第42卷，第5期
2022年5月

光谱学与光谱分析
Spectroscopy and Spectral Analysis

Vol. 42, No. 5, pp1465-1470
May, 2022

GK可能C均值模糊聚类的白菜红外光谱分析

谭　阳[1]，武小红[2,3*]，武　斌[4]，沈砚君[1]，刘锦茂[1]

1. 江苏大学卓越学院，江苏 镇江　212013
2. 江苏大学电气信息工程学院，江苏 镇江　212013
3. 江苏省农业装备与智能化高技术研究重点实验室，江苏 镇江　212013
4. 滁州职业技术学院信息工程学院，安徽 滁州　239000

摘　要　红外光谱分析是基于分子振动与跃迁理论的鉴别物质化学组成的技术。得到的光谱数据常常具有较高的维数和重叠度，这给后续的数据处理带来困难。为此提出一种GK可能C均值聚类算法(GKIPCM)，引入了GK聚类算法的马氏距离测度与改进的可能C均值聚类算法(IPCM)的模糊隶属度与聚类中心更新方程，使样本的距离测度具有自适应性且避免了聚类中心的一致性。GKIPCM算法具有分类精度更高，分类准确率对参数敏感性低的优点。将四组洗净白菜作为光谱分析对象，分别施加三种农药(高效氯氟氰菊酯)配比，采用安捷伦Cary 630 FTIR光谱仪采集白菜的傅里叶中红外光谱(FT-MIR)。首先对样本进行预处理，使用多元散射矫正(MSC)对光谱数据降噪，消除数据偏移量；其次，由于采集到的数据波数范围为4 300～590 cm^{-1}，数据维数达到了971维，故使用主成分分析(PCA)对数据实现降维，降维后的数据维度减小到了23，且23个主成分的累积贡献率高达99.60%；但各类光谱的特征信息依然混杂在一起，故使用线性判别分析(LDA)提取特征鉴别信息，进一步将数据降至3维；最终，运行模糊C-均值聚类算法(FCM)得到较优初始聚类中心，使用GKIPCM算法对四类降维后的光谱数据进行聚类分析，并与GK聚类算法与IPCM聚类算法的运行结果作对比。GKIPCM算法的总迭代时长为0.218 8 s，分类准确率达到了97.22%。相较之下，GK算法与IPCM算法的准确率分别为63.89%和91.67%，运行的总时长为0.093 8与0.062 5 s。从实验结果可看出，GKIPCM算法可以通过分析光谱数据从而完成对不同程度农药残留进行定性分析的任务。

关键词　白菜；农药残留；光谱分析；主成分分析；线性判别分析；模糊聚类
中图分类号：O657.33　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2022)05-1465-06

引　言

民以食为天，食以安为先。我国自古以来就是一个人口数量庞大的农业大国，对于食品质量的重视程度从未放松。为了保证粮食的产量与营养价值，往往需要对作物定期喷洒定量的农药，以消灭害虫。然而，农药的残留对人类健康与自然环境却有所危害[1-3]。当农作物流入市场时，所施加的农药总会或多或少地残留在农作物的表面，其浓度由于初始浓度，喷洒周期，贮存时间等因素而各不相同。我国有严格的进口食品安全管理体制和进口蔬菜农药标准，故准确有效的蔬菜农药残留检测分类与国计民生息息相关，具有重要的研究价值。

国内外学者在农药残留检测方面做了很多的研究，较为常见的方法有传统的化学检测与红外光谱分析等。红外光谱分析的独特优势，推动了农药残留检测方法的进一步发展，实现了快速无损化；Jiang等借助气相色谱法，使用近红外高光谱成像系统预测桑叶中农药残留的分布[4]；Zhou等采用偏振光谱检测技术收集偏振光信息，使用BP神经网络，K近邻和支持向量机建立分类模型，最后得到不同农药残留的校准识别率为100%，预测识别率为97.78%[5]；Sun等提出了一种结合化学分子结构和小波变换的方法提取特征波长，

