第三届“泰迪杯”

全国大学生数据挖掘竞赛

优

秀

作

品

作品名称：城市供水处理混凝投药过程的建模与控制

荣获奖项：一等奖

作品单位：东北大学秦皇岛分校

作品成员：李起旺　唐鑫桂

指导教师：

基于非线性预测建模的混凝投药过程控制

摘　要：

针对水厂混凝投药过程，本文根据广东省某水厂历史数据，建立了进水流量、浊度、PH值、加药量和沉淀池出水浊度之间的四个阶段性数学模型，并对模型进行优化最终建立了最佳投药量的闭环预测系统，实现对投药量的实时控制。

第一阶段：忽略PH对投药量的影响，综合已有文献[3]中的投药量与各因素间的指数关系，建立了基于非线性回归辨识指数模型，并采用取对数方法将其转换为线性回归问题，最终求得出水浊度与投药量、原水浊度及取水量之间的函数表达式（可信度为99%），并通过式（2-5）计算反应到沉淀结束的时间为90min。

第二阶段：考虑时间的滞后性，本文建立了基于改进的BP神经网络的非线性黑箱投药控制模型，将原水PH、原水浊度、取水量以及投药量作为输入参数，出水浊度作为输出参数。通过设置学习参数和动量系数，并控制出水浊度达到标准值，最终确定了4-12-1最佳的神经网络结构，其泛化误差为0.0014。运用此模型，计算了在不同原水条件下的最佳投药量，具体可见表2-6；并分析了原水PH、原水浊度及取水量对最佳投药量的影响，可见图2-7、2-8、2-9。

第三阶段：鉴于混凝投药过程是一个反馈控制过程，在投药时混凝剂量不仅与控制本次出水浊度有关，还与上次的出水浊度相关，模型二只是通过控制出水浊度来寻找最佳的投药量的正向反馈。为使模型更有效，与实际情况更接近，本文在模型二基础上，增加出水浊度作为输入量，投药量作为输出量建立新的BP神经网络模型，并通过样本数据的训练和误差控制确定了最佳网络结构。出水浊度作为反馈参数实现对最佳投药量的反馈控制：以1.0NTU为临界点，大于1.0NTU会使投加量增加，且越大增加越多，小于1.0NTU会使投加量减少，且越小减少越多，具体可见表2-7和图2-12.

第四阶段：由于温度会影响分子的布朗运动和水的动力粘度，最终会影响到最佳PAC量的投加，故需要将温度也作为输入参数，研究温度对最佳PAC量投加影响，其模型是在模型三基础上增加温度作为输入量构建新的神经网络模型，从气象局搜集到2013年8月22日到2014年9月5日的气温数据，并利用文献[11]气温与水温函数关系计算得日均水温数据，同时将附件数据集预处理后的数据进行日均化后与温度数据进行标准化处理后作为输入参数进行网络训练，最终确定较优的神经网络结构，在10-25^{0}C时最佳投药量几乎不变，而在温度升高投药量降低；反之升高，具体可见图2-17.

最后，本文根据 BP 神经网络的缺陷，设计了基于 RBF 神经网络结构的投药控制模型，并基于之前的模型最终设计出最佳投药量的闭环预测系统（图 2-20），实现对投药量在水发生变化情况下的实时控制。

关键词：混凝投药，非线性预测，回归模型，BP 神经网络，RBF 神经网络

Process Control of Coagulant Dosage Based on the Nonlinear Predictive Modeling

Abstract:

As for coagulant dosing process In a water industry, this paper applies the historical data of the waterworks in Guangdong Province to establish the four stages of mathematical model about influent flow rate, turbidity, pH, dosage and effluent turbidity, and further optimizes the model and finally establish the optimal dosage of closed-loop prediction system to realize the real-time control of coagulant dosage.

The first stage: Neglecting the influence of pH on the dosage and mirroring dosage among various factors and the exponential relationship in literatures [3] , this paper establishes the identification index model based on nonlinear regression, and transforms it into a linear regression problem with the logarithmic method, and get the final function between these variables, and reaction precipitation over time is obtained 90min by formula 2-5 .

The second stage: Considering the time lag, this paper establishes the nonlinear black box dosing control model based on Improved BP neural network , where the raw water pH, turbidity, water and dosage are input parameters, the effluent turbidity is output parameter. By setting the learning parameters and momentum coefficient, and controlling the turbidity of the effluent to reach the standard value, the optimal neural network structure of 4-12-1 is determined, and its generalization error is 0.0014. Using this model, the optimal dosage in different raw water conditions were calculated, Tab2-6 and Fig 2-7/8/9 can show you .

The third stage: Considering the process of coagulant dosage is a feedback control process, the dosage of PAC is not only controlled by the effluent turbidity but also the effluent turbidity. Model 2 finds the best dosage just form positive feedback by controlling effluent turbidity. In order to make the model more effective and more close to the actual situation, this paper increases turbidity as another input based on the second model, the dosage as output, to establish new model based on BP neural network, and determined the optimal network structure through training and error control sample data. The detailed optimal dosage changes under different raw water conditions can be seen Tab 2-7 and Fig 2-12.

The fourth stage: Because the dynamic viscosity temperature will affect the molecular Brown movement and water, eventually affect the best amount of PAC. There is a need to consider the temperature as the input parameter, and study the effects of temperature on the optimum PAC. This model is based on the third model by increasing temperature as input and constructs a new neural network model. The temperature data from August 22, 2013 to September 5, 2014 are collected from the Meteorological Bureau, and gets the daily water temperature through the relationship between air temperature and water temperature in literature [11], and normalizes the attachment data set after pretreatment and the temperature data as the input parameters of the network training, and ultimately determines the neural network an optimal structure. The optimal dosage doesn't change between 10^0C and 25^0C,which can be shown in Fig 2-17.

In the end, according to the defects of BP neural network, this paper builds a control model based on the administration of the structure of RBF neural network, and design the optimum dosage of closed-loop prediction system (Fig 2-20), implementation of coagulant dosage in water with changes in the real time .

Key words: Coagulant Dosage, Nonlinear Prediction, Regression Model, BP Neural Network, RBF Neural Network

目 录

1. 挖掘目标

在水资源污染日益严重的今日，如何有效地对水进行净化处理，成为了当今国内外学者研究的热点问题。制水过程中混凝剂投加量决定了自来水质量和制水能耗，研究混凝过程投药控制以及对整个制水过程的优化，有助于企业降低能耗，提高供水企业的经济效益和社会效益。

本次数据挖掘建模主要针对广东某水厂投药控制系统实时采集的数据信息，分析辨识原水水质、出水流量、出水浊度、药物投加量及温度等的相关关系，设计一个对混凝剂投药量进行实时控制系统模型，从而为水厂受不同条件影响时选择最佳的投药量，本题数据挖掘的具体目标如下：

- 分析附件中变量间的数据关系，采用数理统计方法建立数学模型，求解原水添加混凝剂反应到沉淀结束出水所需时间。
- 考虑问题一时间滞后性，将历史原水水质、流量及混凝剂投加量作为模型输入参数，建立有效的数学模型，输出最佳混凝剂投药量结果。
- 在问题二基础上，增加输入参数，即增加沉淀池浊度，修改数学模型控制系统，输出新的最佳混凝剂投药量，实现对投药量的反馈调节。
- 鉴于温度影响化学反映速度，在问题三模型控制系统引入温度作为输入参数，通过仿真实验分析温度对输出结果的影响。

2. 分析方法与过程

2.1. 总体流程

本文对混凝沉淀过程的投药数学模型进行了深入分析和研究，采用机理法和智能方法来确定投药数学模型，建立了基于 RBF 神经网络的最佳投药量的闭环预测系统，具体建模的步骤如下所示：

