基于三维荧光光谱特征的中药药性模式识别研究

樊凤杰[1]，轩凤来[1]，白 洋[1]，纪会芳[2]

1. 燕山大学电气工程学院，河北 秦皇岛 066004
2. 联勤保障部队第九八四医院，北京 100094

摘 要 由于三维荧光光谱技术选择性好、灵敏度高、测试快速等优点目前已在众多领域中被广泛应用。中药药性理论是中药的核心基础理论，是中药学的特色理论之一，中药药性的客观化判别是中医药现代化研究的关键问题。中药中大部分分子具备产生荧光的能力，因而，针对中药三维荧光光谱特征，从中药药性的角度对中药进行分类识别研究。利用FS920型稳态荧光光谱仪测得5组不同浓度的23味寒温类中药溶液制剂的三维荧光光谱数据，获取样本的等高线图和三维荧光光谱图；分析不同样本不同激发波长和发射波长范围存在噪声的基础上，应用集合经验模态分解算法(EEMD)对光谱图进行降噪预处理；基于局部线性嵌入算法(LLE)对光谱数据进行特征提取，分析近邻点数$k=12$，本征维数$d=7$时得到的特征向量，结果表明不同浓度的寒性药在PC4和PC6的特征值变化明显，不同浓度的温性药在PC1，PC2，PC4和PC7的特征值变化明显，且浓度越高特征值都有下降趋势。将提取的特征向量输入到随机森林(RF)中，构建LLE-RF分类模型，分析不同参数时LLE-RF分类模型对寒温类中药荧光光谱数据的分类效果，设置RF分类器中训练集和测试集的样本比例分别为3∶1和2∶1，即训练集的比重r分别为3/4和2/3，分析LLE中近邻点数k取值为7～18，本征维数d分别取值为6，7，8，9和10时分类正确率。当近邻点数$k=12$，本征维数$d=7$时LLE-RF模型对中药药性的分类正确率最高，达到96.6%。最后比较同一比例r情况下，采用不同核函数构造SVM分类器对寒温类中药荧光光谱数据分类效果，当多层感知机作为核函数时，分类效果最差。当$r=3/4$，径向基作为核函数时，寒温类中药荧光光谱数据的分类效果最好，正确率达到82.1%。分析结果表明，通过荧光光谱技术与LLE-RF相结合的方法，能有效的将寒温类中药进行模式识别，并且分类效果比LLE-SVM更理想。

关键词 三维荧光光谱；特征提取；中药药性；局部线性嵌入；随机森林

中图分类号：TB96 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)06-1763-06

引 言

中药是我国中华民族的瑰宝，经过几千年的积累和研究，中药在临床应用中发挥着巨大的作用。中药药性是对中药性质与功能的高度概述，是中医药理论的核心，从整体角度了解和研究中药药性，对中医药理论的发展和传承具有重大意义。近年来，许多学者在药性模式识别、中药药性组合与药性及功效的关系、药性表征模式等领域进行了深入研究[1-3]。王晓燕等采用GC-MS技术对寒热性药物进行检测，通过不同模式识别方法建立了药性的判别模型[4]。吴思媛等采用RF和SVM等方法对寒热类中药进行分类，结果显示，RF分类正确率为87.5%，SVM分类正确率为92.5%，RPART分类正确率为95.2%[5]。陈昭等借助LS-SVM算法建立了以药性为基础的清热药分类模型，其正确率达到79.2%[6]。

本文将三维荧光光谱技术应用到中药药性模式识别领域，针对光谱数据的非线性特征，利用局部线性嵌入算法(local linear embedding，LLE)对寒性和温性中药光谱数据进行特征提取，并结合随机森林(random forest，RF)、支持向量机(support vector machine，SVM)分别建立LLE-RF、LLE-SVM分类识别模型，研究不同分类模型对中药药性的分类识别效果。

收稿日期：2019-05-23，**修订日期**：2019-09-16

基金项目：国家自然科学基金项目(61201111)，燕山大学博士基金项目(BL17026)资助

作者简介：樊凤杰，1977年生，燕山大学电气工程学院副教授 e-mail：ffjmz@126.com

1 基本原理

1.1 LLE 算法

LLE 算法是针对非线性数据的一种降维技术，且能够使降维后的数据保持原有的拓扑结构。该算法是假设高维数据集中的每一个数据点都可以用它临近的若干个数据点近似线性表示，将整个高维数据集分解成若干个具有线性特征的流形区域，并寻求最优权值映射矩阵，来最小化数据集重构后的误差，从而达到降维的目的。LLE 算法主要步骤如下：

