聚知刊出版社
JZK publishing

# 基于数据挖掘技术的大气污染物预测研究综述

琚汪慧 周欣宇

安徽新华学院大数据与人工智能学院，安徽合肥，230088；

**摘要：**随着近几年环境问题的日益严峻，人们对大气污染物数据的需求越来越高。海量数据的产生让传统的信息分析与预测方法以无法满足目前处理信息的需求。随着近十年数据挖掘技术的发展，利用神经网络、支持向量机等进行大气污染预测的研究活动急剧增加。为此，研究者应用开发了许多的预测模型，用于分析未来的大气污染物的变化。本文分析了用于大气污染物预测建模的主流技术极其优缺点，探讨了大气污染物预测建模的研究方向，最后对该应用的前景进行展望。

**关键词：**数据挖掘；大气污染物；神经网络；支持向量机

**DOI：**10.69979/3041-0673.24.9.023

## 引言

大气污染物是指人类活动或自然过程排放到大气中的，对人体和环境造成损害的物质。常见的大气污染物有 NO2，CO，SO2，O3，PM2.5，PM10，这六大污染物与我们的生活息息相关，破坏我们的身体和环境。本文就是基于数据挖掘技术对这 6 种污染物的浓度进行预测的研究综述。

数据挖掘技术预测大气污染物浓度，最早研究者应用了回归模型，其模型简单方便并且只要模型采用的数据一样，多次计算都会得到同样的结果。近年来人工神经网络（ANN）的发展，一些研究者将 ANN 应用到大气污染物的预测中，取得了显著的效果[1-4]。神经网络所具有的自学习、自组织能力和高容错性的优点契合大气污染物的预测[5]。支持向量机（SVM）可以较好的解决小样本、非线性、高维数等优点，已经成为广泛使用的算法应用到各个邻域中[6]。本文探讨各种方法应用到大气污染物浓度预测的优缺点以及改进思路。

## 1 大气污染物预测

近年来，经济的快速发展伴随着大量工业能源的消耗，导致废气排放剧增，严重破坏了人们的生活环境，对身体健康构成威胁。历史上，如 1930 年马斯河谷烟雾事件和 1943 年洛杉矶光化学烟雾事件，都因大气污染造成了巨大的人员伤亡和健康损害。在中国，尽管经济持续增长，但环境污染和生态破坏问题同样严峻，部分增长是以牺牲环境为代价的。自 2000 年起，中国大气环境自动监控技术取得了显著进展。至 2015 年，地级以上城市已建立了 1436 套监视系统，积累了大量宝贵的历史数据。这些数据原本主要用于生成观测报告，但随着大气污染研究的深入，其重点已转向预测，旨在更全面地了解污染物浓度变化及气象条件的影响。

当前的大气污染预测研究聚焦于六大主要污染物，通过综合考虑气象参数、排放数据和交通数据等多种因素来构建预测模型。由于空气质量数据通常具有年度周期性，许多研究采用涵盖一年的数据集。气象参数如风速、风向、相对湿度和大气湍流对污染物的扩散和浓度具有重要影响；排放数据涵盖城市环境中的一次和二次空气污染物，被视为重要预测因子；交通数据在路边空气污染物的形成中扮演关键角色。这些数据为数据挖掘技术在大气污染物浓度预测中的应用提供了坚实基础，有助于更准确地把握污染物的变化规律及影响因素。

## 2 大气污染物浓度预测常用技术

### 2.1 回归分析

回归分析是国内外最早使用的预测模型，其模型简单易于理解，虽然其在长期的预测中不如其它的预测模型，但是在短期预测中效果显著。在使用预测技术的时候，不仅要考虑预测的精度，还要考虑预测实现的难易程度以及成本。短期的预测线性回归综合而言是很好的方法。

