单位代码：10293 密　级：

南京邮电大学

专 业 学 位 硕 士 论 文

论文题目：基于机器学习的空气质量分类判别研究

学　　号	1318024723
姓　　名	林启明
导　　师	叶全意 副教授 / 徐恩明 副教授
专业学位类别	工程硕士
类　　型	非全日制
专业（领域）	电子与通信工程
论文提交日期	2021 年 11 月 12 日

Research on Air Quality Classification and Discrimination Based on Machine Learning

Thesis Submitted to Nanjing University of Posts and Telecommunications for the Degree of

Master of Master of Engineering

By

Lin QiMing

Supervisor: Associate Prof. Ye QuanYi/ Associate Prof. Xu Enming

Nov 12 2021

摘要

当前全球气候剧变，由此引发的空气污染问题日益严重，人们的日常生活受到了严重的影响。随着环保意识的逐渐增强，人们对于空气质量改善的需求越来越高涨，如何做好污染的防治工作成了一个十分急迫的问题。在保持经济高速发展的同时将工业化对环境气候的影响降低到最小，这已经成了各国学术界所共同追求的目标。采取科学的方法进行空气质量监控工作已经成了一项重要议题，准确地从空气质量数据中获取信息是这项议题的关键。掌握污染变化的规律，十分有利于科学地指导空气污染防治工作，对城市健康发展具有十分重要的指导意义。

随着科技的快速发展，新兴的机器学习技术已经融合到社会发展的各个领域，已经在统计样本分类、时间序列预测等方面获得了突破性的进展。本文主要基于机器学习技术，引入了若干典型机器学习算法对空气质量数据分类的判别和预测进行了研究。文章收集了中国空气质量数据监测网上公开的南京市近 6 年来的数据样本，分别进行了空气质量等级分类以及 AQI(Air Quality Index)污染指数的预测判别两项工作，主要工作如下：

(1) 针对 SoftMax 逻辑回归在多分类情境下对空气质量数据集的分类准确率不高的问题，提出了低置信样本的概念并结合了 SVM(Support Vector Machine)模型对低置信样本进行结合分类的方法。在实验数据上的运行结果表明，新模型的准确率要优于 SoftMax 模型的分类准确率。

(2) 针对 SVM 的模型参数使用网格搜索法进行参数搜索的运行时间过长的问题，引入了两种群智能算法 GWO(Grey Wolf Optimization)和 PSO(Particle Swarm Optimization)，将 SVM 模型的参数作为群智能算法的输入、分类错误率作为适应度值，经过迭代搜索适应度最小时的参数。在实验数据上的运行结果表明，两种群智能算法的精确率都得到了提升，同时 PSO-SVM 的精确率要优于 GWO-SVM 方法，但是前者的算法执行时间更长。

(3) 针对另外一种对 AQI 指数进行预测的问题，提出了一种基于时序 EMD(Empirical Mode Decomposition)-SVR 预测算法，对月平均污染数据集的 AQI 指数进行预测，结果表示在月平均数据集上 SVR 模型的预测效果比传统的 ARIMA 模型要好，而相比于基于特征的 SVR 算法上的数据集的优势是依赖的数据样本更少。

关键词： 空气质量指数，支持向量机 ，灰狼优化算法 ，粒子群优化算法，经验模态分解 ，时间序列预测

Abstract

It is becoming more and more serious that the global climate has changed dramatically. With the increasing awareness of environmental protection, people's demand for air quality improvement is increasing. How to do a good job in pollution prevention and control has become a very urgent problem. While maintaining rapid economic development, minimizing the impact of industrialization on the environment and climate has become the common goal of academic circles all over the world. It has become an important issue to adopt scientific methods for air quality monitoring. The key to this issue is to accurately obtain information from air quality data. Mastering the law of pollution change and understanding the impact of air pollution on the environment is very helpful to scientifically guide the prevention and control of air pollution and has very important guiding significance for the healthy development of cities.

(1) Aiming at the problem that the classification accuracy of logistic regression on air quality data set is not high in multi classification situation, the concept of low confidence samples is proposed, and the method of classifying low confidence samples is combined with SVM model. The running results on the experimental data show that the accuracy of the hybrid model is better than that of the softmax model.

(2) Aiming at the problem that the running time of SVM model parameters using grid search method is too long in general, two population intelligent algorithms GWO and PSO are introduced. The parameters of SVM model are taken as the input of swarm intelligence algorithm, and the classification error rate of SVM is taken as the fitness value. The parameters with the lowest fitness are searched iteratively. The running results on the experimental data show that the accuracy of the SVM model combined with the two population intelligent algorithm has been improved, and the accuracy of PSO-SVM is better than gwo-svm, but the execution time of the former algorithm is longer.

(3) Aiming at another problem of AQI index prediction, a time series emd-svr prediction algorithm is proposed to predict the AQI index of monthly average pollution data set. The results show that the prediction effect of SVR model on monthly average data set is better than that of traditional ARIMA model, and the advantage of SVR model over feature-based SVR algorithm is that it depends on fewer data samples.

Key words: Air Quality Index, Support Vector Machine, Grey Wolf Optimization, Particle Swarm Optimization, Empirical Mode Decomposition, Time Series Forecasting

目录

专用术语注释表

缩略词说明：

AQI	Air Quality Index	空气质量指数
EMD	Empirical Mode Decomposition	经验模态分解
SVM	Support Vector Machine	支持向量机
SVR	Support Vector Regression	支持向量回归
GWO	Grey Wolf Optimization	灰狼优化方法
PSO	Particle Swarm Optimization	粒子群优化方法

第一章 绪论

1.1 研究背景及意义

随着绿色可持续发展的理念的认识的不断加深，近年来，人们对环境保护的呼声越来越高。环境污染的治理一直是国家生态文明建设中的一项重要工作环节，而大气污染的治理更是重中之重。被污染的空气中存在着有害人类健康的污染物，如 PM2.5，一氧化碳，二氧化硫，二氧化氮，臭氧等对人们的呼吸系统存在直接危害的有害物质，从而严重影响人们的身体健康。在众多的环境污染保护的相关议题中，空气污染是其中的重中之重[1]。然而在耶鲁大学 2016 年发表的一份关于环境表现的报告中，我国的总体气候质量在全球排名倒数第二[2]。这些调查结果显示我国的空气污染已经严重到了不可忽视的地步[3-4]。因此，针对大气污染的研究已经迫在眉睫。

南京位于长三角经济区的腹地，一直以来都是我国经济发展的最前线，发达的经济环境伴随而来的环境污染问题也同样不容乐观。相关数据表明，1954 年至 2012 年南京市的雾霾天气数呈现出较为明显的上升趋势[5-8]。随着城市化与工业化的加速发展，如何在发展的同时能够兼顾空气污染的治理，同时做到经济社会的绿色健康发展与环境生态的良好保护，已经成为了一项具有现实意义上的重要课题。为了达成上述目标，必须要坚持科学的方法。为此，通过科学方法认识空气污染的形成，了解其危害，采用技术手段预测空气质量的变化，及时对污染进行预测十分有必要。

对空气污染使用数值进行量化的指标是空气质量指数 AQI(Air Quality Index)，它以 1-6 级别进行分级的形式来表示当前空气污染的程度，污染程度随着级别数值的升高而升高。我国进行 AQI 评估的指标目前主要是以下六项：PM2.5，PM10，二氧化硫，二氧化氮，一氧化碳和臭氧[9]。AQI 指数是由上述六项的分指数各自独立计算，最后求分指数中的最大值所求得到。不同的分指数存在着较大差异，其 1 小时平均值、8 小时平均值以及 24 小时平均值的区别，并在计算时使用的平均值不尽相同，并且二氧化硫和臭氧两项分指数还存在不同平均值的大小决定均值选择的问题。可见 AQI 的传统求解存在着计算繁琐，计算量大等问题。人们在日常生活生产中，往往只关心当天的污染级别，对具体的空气质量指数数值并不十分关注。利用科学的技术手段来避免对分指数进行繁琐的计算，快速准确得到当前的空气污染级别，能够极大方便指导人们的生产活动。

近几十年来，机器学习与人工智能技术在理论与应用上都得到了飞速的发展。随着计算

机硬件的性能巨幅提升，人工智能技术已经由上世纪八十年代的理论计算阶段逐渐走向了实践的应用层面，成为了计算机进行智能化人机交互的主要手段。随着学科间的融合趋势的发展，人工智能技术也在数据挖掘、自然语言处理、图像处理等领域大放异彩[10]。人工智能技术的代表性特点就是擅长海量数据的处理，挖掘出数据中的联系与规律。本质上来说，它是依据经典统计数学规律，为已有样本数据进行某种特定的变换，得到某种特定的数学模型来对新的未知数据进行分类和预测的一种方法。可见，在空气质量的分类和预测领域中引入机器学习方法，前景十分广阔。本文将围绕这一议题进行研究。

1.2 研究现状

目前气候变化引发的环境问题已经严重影响了人们身体健康。随着近几年机器学习技术的快速发展，很多学者都关注使用机器学习算法对空气质量数据集进行分析，从而对空气质量进行防控监测。已经有学者提出了基于机器学习和深度学习等理论的方法，对空气污染进行预测和分类。在理论层面，主要分为对空气质量进行分类预测以及数值预测的两大方向。目前国内外已经有不少学者提出基于机器学习方法的空气质量的分类与预测模型[11]，经过实验证明基于机器学习的空气质量指数等级分类研究的有效性。下面就分类和预测这两个方向分别阐述研究现状。

1.2.1 空气质量等级分类问题研究现状

目前已经有不少学者使用机器学习相关理论对空气质量等级进行分类。常见的机器学习方法有 BP 反向传播网络、支持向量机、迁移学习理论、决策树理论等，部分文献使用了 BP 反向传播神经网络对空气质量等级做了判别[12-14]。文献[15]提出了一种改进的 SOFM 神经网络，它对全国 113 个城市的公开 AQI 等级数据进行分类，对各城市的环境治理工作成效做出了评价。文献[16]分别使用了 Lasso 回归、随机森林回归和深度学习 RNN-LSTM 对川渝地区的 PM2.5 数值进行了预测判别，结果达到预期。陈祖云[17]等建立了支持向量机模型进行空气质量等级的判别。在对于分类进行判别的过程中，也有人提出了模糊评价的机制[18-20]。文献[21-23]使用了灰色理论(Grey Theory)完成了对空气质量的评价工作。为了解决空气污染影响因子较多的问题，文献[24]使用了基于 PCA 主成分分析和 k 均值聚类结合的方法降低了污染因子的维数，并通过实验在众多污染因子中找到了对分类结果起着决定性因素的几种污染因子，给提高分类模型准确性提供了相关的依据。在实际应用中，为了确定分类模型的各影响因子的权重对

分类结果造成的影响，文献[25]提出了基于不同熵的多级模糊综合评价方法，经过实验证实，该方法能够有效提高强污染因子对分类结果的影响，但是该方法也削弱了潜在的因子的影响，可能造成结果的偏差。

1.2.2 AQI 指数预测问题研究现状

随着全球性气候的剧变，人们越发关心自身生活环境的状况。提前对未来几天内的空气质量进行预测分析，这对指导人们生活和生产有着现实意义。很多学者对空气污染指数的预测进行了研究。目前对 AQI 指数进行预测大致有两大方向，其一是传统的基于经典时间序列的预测法，其二是基于机器学习的预测方法，它与传统方法的区别是对污染因子加以训练，然后通过训练所得的模型对新数据进行预测。

首先是传统的基于经典时间序列的 AQI 预测方法，该方法主要将 AQI 指数视作纯时序变化的序列，而与其污染物影响因子无关。其代表性理论是经典时间序列预测的 ARIMA 模型[26]。常见的基于时间序列的 AQI 预测方法有多元回归模型[27]，自回归积分滑动平均模型 ARIMA[28]，季节性回归积分滑动平均模型 SRIMA[29]等。这类模型的优点是算法实现较为简单，但缺点也是很明显的。首先，经典模型的思路是将预测结果指拟合为时间序列的多项式函数。使用多项式进行多元拟合回归的方法虽然直接有效，但是容易遇到过拟合问题从而导致预测结果的偏差。其次，ARIMA 理论要求时间序列有着严格的平稳性要求，而实际的 AQI 序列往往是非线性序列非平稳序列，虽然可以使用差分法等将其平稳化，但是这也会导致模型的实际预测精度误差较大，结果不是很理想。因此，很多学者提出了一些改进方法。文献[30]中提出了一种改进基于最小二乘法的 SVM 模型，实验结果表明相对于传统时序预测法，该方法的精度有明显提高。还有学者重点研究对非平稳序列进行平稳化的处理，这一方向的的工作也做了很多[31-36]。

其次是基于机器学习方法的 AQI 指数预测方法，这类方法的特点是可以根据 AQI 指数的污染物因子进行数据分析从而挖掘特征。与机器学习方法结合起来，能够有效找到 AQI 的变化趋势和各污染物因子之间的内在联系。相比于基于时间序列的预测方法，它的预测因子不局限于单一的时间序列，而是可以将污染物因子作为模型的预测因子，从而可以充分训练特征。相比较于基于时间序列的方法，这种基于机器学习的方法效果更好。机器学习进行序列回归预测的方法有很多，最具有代表性的有广义神经神经网络[37]、SVR[38]等。由于使用单一的神经网络容易陷入过拟合问题，学者们提出了很多改进的方法。文献[39]提出了一种基于 PCA 方法的神经网络 AQI 预测模型，该方法有效减少了特征的数量，使得减少网络的深度的

同时还提升了预测任务的效果。文献[40]使用了 NARX 网络来预测昆明的 AQI 指数，实验结果表明该模型的效果要好于 CMAQ 等统计模型。文献[41]在传统的 BP 神经网络变换的基础上引入了小波变化法，用小波变化法与传统的 BP 神经网络结合方式对原始 AQI 序列进行回归预测。从结果上来看，AQI 预测精度有一定提升。

深度学习技术已经随着计算机硬件的迅猛发展得到了广泛的应用。作为人工智能领域最核心的技术之一，深度学习被广泛应用于图像语音和模式识别等领域内。不少学者将深度学习技术应用到空气污染预防治理领域并取得了不错的效果[42-45]。文献[46]提出了一种基于深度学习的新方法深度信念网络 DBN 技术，该技术将数据从原特征空间转换到具有语义特征的新特征空间，能够通过学习获得层次化的特征表示，实验结果表明预测精度得到提升。针对污染因子对于污染结果程度影响的不确定性问题，有学者们提出了一种灰色理论[47-51]，该方法不预先假定某几种影响因子对预测结果的影响，而是通过因子对结果的影响来综合评定所选的污染物种类。灰色模型在 AQI 指数接近指数分布时的预测效果较好，但是缺点是多步预测的精度较低。

尽管在空气质量等级分类和 AQI 指数预测方面已经有了不少的研究工作，但是在基础理论以及应用研究方面仍然存在着很多不足的地方。在对空气质量等级的分类中，现有的研究方法往往都基于某种单一机器学习方法，但是忽略了不同方法之间的结合；现有的对方法模型参数的优化中，往往使用网格搜索法，却没有很好地对优化方法本身进行改进；在对 AQI 指数进行预测时，现在主流的方法都基于前面介绍的两种理论，然而还有不少新型的理论方法对序列进行预测。本文从将这三个方面入手，首先针对研究方法单一的问题，提出了机器学习方法相结合以提高准确率的方式；其次，针对参数选择优化问题，使用了群智能优化的方法来改进传统的网格搜索法；然后对于污染指数的预测问题，使用了经验模态分解算法对原 AQI 序列进行分解后，对分解的子序列进行单独预测，最后叠加子序列的预测结果。实验表明得出的结果相比于经典时序预测理论效果更好。

1.3 论文主要工作和结构安排

本文的数据集来自空气质量检测网(https://www.aqistudy.cn/historydata/)数据库，使用了库中江苏省南京市的 2014-2021 年间的空气质量数据集。根据空气质量数据集中给出的六项参数 PM2.5、PM10、一氧化氮 NO、二氧化硫 SO_2、一氧化碳 CO、臭氧 O_3，通过引入传统机器学习方法(如 SVM)和群智能算法(如灰狼优化方法)等算法模型以及框架，构建设计了空气质量分类算法，并对算法性能进行了分析与讨论。文章的组织结构如图 1.1 所示：

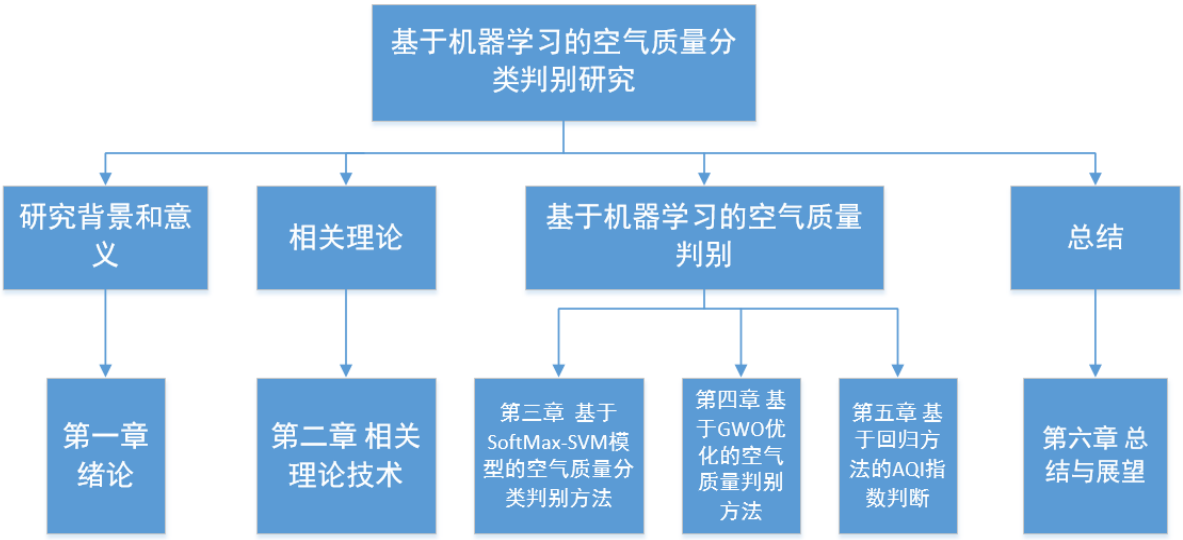

图 1.1 本文的章节相互关系以及结构

其中第一章给出了相关研究背景和研究意义，第二章给出并介绍了全文主要涉及并且所应用的理论，三到五章节列举出具体的分类算法实现细节，同时也作为全文的主旨部分，第六章对前述内容做出总结。

各章的详细结构安排如下：

第一章首先引入论文课题空气质量分类研究的相关背景和意义，然后阐述了现阶段的有关研究成果，简要对照了各相关算法的主要优缺点，接着给出文章的框架结构。

第二章介绍了文章中所使用的基础理论知识和算法模型，包括 SVM 支持向量机、Logistic 逻辑斯蒂回归模型等，主要介绍算法模型的基本原理及特点，为后文主体部分的研究提供了理论基础。

第三章研究了基于多分类 logistic 回归的支持向量机空气质量等级判别算法。首先介绍了 Logistic 回归的算法模型，给出了多分类情况下的理论分类模型。为了解决单一模型的分类精度不高的问题，引入了低置信样本的概念，根据低置信样本集合的特性，提出了支持向量机混合训练低置信样本点的思路。根据所选的样本阈值点筛选出恰当的低置信样本集合，构建混合训练模型。最后进行实验仿真，根据仿真结果表明，相较于已有单一模型的空气质量分类算法，混合模型具有更好的分类精度和性能。

第四章研究了基于 GWO-PSO 灰狼优化的空气质量等级判别算法。针对支持向量的模型参数问题，相比于传统的网格搜索求解法，本章提出了一种基于群智能优化方法的支持向量机模型参数求解方法，将支持向量机的模型参数作为群智能算法的输入，通过迭代在解空间中找到模型准确率最高时候的参数向量。首先引入单一 GWO 以及单一 PSO 优化参数的局限性，然后提出了一种结合两种方法的方式进行 SVM 的参数优化，通过实验表明，精度得到

了一定的提高。

第五章研究了基于时序 EMD-SVR 的空气质量指数 AQI 回归算法。首先，介绍了机器学习回归问题中的常见评价指标，然后给出了使用 EMD 方法进行 AQI 指数预测的流程，然后介绍了支持向量机在回归问题中的形式——支持向量回归机 SVR。接着给出了使用时序 EMD-SVR 进行空气质量指数 AQI 预测的流程。最后在数据集上进行仿真测试并且对比了两种方法的预测结果。