(1)确定近邻域，选取近邻点。采用 k 邻域法，以欧式距离为度量标准，选取样本 x_i 的 k 个近邻点；

(2)样本点局部重构，计算权值矩阵 $\boldsymbol{W}$。确定好 K 邻域后，在此邻域内利用 x_i 的近邻点及 x_i 与近邻点之间的权值 W_{ij} 近似表达 x_i。然后对所有 x_i 做同样的计算，利用 W_{ij} 构造局部重建权值矩阵 $\boldsymbol{W}$，并满足重构误差 $\varepsilon(w)$最小，即

$$\varepsilon(w) = \sum_{i=1}^{n} \left| x_i - \sum_{j=1}^{n} w_{ij} \times x_j \right|^2 \tag{1}$$

当 x_j 不属于 x_i 的 K 个近邻点之一时，会出现 $W_{ij}=0$，W_{ij} 的第二个约束条件为

$$\sum_{j=1}^{n} w_{ij} = 1 \tag{2}$$

(3)寻求最优映射，计算样本点低维空间输出。对每个样本点 x_i 计算高维到低维流行的映射 y_i，最小化加权误差 $\varepsilon(y)$

$$\varepsilon(y) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{n} w_{ij} \times y_j \right|^2 \tag{3}$$

式(3)中具有限制条件

$$\sum_{j=1}^{n} y_i = 0 \tag{4}$$

$$\frac{1}{n} \sum_{i=1}^{n} y_i \times y_i^{\mathrm{T}} = 1 \tag{5}$$

1.2 RF 算法

随机森林(random forest，RF)是 2001 年 Breiman 提出的，其基本思想是用 bootstrap[7]方法从原始样本中抽取多个子样本，对每个子样本进行决策树建模，再利用投票法或平均法组合多棵决策树的预测结果来决定最终预测结果。该方法具有更好的噪声容忍度及更高的预测准确率，且不容易出现过拟合问题[8]。建立 RF 的具体步骤如下[9-10]：

步骤 1：训练数据抽样。设原始样本集的大小为 N，从原始样本集中随机可放回地抽取 n 个样本作为新的训练集。

步骤 2：属性子空间抽样。随机地从 M 个原始属性中选取 m 个属性形成新的属性子空间。

步骤 3：决策树模型建立。根据 CART 算法构建树，每棵决策树都完整生长，直到叶子子节点。

步骤 4：利用“森林”中每棵决策树对测试样本进行测试，得到 T 个对应的分类结果。

步骤 5：采用投票方法，将 T 个对应的分类结果中最多的类别作为该测试样本最终的类别归属。

2 实验部分

光谱数据采集仪器为英国 Edinburgh Instruments 公司生产的 FS920 型稳态荧光光谱仪，测量时将积分时间设为 0.1 s，激发波长 EX 扫描范围 220～550 nm，发射波长 EM 扫描范围 240～570 nm，激发波长和发射波长的采样波长间隔均为 10 nm。选取补虚类中药 23 味，23 味中药中只包含寒性和温性两类药性的药物，因此，本文仅对寒性和温性药物进行分析，其中寒性药物有：百合、玉竹、麦门冬、北沙参、天门冬、桑葚、白芍；温性药物有：人参、大枣、白术、山药、黄芪、党参、益智仁、肉苁蓉、补骨脂、菟丝子、沙苑子、断续、熟地黄、当归、龙眼肉、何首乌，将 23 味中药分别配制成 5 组不同浓度(10，8，6，4 和 2 mg·mL^{-1})的溶液制剂共 115 个样本作为研究对象，利用荧光光谱仪测得每味中药溶液制剂的荧光光谱数据，并获取每个样本的等高线图和三维荧光光谱图，其中部分样本的等高线图和三维荧光光谱图如图 1 和图 2 所示。