步骤一：针对问题进行过程分析、因素分析，确定整体思路

步骤二：根据前面的因素分析，从原始数据集中选取建模数据

步骤三：采用合适的方法对抽取的建模数据进行预处理

步骤四：建立数学模型进行问题求解，并针对模型缺陷进行优化改进

步骤五：针对实验结果进行分析

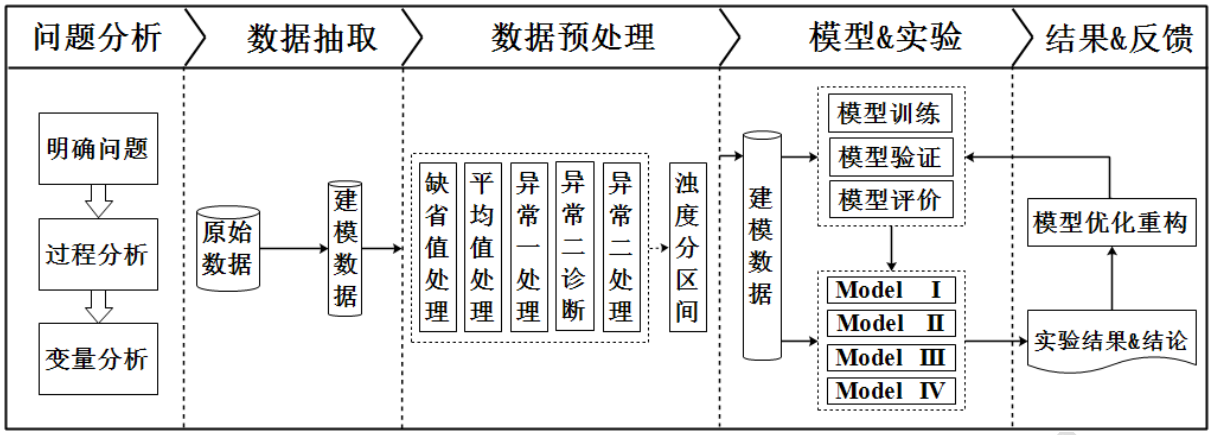

图 2-1 总体流程图

2.2. 具体步骤

2.2.1 问题分析

问题分析旨在全面地理解问题，通过查找文献资料，明确问题细节，梳理解题思路，主要包括明确问题、过程分析及变量分析三部分。

➢ 明确问题

本题属于化学工艺过程控制问题，目的在于控制在不同条件下进行絮凝时混凝剂的投加量。而影响絮凝效果的因素很多，包括取水量、原水水质、原水温度、混凝剂投加量等，投药控制即综合考虑这些因素作混凝剂最少最经济投加，而达到最优絮凝效果。

➢ 过程分析

水处理混凝投加过程包括絮凝反应和沉淀两个过程，絮凝反应发生在反应池中，待反应后进入沉淀池进行沉淀，其具体的过程可以如图 2-2 所示：

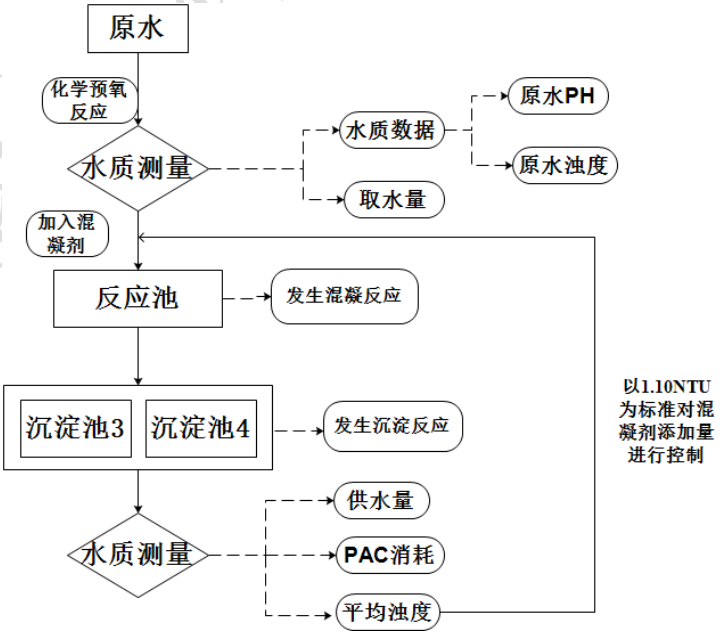

图 2-2 絮凝沉淀过程分析

➢ 变量分析

通过对混凝过程分析，出水浊度是因变量，原水 PH、取水量、原水浊度、PAC 投放量等都属于自变量，可以初步确定题中给出的相关变量对某些变量的影响如表 2-1 所示：

因素变量	原水 PH	取水量	原水浊度	时　　间
对出水浊度影响	有影响[1]	正向影响	正向影响	反向影响
因素变量	供水量	PAC 投放量	温　　度	
对出水浊度影响	无影响	反向影响	有影响[1]	

表 2-1 因素初步分析结果

2.2.2 数据抽取

根据对本题的分析，附件给出了原水 PH、原水浊度、出水浊度、取水量、供水量以及 PAC 消耗量数据，其中取水量是原水的流速，供水量是出厂水的流速（沉淀池后还有部分工艺会造成水的损耗），取水量和供水量的单位是 m^3/h（立方米每小时），PAC 耗是混凝剂 PAC 的消耗，单位是 mg/L（1L 原水消耗 PAC 的量）。由前面变量初步分析结果可知，供水量与出水浊度并无关系，故在选取数据时可以直接删去。

2.2.3 样本数据预处理

本文模型辨识参数所有数据均来自于广告某水厂某水厂投药控制系统实时采集的数据信息，包括混凝剂投量、取水量、原水浊度、出水浊度等多项数据。由于实时数据库中采集得到的生产数据为工艺变量的瞬时值，因而存在许多不合理的野值,加之考虑到仪器的短暂性失灵、采集器清洗等影响，因此必须对实际生产数据进行如下几步预处理。

➢ 缺省值处理

由附件数据集中的历史数据可知，从 8 月 8 日 00:00 到 8 月 19 日 23:00 的数据集中 PAC 消耗量为缺省值，由于该段时间的数据几乎都为缺省值，根据缺省值处理原则，可以直接将该阶段的测量数据去除，不会影响最后的结果，其处理结果为附件 1 的缺省值处理表。

➢ 平均值处理

根据题目给定的信息，对附件中的 3 号和 4 号沉淀池出水浊度数据进行平均化处理，作为沉淀池的出水浊度，处理后的数据可见附件 1 的平均值处理表。

➢ 异常值处理

由附件数据集可知，2013 年 8 月 20 日 1:00 到 8 月 21 日 22:00 的原水浊度数据在 300-800 之间，出水浊度为 1-7 之间，但消耗的 PAC 量在 0-3 之间，同后面的数据进行比较可知 PAC 量数据明显错误。对于这种连续的异常值可以直接删去，不会影响建模，其结果可见附件 1 的异常值处理 1 表。

除了上面存在的极端错误外，原始样本数据中还存在一些样本在某些时间点忽然出

现较大的尖峰或低谷，那么就不能删去，必须进行其他的异常处理。这种异常有可能会存在数据测量错误，可以采用统计方法进行纠正；但如果是因为某些工艺失败而导致数据之间存在某些关系完全被打破，那么对于这些数据就得进行异常值检测[2]。

均值滤波算法[4]是基于统计理论的一种能有效抑制噪声的非线性信号处理技术，其可定义如下：

$$[g(x,y)] = mean\{f(s,t)\}, (s,t) \in S_{xy} \quad (2\text{-}1)$$

异常的判别规则如下：

$$\begin{cases} [\Delta f(x,y)] > std[f(s,t)] \\ [\Delta f(x,y)] > f(x,y) - mean[f(s,t)] \end{cases} \quad (2\text{-}2)$$

其中：$[g(x,y)]$表示$[(x,y)]$为点的输出值，$[S_{xy}]$表示以$[(x,y)]$为中心的邻域，$[f(s,t)]$表示以$[(x,y)]$为中心的邻域内$[(s,t)]$点输入值；若式（2-2）成立，则$[(x,y)]$为异常点。

在对数据进行该滤波检测法时，选取该数据前后两个值作为邻域。在判别异常时需要会出现此种情况：当去掉最这个 5 个数据中最大最小，剩余的 3 个数据比较接近时，该数据会与其平均值接近，因而所得的标准差会特别小，在判别时会将异常数据视作正常数据，为此采用以下加权处理：

$$std(f(s,t)) = 0.2 * f(X(Y(+,)),)std(f(s,t)) * 0.8 \quad (2\text{-}3)$$

其中 $std(f(X,Y))$ 为总体数据标准差，这里为简化计算只取该时间序列的前 100 个数据。通过本种方法处理，最终检测出 70 个异常二的数据情况如下表 2-2 所示，具体可见附件 1 异常 2 检测结果表。

变量	原水 PH	原水浊度	出水浊度	取水流量	PAC 消耗量
异常点个数	0	20	1	29	20

表 2-2 滤波检测算法下异常二检测结果

- 分浊度区间

经前面数据处理步骤，通过对该水厂大量数据分析发现，原水浊度范围变化很大，源水清澈时其最低浊度时仅为 5.13NTU，原水浑浊时其最高浊度高达 868.36NTU，为了使建立的投药量模型适用范围更广，应该对原水浊度划分样本区间集,以便使得建模数据集和验证数据集涵盖所有浊度区间，对数据样本分浊度区间(前闭后开区间)后得到样本数集如表 2-3 所示。