第六章对全文进行了总结，对前文中所使用的方法的不足之处给出了分析归纳，并在此基础上提出下一步研究的方向。

第二章 相关背景知识介绍

2.1 引言

在第一章中，主要介绍了空气质量分类与预测的研究背景和研究意义，并介绍了国内外的相关研究现状和研究方向，接着初步分析了已有的研究方法和技术的一些缺点并介绍了本文将要说明的改进方法。接下来，本章节将首先对本文中涉及的机器学习理论进行阐述和分析。首先介绍文章中的数据集来源以及空气质量相关的概念，然后介绍本文中使用的主要机器学习方法，具体涉及的有逻辑斯蒂回归和支持向量机。

2.2 数据集来源

空气质量指数 AQI 是一种表征空气污染程度的一个指数。影响 AQI 空气污染指数评价指标的主要污染物有细颗粒物 PM2.5、可吸入颗粒物 PM10、二氧化硫 SO_2、二氧化氮 NO_2、臭氧 O_3 以及一氧化碳 CO 六项。为了研究上述六项污染物对空气污染指数 AQI 的具体影响，本文从中国空气质量在线监测分析平台历史数据网(https://www.aqistudy.cn/historydata/)上公开的江苏南京的空气质量历史数据中截取了一部分数据。截取时间是 2014 年 1 月 1 日——2021 年 2 月 28 日，数据采样周期为 24 小时一次。部分数据的展示如下：

表 2.1 显示了部分数据样本：

表 2.1 部分数据样本

日期	AQI	等级	PM2.5	PM10	SO_2	CO	NO_2	O_{3_8h}
2017-02-02	70	良好	50	72	19	1	26	112
2017-02-03	53	良	33	49	11	1.1	44	60
2017-02-04	85	良	63	94	13	1.4	50	36
2017-02-05	125	轻度污染	97	125	16	1.6	50	107
2017-02-06	107	轻度污染	78	111	22	1.3	46	92
2017-02-07	46	优	23	45	13	0.8	30	73

单个数据样本的特征(feature)包含南京市每日的 AQI 指数、PM2.5、PM10、NO_2、SO_2、O_3 和 CO 浓度，同时给出了污染等级。

为了方便研究单一指数对 AQI 指数的影响，绘制每项特征的变化折线图，结果如图 2.1 所示：

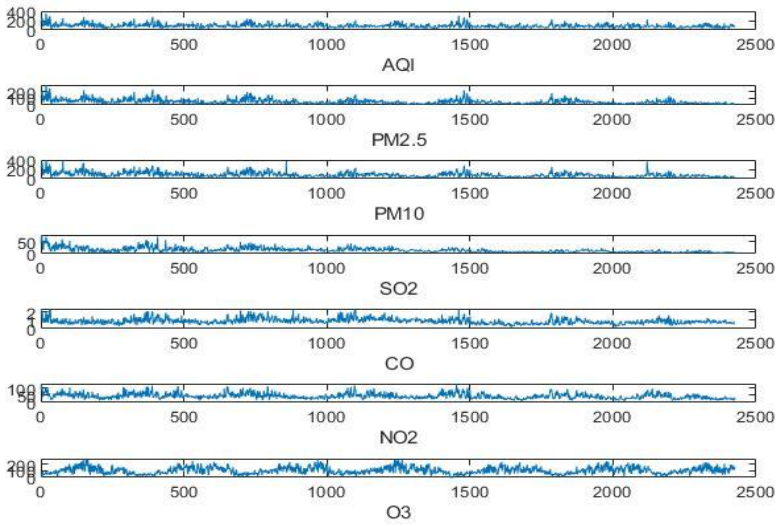

图 2.1 AQI 指数等六项特征变化折线图

由上图的折线的大致走势可以看出，AQI 指数的变化和六项污染物因子的变化趋势基本一致，能够大致看出 AQI 的变化与这六种污染物呈现正相关性；对该样本集合进行数据相关性分析，使用热力图(Heatmap)画出特征之间的相关性程度大小。热力图是一种可以直观观察出样本特征之间两两相关系数的指示图，本文使用 matlab 相关性分析函数做出热力图，如图 2.2 所示：

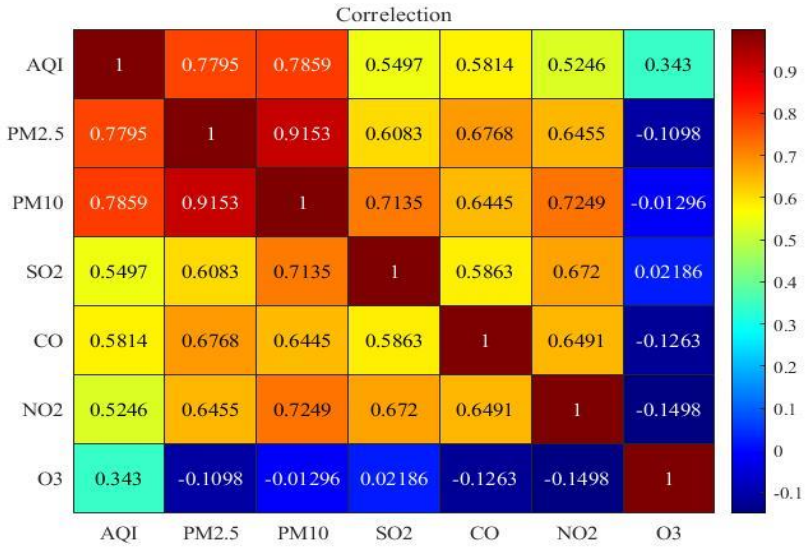

图 2.2 六项特征的相关性分析 Heatmap 热力图

图 2.2 能够通过颜色直观地观察出任意两个样本特征之间的相关性(颜色越深，数字越大，表明其相对应的相关性越大)。结果表明 AQI 空气污染指数和该六项特征都呈现正相关，其中

AQI 与 PM2.5 与 PM10 的相关系数都达到了 0.75 以上，均表现出强相关性，相关系数最低的特征是 O_3，约为 0.34，也表现出相关性。这初步表示南京市的空气污染和上述六项污染物的浓度有着紧密的联系。在大气污染的治理上，建议可以首先从限制这六项染物的排放入手。

由上图 AQI 和其他六项污染物的浓度的相关系数可以看出，数据集中的六项特征都和 AQI 保持了一定的相关性，这表示将这六项特征作为算法模型的输入特征变量，并通过模型做出决策这一方法是可行的。

2.3 逻辑回归

2.3.1 二分类逻辑回归

人工智能是计算机领域的一个研究分支，其核心思想是以数学算法作为工具，对数据的收集和挖掘 [52]，从已知数据中挖掘出有利用价值的信息。机器学习理论是人工智能的重要工具，它通过对数据的样本集合进行特征选取来构造数据集合，然后通过数学算法训练数据集合来建立特定的数学模型，并使用训练模型对新的数据样本进行预测和分类等操作。近些年来，人工智能技术在全球范围内掀起了一股研究热潮。计算机硬件性能近年来得到了大幅度提升，许多在过去限制于计算机硬件性能而难以实现的算法都可以在今天得到实现 [53]。机器学习技术理论的发展，既为人工智能的总体发展带来了理论支撑，同时也为人们利用该技术解决实际问题，指导人们的社会生产实践提供了理论指导。下面对本文中所涉及到的逻辑斯蒂回归及以及支持向量机等方法理论进行阐述。

逻辑斯蒂回归是一种基本的有监督机器学习分类方法。所谓的有监督方法[54]是在有标注(样本的实际分类即为标注)的样本数据集中建立预测模型并训练出该模型的方法。作为一种基本分类理论，它的优点是算法容易实现，缺点是运行时间受数据集规模的约束，处理多分类时需要多次运行单分类步骤，时间复杂度较高。具体来说，对于样本空间$\{x_1, x_2, x_3 \cdots\}$，可以通过 Logistic Regression 算法（以下简称 LR 算法）得到一个映射$f: x \to y$，x 是数据集中样本点集合所组成的样本空间，y 是算法所判别的分类结果，由一系列的离散值组成。

在二分类应用中，往往先在样本空间中设定一个初始判定边界，该判定边界以方程形式给出。若边界方程是线性的，则该判定边界是所谓的线性判定边界。也可以设定高次幂的边界方程来得到非线性判定边界。根据边界方程，计算样本点在当前的判定边界下的损失函数(Lost Function)，然后根据优化方法，比如梯度下降法、牛顿迭代法等得到损失函数值较小的边界方程的参数，同时对现有的判定边界进行更新，最终可以得到一个分类准确率较好的判

定边界。

在二分类问题中，单一样本点的判定通过 sigmoid 函数可以求出判定结果：

$$P(Y=1|x)=\frac{1}{1+e^{-\theta*x}} \tag{2.1}$$

其中w是判定边界，x向量是样本点的各特征值组成的向量。容易得到当$\theta^{\mathrm{T}}*\mathrm{x}$的值为 0 时，结果 P 为 0.5，$\theta^{\mathrm{T}}*x$趋向于正无穷时，结果趋近于 1，反之则趋近于 0。即函数的值域为(0,1)。根据这个特性可以设定接近1为正类，而接近0为负类。

为了找到某个比较理想的参数θ，便于对当前判定边界的好坏进行量化，引入损失函数：

$$J(\theta)=-\frac{1}{m}\cdot\sum_{i=1}^{m}[y^i\cdot log(h_\theta(x))+(1-y^i)\cdot log(1-h_\theta(x))] \tag{2.2}$$

上式中m是训练集的样本总数，y是样本的判定结果，$h_\theta(x)$是判定函数，本文中使用式 2.1 的 sigmoid 函数。损失函数的值表明了在当前选取的参数θ下，分类边界的误差程度。损失函数的计算可以使用梯度下降法，拟牛顿法等方法进行求解。经过多次迭代后可以得到比较理想的损失函数值，该损失函数所对应的分类边界方程就是所求的边界。

2.3.2 多分类 SoftMax 回归

上一小结的过程是针对二分类(即标记只有正类和负类两类)任务而言的。当实现多分类任务(类别数大于 2)的时候，需要在二分类方法的基础之上进行扩展。目前比较主流的方法有 One-Vs-All 法和 SoftMax 法。One-Vs-All 法的思路是依次将所有分类中的一类作为正类，而其他类作为负类，求解出k分类方程 (k为分类的类数)。对于新样本，分别求出其在k个分类模型下的适应度，取其中适应度最大值所属的分类即为判别类别。One-VS-All 方法的缺点是计算步骤较为繁琐，而且不能将整体特征作为训练的影响因子，在计算k个分类方程时会损失一部分精度。与 One-Vs-All 法不同，SoftMax 法将k个模型参数$\theta^{(1)}\sim\theta^{(k)}$作为参数矩阵带入式子 2.2 中直接进行求解得到算法模型。本文中将主要使用 SoftMax 法。

在 SoftMax 回归中，对于任意输入x，输出类别有k个，即$\mathrm{y_i}(i=1,2\ldots k)$，此时的判别函数形式为：

$$h_\theta\left(x^{(i)}\right)=\begin{bmatrix}p\left(y^{(i)}=1\middle|x^{(i)};\theta\right)\\p\left(y^{(i)}=2\middle|x^{(i)};\theta\right)\\\cdots\\p\left(y^{(i)}=k\middle|x^{(i)};\theta\right)\end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^Tx^{(i)}}}\begin{bmatrix}e^{\theta_1^Tx^{(i)}}\\e^{\theta_2^Tx^{(i)}}\\\cdots\\e^{\theta_k^Tx^{(i)}}\end{bmatrix} \tag{2.3}$$

所需求出的$\theta_1,\theta_2,\ldots\theta_k$就是模型的参数，写成矩阵形式就是：

$$\Theta = \begin{bmatrix} \theta_1^T \\ \theta_2^T \\ .. \\ \theta_k^T \end{bmatrix} \tag{2.4}$$

相应的，SoftMax 回归的损失函数也可由基本逻辑斯蒂回归的形式推广而来，具体形式如下：

$$\mathcal{J}(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} sign\left(y^{(i)} == j\right) \log\left(\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}\right)\right] \tag{2.5}$$

上式中m是数据集样本数量；sign是符号函数，当内括号中的表达式为真时，结果为 1，否则为 0。上式所给的损失函数的最小值并没有闭式解，一般采用梯度下降法、LBFGS、拟牛顿法等数值优化方法进行迭代逼近求解。

2.4 支持向量机

支持向量机[55-56] (Support Vector Machine, SVM)同样是一种处理分类问题的常见方法。传统的支持向量机模型的原理是在样本空间中寻找一个分类超平面，使得不同类型的样本尽可能地分布在超平面的两侧，同时分类后的样本点到超平面的距离还要尽可能地大。相比于一般机器学习方法如逻辑斯蒂回归、决策树模型、k 近邻法、k 均值聚类法等，支持向量机有着较高的运行效率和识别精度，并且存在着样本点到判别超平面的距离尽可能大的优点，使得其模型的鲁棒性也较好。正是由于这些优点的存在，使得支持向量机的应用十分广泛，并成为在样本空间容量中等、精度要求较高等需求下的优先考虑方法之一。传统的 SVM 模型是二分类模型，经过 Boser 等人[57]引进了核函数技巧后可以将特征空间映射至高维空间，能够很好地解决非线性分类问题。

2.4.1 线性可分支持向量机

若包含两种不同属性的样本点可以被某超平面完全分隔为两部分，则称该样本组成的集合是线性可分的，且分隔该样本集合的支持向量机被称为线性可分支持向量机。线性可分支持向量机的示意图如下所示：

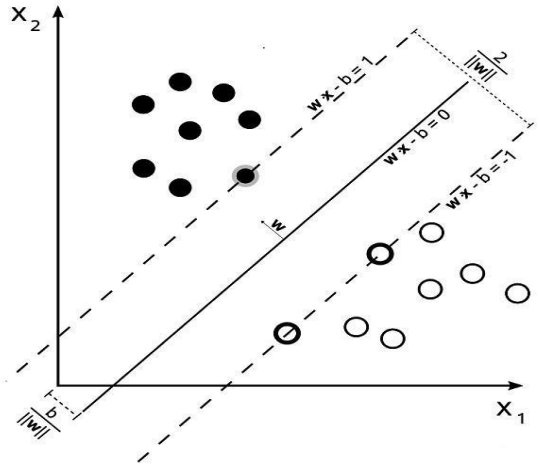

图 2.3 线性可分 SVM 示意图

在图 2.3 中可以看出，超平面$wx-b=0$表示待求的分割面，实心圆和空心圆分别代表着两类不同的样本标签 label。容易发现，存在着无数个分割面，能够将集合划分为两类。而 SVM 的目标是求出使得两类样本到分割平面的距离最大的超平面。图中所示的分割面于实心样本集合和空心样本集合的最大距离都是$\frac{1}{w}$。

SVM 的分隔面可以用如下公式描述：

$$f(x)=sign(w^T-b) \tag{2.6}$$

上式中w是分离超平面的参数，b是偏置，也叫做截距，$sign$是示性函数。SVM 的训练目标就是训练集中不同类别样本间的距离和它们到超平面的距离都应该尽可能地大。距离分隔超平面最近的两个样本与分隔平面平行的截面方程分别是：

$$w^T*x-b=1 \tag{2.7}$$

$$w^T*x-b=-1 \tag{2.8}$$

样本空间中任意两个点的间隔为：

$$\gamma=\frac{(x_+-x_-)*w^T}{|w|} \tag{2.9}$$

上式中x_+和x_-分别表示正类和负类样本的支持向量。在实际应用中，认为规定正类样本的标记y_i为+1，负类样本的标记为-1，可以得到：

$$y_i(w^T*x_i+b)=1 \tag{2.10}$$

将上式中的x_i分别替换成正负样本，并将y_i替换为相对应的标签值，可以得到下式：

$$\begin{cases} w^T*x_+=1-b \\ w^T*x_-=-1-b \end{cases} \tag{2.11}$$

将式（2.11）代入式（2.9），可以得到分类间隔为：

$$\gamma=\frac{1-b+(-1-b)}{|w|}=\frac{2}{|w|} \tag{2.12}$$

在支持向量机模型中，追求的目标是正类和负类之间的间隔尽可能地大，于是，支持向量机的求解可以形式化为求解分类间隔γ的最大值问题。这在最优化理论中可以转化为求解凸二次规划问题，注意到求解$\frac{2}{|w|}$的最大值等同于求|w|的最小值：

$$\min_{w,b}\frac{1}{2}|w|^2 \tag{2.13}$$

$$s.t.\quad y_i(w^T * x_i + b) \geq 1\,, \qquad i = 1,2,\dots N \tag{2.14}$$

上式是典型的凸二次规划问题。为了求解方便，可以添加拉格朗日乘子α，从而将上述问题转化为对偶问题：

$$L(w,b,\alpha) = \frac{1}{2}|w|^2 + \sum_{i=1}^{m}\alpha_i\left(1 - y_i(w^T x_i + b)\right) \tag{2.15}$$

这样做的好处是求解起来更加方便。

为了求解式 2.15 对偶问题，先对参数w和b求偏导数得到：

$$\begin{cases} \dfrac{\partial L}{\partial w} = w - \sum_{i=1}^{m}\alpha_i y_i x_i \\ \dfrac{\partial L}{\partial b} = \sum_{i=1}^{m}\alpha_i y_i \end{cases} \tag{2.16}$$

令上述两项偏导数为 0，可以得到：

$$w = \sum_{i=1}^{m}\alpha_i y_i x_i \tag{2.17}$$

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \tag{2.18}$$

将上式代入拉格朗日对偶问题式子中，可以得到：

$$L(w,b,\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i x_j \tag{2.19}$$

$$s.t.\sum_{i=1}^{m}\alpha_i\, y_i = 0, \qquad \alpha_i \geq 0, i = 1,2,\dots,m \tag{2.20}$$

对式（2.16）求解α的极大值，于是可以得到：

$$\begin{cases} \max\limits_{\alpha}\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i x_j \\ s.t.\sum_{i=1}^{m}\alpha_i\, y_i = 0, \qquad \alpha_i \geq 0, i = 1,2,\dots,m \end{cases} \tag{2.21}$$

上式问题的解就是最终的分隔超平面：

$$\mathrm{f(x)} = \mathrm{w^T}x + b = \sum_{i=1}^{m} \alpha_i y_i x_i^T x + b \tag{2.22}$$

SVM 模型能够在样本空间中求解出正负样本点最大间隔的分离超平面，对线性可分的数据集具有很好的分离精度。可以通过核函数技巧将特征空间映射到高维空间，映射后的模型可以很好地解决非线性分类的问题。

上述二次规划问题的求解受到的限制较大，当样本容量很大的时候，求解该问题的算法往往变得十分低效，存在着训练时间很长、占用计算机较大的内存甚至无法使用的问题。针对这一问题，有人提出了基于小样本的 SVM 模型，其思想是通过预先的筛查机制，从海量数据样本中找出对分离超平面影响最大的样本，用这些样本的子集进行训练来达到减小样本容量的目的。还有学者提出了序列最小最优化算法(Sequential Minimal Optimization，SMO)，其思想是每次迭代只使用两个样本，同时固定其他样本点(将其他样本点视作常数)。将二次规划问题转化为若干个能够得到解析解的子问题，从而大大提升了求解速度和算法的效率。

2.4.2 非线性支持向量机

线性可分支持向量机在线性问题中的效果比较理想，但是在非线性问题中不能有效求解出理想的分隔面。这时可以使用 SVM 在非线性问题中的扩展形式。在非线性问题中，要使用到核技巧使原空间映射到新空间中，在新空间中可以使用线性分类的方法将样本集合分隔开。目前学术界已经证明，对于有限维度的样本空间，必然存在一个高维空间，使得样本在该高维空间内是线性可分的。

首先定义核函数：设 X 为输入空间，H 为特征希尔伯特空间，若存在从 X 到 H 的映射：

$$\phi(\mathrm{x}): \mathrm{X} \to \mathrm{H} \tag{2.23}$$

该映射使得所有x，z∈ X都满足：

$$\mathrm{K(x,z)} = \phi(\mathrm{x})\phi(\mathrm{z}) \tag{2.24}$$

则称$\mathrm{K(x,z)}$为核函数，其中$\phi(\mathrm{x})\phi(\mathrm{z})$是内积。

将核技巧运用到线性支持向量机中，可以得到对偶问题的待优化函数变成：

$$\mathrm{W}(\alpha) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i \tag{2.25}$$

