几种模式识别方法用于中药挥发油红外光谱法鉴别的比较研究Δ

邱新华[1*]，唐铁鑫[1#]，刘　燕[1]，吴美珠[1]，谭雄斯[1]，甘柯林[1]，姚伟生[2]（1.肇庆医学高等专科学校，广东 肇庆 526020；2.肇庆市中医院，广东 肇庆　526020）

中图分类号　R284.1　　**文献标志码**　A　　**文章编号**　1001-0408（2015）21-2986-03
DOI　10.6039/j.issn.1001-0408.2015.21.38

摘　要　目的：比较几种模式识别方法在中药挥发油红外光谱法鉴别中的分类效果。方法：对多种忍冬属和柑橘属中药的挥发油进行红外光谱测定，应用系统聚类、K-均值聚类、人工神经网络、支持向量机方法对样品红外光谱进行分类。结果：系统聚类与K-均值聚类分类效果不佳，人工神经网络和支持向量机方法均取得100%分类正确率。结论：可以将人工神经网络和支持向量机模式识别方法与红外光谱法结合，构建化学计量学指纹图谱技术，用于中药挥发油的鉴别。

关键词　中药；挥发油；红外光谱法；模式识别；化学计量学指纹图谱技术

Comparative Study of Several Pattern Recognition Methods in the Identification of Volatile Oils of Traditional Chinese Medicine by Infrared Spectroscopy

QIU Xin-hua[1], TANG Tie-xin[1], LIU Yan[1], WU Mei-zhu[1], TAN Xiong-si[1], GAN Ke-lin[1], YAO Wei-sheng[2]（1.Zhaoqing Medical College, Guangdong Zhaoqing 526020, China; 2.Zhaoqing Hospital of Traditional Chinese Medicine, Guangdong Zhaoqing 526020, China）

ABSTRACT　OBJECTIVE: To compare the performance of several pattern recognition methods in the identification of volatile oils of traditional Chinese medicine (TCM) by infrared spectroscopy. METHODS: The volatile oils of several *Lonicera* and *Citrus* TCM were determined by infrared spectroscopy. All samples of infrared spectrum were classified by hierarchical clustering, K-mean clustering, artificial neural networks, and support vector machine. RESULTS: The results of hierarchical clustering and K-mean clustering were ineffective. Methods of artificial neural networks and support vector machine achieved correct classification rate of 100%. CONCLUSIONS: Artificial neural networks and support vector machine can be combined with infrared spectroscopy to create chemometric fingerprinting for the identification of volatile oils of TCM.

KEYWORDS　Traditional Chinese medicine; Volatile oils; Infrared spectroscopy; Pattern recognition; Chemometric fingerprinting

红外光谱鉴别法专属性很强，广泛应用于化学药物的鉴别。中药挥发油是混合物，红外光谱信息复杂、变异大，仅凭人工比对很难区分。很多研究报道采用化学计量学方法与红外光谱法相结合，用于中药的鉴别[1-7]，其中最常使用的化学计量学方法是基于欧氏距离、夹角余弦、Pearson（泊松）相关系数等距离（或相似度）量度的相似度分析或聚类分析方法[1-5]，缺少应用人工神经网络和支持向量机模式识别方法的研究。本试验比较了几种模式识别方法在多种忍冬属和柑橘属中药挥发油红外光谱法鉴别中的判别效果，报道如下。

1　材料

1.1　仪器和分析软件

用IRsolution 150SU1工作站软件控制的IRAffinity-1傅里叶红外光谱仪来自日本岛津公司。数据文件的格式转换使用自编程序（用Java语言在Eclipse 4.20集成开发软件下开发）。K-均值聚类、系统聚类、人工神经网络分析使用SPSS 21软件试用版（美国IBM公司）。支持向量机分析使用Libsvm 3.17软

ified by a diafiltration centrifugal device and tangential flow filtration[J]. *Drug Dev Ind Pharm*,2008,34(12):1 331.
[16] 李红茹，李淑芬.脂质体中药物包封率的测定方法[J].药物分析杂志，2008，27(11)：1 844.
[17] 贾乐姣，张典瑞，王言才.水飞蓟宾纳米结构脂质载体中主药含量及包封率测定[J].中国药学杂志，2009，44(18)：1 400.
[18] 张素娟，张永太，申利娜，等.蟾酥固体脂质纳米粒包封率测定[J].中国新药杂志，2013，22(12)：1 465.
[19] 施峰，王岚，施晓琴，等.乳香没药挥发油固体脂质纳米粒包封率的评价[J].中国新药杂志，2013，22(14)：1 709.
[20] 王坤，杨磊，王艳，等.奥扎格雷纳米结构脂质载体包封率的测定方法[J].中国新药杂志，2012，21(22)：2 693.
[21] 周晖，邱立朋，龙苗苗，等.奥沙利铂纳米结构脂质载体中主药含量及包封率测定[J].中国新药杂志，2011，20(11)：1 030.