浊度区间（NTU）	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
样本数量	288	3585	2092	813	550	384	301	191	133
浊度区间（NTU）	90-100	100-200	200-300	300-400	400-500	500-600	600-700	700-800	>800
样本数量	116	472	63	27	17	8	15	16	2

表 2-3 数据样本浊度分区间表

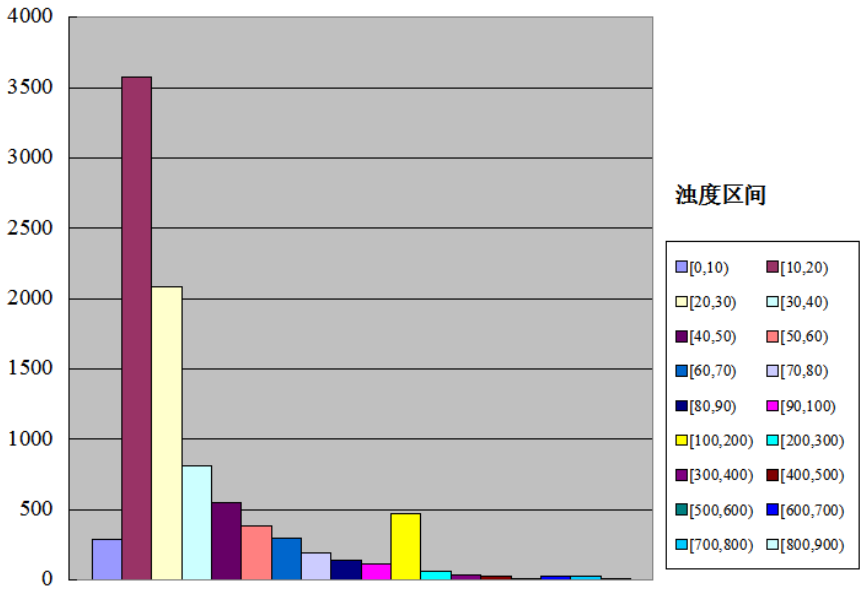

图 2-4 样本数据浊度区间直方图

由表 2-3 和图 2-4 可知:该水厂的原水水质比较稳定,其浊度大部分在 10NTU-400NTU 之间,特别高原水浊度出现非常少,水质情况集中于低浊度区间部分。辨识模型参数时,可将处理后的样本空间划分为训练集和泛化集,训练集用于模型训练和辨识泛化集用于模型的检验和预测。本文按照原水浊度区间划分选取数据，其中三分之二的数据作为训练集，三分之一的数据可以作为验证集。

2.2.4 模型与实验

1) 模型 I：基于非线性回归辨识指数模型

回归分析法是通过大量观察数据利用数理统计方法建立因变量与自变量之间的回归函数表达式的一种方法。回归分析分为线性回归分析和非线性回归分析，通常线性回归分析法是最基本的分析方法,遇到非线性回归问题可以借助数学手段化为线性回归问题处理,然后再用最小二乘法求出参数的估计值,最后再经过适当的变换,得到所求回归方程式。

综合已有研究文献[3]和研究试验结果分析发现：混凝剂投加量与各因素之间可以用如下指数形式表示:

$$M = a_0 C_0^{a_1} Q^{a_2} C_1^{a_3} \quad (2\text{-}4)$$

其中:M 为混凝剂投加量（mg/L）,C_0 为原水浊度（NTU），Q 为取水量（m^3/h）C_0 为沉淀池出水浊度（NTU），$a_0\ a_1\ a_2\ a_3$ 为待估参数。在本部分，由于 PH 值变化不大，对出水浊度的影响可以暂时忽略，以简化本阶段模型。

当求解式（2-3）的各参数，再由题目中给出的反应到结束的时间大约为 70min-120min，所以若设所有出水浊度预测值与对应一个小时后的出水浊度实际值更接近的次数为 h_1，与两个小时后的出水浊度实际值更接近的次数为 h_2，$\Delta t = 1h$ 则原水添加混凝剂反应到沉淀结束所需的时间可以表示为：

$$\Delta T = (1 + \frac{h_1}{h_1 + h_2})\Delta t \quad (2\text{-}5)$$

由式(2-3)可看出投药量与浊度和流量之间为指数非线性关系，这种指数非线性关系相对比较简单,可以通过取对数实现指数非线性到线性化的转换，因此对(2-3)式两边取对数得：

$$\ln M = \ln a_0 + a_1 \ln C_0 + a_2 \ln Q + a_3 \ln C_1 \quad (2\text{-}6)$$

由此可以得到：

$$\ln C_1 = \frac{1}{a_3}(\ln a_0 + a_1 \ln C_0 + a_2 \ln Q - \ln M) \quad (2\text{-}7)$$

若以 $\ln C_1$ 作因变量，$\ln M, \ln Q, \ln C_0$ 作为自变量，则可以建立（2-4）式的多元线性回归模型。

设 $y = \ln C_1, x_1 = \ln M, x_2 = \ln Q, x_1 = \ln C_0$,其中 a_0, a_1, a_2, a_3 为回归系数。对 y 和 x_1, x_2, x_3 分别进行 n 次独立观测，得到 n 组数据样本：$y_i, x_{i1}, x_{i2}, x_{i3}, (i = 1, 2, \cdots, n)$，则有：

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \varepsilon_1 \\ y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \varepsilon_2 \\ \vdots \\ y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \varepsilon_n \end{cases} \quad (2\text{-}8)$$

其中 $\varepsilon_1, \varepsilon_1, \cdots, \varepsilon_n$ 为残差，且相互独立，并服从 $N(0, \sigma^2)$ 分布。

令 $Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$，$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$，$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$，$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$，则式（2-8）可转化为矩阵形式表示：

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 I_n) \end{cases} \quad (2\text{-}9)$$

从经过数据预处理后得到的数据，连续抽取 3000 个数据并将其随机分成 6 个样本表，为确保建立的模型适应于各种浊度区间,在随机分配样本表时,应使每个样本集表中都应包含各种不同源水浊度区间，并任取其中的 5 个数据表用于模型的参数获得，另一组样本表格数据用于检验模型的有效性。采用线性回归辨识,求解出 5 组不同数据样本下的待定参数如表 2-4 所示。

参数	第一组	第二组	第三组	第四组	第五组	平均值
a_0	5.8138	6.0739	5.9439	6.1439	5.8899	5.97308
a_1	0.1918	0.2075	0.2188	0.2195	0.2237	0.21226
a_2	0.3233	0.3624	0.3523	0.3324	0.3421	0.3425
a_3	-0.3677	-0.3324	-0.3429	-0.3596	-0.3389	-0.3483

表 2-4 线性回归辨识模型参数结果表

将表 2-4 最终无阻辨识出的参数取平均后代入式（2-7）中进行线性回归效果的显著性检验得到在 $\alpha = 0.01$ 水平下是显著的，故可信度达到 99%，可以将其作为模型的最后参数，即可得到投药量指数数学模型为：

$$M = 5.97308 C_0^{0.2126} Q^{0.3425} C_1^{-0.3483} \quad (2\text{-}10)$$

由此即有出水浊度与 PAC 投放量、供水流量、原水浊度的函数表达式如下：

$$C_1 = \exp(-0.6094 \ln C_0 - 0.9833 \ln Q - 2.8711 \ln M - 5.1314) \quad (2\text{-}11)$$

则由此根据第六组的样本数据中的水质数据、供水流量、PAC 投放量计算出水浊度预测值（见图 2-5），并与实际样本的出水浊度对比，并统计出 h_1 和 h_2，最终根据式（2-5）可求得絮凝反应到沉淀结束实际为 $\Delta T = 90\,\text{min}$。

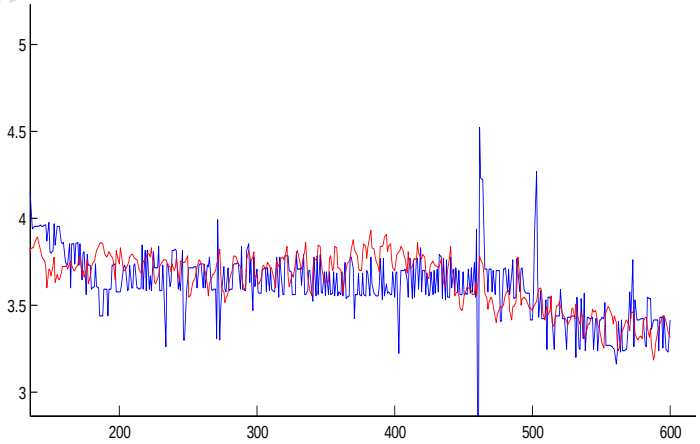

图 2-5 函数预测值与实际值对比图

2) 模型 II：基于改进 BP 神经网络的非线性黑箱投药控制模型

BP 神经网络由输入层、一层或多层隐含层及输出层构成，学习过程由信息的正向传播和误差的反向传播两部分组成，采用梯度最速下降法不断调整权值使网络总误差达到期望值误差以下，能够帮助研究人员简化工作，达到根据输入数据智能推理系统输出的目的。本题要求建立混凝投药过程的变量关系，由模型一即可知变量间的关系属于非线性关系，同时也满足神经网络的构建条件，故可以利用神经网络结构来建立彼此之间的联系。

