Eco-Environmental Knowledge Web

环境工程学报
Chinese Journal of Environmental Engineering

第 14 卷 第 11 期 2020 年 11 月
Vol. 14, No.11 Nov. 2020

http://www.cjee.ac.cn E-mail: cjee@rcees.ac.cn (010) 62941074

文章栏目：相关研究
DOI 10.12030/j.cjee.202007079 中图分类号 X322 文献标识码 A

黄国鑫, 朱守信, 王夏晖, 等. 基于自然语言处理和机器学习的疑似土壤污染企业识别[J]. 环境工程学报，2020, 14(11): 3234-3242.

HUANG Guoxin, ZHU Shouxin, WANG Xiahui, et al. Natural language processing and machine learning-based suspected soil contamination enterprise identification[J]. Chinese Journal of Environmental Engineering, 2020, 14(11): 3234-3242.

基于自然语言处理和机器学习的疑似土壤污染企业识别

黄国鑫[1], 朱守信[1,2], 王夏晖[1,*], 田梓[1], 季国华[1], 卢然[1], 崔轩[1], 陈茜[1]

1. 生态环境部环境规划院，北京 100012
2. 中国地质大学 (北京) 水资源与环境学院，北京 100083

第一作者：黄国鑫(1980—)，男，博士，副研究员。研究方向：土壤和地下水污染防治。E-mail：huanggx@caep.org.cn
*通信作者：王夏晖(1975—)，男，博士，研究员。研究方向：生态保护修复及土壤污染防治。E-mail：wangxh@caep.org.cn

摘　要　针对污染场地识别的精准性不高、科学性不足、全面性不够和数据共享难度大等问题，以南方某地级市为研究区，借助大数据平台，基于自然语言处理和机器学习，通过引入摘要中热词权重构建改进型朴素贝叶斯模型，并对兴趣点 (POI) 数据进行中类行业预测和污染企业识别。结果表明，与随机森林算法和 XGBoost 算法相比，朴素贝叶斯算法的性能最佳；企业名称+经营范围构建有语义词汇库后，朴素贝叶斯算法的准确率、召回率和综合评价指标 (F_1) 值得到大幅提升，分别提高了 0.23、0.23 和 0.23；采用权重 1.27 和平滑参数 α 为 1.10 后，建立了改进型朴素贝叶斯模型，实现了行业类别预测，相应的准确率、召回率和 F_1 值分别为 0.63、0.62 和 0.63；识别出研究区中 26 个疑似土壤污染行业有关 1 774 家企业。改进型朴素贝叶斯模型能够有效地预测疑似土壤污染企业，具有较好的准确率与召回率，能够为场地污染识别与风险管控实践提供理论依据和设计参数。

关键词　土壤污染；自然语言处理；机器学习；中类行业；污染企业识别；改进型朴素贝叶斯模型

近年来，场地土壤污染问题越来越受到公众和社会的关注[1-2]。我国在汲取国外近 40 年治理经验的基础上，提出了“预防为主，保护优先，风险管控”的场地土壤污染防治策略，初步形成了包括法律、法规、导则、指南和规章在内的一整套相对较为完善的场地土壤风险管控体系。尽管如此，我国场地土壤污染风险管理依然处于刚刚起步阶段，尤其是土壤污染底数不清。目前，主要采用现场踏勘、人员访谈、资料分析并结合日常监管等方式进行疑似污染场地识别，但是，这些传统方式的精准性不高、科学性不足、全面性不够，工作效率较低。

近年来，大数据在生态环境保护领域的研究与应用得到了快速发展[3-10]，特别是利用大数据开展土壤污染风险识别与风险管控的研究越来越受到研究者的关注[11-13]。针对非结构化调查报告，利用自然语言处理，自动提取和生成结构化土壤污染信息，实现土壤数据分析已见报道[11]。有学者基于第二次土地调查数据，结合高程、地貌、土地类型等 17 个环境协变量数据，利用随机森林、

收稿日期：2020-07-11；**录用日期**：2020-10-26

基金项目：国家重点研发计划项目 (2018YFC1800205)；生态环境部环境规划院青年科技创新基金 (2018 年度)

