

文章编号:1671-251X(2016)10-0060-05DOI:10.13272/j.issn.1671-251x.2016.10.014

王敏. K 均值改进留一校验法在煤炭近红外光谱异常样本剔除中的应用研究[J]. 工矿自动化, 2016, 42(10): 60-64.

K 均值改进留一校验法在煤炭近红外光谱异常样本剔除中的应用研究

王敏

(山西潞安环保能源开发股份公司 王庄煤矿, 山西 长治 046031)

摘要:针对现有留一校验法存在剔除异常样本耗时长、误判的缺陷,提出一种 K 均值改进留一校验法,并将其用于煤质分析中异常样本的检测与剔除。该方法首先利用 K 均值聚类法对样本进行聚类,得到可疑样本;然后将可疑样本作为验证集,通过留一校验法进行二次判别,剔除异常样本。实验结果表明, K 均值改进留一校验法能快速、准确剔除异常样本,提高了模型的预测精度。

关键词:煤质; 近红外光谱分析; 异常样品; K 均值聚类; 留一校验法
中图分类号:TD67文献标志码:A网络出版时间:
网络出版地址:

Application research of improved K-means leave one out method in rejecting of abnormal samples of coal near infrared spectrum

WANG Min

(Wangzhuang Coal Mine, Lu'an Environmental Protection and Energy Development Co., Ltd., Changzhi 046031, China)

Abstract:In view of problems of time-consuming, misjudgment of rejecting abnormal sample existed in current leave one out method, an improved K -means leave one out method was put forward for detecting and eliminating abnormal sample in coal quality analysis. Firstly, the method uses K -means clustering method to cluster samples, and gets suspicious samples; then it takes suspicious samples as a validation set, adopts leave one out method to do quadratic distinguishing, so as to eliminate abnormal samples. The experimental results show that the K -means leave one out method can eliminate abnormal samples quickly and accurately, and improves prediction accuracy of models.

Key words:coal quality; near infrared spectral analysis; abnormal samples; K -means clustering; leave one out method

0 引言

在煤炭样品近红外无损检测中,分析结果的可靠性首先取决于原始数据的准确性,即煤炭样品的光谱图和化学标准值。煤炭近红外光谱异常样本明

显偏离光谱数据集主体分布,在煤质定量分析过程中,严重影响模型预测精度^[1-2]。因此,需要对异常样本进行研究,改善其光谱数据质量。

目前,异常样本剔除的常用方法有马氏距离等距离判别分析法、拉依达等误差准则法、留一校验法

收稿日期:2016-06-29;修回日期:2016-08-26;责任编辑:张强。
基金项目:江苏省自然科学基金资助项目(BK20140215)。
作者简介:王敏(1984-),男,山西长治人,工程师,现主要从事煤矿机电技术及管理等工作,E-mail:wm1512004@163.com。

等,异常样本一次判别法普遍存在易将正常样本误判为异常样本的问题^[3-4]。留一估计理论上是无偏估计,在实现原理上效果是最佳的,但留一校验法中每个样本均为可疑样本,确定其相对误差需对样本反复训练 N 次,运算量很大^[5]。本文针对留一校验法剔除异常样本耗时长,且存在误判的缺陷,提出一种 K 均值改进留一校验法(Improved Leave One Out Method by K -means,简称 K -means LOO),并将其用于煤质分析中异常样本的检测与剔除。该方法利用 K 均值聚类法对样本进行聚类,得到可疑样本;将可疑样本作为验证集,通过留一校验法进行二次判别,剔除异常样本。

1 煤炭样本选取

在某国家实验室,来自不同地区的煤炭样本中选取具有代表性的146个样本,严格按照GB 474—2008《煤样的制备方法》要求进行制样,粒度均达到0.2 mm级别。在煤炭样本的工业分析中,水分、灰分、挥发分和全硫分含量的测定均严格遵照GB 474—2008规定的步骤和要求,采用Antaris II傅里叶变换近红外光谱仪采集煤炭样本的近红外光谱,其具体参数:扫描范围为3 800~10 000 cm^{-1} ,波长点数为1 609,分辨率为4 cm^{-1} 。对每个煤炭样本重复扫描64次,并取平均光谱作为所测样品的原始光谱。所选煤炭样本的近红外光谱如图1所示。

