移动扫码阅读

景德基,程娜娜,蔡兴农,等. 基于机器学习的典型制药企业工艺过程 VOCs 排放特征因子识别[J].能源环境保护,2022,36(1):77-82.

JING Deji,CHENG Nana,CAI Xingnong,et al. The identification of VOCs emission characteristic factors in production process of a typical pharmaceutical factory based on machine learning[J]. Energy Environmental Protection,2022,36(1):77-82.

基于机器学习的典型制药企业工艺过程 VOCs 排放特征因子识别

景德基[1],程娜娜[1],蔡兴农[1],石展宏[1],杨春亚[1],
李素静[1],王俏丽[2,*],李 伟[1,*]

(1.浙江大学 化学工程与生物工程学院,浙江 杭州 310007;2. 浙江工业大学 环境学院,浙江 杭州 310014)

摘要:制药行业生产工艺复杂,VOCs 排放显著,是实施工业 VOCs 减排的重点行业。为落实制药行业 VOCs 减排策略,需准确识别重点排放企业和工艺过程。基于精细化工园区典型化学合成制药企业 VOCs 污染源成分谱,结合特征选择、分类分析、聚类分析等机器学习手段,进行了 VOCs 特征因子识别。结果表明:该企业 VOCs 排放的特征因子为甲苯、丙酮、乙醛、苯甲醛和正己烷;机器学习手段所识别的特征因子数量精简,在各个生产过程具有相似的浓度分布,体现了污染源 VOCs 排放物种组成上的差异。

关键词:制药企业;VOCs;特征因子;机器学习

中图分类号:X701 **文献标识码**:A **文章编号**:1006-8759(2022)01-0077-06

The identification of VOCs emission characteristic factors in production process of a typical pharmaceutical factory based on machine learning

JING Deji[1],CHENG Nana[1],CAI Xingnong[1],SHI Zhanhong[1],YANG Chunya[1],
LI Sujing[1],WANG Qiaoli[2,*],LI Wei[1,*]

(1.*College of Chemical and Biological Engineering,Zhejiang University,Hangzhou* 310007,*China*;
2.*College of Environment,Zhejiang University of Technology,Hangzhou* 310014,*China*)

Abstract:Pharmaceutical industry,which has complex production processes and a serious VOCs emission problem,is a key industry for implementing VOCs emission reduction. To implement VOCs emission reduction,it is necessary to accurately identify key companies and production processes. Based on the VOCs source profiles of a typical chemical synthetic pharmaceutical factory in a fine chemical industrial park,the VOCs characteristic factors identification were carried out based on machine learning methods such as feature selection,classification analysis,and cluster analysis. The results showed that the VOCs emission characteristic factors were identified as toluene,acetone,acetaldehyde,benzaldehyde and n-hexane. The identified characteristic factors were obviously simplified by machine learning methods,and had similar concentration distribution in each production process. They could reflect the differences in the species composition of VOCs emission from pollution sources.

Key Words:Pharmaceutical factory;VOCs;Characteristic factors;Machine learning

收稿日期:2021-10-27;**责任编辑**:金丽丽

基金项目:浙江省重点研发计划项目(2021C03178,2021C03165)

第一作者简介:景德基(1997-),男,贵州黔东南州人,学士,研究生在读,主要研究方向为小尺度工业地区大气污染溯源研究。E-mail:deji_jing@zju.edu.cn

