重庆邮电大学

硕士学位论文

姓名：刘 文

导师：王国胤

专业：计算机应用技术

二零一四年五月

分类号 TP311 密级 公开

重庆邮电大学硕士学位论文

论文题目 基于支持向量机的水质动态预测方法研究

英文题目 Water Quality Dynamic Prediction Based on Support Vector Machine

硕士研究生 刘 文

指导教师 王国胤 教授

学科专业 计算机应用技术

论文提交日期 2014 年 4 月 论文答辩日期 2014 年 5 月 25 日

论文评阅人 匿 名

答辩委员会主席 房 斌 教授

年 月 日

独 创 性 声 明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。据我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 重庆邮电大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名：刘文　　签字日期：2014 年 5 月 23 日

学位论文版权使用授权书

本学位论文作者完全了解 重庆邮电大学 有关保留、使用学位论文的规定，有权保留并向国家有关部门或机构送交论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 重庆邮电大学 可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名：刘文　　导师签名：

签字日期：2014 年 5 月 23 日　　签字日期：2014 年 5 月 23 日

摘 要

随着水环境科学和其他相关科学不断发展，水质时间序列预测逐渐受到重视。水环境中水质数据具有趋势性的特性，对其预测和模拟，能够为决策者综合规划及管理水环境提供科学依据，也是防治和治理水污染不可缺少的一项基础性工作。

大批学者对水质预测的方法进行了研究，其中许多采用支持向量机进行预测建模的方法，还有一些地方可以改进。本文从解决基于支持向量机进行建模存在的问题出发，主要对支持向量机的参数选择和提高模型预测精度两方面进行研究，以此希望为水质预测的研究者和工程领域的应用提供方法。主要的研究内容包括以下两个方面：

（1）研究并实现了一种基于改进小波变换及支持向量机的水质预测方法

目前，采用传统的单因素参数寻优的方法优化支持向量机的参数，不仅非常耗时，而且不能保证找到的参数最优，本文采用遗传算法优化支持向量机的参数，比单因素寻优的方法更准确、高效；虽然采用小波变换将水质时间序列进行分解，能够降低突发跳变等不规律变化的数据对预测模型精度的影响，但由于分解后的细节序列存在多位小数，数据之间存在数量级的差别，容易影响模型的拟合效果，所以本文进一步利用水质数据时间序列的特性，将小波分解后的细节序列进行平移，降低数据之间的数量级差，提高支持向量机的预测精度。

（2）研究并实现了一种基于灰色理论及支持向量机的水质预测方法

采用小波变换数据处理方法，将水质时间序列变换到多个尺度上，有效地减小了瞬变数据对建模的影响，但是原始序列经小波变换被分解为平稳的尺度子序列和不断变化的细节子序列后，将这些具有不同特性的子序列都采用支持向量机建模的方法预测，大量输入的训练数据会影响模型的预测精度。为进一步提高模型的预测精度，将模型的特性和数据的特性充分结合起来，本文采用灰色理论和支持向量机分别对小波变换分解后的平稳尺度子序列和不断变化的细节子序列建立预测模型，提高模型的预测精度。

关键词：水质预测，支持向量机，遗传算法，数据平移，灰色理论

Abstract

With the rapid development of environmental science and other related sciences, people are increasingly concerned about water quality prediction. It is usually used to master the changes in water quality trends and provides important scientific basis and technical supports for early warning in accidental water pollution.

Currently, a large number of scholars are studying the theoretical prediction of water quality, so the methods of water quality prediction are more. Although many prediction methods based on support vector machine are used to modeling, there are some areas for improvement. The thesis solves the existing problems of them, they are the parameters selecting of support vector machine, improving the prediction accuracies of models. The improvements can give helps to researchers and engineering applications, the main contents include the following two aspects.

Firstly, realized a water quality prediction method based on improved wavelet transformation and support vector machine. The traditional parameter optimization method based on signal factor experiments is not only time-consuming but also can not ensure the most optimal parameters. The thesis uses genetic algorithm to optimize the parameters of support vector machine, which are more accurate and efficient. Although using the method of wavelet transformation, effectively reducing the influence of transient variation on prediction model, the decomposed data mainly are decimals and have different data magnitude. In order to reduce their influence on the accuracies of models, these sequences are unified transformed to the positive zone.

Secondly, realized a water quality prediction method based on grey theory and support vector machine. Although water quality time series are decomposed by wavelet transformation to a smooth scale sequence and changing details subsequences, effectively reducing the influence of transient variation on prediction model, the model of support vector machine predict the sequences of different characteristics, a large number of input training data will affect the prediction accuracies of the model. In this thesis, using grey theory and support vector machine respectively to predict smooth scale subsequences and changing details subsequence.

Key words: Water Quality Prediction, Support Vector Machine, Genetic Algorithm, Data Translation, Grey Theory

目 录

第一章 绪论

1.1 课题研究背景及意义

当代社会，在人口不断增长和社会经济快速发展的同时，生活污水、工业废（退）水和各种固体、气体污染物的排放，导致了日益严峻的水污染问题。水环境污染问题目前已经成为了一个重要的世界性问题。随着逐步加大水环境管理的工作力度，我国已经在水环境保护方面取得了长足的进步，而且水环境的质量整体趋于稳定。但是根据《2012 年中国环境状况公报》中淡水环境统计的结果显示，全国地表水国控断面总体状况表现为轻度污染，形势依然严峻。长江、黄河等十大流域的国控断面中，III 类以下水质断面的比例达到了 31.1%；62 个国控重点湖泊（水库）中，III 类以下水质的湖泊（水库）比例高达 38.7%；我国 198 个地市级行政区的地下水水质监测中，较差级、极差级的所占比例分别为 40.5%和 16.8%[1]。因此我国水环境的安全仍面临着诸多的威胁，各地区存在着不同程度的水环境污染情况，甚至很多城市污染物的排放总量已经远远超过水环境的整体承载容量。由于河流和地下水流动性的原因，有从点源污染演变为带状污染或者面状污染的倾向。

全国河流、湖泊（水库）、地下水等水体的水质状况变化过程不但受降雨量、径流、温度、气候等自然因素的影响，而且受人类活动诸如工业和农业生产、经济发展策略、环保部门监管力度以及生产力水平等众多因素的影响，因而水质变化拥有一定的复杂特征，如呈现趋势性、阶段性和不断变化等特征，是一个复杂非线性的演变过程[2]。为了保障我国水资源利用的可持续发展战略，对水资源的合理利用和对水污染的防治和治理的工作已经刻不容缓。由于人和自然持续影响着环境中水质时间序列的变化，通过水质时间序列的预测，能够为保护水环境和治理水污染提供重要的决策依据。水质数据的推演预测是进行水资源管理、污染防治的重要工作，是进行水质综合污染防治和治理的一种重要手段，也是为顺利实现水环境可持续战略等任务不可缺少的基础工作。水质数据的分析和预测可以准确掌握河流、湖泊及其他水体的现状和将来的发展状况，从而及时发现水质急剧变化的成因，并启动相应的应急措施，能够为水质的预测和预报提供基础数据，为水环境污染防治提供科学依据，因此也引起了越来越多的关注。

本文在国家水体污染控制与治理科技重大专项课题《三峡库区水生态环境感知系统及平台业务化运行》中的子课题《三峡库区水生态环境感知推演技术研究与平台示范运行》即将实施的背景下，将支持向量机模型引入到水质时间序列预

测分析中。支持向量机是一种基于统计学习理论的模型，是近年来研究复杂非线性问题的一种新算法，也是机器学习领域的重要方法，具有突出的非线性回归性能。将它与其他水质预测方法及技术进行研究和改进，进行水质预测推演的模型研究，应用到乌梁素海及浙江嘉兴断面，验证模型的可行性，为感知推演平台集成预测部分进行技术积累。

1.2 国内外研究现状

1.2.1 水质预测的研究现状

水环境中水质数据的变化具有复杂性的特点，对其预测和模拟，是为决策者提供科学依据进行水环境管理、水污染的防治和治理的重要工作。无论是在工程项目进行开发建设阶段，环保部门制定区域的水污染防治方案，政府规范地区工厂、企事业水污染物的排放量和递减率，还是进行生态水系规划保护，解决一些突发污染事件造成的污染问题过程中，都面临水质数据在不同时期，存在阶段性、实时性、突发性等变化问题，因此水质数据的预测工作显得尤为重要。

国内外众多水质预测方法已经被应用到实际的工程项目中。我国水质预测方面的研究工作，最早在八十年代初杨天行、王秉忱等人的《湘江重金属水环境容量研究》中提及[3]。随着水环境科学和其他相关科学不断发展，水质时间序列预测逐渐受到重视，由于其具有动态非线性的时间特性，准确实现非线性时间序列的模拟非常困难，于是更加准确的预测模型不断涌现。水质时间序列受到自然、人类活动等多种因素的影响，其复杂的变化特性和数据信息不全等因素影响了水质预测在中长期尺度上的预测精度，这也给水质预测提出了挑战，同样降低了人类掌控水质变化趋势的能力，为决策水资源的合理利用及规划、水污染的防治和治理带来了困难。目前，根据预测模型的理论基础，大致可以分成数理统计模型、灰色理论模型、神经网络模型、水质模拟模型和混沌理论模型等预测方法[4]。

（1）数理统计模型预测法

数理统计模型方法包括使用单因素建立模型预测和使用多因素建立模型预测两类。单因素预测是用现有的历史数据来预测未来水质变化情况，主要方法有指数平滑及时间序列分析等方法[5][6]；多因素预测是利用影响水质状况的众多影响因子，建立水质状况和影响因子的数学关系，从而间接的反映水质的变化情况，主要有多元线性回归[7][8]等方法。数理统计预测方法，涉及重视历史数据的拟合，需要大量的训练数据等原因，存在对非线性的多因素水质预测难度大的特点。

（2）灰色理论模型预测法

在众多水质预测的方法中，采用灰色理论模型进行水质预测，不需要大量的

水质数据信息，具有能够在缺乏水质数据的情况下进行预测的优势，能够利用少量的水质观测数据建立预测数据和历史水文数据之间的近似函数关系[9][10]。灰色理论系统，可以应用到那些有部分已知的历史水文数据（如水位、各水质数据）和部分未知的水文信息（如未来的水位及水质数据）的水环境系统中，建立一个灰色预测模型。灰色理论模型，是以灰色理论原理中的微分方程为描述方式，对事物的发展过程进行推演，适用于水体参数的变化规律。灰色理论模型，由于理论基础的原因限制了其对大波动水质序列的预测效果，适用于原始平滑的水质数据经过累加呈现指数规律的预测[11]。

（3）神经网络模型预测法

神经网络（Neural Network）模仿了现实世界中人类大脑神经系统的运作过程，具有非线性的数学特性，能够良好的适应大规模的数据环境，对复杂非线性问题也有良好的学习能力。目前应用较多的反向传播神经网络模型是 BP（Back Propagation）网络，它具有神经网络的强大学习能力，能够建立输入、输出数据的非线性映射关系。通过向 BP 神经网络中输入水质训练数据，就能够获得水质数据间的权值关系，训练得到网络模型的权值和阈值可以应用于未来水质数据的预测，可以改变输入节点数、神经元数目、输出节点数等条件训练出不同的预测模型，三层 BP 网络见图 1.1。虽然神经网络预测方法能够刻画出水质时间序列复杂变化的非线性特征，但其本身存在网络收敛较慢、寻找到局部极小值等原因，使得模型的推广能力不强。

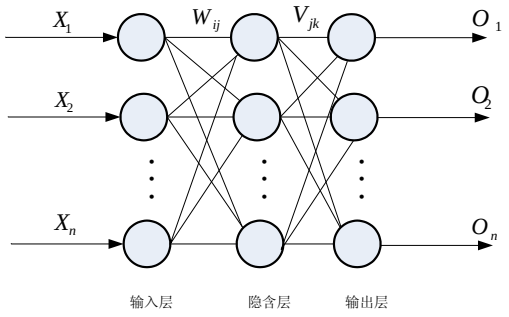

图 1.1 BP 神经网络拓扑

（4）水质模拟模型预测法

20 世纪 20 年代中叶，Streeter-Phelps 提出了一个水质模型，经过多年来的发展，水质模型的研究已经从上世纪 60 年代之前确定模型的角度，转移到水环境的不确定性模型上来了，已经在污染物排放、扩散方面取得了丰硕的成果，建立起来了各种情况下的随机水质预测模型[12]。虽然水质模拟模型可以用来预测水质数据的变化，但是与数理统计预测方法和神经网络预测方法不同，其只适用于较短时间的水质数据预测。

（5）混沌理论模型预测法

混沌理论是从上世纪 80 年代发展起来的，它是研究确定性系统中一些非线性动态问题。混沌理论能够将影响水质状态的物理、化学、环境科学、气象科学、甚至人类活动等多种因素结合起来，综合考虑水质的在时间和空间上变化的复杂关系，从整体上把握了水质变化的复杂性，把复杂的多因素关系转化为简单的单因素关系，是现有基于数学表达式预测方法难以做到的。混沌理论在水质预测领域不断的发展，并且发展起来了多种新的方法[13]，混沌理论预测模型的应用需要丰富的水质资料信息才能得以实现。

1.2.2 支持向量机研究现状

统计学习理论所采用的推理规则考虑了渐进性的要求，而且追求在有限的信息条件下得到最优的结果，是一种专门研究有限样本情况下机器学习规律的理论。最早从 20 世纪 60、70 年代，Vapnik、Lemer 和 Chervonenkis 等人便开始从事统计学习理论的研究。在 70 年代，Vapnik 和 Chervonenkis 在文章《The Necessary and Sufficient Conditions for the Uniforms Convergence of Averages to Expected Values》中，提出了统计学习理论的核心概念——VC 维[14]。在 80 年代 Vapnik 在原来的工作上，进一步提出了统计学习理论历史上具有极为重要意义的原理—风险最小化原理[15]，为今后支持向量机的研究和发展，奠定了坚实的基础。由于早期统计学习理论的研究都停留在经验风险最小化原理的抽象理论当中，学者所提出的 VC 理论并未得到大家的关注。一直到 90 年代中叶，Vapnik 在新的著作中较为完整的提出了支持向量机算法[16]，使得统计学习理论从抽象的理论变为实际的分析回归算法，扩充了统计学习理论的内容。90 年代末，支持向量机在回归及预测方面得到了详细的解释，Vapnik、Gokowich 等人提出了支持向量回归的概念[17]。

支持向量机继承了统计学习和 VC 维理论的精髓，并在结构风险最小化理论上寻找最优解，聚集了最优解、核函数多项选择和具有良好推广能力的特征[18]：

（1）支持向量机无需大量的信息，其追求在有限的信息条件下的最优解，能够解决有限样本情况下的机器学习问题。

（2）支持向量机可以通过二次型优化的方法，避免遇到像神经网络方法收敛于局部极值的情况，能够得到全局的最优解。

（3）支持向量机通过应用 VC 维理论，将复杂的非线性问题转化为高维空间中的线性问题，巧妙地通过核函数简化高维空间中遇到的运算问题，寻求高维特征空间中最优超平面，来解决原始空间中实际非线性问题的回归问题。

