第39卷，第7期
2019年7月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol 39，No. 7，pp2171-2175
July，2019

机器学习算法用于公安一线拉曼实际样本采样学习及其准确度比较

李志豪，沈 俊*，边瑞华，郑 健

公安部第三研究所，上海 200031

摘 要 拉曼光谱设备在公安一线中正逐渐得到普及，主要用于检测易燃易爆及易制毒化学品。但在实际应用中，一线人员不会对拉曼设备进行非常准确的使用和操作，不具备专业知识条件的工作人员无法完全按照最佳条件进行检测，经常会发生离焦、偏移、采样时间过短等一系列问题，而检测结果也不可能完全符合标准测试库的算法，给最终结果比对造成非常大的影响。利用五种主流机器学习算法对实际检查、办案过程中采集到的原始数据进行学习分类，通过比较相应的准确度将最佳算法用于改善一线执法、检查过程中拉曼光谱设备的准确性。采集的数据均来自于公安部第三研究所自行研制的EVA3000型拉曼光谱仪，该光谱仪目前已在全国各省、市、地、县进行了一定的配备，一线检测人员会定期将采集的原始数据回传到EVA3000的后台管理系统。通过该管理系统，在线收集实际检查过程中产生的原始数据，以两类易制毒化学品和易燃易爆化学品为例，随机抽取已定性判定的苯乙酸、二氯甲烷、麻黄碱和硝基苯各40例共计160例，并分别利用决策树、随机森林、AdaBoost、支持向量机和人工神经网络算法各进行40，60，100，150，200，300和500次的交叉训练、预测、求取平均准确度。从实验结果可以看出，在五种学习算法中，对于实际样本的预测准确度排序大致为随机森林≈AdaBoost＞决策树＞SVM＞人工神经网络。实际测试的结果与实验过程中的平均预测准确度大体一致。其中随机森林与AdaBoost的准确度相近，其原因在于两者的算法本质都是不断构建新的训练数据集并提高对于错误样本在下次学习中的权重，而SVM和人工神经网络算法的本质都是基于感知器的算法。可见目前几种主流学习算法中，采用自举汇聚(bootstrap aggregating)方式的算法更适应于对实际样本的采样学习，其准确度也较高。在下一步的工作当中，将继续优化现有的算法，将其实现在后台管理系统上，并测试算法对于目前检测中无法定性物质的在线检测功能。该结果对于进一步将机器学习算法用于实际应用、在线分析，改善一线操作人员非正确使用设备对比对结果造成影响，具有重要意义。

关键词 拉曼光谱；易燃易爆及易制毒化学品；决策树；随机森林；Adaboost；神经网络；支持向量机；公安一线

中图分类号：TP39 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)07-2171-05

引 言

激光拉曼光谱分析技术无需样品制备过程，无损探测、适合水溶液分析、速度快，已在反恐、禁毒领域逐步开始应用，公安、消防、重要场所、盛会安保几乎都配备了拉曼光谱设备来判别爆炸物、毒品等危险物品[1]。但是在实际应用中，一线人员不会与实验室专业人员一样对拉曼设备进行非常准确的使用和操作，经常会发生离焦、偏移、采样时间过短等各种各样的问题，给最终的结果比对造成非常大的影响。

在之前的工作当中，公安部第三研究所已经专门针对易制毒化学品，分别对拉曼设备的最优检测条件进行了研究，包括聚焦位置、分析时间、容器以及外部光源[2]；上海理工大学的蒋林华等人也专门研究了利用主成分分析结合支持向量机的方法[2]，对拉曼光谱进行分析。但实际场景下，在一线卡口、检查站或现场检测当中，不具备专业知识条件的工作人员不可能完全按照最佳条件进行检测，检测的结果也不可能完全符合标准测试库的算法。针对以上问题，通过由公安部第三研究所自行主导研发的“拉曼数据研判平台”，采集

