文章编号:1001-5078(2007)03-0237-03

基于近红外光谱和系统聚类法的蝗虫识别模型研究

熊雪梅,王一鸣
(中国农业大学信息与电气工程学院,北京 100083)

摘 要:以蝗虫的近红外光谱图作为分析的对象,采用聚类分析方法对蝗虫进行快速的识别研究。结果表明:矢量归一化预处理后的光谱数据经最长距离法进行聚类分析,分类结果正确率为100%。此法可为蝗虫自动侦测系统中识别蝗虫提供一种可靠、简便的手段,盲样检测的准确率可达91.67%。

关键词:近红外光谱;聚类分析;蝗虫
中图分类号:TN219;S431 **文献标识码**:A

Locusts Detection Model Based on Near Infrared Spectra and Cluster Analysis

XIONG Xue-mei, WANG Yi-ming
(College of Information and Electrical Engineering, China agricultural university, Beijing 100083, China)

Abstract: The objective of this study was to develop a rapid, real-time method for the detection of locusts. Near infrared spectra (NIR) in the range of 3996 ~ 12489.5cm^{-1} were recorded and the range of 3996 ~ 11502cm^{-1} was selected. Models were established by using the cluster analysis method. Different methods were compared. The study showed that complete-linkage model gave better results than single-linkage and other methods with the correctly classification rate of 91.67% for the test data set. Results indicate that NIR method can be used to on-line detect the locusts.

Key words: near infrared spectra; cluster analysis; locust

1 引 言

近红外光谱可以应用于作物病虫草害的监测,如小麦条锈病,小麦蚜虫发生状况,田间杂草的发生情况[1]等。蝗虫的为害十分严重,造成的直接和间接经济损失,以及治理中化学药物喷洒所带来的环境污染等间接影响都不可忽视。遥感与地理信息系统在蝗虫测报中的应用是通过对蝗虫生境的研究和评价去间接实现对蝗虫可能发生地点和发生情况的监测[2-5]。本文采用近红外光谱与化学模式识别相结合的方法直接侦测出蝗虫,快速准确地预测蝗虫成灾信息,为蝗虫自动侦测系统中识别蝗虫提供技术支持,为开发有效而实用的原位在线、实时动态分析便携式近红外光谱蝗虫识别仪提供理论依据。

包含蝗虫的田间样本是一个复杂的混合物体系,所含各化合物吸收强度的叠加具有难以解析的复杂性,土壤和作物或草化学成分含量各不相同,近红外的光谱有差异,借助各样本近红外光谱图的差异,运用化学模式判别法对谱图进行解析,建立相关的模式识别法,实现蝗虫的快速识别在理论上是可行的。

2 蝗虫近红外光谱识别基本流程

首先收集具有代表性的样品,然后采集样品的

基金项目:国家科技部科研院所社会公益研究专项项目(2004DIB3J076)。

作者简介:熊雪梅(1970 -),女,讲师,博士学位,主要从事人工神经网络和模式识别研究及在农业病虫害预测预报的应用研究工作。

收稿日期:2006-08-24;**修订日期**:2006-10-12

光谱数据,将光谱数据进行预处理,通过数学方法,建立数学分类模型。在分析未知样品时,先对待测样品进行扫描,根据光谱值利用建立的模型可以将待测样品分类识别。

3 实验方法

3.1 样品的收集、制备

本试验所用的分析样品是来自河北沧州蝗虫养殖基地的蝗虫,共79个样品,其中55个样本作为训练集,24个样本作为测试集。训练集中1~43号样本为1~4龄蝗虫,44~47号样本为土壤,48~53号样本为草(包括2个干草样本),54~55号样本为石子。测试集中56~74号样本为1~4龄蝗虫,75号为土壤样本,76~78为草样本。

3.2 光谱的采集

采用Bruke公司生产的Matrix-Ⅰ型傅里叶近红外光谱仪,进行全谱测定,先采集55份样品在近红外光谱整个区域的光谱信息。

把一定量样品放在专用的样品杯中,采用积分球和旋转台测定样品NIR漫反射光谱,在波数为3996~12489.5cm^{-1}范围内,对每份样品进行重复120次扫描,每隔2nm采集反射强度,所有样品均重复2次测定。所采集的蝗虫及土壤等的近红外反射光谱存入计算机中。图1为蝗虫及土壤等的近红外光谱图。纵坐标为吸光度(absorbance units,AU)。

