机器学习算法在不同形态浙贝母与湖北贝母的干法 REIMS 指纹图谱鉴别分析中的应用研究

石岩[1],李宁[2],魏锋[1*]

(1. 中国食品药品检定研究院,北京 102629;2. 北京市药品检验研究院,北京 102206)

摘要 **目的**:使用快速蒸发离子化质谱(REIMS)指纹图谱与机器学习相关技术对不同形态的浙贝母和湖北贝母进行预测和判别。**方法**:通过干法灼烧使样品组分形成气溶胶,引入 REIMS 中,质谱扫描范围 *m/z* 50 ~ 1 200,扫描模式为灵敏模式,扫描时间为 0.2 s。正离子模式采集,数据记录为 continuum 模式,测得样品的 REIMS 指纹图谱数据。通过对数据进行聚类分析、相关性分析、相似度分析、主成分分析,得到数据分布的基本情况,最后建立逻辑回归模型,模型惩罚项参数选择岭回归(l2),优化算法选择拟牛顿法(lbfgs)。**结果**:测得样品的 REIMS 指纹图谱具有品种差异的特征性,逻辑回归模型交叉验证和测试集验证准确率均达到 1.0,可以准确预测和判别样品的品种。**结论**:REIMS 技术结合机器学习在中药领域的潜在应用前景十分广阔。

关键词:浙贝母;湖北贝母;快速蒸发离子化质谱;机器学习;人工智能;逻辑回归;中药分析;真伪鉴别

中图分类号: R 917 **文献标识码**: A **文章编号**: 0254 - 1793(2024)01 - 0134 - 10

doi: 10.16155/j.0254 - 1793.2024.01.14

Research of machine learning in the application of authenticity discrimination of Fritillariae Thunbergii Bulbus and Fritillariae Hupehensis Bulbus in different form with dry - process REIMS fingerpring

SHI Yan[1], LI Ning[2], WEI Feng[1*]

(1. National Institutes for Food and Drug Control, Beijing 102629, China; 2. Beijing Institute for Drug Control, Beijing 102206, China)

Abstract **Objective**: To study and analyze rapid evaporative ionization mass spectrometry (REIMS) fingerprints of samples of Fritillariae Thunbergii Bulbus and Fritillariae Hupehensis Bulbus in different forms for authenticity discrimination with machine learning. **Methods**: Aerosol formations from the samples by high temperature of dry burning method were ionized and determined by REIMS with *m/z* 50 - 1 200 as scanning range in sensitive mode and positive ion mode. The scanning time was 0.2 s and data was recorded as continuous mode. Then the basic situation of REIMS data distribution was studied and analyzed through the methods of cluster analysis, correlation analysis, similarity analysis and principal component analysis. And then logistic regression model with ridge regression (l2) as penalty parameter and quasi - Newton method (lbfgs) as optimization algorithm was established.

* 通信作者 Tel:(010)53852020;E - mail:weifeng@nifdc.org.cn

第一作者 石 岩 Tel:(010)53852081;E - mail:shiyan@nifdc.org.cn

李 宁 Tel:13811671528;E - mail:642781540@qq.com

Results: The REIMS fingerprints of the samples showed the characteristics of variety differences. Both cross validation and test set validation had an accuracy of 1.0, and the logistic regression model could accurately predict and distinguish the varieties of the samples. **Conclusion**: The application prospect of REIMS technique combined with machine learning in the field of traditional Chinese medicine is very broad.

Keywords: Fritillariae Thunbergii Bulbus; Fritillariae Hupehensis Bulbus; REIMS; machine learning; artificial intelligence; logistic regression; analysis of traditional Chinese medicine; authenticity discrimination

百合科贝母属植物有上百种之多,其中有众多的以鳞茎入药的重要药用植物,大多具有清热润肺和止咳祛痰的功效[1]。贝母属药用植物在我国有悠久的药用历史,其药用始载于《神农本草经》;在我国最早的一部诗歌文学总集《诗经》中收载的《卫风·载驰》,其中就有"陟彼阿丘,言采其蝱"的诗句,"蝱"即为贝母,可见贝母属药用植物自古就与我国劳动人民有着紧密的联系[2]。历代本草专著所记载的贝母品种来源虽较为混乱,但都是以贝母属植物为主流,在清代《本草纲目拾遗》一书中将贝母按照功效分为2个大类,即川贝母和浙贝母,这正与临床用药中将贝母分为"川贝"和"浙贝"2个类群的实际情况相契合[1-4]。"川贝"类群是以川贝母为主,而"浙贝"类群则以浙贝母为主,包括了产自长江流域的湖北贝母和东阳贝母等[1]。浙贝母药材是指浙贝母植物的干燥鳞茎,与川贝母功效类似,具有清热化痰止咳和解毒散结消痈的功效[5]。由于价格因素,一些与浙贝母性状相似的贝母类药材在市场上有混为正品浙贝母销售和使用的情况,混伪品多见湖北贝母[6-7]。

