机选取的,因此网络结构不是最优;常用的神经网络为多输入-多输出模式,网络结构复杂。

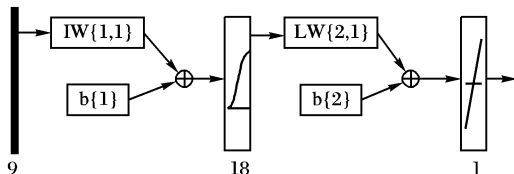


图3 普通神经网络结构

隐层神经元个数的选择直接影响神经网络的训练速度和命中率。通过实验证明隐层神经元数少的网络在训练阶段为了减小误差,需要增加训练次数,训练结束后,虽然模型的误差减小了,可模型却对训练数据集产生过拟合,命中率反而会降低。而隐层神经元数多的网络由于拟合复杂函数的能力更

强,经过相对较少次数的训练,虽然训练数据集没有产生过拟合,但模型的误差相对较大。综合以上考虑选用的神经网络模型如图3所示,输入节点9个,隐层神经元18个,并将此网络结构作为通用网络模型。

理论上三层神经网络就可近似任何函数,因此选择三层网络层数。

2.2 分层神经网络法建模

为提高 BP 神经网络的预测精度,经试验研究发现,当用于建模的训练数据范围减小时,逼近精度有所提高^[10],因此采用二叉树型分层逐渐逼近的思想来构建模型。通过选用二叉树结构将整个预测范围进行分层细化;选用多个神经网络进行递推,减小单个节点模型的预测范围,以提高整个模型逼近的命中率,整个模型结构如图4所示。

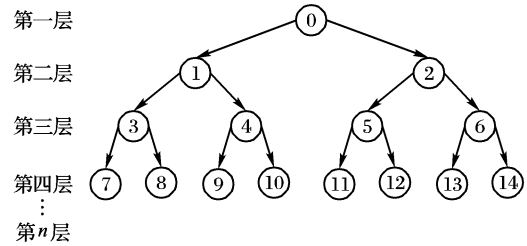


图4 分层神经网络模型结构

图4中每个节点都采用图3所示的网络结构,第一层是在某板形系数值的全范围内建立一个预测模型,第二层开始依次按二叉树结构建模。图4中圆圈数字对应不同预测范围编号,根据优选法中0.618黄金分割定理确定子模型预测范围,各子节点的预测范围如式(3)(4)所示。

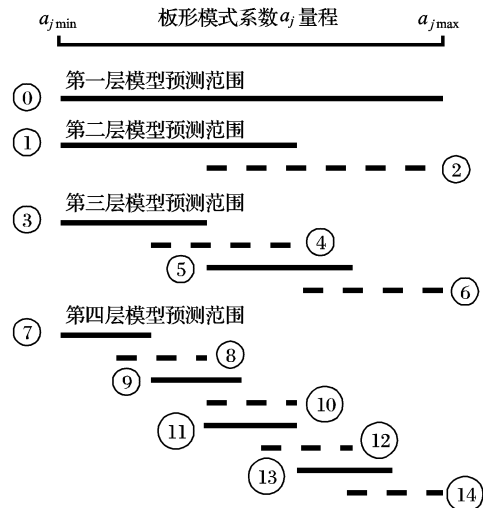


图5 神经网络预测值范围黄金分割

左子树预测范围:

$$\begin{cases} a_{left(n)} = a_{left(n-1)} \\ a_{right(n)} = a_{left(n-1)} + 0.618 \times (a_{right(n-1)} - a_{left(n-1)}) \end{cases} \quad (3)$$

右子树范围:

$$\begin{cases} a_{left(n)} = a_{left(n-1)} + 0.382 \times (a_{right(n-1)} - a_{left(n-1)}) \\ a_{right(n)} = a_{right(n-1)} \end{cases} \quad (4)$$