收稿日期：2021-04-17，**修订日期**：2021-06-12
基金项目：国家自然科学基金项目(31471413)，安徽省教育厅高校自然科学研究重点项目(KJ2019A1129)，滁州职业技术学院校级自科重点项目(YJZ-2020-12)，滁州职业技术学院院级人才项目“优秀骨干教师”(YG2019026，YG2019024)，江苏大学大学生创新训练计划项目(202010299244Y)资助
作者简介：谭　阳，2000年生，江苏大学卓越学院本科生　　e-mail：TaNvYaN@126.com
*通讯作者　　e-mail：wxh_www@163.com

通过提取的特征光谱数据建立支持向量机模型，校准与预测精度达到 100%[6]；Wu 等提出深简网络和近红外透射光谱相结合的检测方法，通过支持向量机、PLS-DA 和 K 近邻法建立分类模型，训练与测试的准确率分别达到了 98.89%和 95.00%[7]；Sun 等基于生菜叶片的高光谱数据，将竞争性自适应重加权采样与连续投影算法应用于最小二乘支持向量回归模型，对生菜叶片中混合农药残留进行定量检测[8]。

聚类分析是一种无监督的机器学习算法，广泛应用于模式识别领域[9]，例如 K 均值聚类算法[10]。模糊聚类分析将模糊理论[11]与聚类算法结合，每一个数据样本可同时隶属于多个聚类，以数据特征相似性度量来识别类群。例如：为了解决模糊 C-均值聚类算法(fuzzy c-means clustering，FCM)[12]对噪声环境的敏感问题，引入可能性的概念代替模糊隶属度，修改目标函数约束条件的可能 C 均值聚类算法[13]；基于模糊马氏距离的自适应改进可能 C 均值聚类算法[14]；使用马氏距离测度预测非椭球形数据的 GK (gustafson-kessel clustering)算法[15]；Wu 等提出的结合模糊协方差矩阵的可能性模糊 C 均值聚类算法[16]。为了得到更适合含农药残留的白菜的中红外光谱聚类分析算法，并在已有的算法上进一步提高分类的准确率且保持算法的优良特性，本文结合 GK 算法与改进的可能 C 均值聚类算法(improved possibilistic c-means clustering，IPCM)，提出一种 GK 可能 C 均值聚类算法(gustafson-kesselimproved possibilistic c-means clustering，GKIPCM)。

首先使用中红外光谱仪采集白菜的中红外光谱数据，其次依次使用多元散射矫正，主成分分析，线性判别分析对数据进行预处理，然后应用 GKIPCM 对数据进行聚类分析，最后对比 GK 聚类与 IPCM 聚类的分类准确率得出结论：GKIPCM 算法可以完成对不同浓度农药残留的定性分析。

1 实验部分

1.1 白菜-高效氯氟氰菊酯光谱数据采集

1.1.1 材料

本实验对象为采购于菜市场的新鲜小白菜，将其以 45 ℃温水洗净后作为施以农药的原材料。选取高效氯氟氰菊酯为实验农药，使用喷雾方式确保农药喷洒均匀。将实验原材料分为四组，第一组不喷洒农药，作为对照实验组；第二、三、四组的农药配比分别为 1∶500，1∶100，1∶20。阴干后将白菜分解成小块，便于进行中红外光谱数据的收集。

1.1.2 光谱仪器与分析软件

仪器选用安捷伦 Cary 630 FTIR 光谱仪，调至 ATR 衰减全反射光谱采集模式。扫描背景与样本各 64 次，分辨率调整为 8 cm^{-1}。配合 Microlab PC、Resolutions Pro 软件，每组样品采集 40 组数据，共计 160 组，每组数据有 971 维，波数变化范围为 4 300～590 cm^{-1}。白菜傅里叶中红外光谱(Fourier transform mid-infrared spectroscopy，FT-MIR)如图 1 所示，采用 Matlab R2016b 对该光谱数据进行定性分析。