（4）支持向量机建立在结构风险的基础上，在经验风险最小化和置信风险两方面做了平衡，具有很好的鲁棒性，推广能力较好。

目前，支持向量机的研究广泛地在人工智能领域、模式识别等领域展开。语

言识别[19]、图像识别[20]、故障检测[21]、文本分类[22]、遥感图像的分析[23]、水质预测等研究方向，通过使用支持向量机都取得了不错的研究成果。支持向量机方法的精度已经可以和传统的方法相媲美，甚至有些情况下远远优于。在工业生产中，支持向量机也得到了很好的应用，如支持向量机被用来进行系统辨识[24]，进行线性或者非线性的识别。支持向量机还可以结合其他的一些方法，如支持向量机与粗糙集理论结合的方法，通过优化支持向量机的预测参数，得到一种新的研究方法[25]；将红外光声光谱技术与支持向量机结合起来，用来鉴别油菜籽品种[26]等。

1.3 支持向量机在水质预测中的应用

水环境系统受到自然因素及人类活动的影响，是一个复杂的非线性系统，并且在时间和空间地域上显示出动态变化、不平衡的特性。物理、化学、环境、气象科学、以及人类活动等这些因素构成了影响水质变化的复杂体系，水质变化便拥有趋势性、阶段性和不断变化等复杂特征，所以水质的预测变得非常困难。

目前，常用的数理统计模型、灰色理论模型、神经网络模型等预测方法应用较为广泛。虽然这些方法在水质预测研究中已经取得了一些重要的成果，但是其自身仍然不够完善。正如 1.2.1 节所描述：数理统计预测方法，涉及重视历史数据的拟合，需要大量的训练数据等原因，存在对非线性的多因素水质预测难度大的特点；灰色理论模型，由于理论基础的原因限制了其对大波动水质数据的预测效果，适用于平滑的水质数据预测；虽然神经网络预测方法能够刻画出水质时间序列复杂变化的非线性特征，对水质预测这项复杂的问题具有良好的适用性，但其本身存在网络收敛较慢、网络结构不确定、寻找到局部极小值等问题，并且神经网络基于取得经验风险最小的原理，在学习的过程中容易陷入过拟合的状况，使得模型的泛化能力不强；水质模拟模型只适用于较短时间的水质数据预测；混沌理论预测模型的应用需要丰富的水质资料信息才能得以实现。因此，寻求更为可靠、高效的水质预测模型，便显得尤为重要。

支持向量机是 Vapnik 等人自上世纪以来，在数理统计和 VC 维理论基础上研究发展起来的，是以结构风险最小为目的一种新的机器学习算法。支持向量机与以往基于经验风险最小化的神经网络不同，其在结构最小化原理的基础上，求解一个二次规划问题的最优解。支持向量机在解决小样本的非线性问题上具有突出的优势，其自身由于解决了模型的过学习、低维空间的非线性问题转化为高维空间的线性问题和高维空间的复杂计算问题，在回归拟合方面表现出极强的优势。

支持向量机在水质预测领域已经有众多的应用。一方面，基于标准的支持向量机方法，郑一华将支持向量机回归理论应用到济南地下水水质数据的预测中[27]；梁勇使用山东省小清河流域某断面的水质监测数据，建立支持向量机预测模型，

验证了支持向量机的良好预测效果[28]。另外一方面，随着支持向量机应用的不断拓展，一些改进的支持向量机方法也不断涌现，如梁学春等解决核函数设计的盲目性和局部最优等非线性优化问题，将一种基于多核加权支持向量机方法应用于水质预测[29]；徐红敏等对标准的支持向量机加以改进，根据训练数据的惩罚系数、拟合误差的情况，不同的训练数据得到不同的权值，得到一种基于加权支持向量机的水质预测方法[30]。

1.4 本文的主要研究内容和成果

支持向量机作为建立在统计学习理论 VC 维的基础上，同时实现结构风险最小化（Structural Risk Minimization，SRM）准则的一种新的机器学习方法，在解决小样本、非线性问题、高维模式识别和局部极小等方面具有优势，目前已被广泛的用于时间序列的预测[31][32]。但目前使用支持向量机对水质时间序列进行预测仍然存在一些问题。

首先，针对支持向量机对水质时间序列的拟合预测精度受其参数选择的影响[33]，已有的基于单因素参数寻优的方法[34]，受各因素之间没有交互作用条件的限制，不仅非常耗时，而且常常存在不能保证找到的参数是最优的问题；再者，针对水质监测参数的时间序列变化受诸多因素的影响，某些监测值表现出突发跳变等不规律变化的情况，虽然采用小波变换能够一定程度上降低跳变值的影响[35]，但是没有充分利用分解后的数据特性，导致在模型训练过程中常常得不到理想的模型。

本文在小波变换进行水质数据分解的基础上，进一步利用分解后的水质数据特性，构建新的预测模型，以降低突变数据对模型构建的影响，以此希望为今后水质预测领域的研究者提供一种新的预测方法，做了如下工作：

（1）采用遗传算法优化支持向量机的参数

遗传算法模拟了生物进化过程中自然选择的过程和群体内部染色体的随机信息交换的机制，基于遗传理论，对支持向量机的参数进行优化，比传统的基于多次单因素实验寻找最优参数的方法更高效。

（2）将小波分解后的细节系数进行平移

水质时间序列经过小波分解为尺度序列和细节序列，得到的细节序列某些数据存在多位小数，数据之间存在数量级的差别，为了减小小数部分输入模型时产生的拟合误差，需要降低数据之间的数量级差，所以将细节序列中的数据统一平移到正数区域。

将上面的方法与支持向量机结合，构建新的预测模型，将该模型用于实际的水质数据预测，并与基于多次单因素实验优化支持向量机参数模型[34]和基于小波

变换的 BP 神经网络模型进行比较分析；再与基于小波变换支持向量机的水质预测模型[35]进行了对比分析。结果表明，这种新的水质时间序列预测模型比原有的支持向量机和 BP 神经网络预测模型，预测精度更高、性能更优越。

其次，前述的水质预测模型采用小波变换数据处理方法，水质时间序列被变换到多个尺度上，有效的减小了瞬变数据对建模的影响，但是原始序列经小波变换被分解为平稳的尺度子序列和不断变化的细节子序列，将这些具有不同的特性的子序列都采用支持向量机建模的方法预测，大量输入的训练数据会影响模型的预测精度。为进一步提高模型的预测精度，将模型的特性和数据的特性充分结合起来，本文继续做了下面的工作：

（3）采用灰色理论对平稳尺度子序列建模

采用灰色理论（最少 4 个数据就能建立预测模型）和支持向量机分别对小波变换分解后的平稳尺度子序列和不断变化的细节子序列建立预测模型。将该预测模型应用于乌梁素海 PH 值和浙江嘉兴自动监测站的溶解氧时间序列预测，通过与文献[36]中的预测结果进行了比较，结果表明：这种新的水质时间序列预测模型比原有的预测模型综合预测精度更高。

1.5 本文的组织结构

本文主要对基于支持向量机的水质动态预测方法进行分析研究，全文共分为五章，各章的内容如下所述：

第一章主要介绍了课题的研究背景与意义，概述了水质预测及支持向量机的研究现状，分析了传统预测模型的不足，并说明了支持向量机在水质预测领域应用具有的优势，最后概述了本文的主要研究内容和成果。

第二章主要介绍了支持向量机的理论。首先介绍了机器学习理论和统计学习理论，包括经验风险、VC 维及结构风险最小化等基本的理论知识。接着介绍了支持向量机线性可分、非线性可分，以及支持向量机的核函数及其参数优化的问题。

第三章主要介绍了基于改进小波变换及支持向量机的水质预测模型。首先，介绍了该模型提出的背景，然后介绍了对原水质预测模型的改进，包括遗传算法寻优及数据平移思想；其次，详细介绍了预测模型的建模过程以及模型评价标准；最后，将该模型与参考文献中模型的最优结果进行了比较和分析。

第四章主要介绍了基于灰色理论及支持向量机的水质预测模型。首先，介绍了该模型是在第三章模型基础上，进一步将水质分解数据的特性和模型更好的结合，分别使用灰色理论和支持向量机对平稳尺度子序列、不断变化的细节子序列建立预测模型；其次，详细介绍了预测模型的建模过程以及模型评价标准；最后，将该模型与第三章中模型的最优结果进行了比较和分析。

第五章对本文的研究工作进行总结，并对未来需要进一步研究的工作进行了分析和展望。

第二章 支持向量机的理论基础

2.1 引言

常见问题中的样本数量通常是有限的，而采用基于样本数目趋于无穷的传统统计学方法就难以取得好的成果。在这种情况下，一种新的基于有限样本的统计学习理论便出现了。支持向量机便是在这种新的统计学习和 VC 维理论的基础上发展起来的，基于结构风险最小化，能够很好的处理小样本情况和解决非线性问题。支持向量机一个重要的应用便是回归，在有限样本空间的情况下，可以良好的平衡样本学习精度和学习程度的关系，其主要的特点是把复杂的问题归为解决一个凸二次规划问题，通过把低维空间数据的非线性问题转化为高维空间中数据的线性问题，以求找到一个合适的超平面，并采用核函数的方法巧妙的解决高维空间的计算问题。可以直接通过训练数据，寻找到支持向量机的最优超平面函数，不需要样本集与支持向量机的任何先验信息，区别于神经网络需要先验知识及依靠人的经验，使得支持向量机模型具有更好的推广能力。目前，在人工智能和机器学习领域，越来越多的研究者将支持向量机应用到实际的问题中，支持向量机理论方法不断的得到拓展，并成为统计学习理论中的重要方法。

机器学习（Machine Learning，ML）是一门涉及概率论、统计学、逼近论等理论的多门学科，其主要是通过计算机模拟人类的学习行为，不断的从观测样本中找出特定的一些规律，获得知识。现实世界中，有大量的可以观测的现象尚不能通过理论进行分析，通过机器学习技术来对这些现象进行学习，找出内部蕴含的规律，智能系统模型不断的自我校正、不断地改善自身性能，来分析这些客观现象，因此机器学习在理论研究、工业生产中得到了广泛的应用。统计学习理论（Statistical Learning Theory，SLT）是机器学习领域的众多研究学科中，解决小样本学习的理论。自上世纪 60 年代提出，统计学习理论经历了没有良好的应用桥梁，实际的应用极少，到 90 年代 Vapnik 提出了支持向量机算法[16]，使得统计学习理论从抽象的理论变为实际的分析回归算法，其在解决有限样本问题的理论才逐渐发展起来，并成为一个较为成熟的理论方法。

2.2 机器学习理论

机器学习是人工智能领域的核心技术，是现代人工智能研究的内容和方向，其主要是通过计算机模拟和学习人类的行为，在学习过程中不断的自我校正、不断地改善自身性能，是实现人工智能的最主要途径[37]。基于现有的训练样本，找

出系统的输入量和输出量之间的依赖关系，并利用这些关系所建立的函数式，来对未来数据量和无法观测的数据量进行模拟预测。这样一种学习可以视为多位函数的逼近问题。

2.2.1 机器学习模型

若给出一个训练样本集$\{(X,Y)\}$，$X \in R^N$，R为实数集，X为N维的输入向量，并且$Y \in R$为一维的输出向量。机器学习的过程就是求解X与Y之间存在未知的函数依赖关系，机器学习的样本学习过程包括以下三部分的函数估计模型，如图 2.1 所示[38]。

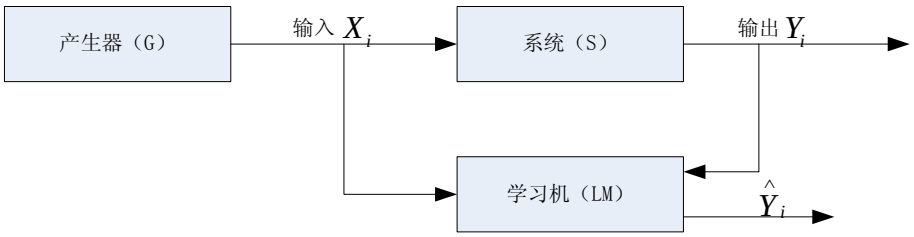

图 2.1 机器学习基本模型

（1）产生器（G），产生任意的系统输入向量$X_i \in R^N$，$i = N^+$，它们都是从不变且未知的概率分布函数$F(X)$中独立抽取。

（2）系统（S），一个输入的向量X_i对应一个输出的值Y_i，产生的依据是同样不变且未知的条件分布$F(Y|X)$。

（3）学习机（LM），在一定条件情况下，存在函数集合$\{f(X,\alpha)\}$，$\alpha \in \Lambda$，且Λ为参数集合，根据输入向量X_i系统输出一个值$\hat{Y}_i$，$\hat{Y}_i$不固定，但是满足函数关系式$\hat{Y}_i = f(X_i,\alpha)$，$\alpha \in \Lambda$，学习机器逼近能力的大小由函数集合中的函数决定。

机器学习的过程就是从给定的函数集合中$\{f(X,\alpha)\}$选择出能够最为逼近系统反馈的一个函数。在系统不断学习的过程中，训练集是根据联合分布$F(X,Y)=F(X)F(X|Y)$取出的l个相互独立分布的训练集合组成：

$$(X_1,Y_1),(X_2,Y_2),(X_3,Y_3),\cdots,(X_l,Y_l) \tag{2.1}$$

为寻求最为逼近系统反馈的函数，需要找到一个损失函数$L(Y,f(X,\alpha))$，$\alpha \in \Lambda$来衡量在相同的输入向量X_i下，输出值Y_i和系统学习得到的反馈值$\hat{Y}_i = f(X_i,\alpha)$之间的误差大小，这种差别用损失函数的数学期望来表示：

$$R(\alpha) = \int L(Y,f(X,\alpha))dF(X,Y) \tag{2.2}$$

式（2.2）中$R(\alpha)$称为函数集在训练数据集合上的期望风险，有时候也被称为风险泛函，α为广义参数。在训练集的联合分布$F(X,Y)$未知的情况下，找到函数集合$\{f(X,\alpha)\}$中期望风险$R(\alpha)$最小的函数$f(X,\alpha_0)$，其便是最为逼近系统反馈的函

数。

可以看出，损失函数在求解最为逼近系统反馈的函数的过程中功不可没，在机器学习中，针对学习目标的不同，必须采用不同的损失函数，模式识别问题、函数逼近问题及概率密度问题是三类基本的机器学习问题。

模式识别问题中，Y 表示类别编号，在两类问题下 Y 的取值集合可以为 $\{0,1\}$ 或者 $\{-1,1\}$，其损失函数定义如下：

$$L(Y, f(X,\alpha)) = \begin{cases} 0, \ if \ \ Y = f(X,\alpha) \\ 1, \ if \ \ Y \neq f(X,\alpha) \end{cases} \tag{2.3}$$

函数逼近问题中，Y 表示连续变量，使用最小平方规则来衡量误差，损失函数如下：

$$L(Y, f(X,\alpha)) = (Y - f(X,\alpha))^2 \tag{2.4}$$

概率密度问题中，系统学习是为了根据训练样本来确定 X 的概率密度，假设 X 的密度函数为 $p(X,\alpha)$ ，损失函数定义如下：

$$L(p(X,\alpha)) = -\log p(X,\alpha) \tag{2.5}$$

机器学习的过程，是从给定的函数集合 $\{f(X,\alpha)\}$ 中选择出能够最为逼近系统反馈的一个函数，这一过程实际上是在训练集的联合分布 $F(X,Y)$ 下，通过最小化风险泛函 $R(\alpha)$，找到学习系统中广义参数 α 的过程。