第一通讯作者简介:王俏丽(1990-),女,浙江义乌人,博士,讲师,主要研究方向为大气污染特征及溯源研究。E-mail:chocowang@zju.edu.cn

第二通讯作者简介:李伟(1965-),男,江西上饶人,博士,教授,主要研究方向为大气污染防控。E-mail:w_li@zju.edu.cn

0 引 言

深入打好污染防治攻坚战,不断改善空气质量,是建设美丽中国的必要前提。地级及以上城市空气质量优良天数比率到 2025 年达到 87.5%,已成为我国“十四五”时期经济社会发展的一项约束性指标[1]。当前,我国大气细颗粒物($PM_{2.5}$)污染形式依然严峻[2-3]且臭氧(O_3)污染日益凸显[4],成为影响空气质量的主要因素。京津冀及周边地区、长三角地区、汾渭平原区域现阶段源解析研究表明,挥发性有机物(VOCs)是 $PM_{2.5}$ 和 O_3 大气复合污染的重要来源[5]。此外,环境空气中部分 VOCs 具有特殊气味并且表现出刺激性、腐蚀性、器官毒性、致癌性,对人体健康造成较大的影响[6-7]。部分 VOCs 可以被传输到平流层,对臭氧层造成破坏,少数 VOCs 属于温室气体[8]。因此,减少 VOCs 的排放对于提高空气质量有着重要意义。

实施 VOCs 减排,要抓好污染严重的重点行业,准确识别重点企业和工艺过程,全面推进工业园区、企业集群等 VOCs 的精准治理和综合治理[9]。随着医药行业的迅速发展,中国已经成为一个医药大国,医药行业的 VOCs 排放成为一个不可忽视的环境问题[10]。随着化工企业“退城入园”工作的推进,化工园区的企业密度日益变大,作为精细化工产业的代表,制药行业在化工园区占据重要的地位。化学合成类制药行业,生产原料使用大量有机溶剂,合成工艺复杂,各类副反应繁多,存在大量间歇性、无组织的 VOCs 排放,使得排放规律不明晰、排放特征不明确,同时还存在监测难度大,污染来源追溯难等问题[11-14]。

针对污染排放源监测构建的污染源成分谱是描述源排放特征的重要数据集之一[15-17]。然而,VOC 污染源成分谱由于数据量大、因子多、信息不完备、数据规则不明显,在其应用过程中难以充分挖掘排放特征。而特征污染物可以简化源成分谱描述,减少数据干扰,以少量的组分表征污染源的排放特征,实现污染源类的定性判定[18-20]。随着科学研究的基本手段从传统的“理论+实验”走向现在的“理论+实验+计算”,乃至出现“数据科学”这样的提法,机器学习的重要性日趋显著。在环境领域,已有部分研究者采用机器学习的手段提取各种类型的特征因子。张云鹏等使用典型相关性分析和空间网格化逻辑回归分析方法获得了影响土地利用变化的全局特征因子和空间特征因子[21]。孙笑笑采用聚类分析和相关性分析提取了浙江近海岸赤潮发生时产生突变的赤潮特征因子[22]。曹丛华等采用主成分分析(PCA)和聚类分析提取了辽东湾鲅鱼圈赤潮的环境特征因子[23]。吴超凡采用回归分析、相关性分析和特征选择方法识别了与森林生物量相关的特征因子[23]。机器学习具备适应复杂数据,能获得预测模型的优点。

本文以长三角地区某精细化工园区内一家典型化学制药企业为研究对象,深入分析其 VOCs 排放特征,利用机器学习的手段开展统计分析,通过数据驱动识别其生产工艺过程的排放特征因子。识别的特征因子种类精简,易于监测,与污染源类能够高度对应,可为化学合成类制药行业实施 VOCs 减排、合理选择排放控制技术及后续地方标准的制定提供基础信息,为实现化工园区大气污染溯源提供了一条新思路。

1 数据与方法

1.1 数据来源

污染源 VOCs 成分谱来自长三角地区某精细化工园区内一家典型化学合成类制药企业,该企业生产的恩诺沙星、阿奇霉素、罗红霉素等产品份额约占海内外市场的 30%。根据环评资料和现场调研,对厂区内 VOCs 排放源开展了全覆盖的样品采集工作,收集了 20 个污染源样本,分析了 116 种 VOCs 组分的浓度,并基于分析结果构建了基于工艺过程的精细化污染源成分谱,参见前期相关成果[25]。采样信息如表 1 所示。将污染源成分谱表示为数据集 $D = \{x_1, x_2, \cdots, x_m\}$,其中 $m = 20$,代表样本数量。$x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ 代表每个样本由各个 VOC 物种浓度组成的特征向量,单位 $\mu g \cdot m^{-3}$;$d = 86$,为所有检测出的 VOCs 物种的数量。