为规范中药临床用药,保障人民群众用药的安全有效,有必要对浙贝母与湖北贝母的特征进行鉴别研究。由于浙贝母与湖北贝母化学成分相近,目前对于二者的鉴别研究并不多,其理化特征鉴别的实验过程多经过色谱分离或较复杂的样品处理,且数据处理和分析方法识别准确度欠缺[8-9]。快速蒸发离子化质谱(rapid evaporative ionization mass spectrometry,REIMS)是一种手持式的、实时原位质谱分析技术,依靠将生物组织快速蒸发所产生的气溶胶引入质谱仪进行样品的检测[10-14]。目前REIMS的应用领域基本集中在手术医疗以及肉类食品检验等范围[15-19],在中药,尤其是中药材及中药饮片的分析领域中几乎为空白,这与其手持式离子化设备iknife通过样品导电产热的原理有关。本研究在前期[20]创新地使用可调恒温电烙铁解决了干燥白头翁及其伪品药材表面缺乏导电性的问题的基础上,开展了浙贝母及其伪品湖北贝母的REIMS指纹图谱的测定,并且根据所获得的数据采用机器学习的相关算法进行了数据处理和机器识别研究。考虑到实际应用情况,为了增加本研究及其所建立模型的泛化能力,在样品数据采集时,兼顾了浙贝母和湖北贝母的较完整鳞茎和粉末状2种常见的形态(见图1),在机器识别模型时,训练集和测试集的数据也都包括了较完整鳞茎和粉末状的测定数据。最终所建立的模型对浙贝母和湖北贝母可以达到完全准确的识别,对浙贝母和湖北贝母的鉴别提供了一个可靠的方法,对于贝母类药材的真伪鉴别研究也有一定的参考意义。

浙贝母 (Fritillariae Thunbergii Bulbus)　　湖北贝母 (Fritillariae Hupehensis Bulbus)

图1　样品形态

Fig. 1　Sample form

1　材料

1.1　仪器

SYNAPT G2-S四极杆飞行时间质谱仪(Waters公司);REIMS离子源(Waters公司);Pump 11 Elite注射泵(Harvard Apparatus公司);WSD71可调温恒温数显焊台(Weller公司)。

1.2　试药

亮氨酸脑啡肽(Sigma-Aldrich公司,批号W13012301,质谱级纯度)。

研究样品共144批,浙贝母为74批(编号为ZBM_

01 ~ ZBM_74),其中浙贝母较完整鳞茎样品 35 批(ZBM_01 ~ ZBM_35),浙贝母粉末样品 39 批(ZBM_36 ~ ZBM_74);湖北贝母为 70 批(编号为 HBBM_01 ~ HBBM_70),其中湖北贝母较完整鳞茎样品 38 批(HBBM_01 ~ HBBM_38),湖北贝母粉末样品 32 批(HBBM_39 ~ HBBM_70)。样品详细信息见表 1。

表 1 样品信息表

Tab. 1 Information of samples

编号(code)	收集地(collection place)	外观形态(form)	品种(variety)
ZBM_01 ~ ZBM_16	河北(Hebei)	较完整鳞茎(complete bulb)	浙贝母(Fritillariae Thunbergii Bulbus)
ZBM_17 ~ ZBM_22	安徽(Anhui)		
ZBM_23 ~ ZBM_35	浙江(Zhejiang)		
ZBM_36 ~ ZBM_45	河北(Hebei)	粉末(powder)	
ZBM_46 ~ ZBM_47	安徽(Anhui)		
ZBM_48 ~ ZBM_52	–		
ZBM_53 ~ ZBM_74	浙江(Zhejiang)		
HBBM_01 ~ HBBM_11	湖北(Hubei)	较完整鳞茎(complete bulb)	湖北贝母(Fritillariae Hupehensis Bulbus)
HBBM_12 ~ HBBM_18	河北(Hebei)		
HBBM_19 ~ HBBM_38	安徽(Anhui)		
HBBM_39 ~ HBBM_44	湖北(Hubei)	粉末(powder)	
HBBM_45 ~ HBBM_52	河北(Hebei)		
HBBM_53 ~ HBBM_70	安徽(Anhui)		

1.3 编程语言

Python 计算机编程语言(Python Software Foundation, version:3.8.8)。

2 方法与结果

2.1 质谱数据采集

2.1.1 质谱条件 100 ng·mL^{-1}亮氨酸脑啡肽异丙醇溶液,流速 100 μL·min^{-1}。焊台烙铁头温度 450 ℃;质谱扫描范围 *m/z* 50 ~ 1 200;扫描模式为灵敏模式;扫描时间为 0.2 s。正离子模式采集,数据记录为 continuum 模式。

2.1.2 样品处理及测定 对于较完整鳞茎的供试样品,使用焊台烙铁头在供试样品表面进行干法烧灼;对于粉末形态的供试样品,取适量供试样品粉末于铝箔上,使用焊台烙铁头干法烧灼。干法烧灼生成的气溶胶使用聚四氟乙烯管引入 REIMS离子源,进行质谱扫描检测,每批样品测定 10 次,烧灼时间约为 2 s。样品的 REIMS 质谱图见图 2。

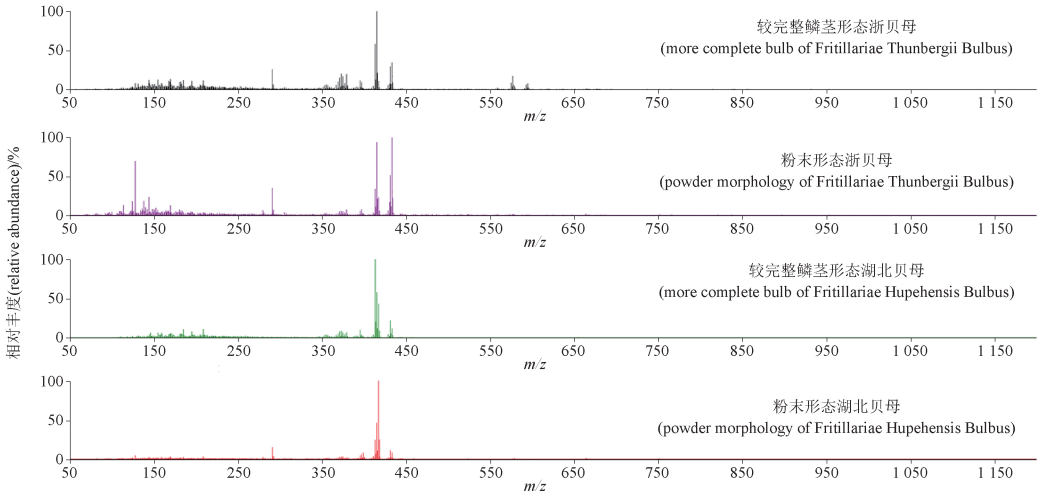

图 2 样品的 REIMS 质谱图

Fig. 2 REIMS spectra of samples

2.2 数据处理与分析

2.2.1 数据处理 质谱扫描范围 *m/z* 50 ~ 1 200，样品总批数为 144 批，数据最终构成 144 × 334 的向量矩阵，经过标准化处理后导出 csv 格式备用。

