DOI:10.19651/j.cnki.emt.2105718

基于Stacking室内污染气体感知与评价系统*

赵艳茹 陈向东 丁 星
(西南交通大学 信息科学与技术学院 成都 611756)

摘 要：设计了一种基于Stacking集成学习的室内污染气体感知与评价系统。用户可以使用手机APP扫描检测终端的二维码,查看当前环境污染气体的实时数据及评价结果。该系统利用Stacking将评价算法与分类算法集成学习,解决评价算法主观性强的问题。首先选用模糊数学综合评价法、决策树和KNN作为基础模型分别进行训练,然后将3个基础模型的输出结果作为特征值,利用逻辑回归作为元模型对3个基础模型进行异质集成。该方法在保证系统实时性的同时有效地提高了评价结果的准确性和客观性。

关键词：Stacking;模糊评价;KNN;决策树
中图分类号：TN98 **文献标识码**：A **国家标准学科分类代码**：510.99

An indoor pollution gas sensing and evaluation system based on Stacking

Zhao Yanru Chen Xiangdong Ding Xing
(School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: An indoor pollution gas sensing and evaluation system based on Stacking integrated learning is designed. Users can use mobile phone APP scan the QR code of the detection terminal to view the real-time data and evaluation results of the current environmental pollution gas. By using Stacking, the system integrates evaluation algorithm with classification algorithm to solve the problem of strong subjectivity of evaluation algorithm. Firstly, fuzzy mathematics comprehensive evaluation method, decision tree and KNN are selected as the basic models for training, and then the output results of the three basic models are taken as eigenvalues, and the logical regression is used as the meta-model to integrate the three basic models heterogeneously. This method can improve the accuracy and objectivity of the evaluation results while ensuring the real-time performance of the system.

Keywords: Stacking; fuzzy mathematics; KNN; decision tree

0 引 言

当代社会室内环境污染问题成为危害人们身体健康安全的主要原因之一。甲醛作为室内环境的主要污染源,危害极大,它会导致呼吸道感染,诱发肺炎,严重时会导致癌症等疾病,被世界卫生组织(IARC)列入致癌物清单[1]。同时,相关研究表明,近几年我国儿童白血病发病率增高的主要原因之一可能是室内环境甲醛暴露。除此之外,Weng等[2]调查了杭州地区公共场所污染情况,调查表明,甲醛在污染物中占比高达61.6%。因此,发展实时检测并客观评价的室内污染气体感知与评价系统有很大的必要性。

巫春玲等[3]设计了基于WiFi的室内空气质量数据采集系统,但是采集指标过少,没有对采集数据进行定性分析;蔡倩等[4]研究了基于WSN的多通道室内环境智能评价方法,结合遗传神经网络算法在MATLAB中建立评价模型,但是评价模型单一,未考虑评价方法的主客观性。因此,改进评价算法主观性强的问题很有必要性。

针对现有的评价类算法主观性强等问题,本文设计了一种基于集成学习的室内污染气体感知与评价系统。集成学习是通过一个元分类器或元回归器来整合多个分类模型或回归模型的集成学习技术[5],它能够降低算法过拟合风险,提高集成的泛化能力以及提高准确率。该系统在医院等大型场所具有一定的研究意义和应用意义。

1 系统总体设计

对于当下学校、办公楼、商场等大型场所缺少对室内环境污染水平的监测、用户无法实时获取当前室内环境污染水平的情况,本系统面向用户提供能够实时检测、数据透

收稿日期:2021-01-19
*基金项目:西南民族大学中央高校基本科研业务费电子信息工程专项(2020PTJS19002)资助

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

明、结果可视化的室内环境污染水平客观评价系统。

1.1　系统组成

系统总共由 4 个部分组成，分别是硬件检测终端、服务器端、Browser 浏览器端和手机 APP 端，如图 1 所示。硬件检测终端采集当前环境内的温湿度、甲醛、PM2.5、PM10、氨气等数据，并使用 WIFI 协议与云服务器通信。服务器端是系统的核心，负责数据的传输、计算和存储，是硬件检测终端和 APP 端进行数据通信的桥梁[6]。Browser 浏览器端管理全部硬件检测终端和历史数据。APP 端可以查阅当前节点的实时数据及污染气体感知等级。

