机器学习差异融合分析运用在空气质量预测中

叶春阳
中国人民大学 北京 100872

摘　要：在社会经济的高速发展之下，工业的发展对我国环境污染带来的危害也逐年增大，其中大气污染问题极为突出。空气质量（AQI）是依据空气中污染物浓度的高低来进行判断的，能够反映城市内空气污染的程度，对空气质量进行预测主要目的就在于帮助环境监测部门根据城市内的空气质量状况来完善环境管理政策。目前对空气质量的预测多数采用大数据与云计算平台的方式，而本文采用机器学习差异融合分析的方法对空气质量进行建模分析，并考虑与空气质量相关的因素对空气质量的发展趋势进行分析预测，其预测精度比传统预测方法更高。
关键词：机器学习；差异融合分析；空气质量预测

Application of machine learning difference fusion analysis in air quality prediction

Chunyang Ye
Renmin University of China, Beijing 100872

Abstract: With the rapid development of social economy, the harm of industrial development to China's environmental pollution is also increasing year by year, among which the problem of air pollution is extremely prominent. Air quality (AQI) is judged according to the concentration of pollutants in the air, which can reflect the degree of air pollution in cities. The main purpose of air quality prediction is to help environmental monitoring departments improve environmental management policies according to the air quality in cities. At present, most of the air quality predictions are based on big data and cloud computing platforms. In this paper, the machine learning difference fusion analysis method is used to model and analyze the air quality, and the factors related to air quality are considered to analyze and predict the development trend of air quality. The prediction accuracy is higher than that of traditional prediction methods.
Key words: machine learning; Difference fusion analysis; Air quality prediction

引言：

近年来，在全球高新技术爆炸式发展的同时，对大气环境带来的污染危害也达到一个峰值，造成了严重的空气污染。我国最为世界上最大的发展中国家，在防控空气污染方面更需加大力度，同时还需加强对环境的治理。在这种状况之下，对空气质量进行预测能够为有关部分制定环境保护决策提供有效可靠的依据。到2022年，我国重点城市内的 AQI 指数下降了 10% 作用，空气质量优良的天数逐步上升。在对空气质量的预测中，常采用结合机器学习的方法，系统迅速地将获取到的空气质量数据整合分析，从环境特征角度探究空气质量的变化，本文在机器学习的基础之上结合了差异融合分析法，根据数据建立模型来进行预测，提高预测精度。

一、机器学习概述

机器学习是建立在数据集合的基础建立样本模型来进行预测或决策，主要依赖于模式和推理算法和统计模型，与计算机统计学密切相关，侧重于对研究项目进行预测，因此机器学习也被成为预测分析。机器学习的任务主要分为监督学习、半监督学习和无监督学习。在监督学习当中，需运用算法从输入与输出的数据当中分析并建立模型，其中分类算法和回归算法都属于监督学习，应用回归算法可建立回归模型，来确定变量之间的定量关系，由此对未知参数进行预测。回归主要就在于预测数值型的目标值。分类算法则在于将数据分类归档，来输出预测值，回归算法所能够预测的范围比分类算法更广泛。在无监督学习当中，重点在于针对输入的数据来建立数学模型，并且对数据点进行分组或聚类。在机器学习的其他算法当中还包括主动学习算法和元学习算法等等，主动学习算法能够根据输入数据来访问所需输出并对输入进行优化，元学习算法则能够进行归纳偏差，生成学习体验序列，更加智能成熟。机器学习的方法在空气质量预测中的应用以比较成熟，但是依然存在较多的未探索领域，应用该方法能够加快对空气质量预测体系的建立，将预测过程标准化，为环境保护提供服务。