极端梯度提升等，绘制了高精度的全国土壤 pH 空间分布地图，并推测了土壤重金属环境容量[12]。值得一提的是，JIA 等[13] 考虑到政府部门间存在数据孤岛、数据共享难度大等问题，以长江三角洲地区为研究区，基于兴趣点 (Point Of Interest) 的非结构化文本数据，利用多项式朴素贝叶斯算法，识别了疑似土壤污染企业，对场地调查评估、风险管控等环境管理提供了良好的决策支撑作用。但是，该研究仅能识别《国民经济行业分类》(GB/T 4754-2017) 中大类行业企业，利用企业名称构建有语义词汇库，且未构建无语义词汇库[13]。识别中类甚至小类行业以提高预测精度、增加有语义词汇库库容以克服朴素贝叶斯算法的过度拟合和零概率现象、构建无语义词汇库以降低维数和提高运算速度等已成为疑似土壤污染企业识别中迫切需要解决的问题。

鉴于此，本研究以南方某地级市为研究区，借助大数据平台，基于自然语言处理和机器学习，尝试利用改进型朴素贝叶斯算法，预测 POI 数据中企业所属中类行业类别，识别疑似土壤污染企业，以期为场地污染识别与风险管控实践提供理论依据和设计参数。

1　材料与方法

1.1　基础数据及预处理

1) 基础数据。国民经济行业分类数据 (1 700 条)：小类行业名称、中类行业名称和分类说明。污染企业数据 (62×10^4 条)：企业名称、行业类别和经营范围。POI 数据 (9 900 条)：企业名称和经纬度坐标。疑似土壤污染行业数据 (38 条)：中类行业名称和特征污染物。日常监管数据 (221 条)：企业名称和经纬度坐标。

2) 数据预处理。剔除标点符号、英文字母、数字等词汇；通过 pynlpir 辅助函数进行降噪；进行唯一性检查、去重、人工补缺和精度归一化处理；利用自行设计的自关联表 (表 1) 对小类行业名称及其分类说明向上聚合至所属中类。

表 1　自关联表

Table 1　Self-correlation table

当前类别标识	类别名称	分类说明	上级类别标识
193	毛皮鞣制及制品加工	—	—
1 931	毛皮鞣制加工	指带毛动物生皮经鞣制等化学和物理方法处理后，保持其绒毛形态及特点的毛皮(又称裘皮)的生产活动	193

注：“毛皮鞣制加工”为小类名称；“毛皮鞣制及制品加工”为中类名称。

1.2　大数据软硬件环境

1) 硬件环境。管理服务器 2 台，用于 CDH Manager 管理和 Zookeeper 分布式协调服务，并作为 Hive 数据仓库入口；计算服务器 4 台，作为 Impala、Spark 的计算节点和 Hbase 节点，其中 2 台还用于 Zookeeper 分布式协调服务，并作为 Redis 数据库。服务器的核心组件为 CPU：12 核心、线程数 2 个/核心、主频 2.2 GHz、三级缓存 16.5 MB。内存：总容量 128 GB、单挑容量 16 GB、规格 DDR4、工作频率 2 400 MHz。磁盘：系统盘容量 600 GB、数据盘容量 2 TB、接口形式 SAS。RAID 卡：支持 RAID0、RAID1、RAID5、RAID10、RAID50、JBOD 等模式。网络：带宽 10 Gbps。系统：CentOS 7.4。

2) 软件环境。核心组件为 JDK 1.8、Python 3.7、Scala 2.11.x、OpenSSL、Nginx、Tomcat、Libgfortran 4.6+、Apache Hadoop 2.x、Apache Zookeeper 3.4.x、Apache Hive 2.1.x、Apache HBase 1.2.x、Hue 3.9.x、Apache Impala 2.12.x、Apache Parquet 2.1.x、Apache Spark 1.6.x、Apache Spark2 2.4.x、Redis 4.x、MongoDB 4.2.x、PostgreSQL 9.4.x、CDH 5.16、ArcGIS 10.2.2、Echart 4.8.0-release。

1.3 大数据技术架构

基于大数据存储和处理的需要，于 CentOS7.4 集群，运用分布式技术，搭建大数据平台架构，主要由数据资源汇聚层、数据平台层、分析处理层、前端展示层和数据访问层等 5 个功能层组成(图 1)，能够满足行业分类预测、污染企业识别、ArcGIS 平台与大数据平台交互、可视化展示等需求。