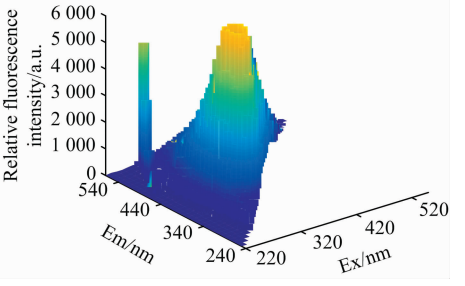

图 1 样本三维荧光光谱图

Fig. 1 Three-dimensional fluorescence spectra of samples

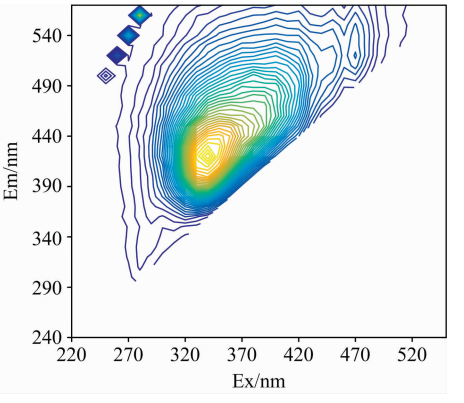

图 2 样本等高线图

Fig. 2 Contour map of samples

由于受仪器以及外界环境等因素的影响，使得采集到的中药三维荧光光谱数据中存在一定的噪声，由图 1 和图 2 可以看出荧光光谱信号在 EX/EM=340 nm/420 nm 处存在主荧光峰，在激发波长 240 nm<EX<280 nm、发射波长 480 nm<EM<560 nm 范围内存在噪声，在 EX/EM=280 nm/560 nm 附近噪声强度较高，对主荧光峰影响较大。因此需要

对数据进行预处理以尽量去除噪声。采用 EEMD 算法(ensemble empirical mode decomposition，EEMD)对每个样本进行降噪处理，该算法降噪后的数据格式为 34×34 矩阵的形式，将每个样本矩阵首尾相接，构成 1×1156 形式，即 115 个中药荧光光谱样本数据为 115×1156 矩阵形式。由于该矩阵维度较大，直接应用模式识别算法进行分类会增加模型的运行时间及计算量，影响分类识别效果，因此，需要采用特征提取算法对中药光谱样本数据进行降维以及特征提取。

本文采用近邻点数 $k=12$，本征维数 $d=7$ 时得到的特征向量进行研究，即将原始中药荧光光谱数据从 1 156 维降到 7 维。LLE 算法得到的部分样本的特征向量如表 1 所示，光谱特征的可视化结果如图 3 所示。图中依次为玉竹、北沙参、白术、龙眼肉的荧光光谱特征。

由图 3 可知，不同浓度的玉竹 PC4，PC6 和 PC7 的特征值变化明显，不同浓度的北沙参 PC4，PC5 和 PC6 的特征值变化明显，不同浓度的白术和龙眼肉 PC1，PC2，PC4 和 PC7 的特征值变化明显，且浓度越高特征值都有下降趋势。将得到的 PC1，PC3 和 PC6 特征向量进行三维聚类，结果如图 4 所示，PC1 代表的特征向量作为 x 轴，PC6 代表的特征向量作为 y 轴，PC3 代表的特征向量作为 z 轴。由图 4 可以看出，仅少数寒性样本与温性样本有重叠，从整体角度分析，寒性样本与温性样本可以较好的识别出来。