二、空气质量预测

2.1 空气质量预测与技术发展

目前，环境污染问题是人们讨论的热点话题，其中讨论最多的便是在工业发展当中所造成的大气污染问题，污染物的排放所造成的污染对自然界和人类健康造成了严重的影响，因此针对大气污染问题的研究迫在眉睫。由于空气质量的变化受到诸多因素的综合影响，传统的研究过程耗时费力，因此在结合大数据技术的条件之下对数据进行分析，能够快速地找出影响空气质量的重要因素和空气质量的变化状况。但是为了能够更好地制定环境保护决策，需要对空气质量进行预测，以便相关工作人员能够做出应对措施进行防控。为了防止出现大面积空气污染的时间，美日德等发达国家早在 1960 年就开始了空气质量预测的相关研究，以高斯模型和拉格朗日模型等为理论基础构建一套空气质量模拟模型。随着对模型建立的研究与发展，目前空气质量模型的建立多以数据集合为基础利用统计学的方式来建立集成模型，并且形成了多模块集成与嵌套的三维模型，从多方面多角度来进行模拟分析，其模型结构复杂但是预测精度却明显提高。随后，空气质量的预测一直以建立统计学模型为主、随着科学技术的发展，基于机器学习的方式为空气质量预测提出了分类回归树模型、回归模型、人工神经网络模型等模型，并在之后的研究当中将深度学习法与集成模型相结合，提高的计算效率与预测精度[1]。

2.2 空气质量预测研究状况

在当前的空气质量预测当中，主要分为数值预测与统计预测，其中以建立统计预测模型为主，能够抽象具体地反映数据变化。国外对于空气质量预测的研究早在 1998 年便是用神经网测模型来对大气污染物进行预测，有效对空气中的化学成分及浓度进行分析，由此来分析造成空气污染的主要因素。在预测当中，神经网络模型比多变量回归模型更加精确。另外，在对数据的统计方面，有采用协议指数、平方相关系数和分数偏差的方式来建立数据库，在数据库的基础上来建立非线性模型来对大气的空气状况进行监测与预测。同样，基于模糊推理原理的神经网络模型的建立以及运用遗传算法对空气质量影响因素进行计算，能够提高空气质量预测精度。在我国，对于空气质量预测的研究在近年来才得到重视，使用粗糙集理论利用 BP 算法来进行神经网络模型，比起传统的神经网络算法能够收拢更多的数据，适于应用在城市空气质量预报中。在城市中大部分空气质量的数据是呈非线性的，也就是说传统的算法无法对数据进行精确的分类分析，需应用智能优化算法对数据进行整理分析，以统计模型的方式来进行预测[2]。差异融合分析的方法在空气质量预测中的应用主要就在于通过循环计算对融合阈值的选取来提高预测精度，能够对空气质量变化较为显著的城市做出更加精确的空气质量预测。

三、数据的分析与处理

3.1 数据处理算法

在应用机器学习来建立模型对空气质量进行预测时，由于影响空气质量的因素诸多，因此需对多方的数据进行运算整理，分析空气相关特征的计算数值与空气质量数据之间的关系，再通过建模来进行预测。

3.1.1KNN 算法

KNN 算法是数据挖掘分类中的常用算法，能够判断未知样本的类别并计算未知与已知之间距离，在对空气质量的数据进行处理时，首先需要对空气质量数据样本的所有因素特征进行量化，并对特征进行归一化处理。由于空气质量样本数据的每一个参数尤其自身的定义域和取值范围，因此需进行 scale 处理，进行归一化，再规定一个距离函数来计算两个相邻样本之间的距离，并按照升序将计算结果排序。KNN 算法的重点就在于能够确定样本数据之间的距离，随后选取距离最小的 K 个点来确定前 K 个点所在类别的出现频数，将出现频数最高的类别来作为当前点的类别分类。在收集好城市内空气质量数据后，将其生成测试集与训练集，读取数据集后计算欧式距离来需按照 K 个相邻的数据计算比例最大的分类，在分类中进行统计，将分类最大的作为预测结果，然后跟实际的进行对比来计算预测的准确率。训练集中城市空气质量数据越多，调节 K 值便能够提高预测准确率。

3.1.2 集成算法

集成算法主要是将子数据集进行结合以及通过多个学习机的结合、权重更新来获得输出值。首先在集成算法中需要对数据进行整合，引入弹性因子来评判样本数据。在随机选取样本数据 D1 后，选择弱学习机算法并设定最大迭代次数，在输出一项值之后将该值进行权重更新作为 D2 输入另一弱学习机算法当中输出，重复以上步骤 n 次后将其输出结果与第一次的输出结果整合进行输出，得到回归模型。在计算过程中需对每个训练样本的绝对误差与相对误差进行计算，并可根据权重来对样本进行选择，由样本采集的时间和训练误差来更新权重对更多的空气质量影响因素进行考虑，降低了预测误差。

