ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.19, No.22, August 2023

E-mail: eduf@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于XGBoost的企业排污违法概率评估模型

邱文韬,黄楠,李俊,李江华

(江西理工大学 信息工程学院,江西 赣州 341000)

摘要:为帮助环保部门监管企业排污违法行为,设计一个能够对企业排污数据进行违法概率评估的模型。基于99家企业的历史排污数据,确定需要监测的排污违法行为,进行数据清洗,使用K-Means和GMM算法对数据进行分类,将预处理后的数据用于构建基于XGBoost算法的排污违法行为概率评估模型,并与随机森林等机器学习算法进行比较,得出XGBoost预测效果优于其他模型的结论。以包含企业排污口编号、污染物种类、浓度、排量等信息的排污数据作为输入数据,正确预测企业污染排放违法行为概率,进而辅助环保部门监管灵活执法,促进生态环境的保护。

关键词: XGBoost;GMM;污染排放;排污违法行为

中图分类号:TP181 **文献标识码:**A

文章编号:1009-3044(2023)22-0018-04

DOI:10.14004/j.cnki.ckt.2023.1206

开放科学(资源服务)标识码(OSID):

0 引言

近年来,我国愈发重视生态环境的保护,环境监管力度不断加强。以江苏省为例,截至2021年3月,全省已发放排污许可证35 000张,基本实现全覆盖。然而,由于排污单位自行监测在固定污染源排污许可制度中占据重要地位,自行监测实施质量直接影响排污许可制度实施效果。目前,根据企业排污数据判别相关违法行为的研究较少,与监督性监测相比,排污单位自行监测的基础相对薄弱[1]。因此,通过建立并训练数据模型,借助模型强大的计算能力对企业污染排放行为进行已有数据的归宗辨识,辅助环保部门识别企业是否可能存在偷排、篡改数据或者伪造数据等违法行为很有必要。

1 研究现状回顾

1.1 机器学习在排污监管中的应用研究

机器学习已经被广泛应用于解决各种各样的现实问题,王军霞等人[1]提出按照数据标识判别、单源数据分析、多源数据分析3个层次对监测数据进行分析,以识别监测数据存在的问题。魏艳等人[2]基于大量污染源自动监测数据的特征分析与异常原因解析,探索建立针对自动监测异常数据的识别规则与标志处理方法,并通过模型训练实现了异常数据的自动标志,该方法按照数据有效性及异常原因进行标志处理,可以为后续数据分析及各类模型训练提供数据基础和保障。陈冲等人[3]对企业催化裂化的监测数据开展数据异常识别研究,从算法的分支步骤与局部度量方面,改进孤立森林算法,提高算法性能,在多个标准数据集上与多个异常识别算法进行对比,验证算法的优越性。但均未深入研究异常数据与可能存在的排污违法行为间的联系。

1.2 XGBoost算法在预测企业排污违法概率中的应用

本文主要研究XGBoost算法用于预测企业排污违法行为概率,该算法由陈天奇[4]在2016年提出,是在GBDT(Gradient Boosting Decision Tree, 梯度下降树)算法基础上进行的改进,在特征粒度上实现了并行,可以利用CPU的多线程进行学习,提高了算法的效率。XGBoost算法在预测准确率、不易过拟合和可扩展性方面明显优于随机森林、决策树、朴素贝叶斯等算法。Boosting分类器[5]属于集成学习模型,张旭春等人[6]基于黄河流域治污企业的历史环保数据与电力数据,构建了基于Boosting的集成算法CatBoost和AdaBoost的模型预测污水处理厂的告警类型。王为久等人[7]采用XGBoost算法构建了非法经营罪的自动量刑预测模型,并与线性回归、决策树、随机森林、最近邻算法等进行比较,结果证明采用XGBoost算法构建的量刑预测模型在多分类问题和回归问题的预测中均高于其他方法。Shrestha等人[7]基于尼泊尔的电信行业流失客户数据,采用XGBoost算法构建客户流失的预测模型,解决了多数机器学习算法无法解决的如何有效考虑数据集不平衡性质的问题。当前研究多使用GBDT及CatBoost等,但性能表现出色的XGBoost算法较少应用于预测企业排污违法行为及概率。

