可能模糊鉴别C均值聚类的茶叶FTNIR分类研究

武 斌[1*]，傅海军[2]，武小红[2,3*]，陈 勇[2]，贾红雯[1]

1. 滁州职业技术学院信息工程系，安徽 滁州 239000
2. 江苏大学电气信息工程学院，江苏 镇江 212013
3. 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

摘 要 茶叶傅里叶近红外光谱(FTNIR)中含有茶叶的有机物化学成分信息，不同品种茶叶的化学成分和含量都有差异，所以利用傅里叶近红外光谱进行茶叶品种分类是可行的。由于茶叶近红外光谱数据具有维数高，有波峰和波谷，光谱重叠交错等特点，所以准确分类光谱数据存在困难。为此，提出一种可能模糊鉴别C均值聚类(PFDCM)算法，将模糊线性判别分析(FLDA)引入到可能模糊C均值聚类(PFCM)算法中，在模糊聚类过程中FLDA可提取茶叶近红外光谱的鉴别信息和进行数据空间的转换。PFDCM在对茶叶光谱进行模糊聚类后得到的模糊隶属度和典型值可实现茶叶近红外光谱的准确聚类，具有聚类速度快，准确率高等优点。由于PFDCM的典型值没有隶属度之和为1的约束条件，因而PFDCM在聚类含噪声的光谱数据方面优于模糊C均值聚类(FCM)。采集岳西翠兰，六安瓜片，施集毛峰和黄山毛峰四种茶叶共260个样本，采用Antaris Ⅱ型傅里叶近红外光谱仪采集茶叶的傅里叶近红外光谱。光谱波数范围为10 000～4 000 cm^{-1}，实验所得近红外光谱为1 557维的高维数据。首先，将光谱数据用多元散射校正(MSC)进行预处理以减少光谱散射和噪声影响和增加信噪比；其次，用主成分分析法(PCA)降低光谱数据空间的维数，经过PCA处理后光谱数据维数为7；然后，用线性判别分析(LDA)提取光谱数据中的鉴别信息并将光谱数据空间的维数进一步降低到3维；最后，分别用FCM，可能模糊C均值聚类(PFCM)和PFDCM进行数据的聚类分析，实现茶叶品种的准确分类。实验结果：权重指数 $m=2.0$，$\eta=2.0$，FCM，PFCM和PFDCM聚类算法的聚类准确率分别为93.60%，93.02%和98.84%；FCM收敛时共迭代25次，而PFCM和PFDCM收敛时分别迭代8次和23次；模糊聚类收敛所消耗的时间，FCM最少，而PFDCM最多。FTNIR技术结合MSC，PCA，LDA和PFDCM提供了一种实现茶叶品种准确鉴别的分类模型。

关键词 茶叶；近红外光谱；主成分分析；鉴别信息提取；模糊聚类

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)02-0512-05

引 言

作为茶的故乡，中国具有历史悠久的茶文化，中国人饮茶可以追溯到神农时代。茶是一种人们喜爱的绿色健康饮品。茶叶中富含有利于人体健康的多种氨基酸，矿物质和维生素等。国内茶叶品种繁多，有普通品种和名优品种。按照茶叶的发酵程度可分为绿茶，红茶和黑茶等；绿茶主要有：杭州西湖龙井，安徽黄山毛峰，安徽六安瓜片、河南信阳毛尖等。不同品种的茶叶，其生长环境不同，组分及含量存在差异，导致其功效也不尽相同。因此，利用当前的先进科学技术和仪器设计出一种简单易行且识别率高的茶叶品种分类模型具有重要的研究价值。

国内外研究人员近年来运用近红外光谱和中红外光谱技术等进行茶叶的定性和定量分析，取得了一定的研究成果[1-2]。例如：武小红等用FTIR-7600傅里叶红外光谱仪检测四川三种茶叶，提出一种联合Gustafson-Kessel (AGK)聚类算法，并用AGK进行茶叶聚类，得到聚类准确率为93.9%[3]。由于发生抹茶中渗入绿茶粉事件，Yu等研究了一种快速无损识别抹茶中绿茶粉的有效方法，检测样本在