1.2 技术路线

污染源成分谱中的每个 VOC 物种被定义为一个特征,构成一个特征集。特征因子的识别过程被转化成机器学习中的一个特征选择过程,对特征子集的评价采用分类器的分类准确率作为标准。识别特征因子的技术路线如图 1 所示。

首先采用 PCA 加聚类分析将所有样本按照 VOCs 的物种组成相似度分为若干类别,并予以标记,实现污染源类别的区分。聚类分析通过对无标

表 1　污染源采样信息

Table 1　Pollution source sampling information

编号	采样点位	所属生产线
1	108 车间	乳酸环丙沙星、罗红霉素
2	402 车间	地克朱利
3	406 车间	恩诺沙星
4	厂界上风向	—
5	103 车间回收釜 A	阿奇霉素
6	103 车间回收釜 B	阿奇霉素
7	606 车间	头孢
8	103 车间肟化工段	阿奇霉素
9	604 车间	头孢
10	厂界下风向	—
11	402 车间加料工段	地克朱利
12	126 车间	红霉素肟、克拉霉素
13	302 车间	福多司坦、米格列奈钙
14	402 车间后段	地克朱利
15	回收车间	—
16	132 车间	环丙羧酸
17	131 车间二楼	环丙羧酸
18	104 车间	环丙沙星
19	403 车间	巴洛沙星、普卢利沙星、甲磺酸达氟沙星
20	129 车间	阿奇霉素

记训练样本的学习，将数据集划分为若干个通常是不相交的子集，每个子集称为一个簇[26]。k 均值聚类作为被广泛使用的聚类算法，是一种基于中心的聚类方法[27]。它通过迭代，将样本分到 k 个类中。通过这样的划分，每个簇可以对应一类排放特征相似的污染源。本研究中的污染源成分谱检测出了 86 种 VOCs 的浓度，属于高维度的样本数据集，将导致聚类分析中的向量相关计算量呈指数增长，并且使样本距离的度量失去意义，大大降低性能。为了使各类样本在 VOCs 组成上的差异更容易区分，PCA 用少数主成分近似表示原有数据集的所有信息，实现降维处理，提高聚类性能。

然后，对标记后的数据集分别使用 PCA 处理后的数据和特征选择处理后的数据训练若干分类器，并计算其分类准确率。分类器是从数据中学习到的一个分类模型或分类决策函数，可以对新的输入进行输出的预测，称为分类[28-29]。从给定的特征集合中选择出相关特征子集的过程，称为特征选择[30]。特征选择在于选取对提高分类器性能有所贡献的特征，即选取能够对污染源类别进行准确分类的 VOCs 物种。比较 PCA 处理和特征选择处理对分类器性能的影响，筛选出初步的特征子集作为预选特征因子。