收稿日期:2023-03-08

作者简介:邱文韬(2002—),男,江西赣州人,本科;黄楠(2002—),女,江西九江人,本科;李俊(2002—),男,江西赣州人,本科;李江华(1976—),男,河南新野人,副教授,博士,研究方向为大数据分析与应用。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2 排污违法行为评估模型构建

2.1 企业排污数据分析

1) 数据来源

所采用的数据集来自河北省99家企业2020—2021年的历史排污数据，该数据集共有549 325条记录，每个记录包含6个字段，分别是企业编号、排污口编号、污染物种类、排放时间、排放浓度/流量以及排放总量。

2) 数据预处理

数据处理包括数据导入、数据清洗和数据存储三个步骤。首先在Linux系统下配置每个步骤所需的环境，并设计数据处理的数据流。数据导入阶段将原始数据导入Hdfs，再写入Hive中。数据清洗时，使用清洗规则清除缺失字段的数据，仅保留常见的五类污染物数据，并清除时间维度很少的数据，将清洗好的数据输出到Linux系统中，准备导出。数据存储时，使用Sqoop将处理后的数据输出到MySQL中。共有468 257条可用数据，选取7:3的比例对数据集进行分割，具体分布见图1：

图1 划分后的数据集

3) 数据分析

参考《中华人民共和国环境保护法》第四十四条，企业应当遵守分解到本单位的重点污染物排放总量控制指标。为贴合实际模拟企业的污染物排放，数据集扩充时只选择普遍排放的五种污染物进行记录：化学需氧量、总氮、总磷、氨氮、污水。根据《中华人民共和国国家标准污水综合排放标准》，确定了69种水污染物的最高排放浓度和行业排水量，将主要污染物的排放指标细化如表1所示：

表1 污染物排放指标

污染物	排放浓度(mg/L)/流量(L/s)
污 水	1600L/s
化学需氧量	100mg/L
总 氮	20mg/L
总 磷	0.5mg/L
氨 氮	4.5mg/L

魏艳等人[2]对数据异常特征与原因进行了分析，将自动监测数据的异常表现总结归纳为6种类型，分别是异常偏高、异常偏低、异常为0以及逻辑错误等。《水污染源在线监测系统(CODCr、NH3-N等)数据有效性判别技术规范》(HJ 356—2019)也为如何进行有效数据的人工判别提供了流程和方法指引。

4) 标准化处理

历史数据的字段名如表2所示。

表2 数据字段信息

列序号	字段名	约束	中文描述
1	bussiness_id	必填	企业编号
2	outfall_id	必填	排污口编号
3	pollu_id	必填	污染物名称
4	rtime	必填	排放时间
5	pollu_am	必填	排放浓度/流量
6	pollu_pl	必填	污染物排量

对数值不连续的business_id和outfall_id字段采取LabelEncoder编码。将rtime字段拆分为年year、月month、日day。对year进行one-hot编码，使用map()函数提取month day。pollu_id有五类，其初始数据类型为字符串，无法体现特征的联系，所以进行one-hot编码。

对pollu_pl和pollu_am特征进行标准化处理。原因：不同企业或不同排污口排放同一种污染物时，排放浓度和排放标准均不相同。例如，污水排放流量的阈值远高于总氮排放的阈值，且数据差异较大。为了避免数据偏差，需要在特征处理时进行标准化处理，使数据符合均值为0，方差为1的正态分布。标准化公式如下所示：

$$x = \frac{x - mean}{std}$$

其中，x为pollu_pl和pollu_am的特征值；$mean$为pollu_pl和pollu_am的平均值；std为pollu_pl和pollu_am的方差；

2.2 模型介绍

根据生态环境部印发的《环境监测数据弄虚作假行为判定及处理办法》以及《排污许可管理条例》，需要重点监控的排污违法行为主要有篡改、伪造自行监测数据以及偷排偷放等。因此根据排污违法行为分类的要求和原始数据的特点，可通过聚类找出正常数据。首先使用kmeans算法进行初始聚类，再用GMM算法进行聚类。由于正常数据占比最多，将聚类后最多的数据归为正常类型，其余数据划分为异常类。使用kmeans聚类对四类异常进行细分：篡改、伪造、偷排、其他。

k均值聚类算法是一种迭代求解的聚类分析算法，其步骤是，将数据分为k组，随机选取k个对象作为初始聚类中心，计算每个对象与各个聚类中心之间的距离，将每个对象分配给距离最近的聚类中心。使用kmeans进行聚类后，可得到如图2所示的效果。

kmeans对不是圆形的簇拟合效果差，所以用GMM模型再进行一遍聚类，将kmeans和GMM聚类后重复的index进行去重，然后将GMM和kmeans聚类后对每一簇的标签进行统一。