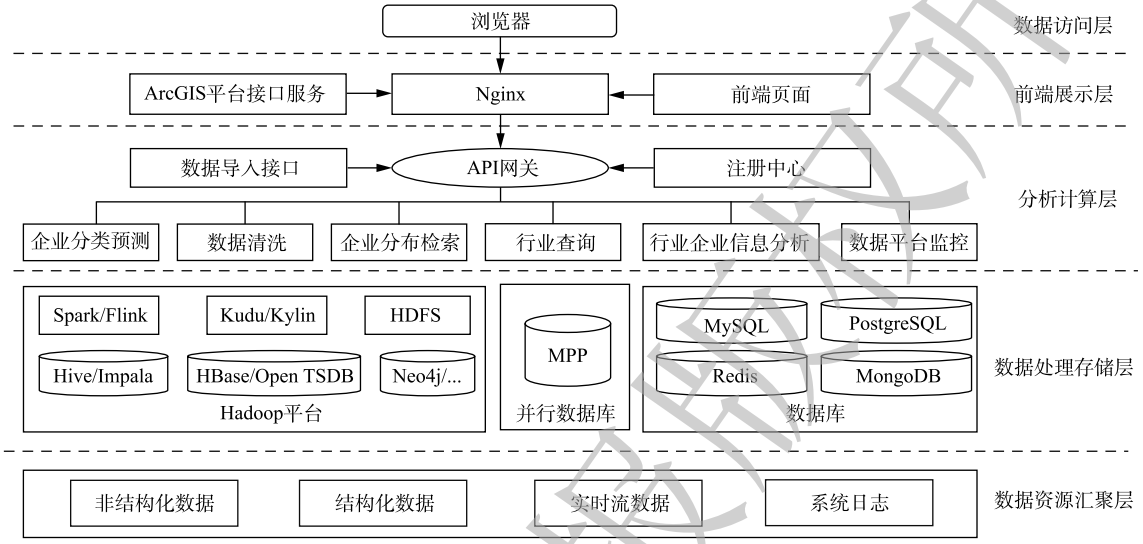


图 1 大数据平台架构

Fig. 1 Big data platform framework

1.4 基于改进型朴素贝叶斯算法的中类行业类别预测与污染企业识别

1) 特征工程处理：针对国民经济行业分类数据、污染企业数据和 POI 数据，首先，采用隐马尔可夫模型[14-15]、Viterbi 算法和 jieba 分词引擎进行中文分词，并采用 cut 函数提取和剔除地名、“公司”“有限”“有限责任”等对行业类别预测无意义的词汇组成无语义词汇库，剩余的词汇组成有语义词汇库；其次，采用词频-逆文本频率算法[16-17]统计各个样本中位于有语义词汇库内词汇词频，其中 min_df 下频率值调整为 0.15、max_df 上频率值调整为 0.90；然后，再次人工过滤并剔除出现次数多且对行业类别预测无意义的词汇，并将其增补进无语义词汇库，同时剩余的词汇作为特征词组成最终的有语义词汇库；最后，采用词频-逆文本频率算法重新统计各个样本中特征词词频 (式 (1)~式 (3))。

特征词正向词频 ($tf_{i,j}$) 计算见式 (1)，特征词逆向文本频率 (idf_j) 计算见式 (2)，特征词词频 ($tf_i df_{i,j}$) 计算见式 (3)。

$$tf_{i,j}=\frac{n_{i,j}}{\sum\limits_{k} n_{i,j}} \tag{1}$$

式中：$tf_{i,j}$为第 i 个特征词在第 j 个污染企业名称中的词频；$n_{i,j}$为第 i 个特征词在第 j 个污染企业名称中的出现次数；$\sum\limits_{k} n_{i,j}$为第 j 个污染企业名称中全部 k 个特征词出现次数的总和。

$$idf_j=\lg\frac{|D|}{\left|\{j:w_i\in d_j\}\right|} \tag{2}$$

式中：idf_j为第 i 个特征词的逆向文本频率；$|D|$为有语义词汇库内所有污染企业名称的总数；d_j为第 j 个污染企业名称；$\left|\{j:w_i\in d_j\}\right|$为包含第 i 个特征词的污染企业名称的总和。

$$tf_i df_{i,j}=tf_{i,j}\cdot idf_{i,j} \tag{3}$$

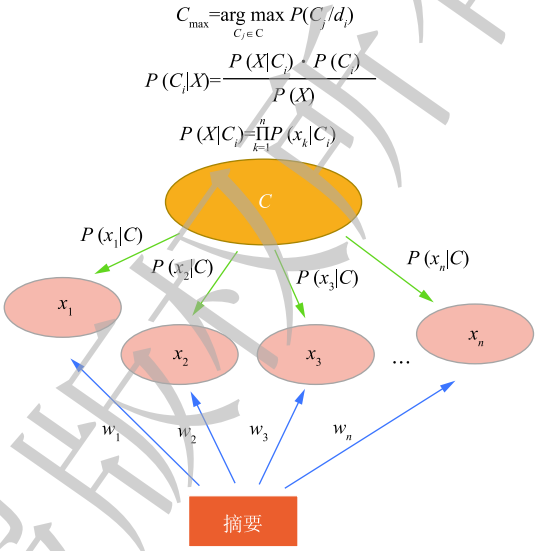
式中：$tfidf_{i,j}$为第 i 个特征词在第 j 个污染企业名称中的权重；$tf_{i,j}$同式 (1)；$tfidf_{i,j}$同式 (2)。