2.2.2 经验风险最小化

实际问题中，由于训练集的联合分布 $F(X,Y)$ 并不知道，因此期望风险并不能直接求出。为了解决这样一个问题，提出采用已知的观测样本 $\{(X,Y)\}$，根据概率论中的大数定律，通过求解损失函数的算术平均得到期望风险的值：

$$R_{emp}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} L(Y_i, f(X_i,\alpha)) \tag{2.6}$$

$R_{emp}(\alpha)$ 采用已知的观测样本 $\{(X,Y)\}$ 来定义，这些已知的观测样本也被称为经验数据，所以这种情况下定义的期望风险也随之被称为经验风险。为了求的最为逼近系统反馈的函数，去寻找学习系统中广义参数 α，将最小化经验风险 $R_{emp}(\alpha)$ 替代传统定义上的最小化期望风险 $R(\alpha)$ 这一过程，称为机器学习中的经验风险最小化。

目前，经验风险存在着一些疑问。在概率论的大数定律下，只有在训练样本集趋于无穷的情况时，$R_{emp}(\alpha)$ 才在概率的意义下近似等于 $R(\alpha)$，且 $R_{emp}(\alpha)$ 和 $R(\alpha)$ 不一定在同一点上取得最小值。另外，在样本数据无穷大时能够取的好的学习效果，但在有限样本下可能就不会出现这种情况。

2.2.3 泛化能力与复杂性

多年来，经验风险最小化原理一定程度上解决了期望风险最小化的问题。在机器学习领域，经典的机器学习算法中都是基于经验风险最小化的，并且大部分的研究者都在寻找最小化经验风险的最优解。如神经网络的研究早期，研究者总是追求 $R_{emp}(\alpha)$ 更小，但是只有在样本趋于无穷大的时候，$R_{emp}(\alpha)$ 才在概率的意义下近似等于 $R(\alpha)$，过度的追求训练误差最小，并不能代表预测效果更好。只要神经网络足够复杂，拟合的时间够长，训练误差可以降到很低，在有限的样本情况下，经验风险甚至可以快速的降低到零，但是这样训练出来的网络几乎不具有泛化能力，更何况对新样本的预测效果，这就是神经网络的“过学习”现象。所以通过对有限的样本学习得到这样一个非常复杂的网络，往往失去良好的推广能力，这就造成了复杂系统及其推广能力之间的矛盾。

因此，一味的追求经验风险最小化，只会让学习系统变得更为复杂，尽管训练误差逐渐降低，但是测试误差难以降低，所以学习系统的泛化能力不高。复杂的学习系统会造成网络学习变慢，对机器处理的性能造成较大的影响。在实际问题中，若采用复杂的学习系统虽然可以使得风险经验最小，但是会导致测试误差过大，如何在系统的学习精度和复杂程度之间平衡，提高学习系统的泛化能力变得尤为重要。在研究过程中，有人不过分追求学习函数的训练误差最小，从而降低了系统的复杂性，采用验证法来选择模型，控制其复杂度等，但由于这种方法缺乏理论支撑，没有从根本上解决问题[3]。

2.3 统计学习理论

统计学习理论主要解决有限样本的估计及预测问题，其研究内容主要分为四部分[39]：

（1）虽然在概率论的大数定律下，样本趋于无穷大的时，经验风险 $R_{emp}(\alpha)$ 在概率的意义下近似等于期望风险 $R(\alpha)$，在实际情况中却不全是这样。所以需要找到在什么条件下，$R_{emp}(\alpha)$ 的最优值是等于 $R(\alpha)$ 最优值的，即统计学习一致性问题。

（2）解决估计经验风险和实际风险之间存在的设计推广的界问题。

（3）根据经验风险和实际风险之间的设计推广的界估计，选择预测函数的问题。

（4）实现新的统计理论学习算法。

在统计学习理论已经有众多衡量函数学习性能的理论，其中最为重要的两个理论是 VC 维和推广的界。

2.3.1 VC 维

VC 维是统计学习理论中的核心概念。其是描述了学习机器或函数集复杂性大小或学习性能强弱的一个重要的指标[40]。VC 维的定义是：假设一个样本集合含有 k 个点，可以用正例和负例（属于或者不属于某个类）将这些点表示为 2^k 种形式。若样本这 2^k 种形式能够被函数集中的学习函数分开或者识别，那么就认为函数集能够将这 k 个样本点打散。函数集的 VC 维就被定义为能够打散的最大样本数目。由此可知，有无穷多个样本点都能够被函数集打散，那么该函数集的 VC 维是无穷大的。通常有界的实值函数的 VC 维的确定，是通过阈值把函数转化成指示函数来得到的[41]。

目前，VC 维是对学习机器或者函数集合学习能力强弱描述的最好指标，通常来说，VC 维值越大，学习机器或者函数集的学习能力就越强，通常学习机器的结构便复杂。现在，还没用通用的算法来求解学习机器的 VC 维，像神经网络这样一些复杂模型的 VC 维，不仅和选择的函数集有关系，而且还与学习算法有关，所以求解 VC 维更为困难。但是一些特殊函数的 VC 维可以知道：在 N 维实数空间中，线性分类器、实函数的 VC 维便是 $N+1$。

2.3.2 推广的界

推广的界是统计学习理论中表示经验风险和实际风险关系的理论，也是用来分析学习机器的性能，以及实现新的统计学习理论算法的重要基础理论[16]。对于指示函数集中的函数，经验风险 $R_{emp}(\alpha)$ 和期望风险 $R(\alpha)$ 之间以概率 $1-\eta$ 满足如下关系式：

$$R(\alpha) \le R_{emp}(\alpha) + \Phi(n/h) \tag{2.7}$$

式（2.7）中 $\Phi(n/h)$ 表示置信范围，h 是指示函数集的 VC 维，n 表示样本数，等式表示学习机器的期望风险等于经验风险 $R_{emp}(\alpha)$ 与置信范围 $\Phi(n/h)$ 的和。该式子表示经验风险及期望风险之间差值的上界，也表示了基于经验最小化原理模型的泛化能力大小，因此将其称为推广的界。且置信范围 $\Phi(n/h)$ 的数学表达式如下：

$$\Phi(n/h) = \sqrt{\left|\frac{h(\ln(2n/h)+1) - \ln(\eta/4)}{n}\right|} \tag{2.8}$$

式（2.8）以 $1-\eta$ 概率成立。由该式子可以推出，固定函数集的 VC 维，$\Phi(n/h)$ 是随着样本数 n 增加的单调减函数；固定样本数 n，$\Phi(n/h)$ 是随着函数集 VC 维增加的单调增函数。

由等式（2.7）可以看出，固定数量的样本集合，当函数集合较大时，能够选择合适的函数得到较小的经验风险，而此时函数集合的 VC 维数颇高，使得置信范

围较大；当函数集合较小时，VC 维没有大函数集合的 VC 维高，即 VC 维较小，置信范围也较小，但是经验风险会增大；所以经验风险 $R_{emp}(\alpha)$ 和函数集的 VC 维（经验风险越小，网络越复杂，VC 维越大）是此消彼长的关系，两个变量不会同时很小，所以需要找到学习机器合理的 VC 维，使得经验风险和置信范围的和最小。某些情况下的“过学习”现象，便是由于在有限样本的情况下，学习机器的复杂程度过高，VC 维偏大，从而置信范围极大，而经验风险的变化又有一个下界，从而导致经验风险与期望风险具有较大差，这便是需要找到合适的 VC 维，使得期望风险和经验风险之间的差距最小的最好例证。需要注意的是，推广性的界在许多情况下是较为松弛的，当 VC 维数较高时更是这样（置信范围大）；通过推广的界可以方便对同一类的函数进行比较，以便从中选择出最优的函数，而不适用于不同类别学习函数之间的比较[42]。

2.3.3 结构风险最小化

传统机器学习算法的学习过程，就是不断的在优化置信范围，通常在模型与训练样本比较相关的时候，才可能取得好的拟合效果，但这种只能依赖先验知识的方法过分的依赖人们的经验。由于 $\Phi(n/h)$ 是随着 n/h 增加的单调减函数，在式（2.6）中：

（1）n/h 较大时，$\Phi(n/h)$ 值较小，所以这时期望风险与经验风险差值较小，所以在这种情况下经验风险小能够保证取得较小的期望风险，但通常这种情况下经验风险也较大，并且实际情况中，样本数量 n 不会太大。

（2）n/h 较小时，$\Phi(n/h)$ 值较大，这时即使经验风险较小，期望风险也不一定会小。所以为了使得 $R(\alpha)$ 最小，必须要让等式右边的经验风险 $R_{emp}(\alpha)$ 和置信范围 $\Phi(n/h)$ 和同时较小。

上面的问题（2）是要着重解决的问题，在 2.3.2 节中也已经提到了，需要找到学习机器合理的 VC 维，来平衡经验风险和置信范围的大小，此时结构风险最小化的理论便是来解决这个问题。

结构风险最小化（SRM）便是寻求经验风险和置信范围之和最小，使得学习机器能够有很好的推广能力，其原理如下：

首先，将函数集 $S=\{f(X,\alpha)\}$，$\alpha\in\Lambda$，分解为一个子函数集的序列：

$$S_1\in S_2\in\cdots\in S_k\in\cdots S$$

其次，根据各子函数集的置信范围 $\Phi(n/h)$ 大小，将其 VC 维按递增的序列排列（VC 维与置信范围成正比，即按照 VC 维的大小排列）：

$$h_1\le h_2\le\cdots h_k\le\cdots$$

在每个子函数集中寻找 $R_{emp}(\alpha)$ 最小的函数，在子集之间平衡经验风险和置信

范围的大小，使得期望风险最小，这便是结构风险最小化过程，如图 2.2。在求解最优函数时，一种方法是，根据各子函数集合的置信范围一样，并且能够在子函数集合中求出各个函数的经验风险，所以通过比较各子函数集合中置信范围和最小经验风险的和值，可以求得最小的期望风险，那么此时求得最小经验风险的函数便是最优函数。通过这种方法，寻找最优函数，尤其是在函数集合非常大的时候，求解最小的期望风险是非常耗时的一个过程。于是，又设计了函数集合的结构，使得各子函数集合都能够取得最小的经验风险，然后将置信范围值最小的子集选择出来。寻找到的最优函数就是选择出来的子集中经验风险取最小值的函数，支持向量机寻找最优函数便是这样的一个过程。

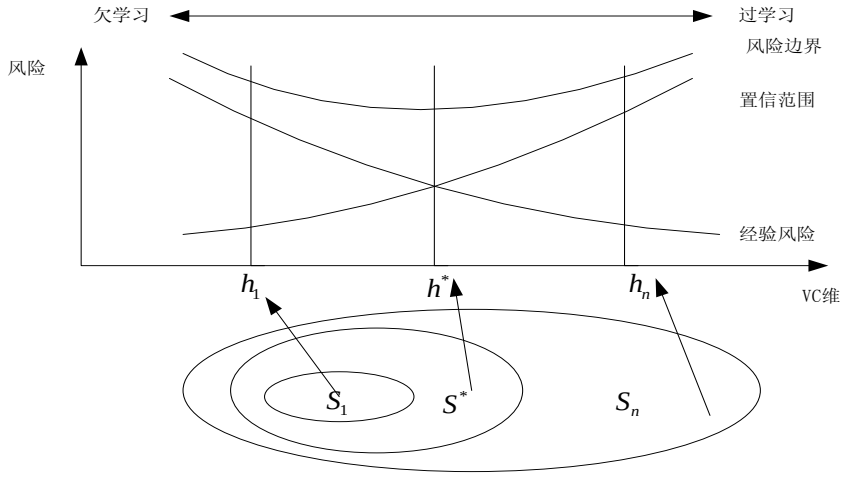

图 2.2 结构风险原理

2.4 支持向量回归机的原理

本文仅使用支持向量机的回归理论，应用于水质时间序列建模，由于篇幅的限制，所以这里仅介绍支持向量机的回归原理。支持向量机最早被应用于数据的分类处理，其在样本空间构造一个最优超平面，将两类正确分开，并使得分类的间隔最大，目的是为了在使得经验风险最小的同时，让置信范围最大。回归分析问题是一个求解函数估计的过程，给出一个训练样本集$\{(X_i,Y_i)|i=1,2,\cdots,l\}$，其中$X_i \in R^N$为$N$维输入向量，$Y_i \in R$为相对应的输出量，寻找出反映这个系统输入输出关系的函数$f(x)$，若找出的函数$f(x)$是线性的，那么这是一个线性回归问题，或$f(x)$是非线性的函数，则这便是一个非线性的回归问题。

支持向量机的回归理论是在解决分类的问题上发展起来的一种回归算法，本文和许多其他文献一样，在名称上，将支持向量机等同于支持向量回归机。支持向量机能够求解线性回归问题，也能求解非线性问题，实际情况中大多是非线性问题，并且支持向量机能够保证结构风险最小，在求解回归问题上比很多传统的

基于经验最小化的学习方法精度更高、泛化能力更强。在分类问题中，使用支持向量机求得的最优超平面将两类样本分的“最开”，而回归问题与分类问题有所区别的是样本都属于一类，此时是让所有的样本点距离最优超平面的“总偏差”最小。其实可以看出，解决回归问题，最终是要让所有的样本点都要落在两条边界线之间，所以求解最优回归超平面就等价于求解分类问题中的最大间隔。

在支持向量机的回归理论中，目前已经有许多的支持向量机模型，本文将ε-支持向量机，应用到水质时间序列的预测建模中，这里对其做一些介绍。

2.4.1 线性可分情况

通过支持向量机寻找最优函数估计，就是要找到如图 2.3 中所示的最优回归超平面，使得让尽可能多的样本落入两条边界线中。所有的样本数据到最优回归超平面的距离视为拟合的误差，而最优回归超平面是所有平面中样本的误差累计和最小的一个平面。

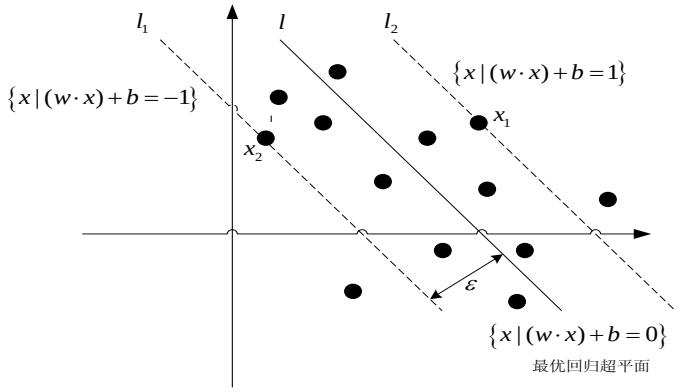

图 2.3 最优回归超平面和ε带

Vapnik 提出了ε不敏感损失值，在图 2.3 中，边界线之间的点距离最优超平面的距离都小于ε，边界线之间的这条带称为ε带，这些点不为目标函数提供损失值，即监测值Y与预测值$\hat{Y}$的差在小于ε时，认为该点到最优超平面的距离为 0，其误差表示函数如下：

$$L(y)=\begin{cases}0 & |f(X)-Y|<\varepsilon \\ |f(X)-Y|-\varepsilon & \text{其他}\end{cases} \tag{2.9}$$