收稿日期：2018-06-10，**修订日期**：2018-10-22

基金项目：国家“十三五”重点研发计划项目(2016YFC0801304)资助

作者简介：李志豪，1985年生，公安部第三研究所助理研究员 e-mail：lizhihao559@hotmail.com

*通讯联系人 e-mail：63368207@qq.com

一线工作人员上传的其使用过程中得到的原始数据，以两类易制毒化学品和易燃易爆化学品(苯乙酸、麻黄碱、二氯甲烷和硝基苯)为例，分别利用决策树、随机森林、Adaboost、支持向量机和神经网络算法对数据结果进行样本学习和分类，并对其准确度进行比较，以期将最适合的算法体系用于研判平台，对于后期一线人员单机设备中难以匹配测试的样本将由后台的学习结果进行在线测试。

1 算法原理

1.1 决策树

决策树算法通常采用二分法划分数据，本文采用 ID3 算法，对划分数据集前后发生的增益进行计算，获得增益最高的代表确定性最高，其特征将作为下一节点继续进行迭代。

具体地讲，在决策树算法中，用一种称为“熵”的测量方法来表示特征的不确定性，其数学表达式如式(1)

$$H = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{1}$$

其中 n 是分类的数目，$p(x_i)$是选择该分类的概率。而信息的增益，则是用来度量熵的减少，数学表达式如式(2)

$$IG(T,a) = H(T) - \sum \frac{|\{x \in T \mid x_a = v\}|}{|T|} \times H(\{x \in T \mid x_a = v\}) \tag{2}$$

其中 $IG(T,a)$为信息增益，T 为一组实例而 a 表示被测试的特征。$H(T)$为父节点的熵，$\sum \frac{|\{x \in T \mid x_a = v\}|}{|T|} H(\{x \in T \mid x_a = v\})$ 为子节点的加权平均熵。

通过遍历当前特征中的所有唯一属性值，对每个特征划分一次数据集，然后计算数据集的新熵值，并对所有唯一特征值得到的熵求和，无序度减少的表征在于熵的减少，也就是信息增益的最佳方向。最后，比较所有特征中的信息增益，返回最佳特征划分。

1.2 随机森林

随机森林属于集成算法或者称为元算法，是在决策树算法上的一种升级。其本质在于对原始训练数据集中的样本进行随机抽取形成一个大小相等的新数据集，并反复将新数据集与原始数据集中的样本进行替换，不断形成大小一致的新数据集，这个过程称为自举汇聚(bootstrap aggregating)[3]。通过自举重采样过程可以得到多组新的训练数据集，分别对其使用决策树算法进行分类，这样就得到了与新数据集数量相当的新分类器。当需要进行分类预测时，随机森林将分别使用训练过程中得到的多组分类器分别进行预测，并选择分类器投票结果中最多的类别作为最后的结果。随机森林中决策树的数量是一个重要参数，通过增加决策树的数量可以提高模型的性能，但将以计算成本作为代价。

1.3 AdaBoost

Boosting 是一种与自举汇聚类似的技术，但其更关注于被已有分类器错分的那些数据，其本质在于提高错分数据的权重，增加其出现在通过自举汇聚形成的新数据集中的概率，并训练以获得新的分类器。AdaBoost(自适应 boosting)算法属于其中一类，并且目前被大多数人认为属于当前已知机器学习算法中的佼佼者[5]。在本文中，第一次迭代过程 AdaBoost 分配相等的权重给所有训练数据集，如同自举汇聚随机形成新的训练数据集，并利用 ID3 决策树进行分类。在随后的迭代中，不断增加之前训练结果中分类错误数据的权重，其将有更大概率出现在自举汇聚形成的新数据集中，并利用 ID3 决策树进行分类。

具体的，首先对每个数据赋予一个权重以 α 表示，计算公式如式(3)

$$\alpha = \frac{1}{2}\ln\left(\frac{1-\varepsilon}{\varepsilon}\right) \tag{3}$$