2) 摘要构建：按小类行业，将行业名称和分类说明中由高至低排在前 100 位的有语义词汇组成热词；然后，利用自关联表对各小类行业的热词向上聚合至所属中类，形成代表中类行业的摘要。

3) 行业类别预测模型构建与训练：首先，结合摘要，将特征词与摘要进行匹配，匹配上时将特征词词频乘以权重作为其特征值，匹配不上时则将特征词词频作为其特征值；其次，使用训练数据集训练基于改进型朴素贝叶斯算法的预测模型[18-19](图 2)，在此过程中，使用 10 折交叉验证的网格搜索方法调整拉普拉斯平滑法[20]中平滑参数 α，使用 5 次验证集的平均准确率最高值作为最佳参数；最后，通过检验数据集的准确率、召回率和 F_1 值评估模型，获取改进型行业类别预测模型。

4) POI 数据的行业类别预测：将 POI 数据输入已经训练好的改进型朴素贝叶斯模型，预测各企业所属行业。

5) 污染企业识别：从 POI 数据的预测结果中提取疑似土壤污染行业数据涉及的中类行业，将其对应的企业作为疑似土壤污染企业。



注：C 为行业类别集合；X 为特征词集合；x_i 为第 i 个特征词；d_i 为第 i 个企业名称；w_i 为第 i 个权重。

图 2　基于改进型朴素贝叶斯算法的行业类别预测模型

Fig. 2　Improved naive Bayesian algorithm-based industry category prediction model

1.5　实验设计

1) 不同行业词云构建：采用隐马尔可夫模型、viterbi 算法和 jieba 分词引擎，对污染企业数据(含企业名称和经营范围)进行中文分词；然后，利用相同词汇累加方法，统计有语义词汇库中词汇出现的次数；最后，使用 Python 中 word cloud 库绘制不同行业词云。

2) 行业分类预测算法筛选：将污染企业数据集按 9∶1 比例划分为训练数据集和检验数据集；在此基础上，比较随机森林、XGBoost 和朴素贝叶斯 3 种算法，通过分别比较准确率、召回率和 F_1 值，确定最佳的行业分类预测算法。

3) 有语义词汇库构建方法比选：利用企业名称和企业名称+经营范围分别构建有语义词汇库，通过分别比较朴素贝叶斯算法的准确率、召回率和 F_1 值，确定最佳的有语义词汇库构建方法。

4) 朴素贝叶斯模型改进：结合摘要，通过比较不同权重和平滑参数 α 引起的朴素贝叶斯算法的准确率、召回率和 F_1 值，确定改进型朴素贝叶斯模型。

5) 行业企业空间分布结果分析：在 ArcGIS 平台上，以南方某地级市作为研究区，将 POI 疑似土壤污染企业和日常监管企业分行业进行空间分布，分析行业分类预测和污染企业识别的实际效果。