当所有样本点落入边界之间的ε带，此时样本到最优回归超平面的总误差和为零，所有样本点到超平面的距离都小于ε，给出一个训练样本集$\{(X_i, Y_i)|i=1,2,\cdots,l\}$，其中$X_i\in R^N$为$N$维向量，$Y_i\in R$为相对应的输出变量，支持向量回归求解最优超平面的问题为一个凸二次规划问题：

$$\min \frac{1}{2}\|\omega\|^2 \tag{2.10}$$

$$s.t.\begin{cases} Y_i-(\omega,X_i)-b\le\varepsilon \\ (\omega,X_i)+b-Y_i\le\varepsilon \end{cases}$$

式（2.10）中$\omega\in R^N$为可变动的权值函数；ε为不敏感损失值，即容忍的误差；b为阈值。

在实际回归问题中，如果让所有的样本点全部落入边界线中，此时经验风险为 0，但是置信范围非常大，造成模型的泛化能力极弱；如果让所有的样本点落入边界线之外，置信范围虽小，经验风险过大，模型的拟合效果不佳。

所以允许一部分样本在边界之外，以保证经验风险和置信范围和最小，此时使用松弛变量ξ_i，表示那些溢出边界的样本点距离边界的距离，如图 2.4。求解超平面的问题转化为下面的二次规划问题：

$$\min\ \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*) \tag{2.11}$$
$$s.t.\begin{cases} Y_i-(\omega,X_i)-b\le\varepsilon+\xi_i \\ (\omega,X_i)+b-Y_i\le\varepsilon+\xi_i^* \\ \xi_i,\xi_i^*\ge 0,\ \ i=1,2,\cdots,l \end{cases}$$

式（2.11）中C为惩罚因子，为模型的复杂度与拟合精度之间的折衷；ξ_i，ξ_i^*是超出误差ε的溢出样本点偏差值，(ω,X_i)表示权值函数ω和向量X_i的内积。

由*Lagrange*优化方法，将式（2.11）转化*Lagrange*函数：

$$\begin{aligned} L(w,b,\xi_i,\xi_i^*)=&\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)-\sum_{i=1}^{l}\alpha_i[(\omega,X_i)+b-Y_i+\varepsilon+\xi_i] \\ &-\sum_{i=1}^{l}\alpha_i^*[Y_i-(\omega,X_i)-b+\varepsilon+\xi_i^*]-\sum_{i=1}^{l}(\eta_i\xi_i+\eta_i^*\xi_i^*) \end{aligned} \tag{2.12}$$

将该等式对w，b，ξ_i和ξ_i^*求偏导数，得到：

$$\begin{cases} \dfrac{\partial L}{\partial w}=0\Rightarrow w=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)X_i \\ \dfrac{\partial L}{\partial b}=0\Rightarrow\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \\ \dfrac{\partial L}{\partial \xi_i}=0\Rightarrow C-\alpha_i-\eta_i=0 \\ \dfrac{\partial L}{\partial \xi_i^*}=0\Rightarrow C-\alpha_i^*-\eta_i^*=0 \end{cases} \tag{2.13}$$

将这些式子带入转化为对偶优化问题：

$$\min\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(X_i,X_j)-\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)+\varepsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*) \tag{2.14}$$
$$s.t.\begin{cases} \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \\ \alpha_i,\alpha_i^*\in[0,C] \end{cases}$$

式（2.14）中α_i、α_i^*为拉格朗日乘子。

最后得到最优回归超平面的近似函数：

$$f(X)=w\cdot x+b=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)(X,X_i)+b \tag{2.15}$$

由 KTT 条件，得到：

$$\begin{cases} a_i[(\omega,X_i)+b-Y_i+\varepsilon+\xi_i]=0 \\ \alpha_i^*[Y_i-(\omega,X_i)-b+\varepsilon+\xi_i^*]=0 \\ (C-\alpha_i)\xi_i=0 \\ (C-\alpha_i^*)\xi_i^*=0 \end{cases} \tag{2.16}$$

求得阈值：

$$\begin{aligned} b=\frac{1}{N_{SV}}\Bigg\{&\sum_{0<a_i<C}[Y_i-\sum_{X_i,X_j\in SV}(\alpha_i-\alpha_i^*)(X_i,X_j)-\varepsilon] \\ &+\sum_{0<a_i^*<C}[Y_i-\sum_{X_i,X_j\in SV}(\alpha_i-\alpha_i^*)(X_i,X_j)-\varepsilon]\Bigg\} \end{aligned} \tag{2.17}$$

式（2.17）中N_{SV}为支持向量总数，SV为支持向量集，说明最优回归超平面的构建是基于支持向量。

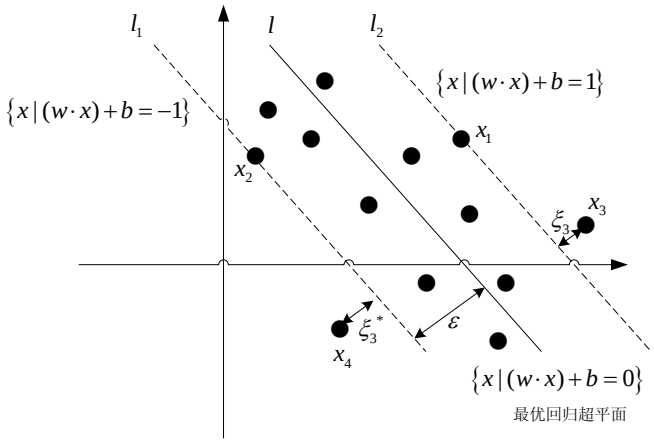

图 2.4 样本点和ξ关系

2.4.2 非线性可分情况

支持向量回归解决非线性问题，是将非线性的原始输入值X_i，$i=1,2,\cdots,l$，映射到高维的特征空间H（*Hibert* 空间），即达到将低维空间的非线性问题转化为高维空间的线性问题[43]。当给出新的测试样本数据时，然后找出一个最优回归超平面（Optimal Hyper-plane），能够通过回归超平面函数得到目标值。

引入非线性的映射函数φ，此时高维特征空间H中最优回归超平面的近似函数为：

$$f(X)=w\cdot x+b=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)[\varphi(X),\varphi(X_i)]+b \tag{2.18}$$

并且阈值为：

$$b=\frac{1}{N_{SV}}\left\{\sum_{0<a_i<C}[Y_i-\sum_{X_i,X_j\in SV}(\alpha_i-\alpha_i^*)[\varphi(X),\varphi(X_i)]-\varepsilon]+\sum_{0<a_i^*<C}[Y_i-\sum_{X_i,X_j\in SV}(\alpha_i-\alpha_i^*)[\varphi(X),\varphi(X_i)]-\varepsilon]\right\} \tag{2.19}$$

高维空间中的向量点积 $\varphi(X_i)\cdot\varphi(X_j)$可以由满足 *Mercer*条件的对称核函数 $K(X_i,X_j)$ 代替，从而巧妙地降低了非线性映射的计算量，所以最优回归超平面的近似函数为：

$$f(X)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(X,X_i)+b \tag{2.20}$$

阈值表示为：

$$b=\frac{1}{N_{SV}}\left\{\sum_{0<a_i<C}[Y_i-\sum_{X_i,X_j\in SV}(\alpha_i-\alpha_i^*)K(X_i,X_j)-\varepsilon]+\sum_{0<a_i^*<C}[Y_i-\sum_{X_i,X_j\in SV}(\alpha_i-\alpha_i^*)K(X_i,X_j)-\varepsilon]\right\} \tag{2.21}$$

2.4.3 支持向量机的核函数

常用的核函数有多项式核函数、径向基核函数和 *Sigmoid* 核函数等。由于径向基核函数在非线性系统识别方面更具有优势[44]，本文因此选取径向基核函数:

$$K(X_i,X_j)=\exp\left[-\frac{\|X_i-X_j\|^2}{2\sigma^2}\right] \tag{2.22}$$

多项式核函数为：

$$K(x_i,x_j)=(x_i.x_j+1)^d \quad d=1,2,\cdots \tag{2.23}$$

*Sigmoid*核函数为：

$$K(x_i,x_j)=\tanh[b(x_i\cdot x_j)+c] \tag{2.24}$$

2.4.4 支持向量机参数优化

在采用支持向量机进行水质预测建模时，将训练数据输入支持向量机进行模型训练，目的就是找到一个最优回归超平面，使得训练数据到最优回归超平面距离尽可能的小，并且使得结构化风险最小。在寻找最优超平面函数的过程中，需要确定不敏感损失值、惩罚因子、核函数参数的具体值。不敏感表示了拟合允许的误差，惩罚因子表示了对落在 ε 带外的样本的惩罚程度的高低，核函数参数影响着数据空间分布的复杂度。不同的参数取值训练得到的模型也不同，对水质时间序列的拟合预测精度具有重要的影响，因此参数的选择十分重要。常用基于单因素参数寻优的方法，需要人工的选择，不仅耗时，而且找到的参数常常不是最优

的，于是采用智能搜寻算法找到合适的参数可以避免这些问题，不仅提高预测建模的效率，而且能够提高模型拟合的精度[45]。

2.5 本章小结

本章首先介绍了机器学习和统计学习的基本理论概念，包括经验风险、泛化能力、VC 维和推广的界等内容。分析了在保证结构风险最小化的情况下，模型具有良好的泛化能力及较好的拟合误差，能够平衡经验风险与置信范围，进而推出支持向量机的概念。接着介绍了支持向量机的线性可分、非线性可分，以及支持向量机的核函数及其参数优化的问题。

第三章 基于改进小波变换及支持向量机的水质预测模型

3.1 引言

武国正等人提出采用支持向量机模型对浅水湖泊乌梁素海的 PH 值进行预测，并将预测结果与传统的线性回归、BP 神经网络和 RBF 神经网络进行了比较，验证了支持向量在水质预测中的优势[34]，但是文章中对支持向量机的参数寻优采用传统的单因素参数寻优的方法，不仅非常耗时，而且不能保证找到参数最优。梁坚等人提出采用小波变换将水质时间序列进行分解，来降低突然跳变等不规律变化的数据对预测模型精度的影响，通过与神经网络的方法进行对比，验证了预测模型的良好效果[35]，由于文中采用小波变换分解后的细节序列存在多位小数，数据之间存在数量级的差别，所以还可以进一步的改进，充分利用水质数据的特性。本文同样采用小波变换，将水质时间序列变换到多个尺度上，以减小瞬变数据对建模的影响，并在此基础上，充分利用水质数据时间序列的特点，进一步降低水质数据特性对模型拟合效果的影响，提高支持向量机的预测精度，做了如下工作：

（1）采用遗传算法优化支持向量机的参数

遗传算法是以自然选择和遗传理论为基础，模拟了生物进化过程中适者生存规则与群体内部染色体的随机信息交换机制，遗传算法对支持向量回归机的参数进行寻优，比传统的基于多次单因素实验寻找最优参数的方法更高效。

（2）将小波分解后的细节系数进行平移

水质时间序列经过小波分解为尺度序列和细节序列，得到的细节序列某些数据存在多位小数，数据之间存在数量级的差别，为了减小小数部分输入模型时产生的拟合误差，需要降低数据之间的数量级差，所以将细节序列中的数据统一平移到正数区域。

将上面两项技术引入新的支持向量机预测模型，将其应用于乌梁素海的水质数据预测，并与文献[34]中提出基于多次单因素实验优化支持向量机参数和基于 BP 神经网络的两种水质预测模型进行比较分析；再和文献[35]中基于小波变换支持向量机的水质预测模型进行了对比分析。

3.2 水质预测模型建模

3.2.1 遗传算法参数寻优

遗传算法是以自然选择和遗传理论为基础，模拟了生物进化过程中适者生存规

则与群体内部染色体的随机信息交换机制，在模式定理和积木块假设[46]保证能够找到全局最优解的条件下，能自适应的控制搜索过程的高效全局搜索算法。本文基于遗传算法的思想，建立适应度函数，来解决支持向量机参数寻优这类多元单峰值的优化问题。

（1）*K*-折交叉验证及遗传算法的适应度函数

在 K -折交叉验证中，首先将训练集分为 K 个数量相同的子集，然后取其中的一个子集作为模型测试集，剩下的 $K-1$ 个子集作为模型训练集。这样重复 K 次，使得训练集中的所有样本都能够被模型预测一次。通过交叉验证分割训练集进行训练测试的方法，充分验证了训练集中各数据的特性，使得预测结果更好地反映了模型实际效果，因此能够防止被选择的模型出现过拟合的情况。目前在进行支持向量回归模型训练的时候，也采用交叉验证的方法[47]。

通过遗传算法优化支持向量回归模型的参数时，需要比较种群个体的适应度大小，这里适应度函数定义为负的均方误差：

$$f(x) = -\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 \tag{3.1}$$

式（3.1）中 N 为训练样本的数目；Y_i 为实际值；$\hat{Y}_i$ 为使用交叉验证方法获得的预测值。

（2）遗传算法参数优化步骤

惩罚因子用 C 表示、损失值用 E 表示、径向基核函数的参数用 G 表示。

步骤 1：设置初始值：最大进化代数、种群规模和交叉验证折数 K 等。随机初始化一个种群 $Init\_Pop$，进行模型训练，求得种群中各个体适应度值 $Valu\_F$ 。

步骤 2：根据父代种群 $Init\_Pop$ 中个体，计算最高适应度，并选取适应度最高时的 C 、 E 、 G 值，然后将它们分别作为当前最佳适应度值 $Best\_Fit$ 、最佳 $Best\_C$ 、 $Best\_E$ 、 $Best\_G$ 值。

步骤 3：选择算子：从种群 $Init\_Pop$ 中选出一部分适应度高的优良个体构成一个新的种群。

步骤 4：交叉算子：基于步骤 3 得到的种群，完成种群中的个体染色体的交叉重组，形成一个新的种群。

步骤 5：变异算子：基于步骤 4 得到的种群，完成种群中个体染色体的基因突变，而增加群体多样性，形成一个新的种群，然后进行模型训练，求得到该种群中各个体适应度值 $Valu\_M$ 。

步骤 6：形成子代种群：比较种群 $Init\_Pop$ 和步骤 5 得到的种群个体适应度值 $Valu\_F$ 和 $Valu\_M$ ，将步骤 5 得到种群中适应度值高的个体替代种群 $Init\_Pop$ 中适应度低的个体，得到适应度水平更高的新种群 $Init\_Pop$ 。此时 $Init\_Pop$ 更新

作为子代。根据子代种群 $Init\_Pop$ 求的此时种群的最高适应度 Max_Fit 和取最高适应度的 C、E、G 值。

步骤 7：比较步骤 6 中的最高适应度 Max_Fit 与当前最佳适应度值 $Best\_Fit$，若 Max_Fit 大于 $Best\_Fit$，则更新最佳适应度值 $Best\_Fit$ 和最佳 $Best\_C$、$Best\_E$、$Best\_G$ 值；否则不做任何操作，转向下一步继续执行。

