第 33 卷 第 5 期
2014 年 9 月

华 中 农 业 大 学 学 报
Journal of Huazhong Agricultural University

Vol. 33 No. 5
Sep. 2014,135～140

基于人工智能的异常地物光谱自适应剔除及分类算法研究

郝建明[1,2] 李宗南[1] 谢 静[3]

1. 华中农业大学资源与环境学院/农业部长江中下游耕地保育重点实验室,武汉 430070;
2. 重庆交通大学土木建筑学院,重庆 404100;3. 华中农业大学理学院,武汉 430070

摘要 针对传统光谱数据预处理与分析的现状,提出一种基于人工智能的光谱异常数据自适应剔除及自动分类算法,通过遗传算法的优化搜索确定马氏距离的阈值,实现异常光谱的自适应剔除,并提出可量化光谱剔除效果的异常一致性指数(ACI)。在此基础上,借助自组织神经网络方法,以各类观测对象的特征光谱作为输入对象,对剔除后的光谱进行自动分类。经过实验验证,算法取得了较好的剔除效果(ACI 达到 86%以上)和分类效果(总体分类精度达到 94%),较好地实现了异常光谱剔除和光谱分类的自动化处理。

关键词 人工智能;遗传算法;马氏距离;自组织神经网络;光谱预处理

中图分类号 TP 751 **文献标识码** A **文章编号** 1000-2421(2014)05-0135-06

光谱分析技术是一种发展迅速的新兴技术,近年来已成熟地应用于有机质预测、食品和药品的品质检测、农作物长势及物种识别等领域中。光谱测量所测定的波谱数据极易受到人为、环境和仪器等因素的影响,导致波谱特征发生异常,影响分析结果的准确性,野外地物光谱测量更是如此[1-2]。另外,光谱分析技术在光谱数据采集的同时,还需要人工进行记录以便对数据进行编号和归类,导致光谱采集过程效率降低。

人工智能作为当前科学技术中正在迅猛发展的一个学科,在光谱数据处理和分析中具有较大的应用潜能。目前,关于异常光谱数据的剔除大多还停留在人工和半自动交互筛选阶段。曹晖等[3]提出了一种基于多种群精英共享遗传算法的异常光谱识别方法,该方法在对光谱样本逐个编码的基础上,采用多个种群同时进行搜索,能适应不同成分异常光谱数据的识别,效果较好,但缺乏通用性,难以适应不同类型的光谱数据处理。Zhu 等[4]基于近红外光谱,提出了一种"二审法",采用回收算子,使得异常光谱识别模型更具有代表性和稳定性。王建义等[5]利用模糊 C 均值聚类法对样品进行聚类,得到可疑样品,再通过 PCA-GA-BP 模型进行识别,该算法过于复杂,效率不高。陈斌等[6]提出一种 PCA 结合马氏距离法剔除近红外光谱中的异常样品,通过设定不同的马氏距离阈值和比对不同阈值下的预测精度来确定剔除阈值。从光谱的定性分析上剔除测量光谱中的异常样品,很好地解决了定量分析和通用性的缺陷,但是需要人工判断剔除的阈值。对于光谱数据的分类,何勇等[7-8]针对苹果测量的光谱数据,建立了基于神经网络的苹果品种识别模型,取得了较好的效果,但是需要事先知道已知类别的光谱样本。

随着各种光谱采集设备的普及,大样本的光谱采集成为可能。光谱数据预处理,特别是异常光谱自动剔除和光谱数据自动分类是影响光谱数据采集效率和质量的关键。因此,笔者结合在光谱采集过程中遇到的问题,提出一种基于人工智能的光谱异常数据自适应剔除及自动分类算法,利用遗传算法的优化搜索能力,结合马氏距离对冠层光谱中的异常数据进行自适应剔除,并采用自组织神经网络的自学习能力实现对处理后的光谱数据进行自动分类,提高光谱采集过程的自动化水平。

收稿日期:2014-02-15

基金项目:国家自然科学基金项目(41201364)、中央高校基本科研业务费专项(2011QC040)、湖北省自然科学基金项目(2010CDB099)和国家大学生创新训练项目(201310504002)

