机器学习在合成大麻素识别鉴定中的应用进展

许 情[1,2], 吕 敏[1,2], 邓虹霄[1], 胡 驰[2], 向 平[1], 陈 航[1*]

([1] 司法鉴定科学研究院, 司法部司法鉴定重点实验室, 上海市法医学重点实验室, 上海市司法鉴定专业技术服务平台, 上海 200063; [2] 中国药科大学药学院, 南京 210009)

摘 要 合成大麻素是一种人工合成的可以引起公共健康风险的精神活性物质,且合成大麻素结构多变,容易被结构修饰,结构未知的合成大麻素的快速出现使得对其鉴识面临了新的挑战。近年来,机器学习已取得很大的进展,已经广泛应用到其他领域,也为结构未知合成大麻素的鉴识以及可能的来源推断提供了新的策略。本文阐述了常用机器学习方法的原理以及机器学习技术在合成大麻素类物质的质谱分析、拉曼光谱分析、代谢组学以及定量构效关系等方面的应用,以期为未知合成大麻素的鉴识提供新的思路。

关键词 合成大麻素;机器学习;非靶向筛查

中图分类号 TP181;R917 **文献标志码** A **文章编号** 1000−5048(2024)03−0316−10

doi: 10.11665/j.issn.1000−5048.2023113003

引用本文 许情,吕敏,邓虹霄,等. 机器学习在合成大麻素识别鉴定中的应用进展 [J]. 中国药科大学学报,2024,55(3):316 − 325.

Cite this article as: XU Qing, LYU Min, DENG Hongxiao, *et al*. Advances in the application of machine learning in the identification and authentication of synthetic cannabinoids[J]. *J China Pharm Univ*, 2024, 55(3): 316 − 325.

Advances in the application of machine learning in the identification and authentication of synthetic cannabinoids

XU Qing[1,2], LYU Min[1,2], DENG Hongxiao[1], HU Chi[2], XIANG Ping[1], CHEN Hang[1*]

[1]Shanghai Forensic Service Platform, Shanghai Key Laboratory of Forensic Medicine, Key Laboratory of Forensic Science of Ministry of Justice, Academy of Forensic Science, Shanghai 200063, China; [2]School of Pharmacy, China Pharmaceutical University, Nanjing 210009, China

Abstract Synthetic cannabinoids (SCs) are synthetic psychoactive substances that can pose a public health risk. The SCs are structurally variable and susceptible to structural modification. The rapid emergence of structurally unknown synthetic cannabinoids has led to new challenges in their identification. In recent years, machine learning has made great progress and has been widely applied to other fields, providing new strategies for the identification of unknown synthetic cannabinoids and the inference of possible sources. This paper describes the principles of commonly used machine learning methods and the application of machine learning techniques to mass spectrometry, Raman spectroscopy, metabolomics and quantitative conformational relationships of synthetic cannabinoids, aiming to provide new ideas for the identification of unknown synthetic cannabinoids.

Key words synthetic cannabinoids; machine learning; non-targeted screening

This study was supported by the National Key Research and Development Program of China (No.2022YFC3300903), the Social Welfare Research Projects of Centralized Research Institutes(No.GY2022D-1), and the Project of Shanghai Key Laboratory of Forensic Medicine(No.21DZ2270800)

收稿日期 2023-11-30 *** 通信作者** Tel: 021-52352955 E-mail: chenh@ssfjd.cn

基金项目 国家重点研发计划项目 (No.2022YFC3300903); 中央级科研院所社会公益研究专项 (No.GY2022D-1); 上海市法医学重点实验室资助项目 (No.21DZ2270800)

合成大麻素[1]是一类新精神活性物质, 与 Δ^9-四氢大麻酚和内源性大麻素类似, 靶向大麻素受体 1 和 2(CB1 和 CB2)[2]。自 2006 年第一代合成大麻素产品在国际上出现以来, 合成大麻素逐渐成为世界上滥用最广泛的药物之一。合成大麻素通常比天然大麻产生更强的不良作用[3-4], 这可能是由于其在 CB1 上的结合亲和力更高。为了逃避检查, 许多不法分子将合成大麻素溶解在有机溶剂中, 并喷洒在香料和草药上出售。因此, 合成大麻素通常被称为“草药”“香料”和“小枝”。合成大麻素具有较大结构多样性, 尽管它们的化学异质性, 但大多数都被一个通用的 Markush 结构[5]所包围, 该结构由 4 个亚基组成: 母核(蓝色)、链接(橙色)、取代基(绿色)和侧链(红色), 其总体结构如图 1 所示。合成大麻素是临床上常用的高效镇痛药, 滥用此类药物可引起心率加快、时间幻觉、恶心、呕吐、注意力难以集中、神经元破坏导致妄想症状等[6−7]。随着相关技术的快速发展, 该类药物的吸食载体越来越多样化, 种类越来越多, 并且在外观和形态上与合法产品越来越相似, 这给该类药物的监管增加了难度。另外, 近年来不法分子通过对合成大麻素进行结构修饰, 改变官能团来逃避监管, 因此, 如何更高效地识别和筛选合成大麻素已成为亟待解决的问题。

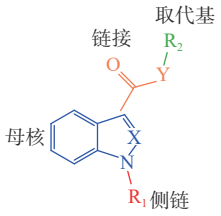

图 1　合成大麻素结构

目前合成大麻素的鉴识多基于靶向筛查策略, 将检出的未知成分质谱图与标准品质谱图比对分析, 确认查获样本中的主要化学成分; 或者与自建或者公开的数据库中已知的合成大麻素的谱图进行比对, 而对于未知的合成大麻素的鉴识却束手无策。机器学习是一种算法模型的总称, 能借助计算机强大的算力支持从大量数据中发现隐含的规律并将其应用于数据的分类和预测。近年来机器学习算法在药物研发领域已经得到较为广泛的应用。借助机器学习算法可以自动对量生物数据以及化学数据进行处理, 以更高效地发现潜在有效药物[8]。在药物发现过程中, 也极大地促进了计算机辅助药物设计的发展, 计算机辅助药物设计基于分子相互作用关系高效用于前期药物靶点研究和先导化合物筛选[9], 也可对药物毒性、耐药性、药物之间相互作用进行预测, 许多机器学习算法已被证明根据药物结构特征预测药物性质方面表现出色[10]。同时, 机器学习算法可以处理来自基因组学、蛋白质组学等多组学和临床试验中大量复杂的数据, 找出潜在的通路、蛋白和机制等与疾病的相关性, 以发现新机制和新靶点。