- 投药量神经网络模型结构选择

本文利用 BP 神经网络建立系统模型时，其输入输出层节点数目由系统输入输出变量数客观确定。通过前部分对投药量的影响因素分析可知,影响投药量的关键因素有取水量、原水浊度、原水 PH 和出水浊度,因此神经网络模型的输入变量为三个，分别为取水量、原水浊度、原水 PH、混凝剂的投加量即投药量，输出变量为出水浊度，其神经网络模型结构如图 2-6 所示：

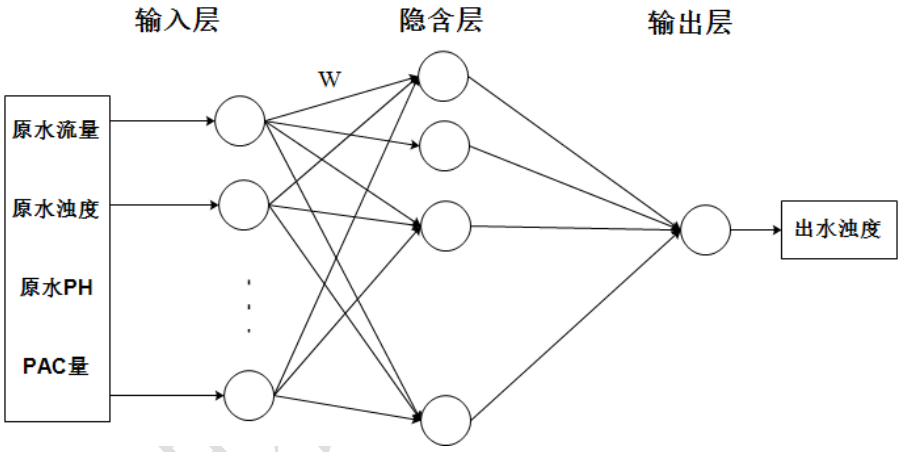

图 2-6 混凝投药神经网络结构图

鉴于模型的准确度和网络结构的复杂性，本文对隐含层神经元节点数分别从 8-17 进行选取,且隐含层神经元节点的激励函数分别选用 TANSIG 和 LOGSIG 函数,从而构成 20 种神经网络模型。对这 20 种模型结构分别进行训练,规定各种模型训练的最大训练次数为 500 次,网络结束训练的条件是当网络相对误差平方和均值误差 E 达到预先设定的精度 0.001,训练结束时得到的网络模型就是所需的混凝剂投加量神经网络模型。

- 网络训练与泛化

本文采用改进的 BP 算法,即在常规 BP 算法中加入决定网络收敛速度和定性的系数：学习速率 η 和动量项系数 α 。学习速率 η 太小，则网络收敛很慢； η 过大,则可能出现麻痹现象甚至使网络出现振荡。冲量系数 α 则可根据误差情况改变网络训练过程中的学习速率。在本模型中,学习速率 $\eta = 0.9$,冲量系数 $\alpha = 0.8$ 。

BP 网络学习的过程实际就是在样本输入数据下根据性能准则函数不断修改各层各节点之间的连接权种,最终获得权值矩阵的过程。本模型中,输入节点数 $i=3$,中间节点数 $j=9\sim16$，输出节点数 $k=1$，读入样本集后,对每个样本 p 作如下计算。

① 前馈计算各层节点输出

对隐含单元：

$$net_{pj}=\sum_{i=1}^{3}w_{pi}x_{ij},o_{pj}=f(net_{pj}) \quad (2\text{-}12)$$

对输出层单元：

$$net_{p}=\sum_{j=1}^{k}w_{pj}w_{j},o_{p}=f(net_{p}) \quad (2\text{-}13)$$

计算每个样本 p 的输出误差 E_p：

$$E_p=\frac{1}{2}(d_p-o_p)^2 \quad (2\text{-}14)$$

若误差达到指定要求，学习结束。否则,进入下面第②步,即从输出层反向传播，逐层修改权值直到误差满足要求为止。

② 反向传播调整各层权值和阈值

输出层权系数调整：

$$\delta_p=o_p(1-o_p)(d_p-o) \quad (2\text{-}15)$$

$$w_j^p(t+1)=w_j^p(t)+\eta\delta_p o_p+\alpha[w_j(t)-w_j(t \quad (2\text{-}16)$$

$$\theta_j^p(t+1)=\theta_j^p(t)+\eta\delta_p o_{pj}+\alpha[\theta_j t()-\theta_j t(- \quad (2\text{-}17)$$

隐含层权值的调整：

$$\delta_{pj}=o_{pj}(1-o_{pj})\sum_k\delta_{pk}w_{kj}^p \quad (2\text{-}18)$$

$$w_{ij}^p(t+1)=w_{ij}^p(t)+\eta\delta_{pj}o_{pi}+\alpha[w_{ij}(t)-w_{ij}(t \quad (2\text{-}19)$$

$$\theta_{ij}^p(t+1)=\theta_{ij}^p(t)+\eta\delta_{pj}+\alpha[\theta t_{ij}()-\theta t_{ij}(- \quad (2\text{-}20)$$

随机选用附件 3 中预处理后数据表中 3000 个数据作为建模，2/3 作为训练数据，1/3 作为测试数据，具体可见附件 3 的建模数据表、训练数据表、测试数据表。在 matlab 中设置不同隐含层激励函数和不同隐含层点个数组成的 16 种网络模型分别进行训练和泛化，得到各种模型的训练误差和泛化误差结果如表 2-5 所示：

隐含层激励函数	隐含层节点个数	训练误差	泛化误差	训练次数

TANSIG	9	0.0019	0.0027	41
	10	4.77×10^{-4}	0.0189	220
	11	7.04×10^{-4}	0.0168	35
	12	5.32×10^{-4}	0.0058	23
	13	0.0012	0.0096	63
	14	9.11×10^{-4}	0.0032	51
	15	0.0025	0.0014	24
	16	0.0023	0.0062	18
LOGSIG	9	1.81×10^{-4}	0.0071	500
	10	0.0022	0.00187	118
	11	3.72×10^{-4}	0.0164	32
	12	1.24×10^{-4}	0.0014	41
	13	0.0013	0.0054	13
	14	1.58×10^{-4}	0.0082	28
	15	6.37×10^{-4}	0.0158	22
	16	5.41×10^{-4}	0.0032	17

表 2-5 各神经网络模型训练结果对比

由上表训练结果可知，隐含层节点个数为 12，隐含层节点激励函数为 LOGSIG 函数时，模型的训练误差最小和校验误差最小，达到训练精度要求时的训练次数也最小，其中训练误差为 1.24×10-4，校验误差为 0.0014，训练次数为 32 次，故该模型为性能最优的神经网络模型，其结构 $4-12-1$。