郝建明,硕士研究生. 研究方向:3S 信息处理与工程应用. E-mail:jaminhoh@hotmail.com

通信作者:谢 静,硕士,讲师. 研究方向:生物光学成像与光谱技术. E-mail:xiejing625@mail.hzau.edu.cn

1 材料与方法

1.1 试验数据

试验数据来源于小麦、水稻、棉花3种农作物冠层光谱，每类作物选取100个采集点，每个点重复采集10条光谱，共计3 000条光谱数据。所用光谱仪是Spectra Vista公司生产的SVC HR-1024野外便携式地物波谱仪，光谱范围为350～2 500 nm，最小积分时间为1 ms，通道数1 024个，具有3个线阵探测器，即512 Si(350～1 000 nm)、256 InGaAs(1 000～1 850 nm)、256扩展的InGaAs(1 850～2 500 nm)。其光谱分辨率：≤3.5 nm(350～1 000 nm)、≤8.5 nm(1 000～1 850 nm)、≤6.5 nm(1 850～2 500 nm)，可以实时计算辐照度和反射率，获得实测物体连续的光谱曲线。

1.2 试验方法

1)自适应马氏距离。异常光谱的剔除主要根据对所测光谱与正常光谱在光谱空间上差异的判断进行，而这种差异又可以通过马氏距离来衡量。本文提出的自适应马氏距离是一种基于最佳马氏距离约束的自动获取剔除农作物冠层异常光谱数据的方法，该方法结合马氏距离和遗传算法的优化搜索能力可以智能地剔除冠层光谱中的异常数据，并且能够适应不同类型的地物光谱数据。

光谱马氏距离是指样本光谱与标准光谱集的平均光谱之间的光谱距离，其计算公式如下：

$$D_i = (T_i - \bar{T})\ \mathrm{cov}^{-1}(T_i - \bar{T})^T \tag{1}$$

其中：D_i为光谱样品i的马氏距离，T_i为样品i的主成分得分矩阵$T_{m\times f}=A_{m\times n}\times P_{n\times f}$，$A_{m\times n}$为采集的原始光谱，$P$为主成分载荷矩阵，$m$为波段数，$n$为样品数量，$\bar{T}$为$n$个光谱样品的平均光谱，$\mathrm{cov}^{-1}$为标准光谱集因子分析中得分阵的协方差阵。

检验光谱数据中的异常样品可通过设定马氏距离的阈值并调整阈值范围的权重系数来实现[3]。如果$D_i=D_0$(D_0为确定的阈值)，则认为光谱数据为正常的光谱数据，否则将被划分为异常光谱数据，并从测量数据中剔除。阈值D_0的确定则利用遗传算法智能获取，以实现马氏距离阈值的自动化，具体实现方法如图1所示。

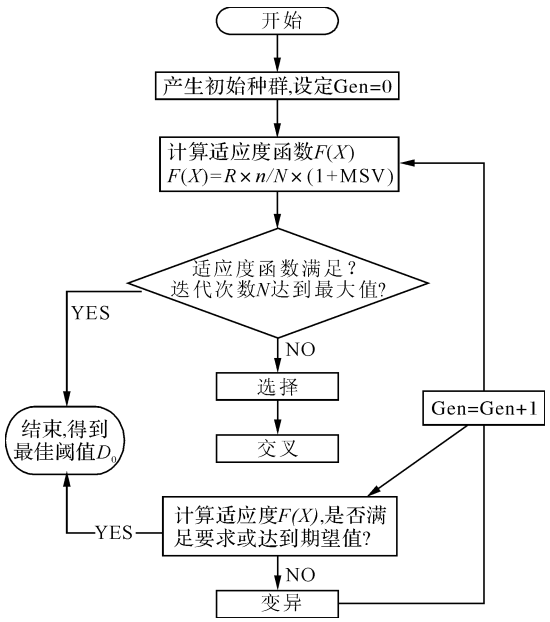

图1 遗传算法优化阈值流程

Fig.1 Figure of GA to optimize threshold

遗传算法优化马氏距离阈值的关键是适应度函数的确定，同一采集点上的作物冠层光谱，具有较高的相关系数，而且光谱曲线的空间特征较为相似，在兼顾异常光谱有效剔除和正常光谱有效保留的原则下，构造了基于遗传算法的适应度函数：

$$\max F(X) = R\times\frac{1}{L+1}\times N_q \tag{2}$$

$$R=\frac{\sum_{i=1}^{n}(m_i-\bar{m})(a_i-\bar{a})}{\sqrt{\sum_{i=1}^{n}(m_i-\bar{m})^2}\sqrt{\sum_{i=1}^{n}(a_i-\bar{a})^2}} \tag{3}$$