3.2 数据采集与预处理

空气质量数据一般可从 BeijingAir 数据发布中心获得，在进行数据采集时需考虑到空气内影响空气质量的物质种类以及影响因素，一般为颗粒物质、可吸入颗粒物、二氧化硫、一氧化碳等等，并且需根据间隔时间来进行数据采集，然后根据不同站点将元数据分类呈各个子数据集。在整理数据集之后可从原数据当中初步对城市的空气质量状况进行判断[3]。在数据的预处理当中，需选取对数据分为训练集与测试集，并从训练集中按照 20：1 的比例分出部分数据作为验证集[4]，另外还需对空气质量数据进行等级分类，用“1”与“0”来评价空气质量，为之后的分析处理做准备。

3.3 数据归一化处理与数据相关性分析

在数据集内，由于所收集的数据种类复杂且根据不同的特征具有不同的数量级，因此需对数据进行归一化处理。所谓归一化处理就是将数据值控制在某一特定区域内，并保持数据的相对特性。数据归一化处理可分为Min-Max 缩放法和 z-score 标准化方法，但是根据所收集的城市内空气质量数据而言，由于具有范围广且离群值多的特点，z-score 标准化方法更能够将数据进行均值转换 [5]。

数据的相关性分析主要针对不符合正态分布特点的数据，对两个标量之间的关系进行评估，一般用相关系数来反映。在进行相关性分析时，一般将相关系数计算值 +1 或 -1 的状况被认为时关联的，而相关系数为 0 时则表示变量之间互相独立。在数据处理中，将各污染物的数据进行相关性分析，比较污染物与空气质量之间的相关程度，能够分析各种污染物对空气质量的影响及影响程度，为建立空气质量模型提供重要分析依据。

四、建立数值模拟模型

4.1CatBoost 模型

在对空气质量进行预测时，需要在对数据进行处理分析后依照算法来对数据建立模型，CatBoost 模型的建立首先需要对模型参数进行选择。针对不同的机器学习问题，对关键参数进行针对性的组合优化，其中参数包括了模型的学习率、迭代次数、选择梯度的增强类型、L2 正则化项的系数、损失函数以及类别特征标识。其次便是结合数据集将参数进行优化。在 KNN 算法当中，将样本数据进行了分类以及计算相邻样本数据的间距。由于在 KNN 算法当中训练集内已经获得了较多的样本量，可采用 TPE 算法对采样过程进行不断的生化，获得最优参数组合，便于对数据进行复盘与参数二次调整。在获得最优参数组合之后，对参数进行交叉验证，观察训练集以及验证集中的计算效果是否相似，一次来判断模型预测性能。在时序交叉验证法当中，需要对样本重新进行等量分布，并按照训练集与验证集的方式来进行反复验证，对验证的评价指标去平均值来获得验证结果，按照验证结果来确定建立一阶 CatBoost 模型时的参数。形成二阶 CatBoost 的重点在于采用 GDBT 算法将一阶 CatBoost 模型进行复用，选取其中的参数来作为新的特征来加深模型，以获得更为精确的预测结果 [3]。二阶 CatBoost 模型在一阶模型的基础之上进行建立，其所得参数更加精简，并且对所收集的数据进行了再次深挖，使得模型更加具体。

4.2 人工神经网络模型

人工神经网络模型是根据生物学中神经网络的原理对数据进行处理而建立的，该模型能够模拟神经系统对信息的处理方式将数据进行分析处理，并且能够进行深度学习。人工神经网络模型所处理的数据常为非线性数据，其神经网络一般分为输入层、隐含层以及输出层，在输入层内主要是将空气质量样本数据输入，然后在隐含层对数据进行归一化处理后，产生训练集与测试集，建立拓扑结构，并得到人工神经网络的权值和阈值。随后利用 MEA 算法对数据进行计算，得到可用参数，产生初始种群。如果在其中的计算误差过大，则会使神经网络内的权值重新返还进行再处理，指导符合参数标准。在对初始种群的分析中，针对其中的重要影响参数的数据将其作为优胜者，以优胜者为中心在神经网络内产生个体，将其作为优胜子群体以及临时子群体，然后对子群体内的每一个体进行计算，进行趋同操作与异化操作，得到最优个体。在输出最优个体后，训练人工神经网络，进行仿真预测 [2]。人工神经网络虽然对空气质量预测能力较高，但是对影响空气质量的污染物浓度的模拟能力较差。这一统计模型的建立能够对空气质量的变化进行准确预测，相应环境保护政策 [1]。