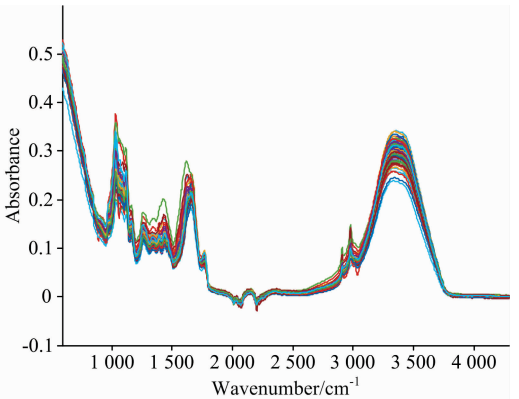

图 1 原始傅里叶变换中红外光谱数据图

Fig. 1 Raw FT-MIR spectral data

1.2 GK 可能 C 均值聚类方法

(1)参数初始化

设置权重指数 $m(m>1)$，测试样本的维数 d，循环迭代次数的初始值 $r=0$，最大允许迭代次数 $r_{\max}=100$，迭代的最大误差参数 $\varepsilon=0.000\ 01$；运行 FCM 算法并将最终模糊隶属度与类中心作为 GKIPCM 算法的初始模糊隶属度 $\boldsymbol{U}_{\mathrm{GKIPCM}}^{(0)}$ 与初始聚类中心 $\boldsymbol{V}_{\mathrm{GKIPCM}}^{(0)}$。

(2)计算第 r 次迭代时的模糊散射矩阵 $\boldsymbol{S}_{fi,r}$

$$\boldsymbol{S}_{fi,r}=\frac{\sum_{k=1}^{n}u_{ik,r-1}^{m}(\boldsymbol{x}_k-\boldsymbol{v}_{i,r-1})(\boldsymbol{x}_k-\boldsymbol{v}_{i,r-1})^{\mathrm{T}}}{\sum_{k=1}^{n}u_{ik,r-1}^{m}} \tag{1}$$

式(1)中，$u_{ik,r-1}$ 为第 $r-1$ 次迭代后第 k 个测试样本对第 i 类的模糊隶属度；n 为样本总量；$\boldsymbol{x}_k$ 为第 k 个白菜中红外光谱测试样本；$\boldsymbol{v}_{i,r-1}$ 为第 $r-1$ 次迭代后第 i 类的类中心($i=1,2,3,4$)；$\boldsymbol{S}_{fi,r}$ 是第 r 次迭代后第 i 类样本的模糊散射矩阵。

(3)计算第 r 次迭代后第 i 个聚类中心的范数矩阵 $\boldsymbol{A}_{i,r}$

$$\boldsymbol{A}_{i,r}=\|\boldsymbol{S}_{fi,r}\|^{\frac{1}{d}}(\boldsymbol{S}_{fi,r}^{-1}) \tag{2}$$

(4)由式(3)计算第 r 次迭代后测试样本 $\boldsymbol{x}_k$ 到类中心 $\boldsymbol{v}_{i,r-1}$ 的距离范数 $D_{ik,r}$

$$D_{ik,r}=\sqrt{(\boldsymbol{x}_k-\boldsymbol{v}_{i,r-1})^{\mathrm{T}}\boldsymbol{A}_{i,r}(\boldsymbol{x}_k-\boldsymbol{v}_{i,r-1})},\ \forall i,k \tag{3}$$

(5)计算第 r 次迭代后第 k 个测试样本隶属于第 i 类的典型值 $t_{ik,r}$

$$t_{ik,r}=\exp\left(-\frac{D_{ik,r}^{2}}{\|\boldsymbol{S}_{fi,r}\|^{\frac{1}{d}}}\right) \tag{4}$$