R为光谱曲线2个特征波段处反射率的相关系数，其中一个为760～1 000 nm范围内的近红外波段，另一个为500～600 nm范围内的绿光波段。R值越大，则适应度越好。其中m_i为线性拟合值，a_i为实际值，$\bar{m}$、$\bar{a}$分别为线性拟合值均值和实际值均值，n为光谱的数量。

$$L=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i-a_i)^2} \tag{4}$$

L为光谱二维平面上点线性拟合的线性残差，残差越小越好，因此$\frac{1}{L+1}$值越大，其适应度越好。其中m_i为线性拟合值，a_i为实际值，n为光谱的数量。

$$N_q = N_n/N_a \tag{5}$$

N_q为正常光谱数据与异常光谱数据的比率，为了使尽可能多的正常光谱参与后续数据分析，保证比率越大，适应度越好。其中 N_n为正常光谱样品的数量，N_a 为异常光谱数量。

2）自组织神经网络。光谱自动分类是对包含不同种类光谱集进行自动分类的处理。本文中采用的自组织神经网络是通过自主寻找样本中光谱之间的内在规律和本质属性，自组织自适应地改变网络参数和结构，在对未知样本进行聚类分析时具有较好的处理能力。本文建立 3 层的自组织神经网络聚类模型，采用欧氏最小距离的竞争机制，对网络权值进行调整。输入层节点数为 20，竞争层节点数为 20，输出层节点数为 3，网络指定参数中学习速率为 0.1，设定训练迭代次数为 5 000 次。采用 Matlab 语言进行算法实现，对 3 种农作物冠层光谱数据进行分析与建模，随机抽取 250 个光谱（其中小麦 79 条，水稻 88 条，棉花 83 条）采集点上的数据进行训练，利用其他 50 个光谱（其中小麦 21 条，水稻 12 条，棉花 17 条）采集点上的数据进行验证，以分析聚类模型的精度。

1.3　算法原理及步骤

算法利用遗传算法的优化搜索功能对马氏距离阈值进行优化，自适应地剔除数据中的异常样品，并结合自组织神经网络的自学习能力对不同光谱采集的光谱数据进行自动分类，改善了数据采集过程中大量的记录工作，提高了数据采集的效率，其具体实现流程（图 2）和步骤如下：

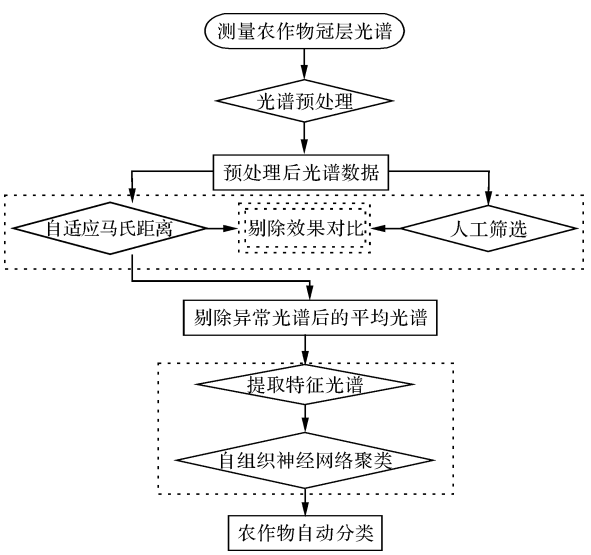

图 2　数据处理整体流程

Fig. 2　Overall flow of data processing

步骤 1：设原始光谱空间为 $A[m\times n]$，对原始光谱进行预处理，通过去除重叠波段，消除噪声，得到预处理光谱空间 $\widetilde{A}[m\times n]$；

步骤 2：对预处理之后的光谱进行降维处理，利用主成分分析得到光谱的 f 个主成分，并计算主成分因子 f 下的载荷矩阵 $P[f\times n]$和得分矩阵 $T[m\times f]$；

步骤 3：通过得分矩阵计算光谱的马氏距离 D_i，$D_i=(T_i-T)\ \mathrm{cov}^{-1}\ (T_i-T)^T$；

步骤 4：设定搜索范围 $(\frac{\max D_i}{2},\max D_i)$，并构造适应度函数 $\max F(X)$，利用遗传算法的优化搜索功能得到剔除异常光谱的最佳阈值 D_0；

