茶叶傅里叶近红外光谱的混合模糊极大熵聚类分析

傅海军[1,2]，周树斌[1,3]，武小红[1,2*]，武 斌[4]，孙 俊[1,2]，戴春霞[1,5]

1. 江苏大学电气信息工程学院，江苏 镇江 212013
2. 江苏大学机械工业设施农业测控技术与装备重点实验室，江苏 镇江 212013
3. 江苏大学科技信息研究所，江苏 镇江 212013
4. 滁州职业技术学院信息工程系，安徽 滁州 239000
5. 江苏大学食品与生物工程学院，江苏 镇江 212013

摘 要 茶作为世界最受欢迎的三大饮料之一，不仅能够提神醒脑，而且还有帮助消化和降低血压等作用。随着人们对茶叶品质要求的日益提高，需要对不同品种的茶叶实现准确的鉴别分析以防止茶叶市场里茶叶品牌名不副实和以次充好等现象的发生。为实现对茶叶快速精准的鉴别分析，设计了一种综合采用傅里叶近红外光谱和新的模糊极大熵聚类(FEC)分析算法的茶叶品种鉴别系统。传统模糊极大熵聚类分析在聚类含噪声数据时，聚类结果往往容易出现错误，即FEC对噪声数据敏感。为解决这个问题，在FEC分析算法的基础上引入可能C均值聚类分析(PCM)，提出了一种混合模糊极大熵聚类(MFEC)分析算法。MFEC可通过迭代计算得到模糊隶属度值，能实现对含噪声的茶叶傅里叶近红外光谱数据的准确聚类分析。首先，使用傅里叶近红外光谱仪(Antaris Ⅱ型)采集岳西翠兰、六安瓜片、施集毛峰三种安徽茶叶的傅里叶近红外光谱数据，光谱波数范围为10 000～4 000 cm^{-1}。其次，对采集到的光谱数据使用多元散射校正(MSC)进行预处理，预处理后先用主成分分析(PCA)将光谱数据维数降至10维，然后再用线性判别分析(LDA)对降维后的近红外光谱数据进行特征提取。最后，通过混合模糊极大熵聚类分析和传统的模糊极大熵聚类分析对三种茶叶的光谱数据进行聚类分析，并对两种聚类分析算法得到的聚类准确率、收敛速度等进行对比分析。实验结果表明：混合模糊极大熵聚类(MFEC)分析算法与传统的模糊极大熵聚类(FEC)分析算法相比较，在相同的权重指数 m 下MFEC具有更高的聚类准确率。在 $m=2$ 条件下，MFEC的聚类准确率达到了100%，而传统的模糊极大熵聚类在相同条件下聚类准确率仅为37.98%。MFEC收敛过程中仅需迭代10次即可达到收敛，而FEC需要迭代100次，因此MFEC可以更高效的进行模糊聚类分析，MFEC相比于FEC聚类性能具有明显的优越性。通过傅里叶近红外光谱技术，混合模糊极大熵聚类分析结合PCA与LDA算法构建的茶叶品种鉴别系统能够高效快速的完成对岳西翠兰、六安瓜片、施集毛峰三种茶叶的准确分类，为茶叶检测领域提供了一种创新的方法与设计思路，具有一定的理论价值和良好的市场应用前景。

关键词 近红外光谱；茶叶；主成分分析；线性判别分析；模糊极大熵聚类分析

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)11-3465-05

引 言

中国人自古以来便喜欢品茶鉴茶，饮茶之风从古代风靡至今，茶成为了最受欢迎的饮用佳品之一。喜好饮茶的人都希望可以品尝到质量可以得到保障的茶叶，而市场上的茶叶却往往鱼龙混杂，良莠不齐，品质不能得到很好地保证，这在很大程度上影响了消费者的利益和优良茶叶品牌的建设和推广。所以，研究出一种高效快速又准确的茶叶品种鉴别方法符合社会和广大消费者的需求[1-2]。