(6)计算第 r 次迭代后的模糊隶属度矩阵 $\boldsymbol{U}_r$，其各元素值计算公式为：

$$u_{ik,r}=\left(\sum_{j=1}^{c}\left(\frac{\|\boldsymbol{S}_{fi,r}\|^{\frac{1}{d}}(1-t_{ik,r})}{\|\boldsymbol{S}_{fj,r}\|^{\frac{1}{d}}(1-t_{jk,r})}\right)^{\frac{1}{m-1}}\right)^{-1} \tag{5}$$

式(5)中，c 为样本类别数。

(7)计算第 r 次迭代后第 i 类的类中心 $\boldsymbol{v}_{i,r}$

$$\boldsymbol{v}_{i,r}=\frac{\sum_{k=1}^{n}u_{ik,r}^{m}t_{ik,r}\boldsymbol{x}_k}{\sum_{k=1}^{n}u_{ik,r}^{m}t_{ik,r}} \tag{6}$$

(8)计算相邻两次迭代后两聚类中心的距离范数 D_r

$$D_r=\|\boldsymbol{v}_{i,r}-\boldsymbol{v}_{i,r-1}\| \tag{7}$$

(9)对已计算数据进行如下判断：

若 $D_r<\varepsilon$ 或 $r\geqslant r_{\max}$，则停止运行，记录迭代结束时模糊聚类中心 $\boldsymbol{V}_{\text{end}}$，模糊隶属度矩阵 $\boldsymbol{U}_{\text{end}}$ 以及典型值矩阵 $\boldsymbol{T}_{\text{end}}$；否则令 $\boldsymbol{v}_{i,r-1}=\boldsymbol{v}_{i,r}$，$u_{ik,r-1}=u_{ik,r}$，回到步骤(2)继续迭代计算。

(10)根据模糊隶属度矩阵与典型值矩阵对样本进行分类：

对于模糊隶属度，若 u_{ij} 为 $\boldsymbol{u}_j$ 中的最大值，则判断第 j 个样本属于第 i 类；对于典型值，若 t_{ij} 为 $\boldsymbol{t}_j$ 中的最大值，则认为第 j 个样本有更大的可能性属于第 i 类。

2　结果与讨论

2.1　中红外光谱数据的预处理(MSC，PCA，LDA)

将样本数据分为训练集与测试集。其中训练集共 4 类，每类 22 个样本；测试集共 4 类，每类 18 个样本。在光谱数据采集的过程中，散射水平的差异往往造成白菜表面实际农药的残留量与光谱波长的数据相关性的下降。为了减小此类差异且提高信噪比，首先使用多元散射矫正(multiplicative scattering correction，MSC)对光谱数据进行预处理。图 2 为使用多元散射矫正后傅里叶中红外光谱数据图。

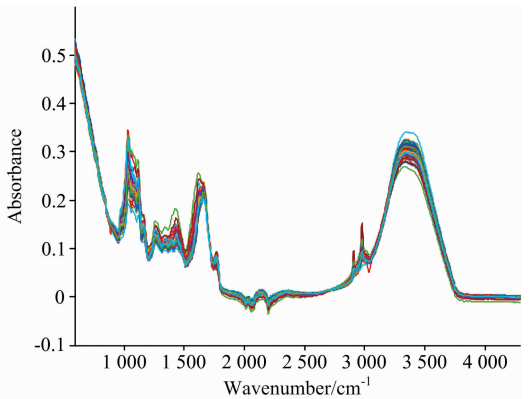

图 2　多元散射矫正后傅里叶中红外光谱数据图

Fig. 2　FT-MIR spectral data processed by MSC

为观察不同农药残留程度的光谱数据的吸光度有所差异，计算了每一类光谱数据的平均值并截取了波数介于 800 与 1 500 之间的部分如图 3 所示。

不同农药浓度样品的中红外光谱吸光度有着明显差值，直接保证了不同类光谱数据的可分性。

尽管 MSC 将基线平移并且消除了偏移量，但数据依然达到了 971 维。为减小运算量，采用主成分分析(principal component analysis，PCA)对数据实现降维并提高分类准确率。通过训练数据集的协方差矩阵得到前 23 个主成分，累计贡献率高达 99.60%。故通过该 23 个主成分将光谱数据缩小至 23 维。为了提取样本数据的相关特征信息并进一步降维，使用线性判别分析(linear discriminant analysis，LDA)处理训练样本。此处求取了 3 个判别向量与特征值。其中特征值的计算结果为：$\lambda_1=7.371\ 3$，$\lambda_2=3.932\ 0$，$\lambda_3=1.931\ 7$。将 PCA 处理后的训练数据投影到判别向量构成的特征空间，得到线性判别分析得分图，见图 4。

