基于时频谱特征的白酒品质分类方法研究

祝海江[1]，唐 昊[1]，孙静娴[1]，杜振霞[2]

1. 北京化工大学信息科学与技术学院，北京 100029
2. 北京化工大学化学学院，北京 100029

摘 要 为了建立快速、准确的白酒品质鉴别方法，利用机器学习方法对不同品质的白酒建模。为了提取不同品质白酒的特征，使用离子迁移谱对不同品质白酒进行分析，构建了基于白酒离子迁移谱信号的特征向量，并对不同品质的白酒进行了识别与分类。白酒样本的离子迁移谱信号通过利用美国 Excellims 公司 GA2100 型电喷雾-离子迁移谱仪(ESI-IMS)采集获得，每一个离子迁移谱信号是强度随时间变化的时间序列信号；提取了原始数据离子迁移谱的时域特征谱峰。为了获得更全面的特征，对离子迁移谱数据进行了傅里叶变换并提取频域内的特征谱峰。同时为了表述信号变化的特征，计算了离子迁移谱的谱熵和过零率，构建 $N\times9$ 维的特征向量矩阵；使用主成分分析(PCA)和线性判别分析(LDA)分别对上述获得的特征进行了特征降维，其中使用 PCA 对特征向量矩阵降维后的前三维特征对整体特征的累计贡献率达到了 95%，而使用 LDA 对特征向量矩阵降维后的前两维特征对整体特征的累计贡献率就达到了 95%。因此，选择了 LDA 作为特征降维方法；最后，利用机器学习中的非线性分类器支持向量机(SVM)对白酒离子迁移谱数据进行分类研究。实验结果表明，在真酒和添加酒精的白酒二分类中，SVM 方法正确分类率达到 100%；而在真酒和分别添加 10%，20%，30%，40%和 50%酒精浓度的五种假酒的六分类中，SVM 方法正确分类率达到 99.7%。比较了逻辑回归(LRM)分类、模糊 C 均值分类(FCM)和 K 近邻分类(KNN)对白酒样本离子迁移谱分类实验结果。研究表明，对于离子迁移谱非常接近的真酒和添加酒精的白酒，基于频谱特征向量的 SVM 方法能够准确的区分开来，为白酒的品质鉴别提供了一种新的检测方法。

关键词 离子迁移谱；白酒品质；支持向量机；时频谱特征

中图分类号：TP181 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2021)09-2962-07

引 言

白酒是中国特产的一种酒类，因其独特的生产原料和工艺，颇受世界各地消费者的喜爱。然而，由于生产工艺、产地、原料等各种因素影响，同一品牌的白酒质量存在不一致的情况，甚至同一酒厂不同生产班次所生产的酒的品质也不一致。因此，白酒品质的分析检测是白酒行业科技发展的一个重要组成部分。

在白酒酿造企业中，鉴别白酒品质常用的方法有：感官品尝法、电子鼻或色谱仪等仪器检测方法。但是，由于生产工艺不断发展，生产原料变得多样化，同时也衍生了不同的制酒工艺，导致现有的鉴定方法存在一定的缺陷。感官评定法虽然便捷，但由于其鉴定模式多基于人为评定，且伴随着白酒种类的多样化，白酒的品鉴结果会受到人为主观因素的影响；电子鼻的价格昂贵，成本较高，同时受到传感器材料和数据处理的限制，检测效果不佳。相对而言，仪器检测的结果会更客观，检测装置的广泛应用为评价白酒提供了标准。

在白酒的频谱分析方面，Yu[1]用近红外光谱和化学计量方法测定了黄酒的酒龄，实验采集了 86 瓶绍兴黄酒的近红外光谱，对原始光谱、平滑处理的光谱以及二阶微分处理的光谱分别使用判别分析法建立酒龄鉴定模型。实验结果表明近红外透射原始光谱结合判别分析法最佳，可作为检验黄酒年龄的一种有效方式。吕海棠等[2]对清香型和浓香型这 2 种不同的白酒去除白酒中的水分，进行干燥萃取，然后通过红外光谱对剩余的干燥物酒类进行定量分析，结果显示，浓香型和清香型白酒干燥物红外光谱差异明显，酯化物在浓香型白酒干燥物中含量较高，而羧酸盐和醇类物质在清香型白酒干燥物中含量较高。这种方法可以用于直接分析图谱中的