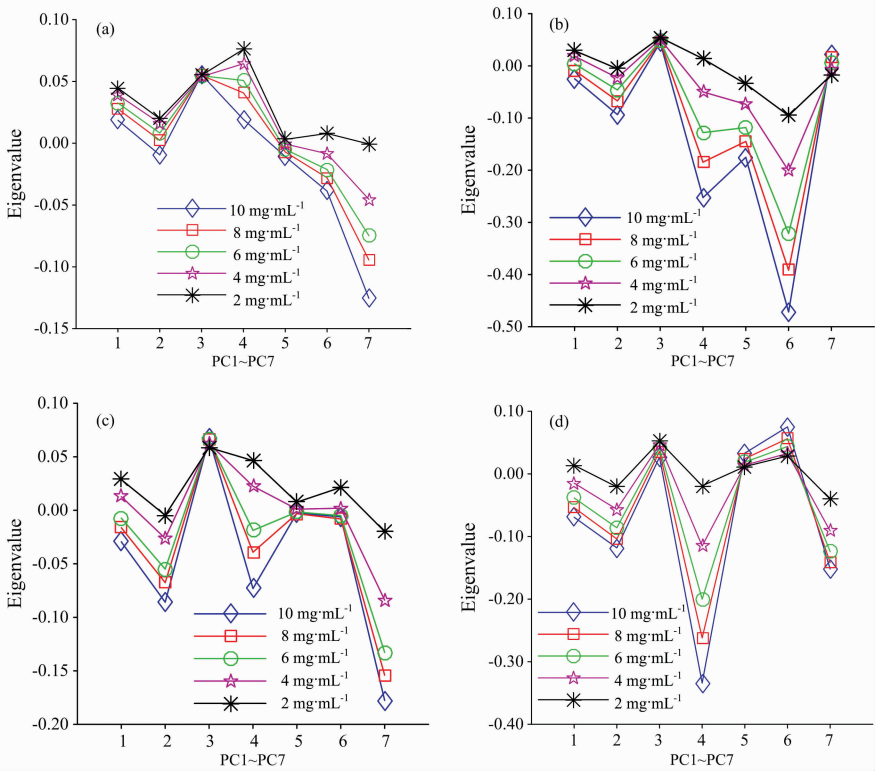

图 3　部分中药荧光光谱特征

(a)：玉竹荧光光谱特征；(b)：北沙参荧光光谱特征；(c)：白术荧光光谱特征；(d)：龙眼肉荧光光谱特征

Fig. 3　Fluorescence spectrum characteristics of some Traditional Chinese Medicine

(a)：Fluorescence spectrum characteristics of *yuzhu*；(b)：Fluorescence spectrum characteristics of *beishashen*；(c)：Fluorescence spectrum characteristics of *baizhu*；(d)：Fluorescence spectrum characteristics of *longyanrou*

表 1　LLE 算法得到的特征向量

Table 1　Features data extracted from LLE

特征量	PC1	PC2	PC3	PC4	PC5	PC6	PC7
样本 1	0.003 94	−0.044 97	0.071 69	0.030 91	0.000 01	−0.003 77	−0.107 14
样本 2	0.011 789	−0.032 95	0.069 22	0.043 25	0.001 78	−0.001 10	−0.087 03
样本 3	0.019 47	−0.021 47	0.066 79	0.054 56	0.004 40	0.008 86	−0.063 10
…	…	…	…	…	…	…	…
样本 115	0.041 83	0.014 66	0.056 34	0.072 790	0.007 90	0.018 91	0.009 48

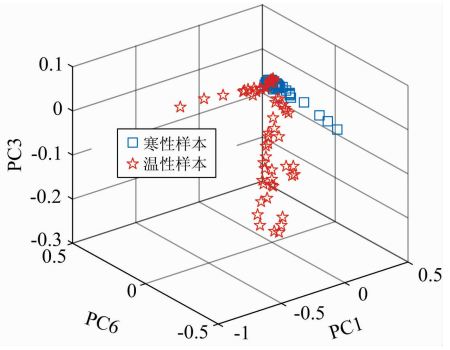

图 4 LLE 部分特征三维聚类

Fig. 4 Three dimensional clustering of partial features based on LLE algorithm

采用 RF 分类器对 LLE 算法提取的寒温类中药荧光光谱特征进行分类，将 LLE 算法得到的特征向量输入到 RF 中，构建 LLE-RF 分类模型，分析不同参数时 LLE-RF 分类模型对寒温类中药荧光光谱数据的分类效果，标记温性药物为第一类，寒性药物为第二类。为了建立较优的中药药性光谱数据分类模型，设置 RF 分类器中训练集和测试集的样本比例分别为 3∶1 和 2∶1，即训练集的比重 r 分别为 3/4 和 2/3。对于 LLE-RF 分类模型，固定分类器中训练集和测试集样本的比例，分析 LLE 中近邻点数 k 取值为 7～18，本征维数 d 分别取值为 6，7，8，9 和 10 时分类正确率变化情况。图 5 为当 r 不同时，LLE-RF 模型分类正确率随近邻点数 k 和本征维数 d 变化而波动情况。

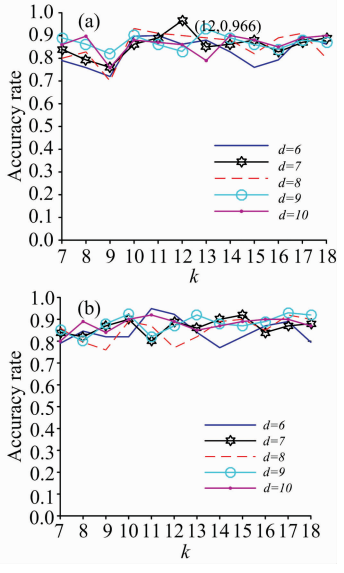

图 5 LLE-RF 不同比例下分类正确率变化情况

(a)：r=3/4；(b)：r=2/3

Fig. 5 The change of LLE-RF classification accuracy rate, when the ratio r equals 3/4 and 2/3