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

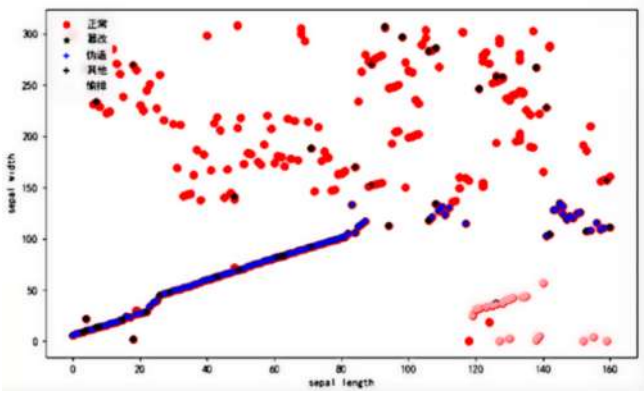

图2　kmeans聚类分布

高斯混合模型(Gaussian Mixture Model)通常简称GMM,是一种业界广泛使用的聚类算法,属于生成式模型,它假设所有的数据样本都是由某一个给定参数的多元高斯分布所生成的。从中心极限定理知,只要给定类个数K足够大、模型足够复杂、样本量足够多,每一块小区域就可以用高斯分布描述。GMM的概率密度函数如下:

$$P_M(x)=\sum_{k=1}^{k}p(k)\,p(x|k)=\sum_{k=1}^{K}\alpha_k\,p\left(x|\mu_k,\sum K\right)$$

其中,k为模型的个数,即聚类的个数;α_k为属于第k个高斯的概率(也称为先验分布),其需要满足大于零,且对一个x而言,α_k之和等于1;$p(x|k)$为第k个高斯的概率密度,其均值向量为μ_k,$\sum K$为协方差矩阵;K由人工给定,其他参数需要通过EM算法进行估计。GMM二次聚类结果如图3所示。

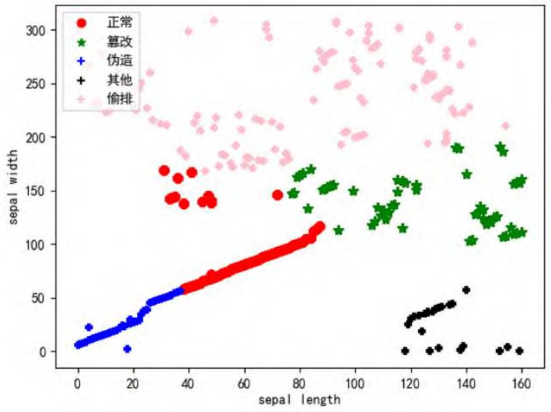

图3 GMM二次聚类分布

企业正常排污数据应接近正常阈值,偷排概率高的数据偏移距离大且分布在质心之上,篡改概率高的数据增长式差异明显,但偏移距离小于偷排数据,伪造数据排放浓度明显下降,分布在质心之下,未分类数据属于其他违法行为。

依照陈天奇博士于2016年提出的XGBoost算法设置训练的目标函数,XGBoost使用了一种新的正则化技术,控制过拟合现象的产生。因此,在模型调整期间,XGBoost会更快、更健壮。此处的数学公式可写为:

$$L(f)=\sum_{n=1}^{N}L\left(\widehat{y_n},y_n\right)+\sum_{m=1}^{N}\Omega\left(\delta_m\right)$$

其中,$\Omega(\delta)=\alpha|\delta|+0.5\beta||\omega||^2$;$|\delta|$为树枝的数量;$\omega$为每片叶子的值;$\Omega(f)$为正则化函数。XGBoost的基本核心算法流程如下:

Step1:在每次迭代中添加一个新树;

Step2:每次迭代计算$\frac{\partial l\left(y_i,\hat{y}_i^{(t-1)}\right)}{\partial\hat{y}_i^{t-1}}$和$\frac{\partial^2 l\left(y_i,\hat{y}_i^{(t-1)}\right)}{[\partial,\hat{y}_i^{(t-1)}]^2}$

Step3:使用新数据来生成树f_t,$Obj=-\frac{1}{2}\sum_{J=1}^{T}\frac{G_j^{\,2}}{H_j+\lambda}+\gamma T$

Step4:使用$\hat{y}_i^{(t)}=\hat{y}_i^{(t-1)}+\varepsilon f(x_i)$,$\varepsilon$选在0.1左右,防止过度拟合。

选取327 780条数据记录,即70%的数据用于训练模型,利用XGBoost模型创建分类器拟合分类器模型,调用训练好的预测数据属于每一类的概率。预测结果符合预期,即大部分数据应该是正常的概率最高,部分预测结果如图4所示:

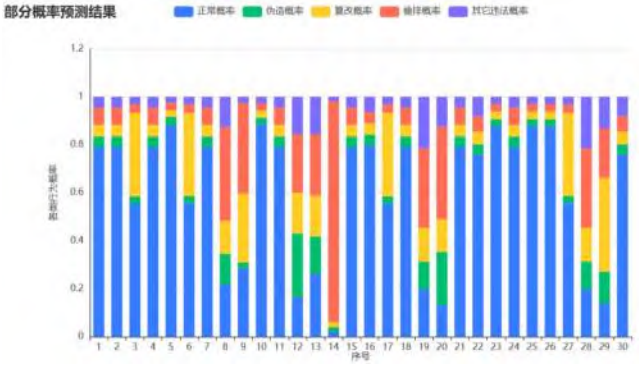

图4 部分概率预测结果

3 模型性能测试与结果分析

在训练模型时,有70%的数据用作训练集,30%用作测试集,通过使用机器学习中评价模型常用的平均绝对误差(MAE)、均方根误差(RMSE)与平均绝对百分比误差(MAPE)作为XGBoost模型、GBDT和随机森林模型的性能评价指标,对企业排污行为违法概率的预测结果进行评价,结果如表3所示,得出XGBoost预测模型在平均绝对误差、均方根误差和平均绝对百分比误差方面均优于传统的随机森林模型,说明XGBoost模型具有更好的可行性。

表3　误差对比

	MAE	RMSE	MAPE/%
XGBoost	1.33	2.58	5.65
GBDT	1.98	3.21	8.28
随机森林	3.72	4.86	12.10

对各模型判定的企业排污数据正常概率进行对比,可以看出XGBoost预测结果最接近真实数据,而GBDT、随机森林次之,如图5所示:

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

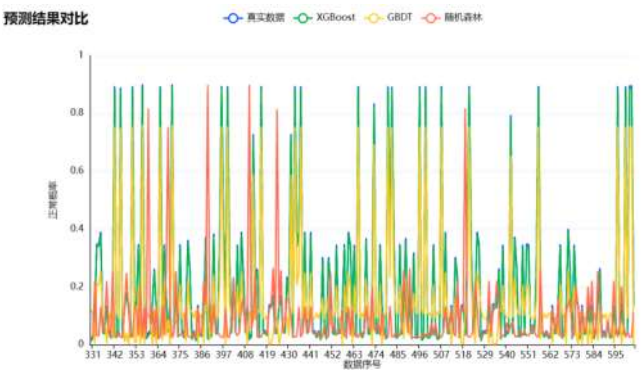

图5　XGBoost与其他模型效果对比

4 结束语

本文填补了XGBoost算法在企业排污违法概率评估方面的应用研究，基于河北省企业历史两年内排污数据，参照相关法律条文及文献制定数据超标及可疑数据的判定标准，运用大数据技术对数据进行预处理，结合机器学习聚类算法无监督特征提取的特点，对大量数据进行分析与分类，提取各类可疑违法数据的特征，尝试将XGBoost算法运用于预测企业排污违法概率中。从实验结果可以看出，XGBoost算法应用于预测排污数据违法概率是可行的，且性能较传统机器学习算法更优，但由于XGBoost算法在评估污染排放违法概率的应用研究较少，本文没有对模型进一步优化，将来可以进一步改进模型来提升模型精度。