BP 网络泛化能够检验训练好的网络模型对未认知的样本数据是否具有较好的推广能力，将泛化样本输入该网络模型得到模型泛化结果如图 2-6 所示：

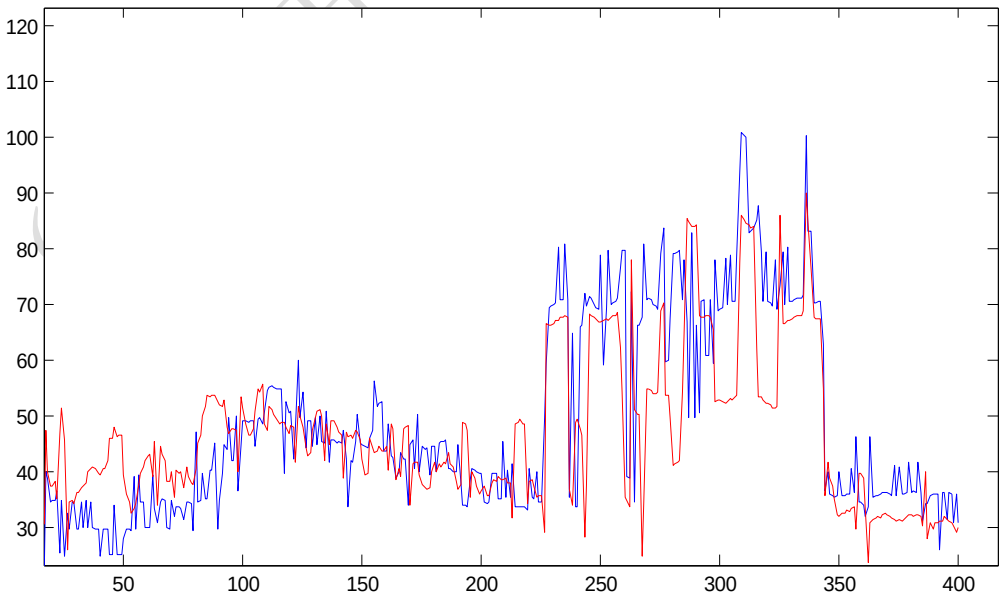

图 2-6 BP 网络模型泛化结果

由上图可以看出，该网络输出的出水浊度预测值与实际值基本一致，说明该网络模型具有较强的推理能力，可以较准确的估算出水厂过程的投药量。故设定原水 PH、原水浊度及取水量，我们可以得到以下对应的最佳 PAC 投药量，如表 2-6 所示：

原水 PH	7.3	7.3	7.3	7.3	7.3	7.4
原水浊度	17.56	17.56	16.4	14.86	65.78	65.78
取水量(m^3/h)	11672	10621	14825	14825	8372	8372
最佳投药量（mg/L）	29.94751	31.74393	23.19359	23.0745	40.56134	40.65781

表 2-6 最佳投药量与原水水质数据及流量对应表

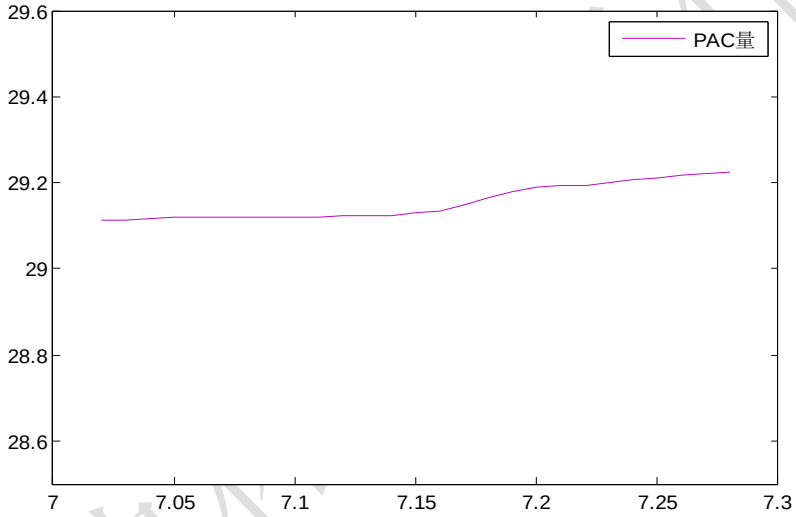

图 2-7 PH 对最佳投药量的影响

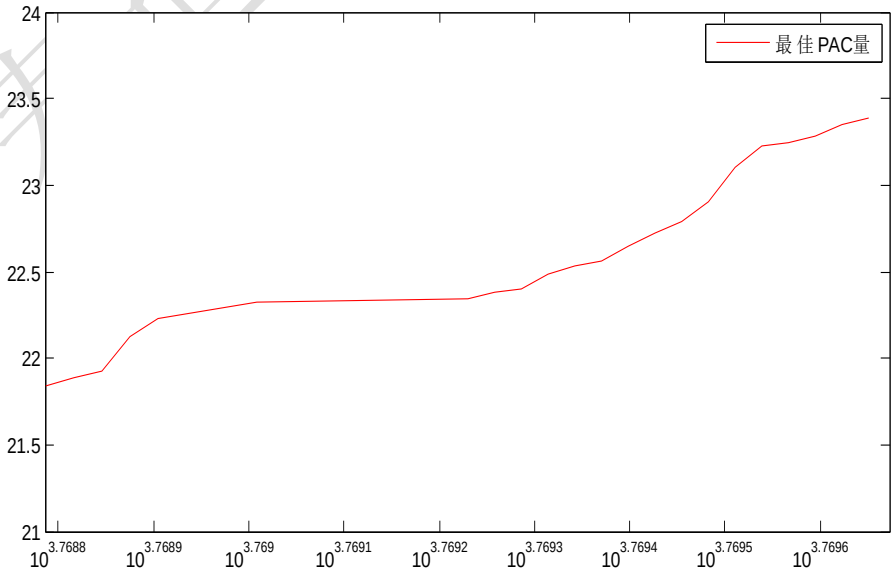

图 2-8 取水流量对最佳投药量的影响

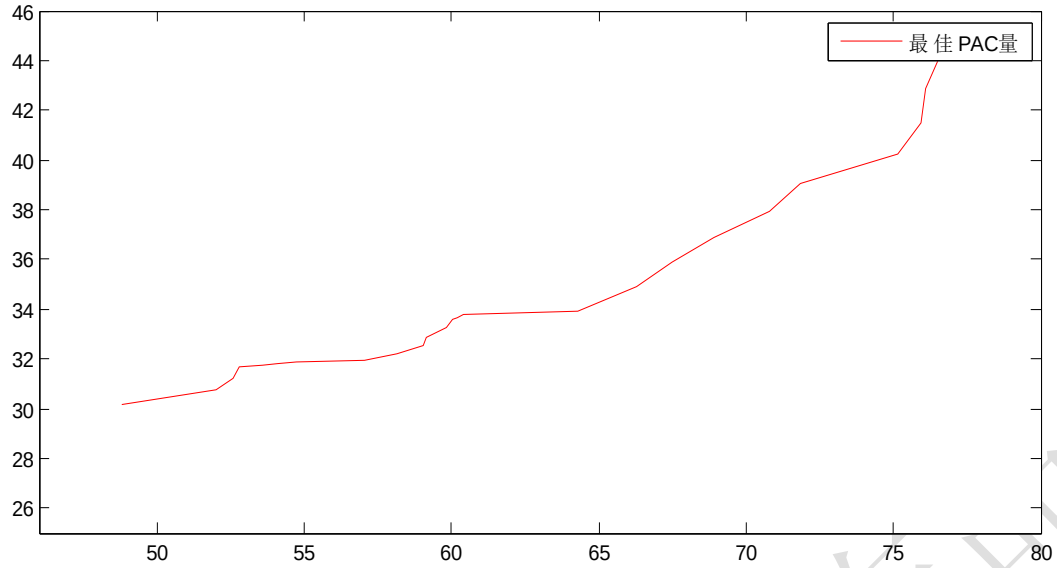

图 2-9 原水浊度对最佳投药量的影响

3) 模型 III：基于改进 BP 神经网络的非线性黑箱投药反馈模型

在模型 II 中，本文主要通过对絮凝沉淀过程进行正向预测，根据给定的水质数据和供水流量通过使出水浊度达到正常标准来控制最佳投药量，而在本文问题分析第一部分就明确指出絮凝沉淀过程属于一个投药反馈控制的过程，由于反应过程的时滞性，在投药时还得考虑上一次投药时的出水浊度，根据对上次的出水浊度来调节本次投入的最佳药物量，因此需要将出水浊度作为输入参数，投药量作为输出结果建立 BP 神经网络反馈模型，其神经网络结构如下所示：

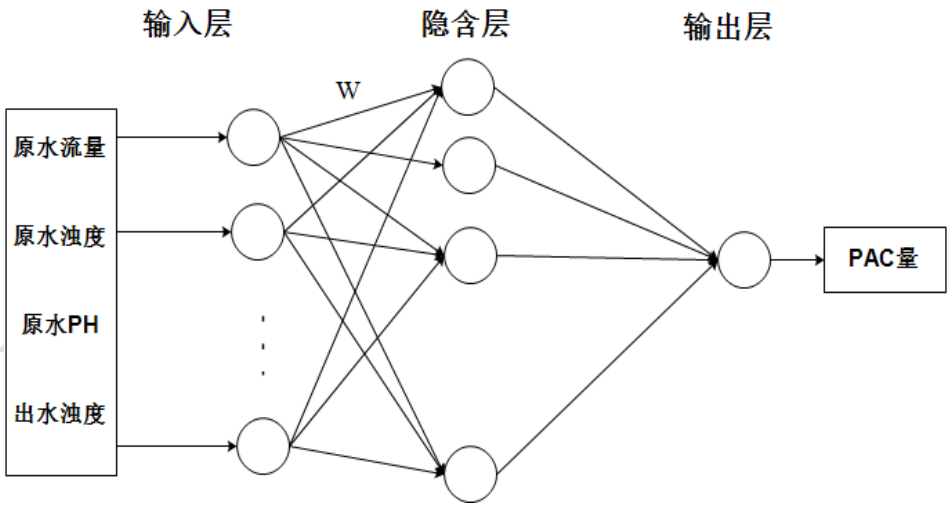

图 2-10 混凝投药反馈神经网络结构

将泛化样本输入该网络模型得到模型泛化结果如图 2-8 所示。由图 2-8 看出,该网络输出的投药量预测值与生产实际中的投药量基本一致,由此可以根据不同的条件下较准确地控制水厂过程的投药量，具体可见表 2-7 所示：