茶叶的鉴别方法从传统的人工鉴别到化学鉴别方法一直在不断发展[3]，但在一定程度上都不能真正满足现代化社会

收稿日期：2018-10-07，**修订日期**：2019-02-16

基金项目：国家自然科学基金项目(31471413)，江苏高校优势学科建设工程项目PAPD，安徽省教育厅高校自然科学研究重点项目(KJ2019A1129)资助

作者简介：傅海军，1976年生，江苏大学电气信息工程学院副教授 e-mail：fuhaijun21@ujs.edu.cn

*通讯联系人 e-mail：wxh_www@163.com

的需求。前者主观因素较强且需要投入较大的人力和时间，不能满足日益扩大的茶叶市场需求；后者则工艺复杂且价格昂贵，不适合进行大规模的实际应用。近年来，傅里叶近红外光谱技术以其绿色高效准确的性能在茶叶鉴别研究中初步崭露头角，从定性到定量，诸多的研究成果论证了这一技术运用于茶叶领域的可行性[4]。例如：Zhuang 等应用近红外光谱可以有效无损地检测山东绿茶的起源地，采用 BP 神经网络，偏最小二乘法(PLS)和支持向量机(SVM)进行回归计算，结果表明运用 PLS 对训练样本和测试样本均可达到 100%的鉴定准确率[5]。Cai 等利用傅里叶红外光谱(FTIR)和模式识别对茶叶品种进行鉴别，采用偏最小二乘法(PLS)与自组织特征映射(SOM)神经网络方法相结合形成了一种相比于 PLS 线性方法更为准确的非线性分类算法，识别率达到 100%[6]。武小红等利用 FTIR 光谱结合 Gustafson-Kessel 聚类方法对茶叶品种进行鉴别分析，为茶叶品种分类提供了一种有效的判别模型[7]。Deng 等利用近红外高光谱成像完成了对浙江龙井茶叶含水量快速准确的无损检测[8]。Li 等提出间隔偏最小二乘法(IPLS)提取和优化全光谱数据的特征，研究了来自 14 种茶树的 160 个茶叶样本的茶多酚(TP)红外光谱快速测定，分别建立基于 PLS，IPLS 和后向间隔偏最小二乘法(BIPLS)的回归预测模型，证明了红外光谱法测定茶叶中 TP 含量的可行性[9]。Xiong 等利用近红外反射光谱和多光谱成像(MSI)系统对铁观音的总多酚含量(TPC)和贮藏期进行无损测定，实验结果表明采用偏最小二乘法的 MSI 系统是无损和快速检测茶叶 TPC 含量的最佳方法[10]，分别采用最小二乘支持向量机(LS-SVM)和 BP 神经网络分类茶叶贮藏期的准确率分别为 95.0%和 97.5%。Xu 等应用傅里叶变换近红外光谱，PLS 和单类偏最小二乘(OCPLS)对中国功能性茶(板蓝根)掺假进行快速鉴别，为板蓝根茶的快速质量控制提供一个有用的新方法[11]。

通过光谱仪采集到的茶叶近红外光谱是一种高维的数据[12]，其中包含了很多复杂的冗余信息，影响了计算结果的准确性，通过对光谱数据的降维处理可减少冗余信息。本工作采用主成分分析(PCA)进行降维处理[13]，然后通过线性判别分析(LDA)进行特征提取[14]。最后在传统模糊极大熵聚类(FEC)[15]算法的基础上引入了可能 C 均值聚类分析(PCM)[16]，在此基础上提出了一种混合模糊极大熵聚类(MFEC)算法，同时用 MFEC 进行聚类分析以实现对茶叶品种的最终鉴别分类。

首先用近红外光谱仪完成对岳西翠兰、六安瓜片、施集毛峰三种茶叶样本的傅里叶近红外光谱(FT-NIR)数据采集，然后经过多元散射校正(MSC)预处理，主成分分析和线性判别分析的数据压缩和特征提取，最后分别通过模糊极大熵聚类分析和混合模糊极大熵聚类分析完成对三种茶叶的分类。结果表明，本工作提出的 FT-NIR 结合 MFEC 算法可以很好的完成对三种安徽品牌茶叶的鉴别分析。