收稿日期：2020-08-27，**修订日期**：2021-01-06

基金项目：国家重点研发计划研究项目(2019YFC1606502)资助

作者简介：祝海江，1971 年生，北京化工大学信息科学与技术学院教授 e-mail：zhuhj@mail.buct.edu.cn

物质含量，并可以有效地分析和确定酒的品质和真假。有研究强调，相对于原始质谱的峰强度，原始质谱的数学转换产生了一个与质谱和分子结构密切相关的新质谱特性。

目前，国内外针对酒的品质鉴别方法有越来越多的学者深入研究。近年来，许多相关的研究都是基于不同的实验数据和分类模式。李建[3]等在碱性加热前提下，根据纯粮白酒的吸光度值在 363 nm 波长处不同的原理，构建白酒标准曲线图来鉴别样品中纯粮白酒的百分比。结果显示 4 个样品的精度在 90%以上，相对标准差在 1%以下；有报道则在三维荧光光谱中，采用主成分分析(principal component analysis，PCA)进行降维处理，使用支持向量机(support vector machine，SVM)算法，通过 k-fold 交叉验证方法发现 SVM 的最佳参数 c 和 γ，建立了高准确率的不同酒品牌的分类模型；姜安[4]等使用多项式插值拟合等方法，将采集的白酒红外光谱数据进行预处理，依据年份、味道等特征构建 SVM 支持向量机分类模型，结果表明该方法较为快速准确。

国外就酒品质的频谱分析多针对葡萄酒或啤酒数据展开研究。Cozzolino 等[5]在白葡萄酒的可见近红外光谱实验中，使用 PCA 主成分分析等方法实现了葡萄酒品种的分类。但是，参与实验的红酒品种数量相对有限，所以在实际应用上必须慎重。Pontes[6]等采用主成分分析法提取白兰地、朗姆酒、其他酒精饮料与其掺假样本的近红外光谱特征，进行酒的分类和验证，现已应用于鉴别假酒，预测准确率可达 100%(95%置信区间)。

近些年，离子迁移谱也被用于分析白酒的风味与品质。朱玲[7]等采用气相-离子迁移谱分析了白酒挥发性风味物质，并通过构建白酒香型风味指纹图谱，实现了三种不同香型白酒分类。李娟[8]等将气相色谱和离子迁移谱相结合，分析了白酒谱图中风味物质的出峰信号，并对 11 种不同香型白酒进行了分类。张志刚[9]等研究了利用离子迁移谱快速检测白酒和红酒中的塑化剂含量。

针对白酒品质鉴定问题，本研究使用真酒和添加不同乙醇浓度的假酒，从中得到样本的离子迁移谱数据，通过频谱分析的方式，结合常用的一维信号特征提取方法，从多角度提取白酒信号特征并作筛选。根据不同的需求，构建二分类和多分类白酒品质分类模型[SVM、K 最邻近(K-nearest neighbor，KNN)分类、LR 逻辑回归(logistic regression analysis，LRM)分类、模糊 C 均值(fuzzy C-means，FCM)算法等]，通过计算多个评价指标给出白酒分类最佳模型。

1　实验部分

1.1　数据获取

使用样本离子迁移谱信号由美国 Excellims 公司 GA2100 型电喷雾-离子迁移谱仪(ESI-IMS)采集获得。Excellims 离子迁移谱，具有快速、准确、高分辨率和高灵敏度的特点。GA2100 离子迁移谱仪具有高分辨率，可以达到 70～120 s，分析时间仅为几秒。

文中白酒样本数据来自于国内某一白酒厂，共 6 类样本，每一类含有若干份样本，其中一类样本为真酒，其余五类分别是添加 10%，20%，30%，40%和 50%酒精浓度的白酒。这六类样本的离子迁移谱如图 1 所示，图中仅显示了部分时间段内离子迁移谱。离子迁移谱中横坐标表示离子的迁移时间，纵坐标表示峰的强度。因为不同品质白酒的离子成分不同，故不同种类的样本含有不同的特征谱峰。由图 1 所示样本的离子迁移谱可知，不同品质白酒谱的形状相似，主要区别在于峰强度有略微的差别，难于直接判断出其白酒品质。因此，可以利用特征提取和机器学习的方法将不同离子迁移谱进行识别和分类。