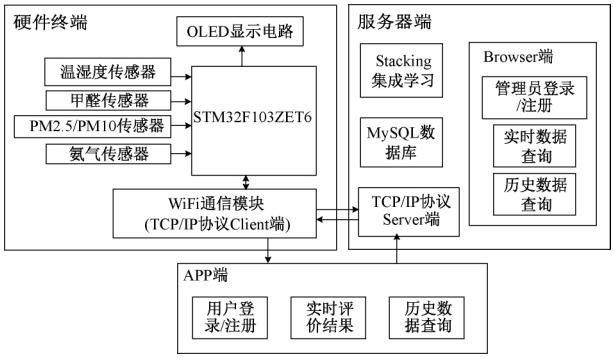

图 1　系统整体结构

1.2　系统工作流程

系统主要工作流程如下：

1）硬件检测终端初始化各模块，连接局域网 WiFi，使用 TCP 通信连接云服务器端，请求发送数据。

2）服务器端接收请求，并向硬件检测终端发送允许连接的指令，将接收的实时数据缓存服务器端的 Redis 数据库。通过实时计算将检测数据及评价结果存储至 MySQL 数据库，并发送至 Browser 端。

3）用户使用 APP 扫描硬件检测终端的二维码，查看当前节点的实时数据及污染气体感知等级。

2　系统设计与实现

2.1　系统硬件设计

系统的硬件部分主要是硬件检测终端的设计，包括以 STM32F103ZET6 芯片为核心的最小系统、OLED 显示电路、无线 WiFi 通信电路和多种检测因子传感器电路[7]，传感器电路分别对温度、湿度、甲醛、PM2.5、PM10 和氨气 6 个指标进行检测。检测终端主要实现多传感器数据的采集和传输功能。

2.2　系统软件设计

1）检测终端软件设计：检测终端的软件部分主要实现模块初始化、各传感器数据采集、原始数据处理、通信模块配置、数据发送与接收等功能。

2）云服务器端软件设计：云服务器端主要实现与 APP 端和硬件检测终端通信功能。使用 Spring 技术实现云服务器，便于服务器管理数据库、控制 Browser 浏览器端页面显示及跳转。并使用 TCP/IP 协议连接多个检测终端。

3）Browser 浏览器端：后台管理系统对局域内的所有设备进行管理，使用可视化图形展示设备的历史检测数据以及评价结果。

4）手机 APP 端：手机 APP 端通过扫描检测终端上的二维码，获取当前终端的检测数据及评价结果，并针对当下的空气质量情况给用户适当的防护建议。

3　基于 Stacking 评价算法的设计与实现

现代综合评价方法有主成分分析法、数据包络分析法和模糊综合评价法等[8]。其中主成分分析法中指标必须要有实际背景和意义。数据包络法无法表明评价指标的实际水平。模糊综合评价法未考虑多指标间信息相互影响的问题。以上的综合评价方法都有一个共同的缺点，即评价结果主观性强。针对这一问题，考虑到该系统的数据组的特征，采用评价算法与分类算法集成的方式，在降低评价算法的主观性的同时，提高评价结果的准确性。

不同的集成模型各有不同的优缺点，Bagging 模型[9]具有容易过拟合的缺点，Boosting 模型具有迭代次数不易设定，并且对离群点较敏感的缺点[10-11]，而 Stacking 集成模型属于异质集成，个体学习器之间没有依赖性，不会出现过拟合的现象，更加符合本系统的设计，因此本文选取 Stacking 集成模型进行集成学习。后续本章将对该算法分成 3 部分介绍：Stacking 集成思想、各分类模型实现过程、算法验证。

