第36卷，第9期
2016年9月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 36, No. 9, pp2793-2797
September, 2016

基于LVQ与SVM算法的近红外光谱煤产地鉴别

李 明，陈 凡，雷 萌*，李 翠

中国矿业大学信息与电气工程学院，江苏 徐州 221116

摘 要 传统煤产地鉴别方法一般以发热量、挥发分、粘结指数、哈氏可磨指数和坩埚膨胀序数作为分类指标，过程复杂耗时较多、耗费巨大的人力、物力并且无法直接快速的得到煤样产地等问题，借助近红外光谱技术快速无损检测的优势，利用基于SVM的留一算法对光谱数据集进行异常样本剔除，得到包含正确光谱信息的煤样光谱数据集，构造基于SVM算法与LVQ算法的定性分析模型，完成基于近红外光谱分析技术的煤产地的快速鉴别，无需对煤样的各种指标进行汇总并且人为预测。针对SVM分析模型中存在随机参数优化问题，引入PSO算法对SVM模型中的损失参数C和核函数半径g进行改进，得到最优参数，最后引入计算准确率的方法对比以上模型并进行评价分析。实验一共收集了加拿大、俄罗斯、澳大利亚、印度尼西亚、中国内蒙等5个地区的煤样光谱数据集，数据集共计305组煤样样本，其中异常样本共计10组，分别选择各国煤炭光谱的前31组作为训练样本，后6组数据作为测试样本，结果表明各分类模型的分类准确率均能达到75%以上，其中基于PSO算法改进的SVM分析模型的准确率可达到96.67%，仅一个样本出现问题，可快速高效地实现基于近红外光谱分析技术的煤产地的鉴别。

关键词 煤产地鉴别；近红外光谱；SVM；LVQ；PSO

中图分类号：TP18 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2016)09-2793-05

引 言

地质变化过程漫长复杂，由于多种地质因素的干扰，煤矿内煤炭成分也千差万别。造成煤炭的多样化的原因很多，主要是由成煤原始物质、成煤年份、还原程度和成因类别上的差异，再加上各种变质作用的影响[1]。

传统煤产地鉴别方法复杂耗时较多，并且要使用特定的实验仪器，如马弗炉、坩埚等，造价昂贵，操作复杂[2]。快速煤成分分析方法，如γ射线法[3]和微波加热法[4]，都只能对单项指标进行测量。陈鹏强[5]等利用近红外光谱分析技术煤炭品质进行定量分析，但是需要按照成分指标对煤炭产地进行鉴别，不能直接测出产地。因此寻找一种快速高效煤产地鉴别方法，是十分必要的。

近红外光谱分析技术[6]应用在煤产地鉴别领域尚属空白。抽取加拿大、俄罗斯、澳大利亚、印度尼西亚和中国内蒙五国港口煤炭样本，并使用近红外光谱分析技术，建立LVQ和SVM分类模型，用PSO算法优化SVM参数。

1 实验部分

1.1 数据

依照国家GB/T213—2008标准，利用Antaris Ⅱ傅立叶变换近红外(FT-NIR)光谱仪测得煤样光谱，化学计量学软件TQ Analyst7.1做光谱处理和软件Matlab编程实现煤炭分类。以上过程均符合国家要求，可保证实验的正确率和准确率。

1.2 LVQ算法

学习向量量化(learning vector quantization，LVQ)[7]神经网络是根据Kohonen竞争算法演变而来。LVQ神经网络由3层组成，分别是输入层、竞争层和输出层神经元。学习算法可以分为两类，分别为LVQ1和LVQ2学习算法。

LVQ1学习算法是根据输入向量的固有结构进行数据压缩的技术。该算法的计算过程如下：首先通过距离公式找到距离输入向量最近的竞争神经元，从而进一步找到与竞争神经元相连接的输出层神经元，如果输入向量的类别和输出层神经元的类别相同，那么相对应竞争层神经元的权值朝输入