2.2.2 聚类分析 使用 Ward 法根据各样品 REIMS 指纹图谱间的欧氏距离进行系统聚类分析，聚类分析树状图结果见图 3。

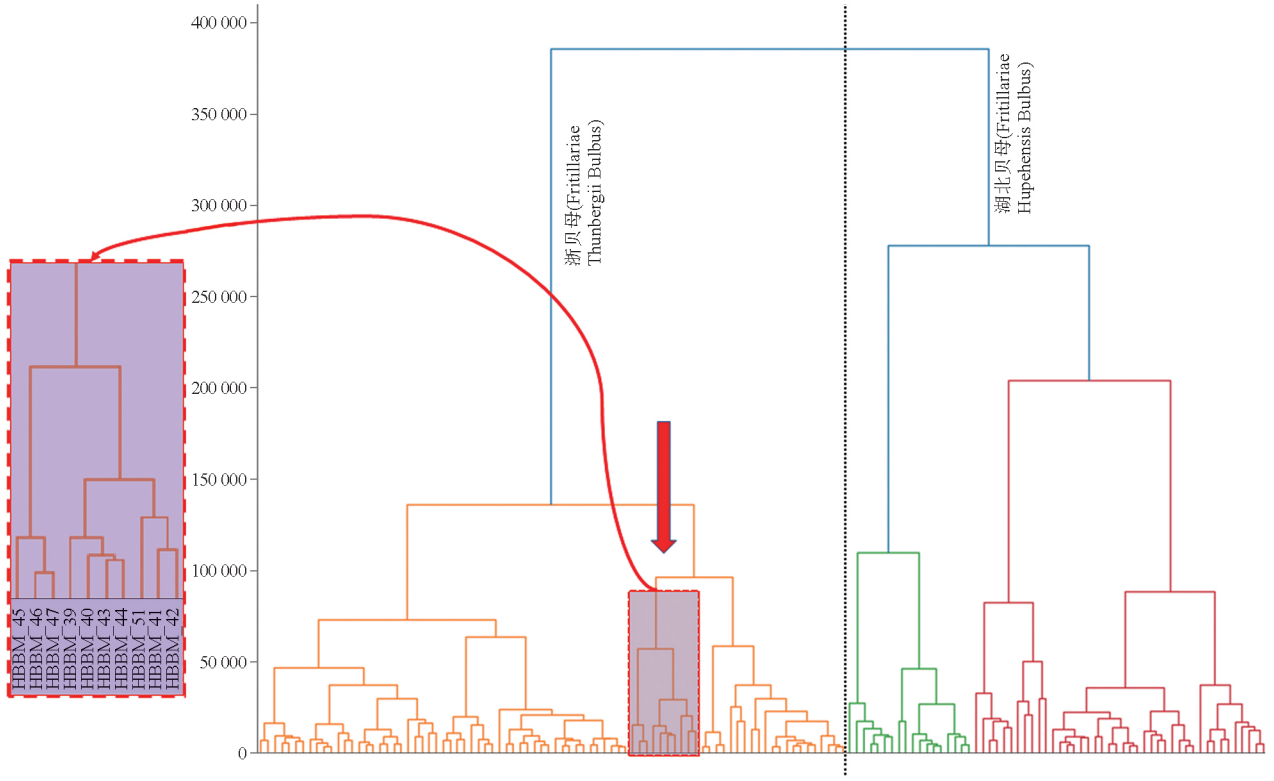

图 3 聚类分析树状图

Fig. 3 Cluster analysis dendrogram

2.2.3 Pearson 相关性分析 对样品的 REIMS 指纹图谱数据按照样品之间的关系进行基于 Pearson 法的相关性分析，按照样品之间 REIMS 指纹图谱相关性强弱绘制矩阵图，颜色越深代表相关性越强，见图 4。

2.2.4 相似度分析 将浙贝母和湖北贝母的 REIMS 指纹图谱数据分别按照样品形态各分为 2 类，可得到样品 4 个分类，即编号为 ZBM_01 ~ ZBM_35 的浙贝母较完整鳞茎样品；编号为 ZBM_36 ~ ZBM_74 的浙贝母粉末样品；编号为 HBBM_01 ~ HBBM_38 的湖北贝母较完整鳞茎样品；编号为 HBBM_39 ~ HBBM_70 的湖北贝母粉末样品。将这 4 个分类样品的 REIMS 指纹图谱数据分别取平均值，可得到代表 4 个类别样品的 REIMS 平均指纹图谱数据，按照夹角余弦法分别计算所有样品与上述 4 个 REIMS 平均指纹图谱的相似度，然后按照样品品种的不同绘制相似度散点分布图及核密度图，即图 5。在图 5 中，A 为样品分别对于浙贝母样品的相似度散点分布图，其中横坐标为样品对于浙贝母较完整鳞茎样品平均指纹图谱的相似度，纵坐标为样品对于浙贝母粉末样品平均指纹图谱的相似度；B 为 A 散点分布图的核密度图；C 为样品分别对于湖北贝母样品的相似度散点分布图，其中横坐标为样品对于湖北贝母较完整鳞茎样品平均指纹图谱的相似度，纵坐标为样品对于湖北贝母粉末样品平均指纹图谱的相似度；D 为 C 散点分布图的核密度图。