4.3RF 模型

RF 模型的建立运用了集成算法，将多个弱学习机算法进行结合，形成强学习机算法，并且通过整个各种不同类型的参数来避免模型参数单一的问题。在 RF 模型当中，进行仿真测试需要对数据进行 MSE 均方差计算，以计算值为衡量标准在数据集中进行搜索来确定 RF 模型参数。对于单一变量的参数，一般在训练集内随机搜索，但是对于连续变量的参数，则需要对数据集进行分布式采样以及交叉验证来获得最优参数解 [4]。RF 模型类似于决策树，对树的最大深度设置为 10 之后对每一个分支以及分支中的子节点数据进行设置，最低数值为 1，然后在训练集当中完成对参数的输出以建立 RF 模型。

五、运用差异融合法对空气质量预测分析

在对空气质量数据进行处理分析之后，尝试建立了 CatBoost 模型、人工神经网络模型与 RF 模型，通过训练后，发现这些模型均能够对空气质量进行高精度的预测，但是当空气质量数据的波动情况大、差异性较大时，模型的预测效果便层次不齐。根据此种情况，在机器学习建立模型的基础上采用差异融合的方法对模型的空气质量预测进行再训练。在 CatBoost 模型、人工神经网络模型和 RF 模型当中，对其分别计算出的预测值建立集合，并且对空气质量数据的波动差值进行运算以确定阈值。在对数值设置完毕后，对预测值集合中的每一个数值根据阈值来进行计算，并取绝对值。在差异融合法中，最重要的便是对融合阈值的选取。融合阈值的选取在设置循环变量的基础上，对不同模型的预测值进行计算，并将对应的阈值存入列表当中进行循环计算，直到获得最终值，取其中 MSE 的最小值和其对应的阈值，该阈值则为融合阈值。选取城市内的空气质量数据进行运算，获得不同模型的融合阈值，分别进行训练的测试来进行比较，并进行时间切片，来评价预测效果。在通过差异融合的运算之后，其中 CatBoost 模型能够对波动较大的空气质量数据进行预测，形成差异融合模型，预测较为准确，且误差低，拟合程度好。

六、结语

综上所述，在机器学习的基础之上将差异融合分析法应用在空气质量预测中，在经历过采用不同算法对数据进行计算，并分别建立 CatBoost 模型、人工神经网络模型与 RF 模型，经过训练检测后发现，针对大范围且波动大的空气质量数据，CatBoost 模型能够在众多样本数据当中在应用差异融合分析法后取得较好的预测结果，并且模型的预测能力十分稳定，具有较好的研究价值与实用价值。

参考文献：

[1] 卢亚灵 , 李勃 , 范朝阳 , 王建童 , 张鸿宇 , 蒋洪强 . 空气质量预测模拟技术演变与发展研究 [J]. 中国环境管理 ,2021,13(04):84-92

[2] 高帅 . 基于机器学习的空气质量评价与预测 [D]. 山西 : 中北大学 ,2019

[3] 钟锦辰 . 基于机器学习几种综合方法的空气质量预测研究 [D]. 四川 : 西南交通大学 ,2020

[4] 高嵩 , 何卓骏 , 刘子岳 , 刘家明 , 王刚 , 李登柯 . 基于机器学习的差异融合分析在空气质量预测中的应用 [J]. 电子测量技术 ,2021,44(18):85-92

[5] 夏起铁 . 基于机器学习技术的城市空气质量预测研究 .[J]. 信息记录材料，2020，21（12）：89-90.

作者简介：叶春阳（1991.2.5）男 汉 河北 本科 职称：数据开发工程师

现主要从事的工作或研究的方向：数据治理、数据仓库、数据挖掘