基于扩散映射和LDA的辛味中药材鉴别研究

陈惠勤，骆德汉
（广东工业大学 信息工程学院，广东 广州 510006）

摘 要：使用PEN3电子鼻对不同种类的辛味中药材进行检测，针对辛味中药材气味数据的高维性与非线性，提出一种基于扩散映射（Diffusion Maps）和线性判别分析（LDA）的中药材鉴别方法。首先将采集的气味数据重构到高维空间中，利用扩散映射方法对高维气味数据进行降维和特征提取，然后利用线性判别分析对提取的低维流形特征进行分类，以可视化的方式显示中药材的分类效果。结果表明，该方法可以很好地区分四种不同种类的辛味中药材，为中药材的分类鉴别提供了一条新的途径。

关键词：电子鼻；辛味中药材；扩散映射；线性判别分析；分类鉴别
中图分类号：TP212.9 文献标识码：A 文章编号：1673-629X(2015)05-0192-04
doi:10.3969/j.issn.1673-629X.2015.05.045

Research on Classification of Pungent Chinese Herbal Medicines Based on Diffusion Maps and LDA

CHEN Hui-qin, LUO De-han
(School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China)

Abstract: An Electronic nose (E-nose) was employed to collect the smell print of different kinds of pungent Chinese herbal medicines. For the high-dimensional and nonlinear of odor data, propose a method based on Diffusion Maps and LDA for the identification of pungent Chinese herbal medicines. Firstly, the collected odor data was reconstructed to the high-dimensional space, using the Diffusion Maps algorithm to extract the low-dimensional manifold features embedded in the high-dimensional odor data. Then the LDA algorithm was employed to implement a linear classifier. The results show that the combination of Diffusion Maps and LDA algorithm can well distinguish four different kinds of pungent Chinese herbal medicines, which provides a new way for the classification and identification of Chinese herbal medicines.

Key words: Electronic nose (E-nose); pungent Chinese herbal medicines; Diffusion Maps; LDA; classification and identification

1 概 述

我国使用中药材治疗疾病已有数千年的历史，在长期的医疗实践中，中医形成了其特有的理论体系，其临床疗效好，副作用小。中药材是中国的国宝，是中华民族繁衍昌盛，防病治病的主要药材，更是我国医疗体系中一个很重要的组成部分。近年来，一些人受利益的驱使，制造和销售假冒伪劣产品，鱼目混珠，扰乱市场，中药材掺假、掺杂、以假乱真等现象比较严重，直接影响到临床用药的有效性和人民的生命安全。所以，中药材的种类鉴别和质量鉴定是十分重要的。

中药材的颜色、气味和味道是其内在化学成分及其内在质量的反映，传统的基于人体感官的经验鉴别能够评价其质量，但与个人的情绪、经验、生理等因素有关，很难准确表述和交流。而气相色谱法（Gas Chromatography，GC）、质谱法（Mass Spectroscopy，MS）等理化方法所得的单一“有效成分”、“指标成分”又不能准确反映和评价中药材的质量。

近年来，随着人类对嗅觉过程理解的深入以及传感器技术的发展，基于模仿人类感觉传导机理开发的仿生嗅觉—电子鼻得到了深入的开发和应用。与传统的经验鉴别法以及理化方法相比，电子鼻具有灵敏、快

收稿日期：2014-06-23 修回日期：2014-09-25 网络出版时间：2015-04-22
基金项目：广东省自然科学基金资助项目（S2011020002906）
作者简介：陈惠勤（1990-），女，硕士研究生，研究方向为机器嗅觉、传感信号处理与模式识别；骆德汉，教授，博士生导师，博士，研究方向为绿色电子与智能信息处理。
网络出版地址：http://www.cnki.net/kcms/detail/61.1450.TP.20150422.0950.001.html

捷、准确、无损的优点,并且能够全面、完整地反映中药材的“整体气味特征”[1]。目前在国内外的研究中,电子鼻被广泛应用于食品、农业、环境监控、医学、生物及安全等领域[2-9]。