收稿日期：2015-04-28，修订日期：2015-08-16

基金项目：国家自然科学基金项目(51304194)，江苏省自然科学基金项目(BK20140215)，中国博士后科学基金项目(2014M551695)资助

作者简介：李 明，1962年生，中国矿业大学信息与电气工程学院教授 e-mail：liming@cumt.edu.cn

*通讯联系人 e-mail：leimengniee@163.com

神经元方向调整，若输入向量和输出层向量不一样，则朝反方向调整。其中计算输入向量和竞争层神经元距离公式如下

$$d_i = \sqrt{\sum_{j=1}^{N}(x_i - w_{ij})^2} \qquad i = 1, 2, \cdots, S^1$$

其中 w_{ij} 为竞争层神经元与输入层神经元之间的权值。

正方向调整权值的公式如下

$$w_{ij\_new} = w_{ij\_old} + \eta(x - w_{ij\_old})$$

反方向调整调整权值的公式如下

$$w_{ij\_new} = w_{ij\_old} - \eta(x - w_{ij\_old})$$

其中 η 是学习率，x 为输入向量。

LVQ2 算法类似于 LVQ1 算法，只是 LVQ1 算法，只有一个竞争层神经元可以获胜，而 LVQ2 算法引入了“次获胜”神经元，使获胜神经元和“次获胜”神经元的权值都得到调整。

1.3 SVM 算法

支持向量机(support vector machine，SVM)[8-9]，主要建立一个超平面作为分类决策面。对 SVM 模型，建模方法一般分为下面几个步骤：首先输入要训练的各国煤样光谱数据 $T=\{(x_1, y_1), \cdots, (x_n, y_n)\}$，其中，$x_i \in R^n$ 为输入的光谱，y_i 为要分的国家类别。然后选取适当的核函数 $K(x_i, x_j)$和损失参数 C，本工作选径向基核函数

$$K(x, x_i) = \exp(-\gamma \| x - x_i \|^2), \gamma > 0$$

构造求解最优化问题

$$\min \frac{1}{2}\sum_{i=1}^{j}\sum_{j=1}^{n} y_i y_j a_i a_j K(x_i, x_j) - \sum_{j=1}^{n} a_j$$

约束条件为

$$\sum_{i=1}^{n} ya_i = 0, 0 \leqslant ai \leqslant C, i = 1, \cdots, l$$

其中 a_i，a_j 是拉格朗日乘子，得到最优解 a^*，取 a^* 的一个正分量，计算阈值 b。

$$b = y_i - \sum_{i=1}^{n} y_i a_i^* K(x_i, x_j)$$

既而，可得到决策函数

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{n} a_i^* y_i K(x, x_i) + b)$$

由于 SVM 模型参数选择存在人为误差，无法凭经验使参数达到最优，针对该问题引入粒子群算法(particle swarm optimization，PSO)[10-11]。PSO 是基于社会群体的思想。在 M 维空间中，每个粒子 i 的位置可以表示为 $x_i = (x_{i1}, x_{i2}, \cdots, x_{iM})$代表 SVM 参数大小，共有 n 个粒子组成群组 $X = (X_1, X_2, \cdots, X_n)$，每个粒子代表一个潜在 SVM 最优参数。其中第 i 个粒子的速度向量为 $v_i = (v_{i1}, v_{i2}, \cdots, v_{iM})$，根据目标函数可以计算出粒子的适用度即分类准确率，粒子 i 搜索解空间时，保存其搜索到的最优参数 $p_i = (p_{i1}, p_{i2}, \cdots, p_{iM})$。和群体的最优参数 $p_g = (p_{g1}, p_{g2}, \cdots, p_{gM})$，每一次迭代都会调整一次粒子的速度和粒子的位置按照下述公式进行更新

$$v_{im}^{t+1} = wv_{im}^t + c_1 r_1 (p_{im}^t - x_{im}^t) + c_2 r_2 (p_{gm}^t - x_{gm}^t)$$

$$x_{im}^{t+1} = x_{im}^t + v_{im}^{t+1}$$

式中，ω 是惯性权值；c_1 和 c_2 是正常数，称之为加速因子，一般在 0 到 2 之间取值；r_1 和 r_2 为中均匀分布的随机数，范围在[0，1]，m 为 M 维中的维数。当迭代结束之后，所得到的全局最优位置解，即需要求出的参数。

2 结果与讨论

2.1 参数与评价指标

实验共采集 305 组光谱样本，其中澳大利亚 58 组，分类标签为 1，俄罗斯 80 组，用 2 作为分类标签，加拿大 37 组，分类标签为 3，印度尼西亚 84 组，分类标签为 4，中国内蒙 46 组，分类标签为 5，然后分别选择各国煤炭光谱的前 31 组作为训练样本，后 6 组作为测试样本，图 1 为各国近红外煤样光谱。