新的决策边界为：

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x_i, x) + b^*\right) \tag{2.26}$$

实际应用中，核函数的选取往往对问题求解效率有一定的影响，一些常用的核函数有：

线性核：$K(\mathrm{x_i}, x_j) = x_i^T * x_j$

高斯核：$K(x,z) = \exp\left(-\frac{|\mathrm{x}-\mathrm{z}|^2}{2\sigma^2}\right)$

多项式核；$K(x,z) = (x \cdot z + 1)^P$

2.4.3 线性支持向量机

在前两节中，主要给出了线性可分的情况的分隔面求解过程，非线性可分时通过核技巧将原空间映射到高维解空间中求出线性分隔面的过程。然而在实际工程应用中，往往并不能直接找到一个完美的分隔超平面，使得两类样本完美地分布在超平面的两侧。原因可以概括为两点：一是对于未知的样本空间，往往难以确定使用何种核函数使得样本集合恰好被该核函数映射空间中的分隔面线性划分；二是在工程问题中并不希望模型在训练集上的准确率过高，因为过高的准确率有时候意味着过拟合，即模型训练地“太好了”以至于将不存在的规律也学习到了。为了解决上面的问题，学界给出了一种软间隔的概念，所谓的软间隔是与前述的“硬间隔”相对应的概念，用来形容严格将样本空间线性划分为两类。

若样本空间线性不可分，即空间中某些点不能满足函数间隔大于等于 1 的约束，此时引入松弛变量。每个样本点x_i都赋予一个松弛变量ζ_i。当样本x_i到超平面的距离大于支持向量到超平面的距离时有$\xi_i = 0$；当样本x_i与超平面的距离小于支持向量与超平面的距离时有$0 \leq \xi_i \leq 1$；若样本分布错误的一侧边，则$\xi_i > 1$。可得引入松弛变量后的约束条件为：

$$\mathrm{y_i}(w * x_i + b) \geq 1 - \xi_i \tag{2.27}$$

对于每一个不满足约束条件的点，都施加一个常数项惩罚$C > 0$，于是目标函数为

$$\frac{1}{2}|w|^2 + C\sum_{i=1}^{m} \xi_i \tag{2.28}$$

拉格朗日对偶问题是：

$$L(w,b,\alpha,\xi,\mu) = \frac{1}{2}|w|^2 + C\sum_{i=1}^{m} \xi_i + \sum_{i=1}^{m} \alpha_i\left(1 - \xi_i - y_i(w^T * x + b)\right) - \sum_{i=1}^{m} \mu_i \xi_i \tag{2.29}$$

上式的求解过程类似于线性可分 SVM 的对偶问题。首先对式（2.29）求偏导并令其偏导为 0，可以得到对偶问题是：

$$\begin{cases} \max\limits_{\alpha} \sum\limits_{i=1}^{m} \alpha_i - \frac{1}{2} \sum\limits_{i=1}^{m} \sum\limits_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i x_j \\ \text{s.t.} \sum\limits_{i=1}^{m} \alpha_i y_i = 0, \qquad \alpha_i \geq 0, \mu_i \geq 0, C = \alpha_i + \mu_i \;\; i = 1,2,\dots,m \end{cases} \tag{2.30}$$

求解上式，可以得到分隔面为：

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i y_i x_i^T x_i + b \right) \tag{2.31}$$

约束条件为：

$$\begin{cases} \alpha_i \geq 0 \\ y_i f(x_i) \geq 1 - \xi_i \\ \alpha_i (y_i f(x_i) - 1 + \xi_i) = 0, \xi_i \geq 0, \mu_i \xi_i = 0 \end{cases} \tag{2.32}$$

2.5 本章小结

本章首先介绍了空气污染指数的相关概念，阐述影响其的六项影响因子。为了验证了六项影响因子对指数的影响程度，绘制了热力图阐述了六项影响因子与 AQI 指数的关系，得出了影响因子与 AQI 呈现正相关的结论。接下来介绍了本文中所使用的数据集，介绍了常见的逻辑回归分类方法以及支持向量机方法及其分别在线性可分、非线性可分、引入软间隔的三种情况下的理论推导，通过引入拉格朗日乘子，将原问题转换为相应的对偶问题进而给出求解，在后文中都将以这两种方法作为分类问题的基础手段进行深入研究。

第三章 基于 SoftMax-SVM 的空气质量分类判别方法

人工智能技术的发展已经与很多周边学科和产业的发展融为一体。传统的空气质量指数计算方法计算较为繁琐。机器学习方法作为人工智能的方向之一，在样本数据分类预测，挖掘数据样本信息特征等方面有着较为出色的表现。将机器学习方法与空气质量预测领域进行结合，前景十分广阔。本章给出了一种基于 SoftMax-SVM 混合模型的空气质量判别算法。首先给出算法的流程，介绍方法的整体思路，接着介绍低置信样本集合的定义并给出求解低置信样本集合的方法。接下来用实验数据进行验证，首先使用 SoftMax 回归对数据集合进行初步训练，为了增强分类精度，由结果给出低置信样本的定义，再对低置信样本使用 SVM 进行二次分类，实验结果表明使用混合模型的情况下的整体准确率要高于单一模型的准确率。

3.1 算法流程

本章提出了一种新型空气质量等级分类算法。首先使用 SoftMax 回归进行分类模型训练，然后根据模型的判别概率筛选低置信样本集合，将原样本集合划分为低置信样本集合与普通集合，并使用 SVM 对经过筛选的低置信样本集合进行再分类，对普通样本集合使用 SoftMax 模型进行分类，最终得到混合分类模型。关于低置信样本的筛选见后面的介绍。算法的流程如图 3.1 所示：

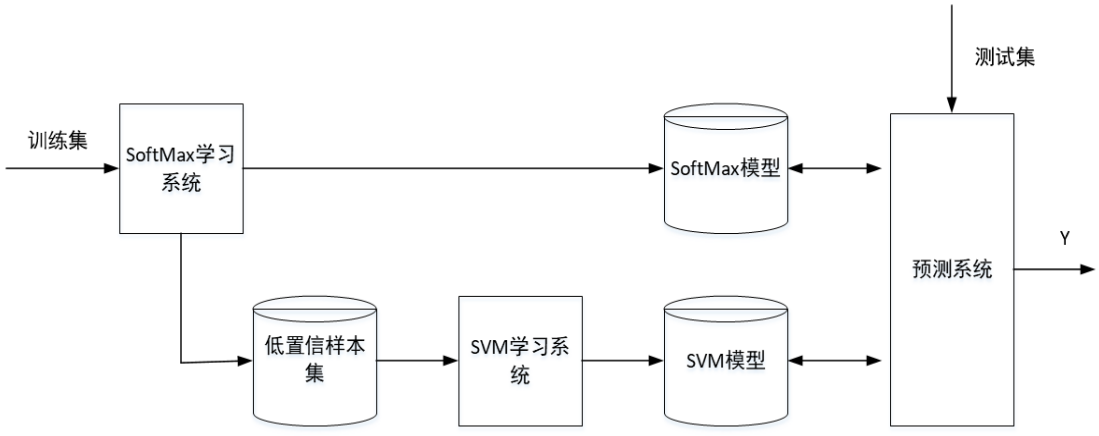

图 3.1　SoftMax-SVM 混合模型的结构图

图 3.1 可以看出，训练集样本先经过 SoftMax 学习系统进行训练，然后得到训练好的 SoftMax 模型，接着根据模型对于训练样本的判别概率的大小设置阈值，定义判别概率低于阈值的样本为低置信样本；之后对低置信样本使用 SVM 算法进行训练，得到低置信样本的 SVM 模型；完成上述的训练过程后，对预测集样本使用同样的方法进行划分低置信样本集合，

最后对两个模型训练的结果进行合并，得到预测的分类输出。

3.2 提取低置信样本集

3.2.1 低置信样本定义

在提取低置信样本定义之前首先介绍 SoftMax 模型的输出。由第二章可知 SoftMax 模型把每个样本的六个污染物因子的浓度作为模型的输入值，SoftMax 模型可以求出样本分别属于每一类的概率，并且最终通过输出值的最大值所属的类别确定该样本的判别类别。

SoftMax 回归模型的输出结果如下：

$$h_\theta\left(x^{(i)}\right)=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x^{(i)}}}\begin{bmatrix}e^{\theta_1^T x^{(i)}}\\e^{\theta_2^T x^{(i)}}\\\cdots\\e^{\theta_k^T x^{(i)}}\end{bmatrix} \tag{3.1}$$

上式表明在k分类问题中，模型对输入样本$\mathrm{x}^{(\mathrm{i})}$的输出的是一个$k$维向量，向量中的第$j$个分量表示模型判别样本$\mathrm{x}^{(\mathrm{i})}$为第$j$类的概率，向量前的系数的作用是将输出归一化。

当某个样本输入 SoftMax 模型之后，求出输出向量的分量最大值，该最大值所对应的下标j就是模型判别该样本属于j类的概率。也就是说，该值的大小决定着模型判定该样本的可能性的大小，如果该值较小，那么 SoftMax 模型判别样本属于类别的程度较低，即判别结果为错误的可能较大；反之则说明判别为该样本的概率较大，相应的判别结果正确可能性较高。

设分量最大值为p，可以知道

$$\mathrm{p}=\max\left(e^{\theta_1^T x^{(i)}},e^{\theta_2^T x^{(i)}},\ldots e^{\theta_k^T x^{(i)}}\right) \tag{3.2}$$

由式 3.1 可知 p 已经归一化处理过，其取值范围在(0,1)之间。在二分类逻辑斯蒂模型中，正类和负类的概率之和为 1。p 的取值大小代表模型判别该类属于当前类的置信程度，p 越接近 1 表示置信程度越高，反之则表示置信程度越低，这个结论也可扩展到多分类 SoftMax 模型。

定义阈值区间$[lower,higher]$，当某个样本$\mathrm{x_i}$的 SoftMax 模型输出分量的最大值p落在区间$[lower,higher]$时，也就是$p\in[lower,higher]$时，就认为模型对该样本$\mathrm{x_i}$的判别准确率较低。定义所有落在该区间下的样本集合$\mathrm{x_i}$为低置信样本，低置信表示模型对于该样本的判别结果的准确率较低；反之则表示 SoftMax 模型对于该样本的判别结果较为可靠，将该样本作为正常样本，结果予以保留。可以看出，通过对阈值区间 lower 和 higher 的灵活设置，能够

主动调整低置信样本的个数，从而可以满足不同应用场景下的需求。

3.2.2 低置信样本提取

按照上一小节中提取低置信样本集的思路，首先构建 SoftMax 模型作为空气质量指数分类判别模型。针对六类空气质量分类数据样本集合，选择数据集中所给的污染物浓度作为输入特征，先在前端进行数据的预处理，然后构建 SoftMax 模型进行训练，以构建空气污染指数分类判别算法。其基本框架结构如图所示：

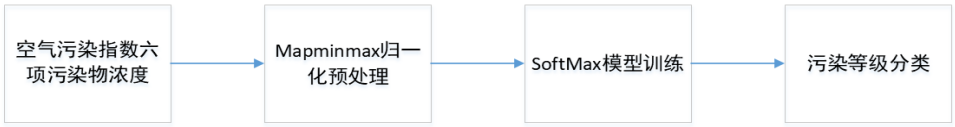

图 3.2 基于 SoftMax 模型的空气质量等级判别流程

该算法分为四个步骤：

步骤 1：获取并导入空气质量数据集；

步骤 2：对输入特征进行均值归一化预处理；

步骤 3：将预处理后的特征作为特征空间输入 SoftMax 模型进行训练；

步骤 4：使用训练好的参数模型在测试数据集上对新样本进行空气质量分类判别。

使用 Matlab 软件实现算法。步骤 1 可使用简单的 load 指令实现数据集的导入操作，六项污染物因子的浓度值作为模型的输入，空气的质量等级作为模型的标签值。

接着进行步骤 2 中的均值归一化处理。常用的均值归一化方法有以下几种：

0-1 标准化：一种常见的数据缩放方法，通过以下的变换可以将每项特征缩放到[0,1]区间中，这也是其名字的又来。其变换方法如下式所示：

$$\mathrm{x_i} = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{3.3}$$

上式中$\mathrm{x_i}(i = 1,2,\dots k)$是特征向量，对每一项特征进行上述变换后，可以将特征的取值范围设置为 [0,1]。这样可以有效避免不同的特征之间的取值差距过大而导致算法的收敛速度过于缓慢从而影响算法的性能。

其次是$\mathrm{Z-Score}$标准化：该方法变换后的特征向量的每一维特征的均值为 0、标准差为 1：

$$\mathrm{x_i} = \frac{x_i - \mu}{\sigma} \tag{3.4}$$

上式中μ, σ为该特征$\mathrm{x_i}$的均值与标准差，上式变换后的特征符合标准正态分布。

除了上面两种常见变换外，还有基于多项式的数据变换等特征变换方法。当数据集存在

缺失遗漏等情况时，可以使用删除法或插值法对缺失的数据进行调整。常见的插值法有均值插值和极大似然估计法等方法。本文中使用的数据集完整性较好，仅有个别天数存在数据缺失现象，故仅使用删除法处理样本数据集，即对于部分因为缺损的数据遗弃不用。PM10 和 SO_2 的取值差异较大，PM10 和 SO_2 的典型取值相差十分明显，对样本进行 0-1 标准化来保证算法的收敛速度。本章中使用 Matlab 中的 mapminmax 函数实现 0-1 标准化。

本文中使用的原始数据集是不平衡数据集，也就是说不同标签下的样本个数差距较大。本文的空气污染数据集中良好和中等的样本的数量较多，而重度污染和严重污染的数量较少，直接用分类算法在数据集上得到的模型会出现误差较大的问题。为了解决这一问题，我们对原数据使用过采样 SMOTE 算法。该技术在 2002 年由 Chawla 提出，是目前工业界处理非平衡数据集的一种常用手段，其基本思想是采用 K 近邻方法对少数类别的样本进行模拟并生成人工样本，使得原始数据中的类别不再失衡。本文中使用这一技术对原南京市空气污染数据集进行处理。

步骤 3 中将经过删除缺损数据以及 0-1 标准化的数据样本输入 SoftMax 模型进行训练，损失函数为式（3.1）定义，随机初始化Θ参数取值全为 0.005，使用 L-BFGS 优化方法进行参数迭代求解，设定迭代次数为 300 次，训练样本集个数为 6024,测试集样本数为 869。具体的参数列表如表 3.1 所示：

表 3.1　SoftMax 参数设置表

参数	设置值
学习率	自适应
优化方法	L-BFGS
迭代次数	300
训练样本数	6024
初始 θ	全 0.005

使用上表给出的参数在训练集的 6024 个样本上进行模型参数训练，并做出损失函数值的变化曲线，如下所示：

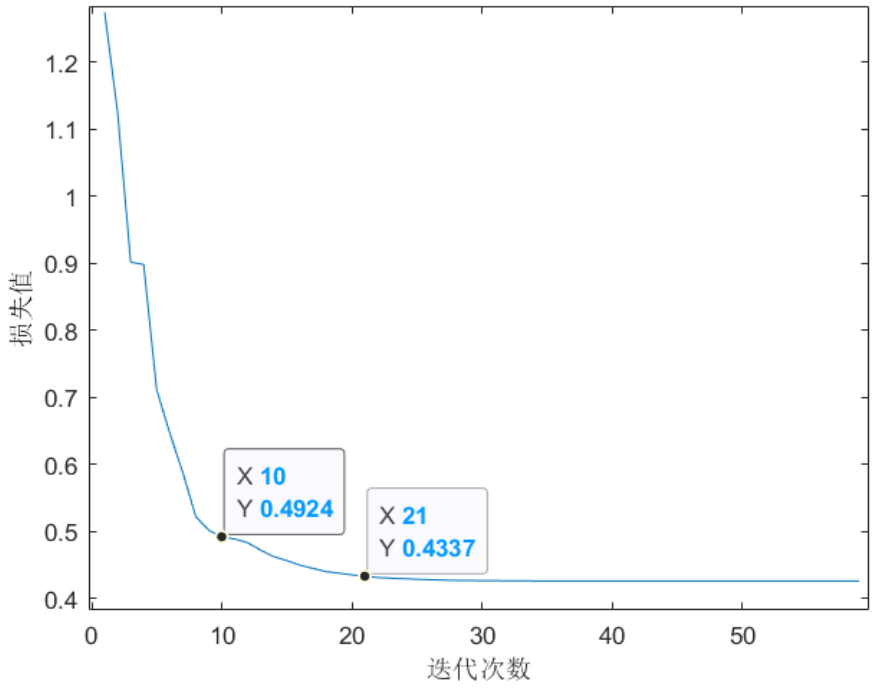

图 3.3　SoftMax 迭代损失变化图

从图 3.3 中能够看出，随着迭代次数的增加，算法的损失值逐渐下降，参数随着迭代对于数据集的拟合程度越来越好。当迭代次数到达 10 左右时，损失值为 0.4924，且速度已经很慢；迭代次数到达 21 时，损失值达到 0.4337，已经逐渐收敛。训练完成时模型的训练集拟合判别准确率约为 86.7%。

得到了参数模型后，验证模型在测试集上的准确率。使用训练得到的模型参数，将测试样本作为测试集进行预测判别。求出测试集中每个样本的预测分类类别，当预测类别与样本的实际类别一致时，模型判别正确，否则判别错误。最后求得预测准确率为 67.43%。能够发现，预测集样本的准确率和训练集样本的准确率存在一定的差别。这表示训练集中存在着一定数量的低置信样本，它们对模型的预测能力产生了一定影响。SoftMax 模型在训练集以及测试集上的分类准确率结果如表 3.2 所示：

表 3.2　SoftMax 预测结果表

训练集准确率	86.7%。
测试集准确率	67.43%

经过实验分析，可以发现经过 SoftMax 模型判别的训练集中存在着一定数量的低置信样本，并且这些样本的分类准确率往往不是很理想。经过多次实验，将样本按照下表的阈值区间进行划分原样本，分别记录各阈值区间的样本个数，以及该区间下的样本的判别准确率，得到训练集各阈值区间的样本数量和准确率的相关数据如表 3.3 所示。

表 3.3　SoftMax 模型在训练集中各置信区间下的样本数量分布及准确率

阈值区间	样本数量	准确率
0-0.6	1018	62.67%
0.6-0.7	1007	72.59%
0.7-0.8	1042	89.54%
0.8-0.9	860	96.4%
0.9-1.0	2097	99.76%

通过上表可以看出，把训练集按照模型判别结果区间[0 , 0.6]、[0.6 , 0.7]、[0.7 , 0.8]、[0.8 , 0.9]、[0.9 ， 1]进行划分。各区间样本的分类准确率随着阈值的提升而上升，这和前面所给出的低置信的划分依据的含义是一致的，即阈值越低，表示模型对于样本的判别“把握”越小，相应的分类准确率越低。

3.3 实验结果与分析

3.3.1 混合模型判别

在 3.2 节中给出了低置信样本的定义以及根据给定阈值区间筛选低置信样本集合的方法。本节将给出混合算法模型的理论依据和原理，并阐述算法的流程并且给出相应的结果。

传统的机器学习方法往往需要大容量数据样本才能训练出比较理想的模型。在实际应用中往往很难获取到海量的数据样本，即使获得了大数据样本，在其中真正有效的数据样本往往只占很少的一部分。如何使用小数据样本训练出理想的模型就具有十分重要的意义，在第二章中介绍的 SVM 支持向量机模型是在的这一背景下提出的。SVM 成立之初就有着能够在有限的小容量样本集合中获得不错的识别精度的优点。在上一小节中在求解 SoftMax 算法模型的同时也得到了“副产物”——低置信样本集合。低置信样本作为原样本的子集，样本的数量小于原集合，符合小数量样本集合的特点。同时，又知道 SoftMax 是线性分类模型，被该模型判别为低置信样本的线性特性一般较差，往往表现出较强的非线性特性，而支持向量机特别擅长处理非线性分类的问题。综上所述，低置信样本具有样本容量小、非线性分布的特点，适合在支持向量机模型的场景下使用。