文中选用不同气味特性的辛味中药材作为实验研究对象,使用 PEN3 电子鼻对不同种类的辛味中药材进行气味信息采集。在对电子鼻获取的高维气味数据的降维与特征提取方面,常用的线性分析方法有主成分分析法(PCA)、独立分量分析法(ICA)、多维尺度变换(MDS)、线性判别分析(LDA)等[10]。但实际上,电子鼻获取的高维气味数据一般是非线性的,这些常用的线性维数约简方法很难发现复杂的非线性数据的内部结构,这就需要研究适用于处理非线性数据的新方法。

美国《Science》杂志于 2000 年发表了三篇与流形学习相关的论文[11-13],由于这种方法在复杂的高维非线性数据处理方面具有强大的优势,一经提出便很快成为一个新的研究热点。与传统的线性分析方法相比,流形学习方法能够很好地揭示高维非线性数据的本质结构,利于进行维数约简和非线性数据分析[14]。扩散映射(Diffusion Maps,DM)[15-18] 正是这样一种典型的流形学习方法,它来源于图的分割方法,将数据集从高维空间映射到低维空间时有聚类效果,所以比较适用于数据分类与识别。

针对中药材气味数据的高维性与非线性,以及因样本数过少而引发的维数灾难问题,文中提出了一种基于 DM 和 LDA 的中药材鉴别方法。首先利用 DM 方法对高维气味数据集进行学习,提取出数据内在的低维流形特征,然后利用 LDA 对提取的低维流形特征进行分类,以可视化的方式显示中药材的分类效果。文中将新兴的流形学习算法 DM 运用于仿生嗅觉气味数据信息的处理,结合传统的线性算法 LDA 共同探索中药材分类鉴别的新方法,在中药材的品种鉴别与质量鉴定研究领域具有广阔的应用前景。

2　实验材料、仪器和方法

2.1　材料与仪器

文中选用的中药材实验样品是由广东药学院提供的,选取了四种不同种类的辛味中药材:肉桂、化橘红、淡竹叶、鸡骨草。

实验采用的 PEN3 便携式电子鼻仪器是由德国 AIRSENSE 公司制造的。该电子鼻包含 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比值 G/G_0,其中,G 为传感器经样品挥发性气体作用后的电导率,G_0 为传感器经过标准活性碳过滤的基准气体作用后的电导率。

2.2　实验条件与方法

在实验中,实验室内温度保持在 24 ~ 27 ℃,相对湿度保持在 50% ~60%。实验样品重量为 10 g,样品静置时间为 60 min,采用静态顶空抽样的方法用 PEN3 进行检测,静态顶空空间大小为 250 ml。样品的采样时间设为 120 s,传感器阵列的清洗时间设为 120 s,采样间隔设为 1 s。四种不同种类的辛味中药材,每种中药材样品连续采样 12 次,共得到 48 个独立的样本。

3　扩散映射和线性判别分析方法的数学原理

3.1　扩散映射(Diffusion Maps)

扩散映射方法来源于动态系统理论,它在数据点相对应的图上定义 Markov 随机路,通过若干时间的随机游走,得到任意两个数据点的距离函数,这个距离函数定义为扩散距离,即它的主要思想是以尽可能保持扩散距离不变来获取低维流形的结构[14]。

对于给定的一个高维数据集 $X = \{x_1, x_2, \cdots, x_N\}$,$x_i \in R^D$,从中提取出一个 d 维 ($d < D$) 的流形特征 Y,扩散映射方法的实现步骤如下。

(1) 根据给定的数据点 x_i 建立一个与之相对应的图,利用高斯核函数定义图中边的权值,从而得到权值矩阵 $\boldsymbol{W}$,其元素为

$$w_{ij} = \exp\left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right) \tag{1}$$

式中,σ 为高斯核的方差。

(2) 通过归一化技术,使矩阵 $\boldsymbol{W}$ 的每行之和单位化为 1,则可以得到归一化的权值矩阵 $\boldsymbol{P}^{(1)}$,其元素为

$$P_{ij}^{(1)} = \frac{w_{ij}}{\sum_k w_{ik}} \tag{2}$$

此矩阵的每行可看作是对应数据点随机游走到其他数据点的概率,则 $\boldsymbol{P}^{(1)}$ 为任意两个数据点之间的一步转移概率矩阵。经过 t 步随机游走,相应的转移概率矩阵为 $\boldsymbol{P}^{(t)} = (\boldsymbol{P}^{(1)})^t$,两点 x_i 与 x_j 之间 t 步后的扩散距离定义为

$$D^{(t)}(x_i, x_j) = \sqrt{\sum_k \frac{(p_{ik}^{(t)} - p_{jk}^{(t)})^2}{\varphi(x_k)}} \tag{3}$$