步骤 5：对每个光谱采集点上的光谱进行判断，若 $D_i>D_0$ 认为是异常光谱数据，否则认为其是正常的；

步骤 6：计算剔除异常光谱后每个光谱采集点上的平均光谱，获取其特征光谱 $\lambda_1,\lambda_2,\ldots,\lambda_n$，使 3 种农作物在空间上能有效地区分开；

步骤 7：将步骤 6 中获得的特征光谱作为自组织神经网络的输入数据，从样品中随机抽取部分样本作为训练集，得到农作物的分类模型，用剩余的样本对训练模型进行验证，得到 $C_i(i=1,2,3\cdots n)$类。

2　结果与分析

2.1　异常光谱数据剔除

为了增强后续光谱处理与分析的效果，本文主要采用小波去噪方法对光谱数据进行平滑去噪。预处理后的光谱数据在 350～2 500 nm 范围内存在 1 024个波段，由于计算量大，不能全部用于光谱马氏距离的计算，而且光谱波段之间存在的相关性会影响马氏距离的计算结果。因此，在计算光谱马氏距离之前需利用主成分分析对数据进行压缩和降维，得到光谱数据的得分矩阵，试验中选用了前 6 个累计贡献率达到 99.8%的主成分代表整个光谱数据进行马氏距离计算。

以 1 个光谱采集点上测量得到的 10 条光谱为例，通过本文构造的遗传算法模型，得到的最佳阈值为 5.812 8（图 3），图中有 4 条测量光谱大于该阈值，即 1 号、2 号、9 号、10 号，这 4 条测量光谱被确定为异常光谱。

整个试验选取 3 种农作物 300 个采集点上的冠层光谱数据进行算法验证。为统计人工筛选和算法

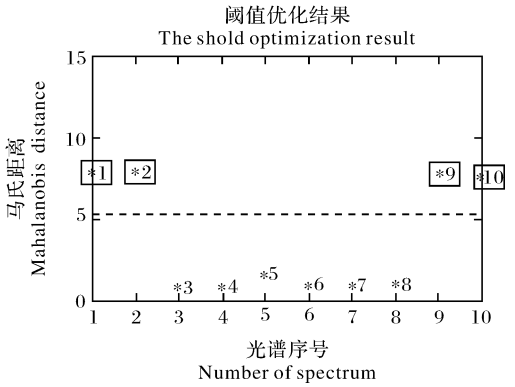

图3 遗传算法优化马氏距离效果图

Fig. 3 Figure of GA optimization of Mahalanobis distance

筛选所剔除异常光谱的一致性，本文提出异常一致性指数(abnormal compliance index，ACI)。ACI是衡量自适应马氏距离处理得到的异常光谱样品(abnormal spectral，AS)与人工筛选的异常光谱样品中相同光谱的比率。其计算公式如下：

$$ACI = \frac{N_s}{N_{AS}} \tag{7}$$

其中 N_s 为自适应马氏距离异常光谱样品与人工筛选结果中相同光谱的数量，N_{AS} 为自适应马氏距离得到的异常光谱的数量。从表1可知，3种农作物ACI指数都达到了86%以上，结果表明所提方法与人工筛选结果有较高的吻合度，具有较好的异常光谱筛选精度。

表1 异常光谱剔除结果对比

Table 1 Comparison of excluding exception spectrum

农作物类型 Types of crops	原始光谱数量 Original spectrum	光谱数据处理方式 Ways of data procession				异常光谱一致性判断 Consistence of abnormal spectrum	
		人工筛选 Manual conduction		自适应马氏距离 Self-adapting Mahalanobis distance			
		正常光谱 Normal spectrum	异常光谱 Abnormal spectrum	正常光谱 Normal spectrum	异常光谱 Abnormal spectrum	一致性异常光谱 Equal abnormal spectrum	ACI/%
小麦 Wheat	1 000	773	227	765	235	206	87.6
棉花 Cotton	1 000	731	269	759	241	211	87.5
水稻 Rice	1 000	733	267	742	258	224	86.7

2.2 冠层光谱自动分类

野外测量农作物冠层光谱时，需要花费人力对采集光谱做编号和类属标记，降低了数据采集的效率。光谱的自动分类是利用自组织神经网络对采集后的光谱数据自动匹配类别，减少了前期复杂的数据记录工作。在剔除异常光谱数据后，求取每个农作物冠层光谱采集点上的平均光谱。得到的平均光谱含有上百个波段的反射率信息，如果全部用作自组织神经网络的输入，必然影响运算效率。因此，在进行神经网络聚类之前，利用主成分分析提取能够将3种作物有效区分的特征波段，提取主成分贡献率较大的前20个特征波段的得分矩阵，选取第一、第二、第三主成分绘制成三维散点图，发现3种农作物的聚类效果较好(图4)。因此，选用前20个特征波段的得分作为自组织神经网络的输入数据，对农作物冠层光谱自动分类。