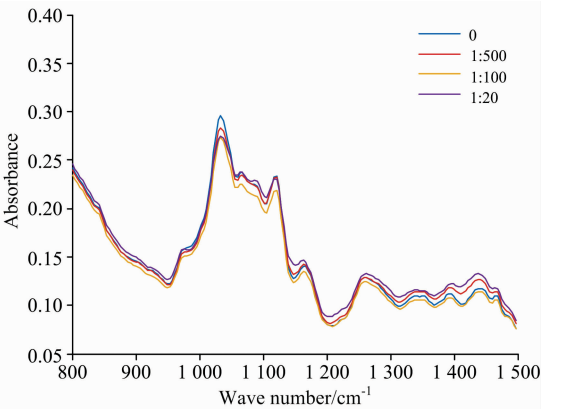

图 3　多元散射矫正后白菜平均中红外光谱图

Fig. 3　The average FT-MIR spectral data processed by MSC

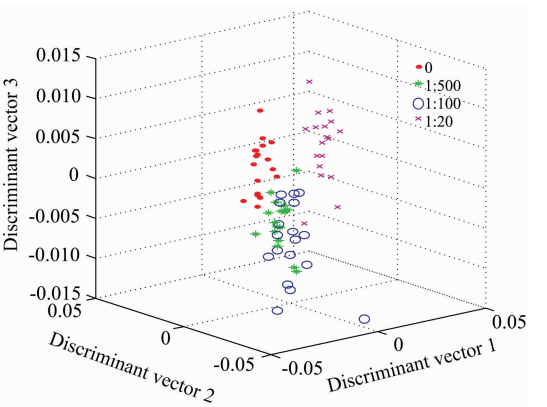

图 4　线性判别分析得分图

Fig. 4　Scores plot of LDA algorithm

图中，“·”，“*”，“○”，“×”分别表示农药配比为无农药，1∶500，1∶100，1∶20 的光谱数据。无农药的样本与配比为 1∶20 的样本区分度最高，相比之下农药配比为 1∶100 和 1∶500 的两类样本区分度稍弱。尽管如此，四类样本整体的区分度较强，说明线性判别分析的特征提取结果是可取的。于是将 PCA 的 23 组线性变换(主成分)与 LDA 的 3 个鉴别向量应用于测试集，为模糊聚类做准备。

2.2　FCM，GK，IPCM，GKIPCM 聚类算法

2.2.1　模糊聚类相关参数的初始化

设置权重指数 $m=2.0(m\geqslant 1)$，测试样本的维数 $d=3$；由于 FCM，GK，IPCM，GKICPM 算法都属于迭代算法，设置迭代次数初始值 $r=1$，最大迭代次数 $r_{\max}=100$；当迭代误差小于 $\varepsilon=0.000\ 01$ 或 $r\geqslant r_{\max}$ 时停止迭代。FCM 的初始聚类中心如式(8)

$$\boldsymbol{V}_{\text{FCM}}^{(0)}=\begin{pmatrix}\boldsymbol{v}_{\text{FCM},1}^{(0)}\\ \boldsymbol{v}_{\text{FCM},2}^{(0)}\\ \boldsymbol{v}_{\text{FCM},3}^{(0)}\\ \boldsymbol{v}_{\text{FCM},4}^{(0)}\end{pmatrix}=\begin{pmatrix}-0.012\ 8 & 0.003\ 6 & -0.002\ 4\\ -0.012\ 3 & -0.001\ 6 & 0.003\ 9\\ -0.006\ 2 & 0.003\ 6 & 0.005\ 3\\ -0.007\ 9 & 0.000\ 5 & 0.002\ 2\end{pmatrix} \tag{8}$$