步骤 8：未达到最大进化代数，转向步骤 3；否则算法停止，求得最优 $Best\_C$、$Best\_E$、$Best\_G$ 值。

遗传算法寻找最优参数原理图 3.1。

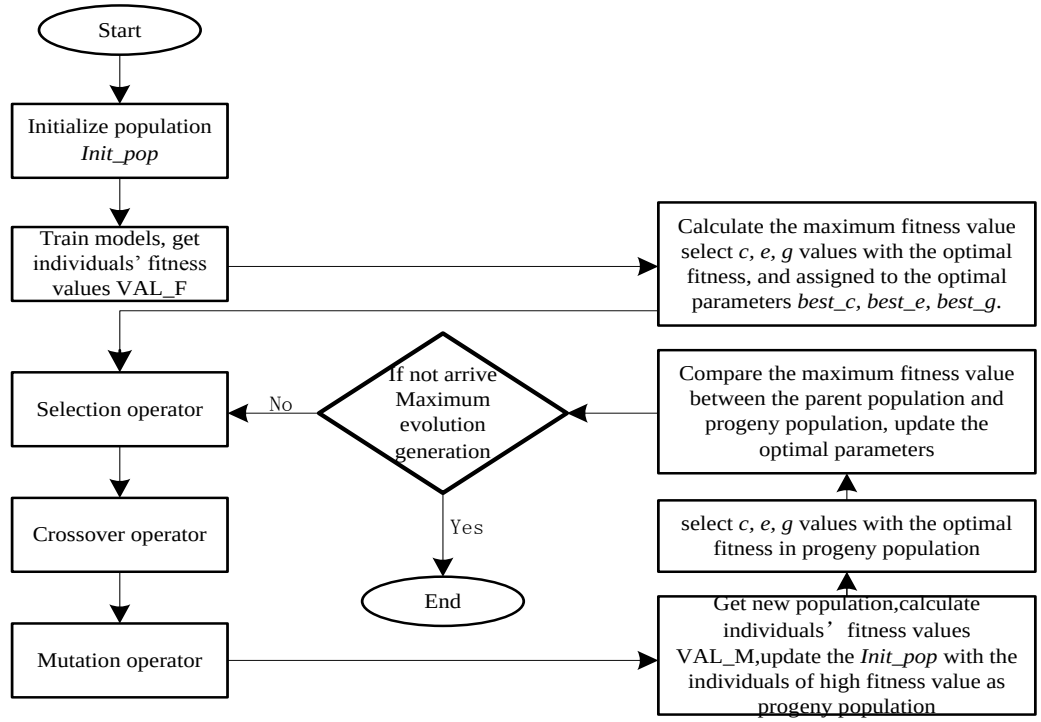

图 3.1　遗传算法原理

3.2.2 小波变换

小波多分辨率分析的主要思想是用不同的分辨率来逐级逼近待分析序列。任意一个初始时间序列为 $S(t)$，$t=1,2,\cdots,n$。通过高通和低通滤波器对 $S(t)$ 进行一步分解，可以将其变换为低频部分 cA_1 和高频部分 cD_1[48]。可以继续对低频部分 cA_1 进行分解，当进行尺度 $1,2,\cdots,N$ 步分解后，就得到一个低频部分 cA_N 和 N 个高频部分 $cD_1,cD_2,\cdots,cD_N$，一步分解过程如图 3.2（a）所示。为使得分解序列与初始时间序列在时间上具有一一对应的关系，根据小波多分辨分析的分解和重构特性，将低频部分 cA_N 和高频部分 $cD_1,cD_2\cdots cD_N$ 进行单支重构，可以得到各尺度下的尺度子序列和细节子序列 $A_N,D_1,D_2,\cdots,D_N$，单支重构过程如图 3.2（b）。

（a）　　（b）

图 3.2 信号的时间-尺度分解（a），各尺度系数的单支重构（b），H_1, L_1 为分解使用的高通和低通滤波器，与输入序列作卷积。H_2, L_2 为重构权重系数。

初始时间序列 $S(t)$ 可以由分解序列 $A_N, D_1, D_2, \cdots, D_N$ 表示：

$$S(t) = A_N + \sum_{i=1}^{N} D_i \tag{3.2}$$

3.2.3 数据准备

由小波变换对原始的时间序列 $S(t)$ 进行多分辨分析，得到各尺度下的序列集合为 $A_N, D_1, D_2, \cdots, D_N$。在分别使用他们的数据序列进行模型训练前，首先根据 *Takens* 理论进行相空间重构，将序列 $\{x_1, x_2, x_3, \cdots, x_n\}$ 转化成矩阵的形式，获得数据之间的关联关系来得到尽可能多的信息量：

$$X = \begin{bmatrix} X_{m+1} \\ X_{m+2} \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \cdots & x_m \\ x_2 & x_3 \cdots & x_{m+1} \\ \vdots & \vdots \ddots & \vdots \\ x_{n-m} & x_{n-m+1} \cdots & x_{n-1} \end{bmatrix}, \quad Y = \begin{Bmatrix} x_{m+1} \\ x_{m+2} \\ \vdots \\ x_n \end{Bmatrix} \tag{3.3}$$

式（3.3）中 m 为重构相空间的嵌入维数。重构相空间的嵌入维数 m 反映重构矩阵的信息量，不同的 m 值对模型的预测效果有不同的影响，可以根据预测误差值的大小来优化选取相空间的嵌入维数。

设 $\hat{Y}_t = \{x_t\}$ 为预测的目标值，将之前的目标值 $X_t = \{x_{t-m}, x_{t-m+1}, x_{t-m+2}, \cdots, x_{t-1}\}$ 作为相关量，建立自相关输入 X_t 与输出 $\hat{Y}_t$ 之间一一映射的关系：$f: R^m \to R$，使得：

$$\hat{Y}_t = f(X_t) = f(x_{t-m}, x_{t-m+1}, x_{t-m+2}, \cdots, x_{t-1}) \tag{3.4}$$

为了降低建模误差，减小数据量纲和加快模型的构建速度，对尺度子序列 A_N 经相空间变换后的矩阵 X, Y 以列向量形式进行数据的归一化：

$$xg_i = (b-a)\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} + a \quad i = k, k+1, \cdots, k+n-m-1 \text{且} k = 1, 2, \cdots, m+1 \tag{3.5}$$

式（3.5）中列向量数据归一化区间为 $[a, b]$，x_i 为归一化前的数据。$x_{\max}$，$x_{\min}$ 分别为列向量中的最大值和最小值。

多尺度分析后得到细节子序列 $D_1, D_2, \cdots, D_N$ 进行相空间重构前，因为序列中数据存在多位小数，且存在数量级的差别，为减小小数部分的拟合预测误差、降低

数量级差，将序列中的数据统一平移到正数区域。若原始序列为 $D_i=\{d_1,d_2,d_3,\cdots,d_n\},i=1,2,\cdots,N$，对序列做平移处理后的序列为：

$$D_i^+=\{d_1+Q,d_2+Q,d_3+Q,\cdots,d_n+Q\}\quad Q\in N+,i=1,2,\cdots,N \tag{3.6}$$

式（3.6）中平移因子 Q 值取为使得序列 D_i^+ 中的所有数据为正数的最小正整数。

3.2.4 模型构建

模型建立的过程如图 3.3 示，步骤为：

步骤 1　若初始的水质时间序列为 $S(t)$，$t=1,2,\cdots,n$，首先将 $S(t)$ 进行小波分解，得到一个低频序列 cA_N 和 N 个高频序列 $cD_1,cD_2,\cdots,cD_N$。再将分解序列进行单支重构，得到重构序列 $A_N,D_1,D_2,\cdots,D_N$。

步骤 2　将细节子序列 $D_1,D_2,\cdots,D_N$ 按公式（3.6）统一平移到正数区域。再根据相空间重构理论，将尺度子序列 A_N 和平移变换后的细节子序列，按（3.3）的形式进行相空间重构，即转化成矩阵的形式。按照公式（3.5），将变换到相空间的尺度子序列按列向量方式进行归一化。

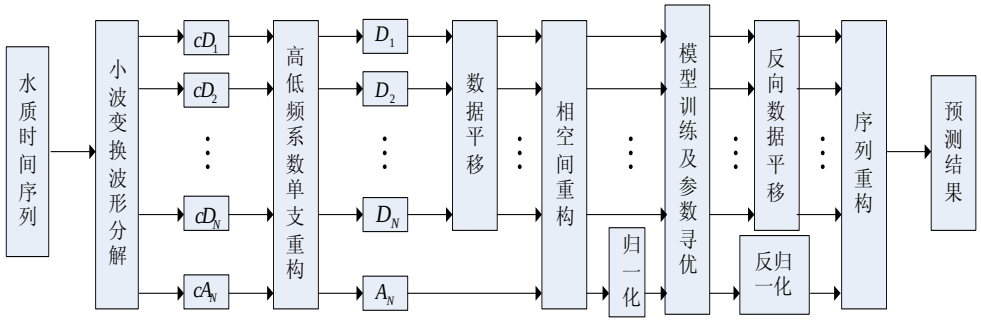

图 3.3　预测模型的建立

步骤 3　结合遗传算法进行参数优化，将经过处理的序列分别进行模型训练，建立各子序列的预测模型。

步骤 4　在进行预测分析时，根据步骤 1 和步骤 2 对数据进行处理，然后使用步骤 3 得到的子序列预测模型进行预测。如输入的向量为 $X_t=\{x_{t-m},x_{t-m+1},x_{t-m+2},\cdots,x_{t-1}\}$ 时，根据式（2.19）的最优回归函数，此时可以得到第 t 点的预测值：

$$\hat{Y}_t=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(X_t,X_i)+b\quad t=m+1,m+2,\cdots,n+1 \tag{3.7}$$

步骤 5　最后，将各模型得到预测的值根据公式（3.2）进行重构，最终得到预测结果。

3.3 模型性能评价标准

本文选取相对误差（Relative Error，RE）、平均相对误差（Average Relative Error，ARE，又被称为平均绝对百分比误差（MAPE））、根均方差（Root Mean Square Error，RMSE）和相关系数（Correlation Coefficient，CC）对模型的拟合预测效果的进行评价：

$$RE = \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \tag{3.8}$$

$$ARE(MAPE) = \frac{1}{n}\sum_{i=1}^{n} \left| \frac{Y_i - \hat{Y}_i}{Y_i} \right| \tag{3.9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{3.10}$$

在式（3.8）（3.9）（3.10）中，Y_i 表示监测值，$\hat{Y}_i$ 表示预测值。RE 代表了单点预测值与监测值的绝对误差占监测值的百分比，ARE 和 $RMSE$ 描述了模型预测的能力，值越小，模型的拟合预测精度越高。

若原始序列为 $X = \{x_1, x_2, \cdots, x_n\}$，模型的拟合序列为 $\tilde{Y} = \{\tilde{Y}_1, \tilde{Y}_2, \cdots, \tilde{Y}_n\}$，相关系数 R 描述了原始序列与拟合序列的相关程度，表示了模型的解释拟合序列的能力：

$$R = \frac{\sum_{i=1}^{n}[(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)(\tilde{Y}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i)]}{\sqrt{\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(\tilde{Y}_i - \frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i)^2}} \tag{3.11}$$

3.4 实验结果与分析

首先基于文献[34]中乌梁素海的水质数据构建本文提出的预测模型，并与引用文献中的 SVR 及 BP 方法进行比较分析，验证模型的性能。

在构建水质预测模型时，首先对数据集进行小波分解，而选择合适的小波基函数是小波分析的重要步骤。研究表明，Daubechies 小波基能较好地分析时间序列问题[49]，故选取 DbN 小波函数。进行小波分解时，不同的分解级数会影响模型构建的消耗与预测的精度。经过实验，对乌梁素海的水质时间序列 $S(t)$ 进行小波分解时，选择 Db6 小波基函数进行 2 级分解和单支重构。在图 3.4 中，序列 $S(t)$ 分解为细节子序列 D_1, D_2 和尺度子序列 A_2。分解后细节序列 D_1 为随机分量，呈现无规律的波动；序列 D_2 数量级较小，有较强的规律性，一定程度上表示原始曲线的峰值；尺度序列 A_2 波形平滑，数值较大，表示原始序列的趋势走向。

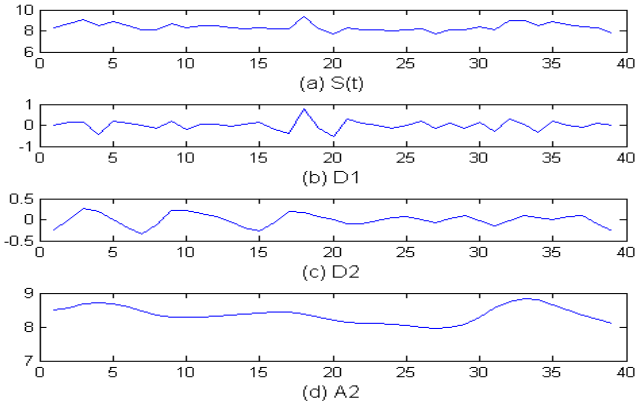

图 3.4　乌梁素海水质曲线及 Db6 小波二级分解量

分解后各序列 D_1, D_2, A_2 均与原始序列 $S(t)$ 在时间上一一对应，相应也有 $n=39$ 个值。将分解得到的细节序列 D_1, D_2 进行按照公式（3.6）进行平移变换，此时对于 D_1, D_2 的序列平移因子 Q 取值均为 1。然后，选择嵌入的维数 m 为 3，按照式（3.3）对 D_1, D_2, A_2 的序列进行相空间重构，各序列分别有 36 个 4 维样本的重构矩阵。根据公式（3.5）将序列尺度序列 A_2 的相空间重构序列进行归一化到空间 $[0,1]$。

为使模型能够预测 2001 年 5、7、10 月和 2002 年 5、7、10 月的 PH 值，选取各序列中的前 33 个值进行模型训练，余下的后 6 个值作为测试值。于是，各序列分别有 30 个 4 维样本用于作为训练集，6 个样本用于作为测试集。

采用上面处理得到的数据，结合遗传优化算法，训练支持向量回归预测模型，模型命名为 GW-SVR。为了增加模型对比，还采用同样的数据处理方法，训练得到一个 3-3-1 的三层 BP 神经网络，网络使用的隐含层为函数 tansig，输出层为 purelin 函数，训练函数为 traingdm，网络训练次数为 5000，训练目标为 0.001，模型命名为 GW-BP。用测试样本得到各子序列的预测值后，将各子序列的预测值按照与式（3.2）相同的形式进行重构得到最终序列的预测值。由 GW-SVR 和 GW-BP 算法，分别得到的拟合序列与监测值序列如图 3.5。

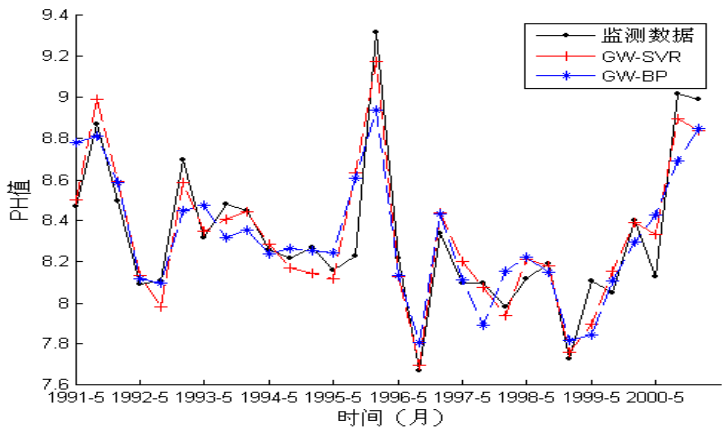

图 3.5　监测值序列和 GW-SVR 算法、GW-BP 算法的拟合序列

从对比文献中，选取最好的 SVR 及 BP 方法的统计结果与 GW-SVR 和 GW-BP 方法的结果进行对比分析，四种算法拟合值与监测值的统计分析如表 3.1 示。从表中可以得到：在相关系数比较上 GW-SVR 和 SVR 相差不大，两者的拟合序列和监测值序列均是极相关的情况；但 GW-SVR 方法的拟合平均相对误差要明显小于 SVR 方法。GW-BP 方法拟合的相关系数比原始的 BP 方法有大幅的提高，平均相对误差也都要小于 BP 方法，但是其相关系数与 GW-SVR 和 SVR 方法相比存在差距，这充分说明了基于结构风险最小化的支持向量机模型在逼近复杂非线性系统上更有优势。综合来看，GW-SVR 方法拟合效果好于原始的 SVR 方法，GW-BP 方法拟合的效果好于原始的 BP 方法。