本章给出的混合 SoftMax-SVM 模型通过 SoftMax 模型筛选出低置信样本，并使用 SVM

支持向量机对低置信样本集合进行训练得到 SVM 模型，与正常样本的 SoftMax 模型进行结合，这样就得到了混合算法模型。

基于混合模型的支持向量机空气质量分类判别的流程如下：

输入：南京市空气质量样本集$T = \{(X_1, Y_1), (X_2, Y_2), \ldots (X_m, Y_m)\}$；阈值区间的上下限；

输出：算法的 SoftMax-SVM 混合模型。

STEP1： 使用样本集合训练 SoftMax 模型，得到模型的参数矩阵Θ。

STEP2: 使用训练所得参数矩阵Θ拟合原数据样本，求出预测概率。

STEP3: 根据阈值上下限所定的阈值区间对样本集进行筛选，将预测概率符合的样本划分为低置信样本集合。

STEP4: 使用上一步中得到的低置信样本集合训练支持向量机模型，得到低置信样本判别模型。

STEP5: 将混合模型作为训练样本模型对新样本进行预测分类。先用 SoftMax 模型拟合新样本特征，判定结果是否符合阈值条件，如果符合低置信条件，则使用支持向量机模型进行分类的预测；如果不属于低置信样本，则正常使用 SoftMax 模型进行判别。最后返回归并后预测的类别结果。

3.3.2 结果比较

SoftMax-SVM 混合模型训练以及测试环境是在 matlab r2019b。按照上述算法的流程，首先在训练集样本上进行参数训练，然后在测试集上进行结果预测。为了对比混合模型与单一模型的差异，分别记录不同阈值区间的训练集以及测试集样本个数、分类准确率，并单独对比在单一 SoftMax 模型和 SVM 模型下的训练以及预测准确率。首先在训练集上进行参数训练，获得的结果如表 3.4 所示：

表 3.4 训练集上单一算法的准确率对比

阈值区间	0-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1.0
样本个数	1018	1007	1042	860	2097
SoftMax 模型准确率	62.67%	72.59%	89.54%	96.4%	99.76%
SVM 模型准确率	75%	78.35%	91.62%	98.36%	100%

由上表可以看出，对于阈值区间[0 , 0.6]下的样本，SoftMax 模型在训练集上的判别准确率只有 62.67%，而支持向量机的拟合判别精度却为 75%，这表示 SVM 对该阈值区间下样本

的判别精度要明显好于前者。从数据结果中可以看出，[0.9 ,1.0]区间的样本数量最多。[0.7,0.8]，[0.8,0.9]，[0.9,1.0]这三个区间下，SVM 比 SoftMax 的提升较小，总体而言差距不大。在本章中对于低置信样本使用 SVM 模型进行训练和预测，对于普通样本使用 SoftMax 模型进行训练和预测。

下一步在测试集上进行空气质量等级的预测判别。为了找到一个最佳的阈值区间，使得该区间下混合模型相比较与单一模型的预测准确率的提升能达到最理想的效果，按照 0.1 的准确率提升作为步长，分析了[0,0.6]，[0,0.7]，[0,0.8]，[0.0.9]的样本作为低置信样本的情况下，混合模型对比单一模型的增长，得到的结果如下：

表 3.5 测试集上混合模型和单一模型的准确率对比

阈值区间	0-0.6	0-0.7	0-0.8	0-0.9
普通样本数	580	365	174	69
低置信样本数	289	504	695	800
低置信样本 SVM 准确率	74.68%	87.68%	80.55%	80.09%
普通样本 SoftMax 准确率	74.48%	77.81%	82.18%	91.3%
SoftMax+SVM 分类准确率	74.57%	83.53%	80.87%	81%
单 SoftMax 模型准确率	67.43%	67.43%	67.43%	67.43%
准确率提升	7.14%	16.1%	13.44%	13.57%

从上面的表中可以看出，总体上本节提出的混合支持向量机模型的预测准确率整体上要比单一 SoftMax 回归判别模型的性能要高。具体地说，单一 SoftMax 模型在预测集上的判别准确率为 67.43%，判别正确样本数为 586 个；而使用了混合 SoftMax-SVM 的模型在不同低置信样本划分的情况下的准确率都要高于前者；在使用[0 ,0.6]区间样本作为低置信样本时，准确率提升 7.14%，在使用[0 ,0.7]区间样本作为低置信样本时，准确率提升 16.1%，提升效果最好。在使用[0 ,0.8]区间样本作为低置信样本时，达到 13.44%。从实际应用的角度来看，[0,0.9]区间相比[0.0.8]的准确率提升并不大，所以得出结论，在本数据集上，使用[0,0.7]区间作为低置信样本的效果较为理想。

3.4 本章小结

本章研究了基于 SoftMax 与 SVM 支持向量机混合模型的空气质量等级判别算法。首先介绍了基本逻辑斯蒂回归方法在多分类任务中的扩展——SoftMax 方法，然后给出了本章中混合算法的求解流程图，接着提出了低置信样本的概念和定义，而后给出不同阈值区间下低置信样本集合的各自训练集判别精确率，验证了判别准确率随着阈值取值而变化的规律。进而提出了使用支持向量机模型对于低置信样本集合进行训练、对普通样本进行 SoftMax 模型训练的混合模型算法。最后按照不同的阈值区间划分测试集样本并送入训练好的混合模型中进行分类判别，并按照实验的结果给出了提升效果最好的阈值区间，同时也指出在其他数据集中使用也可以使用同样的流程求出最佳阈值。实验结果表明，使用 SoftMax 与支持向量机混合模型的算法，预测判别的效果要好于单一逻辑斯蒂模型。但该算法因采用混合模型进行训练与分类，在提升判别性能同时，增加了计算量、多出了样本筛选等步骤，步骤相对繁琐，增加了计算量。

第四章　基于群智能优化的空气质量分类判别方法

在第三章中提出了一种基于混合 SVM 模型的 AQI 质量等级判别算法。在使用 SVM 求分隔面的时候，使用了默认的模型参数，分别是$C=1$以及$g=\frac{1}{n}$，其中n是数据的维度，C是惩项，C越大则分离面对错误分类的惩罚越大，对于训练的精确率要求就越高，反之C越小，则模型对错误的惩罚越小，对于错误分类的容忍度越高，对训练结果的精确率要求越低；g的默认值为$\frac{1}{n}$，其中n是每个数据集样本的特征的数目。可以看出，选取不同的模型参数对于模型训练的结果会产生影响。使用默认参数在大多数情况下可以取得足够的分类精度，但是在实际应用中，对 SVM 模型的参数进行恰当的优化，能获得比默认参数情况下更加好的判别精度。传统的方法是使用网格搜索法，即通过设置参数的搜索区间，按照一定的步长进行依次求解，比较每一个步长对应的结果，使用结果最理想的那一个步长所对应的参数。这样做的缺点是搜索的效率不高，如何高效率寻找最佳参数来优化 SVM 模型的性能十分具有重要的研究意义。本章结合典型的群智能算法作为数值优化方法来给出一种在实际应用中寻找最优化参数的方法。

4.1 群智能算法概述

群智能是指生物主体或者模拟其行为的主体通过合作来进行某种行为从而表现出智能性的特性。它本质上是对生物群体的某种行为(捕食、繁衍等)的模仿，具体地说，是将这些行为的模式转化为数学模型，从而在数学计算问题上进行优化求解的一种方法。也就是说，群智能算法也是受到自然界中动物行为的启示而模拟衍生而出的数值优化算法。它可以模拟非常多物种的生态行为，例如昆虫(蚁群、蜂群等)、群居性脊椎动物(鸟、兽等)等，往往用来求解空间中的数值分布问题。不同于机器学习中需要三要素：模型、算法以及策略，群智能算法往往不需要提供某种特定形式的模型，它仅需要一些特定形式的求解策略，不需要事先建模，对数据进行训练等过程，并依赖于算法本身就可以较好地求解。

群智能算法相对于机器学习方法和深度学习来说较易实现，仅仅涉及各种基础的数学操作步骤，由于不需要实现模型训练的过程，所以对于内存和处理器的要求不高。在逻辑回归中，对模型的求解使用了梯度下降法，该方法对于模型的优化需要保存每一次迭代的梯度信息，而在 SVM 模型的求解过程中，每次迭代都必须遍历全部样本来找出当前的支持向量，

这些都对机器的内存容量以及处理器运行速度有着一定的要求。在大部分群智能算法中，每一次只需要输出待优化函数模型的结果，而不需要求解梯度信息，同时算法求解过程中也不需要遍历全部的样本集合，这些优点都极大地简化了数值优化问题的求解过程。目前学术界所大量提出的群智能理论和方法的研究都证明了：群智能算法能够解决大多数全局最优值问题，是一种十分有效的求解数值优化的方法。随着各种理论的深入研究与充分实践，目前智能优化方法已经在机器学习、自动控制系统、模式识别、路线规划等领域中得到了运用，并且随着理论的不断发展展现出强大的活力。

传统的群智能理论有两种主要的方法：蚁群算法与粒子群算法。蚁群算法模拟了对蚂蚁群体采集食物的过程，后者起初是模拟鸟类觅食的过程，后来也引进了其他的因素后实现了对鸟类的其他社会系统的简单模仿与实现。他们都被证明是很好的数值优化方法。近年来很多国内外学者们在前两种基础算法的思路之上提出了其他的新方法，比如鲸鱼优化算法、人工蜂群算法、布谷鸟算法等。本章使用了灰狼优化算法与粒子群算法，结合支持向量机方法，从而实现对最优化参数的求解。下面分别介绍灰狼算法和粒子群优化方法。

4.1.1 GWO 算法

灰狼优化算法(Grey Wolf Optimization Algorithm, GWO)是由 Mirjalili 等人于 2014 年提出的一种新型启发式算法[58-59]。他们从狼群的捕猎行为以及狼群的社会结构等要素中得到了启发从而给出了一种数值优化的方法。GWO 的核心思想是仿真灰狼群的集体捕食行为，它基于狼群的社会等级结构、结构之间的合作机制，实现了对最佳参数的搜索过程。该算法的描述简洁，具有较强的全局搜索能力以及较快的收敛速度，诞生之后就迅速得到广泛的使用。

灰狼是群居性捕食者，其种群的内部是一种基于等级制的种群结构，基本可以划分为四类等级，按照社会地位从高到低级依次为$\alpha,\beta,\delta,\omega$。

狼群中的第一层为首领，称为α，它决定了整个群体的一切重大问题，比如进行捕食，休息的时间地点等。在算法中，α因子的决策权最高，它负责全局的收敛；处于第二层结构的为β狼，它们的主要任务是其辅助第一层的首领狼做决策以及进行其他的行为，β狼严格听从并执行首领狼的命令，它将狼群中其他个体的行为的结果反馈给首领狼。在算法中，该层的因子负责收集下层的信息，并向上一层提供反馈。位于第三层的是δ狼，它们听从于首领狼以及β狼的命令，它们主要负责放哨以及侦察任务，在团队中更多扮演执行者的角色，它们也负责清除狼群之中的老弱病残者，保持群体的战斗力。在算法中，主要负责收集底层的信息，并向上级提供反馈；处于最后一层的狼为ω狼，它们必须听从其他所有高层狼的命令，在捕食

行为中它们是最后被允许进食的一批狼，其所有行动都要受上层狼的指示，算法中负责向上层提供反馈。灰狼群的捕食行为主要归纳为以下三个步骤：跟踪并接近猎物、包围猎物直到其停止跑动、攻击猎物。可以看出，如果将上述的三个过程运用到数值优化问题中，就可以得到比传统方法更好的效果。将狼群的的等级制度和捕食行为演化为数学模型，就能得到灰狼算法的求解步骤。

首先给出狼群中每类成员在数学模型中的具体意义。对于某个函数，假设解空间为X，输出称为该函数的适应度 fitness。将取得最佳适应度时的解称作“猎物”，定义一个初始解集合，该集合中的所有元素称为“狼群”。对于一般问题，狼群的数量一般为 5-12 之间。将输出为最佳的适应度(也就是最佳参数对应的输出)时的解定义为α狼的空间位置；相应地，将次佳的适应度，即仅比最佳适应度稍差，但优于其他解对应的解定义为β狼的空间位置、将其次适应度，即比最佳和次佳适应度差，但是优于解空间中的其他值对应的解定义为δ狼的空间位置。最后将所有解空间中的其他解，即剩余解定义为ω狼的空间位置。上述过程模拟了灰狼群的等级制度，在实际求解的过程中，前三层的狼个体指导了求解的方向，是重点关注的部分。

狼群对于猎物的追捕过程可以用下面的数学公式描述：

$$D = \left|C * X_p(t) - X(t)\right| \tag{4.1}$$

$$X(t+1) = X_p(t) - A * D \tag{4.2}$$

上式中t为算法迭代的次数，X_P是猎物（也就是最优解）的位置向量，X表示灰狼(解空间中的某个解)的位置向量；A和C是参数向量，参数向量的表达式如下所示，

$$A = 2ar_1 - a \tag{4.3}$$

$$C = 2r_2 \tag{4.4}$$

上式中，a是参数，它的取值是随着迭代次数的进行从 2~0 递减，r_1和r_2在每次迭代时从 0~1 随机变化。

在搜索空间中，一开始，并不知晓最优解的具体位置。假定首领狼α和次级狼β对于猎物的位置敏感，在每轮迭代后保存前三个最好的适应度值并将其依次赋值给α,β,δ，并标记这些狼的位置信息给下一次迭代做准备。下一次迭代时，使用上一轮的前三层狼作为引导更新解空间，具体方式如下：

$$D_\alpha = |C_1 * X_\alpha - X| \tag{4.5}$$

$$D_\beta = \left|C_2 * X_\beta - X\right| \tag{4.6}$$

$$D_\sigma = |C_3 * X_\sigma - X| \tag{4.7}$$

$$X_1 = X_\alpha - A_1 * D_\alpha \tag{4.8}$$

$$X_2 = X_\beta - A_2 * D_\beta \tag{4.9}$$

$$X_3 = X_\sigma - A_3 * D_\sigma \tag{4.10}$$

$$X_{t+1} = \frac{X_1 + X_2 + X_3}{3} \tag{4.11}$$

4.1.2 PSO 算法

粒子群优化方法(Particle Swarm Optimization, PSO)是 Eberhart 等设计的一种全局优化进化方法[60-61],其设计思路和其他的数值优化方法类似,也是基于迭代寻优的思想。类似于 GWO 算法对于狼种群行为的模拟，PSO 算法模拟了粒子群(一般也作鸟群)的合作竞争行为，其思想是通过种群内部各粒子之间信息的共享(即合作行为)，来使得各粒子符合有序排列分布。与 GWO 方法的最大不同之处是引入了速度的概念，每个粒子在迭代中除了更新当前的位置外，还将更新自己的速度，这样在每一轮更新解的时候可以引入上一轮的速度变量，使得适应度的下降和收敛变得更快。

粒子群算法将解空间中的每个解都视为一个基本粒子；算法开始时随机生成初始的解，然后根据当前粒子群计算适应度值。在迭代的过程中,粒子将依据两个值来进行位置的更新，它们分别是$Xbest_p$和$Xgest_g$，前者表示局部最优适应度，即当前代粒子群的最佳适应度值，后者表示包括历史代中种群中目前所找到的最佳适应度值，表示全局最优值。求出上面两个最优适应度后，可以求出速度矢量：

$$V(t+1) = \omega V(t) + \eta_1 r_1 \left(Xbest_p - X(t)\right) + \eta_2 r_2 \left(Xbest_g - X(t)\right) \tag{4.12}$$

上式中$X(t)$表示粒子的适应度值，η_1, η_2为常数，在实际使用中一般都取值 2，r_1, r_2表示(0,1)之间的随机数，ω是惯性权重，一般随着迭代线性减少。

根据当前代的速度向量以及位置，可以进行位置的更新：

$$X(t+1) = X(t) + V(t+1) \tag{4.13}$$

通过迭代的进行，粒子的位置被不断更新，最后找到满足条件的适应度。

4.2 基于 GWO-PSO 优化的支持向量机模型

根据引言中的描述，在支持向量机中，参数C和g的选择对总体模型的准确率存在影响。将支持向量模型的输出分类错误率作为模型的适应度 fitness，注意到错误率等于 1-正确率，

于是上述过程实质上就是用GWO算法找出SVM模型的错误率最小的时候的参数向量$[C, g]$。针对灰狼算法中存在容易陷入局部最小问题以及粒子群算法运行时间较长的问题，本小节研究了一种结合了灰狼优化和粒子群优化两种方法的混合模型。该模型针对 GWO 优化方法的不足之处，结合了 PSO 方法加以改进，使得能够跳出局部最优解从而得到更高的空气质量判别精度，同时相比于粒子群方法在算法的执行时间上所花费的时间更少。

4.2.1 算法思路

在 GWO 算法中，初始化狼群的位置时采用了默认初值的方法，但是研究表明，使用混沌映射初始化的方法往往效果更好。所谓混沌映射，区别于一般随机映射的区别是混沌映射使用了某种确定性的方法产生伪随机序列，而一般随机映射产生的随机数则不能够通过确定性方法产生。常见的混沌映射有 Logistic 映射，Gaussian 映射，Chebyshev 映射等。本章使用 Logistic 映射对灰狼的初始位置进行初始化操作。Logistic 映射通过迭代产生，它的效果类似随机序列，但是具有确定性，可以通过迭代方式给出：

$$a_{\mathrm{n+1}} = k a_n (1 - a_n) \tag{4.14}$$

k 为常数，研究表明 k=4 时生成随机数的分布较为均匀，随机性最好。本章中取 k=4 的 Logistic 映射来初始化灰狼位置。

在 GWO 方法中狼群位置的更新利用了参数$\mathrm{D}_{\alpha}, D_{\beta}, D_{\delta}$，该参数表示狼与猎物之间的位置向量，即每次更新位置都只是使用了狼群和猎物之间的位置；然而在 PSO 算法中，粒子更新位置时不仅依赖上一代的位置，同时也依赖速度向量，从而能够提升收敛速度与精度。

参照 PSO 的位置向量的思想，在 GWO 优化引入方向向量的概念，位置向量更新方法为：

$$\mathrm{V(t+1)} = \omega * \mathrm{V(t)} + \sum_{\mathrm{i=1}}^{3} (\eta_i * k * \left(X_i - X(t)\right) \tag{4.15}$$

η_i为α,β,δ狼的各自权重，本节中取η_1=2，η_2=1.5，η_3=1。k 是协参数，其值为$0.1 * rand(0,1)$。ω为惯性系数，其值随着迭代次数递减，通过下式给出：

$$\omega = \omega_{max} - (\omega_{max} - \omega_{min}) * \frac{n}{n_{max}} \tag{4.16}$$

灰狼算法存在着收敛较慢的特点，为了增加收敛速度，对每代灰狼引入“末位淘汰”机制，思路是每代产生的灰狼通过适应度的筛选，将适应度较低的灰狼直接淘汰，这样进入下一轮迭代的灰狼就具有更“优秀”的适应度，从而在整体上提高收敛速度。具体操作为如下：每轮迭代更新完狼群位置后，根据当前的适应度值由低到高排列，将ω狼中排名末三分之一

的予以淘汰。由于这类个体对最佳参数的寻找的贡献较低，所以将末尾的种群淘汰掉后能够有效提升收敛速度，同时也有助于找到更优秀的个体。淘汰的机制如下：

$$X' = \frac{1}{2} * X + \frac{1}{\sum_{i=1}^{3} \eta_i} \sum_{i=1}^{3} \eta_i * X_i \tag{4.17}$$

