基于最大散度差准则 LDA 的电子鼻中药材鉴别方法

邵雅雯,骆德汉,武 琳,李江勇
(广东工业大学信息工程学院,广东广州 510006)

摘要:在电子鼻的模式识别方法中,线性判别分析(Linear Discriminant Analysis ,LDA)是常用的方法之一。然而,当样本类内散布矩阵奇异时,使用传统的基于 Fisher 准则的 LDA 算法会出现小样本问题。将最大散度差准则引入线性判别分析中,不仅可以解决小样本问题,实现 3 种不同产地中药材的正确鉴别,而且分类效果更好。结果表明:对 3 组样本的最终判别结果达到了 97.8% 的正确判别率,误判的待测样本只发生在安徽白术。

关键词:电子鼻;线性判别分析;最大散度差准则;中药材
中图分类号:TP212.9 **文献标识码**:A **文章编号**:1002 - 1841(2011)11 - 0080 - 03

Classification of Chinese Herbal Medicine Based on LDA with Maximum Scatter Difference Criterion Using E-nose

SHAO Ya-wen, LUO De-han, WU Lin, LI Jiang-yong
(**School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China**)

Abstract: Linear discriminant analysis (LDA) is a popular method among pattern recognition algorithms of electronic nose (E-nose). However, small sample size (SSS) problem would occur while using LDA algorithm with traditional Fisher criterion if the within-class scatter matrix is singular. This paper combined maximum scatter difference criterion and LDA to solve SSS problem, so that three kinds of Chinese herbal medicines from different growing areas were accurately classified. At the same time, the classification result was enhanced. The results show that only a few samples of Anhui Atractylodes are classified incorrectly, however, the classification rate reaches 97.8%.

Key words: electronic nose; linear discriminant analysis; maximum scatter difference criterion; Chinese herbal medicine.

0 引言

白术是一种具有特殊气味的菊科中药材,其品质受产地、采收期、品种等因素的影响,其中产地因素是评判品质的最重要的标准之一[1]。伴随着人们对中药材品质要求的日益提高,对药材的辨别显得尤为重要。

电子鼻的研究始于 20 世纪 90 年代,是一种由具备部分专一性的气敏传感器构成的阵列和适当的模式识别系统组成的仪器,主要用来识别简单和复杂气味。国内外已在食品行业、医疗诊断、环境监测等方面有大量的研究成果和社会应用[2-4],但在中药材领域的研究报道目前并不多。

文中将中药材白术作为研究对象,利用电子鼻进行检测。在电子鼻的模式识别方面,主成分分析(principal component analysis, PCA)、LDA 得到了广泛的应用。LDA 的突出优点就是能够保证投影后的模式样本在新的空间中有最小的类内距离和最大的类间距离,即模式在该空间中有最佳的可分离性,但同时也存在不适用于"小样本问题"等缺点。针对这一缺点,有不少学者采用了 PCA + LDA 的组合方法[5],将 PCA 和 LDA 的优点充分的融合在一起,既能解决 PCA 算法对不同的训练样本数据不敏感的问题,又解决了 LDA 算法中当类内散步矩阵奇异的问题,获得了较好的分类效果。而文中则将最大散度差准则和 LDA 结合起来,不仅解决了小样本问题,而且与 PCA 和 PCA + LDA 相比得到了更好的分类效果。

基金项目:国家自然科学基金资助项目(60971105)
收稿日期:2011 - 03 - 23 **收修改稿日期**:2011 - 08 - 06

1 实验材料、仪器和方法

1.1 材料与仪器

文中采用的中药样品是由广州中医药大学提供的。该样品为不规则的肥厚团块,气清香,味甘、微辛。提供的白术样品分别来自于 3 种产地:河北保定、安徽亳州、浙江绍兴。

所用的仪器是便携式电子鼻 PEN3 (Portable Electronic Nose3),该电子鼻是一种由一组复合化学传感器和识别软件组成的分析仪器,它具有自动调整、自动校准及系统自动富集的功能。PEN3 电子鼻包含 10 个金属氧化物型传感器,它在采样过程中的响应信号记录为某传感器接触到样品挥发性气体后的电导率 G 与该传感器接触经过标准活性碳过滤的基准气体后的电导率 G_0 的比值,即 G/G_0。