在国内，杨小怡和黄世芹等[7]将线性回归技术较早的应用到大气污染物浓度的预测中，建立了自回归模型，将影响空气质量的 3 种污染物分别建立了五阶自回归模型，利用环境监测站的数据得到了污染物浓度的自回归预报方程。朱红芳和王东勇[8]用 1999-2000 年合肥市 4 个探测点的污染物数据，使用多元回归法，从 15 个预报因子中，筛选了 3-4 个因子，建立了探测点污染物的预报方程。蒋明皓和张元茂[9]应用门限自回归模型，对上海市大气污染物做出预测研究，着重研究其与风速之间的关系。得到了风速对污染物浓度有显著的影响，呈负相关。

回归分析在预测大气污染物浓度时面临以下挑战：

（1）众多因素如气象、交通、工厂规模及数量均影响大气污染物浓度，因此难以准确识别关键输入因子。（2）选用不当的输入因子会大幅降低回归模型的精确度，使预测结果失去意义。（3）非线性多项式回归难以构建，无法有效处理高度复杂的数据。

对于上述回归分析的缺点，可以利用逐步回归分析法筛选输入因子。将对回归方程的因变量有明显影响的自变量选入到方程中，没用显著作用的从回归方程中删除，再通过显著性检验计算影响程度，根据其重要性选入回归方程中。在整个过程中，不断的有新的自变量引入，会导致之前入选的影响程度变低，而之前被剔除的变量的影响程度变高，那么前者剔除，后者重新入选，直到再无变化即可。这种方法便于直到显著性较高的输入因子，自变量个数少且方程稳定。赵国君[10]采用权重系数法，对每一个气象因子对某种污染物的权重系数进行计算，得到了 12 个月每个污染物与气象因子相互对应的权重系数，筛选出起主导作用的气象因子。除此之外增加了其他污染物浓度作为自变量，其改进后的多元线性回归的效果优于普通的多元线性回归。而为了解决传统多元回归分析难以解决的非线性多元回归问题。吴今培等[11]将前馈神经网络运用到非线性多元回归分析中，根据结果，综合模型预测的效果比单一模型的效果更佳，这证明了神经网络可以有效解决非线性多元回归问题，势必在未来成为预测大气污染物浓度的主流方法。

## 2.2 神经网络

神经网络因其分步存储、高容错性、大规模并行处理、泛化能力及自适应性等优点，自上世纪 90 年代以来在全球范围内得到广泛应用。在大气污染物浓度预测领域，神经网络已成为最佳的长期预测方法。传统的预测方法，如线性回归，因缺乏学习判断能力，无法从训练集中提取知识并总结经验，而人工神经网络（ANN）则能显著提升预测精度。

在国外，大城市时常面临高浓度污染物的挑战，准确预测污染物浓度趋势仍是一大难题。P.Viotti[12]等人利用 ANN 预测常见大气污染物的短期至中长期浓度，采用 BP 算法和 sigmoid 激励函数，预测结果与监测数据高度吻合，可作为仅需气象条件和交通水平的 24-48 小时预测模型。Heidar Maleki[13]等人则针对伊朗阿瓦兹地区，利用 ANN 预测空气质量指数（AQI），发现模型在阿瓦兹等城市的空气质量预测中具有一定适用性，可有效预防健康影响。此外，ANN 还可作为空气污染空间插值和空气监测网络的有效替代方法，提供未监测位置的数据，描述空气污染空间变异性。

在国内，张定田等[14]利用神经网络预测北京市环境质量，建立数学模型并排序主要影响因素和发展趋势。陆文志则开发了一种改进的神经网络模型，结合主成分分析技术和径向基函数网络，预测污染物趋势，模型具有更简单的网络架构、更快的训练速度和更满意的预测性能。宋耀宇则将卷积神经网络和 LSTM 结合，应用于大气污染预测，实验结果表明这种混合模型在时间序列预测方面效果极佳。