上式中X'是待替换的个体，X是被替换的个体。参数η在式子(4.15)中给出。

4.2.2 算法流程

根据上一小节描述的流程，将改进的 GWO-PSO 混合方法流程总结如下：

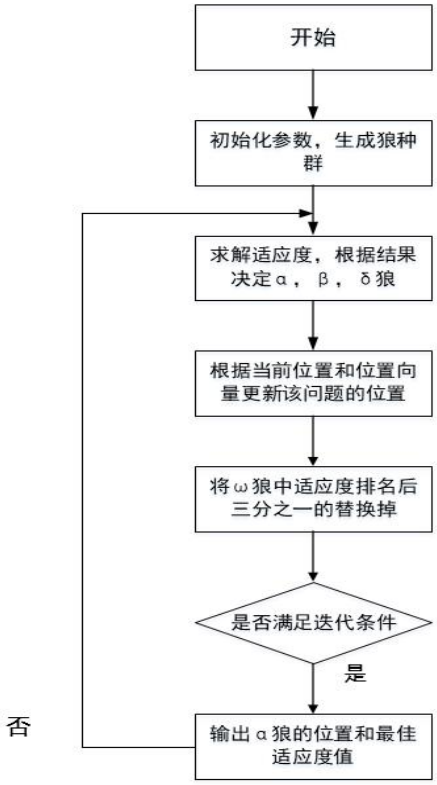

图 4.1　GWO-PSO 方法流程图

步骤 1 种群和参数初始化。使用混沌初值代替随机数初值，使用式 4.14 初始化狼群的解空间X_i(i=1,2,3…n)。

步骤 2 求解当前的适应度值。根据式 4.5-4.11 计算 GWO 算法狼群的适应度，将适应度排名前三的个体依次设置为α,β,δ。

步骤 3 计算狼群个体的位置向量。根据式 4.15，使用当前位置和位置向量 V 更新种群中个体的位置。

步骤 4 执行淘汰过程。对执行过步骤 3 的所有狼群个体使用淘汰机制进行更新，将适应度排名中ω狼的后三分之一进行替换，根据式 4.17 得到新的个体。

步骤 5 判断当前适应度值是否达到要求，如果满足要求，则算法停止并输出当前适应度值；否则执行步骤 2，直到适应度满足要求。

4.3 实验结果分析

在测试集上验证结果，将测试集的 SVM 分类错误率作为适应度 fitness，适应度越小则表示预测集的精确率越高，分别使用 GWO 优化算法、PSO 优化算法、GWO-PSO 混合算法来以降低分类错误率为目标最优值进行最佳适应度值搜索。设定迭代次数 50 次，GWO 优化算法的适应度变化如图 4.2 所示：

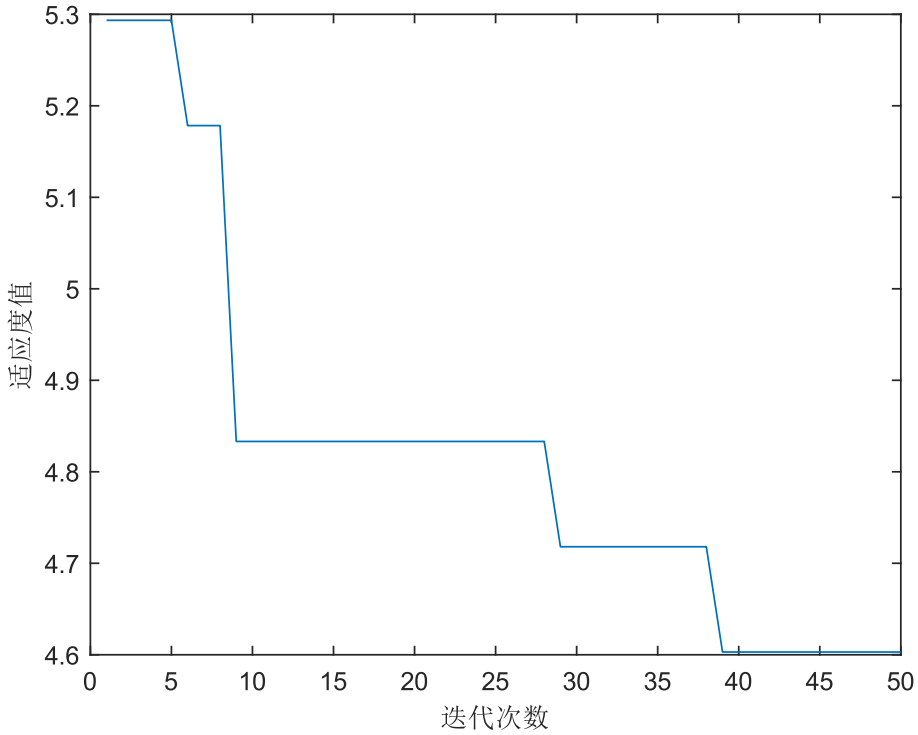

图 4.2 GWO 算法适应度随迭代进行变化图

由图 4.2 可知，随着迭代次数增加，算法整体的适应度逐渐减小，在 39 次时适应度降低到了 4.6%，此时 SVM 在测试集上的准确率为 95.4%。

为了方便对比预测结果与数据集的样本实际标签的差距，我们在平面图上标注出所有样本的模型预测结果和实际标签，横轴作为样本的序号，纵轴为离散值 1-6，表示六类标签值，结果如图 4.3 所示：

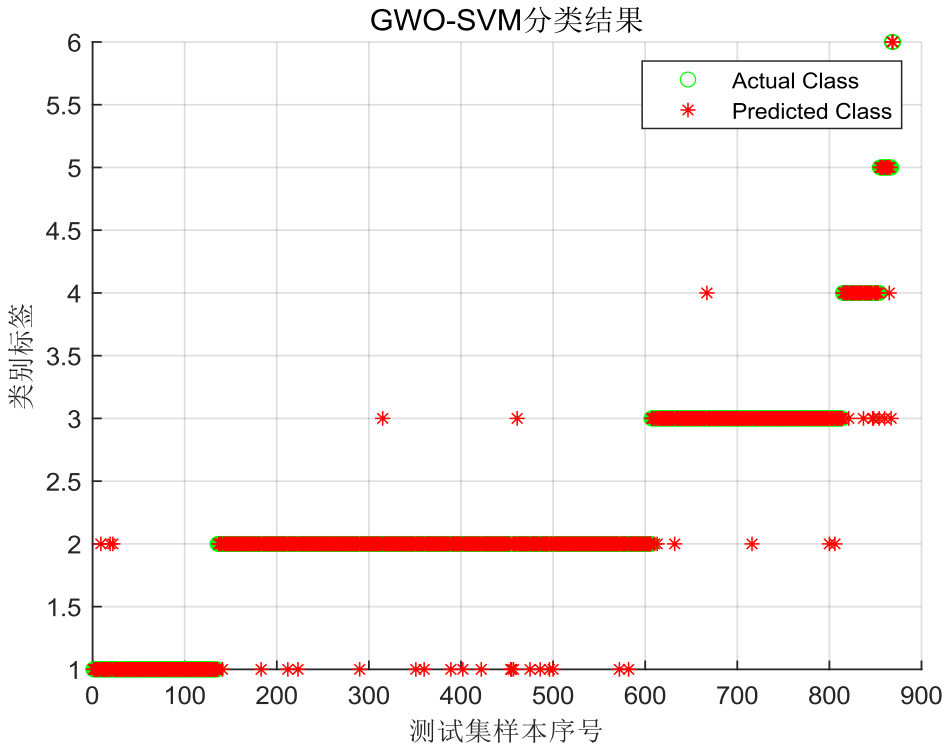

图 4.3 GWO-SVM 模型 AQI 分类效果

图 4.3 中，绿色圆圈表示实际样本，红色星点表示预测样本；可以看出，实际样本和预测样本的重合程度较高。存在着若干预测样本中没有覆盖实际样本的部分，这一部分表示预测分类错误的样本；在 50 次迭代后得到的适应度值为 4.6%，由此求出 GWO-SVM 的分类准确率为 95.4%。

接下来进行 PSO-SVM 的实验。根据 4.1.2 节中粒子群优化 PSO 的运行机制，使用 PSO 优化方法与支持向量机模型结合，得到 PSO-SVM 优化方法。仍然将支持向量模型的输出分类错误率作为模型的适应度 fitness，注意到错误率等于 1-正确率，也就是当求出适应度最小值即为所求的最优解。用 PSO 算法找出 SVM 模型的错误率最小的时候的参数向量$[C, g]$。

在测试集上验证结果，将测试集的 SVM 分类错误率作为适应度 fitness，适应度越小则表示预测集的精确率越高，用 PSO 优化以降低分类错误率为目标最优值进行搜索，设定迭代次数 50 次，适应度变化如图 4.4 所示：

由图 4.4 可知，随着迭代次数增加，算法整体的适应度逐渐减小，在 39 次时适应度降低到了 4.6%，此时 SVM 在测试集上的准确率为 95.4%。

为了方便对比预测结果与数据集的样本实际标签的差距，我们在平面图上标注出所有样本的模型预测结果和实际标签，横轴作为样本的序号，纵轴为离散值 1-6，表示六类标签值，结果如图 4.5 所示：

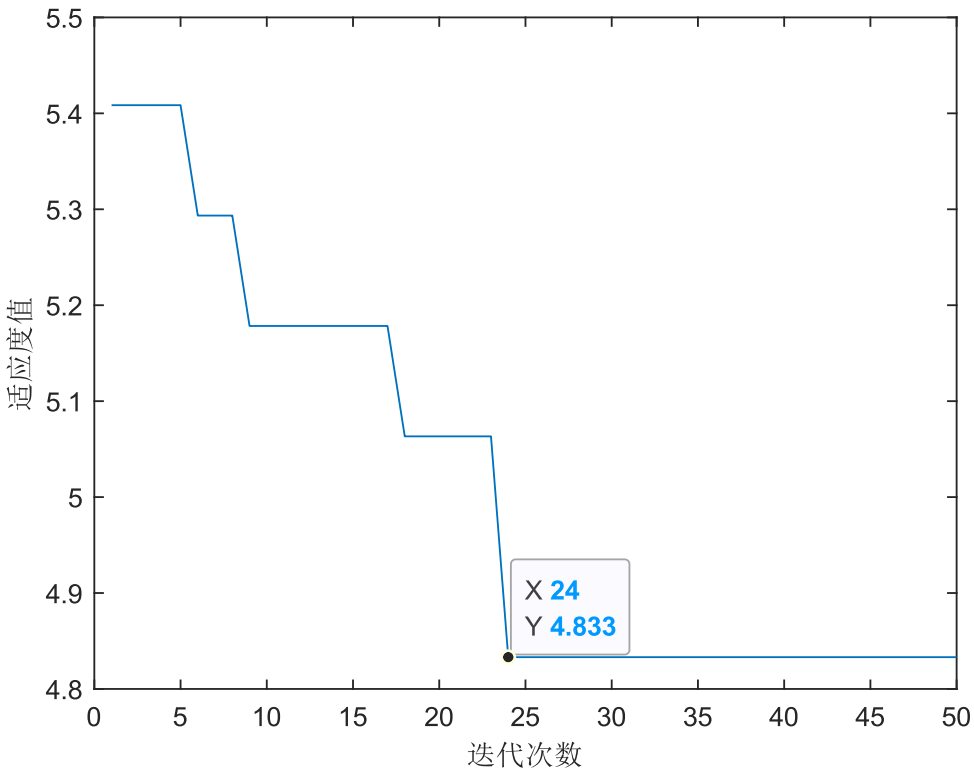

图 4.4　PSO 算法适应度随迭代进行变化流程图

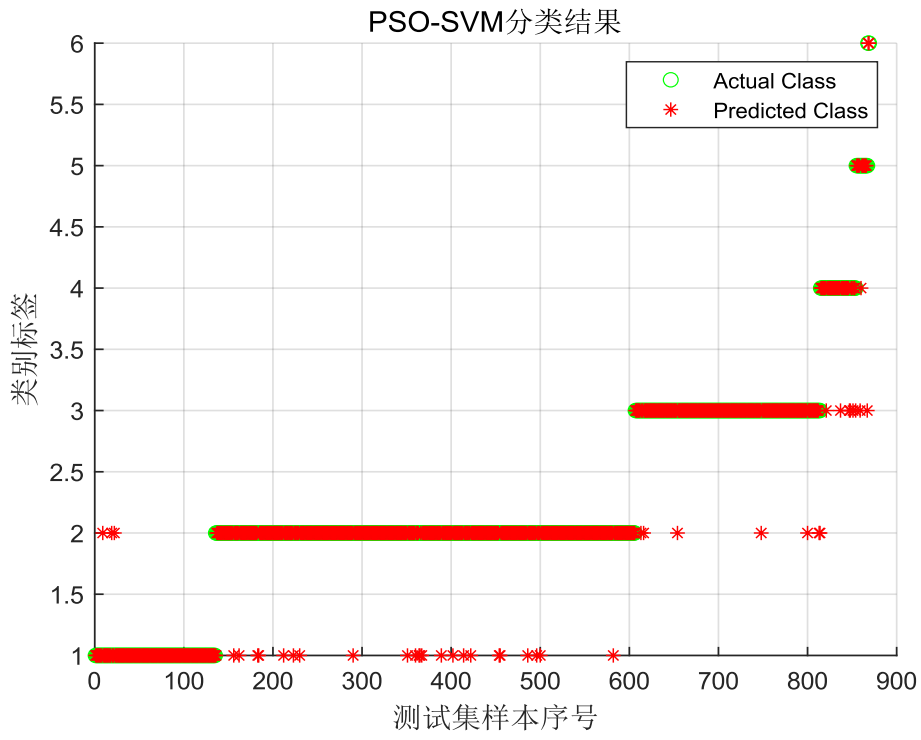

图 4.5　PSO-SVM 算法 AQI 分类效果

图 4.5 中，绿色圆圈表示实际样本，红色星点表示预测样本；可以看出，实际样本和预测样本的重合程度较高。经过计算得出，使用粒子群算法 PSO 结合 SVM 的分类准确率为 95.2%，略低于 GWO-SVM 的结果。

最后，为了验证本章中方法的有效性，下面使用本章中介绍的 GWO-PSO-SVM 方法对最佳适应度进行迭代搜索，将结果与单独使用两种方法的情况进行比较，得到结果如下所示：

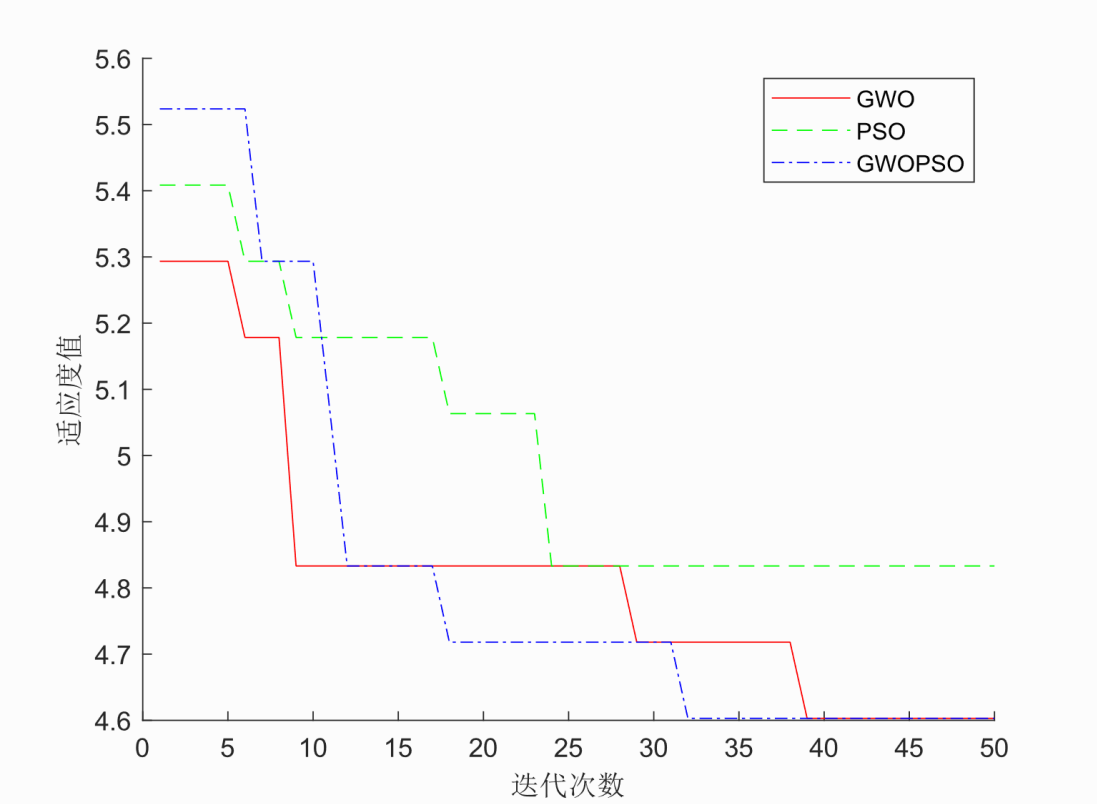

图 4.6 GWO-PSO-SVM 适应度随迭代进行变化图

从图 4.6 可以得到结论：GWO-SVM 算法迭代到 39 次时收敛到最小适应度 4.6；PSO-SVM 算法迭代到 24 次收敛到最小适应度 4.8；GWO-PSO-SVM 算法迭代到 32 次收敛到最小适应度 4.6。可见 PSO-SVM 算法收敛于局部最小值 4.8，而 GWO-SVM 算法和 GWO-PSO-SVM 算法收敛时得到了更佳的适应度 4.6。本章中使用的混合 GWO-PSO-SVM 算法相比 GWO-SVM 提前收敛。

各算法的运行时间和分类准确率汇总如下表所示：

表 4.1　三种优化方法的 SVM 模型结果比较

群智能算法	GWO-SVM	PSO-SVM	GWO-PSO-SVM
预测集准确率	95.4%	95.2%	95.4%
运行时间	70.4 秒	377.5 秒	131 秒

从上表可以看出，GWO-PSO-SVM 混合算法的分类准确率为 95.4%，相比 PSO-SVM 算法的分类准确率 95.2%更高，与 GWO-SVM 算法的分类准确率相同；在运行时间上面看，GWO-PSO-SVM 的运行时间要小于 PSO-SVM 算法。从此可以得出结论，GWO-PSO-SVM 混

合算法的分类准确率高于 PSO-SVM 算法，在收敛时迭代次数要小于 GWO-SVM 算法，GWO-PSO-SVM 混合方法能够获得比 PSO-SVM 方法更低的适应度值。

4.4 本章小结

本章针对机器学习中常见的模型参数调优问题，将群智能算法于经典的支持向量机方法结合起来并给出了一种基于 GWO-PSO-SVM 的混合算法。首先介绍了在数值优化问题中常见的几种优化方法，然后介绍了本章中使用的 GWO 灰狼算法以及 PSO 粒子群算法，针对两种群智能方法的各自优缺点，提出了一种结合两种方法特点的 GWO-PSO 方法；接着将支持向量机模型的预测分类错误率作为群智能算法的适应度值，以 MATLAB R2019b 为实验环境，分别对 AQI 样本数据进行训练和预测，接着通过实验比较三种智能算法的适应度值的变化和测试集分类错误率，结果表明，GWO-PSO-SVM 混合算法的分类准确率高于 PSO-SVM 算法，在收敛时迭代次数要小于 GWO-SVM 算法。

第五章 基于回归方法的 AQI 指数判断

上一章阐述了根据空气质量数据集的污染物浓度特征对当前样本的空气污染等级进行判别的方法。然而在实际应用中，还存在着另一类问题也就是对 AQI 数值进行预测的问题；在分类(Classification Problem)问题中，预测判别的结果是一个离散的数值，也叫做样本标签(Label)，而在本章将要介绍的回归问题中，所预测的数据是连续型数值，这就与前述内容所区分开来。本章将研究对 AQI 指数序列进行预测的方法。

5.1 引言

在对空气质量数据集的处理过程中，除了根据污染物的浓度指标判别当前空气质量以外，依据过去的 AQI 数值来预测当前的 AQI 数值也非常有必要。必须指出在某些应用场景下，预测空气质量指数 AQI 比判别空气污染等级显得更为重要，比如在针对大气环境、工业气体排放污染的科学研究中，人们往往希望得到下一时间周期内的空气质量指数，以此作为依据来求出一些数学指标。这类根据已知数据预测后续数值的问题叫做回归问题(Regression Problem)。本章首先介绍常用的回归算法并使用回归算法在空气质量数据集上进行 AQI 指数的预测的流程，然后给出一种基于时序 EMD-SVR 方法的 AQI 指数预测回归算法，并通过实验验证算法模型的评价指标。

5.2 相关理论介绍

5.2.1 回归问题的评价标准

在回归预测型问题中，对一个模型的预测结果的好坏的评价需要有一定的标准，实际应用中较为常见的评价指标有:均方误差(Mean-square Error, MSE)、均方根误差(Root Mean Squared Error, RMSE)、平均绝对误差(Mean Absolute Deviation, MAE)以及 R 平方值(R-squared)。下面对各评价指标进行介绍。

均方误差又叫平方误差，用每个样本的预测数值和真实数值相减的平方进行累加可以得到。使用均方误差定义模型的损失值较为直观，是回归问题中的常见评价指标。均方误差的计算如式 5.1 所示，其中$\hat{y}_i$表示第i个样本的实际数值，y_i为模型的预测数值。

$$MSE = \frac{1}{\mathrm{m}} \sum_{i=0}^{m-1} (y_i - \hat{y}_i)^2 \tag{5.1}$$

均方根误差即为均方误差开方，表达式为：

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{5.2}$$

平均绝对误差为集合中所有样本的预测值和实际值的差值地绝对值进行线性叠加后除以集合个数。表达式如式 5.3 所示:

$$\mathrm{MAE} = \frac{1}{\mathrm{m}} \sum_{i=0}^{m-1} |y_i - \hat{y}_i| \tag{5.3}$$