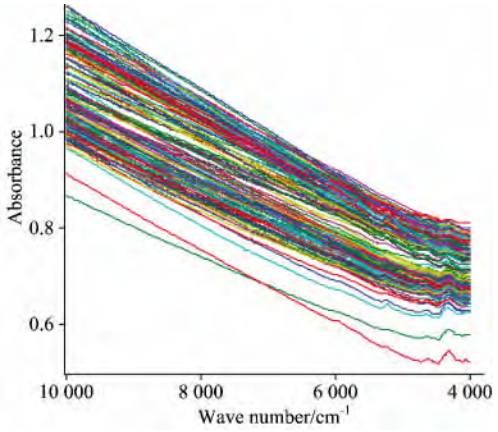

图 1 各国煤样近红外光谱

Fig. 1 NIR of coals from different countries

首先对各国煤的近红外光谱样本进行预处理，剔除异常样本。这里使用的是 SVM 交叉留一法剔除异常样本点，设置的相对误差为 0.008，如图 2 所示。俄罗斯剔除了 5 组，加拿大剔除 1 组，澳大利亚剔除了 0 组，印度尼西亚剔除了 0 组，中国内蒙剔除 4 组异常样本，剔除完成后，以各国后 6 组光谱数据作为测试数据。

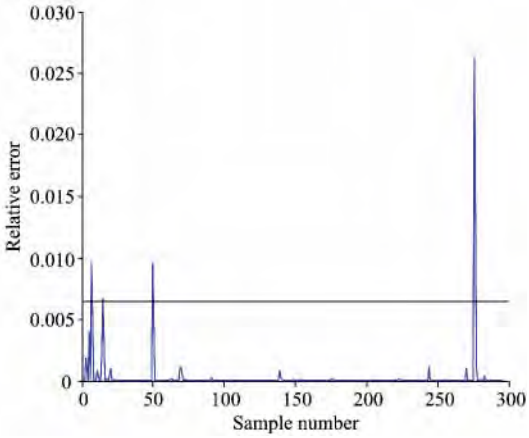

图 2 SVM 网络留一校验法

Fig. 2 Leave-one-out cross validation of SVM

以分类准确率作为评价指标，准确率为：

Accuracy＝分类正确的样本个数/测试样本个数

建立 LVQ1 分类模型，经过 MATLAB 仿真可以得出，LVQ1 误判 7 个，准确率达 76.67%，而 LVQ2 模型，误判 6 个，准确率达 80%，所以在进行各国煤炭分类时，使用 LVQ 神经网络，LVQ2 学习算法要优于 LVQ1 学习算法，其中图 3 为 LVQ1 分类结果，图 4 为 LVQ2 的分类结果图。

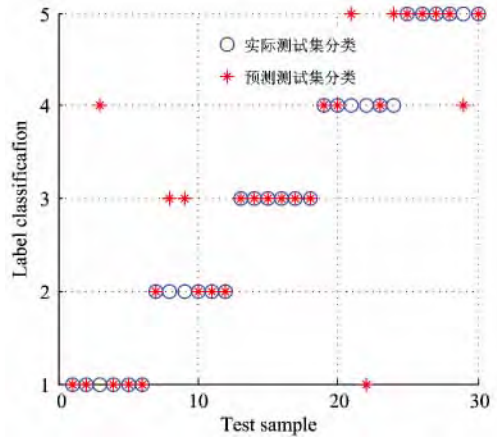

图 3　LVQ(lv1)测试集的实际分类和预测分类图

Fig. 3　The actual and forecast classified figures of LVQ(lv1)

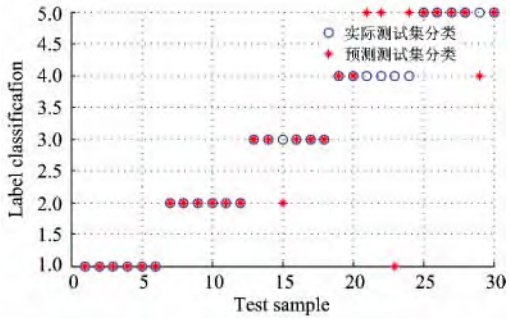

图 4　LVQ(lv2)测试集的实际分类和预测分类图

Fig. 4　The actual and forecast classified figures of LVQ(lv2)