2.2.5 主成分分析 将各批样品的 REIMS 指纹图谱数据按照 95% 累计方差水平进行主成分降维解构，结果得到 5 个主成分（分别为 PC1、PC2、PC3、PC4、PC5），各批样品在这 5 个主成分对应的二维空间的得分分布情况见图 6。

ZBM. 浙贝母(Fritillariae Thunbergii Bulbus)　HBBM. 湖北贝母(Fritillariae Hupehensis Bulbus)

图 4　Pearson 相关性

Fig. 4　Pearson correlation

2.2.6　逻辑回归　将各批样品 REIMS 指纹图谱数据按照浙贝母和湖北贝母 2 种类别建立逻辑回归的分类模型，惩罚项参数选择岭回归(l2)，优化算法选择拟牛顿法(lbfgs)，结果所有样品的预测精确率、召回率以及 F1 分数均为 1.00，初步表明逻辑回归模型表现良好。

为了进一步客观准确地验证模型的预测能力，将 144 批样品按照浙贝母和湖北贝母 2 种类别根据约 8∶2 的比例将样品随机划分为训练集和测试集，即训练集样品为 115 批，其中浙贝母样品 59 批，湖北贝母样品 56 批；测试集样品为 29 批，其中浙贝母样品 15 批，湖北贝母样品 14 批。以训练集按上述参数建立逻辑回归分类模型，并选择 5 折法进行交叉验证，然后以所建立的模型对 29 批测试集样品进行预测，预测精确度、召回率以及 F1 分数均为 1.00。训练集内部交叉验证和测试集外部验证的特异性、召回率、精确率、准确率结果见表 2。

为了考察训练集样品个数对所建立的逻辑回归模型预测能力的影响以及模型的过拟合情况，按照 8∶2比例将样品随机划分训练集和测试集，然后按照训练集样品的个数由 2 至满训练集样品逐个训练并建立模型，并逐个分别计算模型的均方误差(mean square error，MSE)，绘制 MSE 与训练集样品数关系图，见图 7。