(a)：r=3/4；(b)：r=2/3

由图 5 可以看出，当训练集的比重 r 一定时，LLE-RF 模型分类正确率随近邻点和本征维数变化而波动。针对 LLE-RF 分类模型，当训练集和测试集的比例为 3∶1 和 2∶1 时，其分类正确率分别为 96.6%和 95%。其中，当训练集和测试集的比例为 3∶1，近邻点数 k=12，本征维数 d=7 时 LLE-RF 模型正确率最高，为 96.6%，其预测结果如图 6 所示。由图 6 可以看出，LLE-RF 分类模型对寒温类中药荧光光谱数据分类时有 1 个样本识别错误，实际类别为第一类(温性药物)但被错误识别为第二类(寒性药物)。

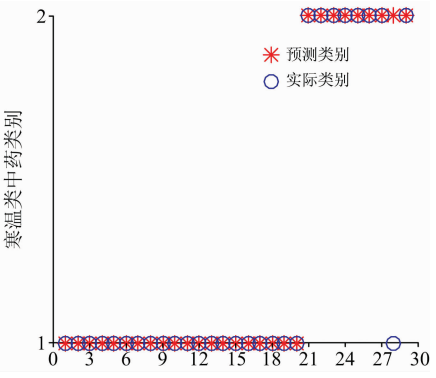

图 6 LLE-RF 模型预测结果

Fig. 6 Prediction results of LLE-RF

为验证 LLE-RF 分类模型的分类效果，当近邻点数 k=12，本征维数 d=7 时，分别采用多项式、径向基以及多层感知机核函数构造 LLE-SVM 分类模型，将该模型与 LLE-RF 分类模型的分类正确率进行比较，记训练集的比重 r 分别为 3/4 和 2/3，三种核函数均在默认参数下构造分类器，分类正确率如图 7 所示。

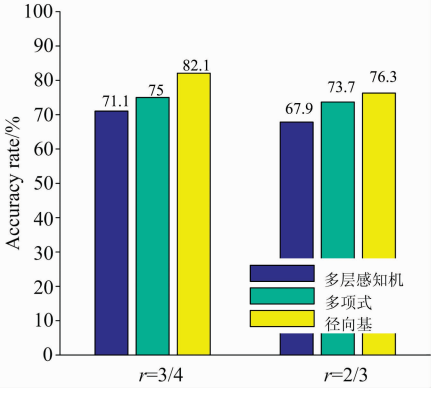

图 7 LLE-SVM 不同比例下分类正确率变化情况

Fig. 7 The change of LLE-SVM classification accuracy rate, when the ratio r equals 3/4 and 2/3

由图 7 可以看出，在同一比例 r 的情况下，采用不同核函数构造 SVM 分类器时，寒温类中药荧光光谱数据分类效果不同。当多层感知机作为核函数时，分类效果最差。针对 LLE-SVM 分类模型，当训练集和测试集的比例分别为 3∶1 和 2∶1 时，其分类正确率分别为 82.1%和 76.3%。因此，当采用 LLE 提取寒温类中药光谱特征，分类器选择 SVM 或 RF 时，设置训练集和测试集的比例为 3∶1 时建立的分类模型效果较好，且 LLE-RF 模型分类正确率高于 LLE-SVM 分

类模型。

3　结　论

三维荧光光谱技术应用到中药药性识别研究领域具有易检测、原材料成本低、省时等优点。本文基于中药药性的荧光光谱特征，将局部线性嵌入算法与随机森林算法相结合，构建 LLE-RF 寒温类中药荧光光谱分类模型，与 LLE-SVM 分类模型比较，LLE-RF 分类正确率高于 LLE- SVM 分类模型，具有较好的分类识别效果。该方法为中药鉴别、中药质量控制以及中药药性研究提供思路。在今后的研究中应借助更先进的仪器设备，从多学科、多角度对中医药理论进行研究，促进中医药现代化发展。

References

[1] ZHANG Xin-xin, LI Yu, JI Yu-jia, et al(张新新，李　雨，纪玉佳，等). Journal of Shandong University·Health Science(山东大学学报·医学版), 2012, 50(1): 143.

[2] LI Jia-hui, CHEN Ren-shou, LI Lu-jie(李加慧，陈仁寿，李陆杰). Journal of Traditional Chinese Medicine(中医杂志), 2019, 60(1): 67.