针对预测分类 SVM，这里随机取值 $C=80$，$g=0.03$，可以得出，误判 3 个，准确率可达 90%，如图 5 所示 SVM 的分类效果图，可看出 SVM 模型分类效果要优于 LVQ 模型。

由于 SVM 中参数选择存在人为主观因素，不能保证所选择的参数 C 和 g 就是最优选择，所以引入 PSO 算法对 C 和 g 参数进行筛选，这里设置 C 最大取值为 2～50，最小取值为 2～－10，g 的取值范围和 C 一致，迭代次数为 50 次，最后得到最优解 C 为 $5.866\ 3\times10^{14}$，g 为 $9.765\ 6\times10^{-4}$。

由图 6 适应度曲线，可以看出当迭代到 15 次左右时，适应度基本稳定，为了使实验更加精确，实验迭代次数设置成 50 次为最佳，PSO-SVM 分类模型，误判 1 个，准确率可达 96.67%，图 7 为 PSO-SVM 实际分类与预测分类图。

最后将四种模型进行比较，表 1 为分类模型比较表。

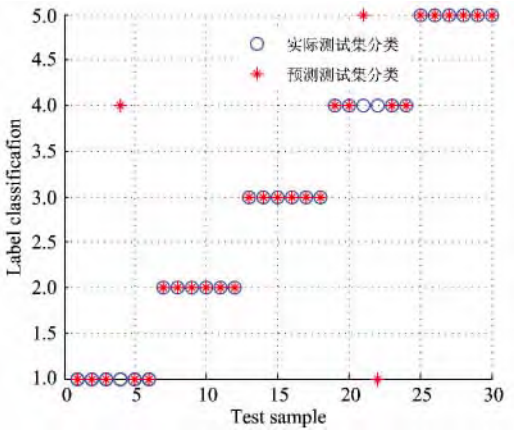

图 5　SVM 测试集的实际分类和预测分类图

Fig. 5　The actual and forecast classified figures of SVM

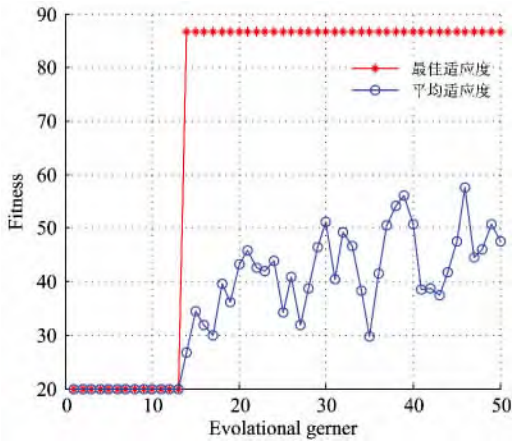

图 6　PSO 适应度曲线

Fig. 6　Fitness curve of PSO ($C1=1.6$, $C2=1.5$)

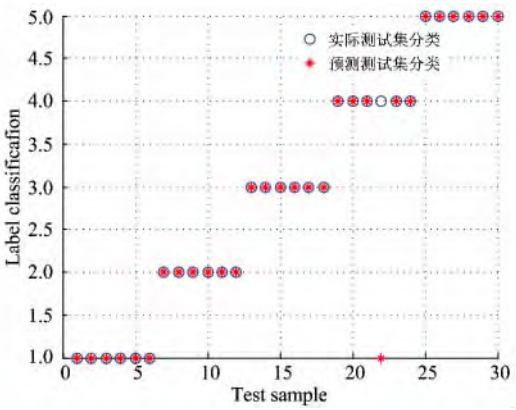

图 7　PSO-SVM 测试集的实际分类和预测分类图

Fig. 7　The actual and forecast classified figures of PSO-SVM

表 1 4 种分类模型比较表

Table 1 Four kinds of model classification results

国家	澳大利亚	俄罗斯	加拿大	印度尼西亚	中国	汇总
LVQ1 错误个数	1	2	1	2	1	7
LVQ1Accuracy	83.33	66.67	83.33	66.67	83.33	76.67
LVQ2 错误个数	0	0	1	4	1	6
LVQ2Accuracy	100	100	83.33	33.33	83.33	80
SVM 错误个数	1	0	0	2	0	3
SVM Accuracy	83.33	100	100	66.67	100	90
PSO-SVM 错误个数	0	0	0	0	1	1
PSO-SVM Accuracy	100	100	100	100	83.33	96.67