其中，ε 为错误率，定义为 ε=(包含该数据并且未正确分类的样本数目)/(所有样本数目)，所有数据的权重 α 构成向量 D。当经过 i 次迭代后，被正确分类数据的权重在向量 D 中更新为

$$D_{i+1} = \frac{D_i e^{-\alpha}}{\mathrm{Sum}(D)} \tag{4}$$

而未被正确分类数据的权重在向量中更新为

$$D_{i+1} = \frac{D_i e^{\alpha}}{\mathrm{Sum}(D)} \tag{5}$$

1.4 支持向量机

本文拉曼光谱数据属于 2 901 维的数据，因此以支持向量机(SVM)进行分类需要使用 2 901 维的超平面。SVM 通过一个非线性映射 ϕ 将样本空间映射到高维特征空间，使样本空间的非线性分类转化为线性分类，并基于结构风险最小化在特征空间中寻找最优超平面，解决线性分类问题。在二维平面上的超平面公式表达为式(6)

$$f(x) = \langle w,x\rangle + b \tag{6}$$

通过求解向量 w 和偏移量 b 可以得到分割平面。进而对应高维向量求解时，式(6)改为

$$f(x) = \sum \alpha_i y_i \langle \phi(x_i),\phi(x)\rangle + b \tag{7}$$

其中 $\phi(x)$为将样本 x 映射到高维特征空间的转换函数，$\langle \phi(x_i),\phi(x)\rangle$为映射的内积，$\alpha_i$ 为支持值代表向量 w 对应映射高维特征空间的权重值，b 为偏移量。但是这种映射的计算成本非常高，注意到将特征向量映射到高维空间时，特征向量仅作为点积出现，而点积是一个标量，一旦计算出标量，就不需要映射的特征向量。为此引入内核的计算，内核是给定原始特征向量的函数，返回与其对应的映射特征向量的点积相同的值，表达式如式(8)

$$K(x,z) = \langle \phi(x),\phi(z)\rangle = \langle (x_1^2,x_2^2,\sqrt{2}x_1x_2),(z_1^2,z_2^2,\sqrt{2}z_1z_2)\rangle \tag{8}$$

其中 x，z 分别为高维空间映射中的特征向量。进一步使用 RBF 核函数[4]，利用最小二乘支持向量机优化计算成本，其判别函数为

$$y(x) = \mathrm{sign}\left[\sum_{k=1}^{N} \alpha_k y_k K(x,z) + b\right] \tag{9}$$

1.5 神经网络

本文使用的神经网络是多层感知器结构，每一个感知器都使用类似于 1.4 节中的超平面分割算法。具体实现过程中，首先定义感知器(神经元)响应函数，并使用梯度下降法

来最小化含有多个变量的实值函数 C。简单起见，假设 C 含有两个变量 v_1 和 v_2，而 Δv_1，Δv_2 和 ΔC 分别表示 v_1，v_2 以及 C 的变化量，那么 ΔC 可以表示为式(10)，而梯度则如式(11)，在每次迭代中，ΔC 为负则能保证成本函数的降低，因此假设$(\Delta v_1,\Delta v_2)^{\mathrm{T}}$ 如式(12)

$$\Delta C \approx \frac{\partial C}{\partial v_1}\Delta v_1 + \frac{\partial C}{\partial v_2}\Delta v_2 \tag{10}$$

$$\Delta C = \nabla C(\Delta v_1,\Delta v_2)^{\mathrm{T}} \tag{11}$$

$$(\Delta v_1,\Delta v_2)^{\mathrm{T}} = -\mu\nabla C \tag{12}$$

μ 为学习速率，并为多层结构中的每一个神经元单元配以权重。权重的计算首先利用反向传播算法来计算神经网络的成本函数相对于其权重的梯度。反向传播是一个迭代算法，每次迭代由两个阶段组成。第一阶段是前向传播，在正向通道中，对神经元层进行输入后传播，直到它们到达输出层，并利用损失函数来计算预测的误差。第二阶段是反向传播，成本函数从后向反向传播入神经元层，修正该层的误差。迭代过程中的误差修正函数如式(13)

$$E = \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{13}$$

其中 n 是神经元单元数量，$\hat{y}_i$ 是神经元的激活输出变量(输出)，y_i 是神经元的响应变量(输入)。通过误差修正函数进行神经元层函数的修正后再次进行传播[6]。