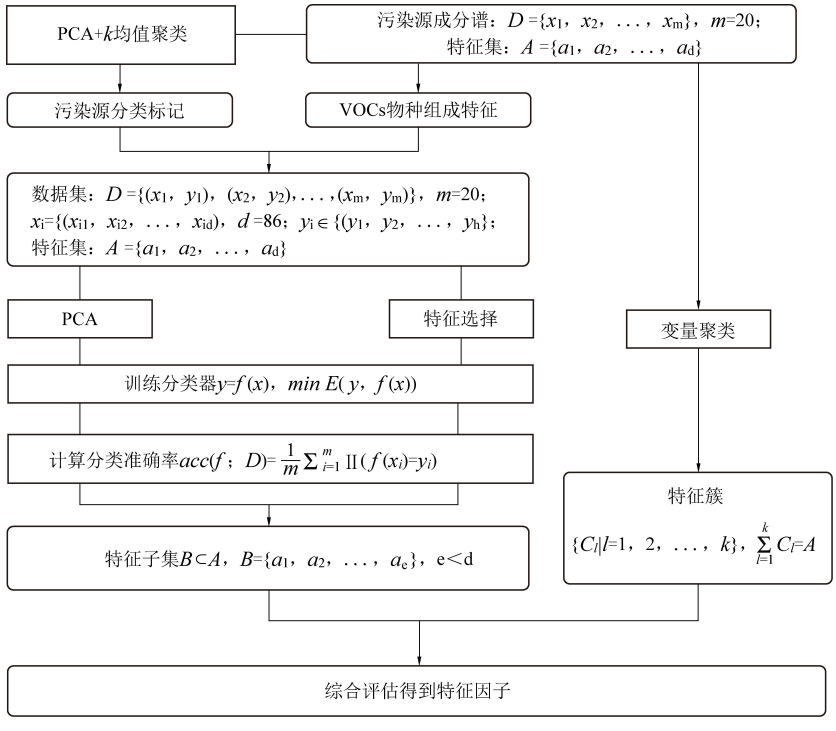

图 1　特征因子识别技术路线

Fig.1　Flowchart of characteristic factors identification

最后对污染源成分谱进行变量聚类处理,将所有 VOCs 物种划分成若干个特征簇。变量聚类根据各个物种在污染源间的浓度分布,将其分为若干个特征簇,构成同一个特征簇的物种拥有相似的污染源间浓度分布。根据综合评估特征选择和变量聚类的结果,确定最终的特征因子。

2 结果与讨论

2.1 特征选择识别特征因子

2.1.1 源样本类别标记

先对原始数据集进行 PCA 的降维处理,选取 95%的解释方差,提取获得前 15 个主成分,如图 2(a)所示。在经过 k 均值聚类后,所有样本被划分为 3 个子集,如图 2(b)所示,将其污染源类别分别标记为 1、2、3。将聚类结果与采样点所属工艺过程进行对比,如表 2 所示。拥有相同聚类标记的样本拥有相似的 VOCs 排放组成,结果显示来自相同工艺流程的样本基本上被划分到了一类。在阿奇霉素生产线,只有 103 肟化车间的样本被赋予了不同标记。同样,在地克朱利生产线,只有 402 车间后段的样本被赋予了不同标记。这说明该企业阿奇霉素生产过程与地克朱利生产过程有着与其它工艺过程显著区分的 VOCs 排放特征,而恩诺沙星、罗红霉素、麻保杀星等生产过程的 VOCs 排放特征则较为相似。聚类标记与工艺过程对应趋势明显,说明通过分析 PCA 提取的主成分信息,该企业的工艺特征得到了明显的区分。然而 PCA 获得的主成分是所有 VOCs 物种的线性组合,无法直接指向具体的物种作为污染源的特征因子,这将给实际的监测工作带来困难,也提高了溯源模型在数据输入方面的难度。