表 3.1　四种算法的拟合值与监测值统计分析

算法	GW-SVR	SVR	GW-BP	BP
平均相对误差（%）	1.1	1.83	1.41	1.92
相关系数	0.94	0.95	0.87	0.62

GW-SVR 及 GW-BP 方法和文献中 SVR 及 BP 方法的预测结果与监测值的曲线如图 3.6 所示。

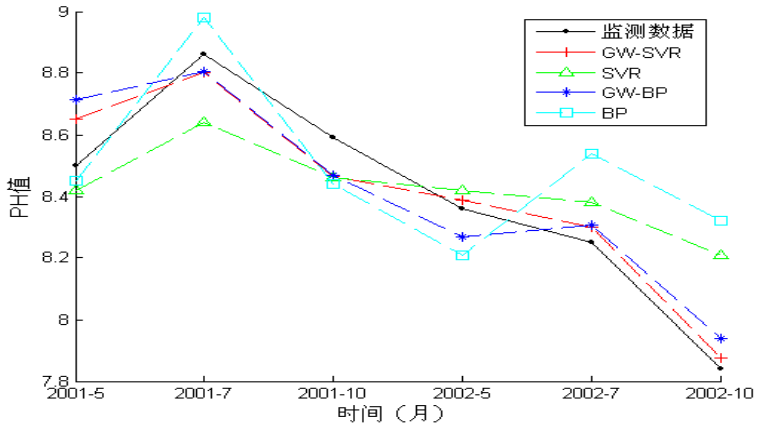

图 3.6　监测值和 GW-SVR、SVR、GW-BP 及 BP 算法预测值曲线图

GW-SVR、GW-BP、SVR 和 BP 算法的预测值与监测值的统计分析见表 3.2。从表中可以看到，GW-SVR 方法和 GW-BP 方法平均相对误差比原始两种算法预测误差要小；在监测日期 2001 年 5 月处的相对误差比较时，SVR 方法预测值的相对误差要小于 GW-SVR 方法，但是在后 5 个监测日期的监测值预测中，GW-SVR 方法预测值的相对误差都要小于 SVR 方法的，说明 GW-SVR 方法在个别点预测效果存在不足，但是总体预测效果明显优于 SVR 方法。同样，GW-BP 算法也在监测日期 2001-05 时的预测精度不及原始的 BP 方法，但是总体预测效果优于原始 BP 算法。

表 3.2　四种算法预测值与监测值的统计分析

日期	实测 PH 值	预测 PH 值							
		GW-SVR	相对误差（%）	SVR	相对误差（%）	GW-BP	相对误差（%）	BP	相对误差（%）
2001-05	8.50	8.6511	1.78	8.42	0.93	8.7139	2.52	8.45	0.62
2001-07	8.86	8.8016	0.66	8.64	2.54	8.8043	0.63	8.98	1.32
2001-10	8.59	8.4660	1.44	8.46	1.49	8.4675	1.43	8.44	1.72
2002-05	8.36	8.3892	0.35	8.42	0.73	8.267	1.11	8.21	1.78
2002-07	8.25	8.2995	0.60	8.38	1.57	8.3077	0.07	8.54	3.51
2002-10	7.84	7.8739	0.43	8.21	4.67	7.9388	1.26	8.32	6.08
平均相对误差（%）			0.88		1.99		1.17		2.51

综上所述，GW-SVR 算法的拟合及预测能力要优于使用神经网络建模的 GW-BP 方法及引用文献中的 SVR、BP 方法。虽然 GW-BP 方法的拟合及预测效果都要优于原始 BP 方法，但是由于神经网络易陷入局部极值等因素的影响，导致与支持向量机建模的方法相比存在模型拟合度不高的问题，并且从与 SVR 方法预测值的相对误差比较中可以发现，GW-BP 方法在一些点上的预测效果仍然不及 SVR 方法。可以说明，基于结构风险最小化原理的支持向量机具备更强的非线性映射的能力，在解决小样本、非线性问题上具有明显的优势，能够解决神经网络无法避免的局部极值等问题。

为进一步验证模型的预测效果，使用本文的方法与文献[35]中的方法作一步预测比较。同样采用引用文献使用的浙江嘉兴自动监测站 2008 年 51 周及 2009 年前 30 周的溶解氧监测数据。这里也采用 Db6 小波基对监测序列进行三级分解，选择嵌入维数 $m=5$，再采用数据处理公式（3.5）（3.6）对分解后的数据进行归一化和平移变换，结合遗传算法优化支持向量机参数进行模型训练。本文提出的预测模型依然用 GW-SVR 表示，与引用文献中使用支持向量机方法（用 SVM 表示）获得的最好预测结果进行比较分析。溶解氧一步预测统计分析结果见表 3.3。

从表 3.3 中对溶解氧 10 周预测值统计，本文提出的 GW-SVR 方法比原文献的 SVM 方法的平均绝对百分比误差和根均方差都要小。从图 3.7 中，GW-SVR 方法和 SVM 方法对溶解氧预测值的相对误差图可以直观看到，在 2009 年第 27 周和第 28 周处，GW-SVR 方法虽不及 SVM 方法预测的效果，但其余 8 周的预测相对误差均要比 SVM 方法小。综上所述，GW-SVR 方法要优于 SVM 方法，从而也说明了 GW-SVR 方法具有很好的推广能力。

表 3.3　溶解氧一步预测值统计分析（单位：mg/l）

日期	实测溶解氧值	预测溶解氧值			
		GW-SVR	相对误差（%）	SVM	相对误差（%）
09-21	5.09	5.0226	1.32	5.0054	1.66
09-22	4.43	4.3235	2.40	4.2979	2.98
09-23	3.59	3.6279	1.06	3.6372	1.31
09-24	3.44	3.6492	6.08	3.6847	7.11
09-25	3.24	3.3422	3.15	3.4267	5.76
09-26	2.87	2.9210	1.78	3.0185	5.17
09-27	2.88	2.7964	2.90	2.9126	1.13
09-28	3.07	2.9669	3.36	3.0048	2.12
09-29	3.18	3.3093	4.07	3.4459	8.36
09-30	3.04	3.0397	0.01	3.2368	6.47
平均绝对百分比误差（%）			2.61		4.21
根均方误差			0.1041		0.1610

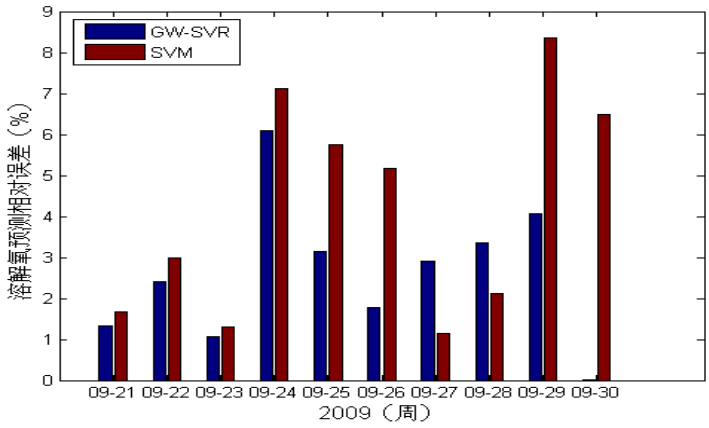

图 3.7　GW-SVR 算法、SVM 算法预测溶解氧相对误差（%）

3.5 本章小结

在水质预测模型的建模中，本文通过使用数据平移等数据处理方法，将水质时间序列变换到多个尺度上，有效的减小了瞬变数据对建模的影响；再结合遗传算法对支持向量回归机的参数进行优化，比传统的基于多次单因素实验寻找最优参数的方法更高效。

从两次实验的对比分析中可以得出以下结论：

（1）通过本文提出的新支持向量机模型与引用文献中支持向量机模型和神经网络模型对比分析，验证了该新水质预测模型具有良好的预测效果及推广能力；

（2）同时，通过向量机模型与神经网络模型的拟合及预测结果对比，说明了

支持向量机具有更好地鲁棒性和复杂非线性映射能力。

第四章 基于灰色理论及支持向量机的水质预测模型

4.1 引言

第三章基于改进小波变换及支持向量机的水质预测模型，采用小波变换数据处理方法，将水质时间序列变换到多个尺度上，有效的减小了瞬变数据对建模的影响。但是原始序列经小波变换被分解为平稳的尺度子序列和不断变化的细节子序列，将这些具有不同的特性的子序列都采用支持向量机建模的方法预测，大量输入的训练数据会影响模型的预测精度。为进一步提高模型的预测精度，将模型的特性和数据的特性充分结合起来，继续做了下面的工作。

采用灰色理论（最少 4 个数据就能建立预测模型）和支持向量机分别对小波变换分解后的平稳尺度子序列和不断变化的细节子序列建立预测模型。将该预测模型应用于乌梁素海 PH 值和浙江嘉兴自动监测站的溶解氧时间序列预测，通过与文献[36]中的预测结果进行了比较，结果表明：这种新的水质时间序列预测模型比原来的预测模型综合预测精度更高。

4.2 GM(1,1)模型

灰色系统理论是由邓聚龙教授创立[50]，其中 GM(1,1)模型是灰色系统理论中的重要内容之一，并在预测领域应用较为广泛。目前国内也有较多的应用，刘子岩等采用灰色理论对长江次级河流水质污染进行短期预测[51]；刘冬君等人将灰色系统理论与神经网络进行结合，将改进的模型应用于密云水库溶解氧的预测，验证了组合模型的效果[52]。下面对 GM(1,1)模型进行简要的介绍。

GM(1,1)模型可以用一个单变量的一阶微分方程表示：

$$\frac{dx}{dt}+ax=u \tag{4.1}$$

已知拥有 n 个时间序列数据的原始序列为：$x^{(0)}=(x^{(0)}(1),x^{(0)}(2),...,x^{(0)}(n))$，首先进行一阶累加生成新的序列：

$$x^{(1)}=(x^{(1)}(1),x^{(1)}(2),...,x^{(1)}(n)) \tag{4.2}$$

式（4.2）中 $x^{(1)}(k)=\sum_{i=1}^{k}x^{(0)}(i),\ k=1,2,...,n$。根据一阶微分方程得到估计参数向量 $\hat{\phi}=\begin{pmatrix}a\\u\end{pmatrix}$，其中 a 是发展灰数参数，u 是控制灰数参数，可以得到：

$$x^{(0)}(k+1)=a[-\frac{1}{2}(x^{(1)}(k)+x^{(1)}(k+1))]+u \quad k=1,2,...,n \tag{4.3}$$

根据方程$Y=B\hat{\phi}$，参数向量$\hat{\phi}$可用最小二乘法求取：

$$\hat{\phi}=(B^TB)^{-1}B^TY \tag{4.4}$$

式（4.4）中 $Y=\begin{pmatrix}x^{(0)}(2)\\x^{(0)}(3)\\\vdots\\x^{(0)}(n)\end{pmatrix}, B=\begin{pmatrix}-\frac{1}{2}[x^{(1)}(1)+x^{(1)}(2)] & 1\\-\frac{1}{2}[x^{(1)}(2)+x^{(1)}(3)] & 1\\\vdots & \vdots\\-\frac{1}{2}[x^{(1)}(n-1)+x^{(1)}(n)] & 1\end{pmatrix}$。微分方程的解为：

$$\hat{x}^{(1)}(k+1)=[x^{(0)}(1)-\frac{u}{a}]e^{-ak}+\frac{u}{a}\quad k=1,2,...,n \tag{4.5}$$

对$\hat{x}^{(1)}(k+1)$进行逆累加生成还原，可得到$\hat{x}^{(0)}(k+1)$预测值，即为 GM(1,1)预测模型：

$$\hat{x}^{(0)}(k+1)=\hat{x}^{(1)}(k+1)-\hat{x}^{(1)}(k)\quad k=1,2,...,n \tag{4.6}$$

4.3 水质预测模型建模

4.3.1 数据准备

由小波变换对原始的时间序列$S(t)$进行多分辨分解，得到 N 阶分解下的尺度子序列A_N，细节子序列$D_1,D_2,\cdots,D_N$。

由于细节子序列$D_1,D_2,\cdots,D_N$中数据由于存在多位小数，且存在数量级的差别，为减小小数部分的拟合预测误差、降低数量级差，将序列中的数据统一平移到正数区域。若原始序列为$D_i=\{d_1,d_2,d_3,\cdots,d_n\},i=1,2,\cdots,N$，对序列做平移处理后的序列为：

$$D_i^+=\{d_1+Q,d_2+Q,d_3+Q,\cdots,d_n+Q\}\quad Q\in N+,i=1,2,\cdots,N \tag{4.7}$$

式（4.7）中平移因子Q的值为使得序列D_i^+中的所有数据为正数的最小正整数。

细节子序列$\{x_1,x_2,x_3,\cdots,x_n\}$进行模型训练前，需根据*Takens*理论进行相空间重构，转化成矩阵的形式，获得数据之间的关联关系来得到尽可能多的信息量：

$$X=\begin{bmatrix}X_{m+1}\\X_{m+2}\\\vdots\\X_n\end{bmatrix}=\begin{bmatrix}x_1 & x_2\cdots & x_m\\x_2 & x_3\cdots & x_{m+1}\\\vdots & \vdots\ \ddots & \vdots\\x_{n-m} & x_{n-m+1}\cdots & x_{n-1}\end{bmatrix},\quad Y=\begin{Bmatrix}x_{m+1}\\x_{m+2}\\\vdots\\x_n\end{Bmatrix} \tag{4.8}$$

式（4.8）中m为重构相空间的嵌入维数。重构相空间的嵌入维数m反映重构矩阵的信息量，不同的m值对模型的预测效果有不同的影响，可以根据预测误差值的

大小来优化选取相空间的嵌入维数。

设$\hat{Y}_t=\{x_t\}$为预测的目标值，将之前的目标值$X_t=\{x_{t-m},x_{t-m+1},x_{t-m+2},\cdots,x_{t-1}\}$作为相关量，建立自相关输入$X_t$与输出$\hat{Y}_t$之间一一映射的关系：$f:R^m\to R$，使得：

$$\hat{Y}_t=f(X_t)=f(x_{t-m},x_{t-m+1},x_{t-m+2},\cdots,x_{t-1}) \tag{4.9}$$

4.3.2 模型构建

模型建立的过程如图 4.1 示，步骤为：

步骤 1　在构建水质预测模型时，首先对数据集进行小波分解，而选择合适的小波基函数是小波分析的重要步骤。研究表明，*Daubechies* 小波基能较好地分析时间序列问题[49]，故本文选取 *DbN* 小波函数。进行小波分解时，不同的分解级数会影响模型构建的消耗与预测的精度，需根据实验以最小预测误差为目标给出。若初始的水质时间序列为$S(t),t=1,2,\cdots,n$，首先将$S(t)$进行 N 阶小波分解，得到一个低频序列cA_N和 N 个高频序列$cD_1,cD_2,\cdots,cD_N$。再将分解序列进行单支重构，得到重构子序列$A_N,D_1,D_2,\cdots,D_N$。