机器学习算法在医药领域中的广泛应用也为未知合成大麻素的鉴识以及可能的来源推断提供了新的策略。本文对机器学习技术在合成大麻素类物质的质谱分析、拉曼光谱分析、代谢组学以及定量构效关系等方面的研究进行了综述。

1　各种机器学习模型的原理及适用范围

机器学习是一类算法模型的总称, 是通过利用数据训练模型, 使用模型进行预测的一种方法[11]。本文简要介绍毒品分析中目前最常用的算法, 根据训练数据是否有标记信息, 机器学习算法可分为“无监督学习”和“有监督学习”两类, 其优缺点如表 1 所示。

1.1　无监督学习(unsupervised learning)

无监督学习的样本没有任何标记, 无监督算法需要自动找到这些没有标记的数据里面的数据结构和特征。在这种情况下, 训练数据不需要任何手工标注的标签, 其中的代表算法包括主成分分析、K-均值聚类和层次聚类。

1.1.1　主成分分析(principal component analysis, PCA)

PCA 算法[12]是最常用的线性降维方法, 它的目标是通过某种线性投影, 将高维的数据映射到低维的空间中, 并期望在所投影的维度上数据的信息量最大(方差最大), 以使用较少的数据维度, 同时保留住较多的原数据的特性。该算法常和其他的机器学习算法相结合, 用于数据前处理过程。

1.1.2　K-均值聚类(k-means clustering, K-means)

K-均值聚类是一种常见的聚类算法, 其算法的思想大致为: 先从样本集中随机选取 K 个样本作为簇中心, 计算所有样本与这 K 个“簇中心”的距离, 对于每一个样本, 将其划分到与其距离最近的“簇中心”所在的簇中, 对于新的簇计算各个簇的新的“簇

表 1 常用机器学习算法模型优缺点

算法名称	优点	缺点
主成分分析	降低数据维度，去除噪声，便于数据可视化和进一步处理，提高计算效率	对异常值敏感，受到样本量和变量个数限制
K-均值聚类	算法简单，容易实现	对数据类型要求较高，适合数值型数据；须事先确定K
层次聚类	可解释性强，无须事先确定聚类数量	计算复杂度高，对噪声和异常值敏感。
K最近邻算法	理论成熟，可用于非线性分类	计算量大，需要大量内存；不适合样本不平衡数据
逻辑回归	实现简单，分类时计算量较小，速度快	容易欠拟合；只能处理二分类问题
支持向量机	泛化能力强，可以解决高维问题	数据样本较大时，计算复杂度升高，训练时长大幅增加
决策树	易于理解和解释，可以可视化分析；比较适合有缺失属性的样本	处理缺失数据困难，容易出现过拟合问题
随机森林	可以用来处理较高维度数据，且不用降维；可以判断特征的重要程度；不容易过拟合；对于不平衡的数据集可以平衡误差	在噪音较大的分类问题上会过拟合
神经网络算法	具有较高非线性拟合能力，可以映射复杂的非线性关系，呈现较高的鲁棒性和自学习能力	数据量较少的情况下，预测准确性降低；缺乏解释模型推理过程和推理能力的能力

中心”，分别计算到簇内其他点距离均值最小的点作为质心。

1.1.3 *层次聚类*（hierarchical clustering） 层次聚类是聚类算法的一种，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。对于给定的样本集合，首先将每个样本分到一个类，然后按照一定规则，例如类间距离最小，将最满足规则条件的两个类进行合并如此反复进行，每次减少一个类，直到满足停止条件。该算法已应用在新精神活性物质谱图分类研究中，如 Gilbert 等[13] 通过对 54 种芬太尼类似物的质谱数据进行 PCA，并结合层次聚类的算法将其分为 9 类，该模型能够根据化学修饰的性质和位置，对 67 种不包括在模型中的芬太尼类似物进行准确分类。

1.2 *有监督学习*（supervised learning）

在有监督学习中，提供算法一组训练数据，该数据包括输入特征和对应的标签。算法通过对这组训练数据的分析来学习特征与标签之间的关系，并使用此学到的关系来预测新数据的标签。常见的监督学习算法有 K 最邻近算法、线性回归和逻辑回归、支持向量机、决策树和随机森林和神经网络算法。

1.2.1 *K 最邻近算法*（k-nearest neighbor, KNN）

KNN 是一个众所周知的简单算法，主要用于判断未知样本的类别[14]，该模型的输入为预先标记类别的数据集，以所有已知样本作为参照，计算未知样本与已知样本的距离，从中选取与未知样本距离最近的 K 个已知样本，根据少数服从多数的投票原则，将未知样本与 K 个最近邻样本中所属类别占比较多的归为一类。

1.2.2 *线性回归*（linear regression）*和逻辑回归*（logistic regression） 线性回归为最为基础的机器学习算法之一，该算法利用大量的样本，通过有监督的学习，学习由 x 到 y 的映射 f，利用该映射关系对未知的数据进行预估，寻找参数 w 和 b，使得对训练集的预测值和真实的回归目标值 y 之间的均方误差最小，因为 y 为连续值，所以是回归问题；线性回归可以预测连续值，但是不能解决分类问题，需要根据预测的结果判定其属于正类还是负类。逻辑回归是一种广义上的线性回归模型，实际上逻辑回归就是将线性回归的结果，通过 sigmoid 函数映射到(0,1)之间。线性回归得到大于 0 的输出，逻辑回归就会得到(0.5, 1)的输出，线性回归得到小于 0 的输出，逻辑回归就会得到(0, 0.5)的输出。总的来说，线性回归解决的是回归的问题，逻辑回归相当于是在线性回归的基础上来解决分类的问题。