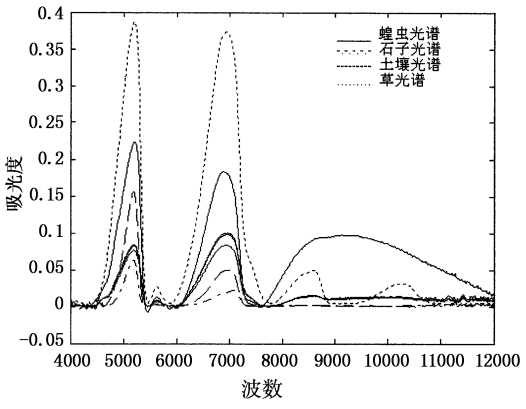

图1 样本的近红外光谱图

采用简易的峰位鉴别法主要是分析组分相差较大的不同种物质,直观、简便,可对不同样品进行鉴别,但包含蝗虫的田间样本是一个复杂的混合物体系,所含各化合物吸收强度的叠加具有难以解析的复杂性,因此必须需要其他的方法,如化学计量学方法等来鉴别。

4 识别软件原理

定性判别分析的重要依据是光谱特征的相似性(或差异性),因此正确获取光谱特征是定性分析的关键一环。目前,在模式识别中应用广泛的特征筛选方法有偏差权重法、K-W检验、主成分分析、偏最小二乘法等。在实际工作中,经常遇到只需要知道样品的类别或等级,并不需要知道样品中含有的组分数与其含量的问题,这时需要应用模式识别法。模式识别法主要用于光谱的定性分析。在近红外光谱定性分析中常用的模式识别方法很多,有聚类分析、判别分析、主成分分析和人工神经网络方法。系统聚类分析是依据一种事先选定的相似性或非相似性(如距离)来度量类在分类空间中的距离,再根据谱系图决定分类结果。

4.1 系统聚类法

系统聚类法是先将每个样品视为一类,然后定义样品间的距离和类与类的距离,聚类过程是首先选择距离最小的两类合并为一类,再按类间距离的定义,计算新类与其他类间的距离,再将距离最近的两类合并,如此继续,直至所有样品归为一类。OPUS软件中的聚类方法有complete-linkage聚类法(最长距离法),single-linkage聚类法(最短距离法),average-linkage聚类法(平均距离法),median算法(中值算法),Ward's算法(沃德法),centroid算法(质心算法)等[6]。其中最长距离法定义为[1]:

设$G_1,G_2,\cdots,G_n$表示n类;d_{kl}表示样品k和l间的距离;D_{ij}表示类G_i和G_j间距离,则:

$$D_{ij}=\max_{\substack{k\in G_i\\ l\in G_j}}\{d_{kl}\}$$

聚类步骤为:

(1)计算所有样品间的距离,得距离矩阵$\boldsymbol{D}_{(0)}$,各样品自成一类,此时$D_{ij}=d_{ij}$;

(2)在$\boldsymbol{D}_{(0)}$非对角线元素中选取最小元素,设为D_{ij},将G_i与G_j合并为一类,记作G_r,则$\{G_i,G_j\}$即G_r中样品为G_i,G_j中全部样品;

(3)计算新类G_r与其他类G_s间的距离D_{rs}:

$$D_{rs}=\max_{\substack{k\in G_r\\ l\in G_s}}\{d_{kl}\}=\max\{\max_{\substack{k\in G_i\\ l\in G_s}}d_{kl},\max_{\substack{k\in G_j\\ l\in G_s}}d_{kl}\}$$

$$D_{rs}=\max\{D_{is},D_{js}\}$$

由此得到距离矩阵$\boldsymbol{D}_{(1)}$。

(4)对$\boldsymbol{D}_{(1)}$重复$\boldsymbol{D}_{(0)}$的步骤得$\boldsymbol{D}_{(2)}$,如此继续下去直到所有样品都归为一类为止。

5 结果和讨论

5.1 判别模型的选择

本文通过采用矢量归一化将光谱数据进行预处理。采用BRUKE公司OPUS近红外软件的聚类分

析方法对这个训练集的近红外光谱数据进行处理，建立判别模型。选择不同的光谱范围和不同的聚类方法建立的模型，其识别率不同。从图1可以看出，光谱范围在3996～11502cm^{-1}内包含丰富信息，因此对比实验时所选光谱范围为3996～11502cm^{-1}。实验结果如表1及图2所示。在表1中，只有经矢量归一化的最长距离法分类结果与实际样本分类情况吻合，即样本1～43为蝗虫，样本44～55为非蝗虫。

表1 训练组分类结果

分类方法	预处理方法	属于类别1的样本	属于类别2的样本
最长距离法	矢量归一化	1～43	44～55
	无预处理	1～28,37,44～48,52～55	29～36,38～43,49～51
最短距离法	矢量归一化	1～43,47～53	44～46,54,55
平均距离法	矢量归一化	1～28,44～55	29～43
中值算法	矢量归一化	1～10,14～23,44～55	11～13, 24～43
沃德法	矢量归一化	1～28,44～55	29～43
质心算法	矢量归一化	1～43,49～53	44～48, 54～55