1.6　数据分析方法

行业分类预测的准确率计算见式 (4)，行业分类预测的召回率计算见式 (5)，行业分类预测的 $F1$ 值计算见式 (6)。

$$P=\frac{n_c}{n} \tag{4}$$

式中：P 为准确率，预测正确的样本占所有样本的比例；n 为所有样本个数；n_c 为预测正确的样本个数。

$$R = \frac{n_c}{m} \tag{5}$$

式中：R 为召回率，预测正确的样本占某个行业所有样本的比例；n_c 同公式 (1)；m 为某个行业所有样本个数。

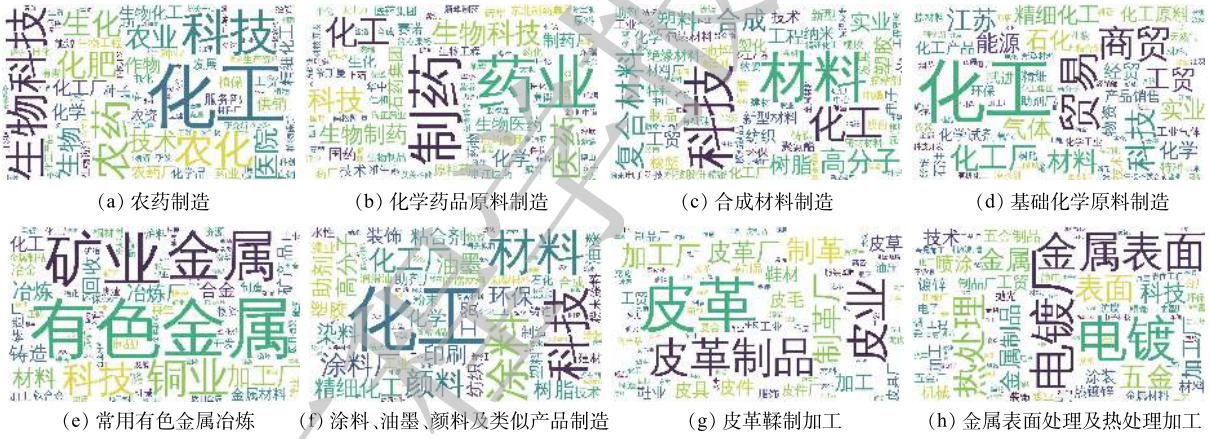
$$F_1 = \frac{2PR}{P+R} \tag{6}$$

式中：F_1 为综合评价指标值；P 同式 (4)；R 同式 (5)。

2 结果与讨论

2.1 不同土壤污染重点行业词云

针对有语义词汇库中多于 40×10^4 个词汇，采用颜色区分词汇，采用字体大小区分出现频率，经统计形成不同土壤污染重点行业词云，部分行业词云见图 3。由图 3 可知，农药制造行业的高频词汇为化工、生物科技、科技；化学药品原料制造行业的高频词汇为制药、药业；合成材料制造行业的高频词汇为科技、材料、化工；基础化学原料制造行业的高频词汇为化工、贸易、商贸；常用有色金属冶炼行业的高频词汇为有色金属、矿业金属；涂料、油墨、颜料及类似产品制造行业的高频词汇为化工、涂料、科技、材料；皮革鞣制加工行业的高频词汇为皮革、皮业、皮革制品；金属表面处理及热处理加工行业的高频词汇为电镀、电镀厂、金属表面。可知，词云有助于初步地认知和感知不同行业特点，并为后续行业分类预测和污染企业识别提供前提基础。



(a) 农药制造 (b) 化学药品原料制造 (c) 合成材料制造 (d) 基础化学原料制造

(e) 常用有色金属冶炼 (f) 涂料、油墨、颜料及类似产品制造 (g) 皮革鞣制加工 (h) 金属表面处理及热处理加工

图 3 8 个基于多源数据的土壤污染重点行业词云

Fig. 3 Eight word clouds based on the multi-source data-based soil contamination key middle-class industry

2.2 行业分类预测算法筛选

随机森林、XGBoost 和朴素贝叶斯等行业分类算法引起的准确率、召回率和 F_1 值变化见表 2。准确率衡量算法分类结果的准确性，召回率衡量算法分类结果的完整性，而 F_1 值则是综合考虑前述 2 个因素衡量算法分类结果效果。由表 2 可知，无论从准确率还是召回率亦或 F_1 值上看，不同算法的分类性能存在一定差异，且朴素贝叶斯算法的性能优于随机森林算法和 XGBoost 算法。其中，前者比后者在准确率上分别提高了 0.07 和 0.04；在召回率上分别提高 0.08 和 0.07；在 F_1 值上分别提高 0.07 和 0.05。因此，采用朴素贝叶斯算法进行行业分类预

表 2 不同行业分类预测算法性能比较

Table 2 Performance comparison of the different industry category prediction algorithms

算法类型	P	R	F_1
随机森林	0.28	0.28	0.28
XGBoost	0.31	0.29	0.30
朴素贝叶斯	0.35	0.36	0.35

测，尽管该算法的性能还有待提高。

2.3　有语义词汇库构建方法

利用企业名称和企业名称+经营范围分别构建有语义词汇库，2 种构建方法引起的朴素贝叶斯算法的准确率、召回率和 F_1 值变化见表 3。由表 3 可知，与仅采用企业名称相比，采用企业名称+经营范围构建有语义词汇库后，朴素贝叶斯算法的准确率、召回率和 F_1 值得到大幅提升，分别提高了 0.23、0.23 和 0.23，这缘于经营范围扩充了有语义词汇库库容，减少了 POI 企业名称向量化时新词汇特征的损失。因此，采用企业名称+经营范围构建有语义词汇库。

表 3　不同有语义词汇库构建方法引起的朴素贝叶斯算法性能比较

Table 3　Performance comparison of the naive Bayesian algorithm by different sematic database construction methods