[3] LIU Min, WU Dong-xue, LI Jing, et al(刘　敏，吴东雪，李　晶，等). China Journal of Chinese Materia Medica(中国中药杂志), 2019, 44(2): 218.

[4] WANG Xiao-yan, LI Feng(王晓燕，李　峰). Liaoning Journal of Traditional Chinese Medicine(辽宁中医杂志), 2015, 42(6): 1303.

[5] WU Si-yuan, HU You-fen, LIU Xiao-wei, et al(吴思媛，胡幼芬，刘晓伟，等). Software Guide(软件导刊), 2014, 13(10): 71.

[6] CHEN Zhao, CAO Yan-feng, HE Shuai-bing, et al(陈　昭，曹燕凤，何帅兵，等). China Journal of Traditional Chinese Medicine and Pharmacy(中华中医药杂志), 2017, 32(5): 2107.

[7] Henrique Z P, Alonso J B, Ferrer M A, et al. IEEE Transactions on Audio Speech and Language Processing, 2009, 17(6): 1186.

[8] QIN Xi-wen, LÜ Si-qi, LI Qiao-ling(秦喜文，吕思奇，李巧玲). Chinese Journal of Biomedical Engineering(中国生物医学工程学报), 2018, 37(6): 665.

[9] Shinnosuke Tomiyama, Mamiko Sakata-yanagimoto, Shigeru Chiba, et al. Electronics and Communications in Japan, 2018, 101(11): 13.

[10] LIU Peng, WU Rui-mei, YANG Pu-xiang, et al(刘　鹏，吴瑞梅，杨普香，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(1): 193.

Pattern Recognition of Traditional Chinese Medicine Property Based on Three-Dimensional Fluorescence Spectrum Characteristics

FAN Feng-jie[1], XUAN Feng-lai[1], BAI Yang[1], JI Hui-fang[2]

1. Institute of Electrical Engineering, Yanshan University, Qinhuangdao　066004, China

2. No. 984 Hospital of the PLA, Beijing　100094, China

Abstract　As three-dimensional fluorescence spectroscopy has many advantages, such as good selectivity, high sensitivity, fast analysis, it has been widely used in many fields. As one of the characteristics of traditional Chinese medicine(TCM), Chinese herbal medicine property (CHMP) is the core of TCM. Objective discrimination of the properties of TCM is the key issues of modernization of TCM. The identification of traditional Chinese medicine property is of great significance in the theoretical study of Chinese medicine. Most of the molecules in traditional Chinese medicine have the ability to generate fluorescence. According to the characteristics of the three-dimensional fluorescence spectrum of traditional Chinese medicines, the classification and recognition were studied from the perspective of the properties of traditional Chinese medicines. Firstly, the three-dimensional fluorescence spectral data of 5 different concentrations of 23 cold and warm Chinese medicinal solutions were acquired by FS920 fluorescence spectrometer. Then, the ensemble empirical mode decomposition (EEMD) algorithm is applied to denoise the spectrogram, based on the analysis of noise in different excitation and emission wavelength ranges of different samples. Based on the local linear embedding (LLE) algorithm, feature extraction of spectral data is carried out. The extracted eigenvectors are input into the random forest (RF) to construct LLE-RF classification model. The classification effect of LLE-RF classification model on fluorescence spectrum data of cold and warm Chinese medicines was analyzed under different parameters. The sample ratio of the training set and test set in RF classifier is set to 3∶1 and 2∶1. The correct rate of LLE classification is analyzed when the nearest neighbor points k is 7～18 and the eigenvalue dimension d is 6, 7, 8, 9 and 10. When the nearest neighbor points k is 12 and

the eigenvalue dimension d is 7, the accuracy of LLE-RF model for classification of Chinese herbal medicines was 96.6%. Finally, the classification effect of SVM classifier constructed with different kernels on fluorescence spectrum data of cold and warm Chinese medicines was compared under the same ratio of r. When multi-layer perceptron is used as the kernel function, the classification effect is the worst. When $r=3/4$ and radial basis function is used as the kernel function, the classification accuracy is 82.1%. The results show that the method of combining fluorescence spectroscopy with LLE-RF can effectively recognize cold and warm Chinese medicines, and the classification effect is better than LLE-SVM.

Keywords Three-dimensional fluorescence spectrum; Feature extraction; Traditional Chinese medicine property; Local linear embedding; Random forest

(Received May 23, 2019; accepted Sep. 16, 2019)