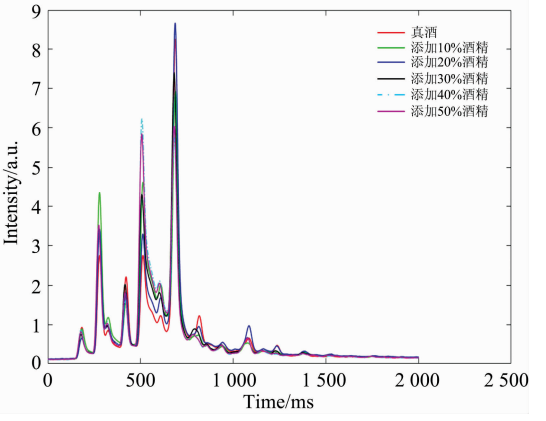

图 1　真酒和添加不同比例酒精的白酒离子迁移谱

Fig. 1　The ion mobility spectra of real liquor and liquor with different proportion of alcohol

1.2　特征提取和选择

通过对白酒离子迁移谱的时域谱峰、频域谱峰、谱熵和过零率等特征进行提取，构建特征向量；分别使用主成分分析(PCA)和线性判别分析(linear discriminate analysis，LDA)进行特征选择。

时域谱峰即离子迁移谱的时域信号不同时刻的峰值。而频域谱峰是根据傅里叶变换对原始离子迁移谱进行处理后，计算出其在不同频率点上的峰值。假设采样频率为 Fs，采样点数为 N，傅里叶变换处理后的结果就是点数为 N 的复数，每一个点就对应着一个频率，而每个点的模值，就是该频率值下的幅度特性，即就是频域谱峰。某一点 n 表示的频率为

$$F_n = \frac{(n-1)\times F_s}{N} \tag{1}$$

由式(1)可以看出，F_n 所能分辨到的最小频率为 $\frac{F_s}{N}$，该点模值 mag(i)乘以 $\frac{2}{N}$，就是对应该频率下信号的频域谱峰 A

$$A = \mathrm{mag}(i)\times\frac{2}{N} \tag{2}$$

在信息论中，熵是信息无序程度的一种度量，也是信息有用程度的一种表现形式。熵越小，不确定性越小，而概率的差别越大，熵就越小。因此，熵可以描述各类别信号的可区分性。谱熵度量了信号的频率分布均匀程度，体现了信号能量分布的频域复杂性。对于一个信号 $x(n)$，功率谱 $S(m)$ 为

$$S(m) = |x(m)|^2 \tag{3}$$

式(3)中，$x(m)$是$x(n)$的离散傅里叶变换。概率分布$P(m)$为

$$P(m) = \frac{S(m)}{\sum_i S(i)} \tag{4}$$

则谱熵H为

$$H = -\sum_{m=1}^{N} P(m)\log_2^{p(m)} \tag{5}$$

正则化后

$$H_n = -\frac{\sum_{m=1}^{N} p(m)\log_2 P(m)}{\log_2 N} \tag{6}$$

其中N是总频率点。$\log_2 N$表示白噪声的最大谱熵，在频域内均匀分布。若已知时频功率谱图$S(t, f)$，则概率分布为

$$P(m) = \frac{\sum_t S(t,m)}{\sum_f \sum_t S(t,f)} \tag{7}$$

为计算给定时频功率谱图$S(t, f)$的瞬时谱熵，t时刻的概率分布为

$$P(t,m) = \frac{S(t,m)}{\sum_f S(t,f)} \tag{8}$$

则t时刻的谱熵为

$$H(t) = -\sum_{m=1}^{N} P(t,m)\log_2 P(t,m) \tag{9}$$

过零率（zero-crossing rate，ZCR)是单位时间内波形通过零点的次数，通常指一个信号的符号变化的比率，例如信号从正数变成负数或反向。在离散时间信号情况下，如果相邻的采样具有不同的代数符号就称为发生了过零，因此可以计算过零的次数。过零率在一定程度上可以反映信号的频率信息。其计算方法如下：首先计算信号绝对值$|x|$，求取信号均值后使其均值变为0，每偏移1个单位计算该状态下信号是否过零点，最后输出过零率F_0。

综上所述，将要提取的时域峰值特征记为U，频域峰值特征记为O、谱熵特征记为P，过零率特征记为Q，则所有的特征合在一起记为$C=\{U, O, P, Q\}$，将矩阵C作为白酒品质分类的特征矩阵。