然而，BP 神经网络在建模过程中也面临挑战。新变量的加入可能导致模型误差增大，预测精度降低。黄世芹[15]等人针对此问题泛化改进了 BP 神经网络，调整性能函数，采用多层 BP 神经网络模型，将影响大气污染物浓度的因子作为输入层，通过神经元处理网络信息，提升预测效果。唐晓城则针对 BP 神经网络学习速度和收敛速度慢的问题，构建了新的 BP 神经网络模型 Tang XC_BPModel，通过批量修改权值提高算法效率，仿真实验表明模型效率和精确度极高。郭庆春则介绍了附加动量法和弹性 BP 算法等改进算法，以解决神经网络的局部极小问题和网络泛化能力差等问题。在大气污染预测研究中，构建高效预测系统的需求迫切。研究者常采用至少一年的数据集以反映季节及节假日影响，但特殊情境变化对预测结果的潜在干扰难以预测，处理不平衡数据的预处理技术尤为关键。数据缺失处理方法的选择需谨慎，以避免产生负面效应。此外，神经网络模型开发需关注预测器选择，全局方法在预测因子选择中的应用应受重视。特殊方法如试错法可能导致结果难以复现，需制定系统化协议确定模型结构。参数不确定性校正也是值得深入研究的领域。

综上所述，神经网络在大气污染物浓度预测中展现出强大潜力。未来研究应继续探索神经网络的优化和改进方法，提高预测精度和效率，为环境保护和健康管理提供可靠支持。同时，需关注预测器选择、数据预处理及模型结构确定等关键问题，制定系统化协议，推动大气污染预测研究的深入发展。

## 2.3 支持向量机

支持向量的概念最早在 1963 年由 Vladimir N.Vapnik 提出的[16]，直到上个世纪 90 年代支持向量机才大规模的应用到计算机科学中。对于大气污染浓度的预测，近些年 SVM 也广泛应用到此领域中，并有着自己独特的优点被环境预测工作者所使用。

支持向量机是一种新的小样本学习方法，但是对于大规模的数据样本。SVM 似乎有些无计可施，近些年的大气污染物浓度使用 SVM 的大多都是一年左右的样本集。但是有些研究需要 10-20 年的大气污染物的数据，显然 SVM 算法是不合适的。这是由于 SVM 求解支持向量需要利用 m（样本数）阶矩阵，大量的样本数对计算和存储，都将耗费大量的内存，预测效率会降低。从本质上讲，

它避免了传统的推理过程，为训练样本到预测样本提供了有效的“传导推理”。支持向量的数量决定计算的复杂程度，而不是样本空间的大小，从而在某种意义上避免了维数灾难。不仅有助于我们捕获关键样本，消除“大量冗余样本”，而且保证了算法的简单性和稳健性。

W.C. Leong,R.O[22]等人提出一个支持向量机来模拟空气污染指数。影响支持向量机模型性能的主要参数有三个：惩罚因子、正则化参数和核函数的类型。在该研究者的论文中只研究了核函数模型参数。利用误差平方和（SSE）、误差平方和平均值（MSSE）和确定系数（R2）对模型结果进行了分析。开发具有适当精度的环境质量模型预测并不是那么容易，特别是在空气质量建模方面。在涉及化学、物理或气象数据的情况下，获得充分反映环境数据动态行为的模型也不容易。因此该研究提出支持向量机模型预测。在此研究中，数据收集自 2009 年至 2014 年，包括缺失数据，其中没有记录任何值。论文作者首先进行筛选，以检查原始样本数据与缺失数据，并识别异常值。在处理缺失数据时，将从原始样本中删除与缺失数据相关的行数据。由于样本已经在离散时间函数中，因此删除此缺失数据不会影响模型预测的总体性能。虽然对缺失数据进行了筛选，但样本中仍存在一些异常值。因此，在模型开发之前，使用简单的回归模型来研究异常值对模型精度的影响。完成数据处理后将支持向量机建模的数据分为：训练数据（用于支持向量机模型训练）、验证数据（用于支持向量机模型的交叉验证）和测试数据/未显示的验证数据（用于最终选定模型的评估），其分布分别为原始数据的 70%、15%和 15%。研究了核函数在支持向量机建模中的作用。研究了三类核函数：线性、多项式和径向基函数（RBF），找出最终支持向量机模型的最佳核函数。文中在评价所建立的模型时，均方误差（SSE）、均方误差（MSSE）和决定系数（R2）作为性能指标。支持向量机模型是在原始数据的基础上，剔除异常点后建立的。该研究成功地建立了一个适用且准确的支持向量机模型，该模型只使用预测因子的值，而不需要最初使用的复杂计算。结果表明，剔除异常值确实提高了模型的性能。国内研究者蔡仁[23]将支持向量机方法和 Elman 神经网络运用到乌鲁木齐的污染物预报之中。将两种模型的污染物浓度误差实验结果进行比对分析，得到了支持向量机比 Elman 网络模型预测值更加的精确的结论。卢玮[24]将粒子群优化算法与支持向量机向结合，为了使支持向量机的核心参数惩罚因子和和函数参数可以寻优，使得达到最好的精确度。同时寻优的过程因为各方面的原因，例如样本集的大小，会对模型的工作效率产生不良的影响。而李光明[25]等人在露天矿区选取 PM10 质量浓度，综合环境数据和气象数据，引入改进型惯性权重的粒子群优化（PSO）算法到支持向量机创建模型。结果分析表明，所提模型的预测精度优于普通模型，且预测精度可达到 98.5%以上。

