

文章编号:1671-251X(2011)11-0075-03

DOI:CNKI:32-1627/TP.20111102.1625.022

近红外光谱煤质分析模型中异常样品的剔除方法

王建义¹, 雷萌²

(1. 大屯煤电(集团)有限责任公司孔庄煤矿, 江苏 徐州 221611;

2. 中国矿业大学信电学院, 江苏 徐州 221116)

摘要:针对建立近红外光谱煤质定量分析模型时训练集中的异常样品严重影响模型预测精度的问题,提出一种二次诊断法剔除异常样品:利用模糊C均值聚类法对样品进行聚类,得到可疑样品;将可疑样品作为验证集,通过PCA-GA-BP模型进行二次诊断,剔除异常样品。实验对比了训练集中异常样品剔除前后,模型对15组待测样品的预测能力,结果表明该方法能够准确剔除异常样品,并有效提高模型的预测精度。

关键词:煤质分析; 近红外光谱; 异常样品; 二次诊断; 模糊C均值聚类; PCA-GA-BP模型

中图分类号:TD672 文献标识码:A 网络出版时间:2011-11-02 16:25

网络出版地址: <http://www.cnki.net/kcms/detail/32.1627.TP.20111102.1625.022.html>

Rejecting Method of Abnormal Samples in Analysis Model of Coal Quality of Near-infrared Spectrum

WANG Jian-yi¹, LEI Meng²

(1. Kongzhuang Coal Mine of Datun Coal and Electricity (Group) Co., Ltd., Xuzhou 221611, China.

2. School of Information and Electrical Engineering of CUMT., Xuzhou 221116, China)

Abstract: In building of quantitative analysis model of coal quality of near-infrared spectrum, abnormal samples in training set seriously influence forecast precision of the model. So the paper proposed a new method of rejecting abnormal samples by twice diagnoses: get suspicious samples by clustering training samples based on fuzzy C-mean algorithm, and taking suspicious samples as validation set, reject abnormal samples by secondary diagnosis through PCA-GA-BP model. An experiment was done to compare predicting ability of the model with and without abnormal samples in training set by 15 groups of samples, and the result showed that the method can reject abnormal samples accurately, and improve forecasting precision effectively.

Key words: coal quality analysis, near-infrared spectrum, abnormal samples, twice diagnoses, fuzzy C-mean algorithm, PCA-GA-BP model

0 引言

随着化学计量学的发展,近红外光谱分析技术被广泛应用于食品、药品、石油等的品质检测、产地鉴别等方面^[1-4]。传统煤质测定方法普遍存在分析时间长、测定精度低、破坏样品结构等缺陷^[5],而近红外光谱分析技术可以快速、无损地对煤炭中各成分的含量进行检测,它将会成为煤质分析未来发展

的主要趋势。

近红外光谱分析技术是一种间接分析技术,其分析结果主要依靠预测模型的准确性和稳定性。煤炭样品的光谱图和标准值的相关性直接影响模型的预测能力,异常样品对分析模型的干扰是最为严重的。产生异常样品的原因:(1)测量仪器、测试方法和环境等因素的影响;(2)技术人员实验操作失误;(3)煤炭的复杂性、多样性。在实际参与训练定量分析模型的样品中常常混有问题样品,因此在训练模型之前,要先对训练集样品进行逐一排查,剔除异常样品。本文首先利用模糊C均值聚类法(Fuzzy

收稿日期:2011-07-04

作者简介:王建义(1962-),男,河北石家庄人,长期从事煤矿机电设备管理工作。E-mail:lrh1960@vip.sina.com

3 PCA-GA-BP 模型二次诊断过程

PCA 即主成分分析,其目的就是把光谱数据降维^[9]。前 5 个主成分的累计可信度(%)分别为 96.191 0、99.289 3、99.645 9、99.813 4、99.882 8,当累加到第 5 个主成分值时,累计可信度接近于 100,也就是说前 5 个主成分几乎包含了所有的光谱信息。然后建立一个典型的三层 BP 神经网络,将前 5 个主成分作为输入向量,利用 GA(遗传算法)的全局搜索能力,搜索 BP 神经网络全局最优的初始权值和阈值,构建一个具有较高预测精度和较快收敛速度的近红外光谱煤质分析模型^[10-11]。