3.1　Stacking 集成思想

本文使用一种评价算法及两种分类算法作为 Stacking 集成学习的分类模型，使用平均法作为结合策略[12]。首先，针对评价类算法具有主观性较强的问题，通过将评价算法与分类算法相结合的方式解决降低评价结果主观性强的问题。其次，使用平均法作为结合策略，是为了在保证算法的准确率同时防止出现过拟合的风险，并提高集成的泛化能力。Stacking 集成学习结构如图 2 所示。

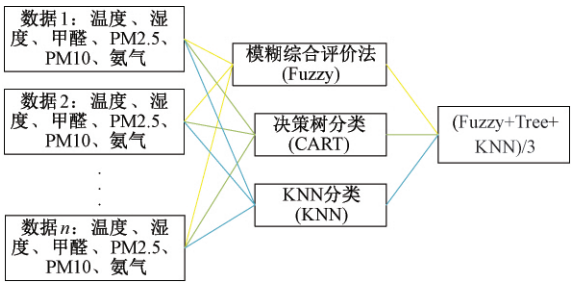

图 2　Stacking 集成学习结构

首先使用模糊综合评价法对检测数据进行计算获得评价结果，然后对训练数据分别使用决策树和 KNN 分类，分类结果即为评价结果，最后使用简单平均法对 3 组评价结

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

果进行集成[13],获得最终的评价结果。

3.2 各分类模型实现过程

本文采用评价算法和分类算法作为 Stacking 集成的分类模型。首先,评价算法采用模糊综合评价法。其次,选取 5 种分类算法分别对 1 000 组训练数据进行训练,参考分类准确率及算法计算时间选取两个较优的分类算法。通过表 1可以看出决策树及 KNN 在保证分类准确率的同时训练时间较短,能够有效保证本系统的实时性。所以本文选用决策树及 KNN 作为分类模型算法。

表 1 不同算法训练性能对比

算法名称	准确率/%	训练时间/s
决策树(CART)	99.0	1.06
KNN	97.9	0.71
AdaBoost	90.7	3.92
Bayes 分类器	87.6	0.96

1)模糊综合评价法

模糊综合评价法基本步骤如下:

(1)确定评价指标 u 及评价指标值 U:$U=\{u1,u2,u3,u4,u5,u6\}=$ {温度,湿度,甲醛,PM2.5,PM10,氨气};

(2)确定评价值集合 V:$V=\{v1,v2,v3,v4,v5\}=$ {优,良,中,差,严重};

(3)确定各评价指标的权重 $\boldsymbol{M}$:使用熵值法、Critic 赋值法对训练数据组进行分析获得权重组,结合层次分析法对 3 组权重做加权平均获得最终权重。

$\boldsymbol{M}=[a1,a2,a3,a4,a5,a6]=[0.137,0.172,0.259,0.127,0.119,0.183]$

(4)采用三角形隶属函数对各评价指标构建隶属度函数。以甲醛为例,甲醛指标的隶属度函数关系如图 3 所示,其中,x 表示甲醛检测值,单位为 mg/m^3,y 表示评价结果的隶属度值。

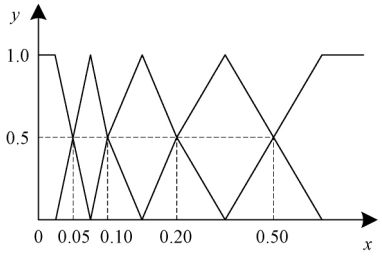

图 3 甲醛指标隶属度函数

(5)将检测数据代入各指标的隶属度函数计算获得模糊矩阵 $\boldsymbol{R}$。$\boldsymbol{R}$ 为 6×5 的矩阵。

(6)计算获得模糊评价 $\boldsymbol{B}=\boldsymbol{M}\cdot\boldsymbol{R}$。

(7)对模糊评价结果 $\boldsymbol{B}$ 解模糊化分析。本系统解模糊化分析使用最大隶属度法,具有计算简单,实时性强的特点。

2)决策树(CART)

构建初始数据集 $D=\sum_{i=1}^{1\,000}d_i$,$d_i=$ {温度,湿度,甲醛,PM2.5,PM10,氨气}。初始属性集 $A=\{a_1,a_2,a_3,a_4,a_5\}=$ {优,良,中,差,严重}。