1.2.3 *支持向量机*（support vector machine, SVM） SVM 算法主要用来执行分类任务，该算法基本原理是：找到一条最佳分割超平面，使得两类数据点尽量分开，同时间距离分割超平面最近的数据点距离最远[15]。该算法适用于数据高维小样本的情况，Broséus 等[16] 基于非大麻素和大麻素叶子成分，构建 SVM 模型，用以区分纤维型和药物型大麻幼苗，整个样本集的分类率在 99% 以上，假阳性率小于 2%。

1.2.4 *决策树*（decision tree）*和随机森林*（random forests, RF） 决策树[17] 是一个属性结构的预测模型，代表对象属性和对象值之间的一种映射关系。它由节点和有向边组成，其节点有两种类型：内节

点和叶节点, 内部节点表示一个特征或属性, 叶节点表示一个类。RF 是一种分类算法, 该算法是通过自助法(bootstrap)重采样技术, 从原始数据中以有放回的方式随机取样得到 n 个训练数据集, 从每个训练数据集中随机选择 k 个特征。反复根据这 k 个特征建立起来 m 棵决策树, 应用每个决策树来预测结果, 并且保存所有预测的结果。对分类模型进行投票, 计算每个预测结果的得票数, 选择得票数最高的模型作为最终决策, 该方法可以通过平均决策树, 可降低过拟合的风险。RF 在分类可解释性及缺失值容忍程度上具有无可比拟的优势。

1.2.5　神经网络算法(artifical neural network, ANN)　ANN 类似于人类大脑解决问题的方式。神经网络最基本的构成元素是神经元, 每个神经元都具有输入、数值处理以及输出的能力。在最简单的情况下, 神经网络[18]由输入层、隐藏层和输出层组成。神经元从输入层通过一个或多个隐藏层链接到输出层, 各层神经元通过激活函数和权重系数相连。样本数据中每个特征属性对应模型输入层中的一个神经元。神经网络模型中可根据实际问题需求包含多个隐藏层。经过隐藏层计算和处理信息, 在输出层输出最终分类或回归分析结果。使用较多的是反向传播神经网络(back propagation, BP), 即通过输入数据的反复训练, 不断修改变量之间影响的系数, 最终达到最优输出结果, 适合解决内部复杂的数据问题。

2　机器学习技术结合质谱技术分析用于合成大麻素识别

质谱是新精神活性物质检测最有力的工具之一。作为经典的识别策略, 通常依赖于质谱数据库和标准品。机器学习具有阅读并理解质谱分析数据结果的能力, 有望成为质谱分析中合成大麻素类物质识别的重要辅助工具。

2.1　机器学习结合质谱分析辅助未知合成大麻素结构推断

RF、SVM、ANN 目前常于质谱结合用于合成大麻素类物质的识别。其原理主要为质谱能获得化合物碎片离子质荷比及其丰度, 碎片离子的相对丰度与分子结构有密切关系, 机器学习可以建立起质谱数据与化学结构之间的特征向量关系。主要是基于这种被学习出来的特征向量关系, 定量的对化合物结构存在的可能性进行预测。目前已有不少研究将机器学习与质谱结合应用于合成大麻素类物质的识别。

Yang 等[19]使用包含 567 个 LC-MS 和 732 个 GC-MS 的数据集生成并评估了 4 种分类模型——KNN、SVM、RF 和 gcForest 来快速筛查新精神活性物质。该研究将收集到的 1 299 个物质质谱数据整合到数据库中, 每种物质用 4 位数字编码(0000~9999), 其中第 1 位数字代表类别最后 3 位数字代表序列号。在用机器学习模型训练前, 对数据进行了预处理, 实现数据的离散化。通过等宽分箱实现峰对齐, 每个箱位置处的特征值对应丰度, 当箱缺失时默认值为零。此外, 为了避免极端特征值并减少由不同碎裂电压引起的偏差, 对丰度特征进行平方根和 L1 正则化。特征空间在 m/z 1~600 范围内构建, 质谱数据的主要特征包含每个 m/z 对应的丰度和二级碎片离子的丰度。按照 8∶2 的比例划分训练集和测试集。使用 4 种算法 KNN、SVM、RF 和 gcForest, 基于该数据集生成分类模型, 模型使用具有 5 倍交叉验证的网格搜索进行优化, 以实现每个模型的最佳学习超参数。4 个模型对两个数据集的芬太尼类物质均达到了较高的准确率和召回率, 说明芬太尼类物质的预警信号具有较高的可信度。对于合成大麻素类物质的识别, gcForest 的表现优于其他 3 个模型。此外, gcForest 对合成卡西酮类物质和阴性样品也具有良好的识别能力。这些模型为合成大麻素、合成卡西酮和芬太尼提供了警告信号。成功建立了一个预警系统, 为识别新精神活性物质提供了一种有用的方法, 实现了未知样品的分类任务, 从而为未知化合物的结构鉴定提供了依据, 并且在几个实际查获的样品上使用了该方法, 已被证明能够快速有效地筛选未知样品中的新精神活性物质。从这个实际应用中, 看到了机器学习技术在合成大麻素结构识别领域的潜力。

Wong 等[20]开发了机器学习模型训练 GC-MS 数据识别未知新精神活性物质。该研究训练和评估多个监督机器学习分类器, 即 ANN、卷积神经网络(CNN)和平衡随机森林(BRF)。能够将 6 个新精神活性物质类别合成卡西酮、合成大麻素、苯乙胺、哌嗪、色胺和芬太尼和其他化合物进行有效地分类。其中 BRF 是表现最好的模型, 该模型的准确性优于经典的库匹配。