1 实验部分

1.1 茶叶近红外光谱的采集

实验用茶叶为岳西翠兰、六安瓜片、施集毛峰等三种安徽品牌茶叶，每种茶叶有 65 个样本，总的茶叶样本数为 195。样本经过研磨粉粹后过 40 目筛。实验室的温度和相对湿度保持相对不变，Antaris Ⅱ型 FT-NIR 光谱仪开机预热 1 h。采用反射积分球模式采集茶叶近红外光谱，每个茶叶样品扫描 32 次。光谱波数范围是 4 000～10 000 cm^{-1}，扫描的光谱波数间隔是 3.857 cm^{-1}，采集的茶叶光谱数据维数为 1 557 维。每个样本采样 3 次，3 次的平均值作为后续实验中样本的光谱数据。采集的 3 种安徽茶叶样本的 FT-NIR 图如图 1 所示。用 Matlab R2014b 编写程序，运行在 Windows 10 系统里，RAM 8GB。

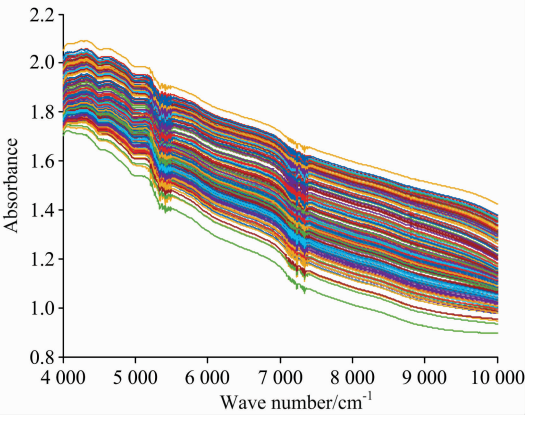

图 1 3 种茶叶样本的近红外光谱图

Fig. 1 FT-NIR spectra of three kinds of tea samples

1.2 混合模糊极大熵聚类分析算法

混合模糊极大熵聚类(MFEC)算法具体描述如下：

(1) 初始化过程：设置权重指数 $m(m>1)$，类别数 c；设置循环计数 r 的初始值和最大迭代次数为 $r_{\max}$；设置迭代最大误差参数 ε；参数 λ 和 β，以每类训练样本的均值作为初始的类中心值 $v_i^{(0)}$；计算测试样本的协方差 σ^2

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{n} \| x_k - \bar{x} \|^2, \ \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{1}$$

式(1)中，n 为测试样本数，$m(m>1)$为权重指数，x_k 为第 k 个茶叶测试样本。$\bar{x}$ 为测试样本的均值。

(2) 计算第 $r(r=1, 2, \cdots, r_{\max})$次迭代时的模糊隶属度值 $u_{ik}^{(r)}$

$$u_{ik}^{(r)} = \left[\sum_{j=1}^{c}\left(\frac{\exp(D_{ik}^2)}{\exp(D_{jk}^2)}\right)^{\frac{1}{\lambda}}\right]^{-\frac{1}{2}}\left[1+\left(\frac{\beta cm^2 D_{ik}^2}{\sigma^2}\right)^{\frac{1}{m-1}}\right]^{-\frac{1}{2}} \forall i, k \tag{2}$$

式(2)中，u_{ik} 是样本 x_k 隶属于类别 i 的模糊隶属度值，$u_{ik}^{(r)}$ 是第 r 次迭代计算的模糊隶属度值；$D_{ik} = \| x_k - v_i^{(r-1)} \|$；$v_i$ 是第 $i(i=1, 2, 3, \cdots, c)$类的类中心值，$v_i^{(r-1)}$ 是第 $r-1$ 次迭代计算的类中心 v_i 的值。