提取完特征向量之后，由于特征的维数较大，部分特征是冗余特征，对分类没有帮助。因此，需要对提取的特征进行降维。采用主成分分析法(PCA)和线性判别分析法(LDA)分别对获得的特征进行降维。PCA是为了去除原始数据集中冗余的维度，让投影子空间的各个维度的方差尽可能大，也就是熵尽可能大。LDA通过数据降维找到那些具有差异性的维度，使得原始数据在这些维度上的投影，不同类别尽可能区分开来。

1.2 分类方法

支持向量机(SVM)是一种在分类与回归分析中常用的监督式学习分类算法。最初用于二分类，也可以通过组合多个二分类器来实现多分类。分类任务为二分类(真酒和添加酒精的白酒分类)和六分类(真酒和添加不同酒精成分的白酒分类)。因此，选择支持向量机作为分类器实现多个类别分类。

选择核函数为径向基函数(radial basis function，RBF)的SVM分类器，在SVM中gamma和C(惩罚系数)是需要人为给定的两个超参数，参数gamma表明单个训练样本的影响大小，值越小影响越大，值越大影响越小；而参数C的值低时使得分界面平滑，而高的C值通过增加模型自由度以选择更多支持向量来确保所有样本都被正确分类。为了确定这两个超参数的最优值，使用网格搜索方法寻找使得模型分类效果最优时两个参数的值。

工作中比较了SVM分类器和逻辑回归分类器(LRM)、模糊C均值分类器(FCM)、K最邻近分类器(KNN)等多种分类算法对样本数据分类的准确率。

2 结果与讨论

2.1 特征提取结果

由于样本数据在采样点1 050到采样点2 500峰值有明显的变化，故保留此时间段的数据。在时域谱峰特征提取中，选择每类样本数据的7个峰值点TDP1(time domain peak)，TDP2，TDP3，TDP4，TDP5，TDP6和TDP7，作为时域谱峰的峰值，如图2所示。

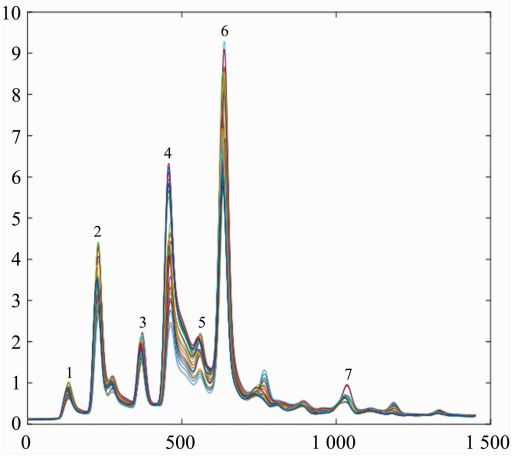

图2 时域特征峰值示意图

Fig. 2 The time domain characteristic peak

由图2可以看到在某些峰上不同类别的数据差别很小，这些峰值不是很好的特征，需要将其剔除。同时在一个类别样本数据内部，希望不同的峰值数据能够集中，以体现样本内数据的稳定性。因此，计算同一峰值下同一类别的样本内方差和不同类别之间同一峰的样本间方差，经过计算发现第1和3峰的各类样本内的方差较大，说明处于第1和3峰的样本内数据不集中；而在第1和2峰值下，各类样本的方差过小，区别不明显。故剔除了第1，2和3峰值，保留TDP4，TDP5，TDP6和TDP7这4个时域谱峰值，即$N\times 4$个特征点，形成时域谱峰特征集合U。

同理，在提取样本离子迁移谱的频域谱峰值时，先对时域的离子迁移谱数据进行快速傅里叶变换，得到其频域响应曲线，如图3所示。根据各个样本的频域响应信号，剔除样

本间方差较小和样本内方差较大的峰值点，最后每个样本信号保留 FDP1(frequency domain peak)，FDP2，FDP3 这 3 个频域峰值，即 $N\times3$ 个特征点，形成频域谱峰特征集合 O。