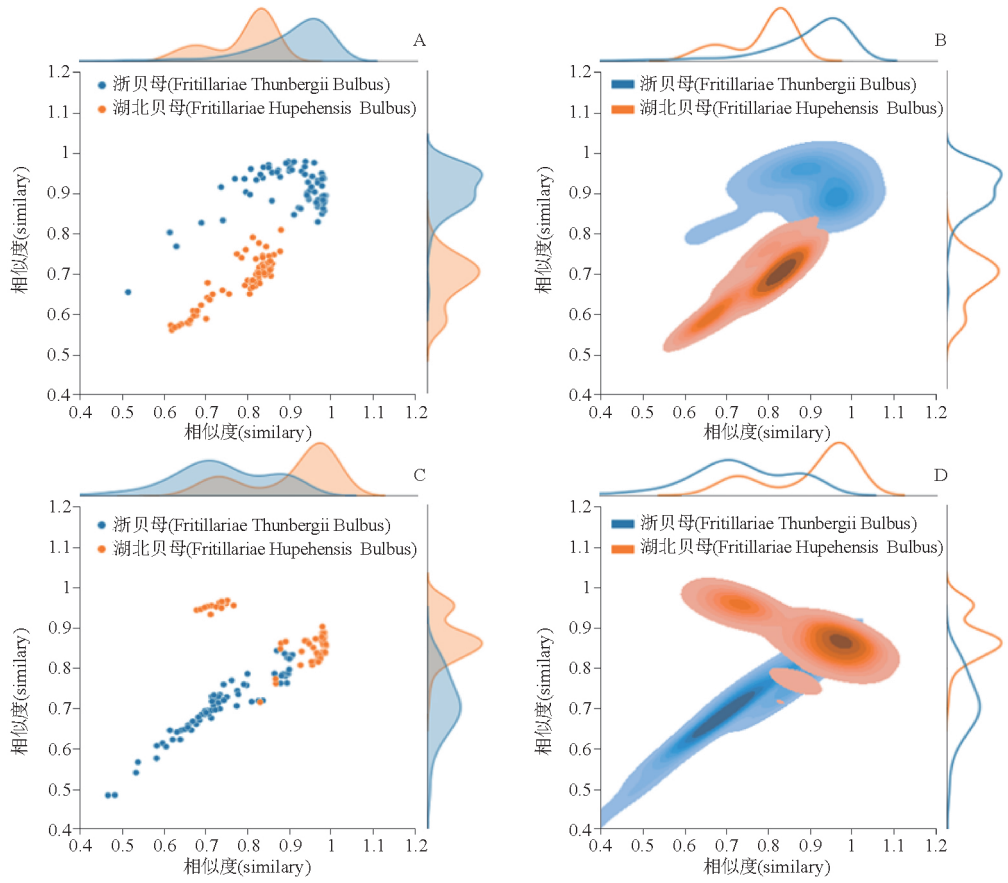

图 5　样品相似度分布

Fig. 5　Similarity of samples

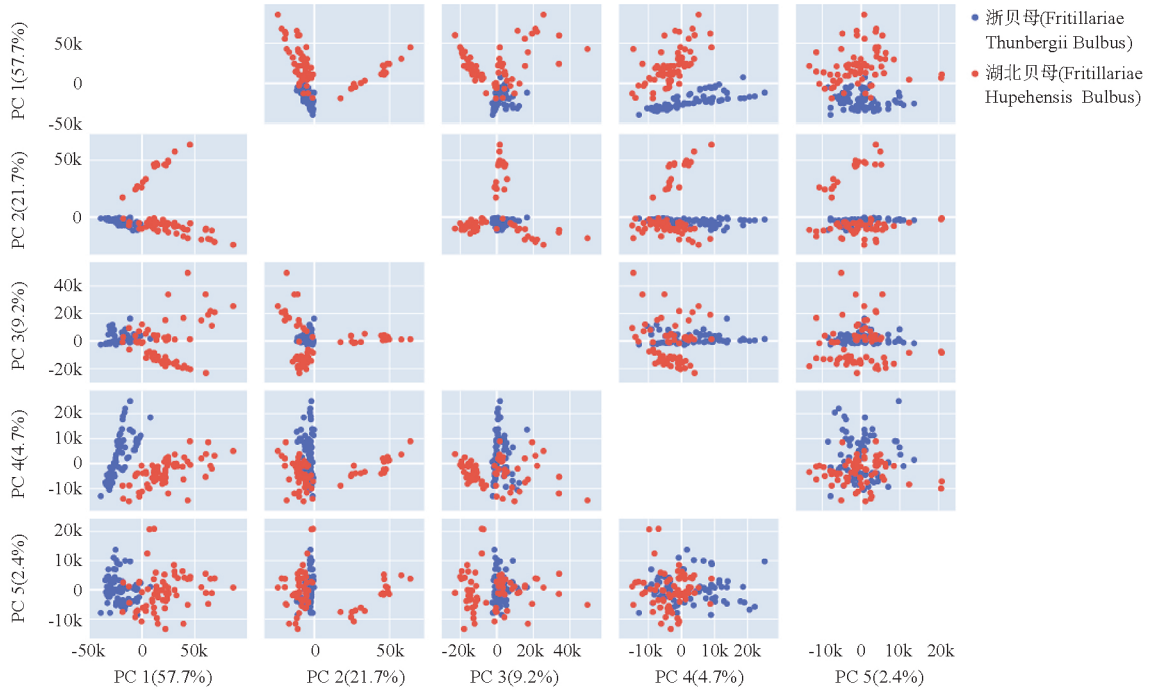

图 6　样品主成分得分分布图

Fig. 6　Principal component score plots of samples

表 2 逻辑回归分类模型预测评价

Tab. 2 Evaluation of classification prediction of Logistic Regression

样品类别 (class)	特异性 (specificity)	召回率 (sensitivity)	精确率 (precision)	准确率 (accurancy)	验证类型 (validation type)
浙贝母(Fritillariae Thunbergii Bulbus)	1.00	1.00	1.00	1.00	交叉验证(cross validation)
湖北贝母(Fritillariae Hupehensis Bulbus)	1.00	0.99	1.00		
浙贝母(Fritillariae Thunbergii Bulbus)	1.00	1.00	1.00	1.00	测试集(test set)
湖北贝母(Fritillariae Hupehensis Bulbus)	1.00	1.00	1.00		

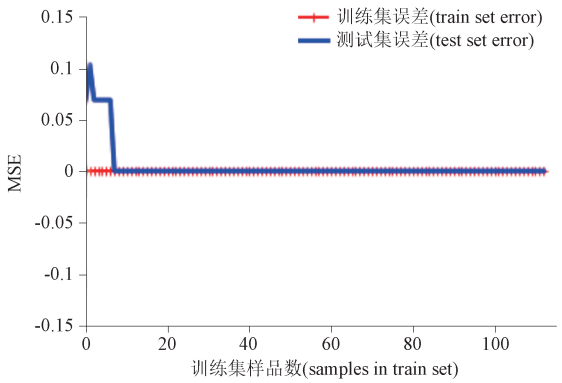

图 7 MSE 与训练集样品数关系

Fig. 7 Relation of MSE and samples in train set

3 讨论

3.1 样品形态的选择

目前市场上的贝母类样品基本以较完整的鳞茎和粉末这 2 种形态存在,为了扩大研究的实际使用意义,增加机器学习模型的泛化能力,同时考察不同形态下的样品对 REIMS 测定的影响,在对浙贝母和湖北贝母进行 REIMS 测定时,同时测定了较完整鳞茎样品和粉末状样品,而在进行数据分析及训练模型时按照样品的品种进行研究和模型效果的验证。结果表明,尽管不同形态样品的数据对于简单的聚类分析或是相似度分析影响较大,但对于属于监督学习的逻辑回归而言则毫无影响,在不同形态的浙贝母和湖北贝母样品的预测方面表现出极佳的鲁棒性。

3.2 REIMS 测定方法

本研究在采集样品中的 REIMS 数据时,考虑到贝母类样品中主要活性成分是生物碱类物质,而该类物质在质谱检测器中通过电离产生的基本为正离子,选择正离子模式采集可得到该类物质更丰富信息的数据,因此选择以正离子模式采集数据[7-9]。在对较完整鳞茎样品进行干法烧灼测定时,为了数据具有更好的代表性,采用了前期研究[20]的采样策略,兼顾了样品表面不同部位,烧灼点分布在大瓣或小瓣鳞叶的上部或下部。REIMS 检测对象基本都是具有导电性的肉类或者湿润的样品,继前期研究[20]之后,本研究继续采用直接使用烙铁头对干燥中药材进行烧灼采集数据的方式,获得了令人十分满意的结果,进一步确证了创新的干法 REIMS 的方法的可行性,对 REIMS 在中药材检测领域的应用具有开创性意义。但值得一提的是,对于研究数据返回到样本成分本身的路线,即烙铁头烧灼的方式是否对浙贝母和湖北贝母的成分有破坏这一点上,尚无研究试验验证或理论指导。