Lee 等[21]构建了基于高分辨率液相色谱串联

质谱的机器学习模型，以解决识别已列管物质和未知新型精神活性物质的分析挑战。利用 770 个高分辨率液相色谱串联质谱条形码光谱组成的训练集，生成并评估了 3 种分类机器学习模型。这 3 种模型分别是 ANN、SVM 和 KNN 模型。在这些模型中，已列管物质和新精神活性物质被划分为 13 个亚组（苯基哌嗪、阿片类药物、苯二氮䓬类药物、安非他明、可卡因、甲卡西酮、经典大麻素、芬太尼、2C 系列、茚唑羰基化合物、吲哚羰基化合物、苯环利定等）。以 193 个 LC-MS-MS 条码光谱作为外部测试集，ANN、SVM 和 KNN 模型的准确率分别为 72.5%、90.0% 和 94.3%，其中 KNN 模型取得了最高的分类准确性，能够识别数据库中没有数据的新精神活性物质。

此外，还有一些基于机器学习技术结合质谱用于芬太尼类物质结构推断的研究，其研究方法具有扩展到合成大麻素类物质的潜力。如 Koshute 等[22] 提出有监督的机器学习分类模型作为库匹配的补充方法，用于从质谱中检测芬太尼类似物。从质谱中提取出了 24 个基于峰值和基于相似性的输入特征。质谱峰相关特征峰包含基峰，平均峰强度和出现最频繁的质谱峰对质量差等。而相似度相关特征主要计算谱图与几种代表性的芬太尼类物质的谱图相似性。考察了 3 种不同的机器学习模型——逻辑回归、ANN 和 RF，模型的选择基于交叉验证集的性能，遵循 10 倍交叉验证策略。Moorthy 等[23] 通过与已知结构的芬太尼类物质计算谱图相似度来确定与可疑芬太尼类物质最为相近的芬太尼类物质化学结构，通过阈值设定判断其为 1 型或者 2 型芬太尼类物质；接着用构建的多维尺度聚类模型判断该可疑芬太尼类物质可能的结构修饰位点。最终综合这两项结果给出关于该可疑芬太尼类物质的化学结构预测。该研究验证了无监督的聚类模型对于未知化合物结构自动归属的可行性。

2.2 机器学习结合质谱解析区分同分异构体

随着越来越多的新精神活性物质进入非法药物市场，需要强有力且有效的异构体识别方法。机器学习技术，如 RF 分类器，可能最适合通过有效地利用质谱中的微小差异来分配正确的同分异构体形式来处理这一分析问题。机器学习技术已用于区分其他新精神活性物质的同分异构体，区分合成大麻素异构体的研究却还未见报道。

Setser 等[24] 建立了基于质谱特征的线性判别分析模型，对合成苯乙胺和色胺进行区分，并对其进行了验证。并比较了采用两种方法选择特征变量。首先，选择已知的每种化合物类别的特征离子，从而产生用于开发 LDA 模型的总共 13 个变量。通过对模型的交叉验证，对苯乙胺和色胺的测试集进行分类，分类成功率为 93%。在第 2 种方法中，PCA 被用作更客观的变量选择方法。该方法共选取了 9 个变量，得到的 LDA 模型分类成功率为 86%。虽然每个 LDA 模型的分类成功率相似，但与需要探测质谱以获取特征的更详细方法相比，PCA 方法用于变量选择的时间要少得多。这里报道的分类模型对于尚未获得参考材料的新兴类似物的类别表征具有潜在的实用性。此外，Bonetti 等[25] 将 DART-TOF 的数据结合 RF 来区分 3 种位置异构体：氟安非他明、氟甲基安非他明和甲基甲卡西酮，分类成功率达到 93.9%，错误率始终保持在 5% 以下，为法医实验室区分同分异构体提供了一种快速可靠的方法。

2.3 机器学习结合质谱解析用于合成大麻素来源识别

研究人员可以通过非法药物特征分析来确定不同缉获的样品之间的联系，以获取贩运路线的信息和收集关于样品来源的背景资料。一些非法药物可以通过评估特征外部参数，如颜色、形状和标志来进行分析。然而，当处理视觉上难以区分的样品时，对样品的化学特性进行更广泛的研究是必要的，在秘密合成过程中形成的杂质是检获的非法药物样本中最重要的鉴别化学特征。在制造过程中，通常除了所需的主要活性成分外，还会产生各种固有杂质，这些杂质是所应用的合成途径及其条件所特有的。研究的目的是确定样品是否可以根据其特征杂质谱分组或区分，以及可以从这些数据中提取哪些来源和生产信息，例如不同的合成途径，反应批次和批次大小。机器学习的引入为这一领域的研究带来新的思路，能帮助研究人员探查各种合成大麻素的地下生产方式。如，Münster-Müller 等[26] 将 4 个缉获的含有合成大麻素 Cumyl - 5F - PINACA 电子烟油样本和 11 个市售烟油样本共 15 个样品处理后，使用超高效液相色谱-质谱测量每个样品的混合杂质分数，通过

HCA 分析目标副产物的相对含量, 根据杂质特征的相对距离对 15 个电子烟油样品进行分组。聚类的结果表明, 购买日期、在线商店的身份和品牌名称是样品聚类的关键因素。

3　机器学习结合拉曼光谱用于合成大麻素的识别

无监督算法 PCA, 有监督算法人工神经网络、RF、SVM、最近邻算法常结合拉曼光谱用于合成大麻素识别。拉曼光谱可以同时测量多种化合物, 但在用其分析复杂的混合物样品时, 会产生庞大的数据, 很难用肉眼可视化, 机器学习算法可以被训练来提取复杂光谱数据中的相关特征, 并预测新化合物的类别, 从而改进检测、鉴定和分类。目前已有研究将曼光谱获得的拉曼特征峰的峰位和峰强度, 作为特征输入到机器学习分类预测模型中进行训练从而实现对未知合成大麻素进行分类识别。