有语义词汇库构建方法	P	R	F_1
企业名称	0.35	0.38	0.36
企业名称+经营范围	0.58	0.61	0.59

2.4　朴素贝叶斯模型优化

不同权重和平滑参数 α 分别引起的朴素贝叶斯算法的准确率、召回率和 F_1 值变化见图 4 和图 5。由图 4 可知，与对照组 (权重为 1) 相比，当权重为 1.15 和 1.30 时准确率、召回率和 F_1 值均变化不大；当权重为 1.27 时三者数值则分别提高了 0.05、0.07 和 0.06，表明权重 1.27 为最佳值。显然，该最佳值明显提升了具有行业分类特征的特征词的特征值，规避了训练集中各行业样本数分布不均造成朴素贝叶斯算法倾向于大类、忽略小类的现象[21]，进而提高了该算法的性能。

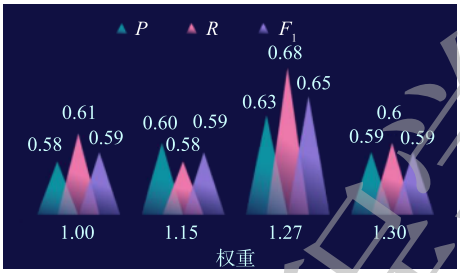


图 4　不同权重引起的朴素贝叶斯算法性能比较

Fig. 4　Performance comparison of the naive Bayesian algorithm by different weights

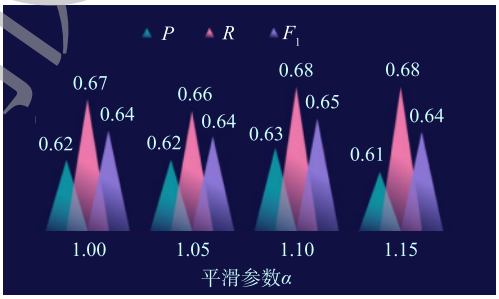


图 5　不同平滑参数 α 引起的朴素贝叶斯算法性能比较

Fig. 5　Performance comparison of the naive Bayesian algorithm by different α parameter values

尽管前述利用经营范围扩充了有语义词汇库，但是依然不可能穷举所有的特征词，故在对 POI 企业名称向量化时仍然会损失新词汇的特征，从而会产生过度拟合现象。另外，在计算先验概率时，若 POI 企业名称的某个特征词在训练数据集中某个行业类别中没有特征值，则会发生零概率现象[20]。据此，在计算后验概率时，利用平滑参数 α 力求缓解过度拟合和零概率现象，从而优化朴素贝叶斯算法。由图 5 可知，当平滑参数 α 为 1.10~1.15 时，准确率、召回率和 F_1 值均变化不大，分别为 0.61~0.63、0.66~0.68、0.64~0.65；而且，平滑参数 α 为 1.10 时，识别效果最好。

2.5　行业企业空间分布

研究区的 POI 数据所属疑似土壤污染行业企业的预测结果见表 4，相应的 POI 企业和日常监管企业的空间分布见图 6。由表 4 和图 6 可知，从行业上看，预测疑似土壤污染行业 26 个，主要为金属表面处理及热处理加工、铁合金冶炼、专用化学产品制造、农药制造、常用有色金属冶炼、基础化学原料制造和合成材料制造 (各行业企业均≥100 家)；同时，现有日常监管中未关注农药制造 (118 家)、化学药品原料药制造 (1 家)、棉纺织及印染精加工 (5 家)、环境治理业 (82 家)、皮革鞣制加工 (47 家)、贵金属冶炼 (23 家) 等行业；从数量上看，识别疑似土壤污染企业 1 774 家，远远

表 4 改进型朴素贝叶斯模型的预测结果

Table 4 Prediction results of the improved naive Bayesian algorithm

序号	中类行业名称	企业数量/家	序号	中类行业名称	企业数量/家
1	金属表面处理及热处理加工	207	14	其他仓储业	51
2	铁合金冶炼	196	15	炼铁	48
3	专用化学产品制造	167	16	电池制造	46
4	农药制造	118	17	皮革鞣制加工	47
5	常用有色金属冶炼	113	18	环境卫生管理	40
6	基础化学原料制造	102	19	贵金属冶炼	23
7	合成材料制造	100	20	炸药、火工及焰火产品制造	11
8	毛皮鞣制及制品加工	94	21	常用有色金属矿采选	10
9	涂料、油墨、颜料及类似产品制造	85	22	铁矿采选	9
10	环境治理业	82	23	棉纺织及印染精加工	5
11	纸浆制造	80	24	稀有稀土金属矿采选	1
12	炼钢	73	25	贵金属矿采选	1
13	稀有稀土金属冶炼	64	26	化学药品原料药制造	1