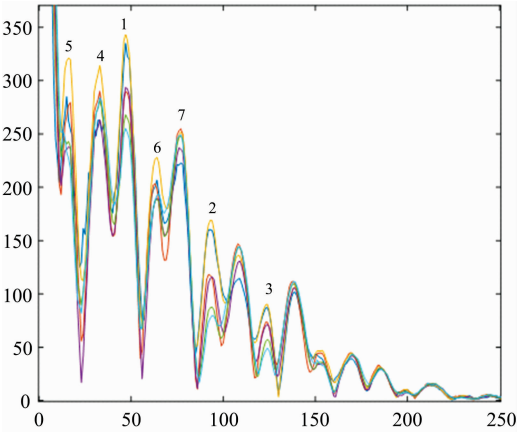

图 3　离子迁移谱的频域响应曲线图

Fig. 3　The frequency domain response curve of ion mobility spectrum

所有样本离子迁移谱的谱熵计算得到后，形成了该类样本的谱熵序列向量，其谱熵-时间关系如图 4 所示。由图 4 可看出，真酒在第 17 ms 左右的谱熵[图 4(a)中黑色虚线椭圆标记]区别于其他添加不同浓度酒精的白酒。真酒在此时的谱熵范围为[0.532 97，0.536 25]，添加酒精后的白酒样本谱熵范围在[0.476 059 584，0.464 387 249]，见图 4(b—f)，且随着添加酒精浓度的提升，谱熵在 15 ms 左右时，区间范围从高逐渐降低。该谱熵特征较为明显，记第 17 ms 的数据为特征矩阵 P。同理，六类样本的过零率特征计算后记为特征矩阵 Q。

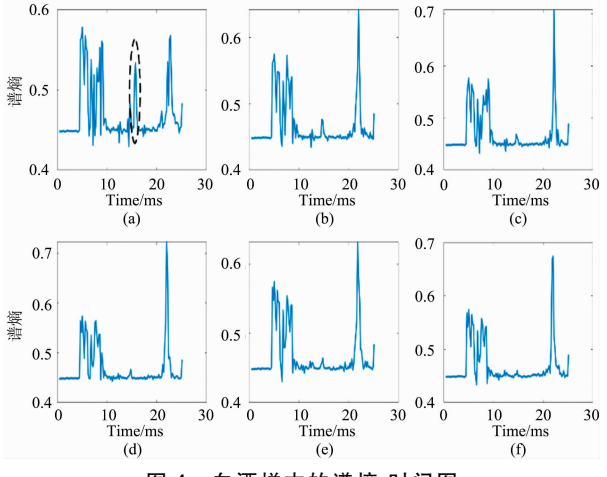

图 4　白酒样本的谱熵-时间图

(a)：真酒；(b)：添加 10%酒精的白酒；(c)：添加 20%酒精的白酒；(d)：添加 30%酒精的白酒；(e)：添加 40%酒精的白酒；(f)：添加 50%酒精的白酒

Fig. 4　The spectral entropy time diagram of liquor samples

(a)：Real liquor；(b)：Liquor added with 10% alcohol；(c)：Liquor added with 20% alcohol；(d)：Liquor added with 30% alcohol；(e)：Liquor added with 40% alcohol；(f)：Liquor added with 50% alcohol

综上所述，已提取的时域峰值特征、频域峰值特征、谱熵特征和过零率特征，可以记为 $C=\{U, O, P, Q\}$，将矩阵 C 作为白酒品质分类的特征矩阵。

在特征选择实验中，利用 PCA 和 LDA 对特征矩阵 C 进行了降维实验。在使用 PCA 对特征矩阵 C 进行降维时，将特征矩阵降为 5 维的特征矩阵，并统计其每个维度的特征对整体特征的贡献率，如图 5 所示。由图 5 可以看出，前三维的特征对整体特征的累计贡献率达到了 95%。因此取前三维特征作为特征矩阵。

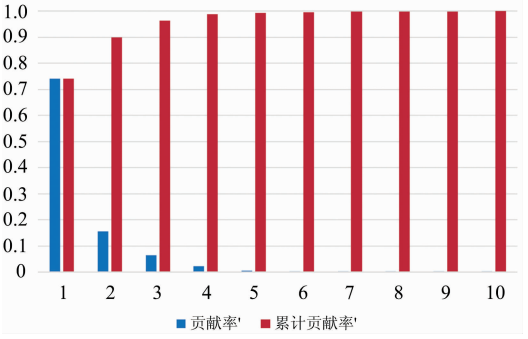

图 5　使用 PCA 降维后特征的贡献率示意图

Fig. 5　The contribution rate of features after dimension reduction using PCA

同理，使用 LDA 对特征矩阵 C 进行降维，并统计每个维度的特征对总体特征的贡献率，如图 6 所示。由图 6 可以看出，前两维特征向量对整体特征的累计贡献率达到了 95%。因此，取前两维特征向量作为特征矩阵。相比 PCA 而言，使用 LDA 降维后，各类样本的特征较为明显，同时特征维度更低。因此，选用 LDA 作为特征降维方法。