Δ 基金项目：广东省中医药局科研课题（No.20122071）

* 实验师。研究方向：药物分析实验方法学。电话：0758-2826327。E-mail：cindyq5@sina.com
# 通信作者：制药工程师，博士。研究方向：天然药物产品开发和质量控制。电话：0758-2826327。E-mail：netscaner@126.com

（收稿日期：2015-03-23　修回日期：2015-06-23）
（编辑：周　箐）

件。

1.2 试剂

无水硫酸钠、氯化钠、乙醚为分析纯；溴化钾为光谱纯；蒸馏水为实验室制备。

1.3 药材

忍冬科忍冬属植物药材样品（均为山银花）共12批，包括3批灰毡毛忍冬、3批黄褐毛忍冬、3批华南忍冬、3批红腺忍冬，分别购自湖南省郴州市、广西省桂林市、广东省肇庆市和广州市等地；芸香科柑橘属植物药材样品共11批，包括4批柚皮（药材名：化橘红）样品、4批陈皮样本、3批青皮样本，分别购自广东省肇庆市、广州市。上述样品经肇庆医学高等专科学校唐铁鑫博士鉴定为各药材正品，标本存于肇庆医学高等专科学校；样品其余部分粉碎、过40目筛，用于挥发油提取。

2 方法与结果

2.1 样品制备

取30.0 g样品粗粉，置于挥发油提取装置中，加适量无水硫酸钠脱水，放置备用[8]。测定前以半径为5 cm、5 000 r/min离心5 min，取上清液混合溴化钾压片。当挥发油量较少时，用乙醚萃取芳香水，乙醚层用无水硫酸钠脱水，同法离心挥干乙醚后，溴化钾压片。

2.2 傅里叶变换红外光谱测定方法

将空白溴化钾片放入红外光谱仪，扫描空白光谱图，扣除空白，再将载有样品的溴化钾片放入红外光谱仪，扫全谱数据，并进行基线扣除、平滑和标准化预处理。将光谱图数据导出成文本数据。

2.3 数据分析

从各样本红外光谱数据中选取1 800～500 cm^{-1}之间的吸光度数值作为有序的一维数组数据集用于计算机模式识别分类。用于支持向量机分析的数据按照Libsvm软件的数据格式要求用自编软件进行转换。各模式识别方法分析中，试验不同参数设置以获得最佳的分类正确率。

2.4 几种中药挥发油的红外光谱

11批柑橘属植物药材、12批忍冬属植物药材的挥发油红外光谱图详见图1。参考文献[9]进行峰归属，图1A中2 924 cm^{-1}左右的吸收峰主要为饱和烷烃C—H的伸缩振动吸收峰，1 456 cm^{-1}左右为饱和烷烃C—H的弯曲振动吸收峰，3 000、1 680、887 cm^{-1}左右呈现了不饱和烷烃的吸收峰，主要来自于柠檬烯，但也混合了其他烯烃成分的吸收峰；图1B中2 924 cm^{-1}左右也呈现了饱和烷烃C—H的伸缩振动吸收峰，1 462 cm^{-1}左右为饱和烷烃C—H的弯曲振动吸收峰，3 000、1 713 cm^{-1}左右呈现了不饱和烷烃的吸收峰。两图在2 357cm^{-1}左右都呈现背景扣除不完全产生的空气中二氧化碳的吸收峰。其他位置处的吸收峰相互混杂、干扰，很难归属。从上述结果可以看出，挥发油成分复杂，但也主要以饱和烷烃和不饱和烷烃为主，而具有特征性的较弱吸收峰互相混杂，难以归属。如果依靠人工对比对各样本进行分类，难以客观地进行分类。因此，笔者研究选取特征性较强区域的红外光谱数据，借助计算机模式识别进行分类。从图中可以看到高于3 000 cm^{-1}和低于500 cm^{-1}都有较大的噪音干扰，2 800～1 800 cm^{-1}之间较为平坦而且有二氧化碳峰干扰，因此笔者选取1 800～500 cm^{-1}之间的红外光谱数据用于计算机模式识别分类。