收稿日期：2019-01-08，**修订日期**：2019-05-23

基金项目：国家自然科学基金项目(31471413)，安徽省高校自然科学研究重点项目(KJ2019A1129)，安徽省质量工程项目(2016ckjh137)，滁州职业技术学院科研规划项目(YJZ-2018-19)，安徽高校自然科学研究重大项目(KJ2018ZD064)资助

作者简介：武 斌，1978年生，滁州职业技术学院副教授 e-mail：wubind2003@163.com

*通讯联系人 e-mail：wubind2003@163.com；wxh_www@163.com

500～700 nm 之间的光谱，数据建模采用主成分分析(PCA)，线性判别分析(LDA)和簇类软独立模型法(SIMCA)[4]。Bartoszek 等用电子顺磁共振(EPR)光谱和半经验数学模型探索绿茶，黑茶和红茶的抗氧化性，结果表明在绿茶中显示出儿茶素的存在而在红茶中缺乏儿茶素的指标，那么儿茶素的存在是绿茶具有抗氧化性能的主要原因，芳香质子含量和总抗氧化力值(TEAC)相关[5]。Meng 等采集福建省三个地区的乌龙茶共 90 个样本，用质子磁共振和近红外光谱检测茶叶，数据采用偏最小二乘判别分析(PLSDA)建立分类模型[6]。Wang 等采集湖北恩施 108 个新鲜茶叶样本，用 Thermo Antaris Ⅱ傅里叶变换近红外光谱仪采集样本的近红外光谱，用前馈人工神经网络和后向间隔偏最小二乘算法建立茶叶购买价格的预测模型[7]。Hu 等收集四种绿茶，共 150 个样本，用光谱仪获取样本的三维激发-发射矩阵荧光光谱，发射光谱范围 330～680 nm 激发光谱范围 300～500 nm，用多线性主成分分析(MPCA)，自权重可变三线性分解(SWATLD)和多线性偏最小二乘判别分析(NPLSDA)处理光谱数据[8]。

模糊聚类是一种模式识别算法，广泛应用于数字图像处理，计算机视觉和模式分类中[9-10]。另一方面，模糊聚类可以用来聚类光谱数据。例如：模糊 C 均值聚类(FCM)[11]，可能 C 均值聚类(PCM)[12]，模糊鉴别 C 均值聚类(FDCM)[13]和 Gustafson-Kessel (GK)聚类可以聚类苹果近红外光谱数据。聚类含噪声的数据是一个重要研究热点[14]，由于 FCM 处理噪声数据存在敏感性，而 PCM 虽然能处理噪声数据但是存在一致性聚类问题，为解决这些问题，Pal 等提出一种可能模糊 C 均值[15]。但是该算法在计算参数 η_i 时需要先运行 FCM，为解决这个问题，武小红等设计了一种新的可能模糊 C 均值(PFCM)[16]。为了进一步提高 PFCM 的聚类准确率，本文结合新的 PFCM 和模糊线性判别分析(FLDA)，提出一种可能模糊鉴别 C 均值聚类(PFDCM)算法，并用该算法进行茶叶近红外光谱数据的聚类分析。

1　实验部分

1.1　茶叶 FTNIR 的采集

从安徽滁州当地超市采购安徽四种品牌茶叶：岳西翠兰、六安瓜片、施集毛峰、黄山毛峰。每种茶叶有 65 个样本，总共样本数为 260。四种茶叶经 DXF－04D 小型高速中药打粉机研磨粉碎，然后使用 40 目筛过滤。实验室温度和相对湿度保持相对不变，将模式设置为反射积分球，Antaris Ⅱ光谱仪扫描每个样本 32 次并计算光谱的均值；光谱的波数变化范围是 4 000～10 000 cm^{-1}，最后得到的茶叶样品光谱数据是 1 557 维；每个样本采样 3 次，取其平均值作为后续模型建立的实验数据。四种茶叶样本的傅里叶变换近红外光谱如图 1 所示。采用 Matlab R2014b 进行光谱数据绘图和模糊聚类算法的编程。