(3) 计算第 r 次迭代时的第 i 类的类中心值 $v_i^{(r)}$

$$v_i^{(r)} = \frac{\sum_{k=1}^{n}[u_{ik}^{(r)}]^m x_k}{\sum_{k=1}^{n}[u_{ik}^{(r)}]^m}, \ \forall i \tag{3}$$

式(3)中，$v_i^{(r)}$ 是第 r 次迭代计算的类中心 v_i 的值，由 c 个类中心值组成类中心矩阵 $V^{(r)}=[v_1^{(r)}, v_2^{(r)}, \cdots, v_c^{(r)}]$。

(4) 循环计数增加，即 $r=r+1$；

若满足条件：($\| V^{(r)}-V^{(r-1)} \| <\varepsilon$)或($r>r_{max}$)则计算终止，否则继续步骤(2)。

2　结果与讨论

2.1　茶叶近红外光谱的预处理

用近红外光谱仪采集到的光谱数据中，除了包含对数据分析有价值的茶叶化学成分的光谱吸收数据信息外，还掺杂着影响数据分析准确率的光散射信息。光散射受多种物理因素(如粒径，形状和分布)的影响，并且在不同样品的光散射信息中可能存在差异。鉴于光散射信息带来的种种不利影响，需要对采集到的原始近红外光谱数据进行预处理。多元散射校正(MSC)是有效的近红外光谱预处理方法，所以本文采用 MSC 方法预处理茶叶近红外光谱的原始数据。该方法需要重新构建待测样品原始数据的理想光谱。这需要近红外光谱变化和样本组成含量符合直接线性的关系。而实际操作中往往难以获得真正的理想光谱，因此根据具体情况，取全部光谱的平均光谱来作为理想光谱是比较合理的。作为多变量散射校正的方法，MSC 方法可使光谱数据去除或减弱光散射所带来的影响，从而使有用光谱信息得到增强。对图 1 的茶叶近红外光谱进行 MSC 处理后的光谱如图 2 所示。

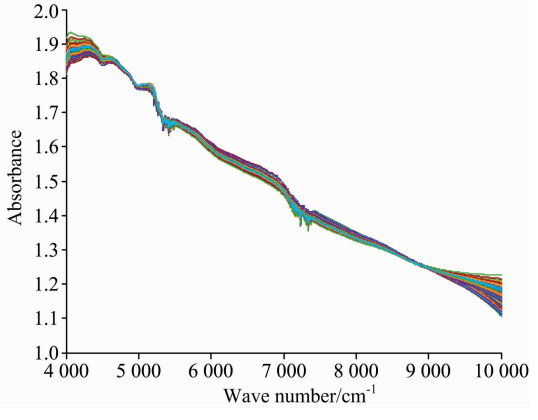

图 2　MSC 预处理后的茶叶近红外光谱图

Fig. 2　FT-NIR spectra pretreated with MSC

2.2　茶叶近红外光谱的降维处理和特征提取

经过 MSC 预处理后的茶叶光谱数据仍然是 1 557 维的高维数据，其中包含了大量的冗余信息，因此需要对光谱数据进行降维处理。这里选取经典降维算法—主成分分析(PCA)算法来实现降维处理。经 PCA 降维处理后前 2 个特征向量的 PCA 得分图如图 3 所示。其中符号“·”，“*”和“。”分别代表了岳西翠兰、六安瓜片和施集毛峰三种茶叶。观察 PCA 得分图可知，岳西翠兰和六安瓜片茶叶样本数据有少部分存在重叠，重叠部分数据在分类时容易出错，而施集毛峰和其余两种茶叶样本数据没有重叠，分类效果好。经 PCA 将光谱数据降至 10 维后，再用线性判别分析(LDA)方法对降维后的数据进行特征提取以提取出有价值的鉴别信息。从每类茶叶样本中选取 22 个样本作为训练样本，即训练集样本数为 66 个，剩下每类 43 个茶叶样本作为测试样本，即测试集样本数为 129 个。