2　实验部分

2.1　测试设备

采集的原始数据均来自于由公安部第三研究所自行研制的 EVA3000 型便携式拉曼光谱仪，光谱范围 200～3 100 cm^{-1}，光源波长(785±0.5) nm，线宽<0.08 nm，波长分辨率 18 cm^{-1}，最大激发功率 500 mW，积分时间 8 ms～2 min，探头工作焦距 7.5 mm。

由于数据完全由一线检查人员自行采集，不存在特定的实验环境、实验操作方法、样品制备等过程，完全根据实际情况而定。

2.2　数据来源

公安部第三研究所为自行研发的 EVA3000 型激光拉曼设备配备了强大的后台支持系统“拉曼数据研判平台”。一线拉曼设备测试后的原始数据将定期进行回传，平台整合了用户数据库、拉曼标准库、用户管理和设备管理功能。系统统计了全国各省市、地、县等配备了 EVA3000 型拉曼设备在查缉期间获得的原始数据，包括冰毒、海洛因、TNT、硝酸甘油、苯乙酸、麻黄碱、二氯甲烷和硝基苯等 160 种易制毒及易燃易爆化学品数据，如图 1(a)和(b)，这些数据完全由一线人员在查缉、办案过程中采集，对于本文希望利用机器学习对实际样本进行训练具有充分的的代表性，从中选取了已完全定性判定且较为典型的四类物质：苯乙酸、二氯甲烷、麻黄碱和硝基苯作为学习目标。

(a) Home page of raman spectra resarch platform

(b) Data uploaded

图 1　拉曼数据平台及上传数据内容

(a)：拉曼数据研判平台主页面；(b)：上传数据内容

Fig. 1　Home page of Raman spectra research platform and data uploaded

(a)：Home page of Ramanspectra research platform；(b)：Data uploaded

2.3　测试结果

分别从上传的实际样本中随机抽取已定性判定上传的苯乙酸、二氯甲烷、麻黄碱和硝基苯各 40 例，分别利用决策树、随机森林、AdaBoost、支持向量机、神经网络进行 40，60，100，150，200，300 和 500 次训练预测并求取平均准确度，每次的训练、测试样本分别从 160 例实际样本中进行随机分割，图 2 是各自的拉曼光谱图，最终的学习与预测结果如表 1 所示。

表 1　5 种算法在不同训练次数下的平均准确率

Table 1　Average accuracy of five algorithms with different training numbers

	决策树	随机森林	AdaBoost	支持向量机	神经网络
40 次训练预测平均准确度	0.836	0.870	0.823	0.793	0.704
60 次训练预测平均准确度	0.839	0.877	0.876	0.730	0.689
100 次训练预测平均准确度	0.855	0.879	0.879	0.757	0.694
150 次训练预测平均准确度	0.891	0.909	0.870	0.792	0.714
200 次训练预测平均准确度	0.872	0.909	0.907	0.846	0.714
300 次训练预测平均准确度	0.874	0.909	0.896	0.732	0.736
500 次训练预测平均准确度	0.897	0.909	0.901	0.789	0.725