将待分类的光谱数据随机分成建模集(250个)和验证集(50个)，建模集用于自组织神经网络的训练数据，验证集用于识别未知样品的种类。将得分矩阵 $T_{20\times250}$ 作为自组织神经网络的输入变量，农作物种类作为输出变量(其中，1、2、3分别代表农作物

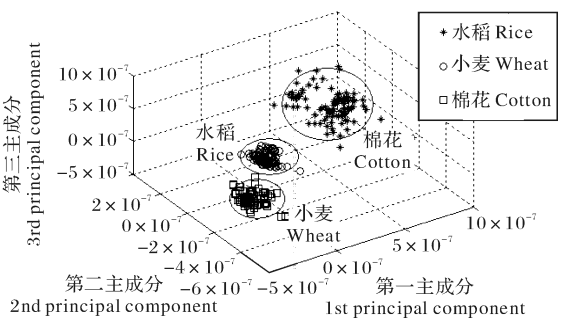

图4 农作物三维聚类效果图

Fig. 4 Three dimensional cluster rendering crops

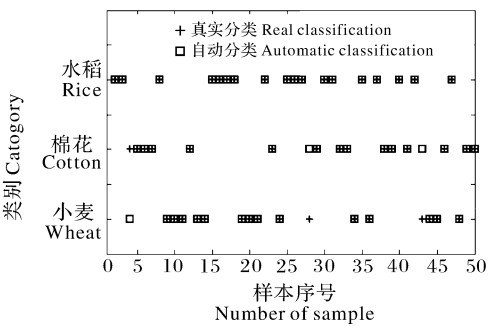

图5 农作物自动分类结果

Fig. 5 Results of automatic classification

种类)，建立一种 3 层神经网络模型的农作物冠层光谱分类模型，通过神经网络的反复训练，选择最佳的网络结构，即输入层 20 个节点，竞争层 20 个节点，输出层 3 个节点。用验证集中的 50 个样品对训练好的模型进行验证(图 5)。

由表 2 中自组织神经网络聚类结果可以看到，16 个光谱样品被划分到小麦类，16 个光谱样品划分到了棉花类，18 个光谱样品划分到了水稻类。和真实分类的结果相比，小麦中 2 个光谱样品错误地划分到棉花类，棉花中 1 个光谱样品错误地划分到了小麦类，总体精度达到 94%，效果较好。每类农作物分类精度见表 3。

表 2　自组织神经网络聚类情况[1)]

Table 2　Clustering for self-organization neural networks

农作物种类 Types of crops	真实分类 Real classify	自组织神经网络聚类 Classify by CLSOM
小麦 Wheat	1、9、10、11、13、14、19、20、21、24、[28]、34、36、[43]、44、45、48	1、[4]、9、10、11、13、14、19、20、21、24、34、36、44、45、48
棉花 Cotton	[4]、5、6、7、12、23、29、32、33、38、39、41、46、49、50	5、6、7、12、23、[28]、29、32、33、38、39、41、[43]、46、49、50
水稻 Rice	2、3、8、15、16、17、18、22、25、26、27、30、31、35、37、40、42、47	2、3、8、15、16、17、18、22、25、26、27、30、31、35、37、40、42、47

1)表中的序号是对 50 个验证集中的光谱样品进行的编号，其中方框中序号表示误分类光谱。The serial number in table is the number of spectral sample in fifty validation set, and the box number indicates the misclassification spectrum.

表 3　神经网络自动分类精度

Table 3　The automatic classify accuracy of BP network

样本 Sample	样本数 Number of sample	误分类数 Number of error claasify	分类精度/% Accuracy of classify
小麦 Wheat	16	1	93.75
棉花 Cotton	16	2	87.50
水稻 Rice	18	0	100.00
总计 Total	50	3	94.00