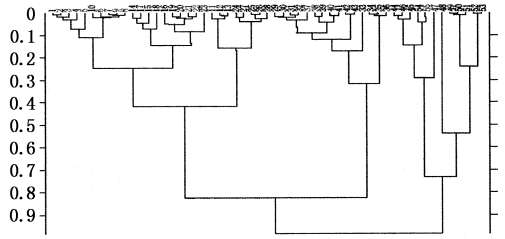

图2 矢量归一化后采用最长距离法的聚类分析结果

（光谱范围为3996～11502cm^{-1}）

对光谱数据不采用任何预处理，很难将样本正确分类；对光谱数据采用矢量归一化预处理后用最长距离法判别模型的分类最正确，若设定分类数为4类，可正确把1、2龄归为一类，3、4龄归为一类，草为一类，土为一类。若设定分类数为2，则能正确把1、2、3、4龄蝗虫归为一类，土和草归为一类，即蝗虫和非蝗虫两类（见图2）。光谱范围取3996～11502cm^{-1}时和全谱范围的聚类分析结果一致。为减少计算量，本文采用光谱范围为3996～11502cm^{-1}并采用最长距离法建立判别模型。

5.2 判别模型预测的可靠性

为检验所建立判别模型的判别能力，用另外一组已知类别24个样品作为盲样，组成测试集，来验证模型的可靠性。测试集所得的正确判别率则称为预测率，预测率对模型好坏的判别比识别率更重要。盲样检验，除1龄和2龄死蝗虫未被检测出为蝗虫而是归为非蝗虫类外，其余未知样本均都识别正确，经计算得出预测正确率为91.67%。图3为预测集中68号样本被正确分类图。由于扫描操作未在同一天进行，背景因素和装样因素影响较大，2006年3月10日扫描的1龄和2龄死蝗虫对聚类结果影响较大，在训练集中剔除了此样本。若将测试集中此样本去除，则模型的盲检率为100%。

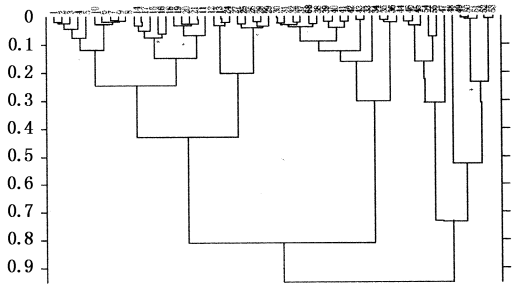

图3 预测集中68号样本被正确分类图

6 结 论

本文运用近红外反射技术与化学模式识别相结合对蝗虫快速识别进行探讨，取得较满意的结果，盲样类别鉴定的准确率可达91.67%，所建立的模型基本上可以准确地判断出蝗虫。对于预测不准确的样品是因为该类样品本身较特别，所挑选的标准样品集没有包括该样品信息。如果能够获得足够多的样品，增加用来建立判别模型的样品集，则判别模型的使用范围将更广，判断的准确率将会大大提高。

参考文献：

[1] 严衍禄. 近红外光谱分析基础与应用[M]. 北京：中国轻工业出版社，2005：526－527.

[2] Wehn H, Rabus B, Wood D, et al. Prediction of locust outbreaks from RADARSAT-1 multi-angle data[C]. Proceedings of IEEE International Symposium on Geoscience and Remote Sensing 2004, IGARSS′04, Anchorage, 2004,5:3543 －3546.

[3] Crooks W T, Archer D J. SAR observations of dryland moisture-towards monitoring outbreak areas of the brown locust in South Africa[C]. Proceedings of IEEE International Symposium on Geoscience and Remote Sensing 2002, IGARSS′02, Toronto, 2002,4:1994－1996.

[4] Jianwen Ma, Hasibagan H X, Devision T. Calibration and verification of remote sensing data for East Asia migratory plague locust reed habitat monitoring[C]. Proceedings of IEEE International Symposium on Geoscience and Remote Sensing 2002, IGARSS′02, Toronto, 2002,5:2868－2870.

[5] Dai Qin, Ma Jianwen, Han Xiu-Zheng, et al. Remote sensing new model for monitor the East Asian migratory locust infections based on its breeding circle[C]. Proceedings of lEEE International Symposium on Geoscience and Remote Sensing 2004, IGARSS′04, Anchorage, 2004,7:4468－4469.

[6] 布鲁克公司. 近红外用户培训手册[Z].