3 结 论

LVQ 神经网络和 PSO-SVM 模型都能很好的区分出煤炭的产地，但从系统稳定性和精确性来说 LVQ 神经要比 PSO-SVM 模型稍差。将 PSO 优化算法应用于煤炭分类，优化了模型参数，使 PSO-SVM 模型在系统精度方面优于 SVM 模型。实验证明：PSO-SVM 网络模型可用于煤产地鉴别，为煤炭分类提供了一种耗时较短、实用性强、简洁高效的分析方法。

References

[1] DONG Da-xiao, SHAO Long-yi(董大啸, 邵龙义). Coal Technology(煤炭技术), 2015, (2): 54.

[2] WANG Jiang-rong, WEN Hui, ZHAO Quan-bin(王江荣, 文 晖, 赵权斌). Coal Preparation Technology(选煤技术), 2014, 5: 64.

[3] Xia Wencheng, Yang Jianguo, Liang Chuan. Powder Technology, 2013, 233: 186.

[4] He L L, Melnichenko Y B, Mastalerz M, et al. Energy & Fuels, 2012, 26(3): 1975.

[5] CHEN Peng-qiang, LU Hui-shan, YAN Hong-wei(陈鹏强, 陆辉山, 闫宏伟). Industry and Mine Automation(工矿自动化), 2013, 39(8): 68.

[6] YANG Kai, CAI Jia-yue, ZHANG Chao-ping, et al(杨 凯, 蔡嘉月, 张朝平, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析) 2014, 34(12): 3277.

[7] Lewis A T, Jones K, Lewis K E, et al. Carbohydrate Polymers, 2013, 92(2): 1294.

[8] Sun Zhanquan. Geoffrey Fox. International Journal of Intelligent Transportation Systems Research, 2014, 12(1): 20.

[9] Bron E E, Smits M, van Swieten J C, et al. Feature Selection Based on SVM Significance Maps for Classification of Dementia, in Machine Learning in Medical Imaging, 2014, LNCS 8679: 272.

[10] Milad Jajarmizadeh, Elham Kakaei Lafdani, Sobri Harun, et al. KSCE Journal of Civil Engineering, 2015, 19(1): 345.

[11] XU Xiao-hua, QUAN Xiao-song, ZHANG Zi-feng(徐小华, 全晓松, 张子锋). Journal of Yunnan Minzu University · Natural Sciences Edision(云南民族大学学报 · 自然科学版), 2014, 23(6): 456.

Near-Infrared Spectrum of Coal Origin Identification Based on LVQ with SVM Algorithm

LI Ming, CHEN Fan, LEI Meng*, LI Cui

School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou 221116, China

Abstract Traditional coal origin identification method generally take the calorific value, volatiles, caking index, hardgrove index and crucible swelling number as the classification index, process complicated, use manpower and material resources and can't get coal sample origin directly, take advantages of the near-infrared spectrum technology fast nondestructive testing, due to be collected in the original spectrum that contains some or false spectral data, using Leave-one-out cross validation based on SVM to eliminate abnormal sample of spectral data set, get the correct spectral information of coal sample spectra data sets, and construct the qualitative analysis model based on SVM algorithm and LVQ algorithm, complete based on near-infrared spectral analysis technology of coal origin identification, don't need to make summary and coal samples of various indicators forecast. In view of the random parameter optimization problems in SVM model, the PSO-SVM model of loss parameters (C) and the radius of ker-

nel function (g) are improved, get the optimal parameters, finally, calculation accuracy of the method above contrast model is introduced to evaluate and analysis. Experiments collect the near infrared spectrum of Canada, Russia, Australia, Indonesia and China's five regions, all the data sets, a total of 305 samples, of which 10 simples is abnormal samples and the first 31 groups of the coal spectra were selected as training samples, 6 sets of data after as test samples. Results show that the classification accuracy of classification model can achieve 75% above, including the analysis of the SVM model based on PSO algorithm to improve the accuracy can reach 96.67%, only a sample appear problem, it will be realized quickly and efficiently based on near-infrared spectral analysis technology of coal origin identification.

Keywords Coal origin identification; Near-infrared spectrum; LVQ; SVM; PSO

(Received Apr. 28, 2015; accepted Apr. 16, 2015)

* Corresponding author