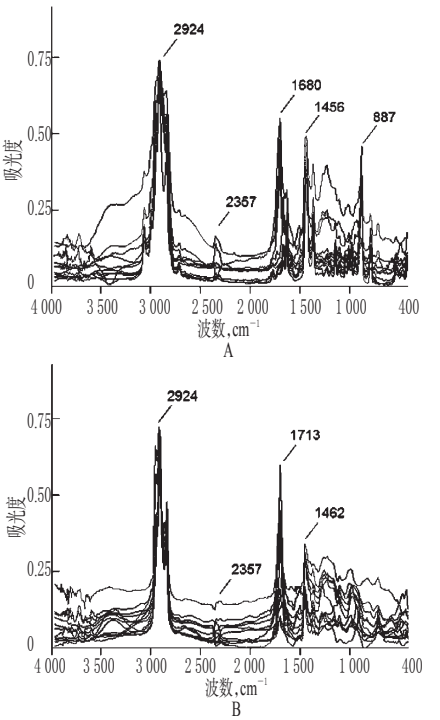

图1 药材样品挥发油的红外光谱图

A.11批柑橘属植物药材；B.12批忍冬属植物药材

Fig 1 Infrared spectroscopy of the volatile oils of TCM

A.11 batches of *Citrus* medicines; B.12 batches of *Lonicera* medicines

2.5 模式识别结果

2.5.1 聚类分析 系统聚类分析是基于样本间相似度量度的间接聚类方法，属于非监督学习方法[10]。使用该方法，一方面观察其能否直接实现两类挥发油的分类，另外一方面也为训练样品集进行监督学习的分类器设计提供参考。本研究系统聚类分析组合尝试了欧氏距离、夹角余弦、Pearson（泊松）相关系数、马氏距离等距离（或相似度）量度方式，组间联接和组内联接聚类方法，都没有能完全将23个样品准确分类。如果分成两类，正确率最高为采用夹角余弦量度和组内联结聚类方法，其他参数采用默认设置，有21个样本分类正确，正确率为91.3%，详见表1、图2。

表1 不同模式识别方法对样品识别的结果

Tab 1 Recognition results of the samples by different pattern recognition methods

模式识别方法	识别正确的样本数	识别错误的样本数	正确率,%
系统聚类分析	21	2	91.3
K-均值聚类分析	13	8	56.5
人工神经网络（多层感知器）	23	0	100
支持向量机	23	0	100

K-均值聚类分析是比系统聚类分析较为简洁和快速的聚类方法。本试验中K-均值聚类分析选择了不同的方法和迭代参数对样本进行分类，将样本分成两类，正确率最高为采用迭代与分类方法，迭代10次或20次，其他采用默认设置；都是仅有13个样本分类正确，正确率为56.5%（详见表1），表明用简洁算法的K-均值聚类分析效果更差。事实上，随着计算机运算速度的提高，进行红外光谱数据的模式识别，即使使用系统聚类分析，分析速度也非常快，没有必要采用K-均值聚类分析

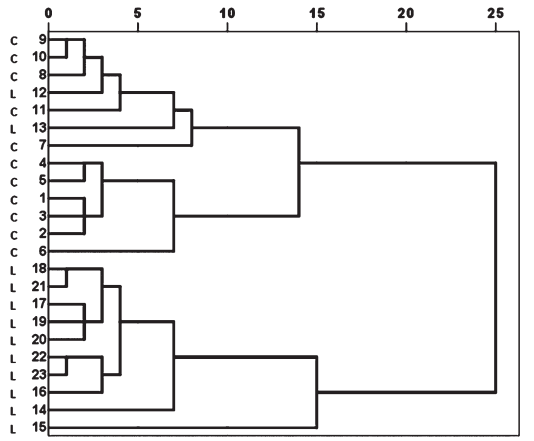

图2　药材样品挥发油的红外光谱数据系统聚类分析图

C.柑橘属植物药材挥发油样品；L.忍冬属植物药材挥发油样品

Fig 2　Cluster analysis chart of infrared spectroscopy data of the volatile oils from samples of TCM

C.volatile oils from samples of *Citrus* medicines; L.volatile oils from samples of *Lonicera* medicines