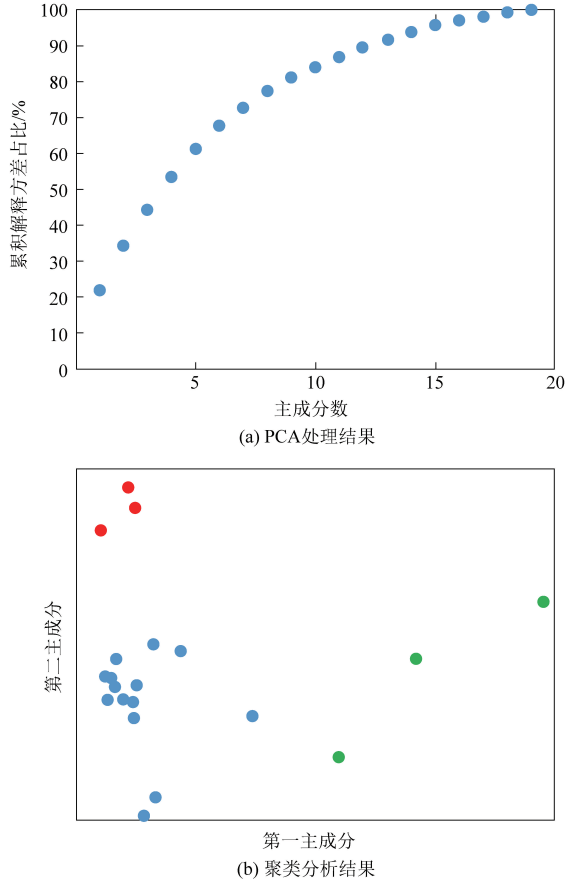

图 2 源样本类别标记

Fig.2 Labels of source samples

表 2 聚类结果与工艺过程对比

Table 2 Comparison between clustering results and production processes

工艺流程	采样点位	聚类标记
阿奇霉素	103 车间回收釜 A,103 车间回收釜 B,129 车间	1
	103 肟化车间	
恩诺沙星,罗红霉素,麻保沙星等	108 车间,406 车间,厂界上风向,606 车间,604 车间,厂界下风向,126 车间,302 车间,回收车间,132 车间,131 车间二楼,403 车间	2
地克珠利	402 车间后段	
	402 车间,402 车间加料工段	3

2.1.2 特征选择结果

对标记后的数据集进行特征选择处理。参考 PCA 选择正交变换组合的原理,特征在不同样本间的方差越大,蕴含的信息越丰富。将 86 个 VOCs 物种按照方差降序排列,对照 2.1.1 中提取的 15 个主成分,选择前 15 个物种特征作为数据输入,训练分类器,并计算其分类准确率。训练过程在 MATLAB 软件自带的机器学习与深度学习工具箱中的 Classification Learner 模块进行,验证方式选择五折交叉验证。训练的分类器类型包括决策树、判别分析、逻辑回归分类器、朴素贝叶斯分类器、支持向量机、最近邻分类器和集成分类器。作为对照,另设一组实验,直接使用 PCA 处理后的带标记数据训练分类器,并计算器分类准确率。观察性能较好的分类器,结果如表 3 所示,经过特征选择处理后的数据,有 2 个分类器的分类准确

率达到了 85.0%,说明通过观察被选择的这 15 个物种可以实现对污染源的准确分类。对比 PCA 处理后数据训练得到的分类器性能,可以发现,特征选择在对污染源进行分类方面,达到了与 PCA 处理同样的效果。因此这 15 个物种被认定为初步识别到的特征因子,如表 4 所示。

表 3　特征选择和分类分析结果

Table 3　Feature selection and classification analysis results　%

特征选择处理数据		PCA 处理数据	
分类器	准确率	分类器	准确率
Naïve Bayes	85.0	Linear Discriminant	85.0
Ensemble	85.0	KNN	85.0
—	—	Ensemble	85.0

表 4　预选特征因子

Table 4　Preselected characteristic factors

类型	特征因子物种
芳香烃	甲苯,乙苯,苯甲醛,间/对二甲苯,邻二甲苯,1,2,4-三甲苯,3-乙基甲苯,连三甲苯
卤代烃	一氯甲烷,二氯甲烷,氯仿
其它	乙醛,正己烷,丙酮,异丙醇

2.2　变量聚类识别特征因子

针对未标记的原始数据集,对 86 个 VOCs 物种进行 k 均值聚类分析。变量聚类根据各个 VOCs 物种在不同样本间的浓度分布将其分成若干个特征簇,被分为同一类的特征拥有相似的样本间浓度分布。结果如表 5 所示,所有物种被分为 3 个特征簇,其中甲苯、丙酮、乙醛、苯甲醛、正己烷、乙酸乙酯被分为一组。除乙酸乙酯外,其余物种均包含在步骤 2.1 识别出的 15 个预选特征因子当中。结合现场调研与污染源成分谱分析,甲苯是该企业多个车间的主要特征污染物,而丙酮、乙醛、苯甲醛和正己烷被划分到与甲苯一类,说明它们在各个车间的浓度分布与甲苯类似。对比特征选择和变量聚类的结果,综合特征因子的特征性和精简性,该企业的特征因子被最终认定为:甲苯、丙酮、乙醛、苯甲醛和正己烷。