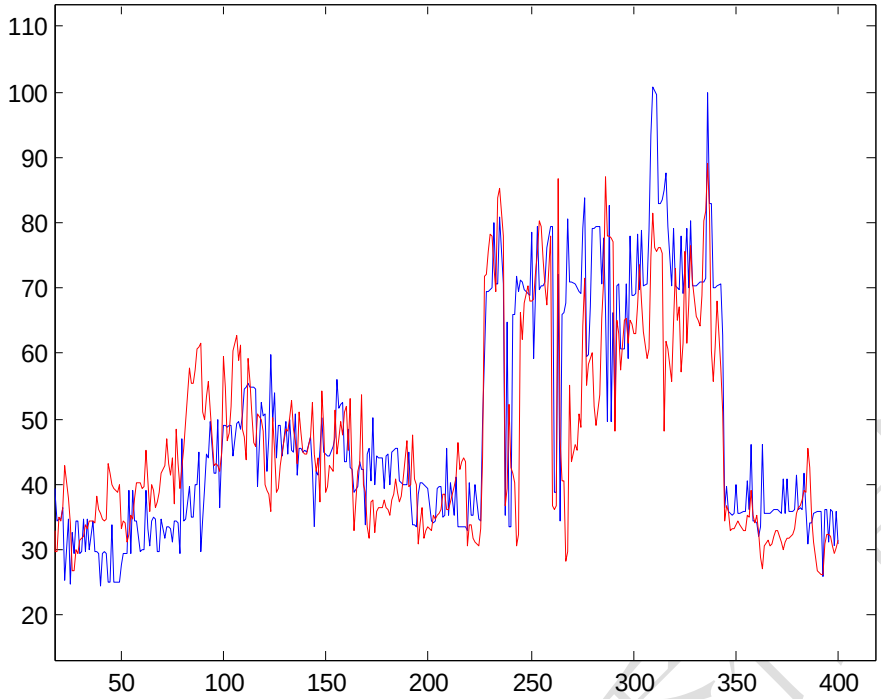

图 2-11　增加出水浊度 BP 网络反馈控制模型泛化结果

原水 PH	7.3	7.3	7.3	7.3	7.4
原水浊度	37.02	37.02	37.02	138.66	138.66
取水量	6696	6690	5880	6690	6690
出水浊度	0.89	0.92	0.92	0.89	0.89
最佳投药量	30.906218	29.80292	28.427013	45.85468	38.34738

表 2-7　增加出水浊度 BP 网络反馈控制模型下的最佳投药量

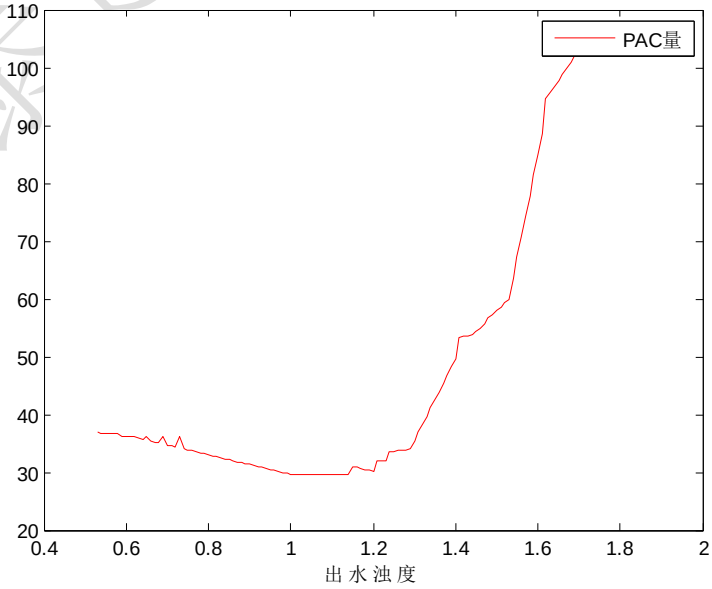

图 2-12 出水浊度对最佳投药量的影响

4) 模型 IV：基于温度因素影响的改进 BP 神经网络模型

由前面可知，温度能影响出水浊度，能影响药剂溶解速度，水温的变化改变了胶体自身的布朗运动动能。要达到目标的脱稳凝聚效果就要求投加更多的混凝剂，使排斥能峰小到足以克服的程度，即温度会影响最佳投药量。水温低时粘度很大，水温高时粘度小，其具体的水动力粘度数据如表 2-8 所示：

温度 ^{0}C	0	5	10	15	20	25	30
粘度	1792	1519	1308	1140	1005	897	801

表 2-8 水的动力粘度(10^{-6}Pa s)

由于附件给出的数据集中并没有温度数据集，本文根据该水厂位于广东广州南沙区，由此可知其水质来源为珠江[10]，则可通过珠江水利网查询获得该水厂从 2013 年 8 月 22 日到 2014 年 9 月 5 日共计 380 个水温数据，具体可见附件 5 的气温表。由文献[11]可知气温和水温存在较强的相关性，日均气温和水温的关系可表示为：

$$T_w = 0.813T_\alpha^* + 3. \qquad (2\text{-}21)$$

其中 T_w 表示水温、T_α 表示气温，由此可由日均气温数据和式（2-21）求得日均水温数据，具体可见附件 5 的水温表.

由于原始数据都是 24 小时的瞬时值，故需要对这些数据进行进行日均化，这些数据为时间序列数据，取 8 月 22 日的数据对部分变量作出趋势图分别如下：

图 2-13 8 月 22 日原水 PH 一天内的变化

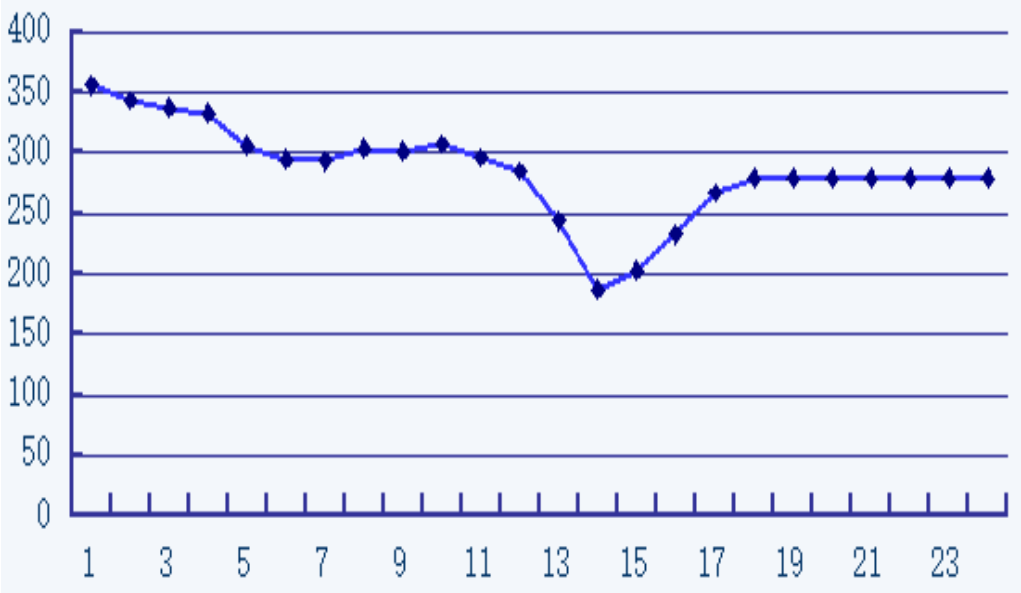

图 2-14 8 月 22 日原水浊度一天内的变化

图 2-15 8 月 22 日出水浊度一天内的变化

由此可知这些变量在一天内的变化有些呈现趋势变化，例如原水 PH，有些则没有规律如出水浊度等，所以这些变量不能采用相同的方法进行日均化。由于时间等间距，故对于起伏较平稳的变量可以采用直接求平均，对于起伏波动较大的则只能采用图像积分面积法求得最终的日均变化，其中由于 PAC 的单位为 mg/L,故还需与取水量数据进行乘积后再进行积分均化，最终经过处理可得到附件 5 的混凝投药综合因素数据表.