R 平方的计算方法为:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=0}^{m-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{m-1}(y_i - \bar{y}_i)^2} \tag{5.4}$$

其中$\bar{\mathrm{y}}_{\mathrm{i}}$表示每个样本的均值，$\mathrm{y}_{\mathrm{i}}$为每个样本的实际值，$m$为样本个数。

上式表明，可知 R 平方取值范围为[0, 1]，当 R 平方值为 1 时，表示每个样本的预测值都和实际值相等，即模型的预测准确率为百分之百；而当数值越接近 1，也说明模型的分类精度越好。当 R 平方取值接近 0 时，预测模型将每个样本的均值当作其预测数值，即没有对数据本身施加任何影响。由上述分析可知，当 R 平方为 1 时表示无差错预测，R 平方为 0 时表示模型为无判别模型，也就是无判别模型是指不考虑样本特征的趋势规律仅仅将样本的均值作为预测回归。可见，使用 R 平方作为模型的指标泛用性较好，在各种回归问题应用场景中都能得到较好的使用，而且 R 平方本身的数值也可以初步反映出模型的优劣。

5.2.2 支持向量回归 SVR 概述

第三章中对空气质量等级判别时使用基于分类的支持向量机 SVM 模型。在回归问题中，支持向量机存在着一种另一种应用方式——支持向量回归机(Support Vector Regression, SVR)。

支持向量回归机是 SVM 理论在回归拟合问题中的一种特殊形式。SVM 在回归问题中的标记 Label 是一系列的离散值，而在回归问题中的结果为连续的预测值。给定样本集$\{(\mathrm{x}^{(1)}, \mathrm{y}^{(1)}), (\mathrm{x}^{(2)}, \mathrm{y}^{(2)}), \dots, (\mathrm{x}^{(\mathrm{m})}, \mathrm{y}^{(\mathrm{m})}) | \mathrm{x}^{\mathrm{i}} \in \mathrm{R}^{\mathrm{n}}, \mathrm{i} = 1,2, \dots, \mathrm{m}\}$，目标为训练出一个模型$\mathrm{h(x)} = \mathrm{w}^{\mathrm{T}} * x + b$，其中$w$是特征的权值，$b$是模型的偏置，使得这个模型能够尽可能准确地拟合出样本特征$\mathrm{x}^{(\mathrm{i})}$的变化规律。在传统的一般数值分析过程中，往往直接通过计算模型的预测数值和实际数值的差值，并将其作为训练损失，当预测值和实际值完全一样时，损失为 0。而在

支持向量回归中，并不追求百分之百的“准确”预测，而是允许预测值和实际值之间存在着一定的误差ε，在误差界限范围之内的预测值视为正常拟合而不计算它的损失，只有预测值和实际值的偏差超过了误差ε时才计算训练损失。

支持向量回归的训练优化可以转化为下列形式：

$$\min \frac{1}{2}|\mathrm{w}|^2 + C\sum_{i=1}^{m} \mathrm{l}_\varepsilon\left(\mathrm{h}\left(\mathrm{x}^{\mathrm{i}}\right) - \mathrm{y}^{\mathrm{i}}\right) \tag{5.5}$$

C 是一个罚项，也叫惩罚系数；l_ε是：

$$\mathrm{l}_\varepsilon(z) = \begin{cases} 0, & |z| < \varepsilon \\ |z| - \varepsilon, & |z| \geq \varepsilon \end{cases} \tag{5.6}$$

引入松弛变量ξ_i，并给出约束条件：

$$\min \frac{1}{2}|\mathrm{w}|^2 + C\sum_{i=1}^{m}\left(\xi_\mathrm{i} + \hat{\xi}_\mathrm{i}\right) \tag{5.7}$$

$$\mathrm{s.t.}\ \mathrm{h}(\mathrm{x}_\mathrm{i}) - \mathrm{y}_\mathrm{i} \leq \varepsilon + \xi_i \tag{5.8}$$

$$\mathrm{y}_\mathrm{i} - \mathrm{h}(\mathrm{x}_\mathrm{i}) \leq \varepsilon + \hat{\xi}_i \tag{5.9}$$

$$\xi_i > 0\ , \hat{\xi}_i > 0, i = 1,2,3 \ldots m \tag{5.10}$$

和支持向量分类问题类似，为了方便求解支持向量回归的最优化问题，一般利用拉格朗日对偶性引进拉格朗日乘子将原问题转换成对偶问题，通过求解对偶问题得到原始问题的最优解。原问题中计算模型参数和特征的内积$\mathrm{w} * \mathrm{x}$，在对偶问题中可以被替换为核函数将特征空间从低维映射到高维空间，进而可以获得良好的非线性分类以及回归性能。

5.2.3 EMD 算法概述

经验模态分解(Empirical Mode Decomposition, EMD)是 Huang 等人于 1998 年提出的一种自适应时域信号分解算法[62-65]，该方法将信号的自身的时域特征进行分解，和频域傅里叶变换以及小波变换等方法有着根本区别。该方法将信号分解为若干有限本征模函数(Intrinsic Mode Function, IMF)，每个 IMF 都表示不同的时间尺度下的局部特征，其最大的特点是可以将非平稳信号进行平稳化，而且不依赖于外部的基函数，具有自适应性。这些特点使得该方法能够用在处理非线性非平稳信号的情景，而且具有较理想的信噪比。目前已经在信号分析等领域得到了非常广泛的应用[66-71]。经过 EMD 方法分解的各分量以及残差表示为下面形式：

$$\mathrm{x(t)} = \sum_{\mathrm{i=1}}^{\mathrm{k}} c_i(t) + r(t) \tag{5.11}$$

上式中$x(t)$为待分解信号，k为分解 IMF 个数，$\mathrm{c_i}(t)$为分解的第 i 阶固有模态分量 IMF,$r(t)$为残差分量，IMF 具有如下特征:

（1）在任何时间点上，由局部极大值和局部极小值定义的上下包络线均值必须为 0，即信号关于时间轴局部对称;

（2）极值点和过零点的数目相等或至多相差 1。

IMF 通过迭代过程产生，每一次迭代的过程也被称为筛分。IMF 算法的每轮筛分都要执行以下的步骤，即在第n轮迭代中：

步骤 1:对于i阶模态分量$\mathrm{c_i^n}(t)$，求出$\mathrm{c_i^n}(t)$的局部极大值和极小值;

步骤 2:用三次样条插值拟合$\mathrm{c_i^n}(t)$的上包络线$\mathrm{e_{up}}(t)$和下包络线$\mathrm{e_{down}}(t)$。

步骤 3:求均值包络线$\mathrm{m(t)} = \frac{[\mathrm{e_{up}(t)+e_{down}(t)}]}{2}$

步骤 4:第 n+l 的 IMF 估计为$\mathrm{c_i^{n+1}}(t) = \mathrm{c_i^n}(t) - m(t)$;

步骤 5:验证$\mathrm{c_i^{n+1}}(t)$是否遵循终止准则，若满足$\mathrm{c_i} = \mathrm{c_i^{n+1}}(t)$，则终止；否则回到步骤 1，以此类推，重复第一步到第四步，直到新得到的序列满足终止原则。这样，原始时间序列就被分解成了多个固有模态分量 IMF 以及一个残差序列，且 IMF 和残差序列满足式 5.11 的约束。

经过上述步骤分解后的固有模态分量 IMF 满足以下两个特征：第一过零点个数与极值点数量之差的绝对值小于等于 1；第二是数据序列中任何一点对应的局部极小值和局部极大值的包络线均值为零，且在时间轴上的局部对称。

5.3 数据集介绍

本章采用的数据集为中国空气质量在线分析平台(https://www.aqistudy.cn/historydata/)上提供的南京市空气质量指数 AQI 的历史记录，与第三章研究空气污染指数 AQI 的分类问题中使用的数据集不同，本章选取的样本数据为南京市 2013 年 12 月一 2021 年 5 月的月均值空气质量指数 AQI 记录，共计 90 组数据，数据的折线图如图 5.1 所示：

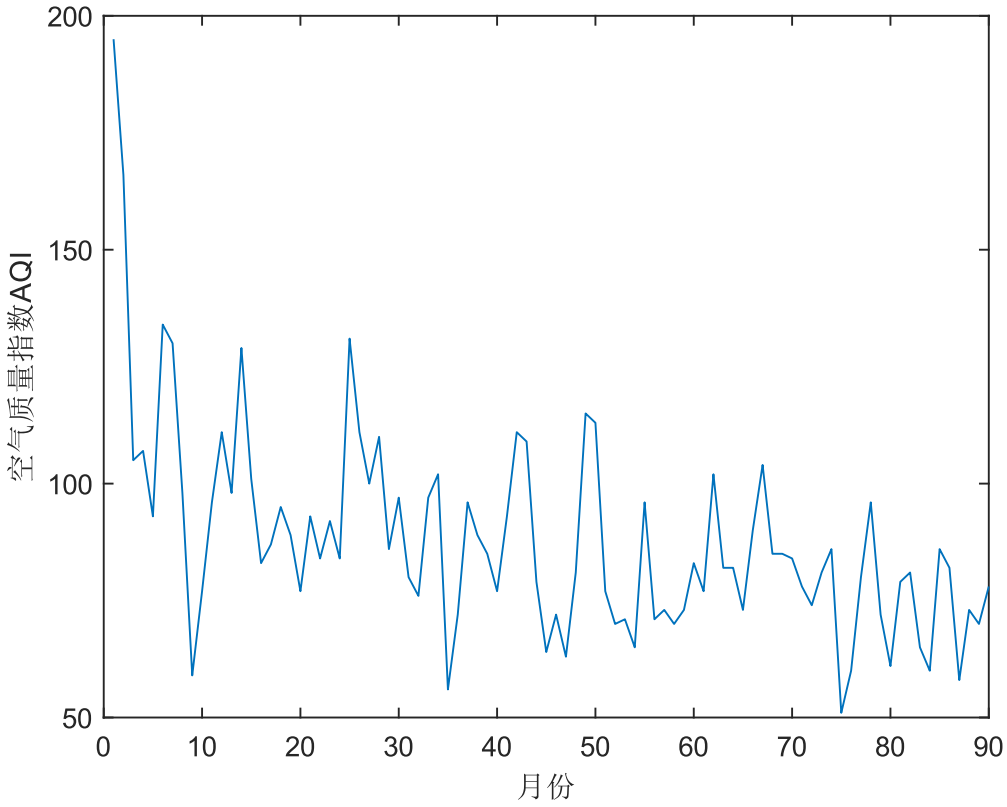

图 5.1 AQI 指数折线图

图 5.1 为南京市月平均空气质量指数变化的折线图，图中可以看出 13 年 12 月的空气污染指数较高，达到了全部数据的峰值，而后续数据呈现出下降趋势，局部上存在着反复，总体上下降。这也从侧面反映出国家近几年来在生态环境治理上获得了显著成果。在图中还可以看出在每年的冬季 12 月至次年的 2 月时空气污染指数呈现出明显的波峰，而到了秋季 9—10 月 AQI 将会有一段明显的下降趋势，这也从侧面表现出空气污染呈现出明显的季节性变化，这也启示人们在进行空气污染治理相关工作的时候，要注意分时用工，在污染相对集中的月份加大治理的投入力度，这样可以便达到事倍功半的效果。

5.4 基于时序 EMD-SVR 的 AQI 指数回归预测算法

EMD-SVR 算法的步骤主要包括三个部分：首先，对时间序列（非线性非平稳序列）实行 EMD 算法，将原序列转化为子序列；接着将分解后的各子序列进行重构，建立起 SVR 时序预测模型，并求出计算预测结果；最后将各自子序列的预测结果叠加后得到原始序列预测结果。

5.4.1 EMD 对 AQI 序列分解

本章选取的样本数据为南京市 2013 年 12 月一 2021 年 5 月的月均值空气质量指数 AQI 记录，共计 90 组数据。下图是对月均值 AQI 数据进行 EMD 分解后的所有子序列。

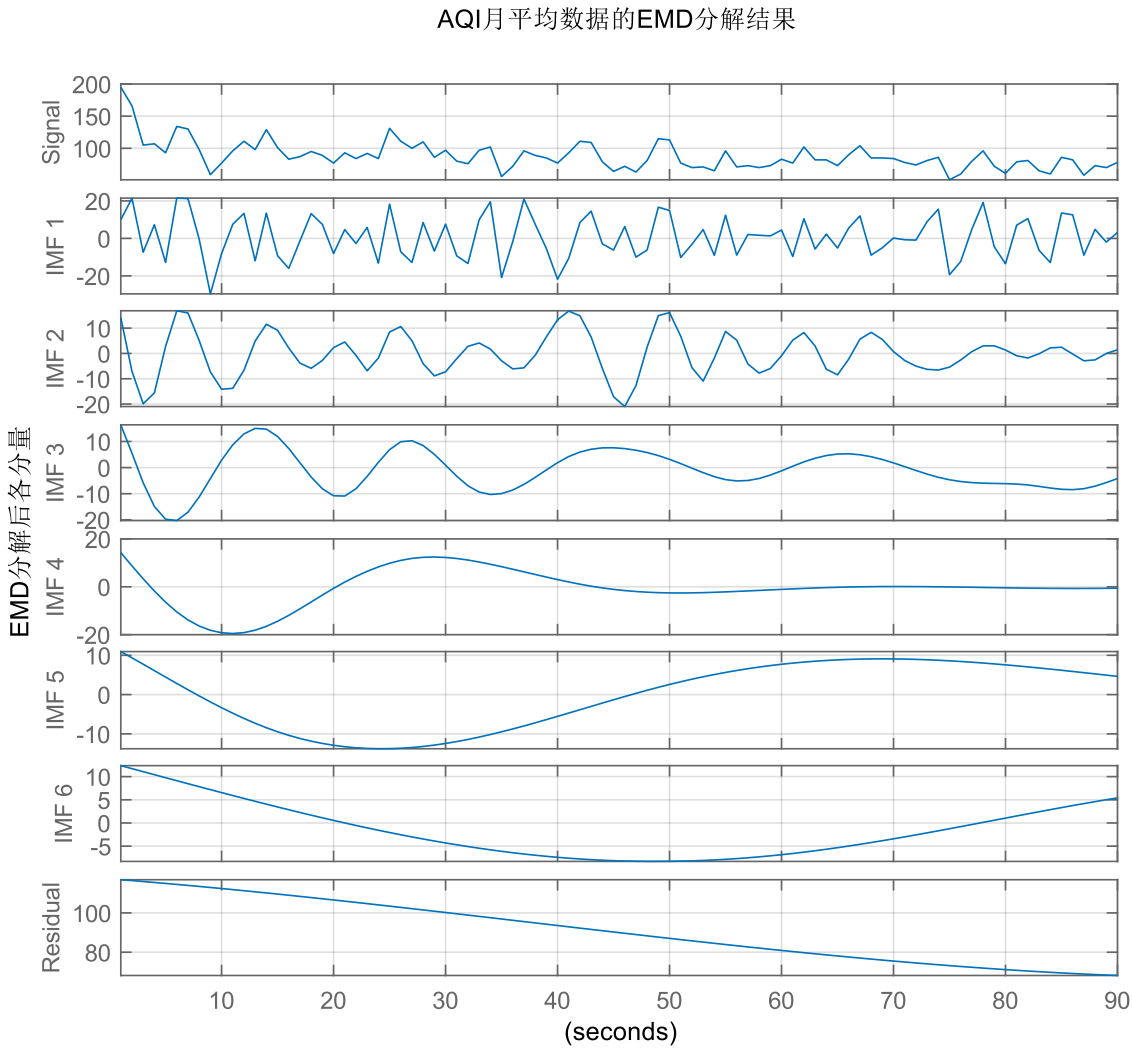

图 5.2 AQI 指数的 EMD 分解图

图 5.2 中 singal 为原始序列，IMF1-IMF6 分别表示 6 阶固有模态分量，Residual 表示残差分量。可以看出，AQI 指数序列可以分解成 6 个子序列以及一个残差序列，并且随着分解的进行，子序列的走势变得更加缓和，即高阶子序列相对于低级子序列的走势更加平缓，保留的线性分量较多。

5.4.2 数据集构造

为了使用 SVR 模型对分解子序列进行回归计算，必须使得时序 AQI 指数数据集满足 SVR

模型的输入格式。为此对原数据样本进行如下处理：

对于时间序列$y_1, y_2, \ldots y_n$，定义输入矩阵：

$$X = \begin{bmatrix} y_1 & \cdots & y_d \\ \vdots & \ddots & \vdots \\ y_{n-d} & \cdots & y_{n-1} \end{bmatrix} \tag{5.12}$$

其中d是步长参数，在本文中取 8。定义输出标签：

$$\mathrm{y} = \begin{bmatrix} y_{d+1} \\ \vdots \\ y_n \end{bmatrix} \tag{5.13}$$

使用上面定义的X和y分别作为 SVR 模型的输入和标签。在实际应用中，按照 3：1 的比例将X和y划分为训练集和测试集。在训练集上对数据集进行训练过程，然后在测试集上进行预测；完成了所有子序列预测后，将所有预测值叠加即可得到原始预测序列。最后通过计算其回归任务的性能指标(比如均方根误差或者平均绝对误差等)来验证算法的性能。

5.5 实验与结果分析

下面对上一小节中介绍的月平均空气质量指数 AQI 数据集进行预测判别任务，使用 EMD-SVR 模型对各子序列的重构数据集进行训练和预测。

5.5.1 各子序列的回归预测结果

首先对 5.3.2 节中介绍的重构数据集按照 3：1 的比例划分训练集和测试集。使用 matlab 中的 EMD 算法对训练集的原始时间序列进行分解，经过分解后得到各子序列，接着使用式 5.12 和式 5.13 对每一个子序列生成重构数据集，然后对重构数据集使用 SVR 方法生成预测序列，各子序列极其预测结果如图 5.3 所示：

图 5.3 是分量 IMF1 子序列的预测结果图，其中洋红色实线是实际 AQI 数值曲线，绿色虚线是 EMD-SVR 算法的预测曲线。可以看出，由于 IMF1 子序列的变化波动较大，导致了算法的预测结果在某些点的偏差较大。

图 5.4 是是分量 IMF2 子序列的预测结果图，其中洋红色实线是实际 AQI 数值曲线，绿色虚线是 EMD-SVR 算法的预测曲线，可以看出，IMF2 子序列的变化相比于 IMF1 要明显缓和，预测序列的变化趋势基本与实际序列一致，而且偏差相对于 IMF1 序列明显较小。