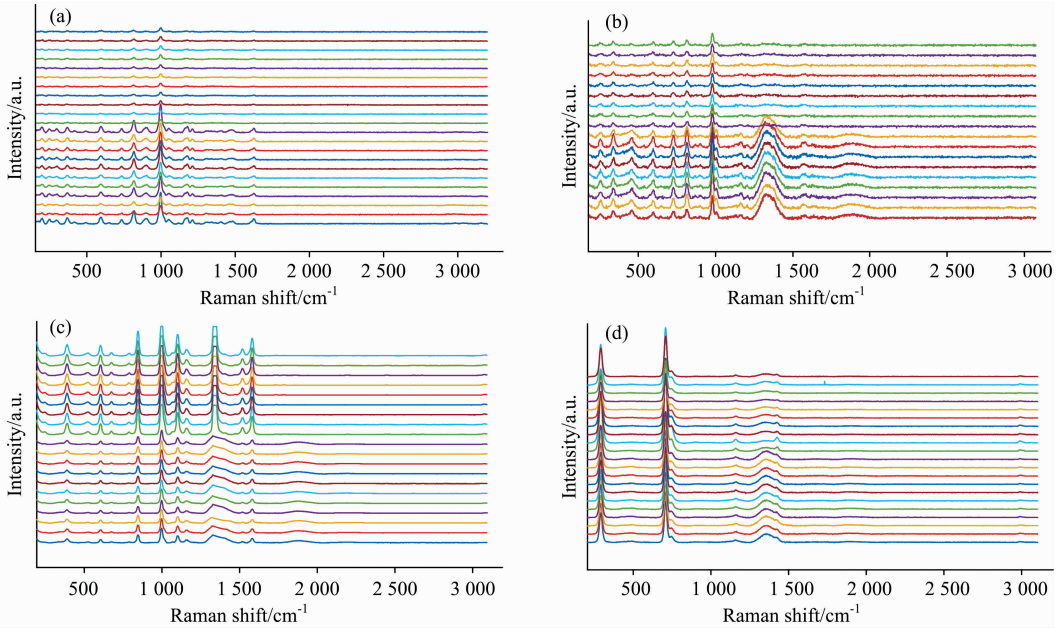

图 2 （a）麻黄碱；（b）苯乙酸；（c）硝基苯；（d）二氯甲烷

Fig. 2 （a）Ephedrine；（b）Phenylacetic acid；（c）Nitrobenzene；（d）Dichloromethane

从测试结果可以看出，在实际操作中的拉曼光谱受到各种不同因素影响会造成光谱的难以识别，利用机器学习算法可以有效提高识别准确度，通过对比 5 种主流算法可以发现利用随机森林进行学习验证的准确度可以接近 90%。

表 2 对 130 种实际样品的预测结果

Table 2 Prediction results of 130 actual samples

	麻黄碱	苯乙酸	硝基苯	二氯甲烷
实际样本	43	35	23	29
决策树	38	32	20	26
随机森林	39	33	21	27
AdaBoost	39	32	21	27
SVM	34	28	18	23
神经网络	31	26	17	21

表 3 对 130 种实际样本的准确率

Table 3 Accuracy of another 130 actual samples

	麻黄碱	苯乙酸	硝基苯	二氯甲烷
决策树	0.884	0.914	0.870	0.897
随机森林	0.907	0.943	0.913	0.931
AdaBoost	0.907	0.914	0.913	0.931
SVM	0.791	0.800	0.783	0.793
神经网络	0.721	0.743	0.739	0.724

2.4 实际效果测试

为进一步验证 5 种算法对于实际测试结果预测的准确度，从数据平台中再次随机抽取 130 例测试样本结果进行检验，翻验抽取的 130 例样本，其中包含麻黄碱 43 例，苯乙酸 35 例，硝基苯 23 例，二氯甲烷 29 例，学习样本为之前提取出的整个 160 例数据集，不再对学习样本进行分割。各种算法的准确预测结果数如表 2 所示，算法准确率如表 3 所示。

3 结 论

采用决策树、随机森林、AdaBoost、SVM 和 ANNs 对实际采集的 160 例样本数据麻黄碱、苯乙酸、硝基苯和二氯甲烷进行学习，并将学习结果用于预测随机抽取的另 130 例实际样本。从实验结果可以看出，在五种学习算法中，对于实际样本的预测准确度排序大致为随机森林≈AdaBoost＞决策树＞SVM＞神经网络。实际测试的结果与实验过程中的平均预测准确度大体一致。其中随机森林与 AdaBoost 的准确度相近，其原因在于两者的算法本质都是不断构建新的训练数据集并提高对于错误样本在下次学习中的权重，而 SVM 和神经网络算法的本质都是基于感知器的算法。