对于 GK，IPCM，GKIPCM，将 FCM 的最终聚类中心与

模糊隶属度作为算法的初始值。

2.2.2 模糊聚类算法的迭代时长

运行 GK，IPCM，GKIPCM，对比迭代次数与收敛时间：GK 迭代了 100 次，运行时间为 0.093 8 s；IPCM 迭代了 13 次，运行时间为 0.062 5 s；GKIPCM 迭代了 87 次，运行时间为 0.218 8 s。处理器：Inter(R) Core(TM) i5-8300H CPU @ 2.30GHz(8 CPUs)；内存：8192MB RAM。

2.2.3 模糊隶属度分类结果

参照 2.2.1 将各参数初始化后，分别运行 GK，IPCM 与 GKIPCM。得到 GK，IPCM，GKIPCM 对应的模糊隶属度分别为 $u_{ik,\mathrm{GK}}$，$u_{ik,\mathrm{IPCM}}$，$u_{ik,\mathrm{GKIPCM}}$，表示第 k 个测试样本对第 i 类的模糊隶属度；同时，IPCM 与 GKIPCM 还具有典型值 $t_{ik,\mathrm{IPCM}}$，$t_{ik,\mathrm{GKIPCM}}$，表示第 k 个测试样本属于第 i 类的可能性，其中 GKIPCM 算法的典型值分布如图 5 所示。典型值取消了测试样本 $\boldsymbol{x}_k$ 对各类的隶属度之和为 1 的约束条件，减弱了聚类过程中噪声对分类的影响。

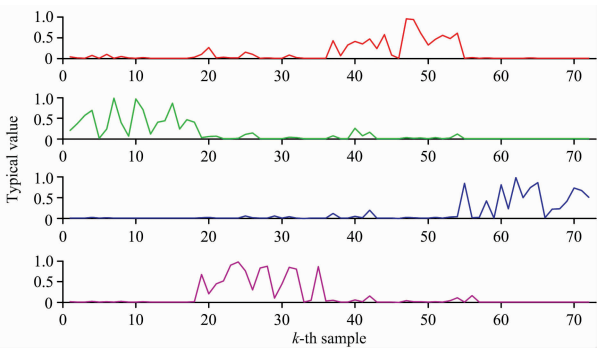

图 5 GKIPCM 典型值

Fig. 5 Typical value of GKIPCM

对于模糊隶属度 u_{ik} 与典型值 t_{ik}，若 u_{ik} 为 $\boldsymbol{u}_k$ 中的最大值或 t_{ik} 为 $\boldsymbol{t}_k$ 中的最大值则认为第 k 个测试样本可以被划分到第 i 个类别当中。最终根据模糊隶属度与典型值得到 GK，IPCM，GKIPCM 的分类精度分别为：63.89%，91.67%，97.22%。

2.3 参数讨论

2.3.1 权重指数 m

在模糊聚类算法中，权重指数 m 对算法的准确程度有着重要影响。对于 GKIPCM，当 $m\geqslant 2$ 时，模糊隶属度矩阵在迭代过程中有着较好的收敛性；但由于同一样本对不同类的典型值没有和为 1 这一约束条件，m 过大时式(5)的计算结果会受到干扰，甚至使下一次迭代时式(1)计算的模糊散射矩阵 $\boldsymbol{S}_{fi,r}$ 接近奇异值，影响分类准确度。此处固定主成分个数为 23，研究权重指数 m 在区间[2，6]上时，各算法的聚类准确率，结果如表 1 所示。