式中

$$\varphi(x_i) = \frac{m_i}{\sum_j m_j},\ m_i = \sum_j p_{ij} \tag{4}$$

$\varphi(x_i)$ 说明了图中高密度区代表的权值更大。从式(3)可以得出,图中的点分布越密集,成对数据点之间的扩散距离越小。扩散距离考虑了连接两点所有边的贡献,所以比测地线距离的鲁棒性更强。

(3) 在保持扩散距离不变的前提下,提取低维流

形特征 $\boldsymbol{Y}$。根据 Markov 随机路的谱图理论，$\boldsymbol{Y}$ 由式(5)的 d 个非平凡主特征向量构成

$$\boldsymbol{P}^{(t)}\boldsymbol{Y}=\lambda\boldsymbol{Y} \tag{5}$$

由于是全连接图，最大的特征值(即 $\lambda_1=1$)是平凡的，其对应的特征向量 $\boldsymbol{\nu}_1$ 应舍弃。低维流形 $\boldsymbol{Y}$ 由其余的 d 个主特征向量给出，即

$$\boldsymbol{Y}=\{\lambda_2\boldsymbol{\nu}_2,\lambda_3\boldsymbol{\nu}_3,\cdots,\lambda_{d+1}\boldsymbol{\nu}_{d+1}\} \tag{6}$$

3.2 Diffusion Maps +LDA 算法

扩散映射方法具有良好的学习高维非线性复杂数据内在流形结构的能力，可以很好地恢复原始高维非线性数据的本征低维空间，但该低维空间未必是最好的识别空间。因此，有必要在低维空间中再对数据进行分析，将数据投影到最好的识别空间中去，在最好的识别空间中进行分类识别。

线性判别分析方法的突出优点是能够保证模式样本投影后在新的低维空间中有最大的样本类间散布距离以及最小的样本类内散布距离，即模式样本在新的空间中有最好的分离性。然而，当选取的特征总数较多或样本总数较少时，直接采用 LDA 算法来处理数据会出现小样本问题，LDA 算法将无法进行下去。

因此，文中综合扩散映射和线性判别分析的特点，采用由扩散映射算法得到的特征矩阵作为线性判别分析算法的输入矩阵，即 Diffusion Maps+LDA 算法来分析中药材的高维非线性气味数据，实现中药材的分类鉴别。

4 结果与讨论

4.1 传感器响应

PEN3 电子鼻对四种不同种类辛味中药材的响应特性如图 1 的柱状图所示。

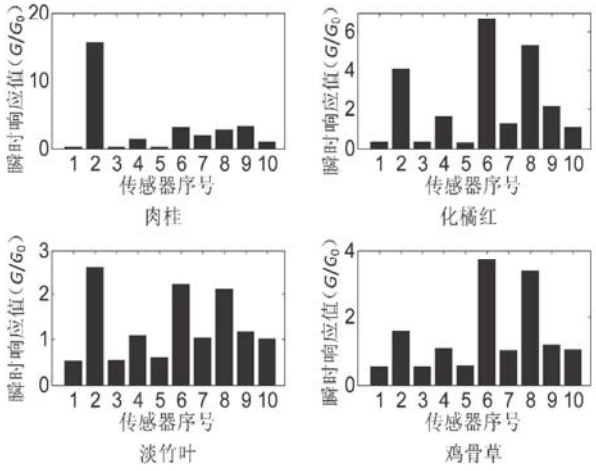

图 1 四种不同种类辛味中药材的响应特性柱状图

从图 1 中可以看出，PEN3 电子鼻的 10 个传感器对不同种类的辛味中药材具有不同的响应特性，每个传感器对不同种类的辛味中药材均有一定的响应，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。由于肉桂和化橘红属于气味特殊的典型辛味中药材，而淡竹叶和鸡骨草属于气味微弱的典型辛味中药材，从图 1 中可以看出传感器 S2 对肉桂的气味最灵敏，响应值明显高于其他的中药材；传感器 S6 和 S8 对化橘红的气味比较灵敏，响应值高于其他的中药材。