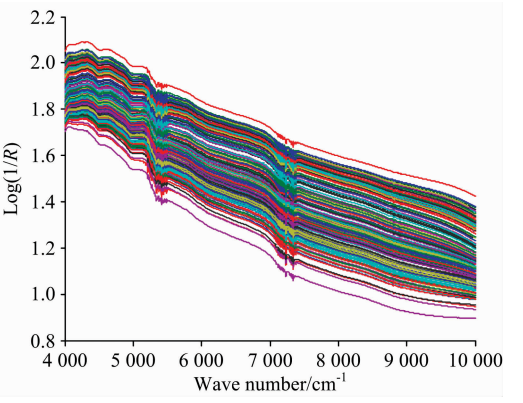

图 1　茶叶的傅里叶变换近红外光谱图

Fig. 1　FTNIR spectra of tea samples

1.2　可能模糊鉴别 C 均值聚类算法

可能模糊鉴别 C 均值聚类算法是一种迭代计算，其主要步骤叙述如下：

(1)初始化：设置权重指数 m 和 $\eta(m>1,\ \eta>1)$，样本数为 n，类别数为 $c(n>c>1)$，参数 a 和 b 的值($a>0,\ b>0$)；设置迭代次数初始值 r 和最大迭代次数 $r_{\max}$；误差参数为 ε；

(2)计算样本的协方差

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

式中，$\bar{x}$ 为样本的均值，x_j 为第 j 个样本。

(3)构造模糊类间散布矩阵 $\boldsymbol{S}_{fB}$

$$\boldsymbol{S}_{fB}=\sum_{i=1}^{c}\sum_{k=1}^{n}[au_{ik}^{(r)}+bt_{ik}^{(r)}]^m(v_i^{(r)}-\bar{x})(v_i^{(r)}-\bar{x})^{\mathrm{T}} \tag{2}$$

其中，$u_{ik}^{(r)}$，$t_{ik}^{(r)}$ 和 $v_i^{(r)}$ 中的 r 表示第 r 次迭代；$u_{ik}^{(r)}$ 为样本 x_k 属于第 i 类的隶属度；$v_i^{(r)}$ 为第 i 类的类中心矢量；$t_{ik}^{(r)}$ 是第 k 个样本 x_k 隶属于第 i 类的典型值。

(4)构造模糊总体散布矩阵 $\boldsymbol{S}_{fT}$

$$\boldsymbol{S}_{fT}=\sum_{i=1}^{c}\sum_{k=1}^{n}[au_{ik}^{(r)}+bt_{ik}^{(r)}]^m(x_k-\bar{x})(x_k-\bar{x})^{\mathrm{T}} \tag{3}$$

其中，x_k 是第 k 个样本。

(5)特征值和特征向量的计算

$$\boldsymbol{S}_{fT}^{-1}\boldsymbol{S}_{fB}\psi=\lambda\psi \tag{4}$$

式(4)中特征值 λ 和其对应的特征向量 Ψ。

(6)样本 $x_k\in R^q$ 投影到特征空间(由 $\Psi_1,\Psi_2,\cdots,\Psi_p$ 组成)

$$y_k=x_k^{\mathrm{T}}[\Psi_1,\Psi_2,\cdots,\Psi_p]\quad(y_k\in R^p) \tag{5}$$

式(5)中，p 和 q 表示维数，Ψ_p 是特征向量组中的第 p 个向量。

(7)同样将 $v_i^{(r)}$ 转化到特征空间(由 $\Psi_1,\Psi_2,\cdots,\Psi_p$ 组成)

$$v_i'^{(r)}=v_i^{(r)}[\Psi_1,\Psi_2,\cdots,\Psi_p] \tag{6}$$

其中，$v_i^{(r)}$ 为第 r 次迭代时第 i 类的类中心值，Ψ_p 为第 p 个特征向量。

(8)在特征空间中计算模糊隶属度函数值

$$u'^{(r+1)}_{ik}=\left[\sum_{j=1}^{c}\left(\frac{\|y_k-v_i'^{(r)}\|}{\|y_k-v_j'^{(r)}\|}\right)^{\frac{2}{m-1}}\right]^{-1},\ \forall i,k \tag{7}$$