CART 算法流程如图 4 所示。

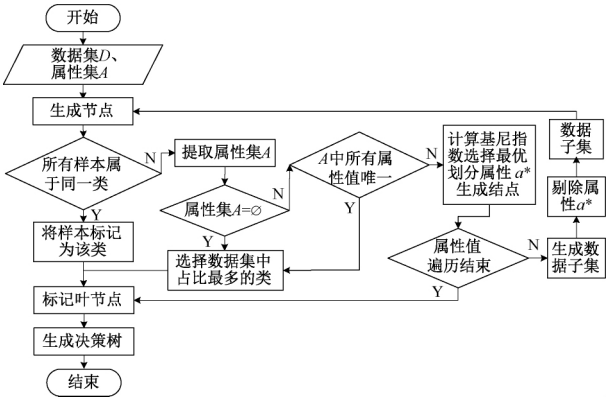

图 4 CART 算法流程

CART 算法的核心是计算当前节点的基尼指数并选择最优划分属性[14-15]。对于数据集 D 有 1 000 组不重复样本数据,属性集 A 中有"优"、"良"、"中"、"差"、"严重"5 个属性值,根据属性集 A 中属性的个数将数据集 D 分为 V 组数据集 $\{D^1,D^2,\cdots,D^V\}$,若 A 中有 5 个属性值则 V 取 5。在属性"优"的条件下,样本 D 的基尼系数表达式如式(1)所示,其中,$|D^V|$ 和 $|D|$ 分别代表对应数据集中的样本数。

$$Gini(D,a_1)=\sum_{V=1}^{V}\frac{|D^V|}{|D|}Gini(D^V) \tag{1}$$

式中:$Gini(D^V)$ 为数据集 D^V 的基尼系数。以数据集 D^1 为例,数据集 D^1 的基尼系数表达式如式(2)所示。其中 k 的取值范围为 1~5。

$$Gini(D^1)=1-\sum_{k=1}^{k}\left(\frac{|C_k|}{|D^1|}\right)^2 \tag{2}$$

假设数据集 D^1 中有 3 种类别,则 k 取 3,$|C_k|$ 为 D^1 中各种类别的样本数。不同的属性值对应不同的基尼系数,基尼指数越小越好,因此选择划分后基尼指数最小的属性作为最优划分属性,即:

$a^*=\min_{a\in A}Gini(D,a)$。

若属性"优"的基尼系数最小,则将属性"优"从 A 中剔除获得新的 A',同时将属性为"优"的数据从数据集 D 中剔除获得数据子集 D'。对 D' 和 A' 再次计算,直到属性集为空集。

3)KNN

KNN 算法模型训练的基本步骤如下:

(1)构建训练集。为保证 KNN 分类结果的准确性,选取 1 000 组数据,有 5 个类别,每个类别需要有 200 组数据来构成训练集。

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

(2)遍历测试集中的数据，计算当前数据 x_0 到训练集中每个数据点的距离。由于该步骤计算量较大，考虑的本系统的实时性，采用计算方式较为简单的欧氏距离进行计算，如式(3)所示。

$$dis(a,b)=\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2} \tag{3}$$

式中：$n=6$，$a_1 \sim a_6$ 分别代表 6 种检测因子对应的检测数据。

(3)找出距离 x_0 最近的 k 个点，将 x_0 分类到这 k 个点中最多的类。k 应取奇数。通过对 1 000 组实验数据做交叉验证，当 $k=3$ 时，KNN 分类准确率最大，为 97.41%，如图 5 所示。