在合成大麻素类物质的检测方面, Lee 等[27] 结合 PCA 和 ANN 对吲哚和吲唑酰胺合成大麻素进行分类。该研究基于 25 个吲哚/吲唑类合成大麻素的标准品的拉曼光谱数据, 根据拉曼特征峰的峰位和峰强度的差异, 首先对样品进行人工分成两类。采用 Fisher 判别分析(FDA)和 PCA 对实验数据进行分析。采用 FDA, 制定两个分类函数对人工分类结果进行判别, 分类总体准确率达到 88%。采用 PCA 对实验数据进行降维处理, 减少冗余数据对实验结果的影响。将原始数据、FDA 处理数据和 PCA 处理数据结合人工神经网络-多层感知器/径向基函数构建分类模型, 在基于多层感知器的人工神经网络模型中, 原始数据、FDA 处理数据和 PCA 处理数据的分类准确率分别为 80%、92% 和 96%, 样本分类的总体准确率为 89.33%。在基于径向基函数的人工神经网络模型中, 样本分类准确率分别为 76%、84% 和 92%, 样本分类总体准确率为 84%。差分拉曼光谱可对 25 种合成大麻素进行区分, 最后将样品分为两类。PCA 结合基于多层感知器的人工神经网络模型对光谱数据的分类效果最好, 总的来说是一种操作简单、检测效率高、结果准确的合成大麻素快速检测方法。

Tian 等[28] 基于移激发拉曼差分光谱(SERDS)结合机器学习算法建立一种快速、无损、准确的新型精神活性物质检测分析分类方法, 可以很好地区分芬太尼、安非他明和合成大麻素。该研究在激发光源(785 和 785.5 nm)的实验条件下, 采用 SERDS 检测芬太尼、安非他明和合成大麻素等 37 种新精神活性物质。提取其特征峰, 并将特征峰归属于物质的结构。同时将 SERDS 与机器学习结合使用, 以寻找最佳的分类预测模型。比较了 SVM、KNN、集成分类器、ANN、决策树、朴素贝叶斯和线性判别分析的分类效果, 并给出了 3 种超参数优化方法。最后, 贝叶斯优化下的 SVM 交叉验证准确率为 97.3%, 能够很好地区分 3 大类样本。可以有效地为海关、医疗、现场警务、大型事件安全、痕迹证据检测等提供解决方案。

4　机器学习结合合成大麻素的代谢组学的鉴定研究

PCA 和 RF 的算法模型已经应用于合成大麻素类物质的鉴定研究。代谢组学旨在捕捉外部刺激对内源性代谢物的影响, 可分为靶向代谢组学和非靶向代谢组学, 其中非靶向代谢组学无偏向性地对所有小分子代谢物同时进行检测分析。非靶向代谢组学研究的主要困难之一是处理产生的庞大数据集, 机器学习算法可以检测和学习大型高维数据集中的模式, 为非靶向代谢组学提供了新的策略。除应用代谢组学技术研究滥用药物的毒性作用机制之外, 也有学者提出了用代谢组学的方法预测新精神活性物质的药理学特征, 创新的代谢组学分析已应用于合成大麻素鉴定研究。

Streun 等[29] 将经典的机器学习算法 RF 和代谢组学相结合来筛选尿液中的合成大麻素, 根据吸食合成大麻素引起的特异性和可测量的尿代谢组学变化对尿液样本进行分类, 可达到 88.1% 的分类准确率。Olesti 等[30] 开发了一种新的药理学分析模型, 该模型采用基于大鼠单胺类神经递质和类固醇激素定量的药物特异性代谢组学指纹来预测新精神活性物质与特定药物类别的相似性, 根据药物与经典滥用药物的药理相似性对新药进行分类。该方法有可能通过促进快速的药物类型分类, 同时减少与滥用这些新兴药物有关的公众可能造成的伤害, 从而有利于风险评估政策。通过比较代谢组学指纹和它们在 PCA 中的接近性, JWH-018 被预测为 Δ^9-THC-like 化合物, 这与药物的药理学一致, 主要与刺激内源性大麻素受体 CB1 和 CB2 有关。

5 机器学习用于合成大麻素构效关系的鉴定研究

定量构效关系(quantitative structure - activity relationship , QSAR)研究[31]是以数学和统计学手段建立起化合物的化学结构和生物活性之间定量关系的模型。多元线性回归算法和偏最小二乘回归算法模型常用于合成大麻素构效关系鉴定研究。通常无法人为判断化合物化学结构和生物活性之间的定量关系,通过机器学习可以形成规律性基础,成为构-效鉴定的关键客观评价指标。通过 QSAR 研究可以对新型未知的合成大麻素的毒性进行预测,化合物保留时间特征和其结构相关关系的 QSAR 研究可以推测未知化合物的色谱和质谱信息,为进一步识别合成大麻素活性物质提供基础。目前已有研究利用 QSAR 模型研究合成大麻素和 CB1 受体和 CB2 受体相结合的化合物的相关性质。

Lee 等[32]建立一个 QSAR 模型,将各种合成大麻素的结构和理化性质与其 CB1 受体结合亲和力联系起来。该研究基于四氢大麻酚和 14 种合成大麻素与 CB1 受体亲和力的数据,使用 R/CDK 工具包计算数据集化合物的分子描述符将化合物的简化分子线性输入规范(simplified molecular input line entry system, SMILES)表示转换为分子指纹,使用多元线性回归算法和偏最小二乘回归算法构建 QSAR 回归模型。通过 Y 随机化检验和外部验证,获得最优模型。该模型可在体内应用于预测非法新的合成大麻素的成瘾性,为预测合成大麻素的滥用提供了一种新的策略。Paulke 等[33]建立了 QSAR 模型,该模型可以在没有参比物质的情况下确定未知化合物对 CB1 的亲和力(以结合常数 K_i 表示)。采用化学高级模板搜索描述符对化合物结构进行向量表示,利用特征对分布相似度计算两个分子之间的相似度。K_i 采用反距离加权法(inverse distance weighting, IDW)计算,使用十倍交叉验证程序对预测模型进行验证。所建立的 QSAR 模型可以作为一种简单、快速、价廉的工具,用于初步了解新的合成大麻素或其他新的精神活性化合物的生物活性。

6 其他用于合成大麻素识别鉴定的技术

近红外光谱技术作为快速、无损地分析和检测复杂基质中不同化合物的首选工具,正在得到广泛应用。Risoluti 等[34]探讨了利用近红外光谱结合 PCA 的化学计量学方法检测缉获样品中新型精神活性物质的可行性,证实了该方法能够很好地对合成大麻素类和苯乙胺类物质进行区分,并成功地应用于“现场”缉获的真实样品中的非法药物,该方法有望成为法医科学中新的精神活性物质初步测定的快速,经济和有用的工具。