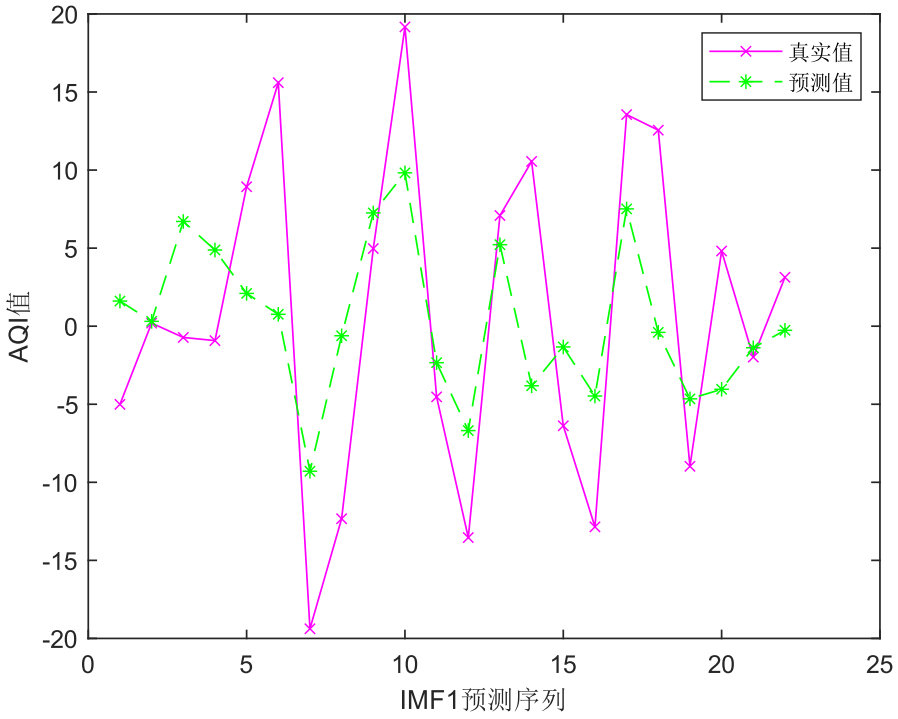

图 5.3 IMF1 的 SVR 预测图

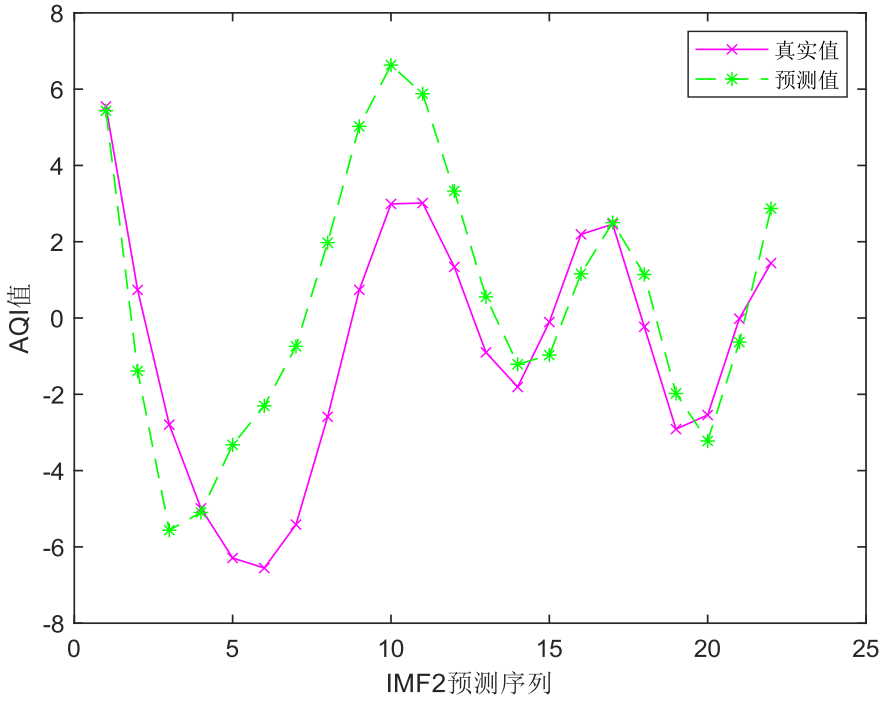

图 5.4 IMF2 的 SVR 预测图

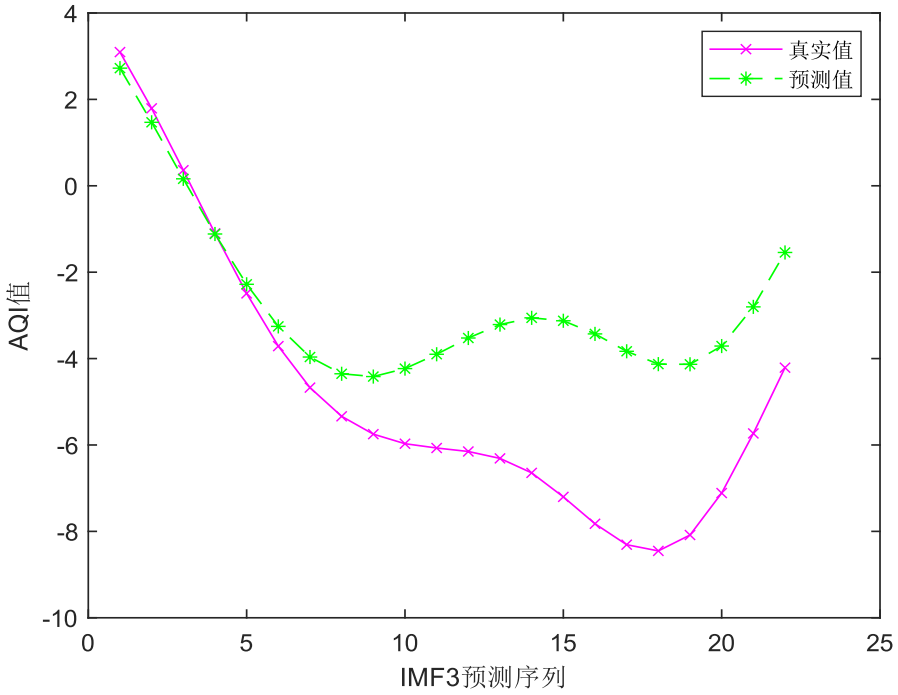

图 5.5　IMF3 的 SVR 预测图

图 5.5 是是分量 IMF3 子序列的预测结果图，其中洋红色实线是实际 AQI 数值曲线，绿色虚线是 EMD-SVR 算法的预测曲线，可以看出，IMF3 子序列的前半段的预测效果很好，但是后半段的偏差较大。预测序列与数据序列的走势在前半段基本一致，但是后半段存在较大的偏差。

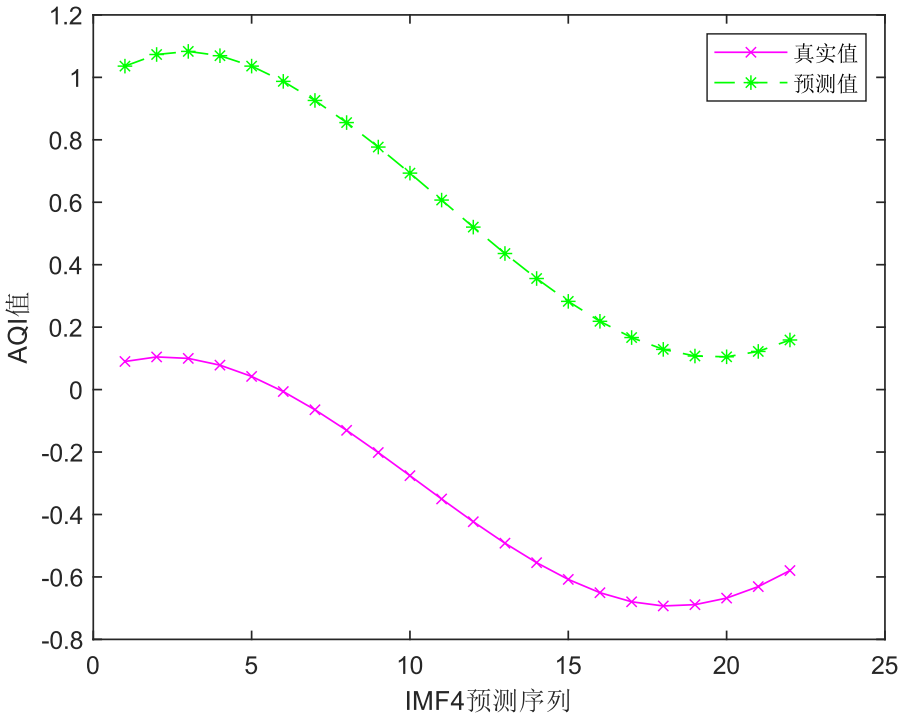

图 5.6　IMF4 的 SVR 预测图

图 5.6 是是分量 IMF4 子序列的预测结果图，其中洋红色实线是实际 AQI 数值曲线，绿色虚线是 EMD-SVR 算法的预测曲线，可以看出，算法的很好预测了子序列的趋势，但是整体上大致偏差了一个常量值。

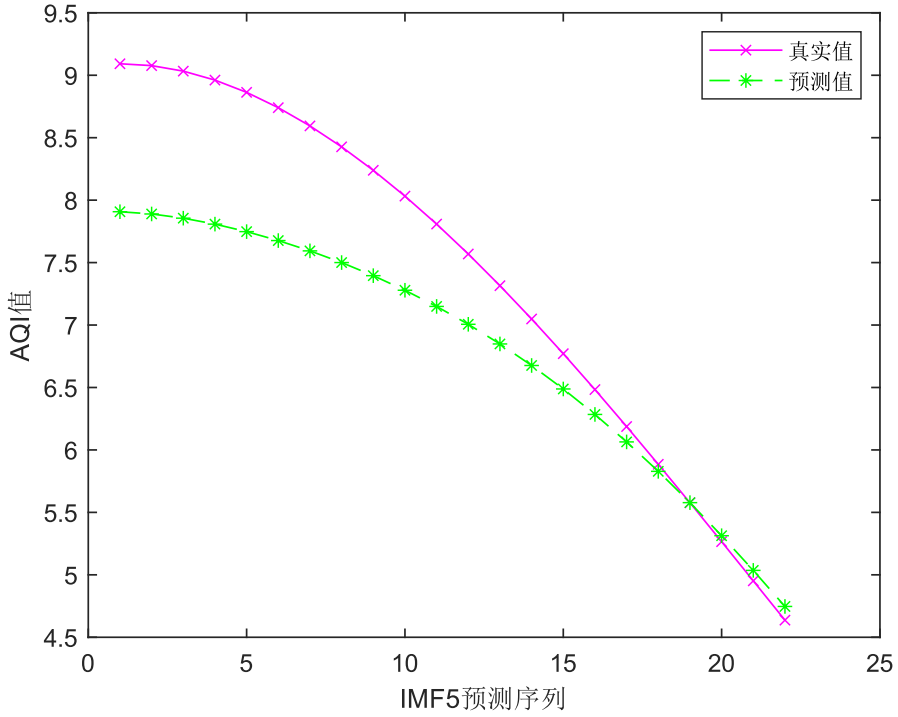

图 5.7　IMF5 的 SVR 预测图

图 5.7 是是分量 IMF5 子序列的预测结果图，其中洋红色实线是实际 AQI 数值曲线，绿色虚线是 EMD-SVR 算法的预测曲线，可以看出，IMF5 子序列的后半段的预测效果较好，但是前半段有一定的偏差。预测值和实际值的整体走势保持一致。

图 5.8 是是分量 IMF6 子序列的预测结果图，其中洋红色实线是实际 AQI 数值曲线，绿色虚线是 EMD-SVR 算法的预测曲线。可以看出，IMF6 子序列的走势预测效果较好，整体上大致存在一个常量偏差。

得到各子序列的预测结果后，将各分量的预测序列线性叠加，就得到整体的最终 AQI 预测序列，预测序列与实际序列的实际效果如图 5.9 所示：

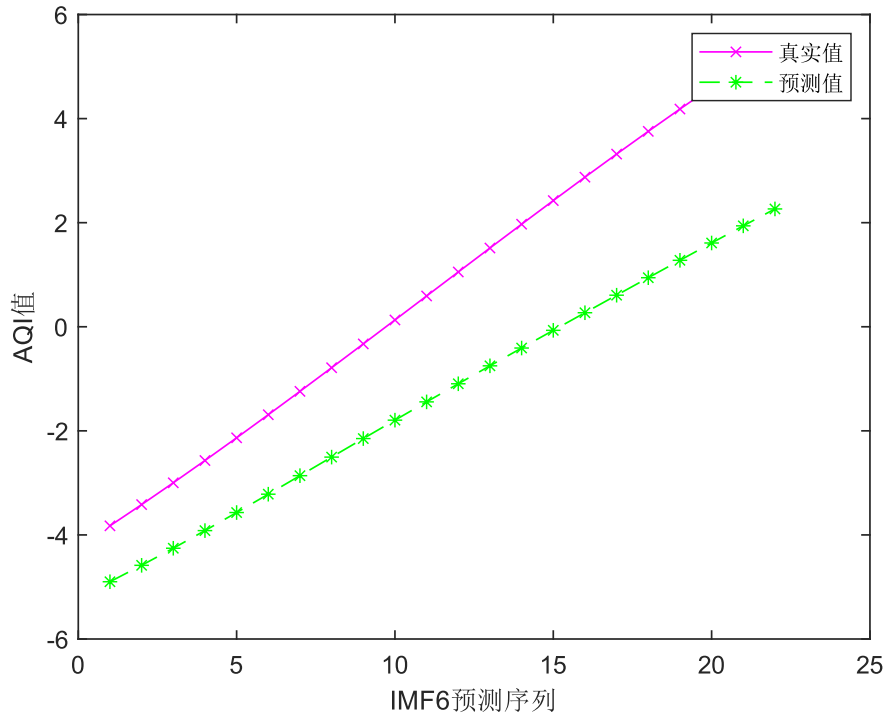

图 5.8　IMF6 的 SVR 预测图

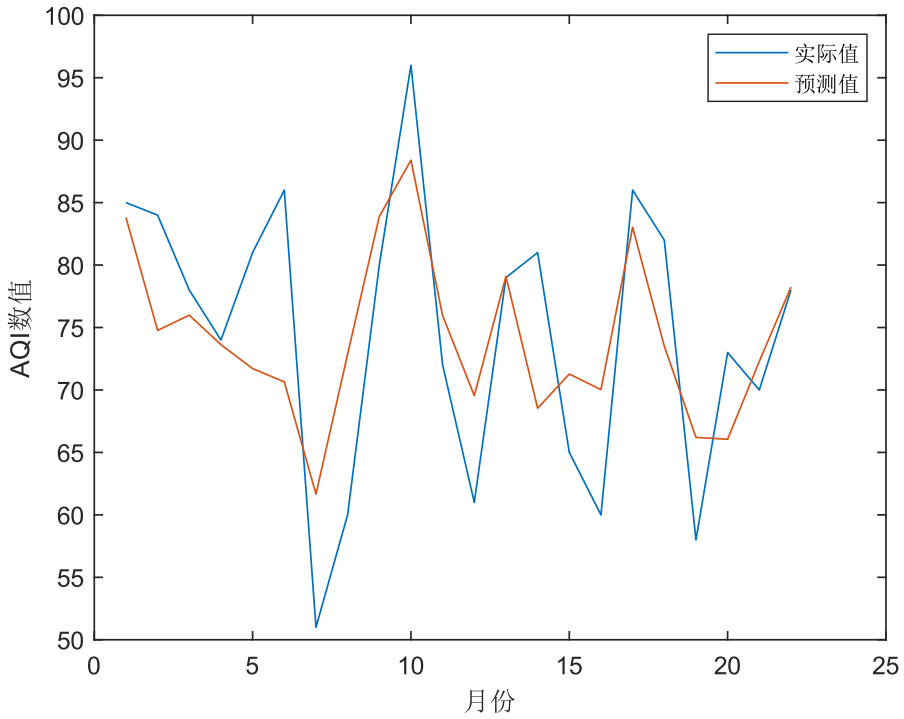

图 5.9　AQI 序列的最终预测结果图

图 5.9 是把各子序列的预测结果相叠加后得到的整体预测结果图。蓝色实线表示实际的 AQI 曲线，橙色实线表示 AQI 预测结果，可以看出，在局部最小和局部最大值出的拟合偏差较大，但是在其他点出的拟合效果较好。为了定量表示时序 EMD-SVR 算法的预测效果，5.5.2

节中将使用基于六项污染物特征的 SVR 预测模型，5.5.3 节将定量计算这两种算法与传统的时序 ARIMA 算法的预测指标。

5.5.2 基于特征 SVR 模型的回归预测结果

本节介绍使用支持向量回归机在六项污染物指标上进行 AQI 预测的结果。在时序 EMD-SVR 模型中，输入是时间序列$x(t)$，即数值是随着时间变化的。这样做的缺点是模型仅能根据历史数据中的 AQI 数据预测未来的数值，但是没有考虑到其他参数特征对结果的影响。相比之下，支持向量回归 SVR 模型就能够充分考虑样本数据的其他潜在特征从而很好地利用这些特征充分训练模型。使用 PM2.5、PM10 等原数据集中的六个污染物浓度作为特征，使用 libsvm 工具箱框架，首先在前 60 组数据集上进行训练，得到 SVR 模型，然后在后 30 组数据测试集上验证回归预测效果，结果如图 5.10 所示：

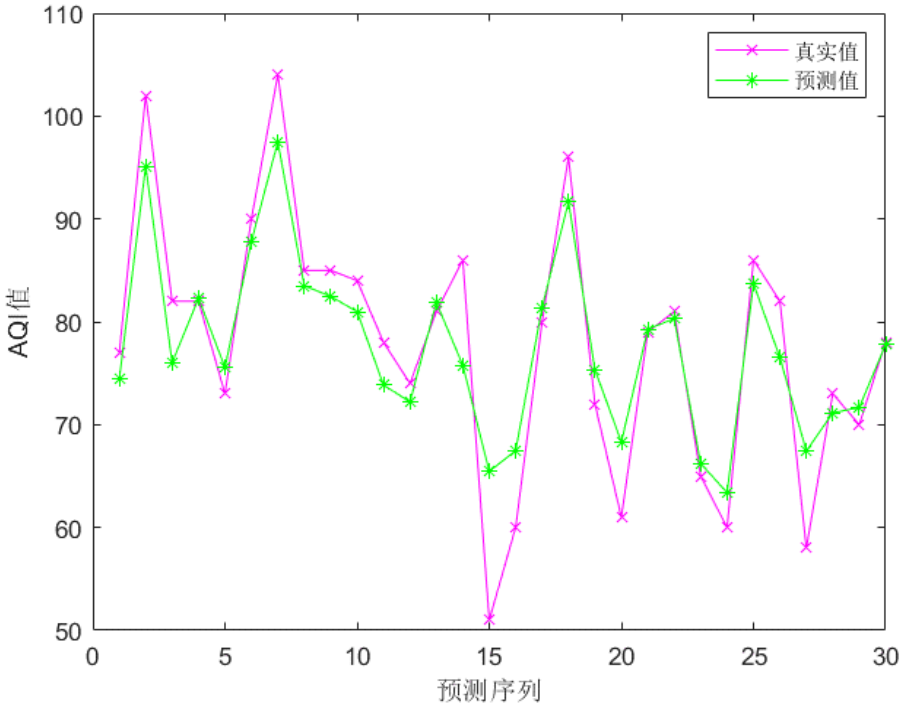

图 5.10 时序 SVR 预测结果图

图 5.10 中洋红色标注的样本是训练集 30 组样本的真实值，浅绿色标注的样本是 SVR 模型的预测结果。可以看出，相对于时序的 EMD 分解算法，基于污染物特征 SVR 模型对于数据的预测偏差较小，并且在各位置都较好地拟合了原数据的变化趋势，但是也能看出在 15 号等样本处的偏差较大。由于支持向量回归机 SVR 是根据 AQI 指数的影响因子也就是污染物的特征进行回归判别的，所以相对于时间序列 EMD-SVR 模型而言，能够接收更多的输入特

征，在拟合效果上要更好，但是由于其必须将污染物特征作为模型输入，相对于基于时序的 EMD-SVR 模型在使用上有着更多的限制。

5.5.3 结果对比

下面将本章所使用算法：基于时序 EMD-SVR、基于特征的 SVR 与时序预测的经典 ARIMA 方法的预测效果进行比较，根据前文介绍的回归模型的评价指标，求出各算法的平均百分误差 MAPE，汇总结果以表格的形式给出结果，如下所示：

表 5.1 三种预测方法结果比较

日期	AQI 值	ARIMA 预测值	基于特征SVR预测值	基于时序 EMD-SVR 预测值
2018.12	77	81.41	74.4	\
2019.1	102	75.94	95.12	\
2019.2	82	98.87	76.02	\
2019.3	82	80.59	82.28	\
2019.4	73	80.6	75.56	\
2019.5	90	72.42	87.81	\
2019.6	104	87.83	97.48	\
2019.7	85	100.77	83.5	\
2019.8	85	83.49	82.47	83.8
2019.9	84	83.52	80.97	74.8
2019.10	78	82.64	73.85	76
2019.11	74	77.21	72.2	73.6
2019.12	81	73.57	81.86	71.7
2020.1	86	79.85	75.77	70.7
2020.2	51	84.38	65.45	61.7
2020.3	60	52.89	67.47	73
2020.4	80	60.72	81.39	84
2020.5	96	78.5	91.65	88.4
2020.6	72	92.97	75.32	75.96
2020.7	61	71.39	68.31	69.5
2020.8	79	61.51	79.28	79.1
2020.9	81	77.47	80.3	68.5
2020.10	65	79.28	66.21	71.3
2020.11	60	64.96	63.39	70.1
2020.12	86	60.43	83.7	83.1
2021.1	82	83.51	76.5	73.5
2021.2	58	79.98	67.43	66.2
2021.3	73	58.59	71.1	66.1
2021.4	70	71.79	71.65	72.3
2021.5	78	69.1	77.78	78.3