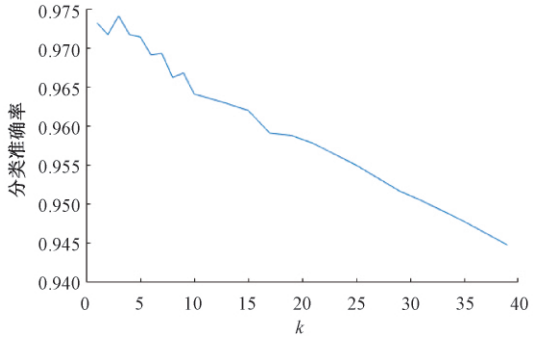

图 5　算法准确率与 k 值变化曲线

3.3　算法验证

多次向实验箱通入不同污染因子，模拟不同的空气质量水平，对 5 种不同的等级环境分别采集 20 组实验数据，共 100 组实验数据作为测试集，并对实验数据进行标记。根据各污染因子的检测值与《室内空气质量标准》中标准值的差值大小，将数据划分为优、良、中、差、严重 5 类，其中优、良代表符合国家标准线以内的 2 种污染水平，中、差、严重代表超标的 3 种污染水平。《室内空气质量标准》部分参数标准如表 2 所示。

表 2　《室内空气质量标准》部分参数

参数	单位	标准值
温度	℃	16～28
湿度	%	30～80
甲醛	mg/m^3	0.10
PM2.5	mg/m^3	0.75
PM10	mg/m^3	0.15
氨气	mg/m^3	0.20