支持向量机模型是有着严谨的理论的数学做支撑，对于一些神经网络模型而言过度依赖于设计者的经验，这就会造成实验精度上的偏差。[26-27]支持向量机模型的建立不需要大量的训练集，少量的数据支持向量机也可以呈现良好的预测性能。同时在样本的敏感性和特异性上的平衡，支持向量机也可以平衡的较好，得到高精确度的预测值和极佳的泛化能力。同时一些改进的支持向量机模型针对传统支持向量机的各种问题进行优化，也让支持向量机更好的运用到大区污染物浓度预测中。

## 3 前景展望

自上世纪末以来，数据挖掘算法在大气污染物浓度预测中得到了广泛应用，国内研究者在本世纪初也开始利用这些算法进行预测，并据此提前治理环境污染，取得了防微杜渐的效果。然而，目前许多国内中小城市仍缺乏适合自己的预测模型，这对环境治理构成了严峻挑战。

在回顾近年来国内外相关期刊论文的基础上，我们发现未来可以从以下方面改进预测模型精度：

首先，输入因子的选择至关重要。当前研究多依赖常规气象数据，但某些非常规因子，如逆温层厚度，对特定污染物浓度（如 PM10）有重大影响。因此，未来研究者需深入探索不同污染物的特殊气象影响因子，以提高预测精度。同时，生活、工业和交通污染源数据也应纳入考虑，作为新的输入因子。

其次，数据处理优化同样关键。除了处理缺失值和离群点，还需对数据量级进行变换，以降低其对预测结果的影响。此外，训练数据长度、数据标准化方案与传递函数的关系，以及训练参数的自适应初始化方案等，均对模型性能有重要影响。在选择数据时，还需关注数据背后的缘由，如政策变化等，以确保预测模型的精确度。

最后，传统算法的优化也不容忽视。未来建模者需根据计算惩罚和运行时间评估各种算法，并进一步研究模型结果的不确定性分析。量化不确定性是提升模型实用性的关键，可通过选择合适的算法或改进算法来实现。此外，还可尝试将传统算法与其他算法结合，如支持向量机与小波神经网络、灰色理论等，以提高预测精度。

综上所述，未来研究者需综合考虑输入因子选择、数据处理优化和传统算法优化等方面，以提升大气污染物浓度预测模型的精度。同时，还需关注模型性能的其他方面，如复制有效性和结构有效性，以确保优秀的预测模型能够适配到需要的城市中。