3.3 数据的预处理

由于采集 REIMS 数据时,人为操作或样品自身形态不同,会造成产生的气溶胶浓度具有差异,进而会造成离子信号强弱有别,因此须对数据进行标准化预处理。

3.4 无监督学习

3.4.1 聚类分析 由聚类分析结果的树状图(图 3)明显可见所有样品首先是分为了 2 大类,即浙贝母和湖北贝母,该处分类明确清晰,大部分的样品聚类准确,但是有 10 批湖北贝母粉末状样品(HBBM39 ~ HBBM47,HBBM51)聚类错误,可见尽管 REIMS 指纹图谱具有一定的品种特征性,但不同形态样品的数据仍然存在一些影响品种聚类的差异之处。

3.4.2 Pearson 相关性分析 Pearson 相关性系数(P_{corr})可以衡量向量间(X,Y)相关程度及相似度关系,其计算方法为 2 个向量的协方差除以向量的标准差,取值范围为 -1 ~ 1。

$$P_{corr}=\frac{\sum_{i}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i}^{n}(y_i-\bar{y})^2}}$$

Pearson 相关性分析多用于变量间相关程度的研

究[21-23]，本次研究将其用作表征不同样品间的相关或相似程度，结果如图 4 所示，蓝色颜色越深代表相关系数数值越大，相关性越强。图 4 显示所有样品均呈现正相关的关系，相同品种的样品 Pearson 相关系数一般较高，但是相同品种和不同品种间相关系数的数据分布有重叠情况，界限不够清晰，无法明确划分。

3.4.3 相似度分析 一般来讲，相似度的计算多采用夹角余弦法、欧氏距离法、Minkowsky 距离法、相关系数法等，一般以夹角余弦法最为常用，因此本研究采用夹角余弦法对样品进行相似度分析。从 Pearson 相关系数与夹角余弦相似度的数学定义可以推导出，在一定条件基础上，二者互为等价，因此在结果数据的分布趋势来看，二者基本保持了一致。在进行相似度分析时，本研究将不同品种且不同形态的样品数据分别进行了相似度分析，即按照各批样品分别对同品种下不同形态样品的平均指纹图谱的相似度分布情况进行了考察，分别以同品种不同形态样品的相似度绘制了分布核密度图，见图 5。

3.4.4 主成分分析 本研究所进行的主成分分析，旨在通过不同主成分的样品分布散点图对浙贝母与湖北贝母的数据特征性差异进行初步的研判。由图 6 可知，样品的 REIMS 指纹图谱数据经过主成分降维处理后，前 5 个主成分累计方差可达 95% 以上，基本可以代表样品数据主要信息。根据各批样品的 5 个主成分得分，可绘制成 10 个二维的样品得分分布图。如图 6 所示，样品的 REIMS 指纹图谱数据的第 1 主成分和第 4 主成分得分分布图中，浙贝母与湖北贝母的样品具有一定的特征性差异。因此，从本研究的数据来看，主成分降维后的样品数据基本可用作不同类别样品的鉴别，但是必须至少以第 1 主成分和第 4 主成分组合，单独以某 1 个主成分作为数据是无法进行准确鉴别的。

3.5 基于逻辑回归的有监督学习及相关评估指标

逻辑回归虽然名为回归，实则是一种分类算法，在机器学习领域内被广泛使用。逻辑回归的分类属性，或者说预测事件发生概率的属性是依托于 sigmoid 函数而实现的。sigmoid 函数可以用公式 $g(z)=\frac{1}{1+e^{-z}}$ 表示，绘制 sigmoid 函数曲线可得到“S”型的曲线，见图 8。当 $z>0$ 时，$g(z)>0.5$，且随着 z 数值的增大而接近 1，当 $z<0$ 时，$g(z)<0.5$，且随着 z 数值的减小而接近 0。

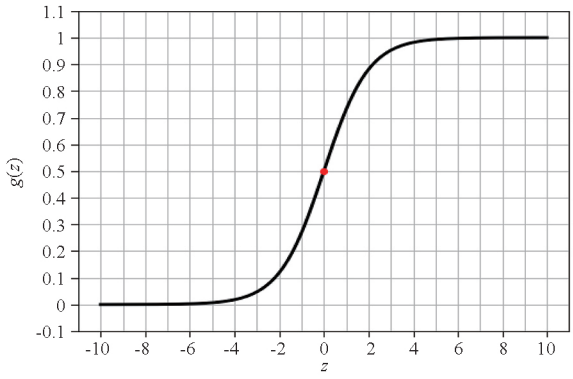

图 8 Sigmoid 函数曲线

Fig. 8 Sigmoid function curve

回到机器学习的研究中，对于常见的二分类问题（$y=0$ 还是 $y=1$），逻辑回归计算的概率即为 y 为 1 的概率，当逻辑回归计算概率 >0.5，则认为 y 为 1，若概率 <0.5，则认为 y 为 0[24]。由于是二分类问题的缘故，因此 $y=0$ 和 $y=1$ 的概率和为 1，即 $P(y=0)+P(y=1)=1$。下式即为逻辑回归的方程，其中 θ 和 X 分别表示参数矩阵和特征矩阵。

$$h_\theta(x)=g(\theta^{T}X)=\frac{1}{1+e^{-\theta^{T}X}}$$

逻辑回归机器学习模型的训练和优化，是基于代价函数（COST）对参数矩阵 θ 进行的最优化学习。逻辑回归的代价函数公式如下式表示，其中根据样品目标矩阵 y 取值为 0 或 1。当 $y=0$ 时，代价函数为 $-\log(1-h_\theta(x))$，当 $y=1$ 时，代价函数为 $-\log(h_\theta(x))$。

$$\mathrm{COST}(h_\theta(x),y)=-y\log(h_\theta(x))-(1-y)\log(1-h_\theta(x))$$

为了解决模型的过拟合问题，需要对代价函数进行正则化处理，即代价函数最小化过程中加入惩罚项。一般来讲，正则化有 2 种方式，lasso 回归和 ridge 回归，即 l1 和 l2。