对测试集中的 100 组数据进行计算，将标记值作为实际值对各算法的评价结果准确率进行判断。所有测试数据测试结果如图 6～9 所示。

测试集部分样本数据如表 3 所示。从表 3 中可以看出，当数据组中各指标数值在某两种等级边界处时，

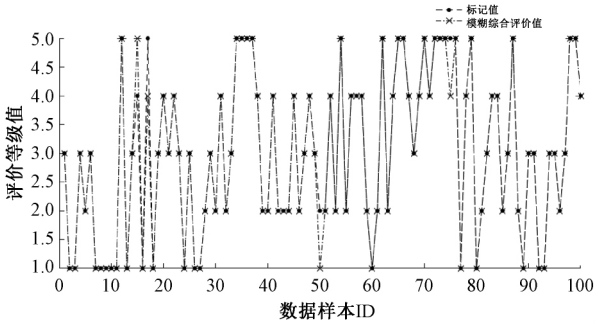

图 6　模糊综合评价结果与标记值对比

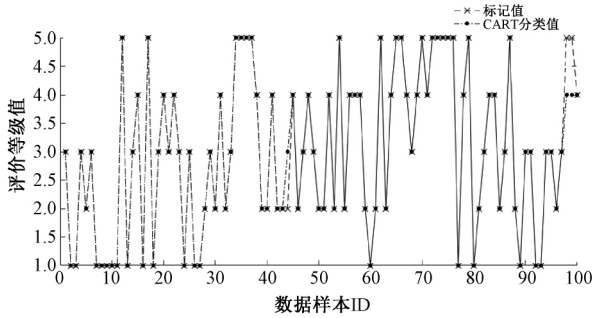

图 7　CART 分类结果与标记值对比

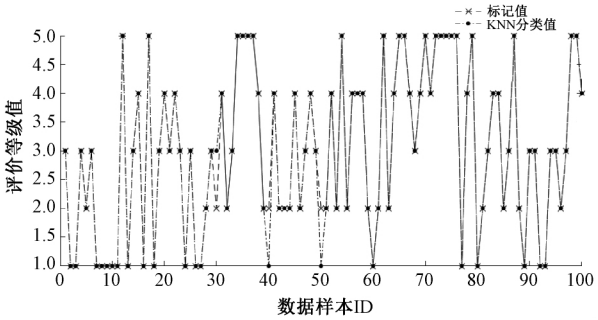

图 8　KNN 分类结果与标记值对比

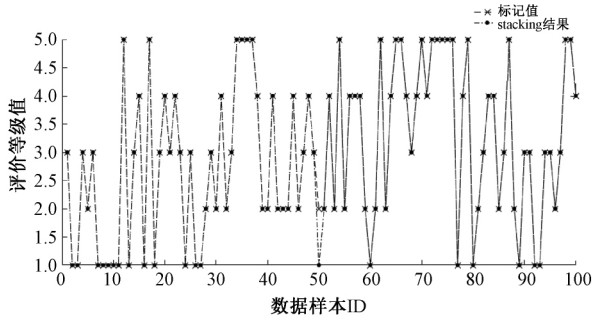

图 9　Stacking 集成结果与标记值对比

Stacking 集成与前 3 种算法比较，能更有效地区分数据组所属类别，同时，Stacking 的评价结果比前3 种算法更具客观性。

对同一组测试集进行实验得到每种算法的运行时间和准确率，模糊综合评价法运行时间大约 0.33 s、准确率为

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

表 3 测试集中部分样本数据

温度/℃	湿度/RH	甲醛/(mg·m^{-3})	PM2.5/(mg·m^{-3})	PM10/(mg·m^{-3})	氨气/(mg·m^{-3})	模糊综合评价法	KNN	CART	Stacking
22.21	53.54	0.035	0.784	1.002	0.297	严重	差	严重	严重
20.23	70.02	0.038	0.780	0.994	0.145	严重	差	差	差
20.50	60.20	0.016	0.066	0.094	0.234	中	良	中	中
21.30	50.21	0.090	0.006	0.012	0.193	良	良	中	良
21.89	50.09	0.011	0.040	0.063	0.154	良	优	良	良
20.44	68.55	0.046	0.045	0.060	0.136	良	优	优	优
20.50	68.10	0.054	0.044	0.055	0.138	良	良	优	良

96%;CART 决策树算法运行时间约为 0.13 s、准确率为 97%;KNN 算法运行时间约为 0.14 s、准确率为 98%;通过 Stacking 集成后,总运行时间约为 0.61 s,准确率提高至 99%。虽然 Stacking 集成付出时间代价略高于模糊综合评价法,但是 Stacking 集成能够大幅提高准确率,并且评价结果更客观。

4 系统测试

本节针对系统功能完整性进行测试。首先打开服务器端,监听检测终端的状态,开启检测终端,连接局域网与服务器端,成功连接后开始进行检测,如图 10 所示。

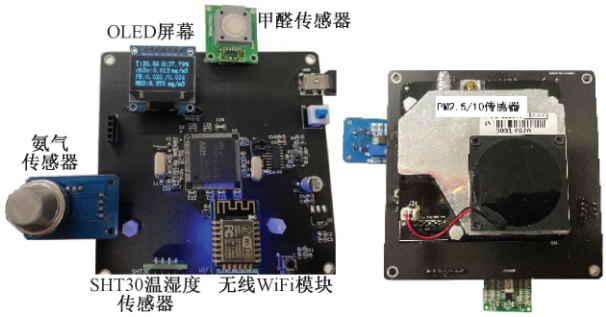

图 10 硬件监测终端工作正面及背面

后台管理系统通过登录管理员账户可以查看所有终端的历史检测数据及评价结果。用户可以通过扫描终端二维码,获取当前检测数据及评价结果。测试中各界面如图 11~13 所示。