式(7)中，$u'^{(r+1)}_{ik}$ 是第 $r+1$ 次迭代运算时样本 y_k 属于第 i 类的隶属度。

(9)在特征空间中计算典型值

$$t'^{(r+1)}_{ik}=\left[1+\left(\frac{bm^2c}{\sigma^2}\parallel x_k-v'^{(r)}_i\parallel^2\right)^{\frac{1}{\eta-1}}\right]^{-1},\ \forall i,k \quad (8)$$

式(8)中，$t'^{(r+1)}_{ik}$ 是第 $r+1$ 次迭代运算时样本 y_k 属于第 i 类的典型值。

(10)在特征空间中计算第 i 类的类中心矢量 $v'^{(r+1)}_i$

$$v'^{(r+1)}_i=\frac{\sum_{k=1}^{n}[a(u'^{(r+1)}_{ik})^m+b(t'^{(r+1)}_{ik})^\eta]y_k}{\sum_{k=1}^{n}[a(u'^{(r+1)}_{ik})^m+b(t'^{(r+1)}_{ik})^\eta]},\ \forall i,k \quad (9)$$

(11)迭代次数 r 值增加，即 $r=r+1$；，直到满足条件：$\parallel v'^{(r+1)}_i-v'^{(r)}_i\parallel<\varepsilon$ 或者 $r>r_{max}$ 则模糊聚类运算停止，否则将 $u'^{(r+1)}_{ik}$ 的值赋给变量 $u^{(r)}_{ik}$，$t'^{(r+1)}_{ik}$ 的值赋给变量 $t^{(r)}_{ik}$，$v'^{(r+1)}_i$ 的值赋给变量 $v^{(r)}_i$，继续从(3)开始重新计算。

(12) 迭代终止后，根据模糊隶属度值对样本进行分类，若 $u'^{(r+1)}_{ik}>0.5$，则判定第 k 个样本 x_k 隶属于第 i 类；反之，若 $u'^{(r+1)}_{ik}<0.5$，则判定第 k 个样本 x_k 不隶属于第 i 类。

2 结果与讨论

2.1 近红外光谱数据的 MSC，PCA 和 LDA 处理

茶叶近红外光谱和茶叶有机物含氢基团的振动信息相关，为检测茶叶的某种有机物含量提供了依据。但是，由于茶叶的样品颗粒度、装填密度等因素造成的散射问题使茶叶近红外光谱和有机物含量的关联性受到影响。为此，需要采用多元散射校正(MSC)进行光谱预处理以减少散射影响和提高信噪比[17-18]。图 1 所示的茶叶傅里叶变换近红外光谱经过多元散射校正处理后，其光谱数据维数仍为 1557 维，需要降维处理以便减少计算量和提高分类准确率。这里用主成分分析(PCA)对光谱数据进行压缩处理。PCA 的前 7 个主成分的累积贡献率达到 99.95%，所以选取 PCA 的前 7 个主成分进行光谱数据压缩后损失的信息少且可以降低数据维数。将 1 557 维的茶叶傅里叶变换近红外光谱投影到 PCA 的 7 个特征向量上可得到 7 维的光谱数据。选取 88 个茶叶样本(每个品种选取 22 个样本)作为训练集，其余的样本组成测试集，即测试样本共 172 个。利用线性判别分析(LDA)计算训练集的鉴别向量和特征值，计算得到的三个特征值为：$\lambda_1=232.29$，$\lambda_2=16.13$，$\lambda_3=2.60$。将茶叶测试样本经过 LDA 特征空间转换后形成三维数据，图 2 是线性判别分析的得分图。如图 2 所示，符号“＊HS”，“○LA”，“□SG”和“◇YX”分别表示黄山毛峰，六安瓜片，施集毛峰和岳西翠兰四种安徽茶叶光谱的测试样本。根据图 2 的数据分布可知，六安瓜片的数据分布比较松散，其余三种茶叶数据分布较为紧密，四种数据的每类之间区分较为明显。