除质谱和光谱技术外,核磁共振技术(nucelear magnetic resonance, NMR)也常用于检测新型毒品以及未知毒物,能够高效地推测出检材中毒品的化学机构及信息,在合成大麻素的检测方面,有文献报道,使用 NMR 技术分析合成大麻素 UR-144 及其代谢物,实现了对其 10 种代谢物的结构解析[35]。另有文献使用 ^{19}F NMR 辅助 51 种含氟合成大麻素类物质的结构解析以及定量[36], ^{19}F NMR 具有无基质干扰的特点,对具有复杂基质的电子液体合成大麻素样品尤其有利,为缉获样品的绝对定量提供了一种合适的分析技术。但截至目前,还未见到运用机器学习结合 NMR 技术分析合成大麻素类物质的研究报道。图 2 对机器学习技术与其他技术结合用于合成大麻素鉴识常用算法方法比较及适用范围进行了比较。

7 总结与展望

机器学习算法能够从大量的数据中自动提取特征,高效地挖掘其中有价值的信息,在合成大麻素的鉴识领域,机器学习技术已经显示出巨大的潜力。本文介绍了在合成大麻素鉴识领域常用机器学习方法、算法。机器学习与质谱、拉曼光谱结合已用于合成大麻素结构和来源识别,与代谢组学和定量构效关系结合用于合成大麻素类物质鉴别。常用的机器学习算法有 RF、SVM、ANN,泛化能力强以及适用于处理高维度数据优势可能是大部分研究乐于使用这些算法的主要原因。在合成大麻素的识别和鉴定领域,基于机器学习的技术仍具有很大的应用潜力,可以有以下两个发展方向:

(1)注重数据资源整理。数据作为机器学习的基础,很大程度上决定了模型的准确性。当前主要的数据来源有数据库和实验室自研数据,对于数据库来说存在数据来源不同缺乏一致性的问题,可尝试将数据归一化、格式化处理,对于实验室自研数据来说,实验室积累了大量的数据,但这些数据隐

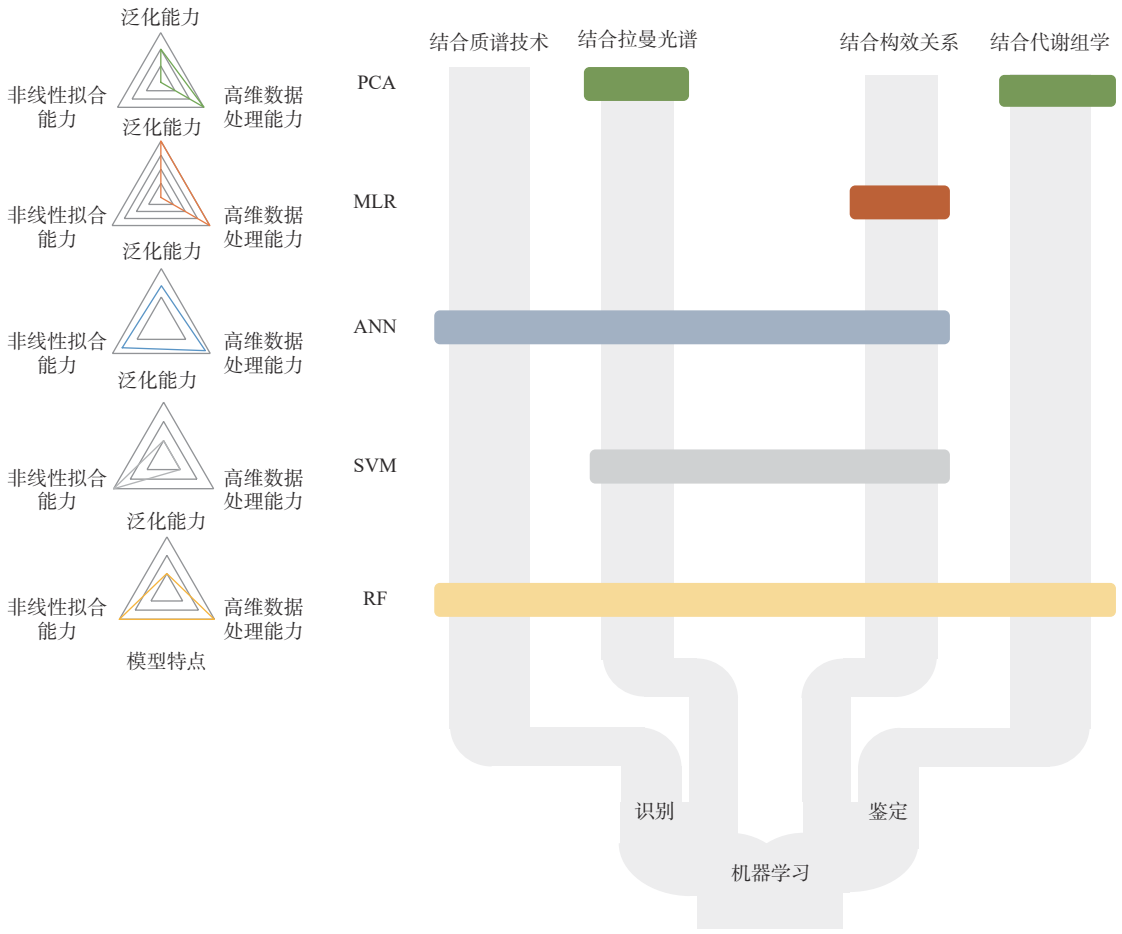

图 2　机器学习结合 4 种其他技术鉴识合成大麻素常用算法方法比较及适用范围

PCA: 主成分分析; MLR: 多元线性回归; ANN: 人工神经网络; SVM: 支持向量机; RF: 随机森林

匿于海量的文献中, 提取困难, 可以借助爬虫工具使用, 在使用的过程中应当注重伦理及数据获取的合法性, 不得侵害实验室利益。

(2)扩大模型的适用范围, 提高模型可解释性。将得到的模型应用到未知的质谱或光谱中检测感兴趣类别的物质, 尝试将机器学习技术与其他的仪器比如近红外光谱, 核磁共振等获得的数据相结合, 也可尝试将各种模型类型组合成一个整体模型, 进一步扩展到其他质谱或光谱技术中。提高模型的可解释性, 更有利于深入理解模型内部的工作原理, 从而提升模型的效果, 也可以更好地理解模型得到的结果。

References

[1] Wiley JL, Marusich JA, Huffman JW. Moving around the molecule: relationship between chemical structure and *in vivo* activity of synthetic cannabinoids[J]. *Life Sci*, 2014, **97**(1): 55-63.