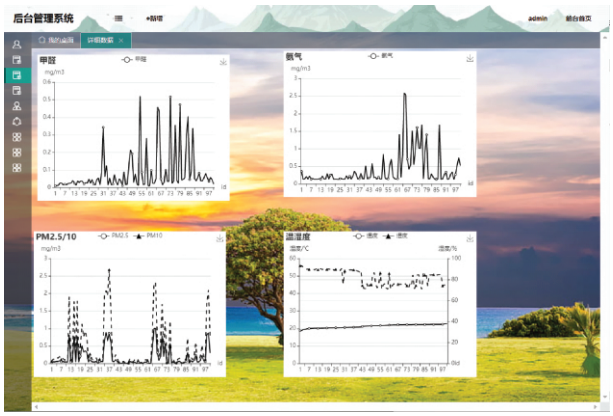

图 11 历史数据详情

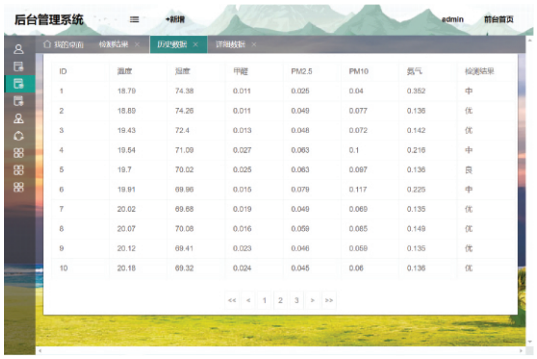

图 12 历史数据及评价结果详情

图 13 APP 端实时数据查看

5 结 论

基于 Stacking 室内污染气体感知与评价系统能够对局域网内的环境进行实时检测,使用 Stacking 集成学习对不同算法的输出结果做二次集成,使系统的评价结果更加客观且准确。管理员和用户能够在线上平台直观地查看当前室内环境污染水平。针对环境污染水平设计了实验对系统进行测试,实验结果表明,该系统检测环境实时性强,数据传输可靠,环境污染水平评价结果客观准确。本文设计的室内污染气体感知与评价系统虽然检测指标个数有限,但充分考虑了指标间的相互影响,具有更客观的评价结果,在大型公共室内场所具有潜在的应用意义。

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

参考文献

[1] 谭和平,马天,方正,等.室内挥发性有害有机物限量标准研究[J].中国测试技术,2006(5):8-15.

[2] WENG M L, ZHU L Z, YANG K, et al. Levels and health risks of carbonyl compounds in selected public places in Hangzhou, China[J]. Journal of Hazardous Materials, 2009, 164(2-3):700-706.

[3] 巫春玲,冯志文,任凯,等.基于 WiFi 的室内空气质量数据采集系统设计[J].建筑电气,2020,39(11):50-53.

[4] 蔡倩,刘奇,顾敏明.基于 WSN 的多通道室内环境智能评价研究[J].物联网技术,2020,10(11):91-93.

[5] 周志华.机器学习[M].北京:清华大学出版社,171-190.

[6] 冉启成,陈向东,张传武,等.支持向量机的端到端共享甲醛检测系统[J].单片机与嵌入式系统应用,2020,20(7):60-64.

[7] 颜鑫,陈向东.基于手机 APP 的 O2O 住房空气质量测评系统[J].单片机与嵌入式系统应用,2019,19(7):51-55.

[8] YUAN J, CHEN Z, ZHONG L, et al. Indoor air quality management based on fuzzy risk assessment and its case study[J]. Sustainable Cities and Society, 2019,50:101654.

[9] 严智,张鹏,谢川,等.一种快速 AdaBoost.RT 集成算法时间序列预测研究[J].电子测量与仪器学报,2019,33(6):82-88.

[10] 金聪,金枢炜.面向图像语义分类的视觉单词集成学习方法[J].电子测量技术,2012,35(8):53-56.

[11] 潘国兵,龚明波,贺民,等.基于 Stacking 模型融合的专变用户电费回收风险识别方法[J].电力自动化设备,2021,41(1):152-160.

[12] 高尚,唐元合,翟明玉,等.基于集成学习的输变电设备数据质量检测方法[J].电子测量技术,2020,43(2):108-112.

[13] 夏雨薇,石美红,贺飞跃,等.基于降维融合特征和集成学习的织物疵点分类[J].国外电子测量技术,2019,38(7):86-91.

[14] 石欣,范智瑞,张杰毅,等.基于 LMS-随机森林的肌电信号下肢动作快速分类[J].仪器仪表学报,2020,41(6):218-224.

[15] LEONG W, KELANI R, AHMAD Z. Prediction of air pollution index(API) using support vector machine(SVM)[J]. Journal of Environmental Chemical Engineering, 2020,8:103208.

作者简介

赵艳茹,硕士研究生,主要研究方向为无线传感网络技术。

E-mail:1774985656@qq.com

陈向东,教授,博士生导师,主要研究方向为新型传感器与智能信息获取。

E-mail:xdchen@swjtu.edu.cn

丁星,助理研究员,主要研究方向为传感器与信息获取技术。

E-mail:xding@swjtu.edu.cn

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