2.2 FCM，PFCM 和 PFDCM 的近红外光谱聚类

2.2.1 模糊聚类的初始化参数设置

在运行 FCM，PFCM 和 PFDCM 聚类算法之前需要设置它们的初始化参数：设置权重指数 $m=2.0$，$\eta=2.0$，参数 a 和 b 的值均为 1，待聚类的样本数为 $n=172$，类别数为 $c=4$；设置迭代次数初始值 $r=1$ 和最大迭代次数 $r_{max}=100$；迭代最大误差参数 $\varepsilon=0.000\ 01$。设置 FCM 的初始聚类中心为

$$V^{(0)}_{FCM}=\begin{bmatrix}v^{(0)}_{1,FCM}\\v^{(0)}_{2,FCM}\\v^{(0)}_{3,FCM}\\v^{(0)}_{4,FCM}\end{bmatrix}=\begin{bmatrix}-0.014\ 6 & 0.007\ 9 & 0.035\ 1\\-0.010\ 7 & -0.002\ 0 & 0.007\ 2\\-0.011\ 7 & 0.007\ 5 & 0.038\ 2\\-0.008\ 2 & 0.014\ 1 & 0.015\ 5\end{bmatrix} \quad (10)$$

运行 FCM，经过 25 轮迭代计算终止后得到的聚类中心作为 PFCM 和 PFDCM 的初始聚类中心。

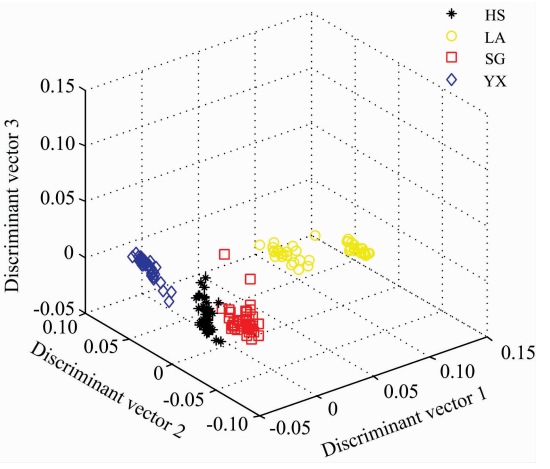

图 2 LDA 的得分图

Fig. 2 Scores plot of linear discriminant analysis

2.2.2 模糊聚类迭代次数和聚类时间

初始化参数设置同 3.2.1 节。运行 FCM，PFCM 和 PFDCM 聚类算法以分析迭代次数及收敛情况。三种模糊聚类算法达到收敛时迭代次数为：FCM 25 次，PFCM 8 次和 PFDCM 23 次。所以，PFCM 和 PFDCM 聚类算法的迭代次数均比 FCM 少。计算机配置：CPU Intel Core i5-2005U 2.20 GHz，RAM 8GB，Windows 10。运行 Matlab R2014b，FCM，PFCM 和 PFDCM 的聚类时间分别为 0.218 7，0.359 4 和 0.843 7 s。由于 PFCM 和 PFDCM 运行前需要运行 FCM 以得到初始聚类中心，因而聚类时间要多于 FCM，且 PFDCM 聚类时间最多。