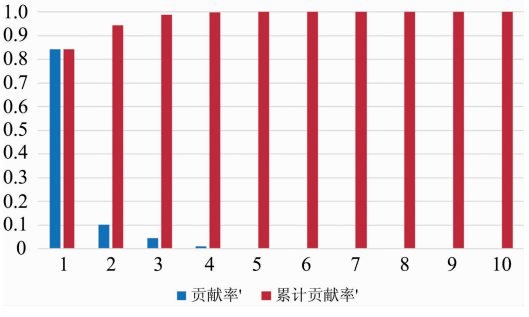

图 6　使用 LDA 降维后特征的贡献率示意图

Fig. 6　The contribution rate of features after dimension reduction using LDA

2.2　分类实验结果及分析

首先，利用 SVM 进行了二分类和六分类实验。在 SVM 二分类实验中，两个参数 C 和 gamma 的寻优结果如图 7，图 8 所示。可以看出，当 C 为 0.435 28，gamma 为 1 时，SVM 的二分类准确率最高，达到了 100%。

而在 SVM 六分类实验中，两个参数 C 和 gamma 的寻优结果如图 9，图 10 所示。可以看出，六分类时最优参数 C 和 gamma 分别为 4 和 11.313 7。此时，SVM 的分类准确率最

高，达到了 99.7%。

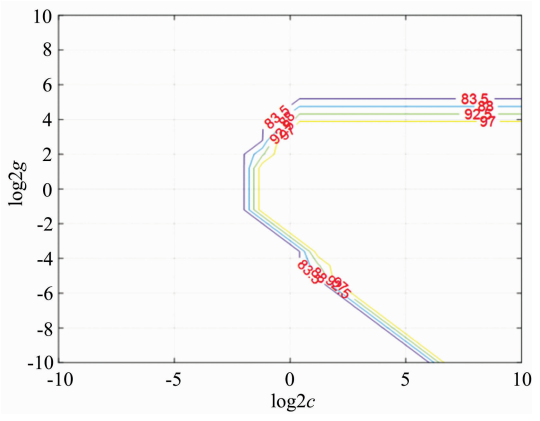

图 7 SVM 二分类实验中参数 *C* 寻优结果示意图

Fig. 7 The optimization results of parameter *C* in SVM binary classification experiment

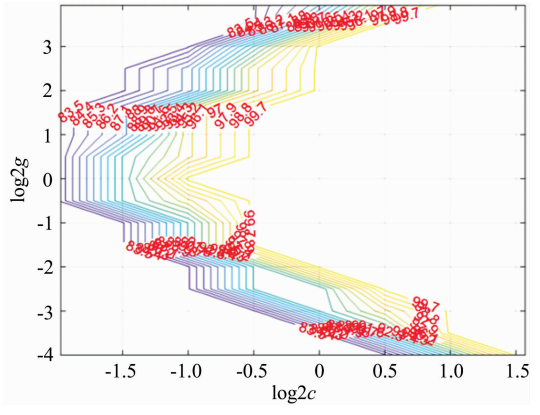

图 8 SVM 二分类实验中参数 gamma 寻优结果示意图

Fig. 8 The optimization results of parameter gamma in SVM binary classification experiment

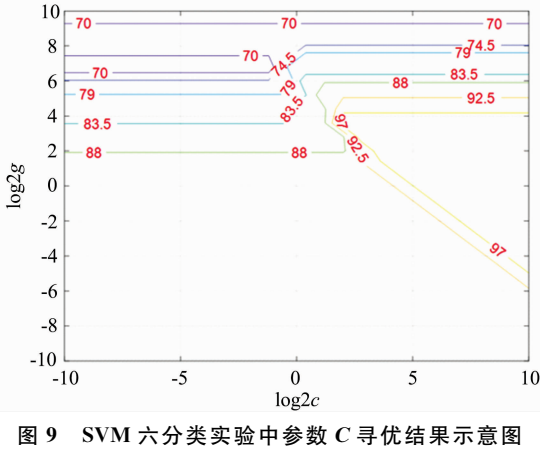

图 9 SVM 六分类实验中参数 *C* 寻优结果示意图

Fig. 9 The optimization results of parameter *C* in SVM six classification experiment