由表 1 可见，m 在有限范围内，GKIPCM 聚类准确率高于 GK 聚类与 IPCM 聚类。

2.3.2 主成分个数

在 PCA 降维过程中，通常使得主成分的累积贡献率达到一个较高的百分比(80%～90%)，否则样本数据维度降低的同时也会损失大量的特征鉴别信息。此处固定 $m=2$，通过调整主成分的个数从而调整累计贡献率，观察对聚类准确率的影响，结果如表 2 所示。

表 1 不同参数 m 的 GK，GKIPCM，IPCM 算法聚类准确率

Table 1 Clustering accuracies of GK，GKIPCM，IPCM with different values of m

m	GK 聚类	GKIPCM 聚类	IPCM 聚类
2.0	63.89%	97.22%	91.67%
2.4	90.28%	97.22%	91.67%
2.8	94.44%	95.83%	91.67%
3.2	94.44%	95.83%	94.44%
3.6	94.44%	94.44%	94.44%
4.0	94.44%	95.83%	94.44%
4.4	94.44%	95.83%	95.83%
4.8	94.44%	97.22%	87.50%
5.2	95.83%	97.22%	86.11%
5.6	95.83%	97.22%	87.50%
6.0	97.22%	93.06%	90.28%

表 2 不同主成分的个数下 GK，GKIPCM，IPCM 算法的准确率

Table 2 Clustering accuracies of GK，GKIPCM，IPCM with different numbers of principal components

主成分个数	GK 聚类	GKIPCM 聚类	IPCM 聚类
11	56.94%	51.39%	59.72%
13	56.94%	87.50%	54.17%
15	88.89%	87.50%	83.33%
17	86.11%	87.50%	86.11%
19	86.11%	94.44%	88.89%
21	47.22%	95.83%	91.67%
23	63.89%	97.22%	91.67%
25	88.89%	94.44%	91.67%
27	93.06%	94.44%	94.44%
29	86.11%	95.83%	95.83%
31	94.44%	95.83%	95.83%
33	93.06%	95.83%	91.67%
35	81.94%	93.06%	91.67%
37	93.06%	94.44%	90.28%
39	94.44%	93.06%	90.28%

由表 2 知：随着主成分个数的增加，GK，GKIPCM，IPCM 的准确率都呈现上升趋势，GK 的准确率相较其他算法对参数敏感一些，GKIPCM 聚类的准确率在主成分个数为 23 时高达到 97.22%。综合表 1、表 2，该 GKIPCM 算法适合处理农药残留程度的定性分析问题。

3 结 论

将 GK 的马氏距离测度与 IPCM 的模糊隶属度和聚类中心更新方程结合，提出了一种新型 GKIPCM 算法。实验证明，对农药残留的白菜傅里叶中红外光谱数据采用 GKIPCM

算法进行分析，准确率达到了 97.22%(例如：$m=2$，主成分个数等于 23)，分类效果优于 GK 与 IPCM 聚类算法，具有较高的实际应用价值。

References

[1] Azam S M R, Ma H l, Xu B G, et al. Trends in Food Science and Technology, 2020, 97: 417.
[2] Choudhary S, Raheja N, Kumar S, et al. Journal of Entomology and Zoology Studies, 2018, 6(3): 330.
[3] Alenyorege E A, Ma H l, Ayim I, et al. Journal of Food Safety and Food Quality, 2018, 69(3): 80.
[4] Jiang S Y, Sun J, Zhou X, et al. Journal of Food Process Engineering, 2017, 40(4): e12510.
[5] Zhou X, Sun J, Lu B, et al. Journal of Food Process Engineering, 2018, 41(8): e12903.
[6] Sun J, Zhou X, Mao H P, et al. Journal of Food Process Engineering, 2016, 40(4): e12509.
[7] Wu M M, Sun J, Lu B, et al. Journal of Food Process Engineering, 2019, 42(3): e13005.
[8] Sun J, Cong S, Mao H P, et al. Journal of Food Process Engineering, 2018, 41(2): e12654.
[9] Sinaga K P, Yang M S. IEEE Access, 2020, 8(1): 80716.
[10] Moradi Fard M, Thonet T, Gaussier E. Pattern Recognition Letters, 2020, 138: 185.
[11] Yuan Z, Chen H M, Li T R, et al. Information Sciences, 2021, 572: 67.
[12] Zhang L L, Luo M N, Liu J, et al. Pattern Recognition Letters, 2020, 130: 275.
[13] Yu H Y, Fan J L, Lan R. Applied Soft Computing Journal, 2019, 80: 845.
[14] Zhou H X, Fu H J, Wu X H, et al. Journal of Food Processing and Preservation, 2020, 44(10): e14795.
[15] Chaomurilige, Yu J, Yang M S. Information Sciences, 2017, 417: 435.
[16] Wu X H, Zhou J, Wu B, et al. Journal of Food Process Engineering, 2019, 42(8): e13298.