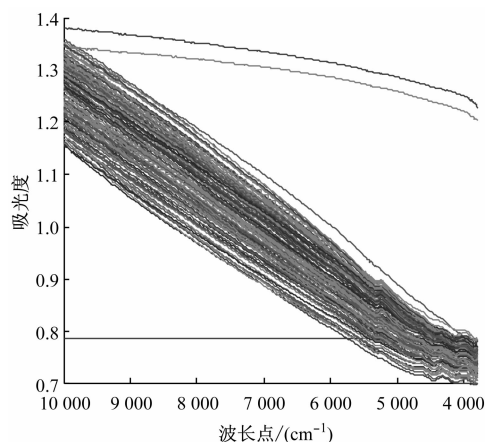


图1 煤炭样本的近红外光谱

2 改进的留一校验法及实验

实验采集146个煤炭近红外光谱样本,由于单一留一校验法中每个样本均为可疑样本,故需对分

类器反复训练146次,耗时长且存在误判的可能,加之异常样本明显偏离光谱正常样本主体,聚类分析法可快速找出可疑样本集,缩小异常样本的搜索范围,基于此,本文提出一种基于 K 均值聚类法和留一校验法相结合的改进留一校校验法。

K 均值聚类法从数据集中随机选取 K 个点作为初始聚类中心,先将样本划分到离聚类中心最近的类,计算各类的均值并作为新聚类中心,更新样本类别划分情况,直至聚类中心没有变化,聚类准则函数收敛。再将 K 均值聚类法与留一校验法相结合,就构成了 K -means LOO。

利用 K -means LOO筛选异常样本的具体过程如下:

输入:将数据集 $M=[X_{1 \times p}, Y]$ 输入模型,其中 $X_{1 \times p}$ 为煤样光谱数据集, Y 为煤样化学测量标准值。

Step1:利用 K 均值聚类法对煤炭光谱样本进行分类,设置分类数 $k=10$ 。

Step2:完成首次筛选。根据正常样本相对集中、异常样本相对分散的原则,将分类结果中小于10的类作为可疑样本类。

Step3:将可疑样本作为预测样本,分别通过留一校验法和BP神经网络算法,得出预测结果。

Step4:设置相对误差阈值 δ ,当 $\delta > 0.5$ 时,判断为异常样本,应予剔除,反之则作为正常样本保留。

3 实验结果与分析

3.1 K 均值聚类结果

将146个煤样数据分成10类,分类结果见表1。根据正常样本相对集中这一特点,可以认为第1、4、8、9四类样本中含有异常样本,即将编号为16、17、18、20、28、29、42、43、66、69、87、92、93、100、104、107、108、110、113、117、118、130、136、139的24个样本作为可疑样本,进行二次判别。

3.2 K -means LOO实验结果

在可疑样本集中每次选取1个样本作为预测样本,其余145个为训练样本,建立BP神经网络模型,重复24次,得到24组可疑样本化学测量值与预测值的相对误差 δ 。实验结果如图2所示,基于留一校验法的一次判别结果如图2(a)所示,判别结果是编号为17、18、23、32、71、87、92的样本为异常样本,应予剔除;基于 K -means LOO的判别结果如

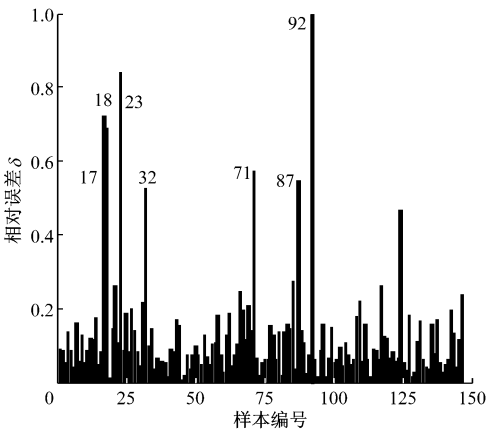
表 1 K 均值聚类法分类结果

类别 编号	样本编号
1	20,66,69,93,100,107,108,117
2	1,15,24,51,56,58,65,68,71,73,74,76,79,86,143
3	2,13,23,32,33,44,46,49,50,55,57,67,77,81,88,101, 102,103,105,106,109, 112,115,116,120,121,122,124, 132,137,138,141,142,146
4	87,104,110,113,130,136,139
5	10,11,12,14,27,38,48,59,60,63,72,90,126,133,140, 144,145
6	9,21,22,25,34,45,53,61,70,96,99,114,119,123,125
7	3,4,6,8,19,30,31,37,41,47,52,54,62,75,78,84,85, 127,135
8	16,28,29,42,43,92,118
9	17,18
10	5,7,26,35,36,39,40,64,80,82,83,89,91,94,95,97,98, 111, 128, 129,131,134