表 5　变量聚类分析结果

Table 5　Results of variable cluster analysis

VOCs 物种	聚类标记
甲苯,丙酮,乙醛,苯甲醛,正己烷,乙酸乙酯	1
乙苯,一氯甲烷,异丙醇,氯仿,苯乙烯,异丁烷等	2
二氯甲烷	3

3　结　　论

本研究以基于工艺过程的精细化污染源成分谱为基础数据,采用特征选择和变量聚类的机器学习方法识别出某典型化学合成制药企业的 VOCs 排放特征因子为:甲苯、丙酮、乙醛、苯甲醛和正己烷。通过这种方法识别的特征因子,拥有相似的污染源浓度分布,并且可以较好地体现各个工艺过程在 VOCs 排放组成上的差异,对精细化的污染源类别实现准确分类。在对污染源成分谱进行分析时,可通过观察这几种物质的 VOCs 浓度组成,分析其所属工艺过程。在实际的生产监管过程中,可采集足够丰富的污染源样本构建成分谱,并训练分类器,通过重点监测特征因子的浓度,输入分类器,得到所属类别以及判别概率,有望实现 VOCs 排放的快速精细化溯源。

参考文献

[1] 中华人民共和国中央人民政府. 中华人民共和国国民经济和社会发展第十四个五年规划和 2035 年远景目标纲要[EB/OL]. (2021-03-13) [2021-09-10]. http://www.gov.cn/xinwen/2021-03/13/content_5592681.htm.

[2] Cheng N, Zhang C, Jing D, et al. An integrated chemical mass balance and source emission inventory model for the source apportionment of $PM_{2.5}$ in typical coastal areas [J]. Journal of Environmental Sciences, 2020, 92: 118-128.

[3] Zhang C, Jing D, Wu C, et al. Integrating chemical mass balance and the community multiscale air quality models for source identification and apportionment of $PM_{2.5}$ [J]. Process Safety and Environmental Protection, 2021, 149: 665-675.

[4] Wang Q, Li S, Dong M, et al. VOCs emission characteristics and priority control analysis based on VOCs emission inventories and ozone formation potentials in Zhoushan [J]. Atmospheric Environment, 2018, 182: 234-241.

[5] 中华人民共和国生态环境部. 重点行业挥发性有机物综合治理方案 [EB/OL]. (2019-06-26) [2021-09-10]. http://www.mee.gov.cn/xxgk2018/xxgk/xxgk03/201907/

t20190703_ 708395. html.

[6] Gong Y, Wei Y, Cheng J, et al. Health risk assessment and personal exposure to volatile organic compounds (VOCs) in metro carriages: A case study in Shanghai, China [J]. Science of the Total Environment, 2017, 574: 1432-1438.

[7] Jaars K, Vestenius M, van Zyl P G, et al. Receptor modelling and risk assessment of volatile organic compounds measured at a regional background site in South Africa [J]. Atmospheric Environment, 2018, 172: 133-148.

[8] Li C, Li Q, Tong D, et al. Environmental impact and health risk assessment of volatile organic compound emissions during different seasons in Beijing [J]. Journal of Environmental Sciences, 2020, 93: 1-12.

[9] 中华人民共和国生态环境部. 关于在疫情防控常态化前提下积极服务落实“六保”任务坚决打赢污染防治攻坚战的意见 [EB/OL]. (2020-06-30) [2021-09-12]. http://www. gov. cn/zhengce/zhengceku/2020 - 06/09/content _ 5518166. htm.