4.2 四种不同种类辛味中药材分类结果

对 PEN3 电子鼻采集的四种不同种类辛味中药材的气味数据分别用 Diffusion Maps 算法和 Diffusion Maps+LDA 算法进行处理，分类结果如图 2 和图 3 所示。

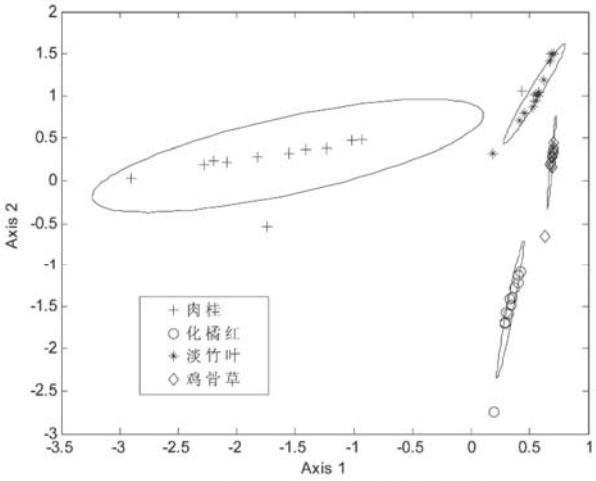

图 2 四种不同种类辛味中药材的 Diffusion Maps 分类结果图

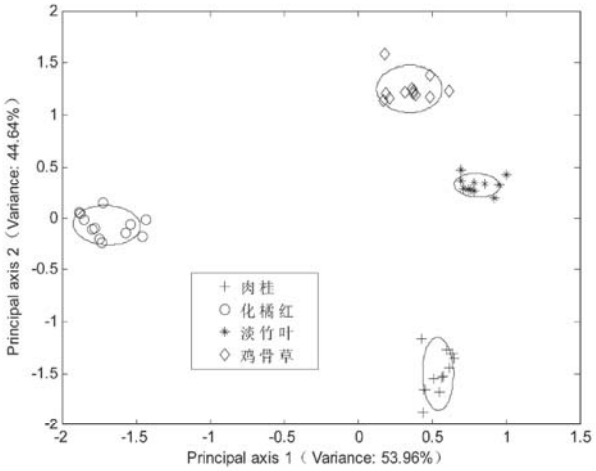

图 3 四种不同种类辛味中药材的 Diffusion Maps+LDA 分类结果图

图 2 中，直接用 Diffusion Maps 算法对高维气味数据进行学习，提取出高维数据内在的低维流形特征，图中横轴表示由 Diffusion Maps 算法得到的第二特征值 λ_2 与对应特征向量 $\boldsymbol{\nu}_2$ 的乘积，纵轴表示第三特征值 λ_3 与对应特征向量 $\boldsymbol{\nu}_3$ 的乘积。在二维空间中可观察到，肉桂的各样本点分布地比较离散，其他三种中药材

有个别样本点远离分布中心。从整体上来说,各类样本类内距离较大,类间距离较小,分类效果不是很理想。

图 3 中,首先利用 Diffusion Maps 算法提取高维气味数据中的低维流形特征,然后利用 LDA 算法对提取的低维流形特征进行分类,图中横轴表示第一主轴,方差贡献率为 53.96%,纵轴表示第二主轴,方差贡献率为 44.64%。在二维空间中可以看到肉桂各样本点间距离明显缩小,其他三种中药材的各样本点也分布地较为集中,各类样本类内距离缩小,类间距离加大,即 Diffusion Maps+LDA 算法可以很好地区分这四种不同种类的辛味中药材,与直接运用 Diffusion Maps 算法相比,分类效果得到了一定的改善,这是因为 LDA 算法的主要优点就是使类间散布距离最大化,类内散布距离最小化。