3　讨　论

本文提出的基于人工智能的异常光谱数据自适应剔除算法，结合了遗传算法优化搜索能力和自组织神经网络的自学习能力，是对传统的野外采集农作物冠层光谱数据处理的一种改进，是一种收敛速度快、剔除精度高、筛选效率高的野外采集光谱数据处理算法。与传统的人为筛选异常光谱数据相比，自动化程度高。以 3 种农作物 300 个野外采集的农作物冠层光谱样本为试验数据，异常一致性指数 ACI 为衡量指标，对算法进行验证，其异常一致性指数 ACI 达到 86%以上，同时，采用了一种自动化马氏距离剔除异常光谱的方法，改进了传统算法中人为确定马氏距离的现状，从而验证了本算法在剔除农作物冠层光谱数据中的可行性。用自组织神经网络对异常光谱处理之后的光谱数据自动分类，该方法具有较强的鲁棒性，改进了以往需要事先知道部分已知样品类别的方法。从试验数据中随机选取 250 个样本作为建模集，对样本进行自学习，利用训练好的分类模型对剩余 50 个光谱样本进行自动分类，其总体分类精度达到 94%，基本上满足光谱数据处理的精度要求，分类效果较好。基于此，本文在改进传统野外采集农作物冠层光谱数据处理的基础上，形成了一套完整的数据自动化处理分类流程，具有很好的实用性。根据我们对农作物冠层光谱数据处理的思路，可以对算法进行改进，使其处理速度更快，处理效率更高。同时，还可以结合信息化时代发展的需求，将算法集成到光谱采集终端上，在采集数据的同时完成数据的自动化处理，以提高光谱数据采集后数据处理的效率和准确度。

参　考　文　献

[1] 孔维豪，祝民强. SVC HR-768 地物光谱仪岩石光谱采集存在的问题与处理[J]. 东华理工大学学报：自然科学版，2012，35(2)：155-159.

[2] 代芬，洪添胜，罗霞，等. 基于可见-近红外光谱的砂糖橘总酸无损检测[J]. 华中农业大学学报，2012，29(4)：518-523.

[3] 曹晖，周延. 多种群精英共享遗传算法在异常光谱识别中的应用[J]. 光谱学与光谱分析，2011，31(7)：1847-1851.

[4] ZHU S P，WANG Y M，WU J Z，et al. Outlier sample elimination criterions and methods for building calibration model of near infrared spectroscopy analysis[J]. Journal of Agriculture Machinery，2004，35(4)：115-118.

[5] 王建义，雷萌. 近红外光谱煤质分析模型中异常样品的剔除方法[J]. 工矿自动化，2011，37(11)：75-77.

[6] 陈斌,邹贤勇,朱文静. PCA 结合马氏距离法剔除近红外异常样品[J]. 江苏大学学报:自然科学版,2008,29(4):277-279.
[7] 何勇,李晓丽,邵咏妮. 基于主成分分析和神经网络的近红外光谱苹果品种鉴别方法研究[J]. 光谱学与光谱分析,2006,26(5):850-853.
[8] HE Y, LI X. Discriminating varieties of waxberry using near infrared spectra[J]. Journal of Infrared and Millimeter Waves, 2006,25(3):192-194.

Artificial intelligence based algorithm for using spectrum to adaptively eliminate exceptional data and automatically classify

HAO Jian-ming[1,2] LI Zong-nan[1] XIE Jing[3]

1. *College of Resources and Environment, Huazhong Agricultural University/ Key Laboratory of Arable Land Conservation (Middle and Lower Reaches of Yangtse River) Ministry of Agriculture, Wuhan* 430070, *China*;
2. *College of Civil Engineering and Architecture, Chongqing Jiaotong University, Chongqing* 404100, *China*;
3. *College of Science, Huazhong Agricultural University, Wuhan* 430070, *China*

Abstract The spectral data measured from spectral measurements are easily affected by human, environmental, equipment and other factors leading to the abnormal spectral characteristics and impacting analyses especially in spectral measurements in the fields. According to the situations of traditional spectral data in preprocessing and analysis, a novel algorithm used for abnormal data excluding adaptively and spectral data classifying automatically based on artificial intelligence was established. The Mahalanobis distance threshold by genetic algorithm searching was determined to exclude abnormal spectral data adaptively and to quantify the effect of excluding abnormal spectral consistency index (ACI). With the self-organizing neural network, spectral characteristics of various types of observing objects were used as input and classified automatically after removing the abnormal. The results showed that the algorithm achieved good excluding (ACI more than 86%) and classification (overall classification accuracy of 94%). It can be used to well automate the handling of excluding spectrum and spectral classification.

Key words artificial intelligence; genetic algorithm; Mahalanobis distance; self-organizing neural network; spectral preprocessing

(责任编辑:陆文昌)