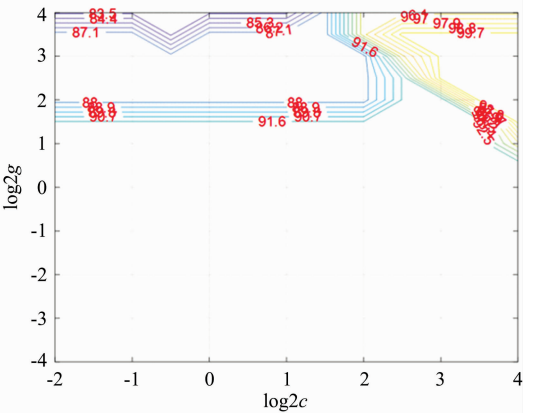

图 10 SVM 六分类实验中参数 gamma 寻优结果示意图

Fig. 10 The optimization results of parameter gamma in SVM six classification experiment

然后，选择逻辑回归(LRM)分类、模糊 C 均值分类(FCM)和 K 近邻分类(KNN)对样本数据进行了二分类和六分类实验，分类准确率如表 1 所示。

表 1 分类方法实验结果比较

Table 1 The comparison of experimental results of classification methods

分类器	二分类准确率/%	六分类准确率/%
逻辑回归(LRM)分类	100	33.33
模糊 C 均值分类(FCM)	100	80.66
K 近邻分类(KNN)	100	91.70
支持向量机(SVM)	100	99.70

由表 1 可以看出，这四种分类方法在对样本数据进行二分类时，分类准确率都可以到达 100%；在进行六分类时，LRM 的分类准确率仅有 33.33%。利用 FCM 进行六分类时，总体每个类别的聚类中心都比较分散，但是类别 4 和类别 5 的聚类中心非常接近，这就会导致在分类上的错误，在六分类任务上表现欠佳，准确率为 80.66%。与 FCM 相似，KNN 也是聚类算法，在实验中 K 取 2 时，分类准确率最高：二分类准确率达 100%，六分类准确率为 91.7%。而 SVM 的六分类准确率最高，达到了 99.7%。

为了进一步评价分类器的性能，实验中对多分类的分类器性能进行了评价。引入 Macro-F1 和 Micro-F1 作为评价指标。指标 Macro F1 由 F1-score 描述得来，首先计算出所有类别总体的准确率 P 和召回率 R，然后计算 F1-score，其计算式为

$$F1 = \frac{2P \times R}{P + R} \tag{10}$$

其中准确率 $P = \frac{TP}{TP + FP}$，召回率 $R = \frac{TP}{TP + FN}$，这里 TP 指“预测为正样本，实际为正样本”，FP 指“预测为正样本，实际为负样本”，FN 指“预测为负样本，实际为正样本”。而指标 Micro-F1 是先计算出每一个类的准确率 P 和召回率 R，然后再计算 F1，最后计算所有 F1 平均值。

四种分类器两个指标 Macro-F1 和 Micro-F1 的计算结果如表 2 所示，综合考虑这两个性能指标结果，可以看出 SVM 具有最佳的分类性能。

表 2　四种分类器性能指标比较结果

Table 2　The performance comparison of four classifiers

分类器	Macro F1	Micro F1
LRM	15.87%	33.33%
FCM	34.14%	80.66%
KNN	26.98%	83.33%
SVM	28.57%	100%

表 3　四种分类器运行时间比较结果

Table 3　The comparison of running time of four classifiers

分类器	计算时间/s
LRM	0.784
FCM	0.558
KNN	0.221
SVM	2.834

研究中对不同分类器运行时间进行了计算，结果如表 3 所示，由表中可以看出，运行时间上 SVM 耗时最长。虽然其他分类器运算速度快，但是六分类准确率确远不如 SVM。因此，在二分类和多分类任务上，SVM 的表现是最出色，同时也说明了本研究对白酒样本提取的特征具有代表性。

3　结　论

利用美国 Excellims 公司 GA2100 型电喷雾-离子迁移谱仪(ESI-IMS)获得白酒样本的离子迁移谱数据，并将离子迁移谱数据的时域特征谱峰，频域特征谱峰，谱熵和过零率作为白酒样本的特征数据。采用主成分分析法(PCA)和线性判别法(LDA)分别对特征数据进行降维研究，通过比较特征贡献率，发现线性判别法有更好的特征提取能力(通过 LDA 降维的前两维特征数据的特征累计贡献率达到了 95%)。通过 SVM 分类器对降维后特征数据分别进行二分类和六分类训练，准确率分别达到了 100%和 99.7%。因此，白酒离子迁移谱数据的时频谱特征结合线性判别法和支持向量机可以作为白酒品质鉴别的一种新的检测方法。

References

[1] Yu H Y, Ying B, Sun T, et al. Journal of Food Science, 2007, 72(3): E125.