[10] 梁小明, 张嘉妮, 陈小方, 等. 我国人为源挥发性有机物反应性排放清单 [J]. 环境科学, 2017, 38 (3): 845-854.

[11] 吴剑, 张玲玲, 章许云, 等. 化工园区大气监控预警溯源排查一体化监管体系的探索与实践 [J]. 环境科技, 2021, 34 (1): 46-50.

[12] 叶菡韵, 田金平, 陈吕军. 精细化工园区工艺过程 VOCs 产生量核算方法 [J]. 环境科学, 2020, 41 (3): 1116-1122.

[13] 贾晓东, 金锡鹏. 我国有机溶剂危害的现状和预防 [J]. 中华劳动卫生职业病杂志, 2000 (2): 5-7.

[14] 戚飞, 徐志斌. 中药废水研究进展 [J]. 广东化工, 2017, 44 (23): 98-99.

[15] Zhong Z, Sha Q, Zheng J, et al. Sector-based VOCs emission factors and source profiles for the surface coating industry in the Pearl River Delta region of China [J]. Science of the Total Environment, 2017, 583: 19-28.

[16] Yuan B, Shao M, Lu S, et al. Source profiles of volatile organic compounds associated with solvent use in Beijing, China [J]. Atmospheric Environment, 2010, 44 (15): 1919-1926.

[17] Shen L, Xiang P, Liang S, et al. Sources profiles of volatile organic compounds (VOCs) measured in a typical industrial process in Wuhan, Central China [J]. Atmosphere, 2018, 9 (8): 297.

[18] Fu S, Guo M, Luo J, et al. Improving VOCs control strategies based on source characteristics and chemical reactivity in a typical coastal city of South China through measurement and emission inventory [J]. Science of the total Environment, 2020, 744: 140825.

[19] Song M, Liu X, Zhang Y, et al. Sources and abatement mechanisms of VOCs in southern China [J]. Atmospheric Environment, 2019, 201: 28-40.

[20] Ma Z, Liu C, Zhang C, et al. The levels, sources and reactivity of volatile organic compounds in a typical urban area of Northeast China [J]. Journal of Environmental Sciences, 2019, 79: 121-134.

[21] 张云鹏, 孙燕, 陈振杰. 基于多智能体的土地利用变化模拟 [J]. 农业工程学报, 2013, 29 (4): 255-265.

[22] 孙笑笑. 联合浮标与卫星数据的赤潮预警与决策服务 [D]. 杭州: 浙江大学, 2017: 33-36.

[23] 曹丛华, 黄娟, 郭明克, 等. 辽东湾鲅鱼圈赤潮与环境因子分析 [J]. 海洋预报, 2005 (2): 1-6.

[24] 吴超凡. 区域森林生物量遥感估测与应用研究 [D]. 杭州: 浙江大学, 2016: 43-47.

[25] Cheng N, Jing D, Zhang C, et al. Process-based VOCs source profiles and contributions to ozone formation and carcinogenic risk in a typical chemical synthesis pharmaceutical industry in China [J]. Science of the Total Environment, 2021, 752: 141899.

[26] Jain A K, Murty M N, Flynn P J. Data clustering: A review [J]. ACM Computing Surveys, 1999, 3 (31): 262-323.

[27] 杨俊闯, 赵超. K-Means 聚类算法研究综述 [J]. 计算机工程与应用, 2019, 23 (55): 7-14.

[28] Krawczyk B, Galar M, Woźniak M, et al. Dynamic ensemble selection for multi-class classification with one-class classifiers [J]. Pattern Recognition, 2018, 83: 34-51.

[29] Krawczyk B, Woźniak M, Herrera F. On the usefulness of one-class classifier ensembles for decomposition of multi-class problems [J]. Pattern Recognition, 2015, 48 (12): 3969-3982.

[30] 李郅琴, 杜建强, 聂斌, 等. 特征选择方法综述 [J]. 计算机工程与应用, 2019, 55 (24): 10-19.