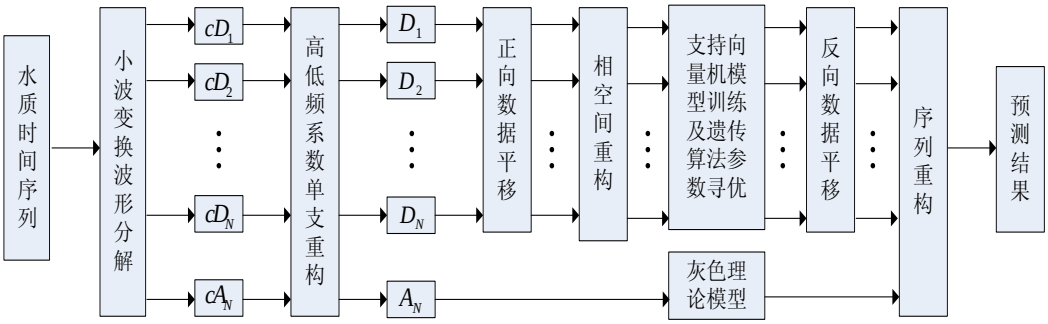

图 4.1　预测模型的建立

步骤 2　将步骤 1 得到的细节子序列$D_1,D_2,\cdots,D_N$按公式（4.7）分别平移到正数区域。再根据相空间重构理论，将其按（4.8）的形式进行相空间重构，即转化成矩阵的形式。

步骤 3　将经过相空间重构的细节子序列分别输入支持向量机模型，结合遗传算法进行参数优化，建立各细节子序列的预测模型。采用窗口滑动的形式，将尺度子序列$x_{t-m},x_{t-m+1},x_{t-m+2},\cdots,x_{t-1}$输入灰色理论模型训练，预测 t 时刻的值，其中 m 为嵌入维数，建立尺度子序列预测模型。

步骤 4　在进行预测分析时，根据步骤 1 和步骤 2 对数据进行处理，然后使用步骤 3 得到的子序列预测模型进行预测。如输入的向量为$X_t=\{x_{t-m},x_{t-m+1},x_{t-m+2},\cdots,x_{t-1}\}$时，根据式（4.6），此时可以得到第 t 时刻的尺度子序列的预测值：

$$\hat{x}^{(0)}(t)=\hat{x}^{(1)}(t)-\hat{x}^{(1)}(t-1)\quad t=5,6...,n \tag{4.10}$$

根据式（2.19）的最优回归函数，此时可以得到第t时刻的细节子序列的预测值：

$$\hat{Y}_t=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(X_t,X_i)+b \quad t=m+1,m+2,\cdots,n+1 \tag{4.11}$$

步骤 5　最后，将细节子序列进行反向平移，将尺度序列模型及细节序列模型得到预测的值根据公式（3.2）进行重构，最终得到预测结果。

4.4 模型性能评价标准

本文选取相对误差（Relative Error，RE）、平均相对误差（Average Relative Error，ARE，又被称为平均绝对百分比误差（MAPE））、根均方差（Root Mean Square Error，RMSE）和相关系数（Correlation Coefficient，CC）对模型的拟合预测效果的进行评价：

$$RE=\left|\frac{Y_i-\hat{Y}_i}{Y_i}\right| \tag{4.12}$$

$$ARE(MAPE)=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i-\hat{Y}_i}{Y_i}\right| \tag{4.13}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2} \tag{4.14}$$

在式（4.12）（4.13）（4.14）中，Y_i表示监测值，$\hat{Y}_i$表示预测值。RE代表了单点预测值与监测值的绝对误差占监测值的百分比，ARE和$RMSE$描述了模型预测的能力，值越小，模型的拟合预测精度越高。

若原始序列为$X=\{x_1,x_2,\cdots,x_n\}$，模型的拟合序列为$\tilde{Y}=\{\tilde{Y}_1,\tilde{Y}_2,\cdots,\tilde{Y}_n\}$，相关系数$R$描述了原始序列与拟合序列的相关程度，表示了模型的解释拟合序列的能力：

$$R=\frac{\sum_{i=1}^{n}[(x_i-\frac{1}{n}\sum_{i=1}^{n}x_i)(\tilde{Y}_i-\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i)]}{\sqrt{\sum_{i=1}^{n}(x_i-\frac{1}{n}\sum_{i=1}^{n}x_i)^2}\cdot\sqrt{\sum_{i=1}^{n}(\tilde{Y}_i-\frac{1}{n}\sum_{i=1}^{n}\tilde{Y}_i)^2}} \tag{4.15}$$

4.5 实验结果与分析

基于文献[34]中乌梁素海 PH 值时间序列的水质数据，构建本文提出的预测模型，并将其拟合及预测结果与文献[36]中的 GW-SVR 模型和 GW-BP 模型进行比较，验证模型的性能。

首先对乌梁素海 PH 值时间序列$S(t)$进行小波分解时，经过多次实验，在选择$Db16$小波基进行 4 级分解和重构时可以取得最小的误差。时间序列$S(t)$被分解为尺度序列A_4和细节序列D_1,D_2,D_3,D_4，各子序列均与原始序列在时间上一一对应，

如图 4.2。根据公式（4.7）平移因子$Q=1$，将细节系数D_1,D_2,D_3,D_4先平移到正数区域，再根据公式（4.8）嵌入维数$m=3$，进行相空间重构得到一个 36×4 的矩阵。将该矩阵的前 30 行 4 列作为训练集，剩下的 6 行 4 列做为测试集。结合遗传算法对支持向量机进行参数优化，取得预测结果。根据公式（4.9）嵌入维数$m=4$，使用 GM(1,1)来建立尺度序列A_4的预测模型。根据公式（3.2）组合尺度子序列和细节子序列的拟合及预测结果，该模型命名为 GW-GMSVR。从对比文献中，选取 GW-SVR 及 GW-BP 模型的拟合结果与本文 GW-GMSVR 模型的拟合结果，如图 4.3.

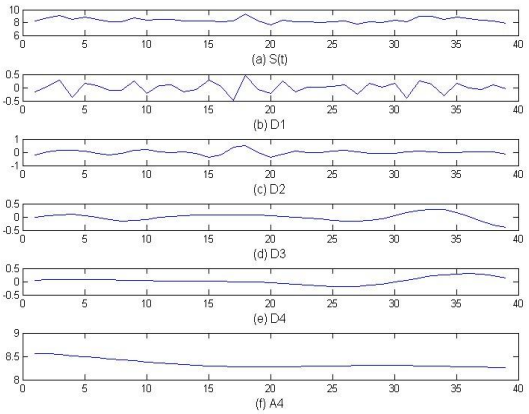

图 4.2　PH 值曲线及 Db16 小波 4 级分解量

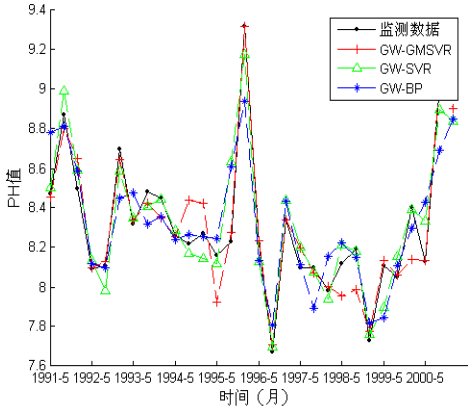

图 4.3　监测值序列和 3 种模型拟合序列

三种模型拟合值与监测值的统计分析如表 4.1 示。从表中可以得到：GW-GMSVR 和 GW-SVR 模型的相关系数值（CC）表示两者的拟合序列和监测值序列均是极相关的情况，但是前者比后者具有更强的相关性；同时，GW-GWSVR 与 GW-SVR 模型的平均相对误差（ARE）相比，明显更小；并且，GW-BP 与 GW-GWSVR 和 GW-SVR 模型相比，相关系数和平均相对误差存在差距，这充分说明了基于结构风险最小化的支持向量机模型在逼近复杂非线性系统上更有优势。综合来看，GW-GMSVR 模型优于原始的 GW-SVR 及 GW-BP 模型。

表 4.1　3 种模型的拟合值与监测值统计分析

模型	GW-GMSVR	GW-SVR	GW-BP
平均相对误差(%)	0.88	1.1	1.41
相关系数	0.96	0.94	0.87

三种模型预测值与监测值的统计分析见表 4.2。从表中可以看到，GW-GMSVR 模型的平均相对误差（ARE）显著小于 GW-SVR 及 GW-BP 模型的。首先比较 GW-GMSVR 和 GW-SVR 模型，在 2001-07，2002-05 和 2002-10 三个监测日期上，GW-GMSVR 的相对误差比 GW-SVR 高出 0.08，0.31 和 0.06，而在其它的三个监测日期上，GW-GMSVR 的相对误差比 GW-SVR 小 1.59，0.87 和 0.29，这说明

GW-GMSVR 模型在个别点预测效果存在不足，但是总体预测效果明显优于 GW-SVR 模型；同样，在监测日期 2001-07 和 2002-07 上，GW-GMSVR 比 GW-BP 的相对误差高，但是在其他四个监测日期的预测中，GW-GMSVR 比 GW-BP 的相对误差要低；GW-BP 与 GW-GWSVR、GW-SVR 模型相比，其平均相对误差要大于后两者，这也说明了支持向量机模型在逼近复杂非线性系统上更有优势。

表 4.2　3 种模型预测值与监测值的统计分析

时间	PH 监测值	PH 预测值					
		GW-GMSVR	相对误差（%）	GW-SVR	相对误差（%）	GW-BP	相对误差（%）
2001-05	8.50	8.5165	0.19	8.6511	1.78	8.7139	2.52
2001-07	8.86	8.7944	0.74	8.8016	0.66	8.8043	0.63
2001-10	8.59	8.6388	0.57	8.4660	1.44	8.4675	1.43
2002-05	8.36	8.4155	0.66	8.3892	0.35	8.267	1.11
2002-07	8.25	8.2161	0.41	8.2995	0.60	8.3077	0.07
2002-10	7.84	7.8783	0.49	7.8739	0.43	7.9388	1.26
平均相对误差（%）			0.51		0.88		1.17

综上所述，GW-GMSVR 模型的拟合及预测能力要优于引用文献中的 GW-SVR、GW-BP 模型。由于神经网络易陷入局部极值等因素的影响，导致与支持向量机模型相比存在模型拟合度不高的问题，并且从与支持向量机模型预测值的相对误差比较中可以发现，GW-BP 模型在一些点上的预测效果仍然不及 SVR 模型。可以说明，基于结构风险最小化原理的支持向量机具备更强的非线性映射的能力，在解决小样本、非线性问题上具有明显的优势，能够解决神经网络无法避免的局部极值等问题。

为进一步验证模型的预测效果，我们将本文的预测方法与文献[36]中采用改进小波变换及支持向量机的方法进行一步预测结果的比较。同样采用引用文献使用的浙江嘉兴自动监测站 2008 年 51 周及 2009 年前 30 周的溶解氧监测数据[35]。这里采用 Db20 小波基对溶解氧序列进行四级分解，首先使用 GM(1,1)模型去拟合预测尺度子序列，根据公式（4.9），选择嵌入维数为 m=4；其次，基于小波变换及数据平移的基础上，使用 SVM 拟合预测细节序列，选择嵌入维数为 m=3，然后结合遗传算法优化支持向量机参数，训练得到一个新的支持向量预测模型，表示为 GW-GMSVR。将该模型的预测结果与文献中 GW-SVR 获得的最好预测结果进行比较分析。溶解氧一步预测统计分析见表 4.3。

从表 4.3 中可以看到，GW-GMSVR 方法比原文献的 GW-SVR 方法的的 ARE 及 RMSE 值都要小。从表中 GW-GMSVR 方法和 GW-SVR 方法的参数预测相对误

差可以看出，在 2009 年 21 周、28 周和 30 周处，GW-GMSVR 方法虽不及 GW-SVR 方法预测的效果，但其余 7 周的预测相对误差均要比 GW-SVR 方法小。综上所述，GW-GMSVR 方法要优于引用文献中使用的 GW-SVR 方法，从而也说明了 GW-GMSVR 方法具有很好的推广能力。

表 4.3　溶解氧一步预测值统计分析表（单位：mg/l）

日期	实测溶解氧值	预测溶解氧值			
		GW-GMSVR	相对误差(%)	GW-SVM	相对误差（%）
09-21	5.09	5.0138	1.50	5.0226	1.32
09-22	4.43	4.3368	2.10	4.3235	2.4
09-23	3.59	3.585	0.14	3.6279	1.06
09-24	3.44	3.5015	1.79	3.6492	6.08
09-25	3.24	3.2917	1.60	3.3422	3.15
09-26	2.87	2.8759	0.20	2.921	1.78
09-27	2.88	2.8858	0.20	2.7964	2.9
09-28	3.07	3.1906	3.93	2.9669	3.36
09-29	3.18	3.2016	0.68	3.3093	4.07
09-30	3.04	2.8956	4.75	3.0397	0.01
平均绝对百分比误差（%）			1.69		2.61
根均方误差			0.0754		0.1041

4.6 本章小结

监测时间序列经小波变换分解为平稳的尺度子序列和不断变化的细节子序列，灰色理论模型对平稳序列具有较好的预测能力，并且支持向量机模型具有复杂非线性映射的能力，本文提出的水质预测模型使用灰色理论建立尺度子序列预测模型，使用支持向量机建立细节子序列预测模型。将该预测模型应用于乌梁素海 PH 值时间序列预测及浙江嘉兴自动监测站溶解氧预测，通过与原来的预测模型比较，结果表明：本文提出的这种新的水质时间序列预测模型比原来的支持向量机和 BP 神经网络预测模型的预测精度更高、性能更优越。

第五章 总结与展望

5.1 总结

水环境中水质数据具有趋势性的特点，对其预测和模拟，是为决策者提供科学依据进行水环境的综合规划及管理、水污染的防治和治理不可缺少的一项基础性工作。随着水环境科学和其他相关科学不断发展，水质时间序列的预测逐渐受到了重视，大量的预测模型不断的涌现，总结起来主要分为数理统计、灰色理论、神经网络模型、水质模拟模型和混沌理论等预测模型，这些模型在水质预测研究中已经取得了一些重要的成果，但是其自身仍然不够完善。支持向量机是在统计学习理论的基础上，实现结构风险最小化准则的一种新的机器学习方法，在解决小样本、非线性问题、高维模式识别和局部极小等方面具有优势。为了解决基于支持向量机进行水质时间序列预测存在的一些问题。

首先，本文采用小波变换将水质时间序列变换到多个尺度上，以减小瞬变数据对建模的影响，并在此基础上，进一步利用分解后的水质数据特点，构建新的预测模型，以降低突变数据对模型构建的影响，做了如下工作：

（1）采用遗传算法优化支持向量机的参数

针对支持向量机的单因素参数寻优非常耗时，而且常常不能保证找到的参数是最优的问题，采用遗传遗传算法优化支持向量机的参数。遗传算法是以自然选择和遗传理论为基础，模拟了生物进化过程中适者生存规则与群体内部染色体的随机信息交换机制，遗传算法对支持向量机的参数进行优化，比传统的基于多次单因素实验寻找最优参数的方法更高效。