1.2 实验方法

1.2.1 实验参数

实验环境参数如下:实验室内温度保持在 24 ~ 26 ℃,相对湿度保持在 52% ~ 56%,采用静态顶空抽样的方法用 PEN3 进行检测。

电子鼻参数设定如下:采样时间设为 60 s,传感器阵列的清洗时间设为 110 s,每种样品采样 30 次,共 90 次。采样数据一半用作训练样本,一半用作待测样本。

1.2.2　采样方法

分别用镊子和天平称取各类样品 100 g,放入容量为 500 mL 的烧杯中,3 个烧杯按产地的不同分别贴上标签:河北、安徽、浙江。为使瓶内顶空气体达到饱和状态,3 个样品瓶分别用保鲜薄膜密封静置 70 min. 先用干净空气清洗传感器阵列,接着用一接在 PEN3 特氟纶管上的探针透过保鲜薄膜插入样品瓶,准备采样。

2　基于最大散度差准则的 LDA 方法的数学原理

LDA 是一种常用的模式分类算法。然而,当样本总数较少或选取的特征数较多时,直接采用 LDA 算法会出现小样本问题,即导致样本类内散布矩阵 S_ω 奇异,LDA 算法将无法进行下去。而解决小样本问题可通过降维的方法使 S_ω 非奇异或者避免对 S_ω 求逆[6]。因此,文中将采用基于最大散度差准则的 LDA 方法,绕过了传统 Fisher 鉴别准则中需要对 S_ω 求逆的步骤,避免了可能出现的"小样本问题"。

设 $G_1,G_2,\cdots,G_N$ 是 N 个模式类,模式 $x\in R^d$ 为 d 维实向量,第 i 类训练样本的个数为 N_i,样本均值 m_i、总体类内散布矩阵 S_ω、类间散布矩阵 S_b 分别定义为:

样本均值 m_i:

$$m_i=\frac{1}{N_i}\sum_{x\in G_i}x,i=1,2,\cdots,N \tag{1}$$

总体类内散布矩阵 S_ω:

$$S_\omega=\sum_{i=1}^{N}\sum_{x\in G_i}(x-\mu_i)(x-\mu_i)^T \tag{2}$$

式中 $i=1,2,\cdots,N$.

类间散布矩阵 S_b:

$$S_b=\sum_{i=1}^{N}(m_i-m)(m_i-m)^T \tag{3}$$

式中

$$m=\frac{1}{N}\sum_{i=1}^{N}m_i \tag{4}$$

Fisher 鉴别准则是,选择使得广义 Rayleigh 商

$$J_F(\omega)=\frac{\omega^TS_b\omega}{\omega^TS_\omega\omega} \tag{5}$$

达到最大值的向量作为投影方向。

最大散度差准则的目的是寻找一个最优的投影方向 ω,使得在低维空间里各类样本能被最大限度分离开来。但它与 Fisher 鉴别准则不同的是,它不是沿用广义 Rayleigh 商,而是将类间散度与指定倍数的类内散度的差作为投影后数据的可分性度量。因此可以定义最大散度差准则函数为:

$$J_M(\omega)=\frac{\omega^T(S_b-C\cdot S_\omega)\omega}{\omega^T\omega} \tag{6}$$

式中:C 是一个正常数,为了方便,文中设定为 1,用来平衡最大化类间散度和最小化类间散度;$S_b-C\times S_\omega$ 称为参数为 C 的广义散度差矩阵。

可以证明,这个最优投影方向 ω 就是使最大散度差准则函数 $J_M(\omega)$ 取极大值的解,即下列广义特征值问题的解:

$$(S_b-C\times S_\omega)\omega=\lambda\times\omega \tag{7}$$

因此,最大散度差鉴别准则可以归结为求广义散度差矩阵 $S_b-C\times S_\omega$ 的特征向量问题。

3　结果与讨论

3.1　传感器响应

PEN3 对 3 种不同产地白术样品的响应特性曲线如图 1 所示,图中横轴为采样时间,纵轴为各传感器的响应值。

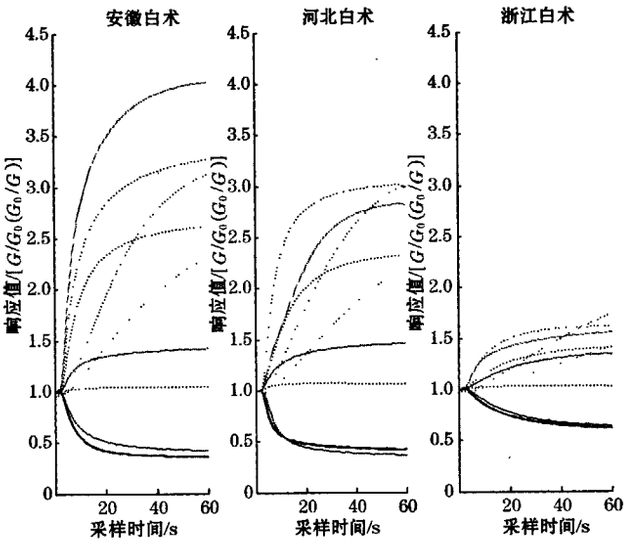


图 1　3 种不同产地白术的响应特性曲线

可以看到,响应值刚开始较小,随着挥发物在传感器表面富集,传感器响应值不断增大,最后趋于平缓,在 60 s 时达到一个较稳定的状态。PEN3 的传感器阵列对不同产地的白术样品具有不同的响应特性曲线,各个传感器对不同品种的中药材样品均有响应,表明各传感器对同一气味具有交叉敏感特性,并且各个传感器的响应特性不完全相同。

3.2　特征参数提取

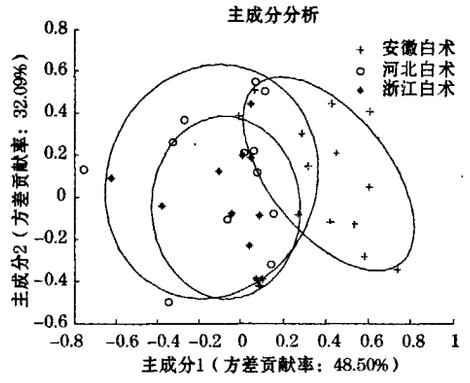
文中将尽可能多地选择能反映原始数据信息的特征变量来组成原始特征向量,以保留原始响应数据中的信息。根据传感器响应特性曲线的趋势及变化情况,分别选择了第 10 s 响应值(传感器响应曲线急速上升阶段的响应值)、第 40 s 响应值(传感器响应曲线缓慢上升阶段的响应值)、第 60 s 响应值(信号相对稳定值)、均值、峰值、方差、标准差以及微分值合共 8 个特征子集组成原始特征向量 T,表示如下:

$$T=[f_{10},f_{40},f_{60},\mathrm{avg},\mathrm{max},\mathrm{var},\mathrm{std},\mathrm{diff}] \tag{8}$$

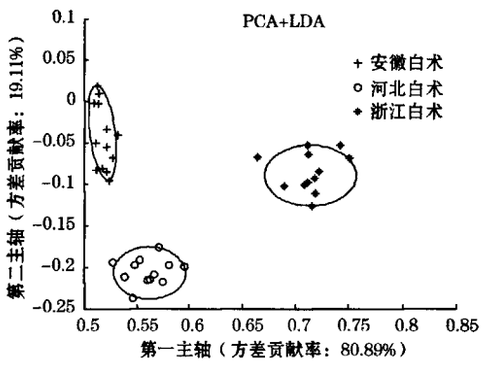
式中:f_i 是传感器阵列在第 i 秒的响应值($i=10,40,60$);avg 是传感器在 60 s 内相应数据的算术平均值;max 是传感器在 60 s 内的最大响应值;var 是传感器在 60 s 内响应数据的方差;std 是传感器阵列在 60 s 内的标准差;diff 是各曲线的微分值。

3.3　分类判别

文中每种产地的白术训练样本采样了 15 次,则 3 种产地的白术训练样本总数为 45 个,而 PEN3 有 10 个传感器,每个传感器的测量值均提取了 8 个特征参数,因此总特征向量维数为 80 维,这时,显然训练样本总数小于特征向量维数,出现小样本问题,此时 LDA 算法无法进行下去。下图 2 是 3 组白术训练样本的 PCA 和 PCA + LDA 分析结果图。