2.3　模糊聚类分析

2.3.1　模糊聚类分析初始参数的设置

FEC 与 MFEC 的初始参数设置为：权重指数 $m=2$，品种数 $c=3$，参数 $\lambda=10$，$\beta=10$，初始迭代次数 $r=1$，最大迭代次数 $r_{max}=100$，迭代最大误差参数为 $\varepsilon=0.000\ 01$，测试样本数 $n=129$，经过 LDA 后得到的训练样本的均值即为初始聚类中心，则得到的初始聚类中心如式(4)所示

$$\begin{bmatrix} v_1^{(0)} \\ v_2^{(0)} \\ v_3^{(0)} \end{bmatrix} = \begin{bmatrix} -0.015\ 4 & 0.001\ 1 \\ 0.027\ 1 & 0.002\ 3 \\ -0.010\ 9 & -0.010\ 0 \end{bmatrix} \tag{4}$$

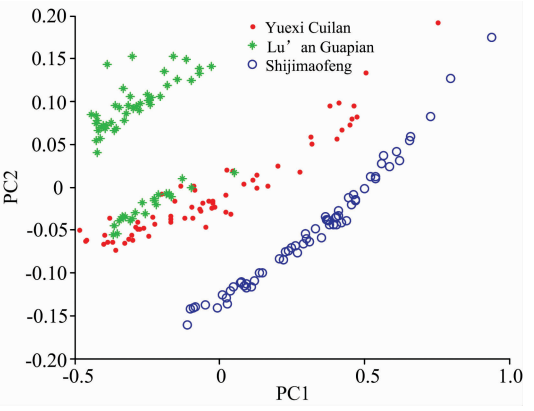

图 3　PCA 得分图

Fig. 3　Scores plot of PCA

2.3.2　模糊聚类准确率

通过运行 FEC 和 MFEC 两种聚类算法，改变 MFEC 算法权重指数 m 的值后观察两种聚类分析算法的聚类准确率。在 m 分别取值 1.2，1.4，…，3.0 下观察聚类准确率的变化如图 4 所示。从图 4 中不难看出，MFEC 的准确率明显高于传统的 FEC 的准确率。在 m 分别取值 1.2，1.4，…，2.6 的情况下，MFEC 的聚类准确率达到 100%，并且在 $m=2.8$，3.0 的情况下聚类准确率达到 96.9%。传统的 FEC 没有参数 m，其聚类准确率仅为 37.98%。当 $m=2$ 时，MFEC 分析算法在经过 10 次迭代计算后收敛，而传统的 FEC 需经过 100 次迭代计算才可达到收敛，所以，MFEC 在聚类收敛上要优于传统的 FEC。

2.3.3　茶叶种类判别

以经过 LDA 处理后的数据样本作为本节所使用的训练样本和测试样本。计算岳西翠兰、六安瓜片以及施集毛峰三种茶叶训练样本的平均值：岳西翠兰平均值 $\bar{x}_1=[-0.015\ 4 \quad 0.011\ 1]$；六安瓜片平均值 $\bar{x}_2=[0.027\ 1 \quad 0.002\ 3]$；施集毛峰平均值 $\bar{x}_3=[-0.010\ 9 \quad 0.010\ 0]$。以训练样本的均值作为初始聚类中心，运行 MFEC 分析至迭代终止后得到的类中心如式(5)

$$\begin{bmatrix} v_1^{(10)} \\ v_2^{(10)} \\ v_3^{(10)} \end{bmatrix} = \begin{bmatrix} -0.015\ 6 & 0.010\ 6 \\ 0.024\ 7 & 0.000\ 0 \\ -0.010\ 5 & -0.010\ 1 \end{bmatrix} \quad (5)$$