表 5.2 三种预测模型的 MAE 和 MAPE

算法	MAE	MAPE
ARIMA	11.74	15.7%
基于特征 SVR	3.86	5.4%
基于时序 EMD-SVR	6.5	9%

由表 5.1 和表 5.2 可以看出，ARIMA 模型对 AQI 指数的变化进行特征规律捕捉以及预测描述的效果最差，从结果上来看，ARIMA 模型进行预测的平均绝对百分误差 MAPE 是 15.7%，基于特征的 SVR 模型对原始 AQI 数据的预测回归能力较强，其模型的平均绝对百分误差是 5.4%；排在中间的是本章中提出的基于时序的 EMD-SVR 算法，其模型的绝对百分误差是 9%；于是得出结论：对于时间序列作为输入的 ARIMA 算法，本章中提出的 EMD-SVR 方法在预测的误差提升较为明显；而相对于污染物特征作为输入的 SVR 方法，本章的 EMD-SVR 方法虽然准确率稍差，但存在着数据的获取流程较为简便的优点。特征 SVR 方法需要获取所有的污染物的六项特征，而时序 EMD-SVR 方法只需要获取 AQI 时间序列就可以进行预测回归操作，能够在污染特征数据集缺失的情况下使用，而且预测误差要明显好于传统的时序预测 ARIMA 方法。

5.6 本章小结

针对大气环境研究领域中常见的 AQI 指数预测问题，本章提出了一种基于时间序列特征的 EMD-SVR 方法。首先介绍了预测回归问题中的评价指标，然后阐述了 EMD 算法的原理，接着给出了 EMD-SVR 方法的序列分解、数据重构、建立模型等流程的具体方法，最后在 AQI 预测数据集上分别使用污染物特征 SVR、时间序列特征 EMD-SVR、时间序列经典 ARIMA 三种法来计算相关评价指标 MAPE 和 MAE，定量分析不同算法对模型的性能进行量化和比较。结果表明，从评价指标上看，时序 EMD-SVR 预测结果要优于传统的时间序列 ARIMA 模型但差于基于特征的 SVR 方法，但是相对于基于特征的 SVR 方法，时序特征方法对于数据集的要求较低，应用场景和领域更为广泛。

第六章 总结与展望

6.1 工作总结

当前全球气候剧变，由此引发的空气污染问题日益严重，人们的日常生活受到了日益严重的影响。随着环保意识的逐渐增强，人们对于空气质量改善的需求越来越高涨，如何做好污染的防治工作成了一个十分急迫的问题。在保持经济高速发展的同时将工业化对环境气候的影响降低到最小，这已经成了各国学术界所共同追求的目标。采取科学的方法进行空气质量监控工作已经成了一项重要议题，准确地从空气质量数据中获取信息是这项议题的关键。掌握污染变化的规律，了解大气污染程度对于环境造成的影响，十分有利于科学地指导空气污染防治工作，对城市健康发展具有十分重要的指导意义。

随着科技的快速发展，新兴的机器学习技术已经融合到社会发展的各个领域，已经在统计样本分类、时间序列预测等方面获得了突破性的进展。本文主要基于机器学习技术，引入了若干典型机器学习算法对空气质量数据分类的判别和预测进行了研究。文章收集了中国空气质量数据监测网上公开的南京市近 6 年来的数据样本，分别进行了空气质量等级分类以及 AQI 污染指数的预测判别两项工作，采用 SoftMax、SVM、GWO、PSO 等机器学习方法以及群智能算法训练构造分类器，通过信号处理领域中的 EMD 算法以及 SVR 算法对 AQI 时间序列进行预测。主要工作如下：

(1) 针对 SoftMax 逻辑回归在多分类情境下对空气质量数据集的分类准确率不高的问题，提出了低置信样本的概念并结合了 SVM 模型对低置信样本进行结合分类的方法。在实验数据上的运行结果表明，新模型的准确率要优于 SoftMax 模型的分类准确率。

(2) 针对 SVM 模型的参数对分类结果的影响问题，引入了两种群智能算法 GWO 和 PSO，将 SVM 模型的参数作为群智能算法的输入，SVM 的分类错误率作为适应度值，经过迭代搜索适应度最小时的参数。在实验数据上的运行结果表明，结合了两种群智能算法的 SVM 模型的精确率都得到了提升，同时 PSO-SVM 的精确率要优于 GWO-SVM 方法，但是前者的算法执行时间更长。结合了 PSO 算法中方向矢量的概念，在 GWO 中引入速度矢量和提前淘汰机制，经过结合的 GWO-PSO-SVM 混合算法能够获得比 PSO-SVM 更小的适应度值以及更高的判别准确率，同时在算法的收敛次数上相比于 GWO-SVM 算法更少。

(3) 针对另外一种对 AQI 指数进行预测的问题，提出了一种基于时序 EMD-SVR 预测算法，对月平均污染数据集的 AQI 指数进行预测，结果表明在月平均数据集上 SVR 模型的预测效果比传统的 ARIMA 模型要好，相比于基于特征的 SVR 算法，时序 EMD-SVR 方法对于数据集的依赖更小。

6.2 研究展望

本文使用了机器学习方法对空气污染指数数据集进行了污染等级判别以及 AQI 指数预测的工作，且对每种算法以及其结合方法进行了结果分析，但是仍然有一些地方存在不足指出值得进一步讨论和优化，具体可以总结为下面几点：

(1) 本文中各算法的优化对象均为六种污染物因子为主要内容的空气质量数据集，所有的研究工作都是围绕该数据集进行展开的，但是众所周知，空气污染的影响因子成分原因十分复杂，除了受到污染物浓度的影响外，还受到气压、风速风向、地形地貌、气候温度、湿度等十分复杂的其他原因影响，而本文并没有考虑到这些因素对结果的影响。

(2) 在第三章中筛选低置信样本中，使用的分类区间间隔为 0.1，后续研究中可以细化阈值区间的选取，在更加小范围的区间中找到对准确率提升最高的阈值；为此，可以引入第四章中所使用的群智能优化算法，将低置信阈值作为优化方法的输入，将训练集出错误率作为适应度值，找到整体到效果最优的阈值，从而使得最终的结果更精确。

(3) 在第四章中提出了基于 GWO 和 PSO 算法的 GWO-PSO-SVM 模型，但是群智能优化算法的发展十分迅速，在后续研究中可以使用更多的分类算法如 BP 神经网络、随机森林法、Adaboost 法等与群智能算法进行结合，通过参数选取得到最佳的分类模型。

(4) 在第五章的时序 EMD-SVR 模型算法中对子序列的分解过程中，使用了重构数据集作为 SVR 的输入，在后期可以对数据集的重构过程进行改进，可以使用卷积序列、小波变换等处理技术对时间序列进行处理。

(5) 本文中选用的机器学习方法较为传统，在后续研究中，可以通过考察调研获取有更多影响因子的数据样本，同时分类算法可以更加多样化，比如增加深度学习中更深层的卷积神经网络等算法以提高分类和预测的精度。

参考文献

[1] 刘建国, 桂华侨, 谢品华. 大气灰霾监测技术研究进展[J]. 大气与环境光学学报, 2015, 10(2): 93-101.

[2] 董战峰, 郝春旭, 李红祥. 2016年全球环境绩效指数报告分析[J]. 环境保护, 2018, 44(020): 52-57.

[3] Zhang Y L. Dynamic effect analysis of meteorological conditions on air pollution：a case study from Beijing[J]. Science of the Total Environment, 2019, 684: 178-185.

[4] Chai F H, Gao J, Chen Z X, et al. Spatial and temporal variation of particulate matter and gaseous pollutants in 26 cities in China[J]. Environmental Sciences, 2014, 26(1): 75–82.

[5] Wu Dan, Yu Yaxin, Xia Junrong, et al. Long-term variation in haze days and related climatic factors in Nanjing[J]. Trans Atmos Sci, 2016, 39(2): 232-242.

[6] Shi Yingying, Zhu Shuhui, Li Li, et al. Historical trends and spatial distributions of major air pollutants in the Yangtze River Delta[J]. Journal of Lanzhou University, 2018, 54(3): 184-191.

[7] Cui Yuhang, Tang Lili, Pan Liangbao, et al. Comparative study on how different meteorological conditions affect atmospheric environment in January in Nanjing[J]. Environmental Science &Technology, 2017, 40(2): 44-52.

[8] Liao Zhiheng, Sun Jiaren, Fan Shaojia, et al. Variation characteristics and influencing factors of air pollution in Pearl River Delta area from 2006 to 2012[J]. China Environmental Science, 2015, 35(2): 329-336.

[9] Zhang Xintong, Xu shan, Jin Huaxing, et al. Analysis of relationship between air quality index and meteorological conditions in Chuzhou City[J]. Anhui Agricultural Science Bulletin, 2017, 23(14): 161-165.

[10] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer. 2001.

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

[12] 李志峰, 张二艳. BP 神经网络在空气质量评价分级中的探索与应用[J]. 电子技术与软件工程, 2014(05): 43-44.

[13] 王琨, 王文帅, 张夏, 孙丽欣. 基于 BP 神经网络模型的森林空气质量评价[J]. 哈尔滨工业大学学报, 2010,42(08): 1278-1281.

[14] 刘杰, 杨鹏, 吕文生, 刘阿古达木. 城市空气质量的 BP 和 RBF 人工神经网络建模及分类评价[J]. 安全与环境工程, 2014, 21(06): 129-134.

[15] 王纪利, 王林山. SOFM 神经网络在城市空气质量信息公开指数评价中的应用[J]. 河北师范大学学报(自然科学版), 2014, 38(05): 456-462.

[16] 芦华, 谢旻, 吴钰, 刘伯骏, 高阳华, 陈贵川, 李振亮. 基于机器学习的成渝地区空气质量数值预报PM_(2.5)订正方法研究[J]. 环境科学学报, 2020, 40(12): 4419-4431.

[17] 陈祖云, 金波, 邬长福. 支持向量机在环境空气质量评价中的应用[J]. 环境科学与技术, 2012, 35(S1): 395-398.

[18] 董胜伟, 苏婷, 黄玲玲, 吕志伟. 空气质量的模糊综合评价与曲线拟合预测模型[J]. 数学的实践与认识, 2018, 48(06): 102-108.

[19] Zhuhong YUAN, Zhenzhen ZHANG, Liu YANG, Jing LI, Degang JI. Fuzzy Comprehensive Evaluation of Air Quality in Shijiazhuang City Based on Grey Correlation[J]. Agricultural Biotechnology, 2020, 9(03): 119-120+123.

[20] Science - Geoscience; Reports on Geoscience from Jilin University Provide New Insights (Fuzzy Comprehensive Evaluation of Debris Flow In Matun Village, Laomao Mountain Area, Dalian City[J]. Science Letter, 2020.

[21] Tzu-Yi Pai, Keisuke Hanaki, Hsin-Hsien Ho, et al. Using grey system theory to evaluate transportation effects on air quality trends in Japan[J]. Transportation Research Part D, 2007, 12(3): 158-166.

[22] 刘朋. 灰色理论在西北地区城市大气污染物浓度预测及质量评价中的应用[D]. 兰州大学, 2010.

[23] 丁鹏. 基于灰色理论的城市空气质量评价及污染物浓度预测研究[D]. 江西理工大学, 2018.

[24] 魏臻, 林芳. 基于聚类分析和主成分分析城市空气质量评价[J]. 淮阴工学院学报, 2018, 27(03): 86-96.

[25] 王恒, 颜斌, 崔鑫, 刘小峰. 基于熵值多级模糊综合评判的空气质量综合评价[J]. 山东科技大学学报(自然科学版), 2016, 35(05): 102-108.

[26] 易丹辉. 时间序列分析:预测与控制[J]. 数理统计与管理, 2000(03): 51.

[27] Vlachogianni A, Kassomenos P, Karppinen A. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki [J]. Science of the Total Environment, 2011, 409(8): 1559-1571.

[28] Jian L, Zhao Y, Zhu Y P. An application of ARIMA model to predict submicron particle concentrations from meteorological factors at a busy roadside in Hangzhou, China [J]. Science of the Total Environment, 2012, 426(1): 336-345.

[29] Slini T, Karatzas K, Moussiopoulos N. Statistical analysis of environmental data as the basis of forecasting: an air quality application[J]. Science of the Total Environment, 2002, 288(3): 227- 237.

[30] Yun zhenXu, PeiDu, JianzhouWang. Research and application of an air quality early warning system based on a modified least squares support vector machine and a cloud model[J]. Environmental Pollution, 2017,

223: 435-448.

[31] 敖希琴, 张怡文, 陈家丽, 费久龙. 基于季节性时间序列模型的合肥地区空气质量分析及预测[J]. 合肥学院学报(综合版), 2018, 35(05): 33-39.

[32] 王坤, 阮金梅, 邓妮. 基于 SARIMA 模型的曲靖市空气质量指数预测[J]. 曲靖师范学院学报, 2018, 37(03): 25-29.

[33] Baptista Ventura Luciana Maria,Pinto Fellipe de Oliveira, Soares Laiza Molezon,et a1.Forecast of daily PM2.5 concentrations applying artificial neural networks and Holt-Winters models[J]. AIR QUALITY ATMOSPHERE AND HEALTH, 2019, 12(03): 317-3 25.

[34] 孟庆云, 张若晴, 袁朱红, 李智坤, 冀德刚. 基于 ARIMA 模型的天津市空气质量各项指标的预测分析[J]. 农业灾害研究, 2018, 8(05): 44-45.

[35] 蔡欣悦, 江建伟, 汪凯. 基于微分方程和时间序列的 PM2.5 预测模型[J]. 辽宁工业大学学报(自然科学版), 2019, 39(04): 270-272.

[36] Lifeng Wu, Xiaohui Gao, Yanli Xiao, Sifeng Liu, Yingjie Yang. Using grey Holt–Winters model to predict the air quality index for cities in China[J]. Natural Hazards, 2017, 88(2).

[37] 李翔. 基于 GAB 和模糊 BP 神经网络的空气质量预测[J]. 华中科技大学学报, 2013,41（z1）：63-65.

[38] 尹琪, 胡红萍, 白艳萍. 基于 GA-SVM 的太原市空气质量指数预测[J]. 数学的实践与认识, 2017, 47（12）：113-120.

[39] Voukantsis D, Karatzas K, Kukkonen J. Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki [J]. Science of the Total Environment, 2011, 409(7): 1266-1276.

[40] 赵琦琳, 邱飞, 杨健. NARX 神经网络模型在昆明市环境空气质量预测中的应用[J]. 中国环境监测, 2019, 35(03): 42-48.

[41] 高鹏, 周松林. 基于小波 Mallat 算法和 BP 神经网络的空气质量指数预测的研究[J]. 池州学院学报, 2017, 31(03): 42-44.

[42] Viotti P, Liuti G, di Genova P. Atmospheric urban pollution: applications of an artificial neural network (ANN) to the city of Perugia[J]. Ecological Modelling, 2002, 148 (1): 27-46.

[43] 周秀杰, 苏小红, 袁美英. 基于 BP 网络的空气污染指数预报研究 [J]. 哈尔滨工业大学学报, 2004, 36 (5): 582-585.

[44] 俞卫忠, 陈建. BP 人工神经网络模型在城市空气污染预报中的应用 [J]. 污染防治技术, 2013, 26 (3): 55-57.

[45] 王国胜，郭联金，董晓清．深圳市区空气污染的人工神经网络预测[J]．环境工程学报，2015，9（7）：3393-3399.

[46] 尹文君，张大伟，严京海，张超，李云婷，芮晓光．基于深度学习的大数据空气污染预报[J]．中国环境管理, 2015, 7(06): 46-52.

[47] 江晓晴，鲁明浩，鲁园园，王欢，焦毛毛，刘鹏妮．基于灰色模型的大气环境质量分析与评价[J]．环境与可持续发展, 2015, 40(04): 114-115.

[48] 司志娟．基于灰色神经网络组合模型的空气质量预测[D].天津大学, 2012.

[49] 司志娟，孙宝盛，李小芳．基于改进型灰色神经网络组合模型的空气质量预测[J]．环境工程学报, 2013, 7(09): 3543-3547.

[50] 韩晓光，李博宇，管智赟．基于灰色关联分析指标筛选的 RBF 神经网络-马尔可夫链的空气质量预测模型[J]．南开大学学报(自然科学版), 2013, 46(02): 22-27.

[51] 赵美玲，薛锐．灰色模型理论在环境空气质量趋势分析中的应用[J]．北方环境, 2013, 25(02): 76-79.

[52] Ian Goodfellow, Yoshua Bengio, Aaron Courville, et al. Deep learning[M]. MIT Press, 2016.

[53] Tom M. Michelle. Machine learning[M]. McGraw-Hill Companies, Inc. 1997.

[54] Bishop M. Pattern recognition and machine learning[M]. Springer, 2006.

[55] Cortes C, Vapnik V. Support-vector networks. Machine Learning[M], 1995, 20(3): 273-297.

[56] Vapnik Vladimir N. The nature of statistical learning theory[M]. Berlin: Springer-Verlag, 1995.

[57] Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers[J]. In: Haussler D, ed. Proc of the 5th Annual ACM Workshop on COLT. Pittsburgh, PA, 1992, 144-152.

[58] S. Mirjalili, S. M. Mirjalili, A. Lewis, Grey Wolf Optimizer[J], Advances in Engineering Software, vol. 69, pp. 46-61, 2014.

[59] 张晓凤，王秀英．灰狼优化算法研究综述[J]．计算机科学, 2019, 46(03): 30-38.

[60] 唐俊. PSO 算法原理及应用[Jl．计算机技术与发展, 2010, 20, (2):213-216.

[61] 尹徐珊．基于改进 PSO 算法的投资组合优化方法的设计和实现[D]．东南大学, 2015.

[62] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, 1998, 454(1971): 903-995.

[63] 郭喜平，王立东．经验模态分解(EMD)新算法及应用[J]．噪声与振动控制, 2008(05): 70-72.

[64] 岳相臣．经验模态分解算法应用研究[D]．西安电子科技大学, 2013.

[65] 孔国杰，张培林，徐龙堂，吴烽．基于经验模态分解的自适应滤波算法及其应用[J]．信号处理，2009,

25(06): 958-962.

[66] 刘慧婷, 张旻, 程家兴. 基于多项式拟合算法的 EMD 端点问题的处理[J]. 计算机工程与应用, 2004(16): 84-86+100.

[67] 邓拥军, 王伟, 钱成春, 王忠, 戴德君. EMD 方法及 Hilbert 变换中边界问题的处理[J]. 科学通报, 2001(03): 257-263.

[68] 杨楚琪. 基于经验模式分解的非线性自适应信号下采样理论研究[D]. 广东工业大学,2019.

[69] 南政年. 基于经验模态分解的样本扩容新方法及其在水文气象领域的应用[D]. 长安大学, 2019.

[70] 宋轲. 基于集合经验模态分解和 D-S 证据理论的异质数据融合[D]. 上海交通大学, 2018.

[71] 李国汉, 王可人, 张颂. 一种基于经验模态分解的信噪比盲估计新算法[J]. 电讯技术,2012,52(05): 663-667.

附录 1 攻读硕士学位期间撰写的论文

[1] 林启明，支持向量机混合模型在窄带物联网领域的应用研究[J]，无线互联科技，已发表。

致谢

时光飞逝，三年的硕士生涯一转眼就快结束了，三年前刚来到学校的情景我还历历在目，三年来的学习让我受益匪浅，实验室积极向上的风气、拼搏努力的氛围对我的影响是极其深远的，在本文即将完成之际，谨向我的导师、同学、实验室的全体同学们表示由衷的感谢。

本文是在我的导师叶全意副教授的指导下完成的，老师那严谨的做事风格和一丝不苟的工作态度极大地感染了我，对我在三年硕士期间产生了积极向上的影响，相信也会在今后的工作生活中继续对我产生警醒，并使我终身受益，再次向我导师表示最真诚的敬意。

我也要感谢我的师兄，他对我的工作学习生活的并帮助很大，他对学术研究的敏锐思维对开拓我的科研思路也产生了极大的助力，同时也让我收获了友谊，我相信在今后的工作生活中我也会继续这份友谊。

我还要感谢我的家人，没有他们的支持，我是无法完成科研任务的，每当在我遇到困难挫折的时候，总是家人给我温暖的关怀让我获得重新前进的动力。无论我面对什么样的艰难，他们都无条件地支持我，这份恩情我将用余生来报答。

最后，我要感谢评审老师以及答辩组的全体老师们，你们能在百忙之中查阅我的论文并且给出指导和建议，我十分感恩。