[2] Schurman LD, Lu D, Kendall DA, *et al*. Molecular mechanism and cannabinoid pharmacology[J]. *Handb Exp Pharmacol*, 2020, **258**: 323-353.

[3] Alves VL, Gonçalves JL, Aguiar J, *et al*. The synthetic cannabinoids phenomenon: from structure to toxicological properties. A review[J]. *Crit Rev Toxicol*, 2020, **50**(5): 359-382.

[4] Alzu'bi A, Almahasneh F, Khasawneh R, *et al*. The synthetic cannabinoids menace: a review of health risks and toxicity[J]. *Eur J Med Res*, 2024, **29**(1): 49.

[5] Banister SD, Connor M. The chemistry and pharmacology of synthetic cannabinoid receptor agonist new psychoactive substances: evolution[J]. *Handb Exp Pharmacol*, 2018, **252**: 191-226.

[6] Tai S, Fantegrossi WE. Pharmacological and toxicological effects of synthetic cannabinoids and their metabolites[J]. *Curr Top Behav Neurosci*, 2017, **32**: 249-262.

[7] Fantegrossi WE, Moran JH, Radominska-Pandya A, *et al*. Distinct pharmacology and metabolism of K2 synthetic cannabinoids compared to Δ(9)-THC: mechanism underlying greater toxicity[J]. [J]? *Life Sci*, 2014, **97**(1): 45-54.

[8] Yan FR. Application and advance of artificial intelligence in biomedical field[J]. *J China Pharm Univ* (中国药科大学学报), 2023, **54**(3): 263-268.

[9] Wang C, Xiao F, Li M, *et al*. Application progress of artificial intelligence in the screening and identification of drug targets[J]. *J China Pharm Univ* (中国药科大学学报), 2023, **54**(3): 269-281.

[10] Yu ZH, Zhang LM, Zhang MN, *et al*. Artificial intelligence-based drug development: current progress and future challenges[J]. *J China Pharm Univ* (中国药科大学学报), 2023, **54**(3): 282-293.

[11] Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: a brief primer[J]. *Behav Ther*, 2020, **51**(5): 675-687.

[12] Ringnér M. What is principal component analysis[J]? *Nat Biotechnol*, 2008, **26**(3): 303-304.

[13] Gilbert N, Mewis RE, Sutcliffe OB. Classification of fentanyl analogues through principal component analysis (PCA) and hierarchical clustering of GC–MS data[J]. *Forensic Chem*, 2020, **21**: 100287.

[14] Jiménez-Carvelo AM, González-Casado A, Bagur-González MG, *et al*. Alternative data mining/machine learning methods for the analytical evaluation of food quality and authenticity - A review[J]. *Food Res Int*, 2019, **122**: 25-39.

[15] Amendolia SR, Cossu G, Ganadu ML, *et al*. A comparative study of K-nearest neighbour, support vector machine and multilayer perceptron for thalassemia screening[J]. *Chemom Intell Lab Syst*, 2003, **69**(1/2): 13-20.

[16] Broséus J, Anglada F, Esseiva P. The differentiation of fibre- and drug type *Cannabis* seedlings by gas chromatography/mass spectrometry and chemometric tools[J]. *Forensic Sci Int*, 2010, **200**(1/2/3): 87-92.

[17] Thijs B, AxelJan R, Melvin G, *et al*. Decision trees and random forests[J]. *Am J Orthod Dentofac Orthop Off Publ Am Assoc Orthod Const Soc Am Board Orthod*, 2023, **164**(6): 894-897.

[18] Winkler DA, Le TC. Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR[J]. *Mol Inform*, 2017, **36**(1/2): 10.1002/minf.201600118.

[19] Yang YQ, Liu DP, Hua ZD, *et al*. Machine learning-assisted rapid screening of four types of new psychoactive substances in drug seizures[J]. *J Chem Inf Model*, 2023, **63**(3): 815-825.

[20] Wong SL, Ng LT, Tan J, *et al*. Screening unknown novel psychoactive substances using GC-MS based machine learning[J]. *Forensic Chem*, 2023, **34**: 100499.

[21] Lee SY, Lee ST, Suh S, *et al*. Revealing unknown controlled substances and new psychoactive substances using high-resolution LC-MS-MS machine learning models and the hybrid similarity search algorithm[J]. *J Anal Toxicol*, 2022, **46**(7): 732-742.

[22] Koshute P, Hagan N, Jameson NJ. Machine learning model for detecting fentanyl analogs from mass spectra[J]. *Forensic Chem*, 2022, **27**: 100379.

[23] Moorthy AS, Kearsley AJ, Mallard WG, *et al*. Mass spectral similarity mapping applied to fentanyl analogs[J]. *Forensic Chem*, 2020, **19**. doi: 10.1016/j.forc.2020.100237.

[24] Setser AL, Waddell Smith R. Comparison of variable selection methods prior to linear discriminant analysis classification of synthetic phenethylamines and tryptamines[J]. *Forensic Chem*, 2018, **11**: 77-86.

[25] Bonetti JL, Samanipour S, van Asten AC. Utilization of machine learning for the differentiation of positional NPS isomers with direct analysis in real time mass spectrometry[J]. *Anal Chem*, 2022, **94**(12): 5029-5040.