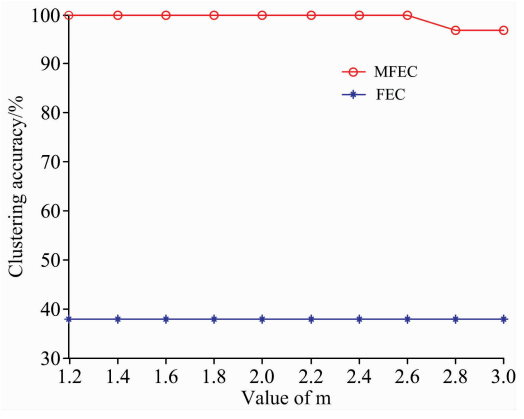

图 4 FEC 和 MFEC 的聚类准确率

Fig. 4 The clustering accuracies of FEC and MFEC

测试样本经过 MFEC 计算后得到三个聚类中心 $v_1^{(10)}$，$v_2^{(10)}$ 和 $v_3^{(10)}$，先判断三个聚类中心属于哪个品种茶叶，判断方法是分别计算某个聚类中心到三个初始聚类中心[见式(4)]的欧式距离，该聚类中心所属的茶叶品种和距离最小的初始聚类中心的茶叶品种相同。

在聚类分析算法达到收敛后，对于得到的样本的模糊隶属度值进行分析。对于 MFEC 的模糊隶属度值而言，通过分析判断某样本模糊隶属度值在三种类别下的情况可判断该样本隶属于哪个品种茶叶。当某样本在某一类模糊隶属度值高于另外两类的模糊隶属度值时则判定该样本属于这一类品种茶叶。对于测试样本 x_k 的模糊隶属度值 u_{ik}，若判定其类别属于第 i 类，则 u_{ik} 的值要大于其他类别的模糊隶属度值。MFEC 迭代收敛后的模糊隶属度图如图 5 所示。

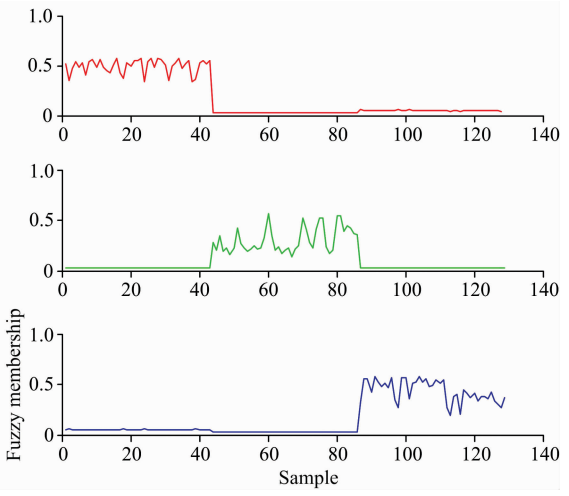

图 5 模糊隶属度值

Fig. 5 Fuzzy membership of MFEC

3 结 论

为解决 FEC 对噪声数据敏感问题，在 FEC 基础上，结合可能 C 均值聚类分析(PCM)，提出了一种混合模糊极大熵聚类(MFEC)分析。将 MFEC 和 FEC 运用于茶叶傅里叶近红外光谱的模糊聚类分析，聚类结果表明，MFEC 算法相比于传统的 FEC 算法，具有更快的收敛速度，更高的聚类准确率。通过使用茶叶的傅里叶近红外光谱数据，结合主成分分析，线性判别分析和 MFEC 算法可对三种安徽品牌茶叶实现快速、准确的分类，MFEC 具有明显更高的聚类准确率。