2.2.3 模糊隶属度

设置 FCM，PFCM 和 PFDCM 聚类算法的初始化参数同 2.2.1 节。分别运行 FCM，PFCM 和 PFDCM 聚类算法至迭代收敛后，可分别得到 FCM，PFCM 和 PFDCM 聚类算法的模糊隶属度值 $u_{ik,FCM}$，$u_{ik,PFCM}$ 和 $u_{ik,PFDCM}$。$u_{ik,PFDCM}$ 如图 3 所示。若第 k 个测试样本 x_k 的 FCM 模糊隶属度为 $u_{ik,FCM}$，则 x_k 的第 i 类所有模糊隶属度之和为 1，即 $\sum_{i=1}^{c}u_{ik,FCM}=1$。如果 $u_{ik,FCM}>0.5$，则判定 x_k 隶属于第 i 类；反之，如果 $u_{ik,FCM}<0.5$，则判定 x_k 不隶属于第 i 类。PFCM 和 PFDCM 聚类算法的模糊隶属度与 FCM 模糊隶属度在判定样本 x_k 属于哪一类的方法相同。但是，PFCM 和 PFDCM 聚类算法除了模糊隶属度外还有典型值，PFDCM 的典型值没有 x_k 的第 i($i=1,2,3,4$)类所有模糊隶属度之和不为 1 这个约束条件，即 $\sum_{i=1}^{c}t_{ik,PFCM}\neq1$ 和 $\sum_{i=1}^{c}t_{ik,PFDCM}\neq1$。对于给定 x_k，x_k 隶属于第

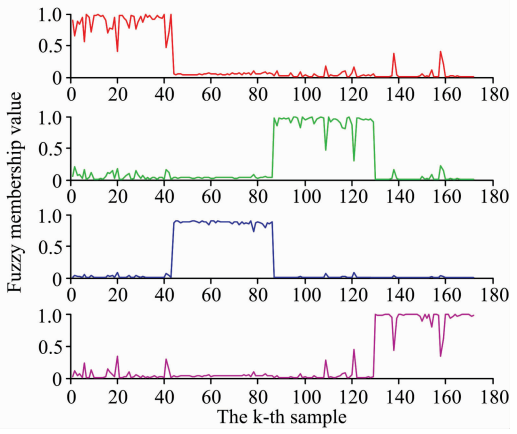

图 3　PFDCM 模糊隶属度值

Fig. 3　Fuzzy membership values of PFDCM

i类的典型值有4个，其中最大典型值所在的类别即为所属类别。根据 $u_{i\text{k}.\,\text{FCM}}$，$u_{i\text{k}.\,\text{PFCM}}$ 和 $u_{i\text{k}.\,\text{PFDCM}}$ 可计算出 FCM，PFCM 和 PFDCM 的聚类准确率分别为 93.60%，93.02% 和 98.84%。PFDCM 的聚类准确率最高。

3　结　论

可能模糊 C 均值聚类(PFCM)基础上结合模糊线性判别分析(FLDA)，提出一种可能模糊鉴别 C 均值聚类(PFDCM)算法。PFDCM 聚类算法可实现在模糊聚类过程中提取样本鉴别信息，进一步提高了聚类准确率。通过对茶叶傅里叶近红外光谱进行 FCM，PFCM 和 PFDCM 三种聚类算法进行光谱聚类分析。结果表明：当权重指数 m 为 2 时，PFDCM 聚类算法的聚类准确率最高，达到 98.84%；PFCM 和 PFDCM 的聚类迭代次数均少于 FCM；PFDCM 聚类时间要多于 FCM 和 PFCM。

References

[1] Zhuang X G, Wang L L, Chen Q, et al. Science China-Technological Sciences, 2017, 60(1): 84.
[2] Ouyang Q, Liu Y, Chen Q, et al. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2017, 180: 91.
[3] Wu X H, Zhu J, Wu B, et al. Computers and Electronics in Agriculture, 2018, 147: 64.
[4] Yu X L, He Y. Spectroscopy Letters, 2018, 51(2): 112.
[5] Bartoszek M, Polak J, Chorazewski M. European Food Research and Technology, 2018, 244(4): 595.
[6] Meng W J, Xu X N, Cheng K K, et al. Food Analytical Methods, 2017, 10(11): 3508.
[7] Wang S P, Gong Z M, Su X Z, et al. Journal of Applied Spectroscopy, 2017, 84(4): 704.
[8] Hu L Q, Yin C L. Food Analytical Methods, 2017, 10(7): 2281.
[9] Zhang T. International Journal of Pattern Recognition and Artificial Intelligence, 2018, 32(9): 1857005.
[10] Kim E H, Oh S K, Pedrycz W. Neural Networks, 2018, 104: 1.
[11] Ghosh P, Mali K, Das S K. Journal of Visual Communication and Image Representation, 2018, 54: 63.
[12] Koutroumbas K D, Xenaki S D, Rontogiannis A A. IEEE Transaction on Fuzzy System, 2018, 26(1): 324.
[13] Wu X H, Wu B, Sun J, et al. Journal of Food Process Engineering, 2017, 40(2): e12355.
[14] Wu B, Wilamowski B M. IEEE Transactions on Industrial Informatics, 2017, 13(4): 1620.
[15] Askari S, Montazerin N, Fazel Zarandi M H, et al. Neurocomputing, 2017, 219: 186.
[16] WU Xiao-hong, ZHOU Jian-jiang(武小红，周建江). Acta Electronica Sinica(电子学报), 2008, 36(10): 1996.
[17] Ma J, Pu H B, Sun D W. LWT—Food Science and Technology, 2018, 94: 119.
[18] 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.