来提高分析速度。

聚类分析结果提示，基于样本间相似度量度的模式识别方法，在红外光谱特征受到较大干扰时难以有效地区分不同类的中药挥发油。因此，需要用适合复杂样品分类的模式识别方法。

2.5.2　人工神经网络　人工神经网络是基于经验风险最小化的监督学习方法，适用于复杂样本的分类[10]。其原理是通过用已知样本培训出识别模型，再用于未知样本的识别，可以在不知道样本特征的前提下实现样本分类。本研究中，采用多层感知器方式，在忍冬属植物药材样本中设置6个样本用于培训、3个样本用于测试、3个样本用于支持（避免神经网络模型记住所有的样本的数据特征造成虚假正确率）；柑橘属药材样本中设置5个样本用于培训、3个样本用于测试、3个样本用于支持；体系结构采用自动体系结构选择、隐藏层中单位数为1～50个，其他参数采用缺省设置，将样本分成两类，得到分类正确率为100%（见表1）。笔者又尝试了人工设置体系结构和算法参数，均未能得到100%的正确率。采用径向基函数方式时，无论采用缺省参数设置或者自己调整参数设置，均无法得到100%分类正确率。结果表明，采用合理的人工神经网络方法比系统聚类法的分类效果更好。

2.5.3　支持向量机　支持向量机是基于结构风险最小化的监督学习方法，也适用于复杂样本的分类[10]。笔者将8个山银花样本和7个柑橘属药材样本用于培训，得到培训模型用于对所有样品进行分类。并尝试了不同的SVM（支持向量机）类型和和函数类型的组合，对-g[核函数中的函数设置（gama）]、-c（C-SVC的惩罚系数C）进行过优化。当SVM类型为nu-SVC，核函数类型为sigmoid核，并将nu-SVC的nu参数设置为0.25，其他参数采用缺省值的时候，培训得到的模型对样品的分类正确率达到了100%（见表1）。以上结果表明，采用合适的支持向量机方法可以得到与人工神经网络方法一样的判别效果。

3　讨论

基于相似度量度的聚类分析，数学算法简单、容易应用，但是中药挥发油红外光谱数据信息复杂，仅仅将复杂的数据通过相似度计算投射到一维数轴，再以其数值大小为基础进行聚类分析识别样品，不能获得最佳的分类正确率。而K-均值聚类分析为提高运算速度而优化了聚类分析过程，识别结果更加不理想。

人工神经网络和支持向量机方法算法复杂、较难掌握和运用，但是它们对复杂数据的数学建模和识别能力更强、更准确。人工神经网络要求用尽量多的样本来训练模型以减少经验风险，当样本数量较少时容易陷入局部最优化而影响对未知样本的识别；支持向量机是基于结构风险最小化设计的，在样本数量较少时通常也能获得较好的效果。在本研究中，用人工神经网络和支持向量机方法进行的模式识别，正确率都达到了100%，显示两种方法效果一致。

因此，对于利用红外光谱法构建中药指纹图谱，用相似度计算以及聚类分析的方法较难获得满意效果；将人工神经网络和支持向量机模式识别方法与红外光谱法结合，构建中药挥发油化学计量学指纹图谱技术，能够提高识别正确率，更适合用于建立中药挥发油的红外快速鉴别方法。

参考文献

[1] 张石楠，张桂芝，张立.中药饮片挥发油的红外指纹图谱研究[J].现代中药研究与实践，2009，23(1)：25.

[2] 周晔，李佩孚，张庆伟，等.傅里叶红外光谱法鉴别部分黄精属生药的研究[J].光谱学与光谱分析，2013，33(7)：1 791.

[3] 陈灶鑫，徐永群，陈小康，等.灵芝及灵芝提取物红外光谱规律的研究[J].光谱学与光谱分析，2013，33(5)：1 206.

[4] 何翠薇，谢艳林，欧庆勇.傅里叶变换红外光谱法鉴别肉桂药材模型研究[J].时珍国医国药，2013，24(10)：2 430.

[5] 王孝勋，李丹丹，姚德惠，等.广西不同产地对叶百部红外光谱学研究[J].时珍国医国药，2013，24(2)：318.

[6] 袁玉峰，陶站华，刘军贤，等.红外光谱结合主成分分析鉴别不同产地黄柏[J].光谱学与光谱分析，2011，31(5)：1 258.

[7] 孙仁爽，金哲雄，张哲鹏，等.犁牛儿苗科11种中药材红外光谱鉴定及聚类分析[J].光谱学与光谱分析，2013，33(2)：371.

[8] 国家药典委员会.中华人民共和国药典：一部[S].2010年版.北京：中国医药科技出版社，2010：附录63-64.

[9] 周欣，孙素琴，黄庆华.FTIR对不同产地陈皮的鉴别研究[J].光谱学与光谱分析，2007，27(12)：2 453.

[10] 张学工.模式识别[M].北京：清华大学出版社，2010：3-20.

（收稿日期：2014-10-28　修回日期：2015-01-11）

（编辑：余庆华）