初次诊断后,用 58 组正常的煤炭样品训练 PCA-GA-BP 模型,12 组可疑样品作为验证集。预测结果如表 2 所示,第 18、31、45、52、53、54、59 组样品的预测值与标准测量值的误差较大(>0.03),属于异常样品,应将其剔除,而剩余的 5 组样品应归入训练集样品中。因此,最终适合参与近红外光谱煤质分析建模的训练集样品有 63 组。

表 2 可疑样品测量值与预测值

组别	测量值	预测值	误差值
11	0.248	0.231	0.017
18	0.253	0.161	0.092
27	0.234	0.25	0.016
31	0.053	0.291	0.238
33	0.05	0.077	0.027
40	0.087	0.098	0.011
45	0.084	0.117	0.033
49	0.062	0.071	0.009
52	0.106	0.143	0.037
53	0.257	0.07	0.187
54	0.08	0.141	0.061
59	0.045	0.217	0.172

异常样品剔除前后,各验证集样品的预测值与标准测量值的误差如图 3 所示。

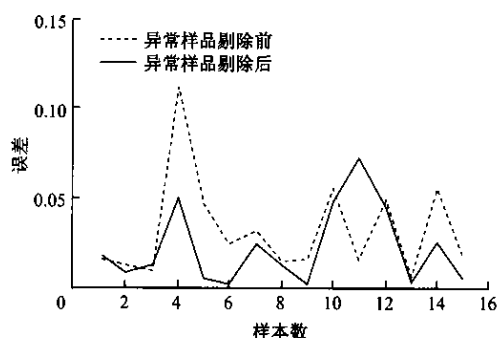


图 3 验证集样品预测值与标准测量值的误差

从图 3 可以看出,大部分样品预测值的精度较异常样品剔除前都有显著的提高。

4 结语

在建立近红外光谱煤质分析模型时,预测值与标准值之间决定系数的大小反映了模型的准确性及可靠性。实验证明,二次诊断法可以剔除异常样品,减少对异常样品的误判,有效提高模型的准确性,为剔除异常样品提供了一个有效、可靠的新方法。由于 FCM 算法对大样本聚类时将耗费大量的存储空间和运算时间,且有时会收敛到局部极小点上,因此在剔除大样品中的异常样品时,需进一步改进聚类算法。

参考文献:

- [1] CEN Haiyan, HE Yong, HUANG Min. Measurement of Soluble Solids Contents and PH in Orange Juice Using Chemometrics and Vis-NIRS[J]. Journal of Agricultural and Food Chemistry, 2006, 54(20): 7437-7443.
- [2] 王加华, 韩东海. 基于遗传算法的苹果糖度近红外光谱分析[J]. 光谱学与光谱分析, 2008, 28(10): 2308-2311.
- [3] 陈永明, 林萍, 何勇. 基于遗传算法的近红外光谱橄榄油产地鉴别方法研究[J]. 光谱学与光谱分析, 2009, 29(3): 671-674.
- [4] 陆婉珍, 袁洪福, 徐广通, 等. 现代近红外光谱分析技术[M]. 北京: 中国石化出版社, 2000.
- [5] ANDRES J M, BONA M T. Analysis of Coal by Diffuse Reflectance Near-infrared Spectroscopy [J]. Analytica Chimica Acta, 2005, 535(1-2): 123-132.
- [6] 张双全, 吴国光. 煤化学[M]. 徐州: 中国矿业大学出版社, 2004.
- [7] 李雷, 罗红旗, 丁亚丽. 一种改进的模糊 C 均值聚类算法[J]. 计算机技术与发展, 2009, 19(12): 71-73.
- [8] WANG Weina, ZHANG Yunjie, LI Yi, et al. The Global Fuzzy C-Means Clustering Algorithm [C] // Proceedings of the 6th World Congress on Intelligent Control and Automation, 2006, Dalian: 3604-3607.
- [9] 何晓群. 多元统计分析[M]. 北京: 中国人民大学出版社, 2004.
- [10] 靳建彬, 王元钦, 陈源. 基于遗传算法的 BP 神经网络优化策略研究[J]. 计算机与现代化, 2010(9): 88-91.
- [11] 李伟超, 李志刚, 杨旭海. 遗传算法在人工神经网络中的应用[J]. 电子测量与仪器学报, 2008(增刊 2): 170-173.