[2] LÜ Hai-tang, REN Yan-rong, LI Chun-hua(吕海棠, 任彦蓉, 李春花). China Brewing(中国酿造), 2010, (10): 175.

[3] LI Jian, JIANG Xue(李　建, 姜　雪). China Brewing(中国酿造), 2015, 34(1): 118.

[4] JIANG An, PENG Jiang-tao, PENG Si-long, et al(姜　安, 彭江涛, 彭思龙, 等). Computers and Applied Chemistry(计算机与应用化学), 2010, 27(2): 233.

[5] Cozzolino D, Smyth H E, Gishen M. Journal of Agricultural and Food Chemistry, 2003, 51(26): 7703.

[6] Pontes M J C, Santos S R B, Araujo M C U, et al. Food Research International, 2006, 39(2): 182.

[7] ZHU Ling, CAI Jin-zhong, LIU Ben(朱　玲, 蔡尽忠, 刘　奔). Guangdong Chemical Industry(广东化工), 2020, (5): 22.

[8] LI Juan, GUO Ya-jie, ZHAO Wei-jun, et al(李　娟, 郭亚洁, 赵伟军, 等). Journal of Food Safety & Quality(食品安全质量检测学报), 2014, 5(6): 1687.

[9] ZHANG Zhi-gang, LIN Li-yi, HAO Yu-lei, et al(张志刚, 林立毅, 郝玉蕾, 等). Journal of Inspection and Quarantine(检验检疫学刊), 2016, 26(3): 18.

Classification Method of Liquor Quality Based on Time and Frequency Spectrum Characteristics

ZHU Hai-jiang[1], TANG Hao[1], SUN Jing-xian[1], DU Zhen-xia[2]

1. College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China
2. College of Chemistry, Beijing University of Chemical Technology, Beijing 100029, China

Abstract In order to establish a fast and accurate method for liquor quality identification, this paper uses the machine learning method to model liquor of different quality. To extract the characteristics of different quality liquor, we analyzed different quality liquor with the ion mobility spectroscopy, constructed the feature vectors based on the signal of ion mobility spectroscopy, and classified different quality liquor. The ion mobility spectroscopy signals of liquor samples were obtained using the Excellims GA2100 electrospray ionization mobility spectrometry (ESI-IMS). Each ion mobility spectroscopy signal is a time series signal with its intensity varying with time. In the aspect of feature extraction, the time-domain characteristic peaks of the original data were extracted. Fourier transform is performed on the data of ion mobility spectroscopyfor more comprehensive characteristics, and the characteristic peaks in the frequency domain were extracted. At the same time, in order to express the characteristics of signal change, the spectral entropy and zero-crossing rate of ion mobility spectroscopy were calculated, and the $N \times 9$ dimensional feature matrix was constructed; Then, principal component analysis (PCA) and linear discriminant analysis (LDA) were used to reduce the dimensions of the features. The cumulative contribution rate of the first three-dimensional features of PCA to the overall features is 95%. By contrast, the cumulative contribution rate of the first two-dimensional features of LDA is 95%. Therefore, LDA is chosen as the feature dimension reduction method; Finally, a support vector machine (SVM), a nonlinear classifier in machine learning, was used to classify liquor ion mobility spectrum data. The experimental results show that the correct classification rate of the SVM method is 100% in the classification of real liquor and liquor with added alcohol; The correct classification rate of the SVM method is 99.7% in the six classifications of real liquor and five kinds of fake liquor with 10%, 20%, 30%, 40% and 50% alcohol concentration respectively. In addition, this paper compared the results of classification of ion mobility spectroscopy of liquor samples by logistic regression analysis (LRM), fuzzy C-means (FCM) and *k*-nearest neighbor (KNN). The results show that the SVM method based on spectrum feature vector can accurately distinguish the real liquor and the liquor with added alcohol, which provides a new detection method for identifying liquor quality.

Keywords Ion mobility spectroscopy; Chinese liquor quality; Support vector machine (SVM); Time and frequency feature

(Received Aug. 27, 2020; accepted Jan. 6, 2021)