本部分在模型三基础上增加温度输入参数，构建新的三层结构 BP 神经网络模型实来研究温度对投药量的影响。此时在输入层中，输入向量有 5 个元素：原水温度、原水 pH 值、原水浊度、取水量和出水浊度，输出层则只有投药量，其具体的神经网络结构如下图 2-12 所示：

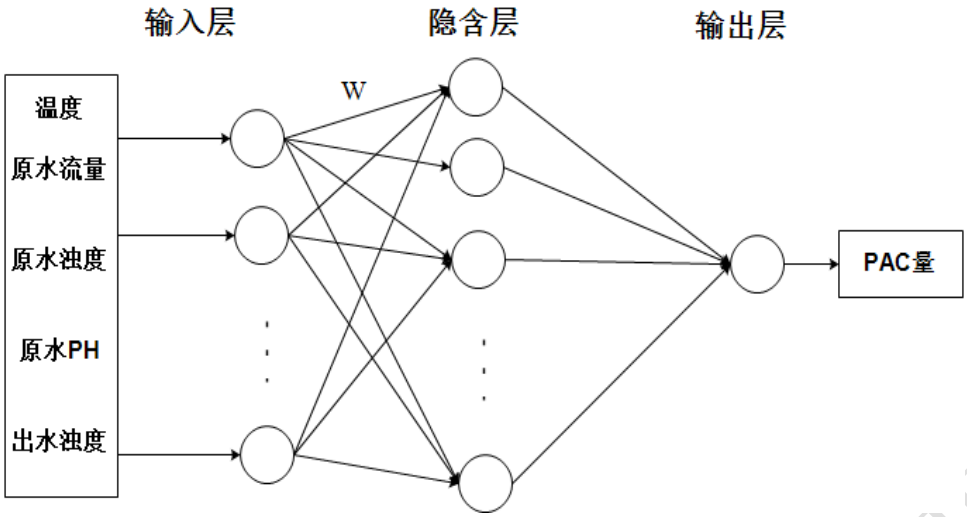

图 2-16 增加温度的神经网络结构

根据前面的神经网络结构，由于输入参数增加，所以需要重新寻找最优的隐层节点数，同样采用相同的方法最终确定隐层节点为 12 最佳。通过本模型改变温度输入量，最终得到最佳投药量随温度的变化图 2-17：

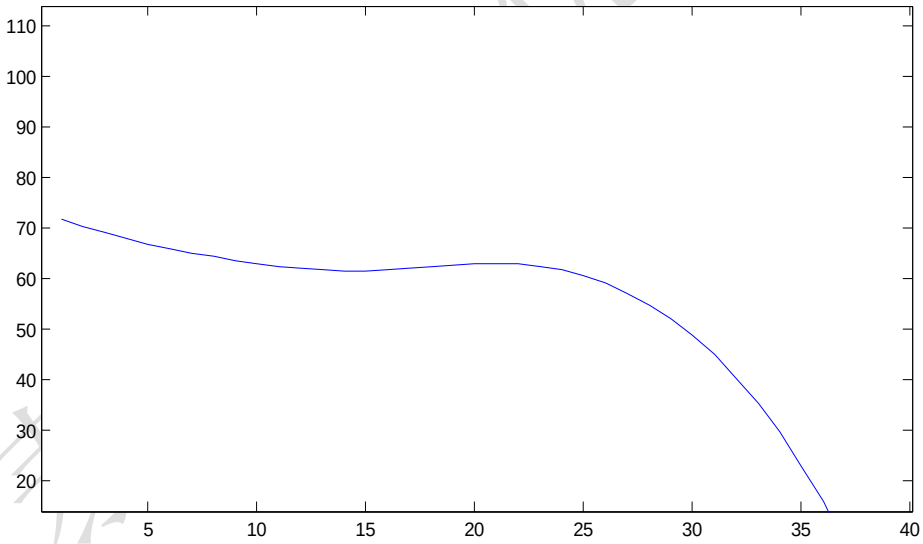

图 2-17 温度对最佳投药量的影响

5) 模型评价与总结

本文采用分阶段建立模型使模型由简到繁，不断优化最终建立最全面、最合理、最符号现实的模型，实现从投药-絮凝-沉淀-投药的整个反馈控制过程。

该四阶段的四个模型层层递进：模型一参考文献[1]，并考虑到 PH 的数据变化并不大，故在建模时忽略了 PH 对出水浊度的影响，使模型变得简单；模型二采用神经网络建模，分析了原水 PH、原水浊度、取水量以及 PAC 投放量与出水浊度的关系，通过控制出水浊度来控制 PAC 投放量，比模型一更优，属于正向控制；模型三在模型二基础

上考虑出水浊度对 PAC 的反馈调节建立了新的神经网络，使模型接近现实中的投药控制系统；模型三则引入温度因素，丰富了神经网络的结构，使模型对现实控药系统更接近、输出结果更准确。四种模型的关系可以表示如下：

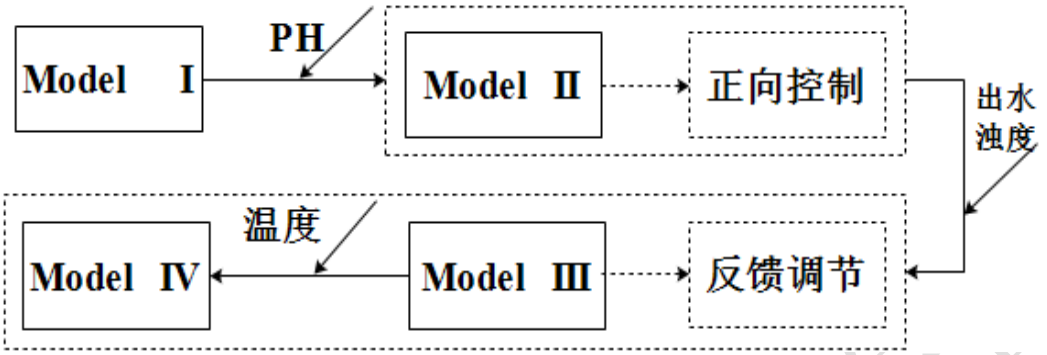

图 2-18 四种模型关系图

当然模型也存在一些不足：BP 神经网络梯度下降的学习方法可能会陷入局部最优，所以建立的 BP 网络模型会存在一些误差，影响最终的投药量确定。

6) 模型优化

RBF 神经网络结构模型较 BP 神经网络有精度高、训练速度快、泛化能力强等优点，故可以将模型二和模型四均改进为用 RBF 神经网络算法。

RBF 神经网络为典型的三层前馈式网络，包括输入层、中间非线性处理层和线性加权输出层，隐层节点作用函数为 RBF 函数。针对本题的混凝投药过程，输入、输出参数与前模型相同，RBF 采用常用的高斯基函数，根据文献[12]选用 10 个隐层节点如表 2-9 所示，阈值取 6.4241，隐层高斯基函数中心采用 K-均值聚类算法确定，输出层权值通过 RLS 算法调节。

w_1	w_2	w_3	w_4	w_5
0.6323	0.7058	0.7440	1.1819	0.3193
w_6	w_7	w_8	w_9	w_{10}
0.2016	0.3295	-0.0435	0.3854	0.5244

表 2-9 隐含层节点权系数

在 SPSS 中通过 RBF 训练后形成的神经网络结构模型所作出的预测控制量能够与实际值得到较好的拟合，能够实现对投药量更好地控制，其训练的泛化结果如图 2-15 所示：