在四种不同种类辛味中药材样本的分类实验中,每种中药材各有 12 个样本,即共有 48 个样本。每个样本是 120* 10 的矩阵,可以看作是一个维数为 1 200 的样本点。在分别使用 Diffusion Maps 算法和 Diffusion Maps+LDA 算法对样本高维气味数据进行处理时,需要设置两个重要的参数,一个是高斯核函数的方差 σ,另一个是本征维数即嵌入空间的维数 d。σ 和 d 的选取在算法中是关键因素,对嵌入结果有很大的影响。文中是通过 MATLAB V.7.10 (R2010a) 来分析和处理中药材的高维气味数据,实验经过多次仿真、比较和分析,最终确定了在单独使用 Diffusion Maps 算法时,选择的最优参数为 $\sigma=100$, $d=2$;在使用 Diffusion Maps +LDA 算法时,选择的最优参数为 $\sigma=100$, $d=10$。

5　结束语

文中利用 PEN3 电子鼻对不同种类的辛味中药材进行了检测,针对高维非线性气味数据的处理困难,提出了一种基于扩散映射和线性判别分析的中药材鉴别方法。

在四种不同种类辛味中药材的分类中,Diffusion Maps+LDA 算法可以很好地区分这四种不同种类的辛味中药材,并且分类效果要好于单独使用扩散映射算法。结果表明了扩散映射算法的可行性和有效性,体现了线性判别分析方法能使样本类内散布最小化以及类间散布最大化的优点,更充分说明了 Diffusion Maps +LDA 算法的优越性,该算法为辛味中药材的分类鉴别提供了一条新的和有效的途径。

参考文献:

[1] 邹慧琴,韩　玉,邢　姝,等. 电子鼻技术及其在中药领域中的应用 [J]. 世界科学技术-中医药现代化,2012,14(6):2120-2125.

[2] Santonico M, Bellincontro A, de Santis D, et al. Electronic nose to study postharvest dehydration of wine grapes [J]. Food Chem,2010,121(3):789-796.

[3] 贾洪锋,王　鑫,邓　红,等. 电子鼻在食品气味分析中的应用 [J]. 粮油食品科技,2013,21(1):38-42.

[4] 江琳琳,潘磊庆,杨虹贤,等. 电子鼻在果蔬品质检测中的研究进展 [J]. 安徽农业科学,2010,38(23):12918-12920.

[5] Lamagna A, Reich S, Rodriguez D, et al. The use of an electronic nose to characterize emissions from a highly polluted river [J]. Sensors and Actuators B: Chemical,2008,131(1): 121-124.

[6] Steiner L A, Pfister D, Strebel S P, et al. Near-infrared spectroscopy can monitor dynamic cerebral autoregulation in adults [J]. Neurocritical Care,2009,10(1):122-128.

[7] Sorend F A, Serrador J M, Jones R N, et al. The sit-to-stand technique for the measurement of dynamic cerevral autoregulation [J]. Ultrasound Med Biol,2009,35(1):21-29.

[8] Yinon J. Field detection and monitoring of explosives [J]. Trends in Analytical Chemistry,2002,21(4):292-301.

[9] Alizadeh T, Zeynali S. Electronic nose based on the polymer coated SAW sensors array for the warfare agent simulants classification [J]. Sensors and Actuators B: Chemical, 2008, 129 (1):412-423.

[10] 刘红秀,骆德汉,张泽勇. 机器嗅觉系统气味识别算法 [J]. 传感技术学报,2006,19(6):2518-2522.

[11] Seung H S, Daniel D L. The manifold ways of perception [J]. Science,2000,290(5500):2268-2269.

[12] Tenenbaum J B, de Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction [J]. Science,2000,290(5500):2319-2323.

[13] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding [J]. Science,2000,290(5500):2323 -2326.

[14] 夏鲁瑞,胡笃庆,秦国军. 基于流形学习的涡轮泵海量数据异常识别算法 [J]. 航空动力学报,2011,26(3):698-703.

[15] Coifman R R, Lafon S. Diffusion maps [J]. Applied and Computational Harmonic Analysis,2006,21(1):5-30.

[16] Nadler B, Lafon S, Coifman R R, et al. Diffusion maps, spectral clustering and the reaction coordinates of dynamical systems [J]. Applied and Computational Harmonic Analysis,2006,21 (1):113-127.

[17] Lafon S, Lee A B. Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2006,28(9):1393 -1403.

[18] Coifman R R, Kevrekidis I G, Lafon S, et al. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems [J]. Multiscale Model and Simulation, 2008,7(2):842-864.