参考文献：

[1] 王军霞,刘通浩,张守斌,等.排污单位自行监测监督检查技术研究[J].中国环境监测,2019,35(2):23-28.
[2] 魏艳,赖静娴,周启龙,等.污染源自动监测异常数据识别规则及处理方法探索[J].环境监测管理与技术,2022,34(2):56-59.
[3] 陈冲,何为,钟田福,等.基于孤立森林方法的催化裂化装置排污数据异常识别[J].西安石油大学学报(自然科学版),2021,36(4):119-126.
[4] Chen T Q,Guestrin C.XGBoost:a scalable tree boosting system [EB/OL].2016:arXiv:1603.02754.https://arxiv.org/abs/1603.02754.
[5] Rosa G J M.The elements of statistical learning:data mining,inference,and prediction by HASTIE,T.,TIBSHIRANI,R.,and FRIEDMAN,J[J].Biometrics,2010,66(4):1315.
[6] 张旭春.基于CatBoost模型实现对污水处理厂排污情况的监测预警[D].兰州:兰州大学,2021.
[7] 王为久,徐敏亚,徐博希,等.基于XGBoost算法的非法经营罪量刑预测模型构建及应用[J].情报探索,2022(9):20-28.
[8] Shrestha S M,Shakya A.A customer churn prediction model using XGBoost for the telecommunication industry in Nepal[J].Procedia Computer Science,2022(215):652-661.

【通联编辑：唐一东】

（上接第17页）

[6] 程志明.基于UNET的医学图像分割研究[D].衡阳:南华大学,2021.
[7] Carballido-Gamio J,Belongie S J,Majumdar S.Normalized cuts in 3-D for spinal MRI segmentation[J].IEEE Transactions on Medical Imaging,2004,23(1):36-44.
[8] Chung D H,Sapiro G.Segmenting skin lesions with partial differential equations based image processing algorithms[C]//Proceedings 2000 International Conference on Image Processing (Cat.No.00CH37101).September 10-13,2000,Vancouver,BC,Canada.IEEE,2002:404-407.
[9] Nguyen H T,Worring M,van den Boomgaard R.Watersnakes:energy-driven watershed segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2003,25(3):330-342.
[10] Xie J,Jiang Y F,Tsui H T.Segmentation of kidney from ultrasound images based on texture and shape priors[J].IEEE Transactions on Medical Imaging,2005,24(1):45-57.
[11] 石磊,籍庆余,陈清威,等.视觉Transformer在医学图像分析中的应用研究综述[J].计算机工程与应用,2023,59(8):41-55.
[12] Shelhamer E,Long J,Darrell T.Fully convolutional networks for semantic segmentation[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence.May 24,2016,IEEE,2016:640-651.
[13] Ronneberger O,Fischer P,Brox T.U-net:convolutional networks for biomedical image segmentation[M]//Lecture Notes in Computer Science.Cham:Springer International Publishing,2015:234-241.
[14] Zhou Z W,Rahman Siddiquee M M,Tajbakhsh N,et al.UNet++:A nested U-net architecture for medical image segmentation [M]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support.Cham:Springer International Publishing,2018:3-11.
[15] Piao S Y,Liu J M.Accuracy improvement of UNet based on dilated convolution[J].Journal of Physics:Conference Series,2019,1345(5):052066.
[16] Wunderling T,Golla B,Poudel P,et al.Comparison of thyroid segmentation techniques for 3D ultrasound[C]//SPIE Proceedings,Medical Imaging 2017:Image Processing.Orlando,Florida,USA.SPIE,2017.
[17] Zhou Z W,Siddiquee M M R,Tajbakhsh N,et al.UNet:redesigning skip connections to exploit multiscale features in image segmentation[J].IEEE Transactions on Medical Imaging,2020,39(6):1856-1867.
[18] 李春林,赵翠,司迁,等.智慧医疗的发展现状与未来[J].生命科学仪器,2021,19(2):4-13.

【通联编辑：唐一东】

(C)1994-2023 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net