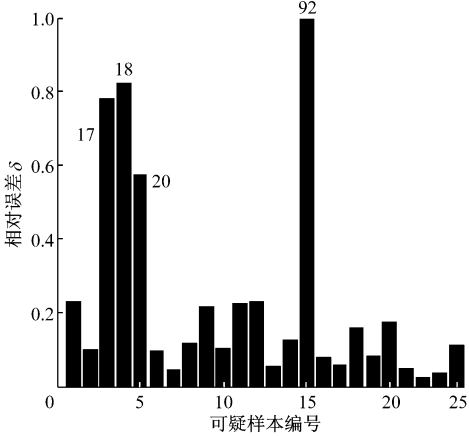
图 2(b)所示,判定 17、18、20、92 样本为异常样本,应予剔除。在实验过程中,留一校验法剔除异常样本用时 215.75 s,K-means LOO 用时 47.00 s。K-means LOO 大幅度减少了判别时间,且为实验保留了较多的样本数据。

将留一校验法和 K-means LOO 剔除异常样本后的光谱数据和化学测量值作为输入和输出,分别建立 3 层 BP 神经网络模型和 PLS(Partial Least Squares,偏最小二乘)模型,将 K-means LOO 中未被剔除的 19 个组可疑样本作为验证集。异常样品

剔除前、留一校验法剔除后和 K-means LOO 剔除后,PLS 模型和 BP 神经网络模型的输出结果,即煤样的水分、灰分、挥发分和全硫分的预测误差如图 3 所示,均方根误差 RMSE 和相关系数 R 见表 2。



(a) 基于留一校验法的一次判别结果



(b) 基于 K-means LOO 的判别结果

图 2 2 种方法的异常样本判别结果

表 2 留一法改进前后 BP 神经网络模型和 PLS 模型的均方根误差和相关系数

评价参数	BP 神经网络模型			PLS 模型		
	原始值	留一校验法	K-means LOO	原始值	留一校验法	K-means LOO
RMSE	0.046 925	0.028 127	0.017 338	0.047 087	0.030 835	0.019 975
R	0.928 666	0.982 336	0.983 195	0.931 044	0.984 203	0.980 958

由图 3 和表 2 可知,剔除异常样本后,煤样的水分、灰分、挥发分和全硫分的预测误差大幅降低。BP 神经网络模型的均方根误差由 0.046 925 降低至 0.03 以下,相关系数由 0.928 66 升高至 0.98 以上,经留一校验法处理后,RMSE 降至 0.028 127,较异常样本剔除前降低了约 40.4%,经 K-means LOO 处理后,RMSE 降至 0.017 338,较异常样本剔除前降低了约 63.8%;PLS 模型的均方根误差由

0.047 087 降低至 0.031 以下,相关系数由 0.931 044 升高至 0.98 以上,经留一校验法处理后,RMSE 降至 0.030 835,较异常样本剔除前降低了约 34%,经 K-means LOO 处理后,RMSE 降至 0.019 975,较异常样本剔除前降低了约 57.4%。上述结果表明,K-means LOO 可有效剔除建模中的异常样本,较传统的留一校验法,模型的预测精度得到大幅提高。