Classification of FTNIR Spectra of Tea via Possibilistic Fuzzy Discriminant C-Means Clustering

WU Bin[1*], FU Hai-jun[2], WU Xiao-hong[2,3*], CHEN Yong[2], JIA Hong-wen[1]

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

Abstract Fourier transform near-infrared spectroscopy (FTNIR) spectra contain valuable information about the chemical constituents of tea. Furthermore, the chemical constituents and their content of tea reveal differences concerning different kinds of tea and, therefore, it is feasible to classify tea varieties by FTNIR. FTNIR spectra have the characteristics of high dimension, crests and troughs, spectral overlapping and staggering, so it is difficult to classify spectra. In order to solve this problem, possibilistic fuzzy discriminant c-means clustering (PFDCM) was proposed by introducing fuzzy linear discriminant analysis (FLDA) into possibilistic fuzzy c-means clustering (PFCM) for purpose of discriminating FTNIR spectra correctly. Interestingly, during fuzzy clustering FLDA can not only extract discriminant information from FTNIR spectra but can transform the data space. PFDCM can achieve the accurate classification of FTNIR spectra according to its fuzzy membership and typicality values, and it has some advantages such as fast speed and high accuracy. PFDCM is superior to fuzzy c-means (FCM) clustering in clustering spectra containing noisy data because the typicality values of PFDCM are no constraint that the sum of the membership degrees is one. Four varieties of tea samples, called Yuexi Cuilan, Lu'an Guapian, Shiji Maofeng and Huangshan Maofeng, were collected in this study, and a total of 260 tea samples were scanned over the range of 10 000～4 000 cm^{-1} by FTNIR spectrometer, and in the end the 1 557-dimensional data were acquired for further processing. For a start, spectral data were pretreated with multiplicative scatter correction (MSC) to reduce spectra scattering and noise effect and increase signal-to-noise ratio. Secondly, principal component analysis (PCA) was used to reduce the dimensionality of FTNIR spectra to seven. Thirdly, discriminant information was extracted from spectra and the dimensionality of data was transformed from seven to three by linear discriminant analysis (LDA). Finally, fuzzy c-means (FCM) clustering, PFCM and PFDCM were put into use, clustering data to classify tea variety correctly. The experimental results showed that under the condition of the weight index $m=2.0$ and $\eta=2.0$, the clustering accuracy rates of FCM, PFCM and PFDCM achieved 93.60%, 93.02% and 98.84%, respectively. After 25 iterations, FCM converged, but PFCM and PFDCM achieved 8 iterations and 23 iterations, respectively, and converged. As fuzzy clustering algorithms converged, FCM consumed the least time but the most time-consuming clustering was PFDCM. In conclusion, FTNIR coupled with MSC, PCA, LDA and PFDCM presented a classification model for the accurate identification of tea varieties.

Keywords Tea; Near-infrared spectroscopy; Principal component analysis; Discriminant information extraction; Fuzzy clustering

(Received Jan. 8, 2019; accepted May 23, 2019)

* Corresponding authors