式中: $a_{left(n)}$ 为第 n 层板形模式系数预测范围的左边界; $a_{right(n-1)}$ 为第 $n-1$ 层板形模式系数预测范围的右边界,如图5所示。其中各子节点的预测范围和其所在层数的关系如式(5):

$$R_n = (0.618)^{n-1} \quad (5)$$

式中: R_n 为神经网络模型第 n 层各节点的预测范围。根据实际应用对板形识别精度要求来确定整个网络的层数。根据式(5)可知,当模型为7层时,最底层模型预测值范围已缩小为

总量程范围的 5.6%, 在此范围内预测命中率已经很高。

用板形应力偏差集 P 和板形模式系数集 O 数据对分层网络模型进行训练, 并将以上各层子模型组合起来, 即构成板形模式数据挖掘总模型。

3 实验结果

根据来自生产实际中 300 组数据对分层 BP、普通 BP 进行训练。在网络模型训练过程中, 首先将待预测板形应力偏差集 P 送入第一层编号为 0 的模型进行预测, 将预测结果与该模型预测范围中心点值进行比较, 若该节点的模型预测值小于中心点值, 则将 P 送入该节点的下一层左子节点模型再进行预测; 若大于中心点值, 则将 P 送入该节点的下一层右子节点模型进行预测, 每个子模型预测值范围中心点的计算方法如式(6):

$$x_{mid(k)} = \frac{a_{left(k)} + a_{right(k)}}{2} \quad (6)$$

式中: $x_{mid(k)}$ 为编号为 k 的子模型的左、右边界中点; $a_{left(k)}$ 为编号为 k 的子模型的左边界点; $a_{right(k)}$ 为编号为 k 的子模型的右边界点。

如此重复上述过程, 板形应力偏差集 P 不断被送入更深一层模型进行预测, 随着模型深度的增加, 预测命中率不断提高。将 100 测试样本输入训练好的两个模型, 普通神经 BP 模型的预测结果如表 1 所示; 当采用 7 层分层 BP 模型时, 其预测结果如表 2 所示。

表 1 普通 BP 模型板形预测结果

板形模式系数	预测值偏差 $err/\%$			
	$ err \leq 5$	$5 < err \leq 10$	$10 < err \leq 20$	$ err > 20$
a_1	62	17	15	6
a_2	59	16	16	9
a_4	52	21	15	12

表 1 中: $|err|$ 表示相对误差的绝对值, 不同板形模式系数的误差分布不同, 其相对误差主要集中在 $\pm 5\%$ 范围内, 其中 a_1 的 $\pm 5\%$ 命中率为 62%, a_2 次之, a_4 的 $\pm 5\%$ 命中率最低为 52%; 各板形模式系数 10% 的命中率都在 80% 左右。虽然理论上三层神经网络可以逼近任意非线性, 但由于训练样本有限, 实际上命中率随着勒让德多项式阶次的增加其系数 a_i 的逼近精度逐渐降低, a_4 有 12% 的测试样本偏差超过 20%。由此可见, 如果单纯采用三层神经网络, 其逼近的精度不太理想。

采用分层的方法建立新的模型, 将 100 组测试数据送入训练好的模型中进行预测, 其预测结果如表 2 所示。

表 2 分层 BP 模型板形预测结果

板形模式系数	预测值偏差 $err/\%$			
	$ err \leq 5$	$5 < err \leq 10$	$10 < err \leq 20$	$ err > 20$
a_1	75	16	7	2
a_2	72	15	8	5
a_4	68	21	8	3

从表 2 可以看出采用分层法建立的神经网络模型 a_1 的误差在 $\pm 5\%$ 以内的命中率为 75%, 比普通三层神经网络有了很大提高; 各模型系数的预测精度都有很大提高, a_4 的 10% 命中率从表 1 中的 73% 达到 89%, 完全可以满足生产需求。

参考文献:

- [1] FAYYAD U, SHAPIRO P, UTHURUSAMY S. Advances in Knowledge Discovery and Data Mining[M]. Cambridge, MA: MIT Press, 1996.
- [2] 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.
- [3] 鲁红英, 肖思和. 基于改进的遗传神经网络数据挖掘方法研究[J]. 计算机应用, 2006, 26(4): 878-879.
- [4] SHIN C K, KIM H K, PARK S C, *et al.* A hybrid approach of neural network and memory-based learning to data mining[J]. IEEE Transactions on Neural Networks, 2000, 11(3): 637-646.
- [5] MOHAMADI H, HABIBI J, ABADEH M S, *et al.* Data mining with a simulated annealing based fuzzy classification system[J]. Pattern Recognition, 2008, 41(5): 1824-1833.
- [6] WU, R-C, CHEN R-S. The application of data mining technology for intelligent product quality analysis improvement system[J]. WSEAS Transactions on Information Science and Applications, 2007, 4(7): 693-699.
- [7] HSIES K-L. Applying data mining techniques into achieving process improvement[J]. WSEAS Transactions on Systems, 2006, 5(12): 2774-2780.
- [8] 张秀玲, 刘宏民. 变结构神经网络在板形信号模式识别方面的应用[J]. 钢铁研究学报, 2001, 13(2): 62-66.
- [9] 张材, 谭建平. 基于遗传算法-反向传播模型的板形模式识别[J]. 中南大学学报: 自然科学版, 2006, 37(2): 294-299.
- [10] 王健, 张香燕, 张乃尧. 一类二叉树型分层模糊系统的等效性条件[J]. 清华大学学报: 自然科学版, 2007, 47(7): 1237-1240.

(上接第 780 页)

参考文献:

- [1] 杨桃, 刘湘南, 张柏, 等. 基于多特征空间的遥感信息自动提取方法[J]. 吉林大学学报: 地球科学版, 2005, 35(2): 257-260.
- [2] 李述, 刘勇. 基于多特征的遥感影像土地利用/覆盖分类——以腾格里沙漠东南边缘地区为例[J]. 遥感技术与应用, 2006, 21(2): 154-158.
- [3] 罗亚, 徐建华, 岳文泽, 等. 植被指数在城市绿地信息提取中的比较研究[J]. 遥感技术与应用, 2006, 21(3): 212-219.
- [4] 甄计国, 王义德, 赵军. 兴隆山国家级自然保护区的植被指数及其变化特征[J]. 遥感技术与应用, 2006, 21(4): 294-301.
- [5] JOHN G L, DING YUAN, ROSS S L, *et al.* A change detection experiment using vegetation indices[J]. Photogrammetric Engineering & Remote Sensing, 1998, 64(2): 143-150.
- [6] 王正兴, 刘闯, ALFREDO H. 植被指数研究进展: 从 AVHRR-NDVI 到 MODIS-EVI[J]. 生态学报, 2003, 23(5): 980-987.
- [7] 罗亚, 徐建华, 岳文泽. 基于遥感影像的植被指数研究方法述评[J]. 生态科学, 2005, 24(1): 75-79.

- [8] QI J, CHEHBOUNI A, HUETE A R, *et al.* A modified soil adjusted vegetation index[J]. Remote Sensing of Environment, 1994, 48: 119-126.
- [9] 徐希蘅. 遥感物理[M]. 北京: 北京大学出版社, 2005.
- [10] 赵英时. 遥感应用分析原理与方法[M]. 北京: 科学出版社, 2003.
- [11] 胡良军, 邵明安. 论水土流失研究中的植被覆盖度量指标[J]. 西北林学院学报, 2001, 16(1): 40-43.
- [12] 张爽, 刘雪华, 靳强. 决策树学习方法应用于生境景观分类[J]. 清华大学学报: 自然科学版, 2006, 46(9): 1564-1567.
- [13] 祁治. 基于模糊集理论的关联规则研究[D]. 武汉: 武汉科技大学, 2005.
- [14] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.
- [15] 张青. 决策树分类算法的研究与改进[D]. 郑州: 郑州大学, 2002.
- [16] 秦文. 分类技术中的决策树算法分析[J]. 深圳信息职业技术学院学报, 2004, 2(1): 54-58.