(a)3 组训练样本的 PCA 分析图



(b)3 组训练样本的 PCA + LDA 分析图

图2 3组训练样本的分析结果图

从图2(a)中可以看出，单独采用 PCA 算法对 3 组训练样本的分类效果非常不理想，各批次的样本点交错在一起，难以区分。原因是当样品等级质量差别较小时，电子鼻中各传感器所能反映样品差异的响应信息存在较大的重叠性或相关性，PCA 算法寻找的只是数据分布的主轴方向，经降维后保留下来的信息不一定对分类最有效[7]。从图2(b)中可以看出，采用 PCA + LDA 方法对 3 组训练样本的区分效果明显优于单独采用 PCA 算法，原先交错在一起的训练样本点已经全部被明显地区分开来。这是因为 LDA 算法的主要思路就是使类内散布最小化、类间散布最大化。

为避免小样本问题，文中采用基于最大散度差准则的 LDA 算法。分析结果如图 3 所示。

从图 3 中可以看到，采用基于最大散度差准则的 LDA 方法的分类效果明显优于单独使用 PCA 算法和 PCA + LDA 算法。河北白术样本点主要集中在特征空间的下半部分，浙江白术样本点主要分布在左上部分，而安徽白术样本点则集中在右上部分。各类训练样本点都能明显地被区分开，而且相对于 PCA + LDA 方法，类内样本点的分布显得更为集中，类间分界面更加明显。

3.4 待测样本的鉴别

表 1 是在二维特征空间中实现的对各待测样本的分类鉴别结果，最终识别率为正确识别的待测样本数与待测样本总数之比。

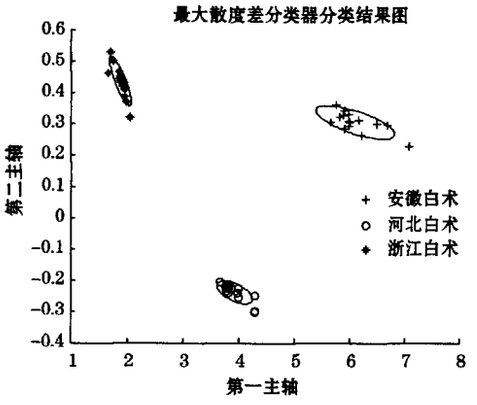


图3 3组训练样本的基于最大散度差准则的 LDA 分析结果图

表1 3组白术待测样本的鉴别结果

	待测样本总数	正确识别数	错误识别数	识别率/%
河北白术	15	15	0	100
安徽白术	15	14	1	93.3
浙江白术	15	15	0	100

结果表明：对于 45 个待测样本，有 1 个被错误地进行了判断，安徽白术的识别率是 93.3%，河北白术和浙江白术的识别率均为 100%，对所有待测样本的最终判别结果达到了 97.8% 的正确判别率。

4 结束语

运用电子鼻技术进行气味分析，不仅客观、准确，而且重复性好、方便快捷。文中利用 PEN3 型电子鼻对 3 种产地的白术样品进行了检测，采用了基于最大散度差准则的 LDA 分析方法对数据进行分析，解决了小样本问题，实现了 3 种产地白术的正确鉴别，对所有待测样本的正确识别率达到 97.8%，并且分类效果优于单独使用 PCA 或 PCA + LDA 算法。这为中药材品质的保证提供了一种有效的方法。

参考文献：

[1] 胡平. 中药材质量评价体系的方法学研究：[学位论文]. 上海：华东理工大学，2006.

[2] GHASEMI-VARNAMKHASTI M, MOHTASEBI S S, SIADAT M, et al. Meat quality assessment by electronic nose (Machine Olfaction Technology). Sensors, 2009, 9(8): 6058 - 6083.

[3] MAZZONE, PJ. Analysis of volatile organic compounds in the exhaled breath for the diagnosis of lung cancer. Journal of Thoracic Oncology, 2008, 3(7): 774 - 780.

[4] 黄小燕，赵向阳，方智勇. 电子鼻在气体检测中的应用研究. 传感器与微系统，2008，27(6)：47 - 50.

[5] 陈洪波，汤井田，陈真诚. 基于 PCA - LDA 的 HIFU 治疗中组织损伤无损检测方法. 中国生物医学工程学报，2008，27(6)：812 - 816.

[6] 宋枫溪，杨静宇，刘树海. 基于多类最大散度差的人脸表示方法. 自动化学报，2006，32(3)：378 - 385.

[7] 殷勇，田先亮. 基于 PCA 与 Wilks 准则的电子鼻酒类鉴别方法研究. 仪器仪表学报，2007，28(5)：39 - 43.

作者简介：邵雅雯(1985—)，硕士研究生，主要从事电子鼻应用的研究。
E-mail: yawen329@163.com