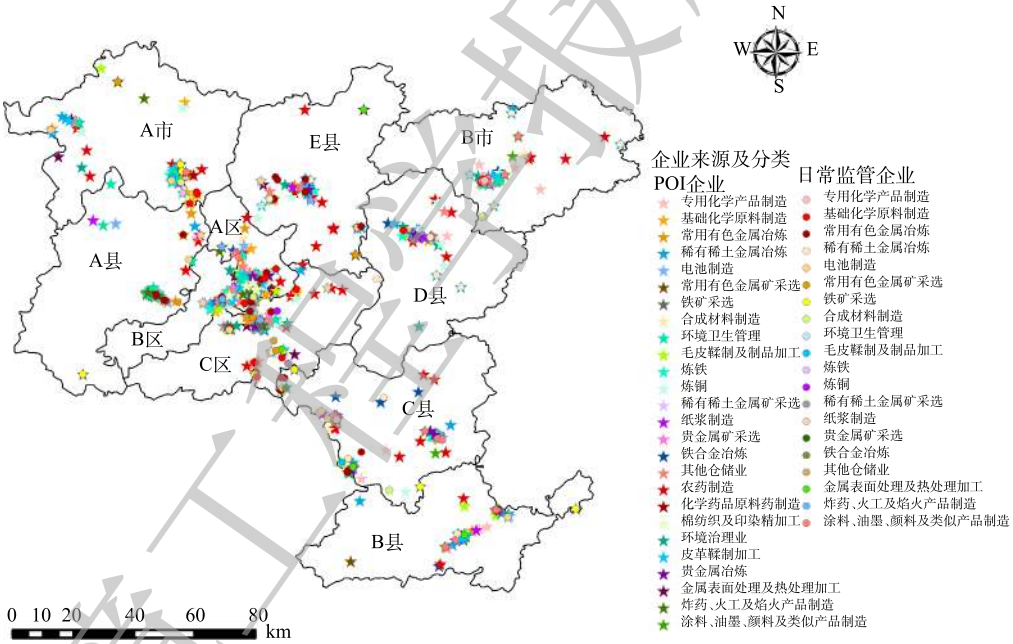


图 6 研究区中行业企业空间分布

Fig. 6 Spatial distribution of the industry enterprises in the study area

多于日常监管掌握的 221 家企业；从空间分布上看，各区 (市、县) 均存在企业集聚区，特别是在 A、B、C 区的企业分布最为集中。以上结果表明，后续应强化对新识别的行业、企业及其集聚区的土壤污染隐患排查与风险管理。另外，本研究未考虑企业生产规模、生产年限、地块利用历史等因素，对于零星分布的企业同样应做好监管。

3 结论

1) 在行业分类预测时，朴素贝叶斯算法的性能优于随机森林算法和 XGBoost 算法的性能。

2) 与仅采用企业名称相比，采用企业名称+经营范围构建有语义词汇库后，朴素贝叶斯算法的

准确率、召回率和 F_1 值均得到大幅提升，可将其作为最佳的有语义词汇库构建方法。

3) 采用权重 1.27 和平滑参数 α 为 1.10 后，建立了改进型朴素贝叶斯模型，相应的准确率、召回率和 F_1 值分别为 0.63、0.62 和 0.63，进而获得了最佳的分类预测性能。

4) 利用改性型朴素贝叶斯模型识别出研究区中 28 个疑似土壤污染行业有关 1 774 家企业，其在各区 (市、县) 均存在集聚区，特别是在 A、B、C 区最为集中。

参 考 文 献

[1] 宋昕, 林娜, 殷鹏华. 中国污染场地修复现状及产业前景分析[J]. 土壤, 2015, 47(1): 1-7.

[2] 李梦瑶. 中国污染场地环境管理存在的问题及对策[J]. 中国农学通报, 2010, 26(24): 338-342.

[3] 王夏晖. 大数据: 场地污染智能识别与风险精准管控驱动力[J]. 环境保护, 2019, 47(3): 14-16.

[4] FAZIO M, CELESTI A, PULIAFITO A, et al. Big data storage in the cloud for smart environment monitoring[J]. Procedia Computer Science, 2015, 52: 500-506.

[5] 李赛. 大数据环境下突发事件应急决策支持系统研究[D]. 武汉: 华中师范大学, 2016.

[6] 周煜申, 康望星, 沈存, 等. 大数据在水环境综合评价预警中的应用研究[J]. 江苏科技信息, 2017, 34(35): 52-54.

[7] HENGL T, DE JESUS J M, HEUVELINK G B M, et al. SoilGrids250m: Global gridded soil information based on machine learning[J]. Plos One, 2017, 12(2): 1-40.