可见目前几种主流学习算法中，采用自举汇聚(bootstrap aggregating)方式的算法更适应于对实际样本的采样学习，其准确度也较高，在下一步的工作当中，将继续优化现有的算法，将其实现在“拉曼数据研判系统”上，并测试算法对于目前检测中无法定性物质的在线检测功能。

References

[1] CHENG Cheng(成 诚). 2016 Paper Collection for Seminar on Infrared and Remote Sensing Technology and Applications(2016 年红外、遥感技术与应用研讨会暨交叉学科论坛论文集), 2016. 84.

[2] JIANG Lin-hua, SHEN Jun, YU Zhi-hao, et al(蒋林华, 沈 俊, 余治昊, 等). Optical Instruments(光学仪器), 2018, 40(2): 31.

[3] Rafael Pino-Mejías, María-Dolores Jiménez-Gamero, María-Dolores Cubiles-de-la-Vega, et al. Pattern Recognition Letters, 2008, 29(3): 265.

[4] Suykens J A K, Vandewalle J. Neural Processing Letters, 1999, 9(3): 293.

[5] Rubin Daniel B. Statistical Applications in Genetics and Molecular Biology, 2011, 10(1): 54.

[6] Srinivas Y, Stanley Raj A, Hudson Oliver D, et al. Geoscience Frontiers, 2012, 3(5): 729.

Accuracy Comparison of the Machine Learning Algorithm Used to Raman Real Sample Collection in the Front Line of Public Security

LI Zhi-hao, SHEN Jun*, BIAN Rui-hua, ZHENG Jian

The Third Research Institute of Ministry of Public Security, Shanghai 200031, China

Abstract Raman spectroscopy equipment comes into use in the front line of public security gradually, which is mainly used for the detection of inflammable,explosive and easily-made drug chemicals. However, workers without professional knowledge may not be able to perform detection in full accordance with the best conditions. Frequent problems such as defocusing, offsetting and short sampling time may cause a great influence on the final comparison. In this article, five mainstream machine learning algorithms were used to train and classify the original data collected during the actual inspection and handling of the case. Also, the accuracy comparisons was given in this paper. According to the result, algorithm with the best accuracy will be used to improve the Raman spectroscopy in the future. The collected data were all from the EVA3000 Raman spectrometer developed by the Third Research Institute of the Ministry of Public Security. The spectrometer had been equipped in certain provinces, cities, prefectures and counties across the country. Front-line inspection personnel would periodically transmit the raw data back to the EVA3000's back office management system. Through the management system, the raw data generated during the actual inspection was collected. A total of 160 cases including phenylacetic acid, methylene chloride, ephedrine and nitrobenzene, which had been qualitatively determined, were randomly extracted from the uploaded database. The 40-, 60-, 100-, 150-, 200-, 300-, 500-time trainings and predictions with decision trees, random forests, AdaBoost, support vector machines and artificial neural networks were executed to calculate average accuracy respectively. From the experimental results, we can see that among the five learning algorithms, the ranking of the prediction accuracy to actual samples is roughly random forest≈AdaBoost>decision tree >SVM>ANNs. The verification results are generally consistent with the experimental ones. The accuracy of random forest is similar to AdaBoost because both algorithms constantly build new training data sets from the original ones and improve the weight of the wrong samples in the next training. On the other hand, SVM and ANNs are perceptron-based algorithms. It can be seen that in the current mainstream algorithms, bootstrap aggregating method is more suitable for the sampling training of actual samples. In the next step, the research team will continue to optimize existing algorithms and implement them in the back office management system for on-line detection. The results of this paper are of great significance for further using machine learning algorithms to the practical applications in the field of the front line of public security.

Keywords Raman spectroscopy; Flammable and explosive chemicals; Easily-made drug chemicals; Decision tree; Random forest; Adaboost; SVM; ANNs; Public security

(Received Jun. 10, 2018; accepted Oct. 22, 2018)

* Corresponding author