[26] Münster-Müller S, Matzenbach I, Knepper T, *et al*. Profiling of synthesis-related impurities of the synthetic cannabinoid Cumyl-5F-PINACA in seized samples of e-liquids via multivariate analysis of UHPLC-MS^n data[J]. *Drug Test Anal*, 2020, **12**(1): 119-126.

[27] Lee J, Jiang H. Analysis of indole and indazole amides synthetic cannabinoids by differential Raman spectroscopy based on ANN[J]. *J Forensic Sci*, 2022, **67**(6): 2242-2252.

[28] Tian LC, Jiang H, Chen TZ. A rapid and nondestructive approach for forensic identification of novel psychoactive substances using shifted-excitation Raman difference spectroscopy- and machine learning[J]. *J Raman Spectrosc*, 2023, **54**(5): 540-550.

[29] Streun GL, Steuer AE, Poetzsch SN, *et al*. Towards a new qualitative screening assay for synthetic cannabinoids using metabolomics and machine learning[J]. *Clin Chem*, 2022, **68**(6): 848-855.

[30] Olesti E, De Toma I, Ramaekers JG, *et al*. Metabolomics predicts the pharmacological profile of new psychoactive substances[J]. *J Psychopharmacol*, 2019, **33**(3): 347-354.

[31] Khan K, Benfenati E, Roy K. Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds[J]. *Ecotoxicol Environ Saf*, 2019, **168**: 287-297.

[32] Lee W, Park SJ, Hwang JY, *et al*. QSAR model for predicting the cannabinoid receptor 1 binding affinity and dependence potential of synthetic cannabinoids[J]. *Molecules*, 2020, **25**(24): 6057.

[33] Paulke A, Proschak E, Sommer K, *et al*. Synthetic cannabinoids: *in silico* prediction of the cannabinoid receptor 1 affinity by a quantitative structure-activity relationship model[J]. *Toxicol Lett*, 2016, **245**: 1-6.

[34] Risoluti R, Materazzi S, Gregori A, *et al*. Early detection of emerging street drugs by near infrared spectroscopy and chemometrics[J]. *Talanta*, 2016, **153**: 407-413.

[35] de Castro JS, Rodrigues CHP, Bruni AT. *In silico* infrared char-

acterization of synthetic cannabinoids by quantum chemistry and chemometrics[J]. *J Chem Inf Model*, 2020, **60**(4): 2100-2114.

[36] Liu CM, Song CH, Jia W, *et al*. The application of ^{19}F NMR spectroscopy for the analysis of fluorinated new psychoactive substances (NPS)[J]. *Forensic Sci Int*, 2022, **340**: 111450.

［专家介绍］陈航, 司法鉴定科学研究院副主任法医师, 硕士生导师, 国际法医毒理家协会(The International Association of Forensic Toxicologists, TIAFT)会员, 入选上海市青年科技英才计划。主要从事法医毒物学研究及基于应用研究的司法鉴定公共法律服务。主持或参与含“十二五”“十三五”国家重点研发专项在内的多项国家级、省部级科研项目, 曾作为学术秘书参与编制“十三五”国家规划高等院校教材《法医毒物学》及配套材料, 参编《法医毒物学手册》《法医毒物鉴定理论与实践》《滥用物质分析与应用》《毛发分析基础及应用》《新精神活性物质分析与应用》等专著, 开发并登记包括《司法鉴定材料管理信息化系统(FSMS V1.0)》《法医毒物学化合物知识库系统 V2.0》《法医毒物数字化平台 V2.0》等数字化软件。

· 校园信息 ·

本刊副主编孔令义/张超团队在*Molecular Cancer*发表最新研究成果

近日, 本刊副主编孔令义/张超团队在学科顶尖期刊 *Molecular Cancer* 在线发表了题为“The incorporation of acetylated LAP-TGF-β1 proteins into exosomes promotes TNBC cell dissemination in lung micro-metastasis”的最新研究成果。中药学院博士后余培、2019 级硕士毕业生韩玉豹为本文共同第一作者, 孔令义教授和张超副研究员为本文共同通讯作者, 中国药科大学为本文第一通讯单位。

三阴性乳腺癌(TNBC)是复发率和死亡率最高的乳腺癌亚型, 肺部是其常见的转移部位。大量证据表明, TNBC 在肺部的转移经常以 TNBC 肺部微转移灶细胞的重复扩散形成(由一变多), 而不是直接来源于乳腺癌原发部位的肿瘤细胞的多次转移。TNBC 肺部微转移性肿瘤灶能够通过改变血管微环境为大量肺转移性肿瘤灶的形成提供有利的“土壤”。外泌体是细胞间通讯的重要介质之一, 肿瘤来源外泌体的“分子货物”(包括蛋白质和核酸)可以调节细胞行为, 但 TNBC 外泌体的“分子货物”是如何影响血管微环境并促进肺部大量转移性肿瘤灶形成的分子细节仍然难以捉摸。

该研究揭示了 TNBC 外泌体负载 LAP-TGF-β1 在重塑肺血管生态位, 促进 TNBC 肺转移中的关键作用。尽管各种阻断 TGF-β 信号通路的策略已经开发出来并取得了临床进展, 但其严重的副作用限制了其临床应用。该研究发现, 在肺转移部位, TNBC 外泌体上的 LAP-TGF-β1 可以在远低于游离型 TGF-β1 的剂量下显著重塑肺血管生态位, 促进 TNBC 细胞在肺部的外渗和定植。进一步的研究发现在 LAP-TGF-β1 的 TGFB1 区域存在一个非经典的 KFERQ 样序列, 其 K304 位点会发生乙酰转移酶 TIP60 介导的乙酰化, 促进与 HSP90A 的相互作用而转运至外泌体。同时抑制 HSP90A 和 TIP60 可显著降低 LAP-TGF-β1 的外泌体负载量, 有效抑制 TNBC 肺转移。该研究不仅为 TNBC 肺转移的分子基础提供了新的见解, 而且为开发新型的治疗策略奠定了基础。

该研究工作得到了国家自然科学基金项目和江苏省卓越博士后计划的资助。

(供稿单位: 中药学院)