逻辑回归中常用的优化算法有 lbfgs、liblinear、newton-cg、newton-cholesky、sag 和 saga，这些算法一般是基于牛顿法、拟牛顿法以及梯度下降法等经典算法原理。在选择使用优化算法时，须根据算法自身特点，结合模型是二分类还是多分类问题，惩罚项的不同等进行选择。lbfgs 法是一种基于拟牛顿算法的经典而流行的优化算法，所谓拟牛顿法，是指不

直接计算海森矩阵，而是通过一定方式构建近似海森矩阵或海森矩阵的逆，来迭代优化代价函数或损失函数，这样既能保证较为高效的收敛，又能保证模型的鲁棒性。lbfgs 法是通过构建近似海森矩阵的方式，在 bfgs 法基础上采用优先内存空间的方式实现的，其在各个领域都表现出了强大的生命力，目前已成为无约束优化领域的核心算法。本研究为二分类问题，采用的惩罚项为 l2，最终选择使用 lbfgs 法，取得令人满意的分类结果。

特异性、召回率、精确率、准确率是评估机器学习模型学习和预测分类效果的常用经典指标。样品经过机器学习模型学习和预测分类之后，可以根据样品的真实类别情况以及模型分类结果分为 4 类，即真阳性（TP）、假阴性（FN）、真阴性（TN）和假阳性（FP）。那么特异性可以用公式$\frac{TN}{TN+FP}$表示，召回率可以用公式$\frac{TP}{TP+FN}$表示，精确率可以用公式$\frac{TP}{TP+FP}$表示，准确率可以用公式$\frac{TP+TN}{TP+FN+TN+FP}$表示。根据以上公式可知，特异性、召回率、精确率、准确率的数值范围为 0 ~ 1，数值越大越代表模型的学习和预测分类效果优秀。如表 2 所示，本研究所建立的逻辑回归模型在样品训练集内部 5 折法交叉验证和测试集外部验证的特异性、召回率、精确率、准确率，除对湖北贝母的召回率为 0.99 外，其他指标数值均为 1，可见逻辑回归模型应用于浙贝母与湖北贝母干法 REIMS 指纹图谱数据预测分类的效果非常准确，尤其是更为重要的针对测试集样品的 4 项判别指标都为 1，判别完全准确。

机器学习不仅依赖于单纯的算法，样品的数量也直接关系到机器学习的效果。在目前的大数据时代，人类活动会产生海量的各种数据，对于药品分析而言，其样品数据的获得则依赖于样品的收集情况和实验测定活动。在一定的阶段中，样品批数可能会有所受限，那么在现有的样品批数情况下，数据数量对所建立的机器学习模型影响程度就应该需要纳入考量。本研究所绘制的图 7 表示了在训练集样品数量逐渐增加的情况下，逻辑回归模型的内部交叉验证和测试集的 MSE 的变化趋势。由该图可以得知，最初内部交叉验证的 MSE 为 0，而测试集则有较大 MSE，表示模型在训练集样品批数较少时预测分类效果一般，而随着训练集中样品批数的增加，测试集 MSE 迅速为 0，并且随着训练集样品批数的持续增加，测试集 MSE 始终保持为 0，表示模型并未出现过拟合情况。

在本研究中，通过对 115 批训练集样品的 REIMS 指纹图谱数据进行学习训练而优化得到的逻辑回归模型在内部交叉验证和测试集样品验证中，都获得了准确的结果，尤其是对于 29 批测试集样品的准确预测分类，充分表明了该模型的有效性。此外，考虑到具体应用情况，在采集得到 REIMS 指纹图谱数据时混合采集了较完整鳞茎和粉末状 2 种常见形态的浙贝母和湖北贝母，逻辑回归验证结果也充分说明了模型预测能力对于样品形态变化的良好的鲁棒性。

3.6 展望

本研究是在前期开创性采用干法 REIMS 对中药材及饮片进行研究[20]基础之上，对不同形态浙贝母和湖北贝母进行的一次有益尝试。使用首创的电烙铁直接对干燥中药材取样测定法，对贝母类药材 2 种常见的形态，即较完整鳞茎和粉末状分别进行质谱数据采集，并使用机器学习领域的多种经典算法对所获得的 REIMS 数据进行分析与研究，其中建立的逻辑回归监督学习模型对 2 类样品的分类可以达到极高的正确率。本研究不仅表明 REIMS 对贝母类药材分类预测的强大能力，同时也再一次验证了 REIMS 结合机器学习在中药材及饮片研究领域的良好适用性。

参考文献

[1] 王书军，高文远，于琳，等. 百合科贝母属药用植物分类研究进展[J]. 中国中药杂志，2007，32(16)：1609
WANG SJ，GAO WY，YU L，*et al*. Progress of taxonomic study on *Fritillaria* (Liliaceae) medicinal plant[J]. China J Chin Mater Med，2007，32(16)：1609

[2] 肖培根. 湖北贝母的研究进展[J]. 中国中药杂志，2002，27(10)：726
XIAO PG. A review on the study of Hubeibeimu[J]. China J Chin Mater Med，2002，27(10)：726

[3] YU M，WANG S，CAI R，*et al*. Discrimination and content analysis of *Fritillaria* using near infrared spectroscopy[J]. J Anal Methods Chem，2015：752162

[4] ZHOU M，MA X，DING G，*et al*. Comparison and evaluation of antimuscarinic and anti-inflammatory effects of five Bulbus Fritillariae species based on UPLC-Q/TOF integrated dual-luciferase

reporter assay, PCA and ANN analysis[J]. J Chromatogr B Analyt Technol Biomed Life Sci, 2017, 1041 – 1042: 60