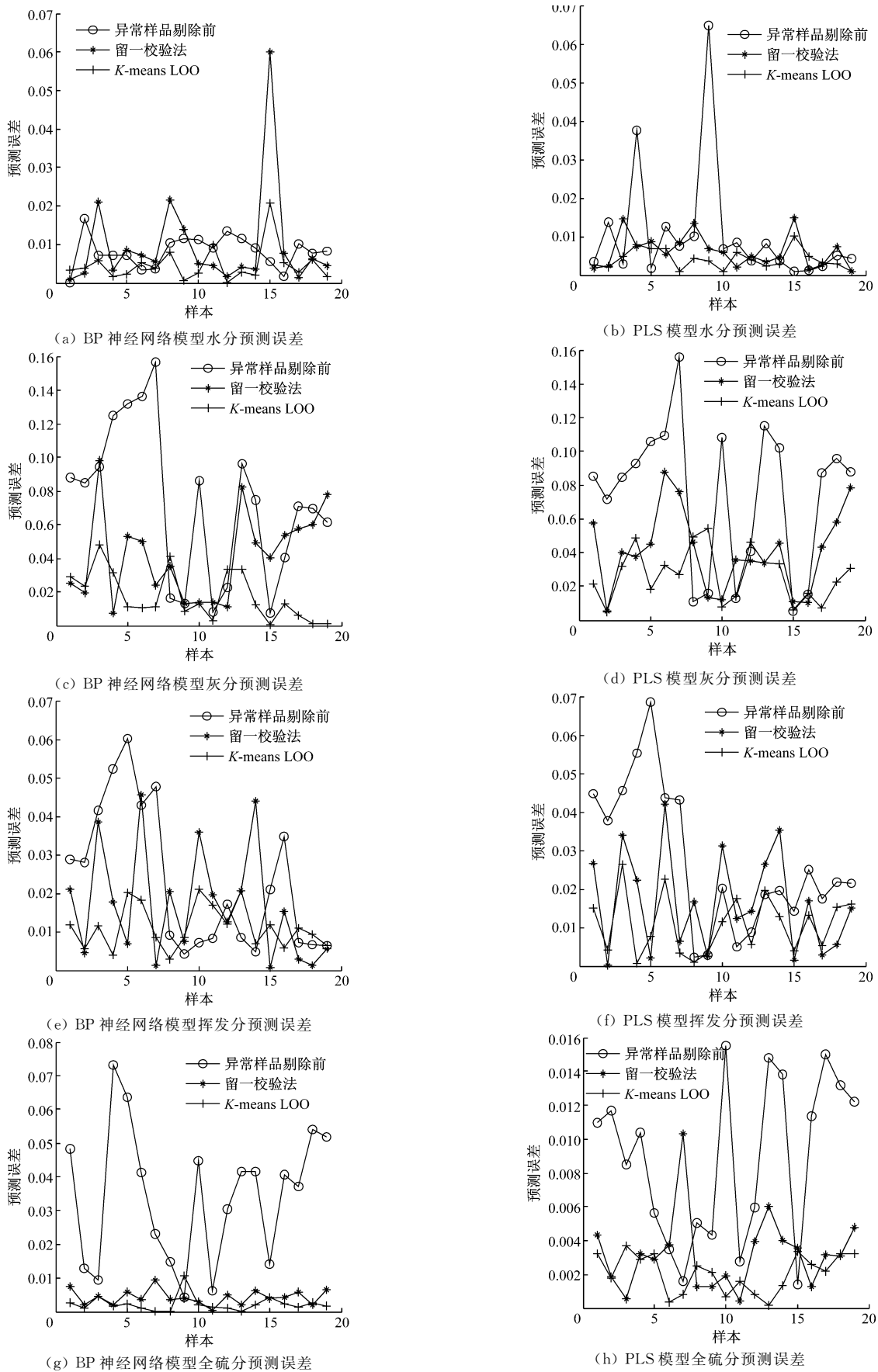


图3 模型预测效果

4 结语

在煤质近红外光谱分析中,异常的光谱样本严重影响定量分析模型的预测精度。针对单一留一校验法去除光谱异常样本的不足,提出了一种 K 均值聚类法与留一校验法相结合的改进留一校验法,并进行了实验研究。实验结果表明,K 均值改进留一校验法能快速、准确剔除煤炭近红外光谱异常样本,提高了模型的预测精度。

参考文献:

[1] 雷萌.基于机器学习的煤质近红外光谱分析[D]. 徐州: 中国矿业大学,2011.

[2] 赵振英,林君,张怀柱. 近红外光谱法分析油页岩含油率中异常样品识别和剔除方法的研究[J]. 光谱学与光谱分析,2014,34(6):1707-1710.

[3] HUANG Z R, SHA S. Feasibility study of near infrared spectroscopy with variable selection for non-destructive determination of quality parameters in shell-intact cottonseed [J]. Industrial Crops and Products, 2013, 43(1): 654-660.

[4] LEWIS A T, JONES K,LEWIS K E, et al. Detection of Lewis antigen structural change by FTIR spectroscopy[J]. Carbohydrate Polymers,2013,92(2): 1294-1301.

[5] 刘翠玲,胡玉君,吴胜男,等. 近红外光谱奇异样本剔除方法研究[J]. 食品科学技术学报,2014,32(5):74-79.