References

[1] Wu X H, Wu B, Sun J, et al. International Journal of Food Properties, 2016, 19: 1016.
[2] WU Xiao-hong, ZHAI Yan-li, WU Bin, et al(武小红, 翟艳丽, 武 斌, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(6): 1719.
[3] Barbin D F, Sun D W, Nixdorf S L, et al. Food Research International, 2014, 61(7): 23.
[4] Cebi N, Yilmaz M T, Sagdic O. Food Chemistry, 2017, 229: 517.
[5] Zhuang X G, Wang L L, Chen Q, et al. Science China Technological Sciences, 2017, 60(1): 84.
[6] Cai J X, Wang Y F, Xi X G, et al. International Journal of Biological Macromolecules, 2015, 78: 439.
[7] Wu X H, Zhu J, Wu B, et al. Computers & Electronics in Agriculture, 2018, 147: 64.
[8] Deng S, Xu Y, Li X, et al. Computers & Electronics in Agriculture, 2015, 118(C): 38.
[9] Li X, Sun C, Luo L, et al. Computers & Electronics in Agriculture, 2015, 112: 28.
[10] Xiong C, Liu C, Pan W, et al. Food Chemistry, 2015, 176: 130.
[11] Xu L, Fu X, Fu H, et al. Journal of Food Quality, 2016, 38(6): 450.
[12] Li C, Guo H, Zong B, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2019, 206: 254.
[13] Jaiswal P, Jha S N, Kaur J, et al. Food Chemistry, 2018, 238: 209.
[14] Salman A, Shufan E, Sahu R K, et al. Vibrational Spectroscopy, 2016, 83: 17.
[15] Li R P, Mukaidono M. Fuzzy Sets and Systems, 1999, 102(2): 253.
[16] Krishnapuram R, Keller J. IEEE Transactions on Fuzzy Systems, 1993, 1(2): 98.

Mixed Fuzzy Maximum Entropy Clustering Analysis of FT-NIR Spectra of Tea

FU Hai-jun[1,2], ZHOU Shu-bin[1,3], WU Xiao-hong[1,2*], WU Bin[4], SUN Jun[1,2], DAI Chun-xia[1,5]

1. School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China
2. Key Laboratory of Facility Agriculture Measurement and Control Technology and Equipment of Machinery Industry, Jiangsu University, Zhenjiang 212013, China
3. Institute of Scientific and Technical Information, Jiangsu University, Zhenjiang 212013, China
4. Department of Information Engineering, Chuzhou Vocational Technology College, Chuzhou 239000, China
5. School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China

Abstract Tea is one of the three most popular drinks in the world. It can not only refresh the mind, but also help digestion and lower blood pressure. With the increasing advance of requirements of tea quality by people, it is necessary to achieve accurate identification of different varieties of tea to prevent the false tea brands and adulteration in the tea market from happening. In order to identify tea varieties quickly and accurately, a tea variety identification system was designed with a combination of Fourier transform near-infrared spectroscopy (FT-NIR) and a novel fuzzy maximum entropy clustering. When traditional fuzzy maximum entropy clustering (FEC) clusters the data with noise, clustering results are often prone to errors, that is to say, FEC is sensitive to noise. To solve this problem, a mixed fuzzy maximum entropy clustering (MFEC) was proposed by introducing possibilistic c-means (PCM) clustering into traditional FEC. MFEC has fuzzy membership and typicality values by iterative computing, and it can cluster FT-NIR data mixed with noise accurately. Firstly, three kinds of Anhui tea samples (i. e. Yuexi Cuilan, Lu'an Guapian and Shiji Maofeng) were prepared for FT-NIR data collection with Antaris Ⅱ spectrometer in the wave number range of 10 000～4 000 cm^{-1}. Secondly, spectral data were preprocessed by multiple scattering correction (MSC), and then the dimensionality of the data was reduced to 10 by principal component analysis (PCA), and then the discriminant information of the data was extracted by linear discriminant analysis (LDA). Finally, MFEC and FEC were applied to perform clustering analysis on the data, respectively, and they were compared in the clustering accuracy and convergence speed. The results of this study indicated that in the condition of $m=2$, the clustering accuracy rate of MFEC was 100%, while that of FEC was 37.98%. MFEC achieved convergence after four iterations while FEC converged after 100 iterations. Therefore, MFEC could cluster spectral data more efficiently than FEC, and MFEC had the obvious superiority. Three types of Anhui tea samples could be classified correctly and efficiently by combining FT-NIR technology with PCA, LDA and MFEC. This method provided an innovative method and design idea for the identification analysis in the tea testing field, and it has certain theoretical value and good market application prospect.

Keywords Near-infrared spectroscopy; Tea; Principal component analysis; Linear discriminant analysis; Fuzzy maximum entropy clustering

(Received Oct. 7, 2018; accepted Feb. 16, 2019)

* Corresponding author