## 参考文献

[1]田启凡. 基于空间分析技术的桂北地区大气污染物传输分析与预测研究[D]. 桂林：桂林电子科技大学.
[2]陈柳. 小波网络在大气污染物浓度预测中的应用[J]. 环境科学与技术, 2007, 30(1):3.
[3]陈柳. 小波分析和神经网络应用于大气污染预测的研究[D]. 西安：西安建筑科技大学, 2006.
[4]吴亚平, 张琦, 王炳赟,等. 四川雅安三种主要大气污染物浓度与气象条件的关系及其预测研究[J]. 高原气象, 2020, 39(4):10.
[5]张卓然. BP 神经网络和自适应模糊推理系统在多传感器粮情信息融合系统中的研究及应用[D]. 武汉：武汉工业学院.
[6]张学工. 关于统计学习理论与支持向量机[J]. 自动化学报, 2000, 26(1): 32-42.
[7]杨小怡,黄世芹,姚雷.用线性回归方法建立贵阳市空气质量预报模式[J].贵州气象,2001(04):10-12.
[8]朱红芳,王东勇.合肥市空气质量预报方法[J].气象,2002(05):40-4.
[9]蒋明皓, 张元茂. 采用门限自回归模型预测环境空气质量[J]. 上海环境科学, 2001, 020(008):375-377.
[10]赵国君. 长春市空气质量预报系统的建立及应用[D].长春：吉林大学,2004.
[11]吴今培,黄磊.非线性多元回归分析的神经网络方法(英文)[J].长沙电力学院学报(自然科学版),2002(02):19-22.
[12]P Viotti,G Liuti,P Di Genova. Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia [J]. Ecological Modelling,2002,148(1):10-12.
[13]Maleki, H., Sorooshian, A., Goudarzi, G. et al. Air pollution prediction by using an artificial neural network model. Clean Techn Environ Policy,2019,21(1):1341 - 1352.
[14]Zhang Dingtian,Zhang Xiaoxi. A Neural Network Forecasting Model of Beijing Environment Quality Based on Set Pare Analysis[J]. Energy Procedia,2011,5(1):12-18.
[15]黄世芹.改进 BP 神经网络在城市环境大气污染分季节预报中的应用[J].贵州气象,2005(03):6-8.
[16]Vapnik V N, Chervonenkis A Y. On a perceptron class[J]. Avtomat. i Telemekh, 1964, 25(1): 112-120.
[17]Leong W C, Kelani R O, Ahmad Z. Prediction of air pollution index (API) using support vector machine (SVM)[J]. Journal of Environmental Chemical Engineering, 2020, 8(3): 103208.
[18]蔡仁,李如琦,唐冶,路光辉.支持向量机在乌鲁木齐污染物预报中的应用研究[J].沙漠与绿洲气象,2014,8(03):61-67.
[19]卢玮. 基于 PSO-SVM 模型的地下建筑空气质量预测研究[D].武汉：华中科技大学,2017.
[20]李光明,王军,李颀.改进的PSOGA-SVM模型应用于露天矿区空气质量预测[J].中国科技论文,2019,14(12):1348-1355.
[21] 谷瑞, 顾家乐, 宋翠玲. 基于金字塔分割注意力和联合损失的表情识别模型[J]. Journal of Data Acquisition & Processing/Shu Ju Cai Ji Yu Chu Li, 2024, 39(6):67-72.
[22]罗亚波,罗健.基于阻滞增长神经网络的双目视觉标定方法[J].华中科技大学学报(自然科学版),2021,49(12):71-75+82.

本文得到安徽省教育厅高校自然科学重点研究项目：软件工程专业改造提升工项目项目(No. 2023zygzts105)、混合型机器学习方法在精准医疗脑疾病中的预测性研究(No. 2024AH050621)、安徽省大学生创新训练计划项目：深度学习网络模型结合人工合成肿瘤提高肝脏肿瘤分割精度(No. 2024AH050621)、安徽省大学生创新训练计划项目：大语言模型在神经外科门诊检查中的对比研究(No.S202412216215)资助.

作者简介：琚汪慧（1988-）女，汉，安徽池州，教师/助教，硕士研究生，研究方向：软件设计，大数据挖掘，安徽新华学院，安徽省合肥市，231000
周欣宇，性别，（2006-），研究方向：数据挖掘。