[8] 马丽萍, 曹国良, 郝国朝. 基于大数据的大气污染防治方式优化探究-以西安市为例[J]. 环境与可持续发展, 2018, 43(2): 54-56.

[9] 铁晓波. 大数据平台下基于人工免疫系统的MBR膜污染研究[D]. 天津: 天津工业大学, 2017.

[10] 赵苗苗, 赵师成, 张丽云, 等. 大数据在生态环境领域的应用进展与展望[J]. 应用生态学报, 2017, 28(5): 1727-1734.

[11] WANG D S, LIU J Z, ZHU A X, et al. Automatic extraction and structuration of soil-environment relationship information from soil survey reports[J]. Journal of Integrative Agriculture, 2019, 18(2): 328-339.

[12] CHEN S, LIANG Z, WEBSTER R, et al. A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution[J]. Science of the Total Environment, 2019, 655: 273-283.

[13] JIA X, HU B, MARCHANT B P, et al. A methodological framework for identifying potential sources of soil heavy metal pollution based on machine learning: A case study in the Yangtze Delta, China[J]. Environmental Pollution, 2019, 250: 601-609.

[14] NASFI R, AMAYRI M, BOUGUILA N. A novel approach for modeling positive vectors with inverted Dirichlet-based hidden Markov models[J]. Knowledge-Based Systems, 2020, 192: 1-17.

[15] ARPAIA P, CESARO U, CHADLI M, et al. Fault detection on fluid machinery using Hidden Markov Models[J]. Measurement, 2020, 151: 1-7.

[16] 黄春梅, 王松磊. 基于词袋模型和TF-IDF的短文本分类研究[J]. 软件工程, 2020, 23(3): 1-3.

[17] 王方伟, 杨少杰, 赵冬梅, 等. 基于改进TF-IDF的多态蠕虫特征自动提取算法[J]. 华中科技大学学报(自然科学版), 2020, 48(2): 79-84.

[18] 何敏, 武德安, 吴磊. 基于MapReduce的平均多项朴素贝叶斯文本分类[J]. 计算机应用研究, 2016, 33(1): 115-117.

[19] 赵博文, 王灵矫, 郭华. 基于泊松分布的加权朴素贝叶斯文本分类算法[J]. 计算机工程, 2020, 46(4): 91-96.

[20] 徐光美, 刘宏哲, 张敬尊, 等. 用平滑方法改进多关系朴素贝叶斯分类[J]. 计算机工程与应用, 2017, 53(5): 69-72.

[21] 陈凯, 黄英来, 高文韬, 等. 一种基于属性加权补集的朴素贝叶斯文本分类算法[J]. 哈尔滨理工大学学报, 2018, 23(4): 69-74.

(本文编辑: 金曙光)

Natural language processing and machine learning-based suspected soil contamination enterprise identification

HUANG Guoxin[1], ZHU Shouxin[1,2], WANG Xiahui[1,*], TIAN Zi[1], JI Guohua[1], LU Ran[1], CUI Xuan[1], Chen Xi[1]

1. Chinese Academy for Environmental Planning, Beijing 100012, China

2. School of Water Resources and Environment, China University of Geosciences (Beijing), Beijing 100083, China

*Corresponding author, E-mail: wangxh@caep.org.cn

Abstract Aiming at the problems of low accuracy, inadequate scientific basis, bad wholeness and the difficulty in data sharing of soil contamination identification, a typical city in South China was selected as the research area. Based on the natural language processing and machine learning, an improved naive Bayesian model was constructed by the weights of hot words from an abstract and then utilized to predict the middle-class industries and identify the relevant contamination enterprises from point of interest (POI) data with a big data platform. The results showed that the performance of the naive Bayesian aggregation was better than that of random forest and XGBoost aggregations; the precision, recall and F_1 values of the naive Bayesian aggregation were improved by 0.23, 0.23 and 0.23 after the semantic vocabulary database was constructed by enterprise name and business scope; the naive Bayesian model that constructed under the weight of 1.27 and smoothing parameter α value of 1.10 could be used for the prediction of the middle-class industries with the precision, recall and F_1 value of 0.63, 0.62 and 0.63, respectively, and 1 774 suspected soil contamination enterprises affiliated to 26 industry categories were identified in the research area. Therefore, the improved naive Bayesian model with the good precision and recall values can be effectively used to predict the suspected contamination enterprises, and provides the theoretical bases and design parameters for site contamination identification and risk management.

Keywords soil contamination; natural language processing; machine learning; middle-class industries; contamination enterprise identification; improved naive Bayesian model