Qualitative Analysis of Pesticide Residues on Chinese Cabbage Based on GK Improved Possibilistic C-Means Clustering

TAN Yang[1], WU Xiao-hong[2,3*], WU Bin[4], SHEN Yan-jun[1], LIU Jin-mao[1]

1. Institute of Talented Engineering Students, Jiangsu University, Zhenjiang 212013, China
2. School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China
3. High-tech Key Laboratory of Agricultural Equipment and Intelligence of Jiangsu Province, Jiangsu University, Zhenjiang 212013, China
4. Department of Information Engineering, Chuzhou Polytechnic, Chuzhou 239000, China

Abstract Infrared spectroscopy is a technology used to identify the chemical composition of substances based on molecular vibration and quantum-jump theory. Due to the unique absorbance of different functional groups, the spectral data related to the absorbance and the wavelength (or wavenumber) can be obtained when the infrared beam irradiates the molecular. However, the spectral data from experiments always have high dimensions and overlap, making it difficult to process the data. Thus, this paper proposed an improved Gustafson-Kessel possibilistic c-means clustering (GKIPCM), introducing the Mahalanobis distance from GK clustering and the iterative equations of fuzzy membership values and cluster centers from improved possibilistic c-means clustering (IPCM). GKIPCM makes the data adapt to different mathematical distance measures and avoids identical cluster centers. Furthermore, GKIPCM has higher classification accuracy, which is less sensitive to parameters. In the experiments, four groups of washed Chinese cabbage were the objects of spectral analysis and different concentrations of lambda-cyhalothrin pesticide were sprayed on the Chinese cabbages. Spectral data of Chinese cabbages were collected with Agilent Cary 630 FTIR spectrometer. Firstly, multiplicative scatter correction (MSC) was applied to reduce the noise and eliminate data offset when pre-processing the data. Secondly, principal component analysis (PCA) was utilized to reduce dimensions due to the wide wavenumber range (4 300～590 cm^{-1}) and the high data dimensions (971). After conducting PAC, the dimensionality of data was reduced to 23, and the total contribution of 23 principal components reached 99.60%. Nonetheless, the feature information was still mixed. So the linear discriminant analysis (LDA) was used to extract features of the spectral data, and the LDA algorithm reduced the dimensionality of the spectral data to 3. Finally, the fuzzy c-means clustering (FCM) was employed to obtain the optimal initial cluster centers. Then, the GKIPCM algorithm was applied to cluster four different groups of spectral data. Comparisons were made among the clustering results of GKIPCM, GK and IPCM. The running time and accuracy of

GKIPCM were 0.218 8 seconds and 97.22%, and those of GK and IPCM were 0.093 8 seconds and 63.89%, 0.062 5 seconds and 91.67%, respectively. According to the results of the experiments, the GKIPCM algorithm finished the qualitative analysis of different concentrations of pesticide residues by analyzing the spectral data.

Keywords Chinese cabbage; Pesticide residues; Infrared spectroscopy; Principal component analysis; Linear discriminant analysis; Fuzzy clustering

(Received Apr. 17, 2021; accepted Jun. 12, 2021)

* Corresponding author