[5] 中华人民共和国药典 2020 年版. 一部[S]. 2020: 304
ChP 2020. Vol Ⅰ[S]. 2020: 304

[6] 马秀芹, 秦伟华, 李学斌, 等. 浙贝母与其混淆品湖北贝母的鉴别[J]. 时珍国药研究, 1997, 8(1): 37
MA XQ, QIN WH, LI XB, *et al*. Identification of *Fritillaria thunbergii* and its adulterant *Fritillaria hupehensis*[J]. Lishizhen Med Mater Med Res, 1997, 8(1): 37

[7] 周建良, 刘伟, 郭增喜, 等. 基于快速液相色谱 – 四极杆飞行时间串联质谱的浙贝母特征图谱研究[J]. 中国中药杂志, 2013, 38(17): 2832
ZHOU JL, LIU W, GUO ZX, *et al*. Fingerprint analysis of *Fritillaria thunbergii* using rapid resolution liquid chromatography coupled with electrospray ionization quadrupole time – of – flight tandem mass spectrometry[J]. China J Chin Mater Med, 2013, 38(17): 2832

[8] ZHANG ZF, LU LY, LIU Y. Comparing major alkaloids of Fritillariae Hupehensis Bulbs (FHB) and congeneric plants by HPLC – ELSD and HPLC – ESI – MS^n[J]. Nat Prod Res, 2014, 28(15): 1171

[9] WANG Z, XIE H, REN J, *et al*. Metabolomic approach for rapid differentiation of *Fritillaria* bulbs by matrix – assisted laser desorption/ionization mass spectrometry and multivariate statistical analysis[J]. J Pharm Biomed Anal, 2020, 185: 113177

[10] SCHÄFERFER KC, DÉNES J, ALBRECHT K, *et al*. *In vivo*, *in situ* tissue analysis using rapid evaporative ionization mass spectrometry[J]. Angew Chem Int Ed, 2009, 48(44): 8240

[11] BALOG J, SZANISZLO T, SCHAEFER KC, *et al*. Identification of biological tissues by rapid evaporative ionization mass spectrometry[J]. Anal Chem, 2010, 82(17): 7343

[12] STRITTMATTER N, REBEC M, JONES EA, *et al*. Characterization and identification of clinically relevant microorganisms using rapid evaporative ionization mass spectrometry[J]. Anal Chem, 2014, 86(13): 6555

[13] BALOG J, KUMAR S, ALEXANDER J, *et al*. *In vivo* endoscopic tissue identification by rapid evaporative ionization mass spectrometry (REIMS)[J]. Angew Chem Int Ed, 2015, 54(38): 11059

[14] 高海燕, 孟宪双, 付珺珺, 等. 快速蒸发电离质谱技术及其应用研究进展[J]. 分析测试学报, 2021, 40(2): 159
GAO HY, MENG XS, FU JJ, *et al*. Research progress on rapid evaporative ionization mass spectrometry and its applications[J]. J Instrum Anal, 2021, 40(2):159

[15] SONG G, CHEN K, WANG H, *et al*. *In situ* and real – time authentication of *Thunnus* species by iKnife rapid evaporative ionization mass spectrometry based lipidomics without sample pretreatment[J]. Food Chem, 2020, 318: 126504

[16] CUI Y, WANG H, ZHAO Q, *et al*. Real – time detection of authenticity and adulteration of krill phospholipids with soybean phospholipids using rapid evaporative ionization mass spectrometry: application on commercial samples[J]. Food Control, 2021, 121: 107680

[17] MANOLI E, MASON S, FORD L, *et al*. Validation of ultrasonic harmonic scalpel for real – time tissue identification using rapid evaporative ionization mass spectrometry[J]. Anal Chem, 2021, 93: 5906

[18] 刘鸣畅, 林继红, 刘哲硕, 等. 快速蒸发电离质谱技术(REIMS)鉴别肉品[J]. 质谱学报, 2020, 41(5): 470
LIU MC, LIN JH, LIU ZS, *et al*. Identification of meat by rapid evaporation ionization mass spectrometry (REIMS)[J]. J Chin Mass Spectrom Soc, 2020, 41(5): 470

[19] 张燕平, 王海星, 陈康, 等. 实时质谱新技术分析南极磷虾油的脂质组学轮廓[J]. 中国食品学报, 2020, 20(9): 226
ZHANG YP, WANG HX, CHEN K, *et al*. Novel real – time mass spectrometry for detecting the lipidomics profile of antarctic krill oil[J]. J Chin Inst Food Sci Technol, 2020, 20(9): 226

[20] 石岩, 姚令文, 魏锋, 等. 基于干法快速蒸发离子化质谱(REIMS)指纹图谱与机器学习算法联用的白头翁真伪判别研究[J]. 中国中药杂志, 2023, 48(4): 921
SHI Y, YAO LW, WEI F, *et al*. Authenticity discrimination of Pulsatillae Radix based on dry – process REIMS fingerprinting combined with machine learning[J]. China J Chin Mater Med, 2023, 48(4): 921

[21] SANDUSKY P, RAFTERY D. Use of semiselective TOCSY and the pearson correlation for the metabonomic analysis of biofluid mixtures: application to urine[J]. Anal Chem, 2005, 77(23): 7717

[22] WANG Z, LUO P, CHENG L, *et al*. Hapten – antibody recognition studies in competitive immunoassay of α – zearalanol analogs by computational chemistry and pearson correlation analysis[J]. J Mol Recognit, 2011, 24(5): 815

[23] LUNSFORD R, GILLIS D, GRUN J. Automated identification of components in a chemical mixture utilizing multi – wavelength resonant – Raman spectroscopy and a Pearson correlation algorithm[J]. J Raman Spectrosc, 2012, 43(10): 1472

[24] GÉRON A. Hands – on Machine Learning with Sciket – Learn, Keras & Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems[M]. 2nd Ed. Sebastopol: Reilly Media, 2019

(本文于 2023 年 11 月 17 日修改回)