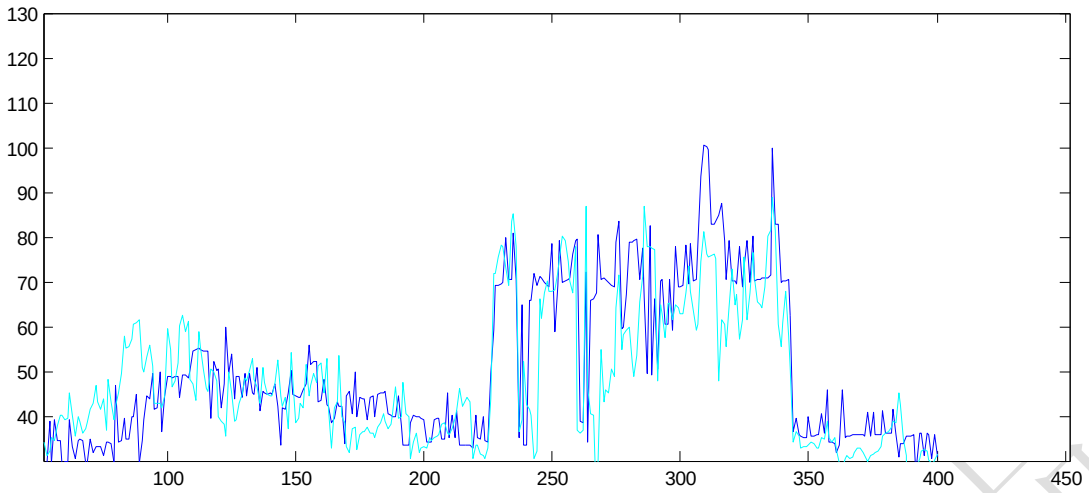

图 2-19 RBF 神经网络泛化结果

将 RBF 神经网络模型与 BP 神经网络模型检验结果进行对比可知，RBF 神经网络模型可以较好地逼近水处理过程的非线性关系，建模精度高，训练时间段，适用于在线要求较高、数据量角度的系统模型辨识，其具体差异可以见表 2-10：

指标	最大绝对误差	平均绝对误差	最大相对误差	平均相对误差
RBF	0.31	0.13	6.36	2.79
BP	0.95	0.21	16.78	4.12

表 2-10 RBF 与 BP 的误差比较结果

根据 RBF 神经网络最终可建立如下的闭环预测控制系统，其工作原理为：基于出水浊度的 RBF 神经模型对未来的出水浊度作出预测，利用 RBF 正向控制投药量，通过系统实际输出和模型预测输出的出水浊度差异进行反馈校正。该控制系统在水质突变时可以快速响应，在保证出水稳定、合格前提下迅速改变 PAC 投加，实现了投药量的实时控制，其具体结构如下所示：

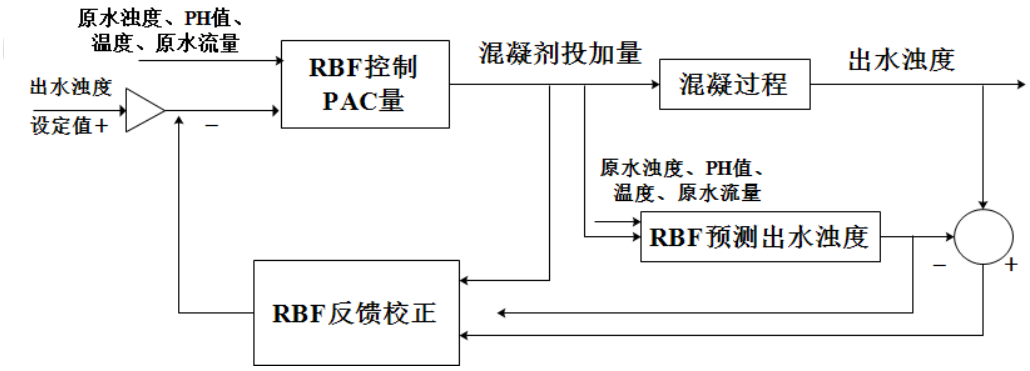

图 2-20 基于 RBF 神经网络的最佳投药量的闭环预测系统

2.3. 结果分析

对于问题一，本文采用非线性回归模型对出水浊度进行预测，并通过统计预测值与下一时刻和下两个时刻的实际值更接近的次数求得其反应到沉淀结束的时间为 90min，由于没有考虑到 PH 的影响，故其时间不够准确，由后面的 PH 对最佳投药量和出水浊度的影响，可知该时间应该过大，这也正反映出时间的滞后性。

对于问题二，本文采用 BP 神经网络模型对出水浊度进行标准控制，最终得到在不同条件下的最佳投药量。从表 2-6 和图 2-7、2-8、2-9 可知，PH、原水浊度以及取水量与最佳投药量呈正相关，其中 PH 对最佳投药量影响非常小，取水量对最佳投药量影响较大，原水浊度对药量影响最大。

对于问题三，本文增加出水浊度作为神经网络的输入量，得到在不同条件下最佳投药量的控制，此时出水浊度作为投药量的反馈控制参数，从表 2-7 和图 2-12 所示，在其他参数不变得情况下出水浊度在 1.0NTU 标准以下时，最佳投药量都会减少，但是出水浊度越小则药量减少得越多；出水浊度在 1.0NTU 标准以上时，最佳投药量都会增加，但是出水浊度越大则药量增加得越多。

对于问题四，本文增加了温度参数作为输入量，由最终结果可以得知在 $0\text{-}10^0$C 之间，温度升高时最佳投药量会有所增加；在 $10\text{-}25^0$C 时，温度对最佳投药量几乎没有影响，而在 25^0C 后，随着温度的增加，由于温度加快了絮凝反应和减小了水的粘度使投药量减少，其具体的结果可见图 2-17.

3. 结论

本次数据挖掘建模在于利用附件数据信息，建立原水水质、取水量、出水流量、出水浊度、药物投加量及温度的数学模型，实现对混凝剂投药量进行实时控制。

第一阶段综合文献[2]运用非线性回归模型建立出水浊度与原水流量、原水浊度及投药量之间的函数关系，通过利用回归模型对样本输出进行预测，将输出与样本实际值进行对比，通过统计预测值与实际值或下一时刻实际值接近的次数，即式（2-5）求得反应到沉淀结束的时间位 90min。

第二阶段以原水 PH、原水浊度、投药量、原水流量作为输入，出水浊度作为输出，通过控制出水浊度的标准训练样本数据建立改进 BP 神经网络控制模型，由此模型求得在不同原水水质和原水流量情况下最佳投药量，结果如表 2-6 所示；并分析了原水 PH、原水浊度、原水流量对最佳投药量的影响，具体可见图 2-7、2-8 和 2-9。

第三阶段以原水 PH、原水浊度、出水浊度、原水流量作为输入，投药量作为输出，通过控制出水浊度的标准训练样本数据建立改进 BP 神经网络反馈模型，出水浊度在本模型中作为反馈参量实现对投药量的反馈控制，结果可见表 2-7 及图 2-12.

第四阶段以原水 PH、原水浊度、出水浊度、原水流量以及温度作为输入，PAC

投放量作为输出，通过控制出水浊度的标准训练样本数据建立改进 BP 神经网络控制模型，由此模型研究了温度对最佳投药量的影响，具体可见图 2-17，并且该模型能够对不同水质条件、水流量、温度等因素下最佳投药量的控制。

最后根据 BP 神经网络的缺陷，本文设计了基于 RBF 神经网络结构的投药控制模型（见图 2-14），并基于之前的模型最终设计出最佳投药量的闭环预测系统，实现了对投药量在水发生变化情况下的实时控制。

4. 参考文献

[1] 白桦. 智能控制在净水厂混凝投药过程中的应用研究[D]. 哈尔滨工业大学，2002

[2] 马勇，朋永臻等编著. 城市污水处理系统运行及过程控制. 北京：科学出版社，2007

[3] 田一梅，张宏伟，齐庚中，罗津悦.水处理系统运行状态数学模拟的研究[J]. 中国给水排水，1998,14(4):10-12

[4] 辛欣,周娜,王震. 数据异常值检测及修正方法研究[J]. 现代电子技术,2013,11:5-7+11.

[5] 崔琳琳. 自来水混凝投药过程建模与模糊控制研究[D]. 昆明理工大学,2014.

[6] 王军栋. 混凝投药过程非线性预测控制研究[D]. 哈尔滨工业大学,2011.

[7] 唐德翠. 城市供水水处理系统的建模、控制与运行优化研究[D]. 华南理工大学,2013.

[8] 常永滑. 净水厂混凝投药控制的研究[D]. 天津大学,2007.

[9] 白桦,李圭白. 混凝投药的神经网络控制方法[J]. 给水排水,2001,11:83-86+0.

[10]李亮,周建忠,汪麟,赵忠富. 广州南沙水厂的工程设计及特点[J]. 中国给水排水,2011,14:41-45.

[11]董林垚,陈建耀,付丛生,蒋华波.珠海小规模溪流水温与气温关系研究[J]. 水文,2011,01:81-87.

[12]沈捷,王莉,林锦国. 水处理过程的 RBF 和 BP 神经网络建模[J]. 微计算机信息,2007,34:294-296.