（2）将小波分解后的细节系数进行平移

针对小波变换一定程度上能够降低跳变值的影响，但是水质时间序列经过小波分解得到的细节序列某些数据存在多位小数，数据之间存在数量级的差别，导致在模型训练过程中常常得不到理想的模型问题。为了减小小数部分输入模型时产生的拟合误差，需要降低数据之间的数量级差，所以将细节序列中的数据统一平移到正数区域。

将上面的方法应用到新的支持向量机预测模型中，将其用于乌梁素海的水质数据预测，并与基于多次单因素实验优化支持向量机参数模型和基于小波变换的 BP 神经网络模型进行比较分析；再与基于小波变换支持向量机的水质预测模型进行了对比分析。结果表明，这种新的水质时间序列预测模型比传统的支持向量机和 BP 神经网络预测模型，精度更高、性能更优越。

其次，前述的水质预测模型采用小波变换数据处理方法，将水质时间序列变换到多个尺度上，有效的减小了瞬变数据对建模的影响，但是原始序列经小波变换被分解为平稳的尺度子序列和不断变化的细节子序列，将这些具有不同的特性的子序列都采用支持向量机建模的方法预测，大量输入的训练数据会影响模型的预测精度。为进一步提高模型的预测精度，将模型的特性和数据的特性充分结合起来，做了下面的工作：

（3）采用灰色理论对平稳尺度子序列建模

采用灰色理论（最少 4 个数据就能建立预测模型）和支持向量机分别对小波变换分解后的平稳尺度子序列和不断变化的细节子序列建立预测模型。将该预测模型应用于乌梁素海 PH 值和浙江嘉兴自动监测站的溶解氧时间序列预测，通过与文献中的预测结果进行了比较，结果表明：这种新的水质时间序列预测模型比原来的预测模型综合预测精度更高。

5.2 展望

水环境系统受到自然因素及人类活动的影响，是一个不断变化的综合系统，在地域、时空上显示出不断变化的动态特性。因此，流域中的水质状况便受制于这样一个复杂的系统，表现出有趋势性、阶段性、季节性和突变性等特征，研究人员对水质进行预测便变得尤为困难，如何进行高精度的预测建模是今后很长一段时间内主要目标和核心内容。目前针对不同的流域、水质特性，国内外研究者提出了大量的预测方法，成果丰硕。本文采用支持向量机进行水质预测，在支持向量机参数选择和水质数据特性分析方面进行了研究和探索，并取得了一些初步的研究成果，但是由于时间和水平所限，很多工作还需不断改进和完善，具体体现在以下几个方面：

（1）在支持向量机的回归理论中，目前已有较多的支持向量回归模型，本文采用ε-支持向量回归机，应用到水质时间序列的预测建模中，得到了较好的预测精度，但是对支持向量机模型的特性研究还存在不足，还可以进一步研究。

（2）本文采用支持向量机模型与其他的方法相结合的方式，建立水质预测模型，在采用训练数据进行支持向量机模型训练的过程中，需要找到合适的支持向量机参数，使用遗传算法进行参数优化提高了准确性和效率，但遗传算法初始种群的不同，常常对参数寻优结果有着重要的影响，并且目前还有其他的一些参数优化方法可供参考，寻找到准确的模型参数是一项重要而艰巨的任务。

致 谢

时光荏苒，硕士研究生的学习生活伴随着校园里四月盛开的樱花渐近尾声。三年的学习生活使我受益匪浅，在学习知识、提升技能的同时，结识了一群良师益友。一幕幕，回忆起来历历在目，心存不舍与感激。

首先，非常感谢我的研究生导师王国胤教授。王老师严谨、严格的治学态度，积极、探索的治学精神使我明白：在做科研方面没有任何的投机取巧，唯有自己的坚持不懈。在过去三年中，王老师悉心地教导我们，关心大家的学习和生活、与大家探讨学术问题。王老师还教会我们身体是革命的本钱，要注意劳逸结合。

同时，本论文的完成还离不开对我们悉心指导的书记曾跃老师，办公室主任曾立梅老师，研究生辅导员赵莉莉老师、夏淑芳老师，科研秘书许可老师。五位老师不论是在生活中，还是在工作上都给予了我很大的关心和帮助，也是我能顺利完成学位论文的动力之一。

感谢中国科学院重庆绿色智能技术研究院的傅剑宇博士、利节博士、吴迪师兄、周博天博士、董建华博士，对我在项目和学习上的指导和帮助。感谢在联合培养两年学习期间，一起共同学习和生活的邹轩、李鸿、刘永福、赵健、曹磊、苏丹等同学，感谢罗强、凡少强、王金芝、张艳桃、贺海鹏等实验室的同门同学，同时还要感谢我的室友刘森森、林炎钦，正是在他们的陪伴下，我愉快的度过了三年的研究生生活。衷心的感谢你们！

另外，我还要感谢我的家人，是他们一直鼓励和支持着我。

最后，我还要特别感谢重庆邮电大学研究生院、计算机学院的所有领导和老师，是他们为我们研究生的培养工作付出了数不清的艰辛和汗水。衷心的感谢你们！

感谢参与评审和答辩以及所有为本文提出宝贵意见的各位老师和专家！

签名：刘文

签字日期：2014 年 4 月 21 日

攻硕期间从事的科研工作及取得的研究成果

- **从事的主要科研工作**

[1] 三峡库区水生态环境感知系统及平台业务化运行，国家水体污染控制与治理科技重大专项，课题号: 2014ZX07104-006.

[2] 不确定性概念内涵与外延的双向认知计算理论模型与方法，国家自然科学基金项目，项目编号:61272060.

- **科研成果**

[1] W. Liu, G. Y. Wang, J. Y. Fu, X. Zou. Water Quality Prediction Based on Improved Wavelet Transformation and Support Vector Machine[A]. // The 2013/2nd International Conference on Energy and Environmental Protection[C]. Trans Tech Publications Inc, 2013, 726-731:3547-3553.

[2] 刘文，王国胤，宋应文，傅剑宇，利节. 基于灰色理论及支持向量机的水质预测[A]. // 2013 年水资源生态保护与水污染控制研讨会[C]. 中国环境科学学会编, 2013:604-609.

[3] 刘文，王国胤，傅剑宇，苟光磊，李鸿，邹轩. 一种基于灰色理论及支持向量机的水质预测的方法[P]. 中国专利:201310658980.X.

[4] H. Li, G. Y. Wang, G. L. Gou, W. Liu. Boundary variable precision dominance-based rough set approach in multicriteria sorting problems[A]. // The 2013/2nd International Conference on Energy and Environmental Protection[C]. Trans Tech Publications Inc, 2013, 734-737:3102-3106.

- **联合培养经历**

2012.7-2014.6 中国科学院重庆绿色智能技术研究院 数据挖掘与认知中心

- **会议报告**

[1] 在2013/2nd International Conference on Energy and Environmental Protection (ICEEP 2013)，作分组报告“Water Quality Prediction Based on Improved Wavelet Transformation and Support Vector Machine”.

[2] 在2013年水资源生态保护与水污染控制研讨会，作分组报告“感知智慧三峡的魅力”.

参考文献

[1] 国家环保部. 2012 年中国环境状况公报-淡水环境[EB/OL]. http://jcs.mep.gov.cn /hjzl/zkgb /2012zkgb/201306/t20130606_253418.htm.

[2] 孙国红, 沈跃, 徐应明, 等. 基于 Box_Jenkins 方法的黄河水质时间序列分析与预测[J]. 农业环境科学学报, 2011, 30(9):1888-1895

[3] 武国正.支持向量机在湖泊富营养化评价及水质预测中的应用研究[D]. 呼和浩特:内蒙古农业大学, 2008.

[4] 李如忠. 水质预测模式理论研究进展与趋势分析[J]. 合肥工业大学学报, 2006, 29(1):26-30.

[5] 史根香, 郭海生. 指数平滑法在地下水水质预测中的尝试[J]. 湖北地矿, 1998, 12(1):35-40.

[6] 唐宗鑫, 简文彬. 闽江下游水质预测的时间序列模型[J]. 水利科技, 2002, (2): 7-9.

[7] 李博之. 鄱阳湖水体污染现状与水质预测、规划研究[J]. 长江流域资源与环境, 1996, 5(1): 60-66.

[8] 刘倩纯, 余潮, 张杰, 等. 鄱阳湖水体水质变化特性分析[J]. 农业环境科学学报, 2013, 32(6):1232-1237.

[9] 张军. 灰色预测模型的改进及其应用[D]. 西安:西安理工大学, 2008.

[10] 杨华龙, 刘金霞, 郑斌. 灰色预测 GM(1,1)模型的改进及应用[J]. 数学实践与认识, 2011, 41(23):40-46.

[11] 刘东君, 邹志红. 灰色和神经网络组合模型在水质预测中的应用[J]. 系统工程, 2011, 29(9):105-109.

[12] H. E. Jobson. Predicting Travel Time and Dispersion in Rivers and Streams[J]. Journal of Hydraulic Engineering, 1997, 123(11): 971-977.

[13] J. M. Zaldívar, E. Gutiérrez, I. M. Galván, etc. Forecasting High Waters at Venice Lagoon Using Chaotic Time Series Analysis and Nonlinear Neural Networks[J]. Journal of Hydroinformatics, 2000, 2:61-84.

[14] V. Vapnik, A. Y. Chervoknenkis. The Necessary and Sufficient Conditions for the Uniform Convergence of Averages to Their Expected Values[J]. Teoriya Veroyatnostei I ee Primeniniya, 1981, 26(3):543-564.

[15] V. Vapnik, S. Kotz. Estimation of Dependences Based on Empirical Data[M]. New York: Springer, 1982.

[16] V. Vapnik. The Nature of Statistical Learning Theory[M]. New York: Springer-Verlag, 1995.

[17] V. Vapnik, S. Gokowich and A. Smola. Adavances in Neural Information Processing Systems 9[M]. Massachusetts:MIT Press, 1997:281-287.

[18] 徐红敏. 基于支持向量机理论的水环境质量预测与评价方法研究[D]. 吉林:吉林大学, 2007.

[19] O. L. Mangasarian. Advances in Large Margin Classifiers[M]. Massachusetts:MIT Press, 2000:135-146.

[20] 谢赛琴, 沈福明, 邱雪娜. 基于支持向量机的人脸识别方法[J]. 计算机工程, 2009, 35(16):186-188.

[21] 胡良谋, 曹克强, 徐浩军. 基于回归型支持向量机的液压舵机故障诊断[J]. 系统仿真学报, 2007, 19(23), 5509-5512.

[22] C. C. Chang, C. J. Lin. Training V-support Vector Classifiers: Theory and Algorithms [J], Neural Computation, 2001, 13(9):2119-2147.

[23] P. Laskov. Feasible Direction Decomposition Algorithms for Training Support Vector [J]. Machine Learning, 2002, 46(1):315-349.

[24] 朱家元, 陈开陶, 张恒喜. 最小二乘支持向量机算法研究[J]. 计算机科学, 2003, 30(7):157-159.

[25] 邓燕. 基于粗糙集—支持向量机的油气储层参数预测方法研究[D]. 北京:中国地质大学, 2013.

[26] 陆宇振, 杜昌文, 余长兵, 等. 红外光声光谱技术结合支持向量机鉴别油菜籽品种[J]. 计算机应用与化学, 2014, 31(1):117-120.

[27] 郑一华. 基于支持向量机的水质评价和预测研究[D]. 江苏:河海大学, 2006.

[28] 梁勇. 基于支持向量回归的水质预测研究[D]. 武汉:武汉理工大学, 2012.

[29] 梁雪春, 龚彦冰, 肖迪. 一种多核加权支持向量机的水质预测方法[J]. 东南大学学报, 2011, 41:14-17.

[30] 徐红敏, 王继广. 加权支持向量回归机及其在水质预测中的应用[J]. 世界地质, 2007, 26(1):58-61.

[31] U. Thissen, R. van Brakel， A. P. de Weijer, etc. Using Support Vector Machines for Time Series Prediction[J]. Chemometrics and Intelligent Laboratory Systems, 2003, 69:35-49.

[32] G. H. Tan, J. Z. Yan, C. Gao, etc. Prediction Water Quality Times Series Data Based on Least Squares Support Vector[J]. Proeedia Enihineering, 2012, 31:1194-1199.

[33] 杨虞微, 左洪福, 陈果. 支持向量机时间序列预测模型的参数影响分析与自适

应优化[J]. 航空动力学报, 2006, 21(4):767-772.

[34]武国正, 徐宗学, 李畅游. 支持向量回归机在水质预测中的应用与验证[J]. 中国农村水利水电, 2012, 1:25-29.

[35]T. N. He, P. J. Chen. Prediction of Water-quality Based on Wavelet Transform Using Vector Machine[A]. // International Symposium on Distributed Computing and Applications to Business, Engineering and Science[C]. IEEE, 2010:76-81.

[36]W. Liu, G. Y. Wang, J. Y. Fu, X. Zou. Water Quality Prediction Based on Improved Wavelet Transformation and Support Vector Machine[A]. // The 2013/2nd International Conference on Energy and Environmental Protection[C]. Trans Tech Publications Inc, 2013, 726-731:3547-3553.

[37]T. M. Mitchell. 曾华军译. 机器学习[M]. 北京:机械工业出版社,2008.

[38]V. Vapnik. 许建华, 张学工译. 统计学习理论[M]. 北京:电子工业出版社, 2004.

[39]黄华江. 实用化工计算机模拟-MATLAB 在化学工程中的应用汇[M]. 北京:化学工业出版社, 2004:273-296.

[40]袁玉萍, 陈庆华, 汪洪艳. 关于支持向量机 VC 维问题证明的研究[J]. 农业与技术, 2006, 26(4): 210-211.

[41]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.

[42]梁坚. 支持向量机在水质评价及预测中的应用研究[D]. 浙江:浙江工业大学, 2009.

[43]杨志民, 刘广利. 不确定性支持向量机:算法及应用[M]. 北京:科学出版社, 2012.

[44]海娜, 张葛祥, 金炜东. 系统辨识中支持向量机核函数及其参数的研究[J]. 系统仿真学报, 2006, 18(11):3204-3208.

[45]K. P. Wu, S. D. Wang. Choosing the Kernel Parameters for Support Vector Machine by the Inter-cluster Distance in the Feature Space[J]. Pattern Recognition, 2009, 42(5):710-717.

[46]雷英杰, 张善文. MATLAB 遗传算法工具箱及应用[M]. 西安：西安电子科技大学出版社, 2005:34-36.

[47]C. W. Hsu, C. C. Chang, C. J. Lin. A Practical Guide to Support Vector Classification[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/papers/ guide/ guide.pdf

[48]S. Mallat. A Theory for Multiresolution Signal Decomposition:the Wavelet Representation[J]. IEEE Pattern Analysis and Machine Intelligence, 1989, 11(7):674-693.

[49]周伟. MATLAB 小波分析高级技术[M]. 西安: 西安电子科技大学出版社,

2006:51~ 63.

[50] 邓聚龙. 灰色理论系统[M]. 武汉:华中科技大学出版社, 2002.

[51] 刘子岩, 罗固源, 吕青峰. 灰色模型理论对长江次级河流水质污染的短期预测模拟[J]. 三峡环境与生态, 2011, 33(3):47-50.

[52] 刘东君, 邹志红. 灰色和神经网络组合模型在水质预测中的应用[J]. 系统工程, 2011, 29(9):105-109.