关于水质预测的重要性：如果能更好的预测水质的成分变化，那么对于后期的水质处理，可以极大的提高效率，并且在处理水质的成本上，也可以使得我们可以及时的选择最低成本，最大程度的处理水污染问题。

摘要： 作者余等，针对于水质检测常见技术ELM、BPNN、LSSVM进行了实验和比较，并且结合群智能领域的GA和PSO，对ELM的权值和阈值的选取做了优化，使得ELM的效果和拟合程度更好。

一、国内外常用的水质预测法

1. 统计预测法

该方法主要是通过过去已有的数据进行训练，然后对未来的数据进行预测，常用的方法有：**回归分析预测法、指数平滑预测法、灰色系统理论预测法。**

但是该方法建立的模型都是针对于过去陈旧的数据，在对于整体水质预测的时候，对于水质这种变化性大、及时性强的数据信息来说，最终预测的结果并不总是准确的，且当水质在某个过程加入某个物质导致我们数据和历史数据不一致时，从这个节点开始往后的所有数据都会是错误的，直到水质数据接近历史数据的时候。

回归分析预测法在水质预测领域，主要用于对水质中某几个物质进行预测，通过他们之间的关联性进行预测，那么在中途水质即使发生变化，由于我们观察的仅仅是这几种物质，在我们的数据模型里面，对应的预测值也会随之变化，这种方式便于我们观察各个物质的变化情况，优点在于我们检测的数据项变少，可以通过观察其中几个的变化，然后通过关联关系，得到其他相关元素的大概数值。

指数平滑预测法 结合了全期平均法和移动平均法，对时效性进行了加强，同时也没有忽略全局的数据影响，且加入了加权因子进行简化和动态调整，使得数据的预测更加理想化且可控性更强。

灰色系统理论预测法 主要用于信息不全的情况，利用仅有的信息，提取出我们需要的有价值的信息，再对数据进行预测处理。灰色系统对数据要求低，不需要太多数据，也不需要数据分布规律就可以进行预测，而且计算简便，较为准确。

2. 智能预测法

人工神经网络预测法（ANN） 是一种非线性的，通过一层一层的过滤和预测，从而得到更准确的结果，并且每一层都是基于前一层的结果而新建的，对最新的数据结果有着及时的反馈和处理。它具有较强的泛化能力自组织的能力，与传统的统计预测方法相比，其预测结果具有较高的精度，能够较准确的反应出水质指标内在的变化规律。

支持向量机（SVM） 是基于统计学理论的VC理论和结构风险最小化为原理的一种方法。在进行分类和回归的时候，可视性清晰明了。SVM能够较好地处理具有复杂性、非线性、高维数、局部极小点、小样本等特点的水环境水质指标的预测，具有广泛的推广能力，已成为水质预测研究热点之一。

3. 机理模型预测法

WASP 水质模型预测法 是一种模拟水文动力学、湖泊、河流一维不稳定流以及河口三维不稳定流，研究常规污染物和有毒污染物质在水中迁移转化的**规律**。

S-P水质预测法 主要适用于河流处于充分混合且稳定流动状态，预测相关指标的变化情况。

QUAL-II模型预测法 属于溶解综合水质模型，通过对水生生态系统和各个污染物之间的关系进行分析，更加深入的研究水质问题。该模型已经规模性为众多的河流解决了水质规划、水环境容量等问题。

小结

- 统计预测法：整体的实现较为简单，关于回归的难点在于如何找到回归关系，当有了对应的关系后，就可以通过几个数据项，从而预测其他数据项的数据，从而实现记录少量数据项，得到更多数据项。
- 智能预测法：目前已经有很多的相关机器学习、深度学习的方法，已经有了很多前人实现的模型，在使用上较为简单，但是在对数据的处理上需要结合实际情况，和其他的模型进行合理的结合使用。
- 机理模型预测法：因为这种预测法主要是用于总结变化的规律，然后根据规律建立合适的模型，所以实现难度较大，但是由于其更贴近事实，是从事实出发建立的模型，所以在预测结果上更好。
- 在预测结果不追求非常好的时候，建议使用前面两种，因为实现和使用更简单；否则使用机理模型。

二、基础知识

1. 污水排放标准

表 2.3 城镇污水处理厂污染物排放标准（GB8978—2002）

Tab. 2.3 Pollutant discharge standards for urban sewage treatment plants

水质因子	单位	判定	Classes			
			I		II	III
			A	B		
COD	mg/L	≤	50	60	100	120
BOD	mg/L	≤	10	20	30	60
NH_3-N	mg/L	≤	5	8	25	-
SS	mg/L	≤	10	20	30	50
TN	mg/L	≤	15	20	-	-
TP	mg/L	≤	0.5	1	3	5

2. COD(in)和其他因子的关系

选取Pearson系数和双侧显著性检验。

表 2.4 Pearson 系数和双侧显著性检验对进水 $COD_{(in)}$ 与其他变量之间的相关性分析

Tab. 2.4 Pearson coefficient and bilateral significance test for correlation analysis between inlet $COD_{(in)}$ and other variables

	$COD_{(in)}$	$BOD_{(in)}$	NH_3-$N_{(in)}$	$SS_{(in)}$	$TP_{(in)}$	$TN_{(in)}$
$COD_{(in)}$	1	0.944	0.495	0.232	0.299	0.449
$BOD_{(in)}$	0.944	1	0.480	0.235	0.343	0.443
NH_3-$N_{(in)}$	0.495	0.480	1	0.229	0.368	0.917
$SS_{(in)}$	0.232	0.235	0.229	1	0.509	0.215
$TP_{(in)}$	0.299	0.343	0.368	0.509	1	0.394
$TN_{(in)}$	0.449	0.443	0.917	0.215	0.394	1

下面是出水COD和进水其他元素的关系。

表 2.5 Pearson 系数和双侧显著性检验对出水 COD 与进水水质因子之间的相关性分析

Table. 2.5 Pearson coefficient and bilateral significance test for correlation analysis between effluent $COD_{(out)}$ and inlet water quality factors

	$COD_{(out)}$	$COD_{(in)}$	$BOD_{(in)}$	$NH_3\text{-}N_{(in)}$	$SS_{(in)}$	$TP_{(in)}$	$TN_{(in)}$
$COD_{(out)}$	1	0.523	0.311	0.229	0.221	0.250	0.206
$COD_{(in)}$	0.523	1	0.944	0.495	0.232	0.299	0.449
$BOD_{(in)}$	0.311	0.944	1	0.480	0.235	0.343	0.443
$NH_3\text{-}N_{(in)}$	0.229	0.495	0.480	1	0.229	0.368	0.917
$SS_{(in)}$	0.221	0.232	0.235	0.229	1	0.509	0.215
$TP_{(in)}$	0.250	0.299	0.343	0.368	0.509	1	0.394
$TN_{(in)}$	0.206	0.449	0.443	0.917	0.215	0.394	1

3. 全文的数据处理分析大致框架

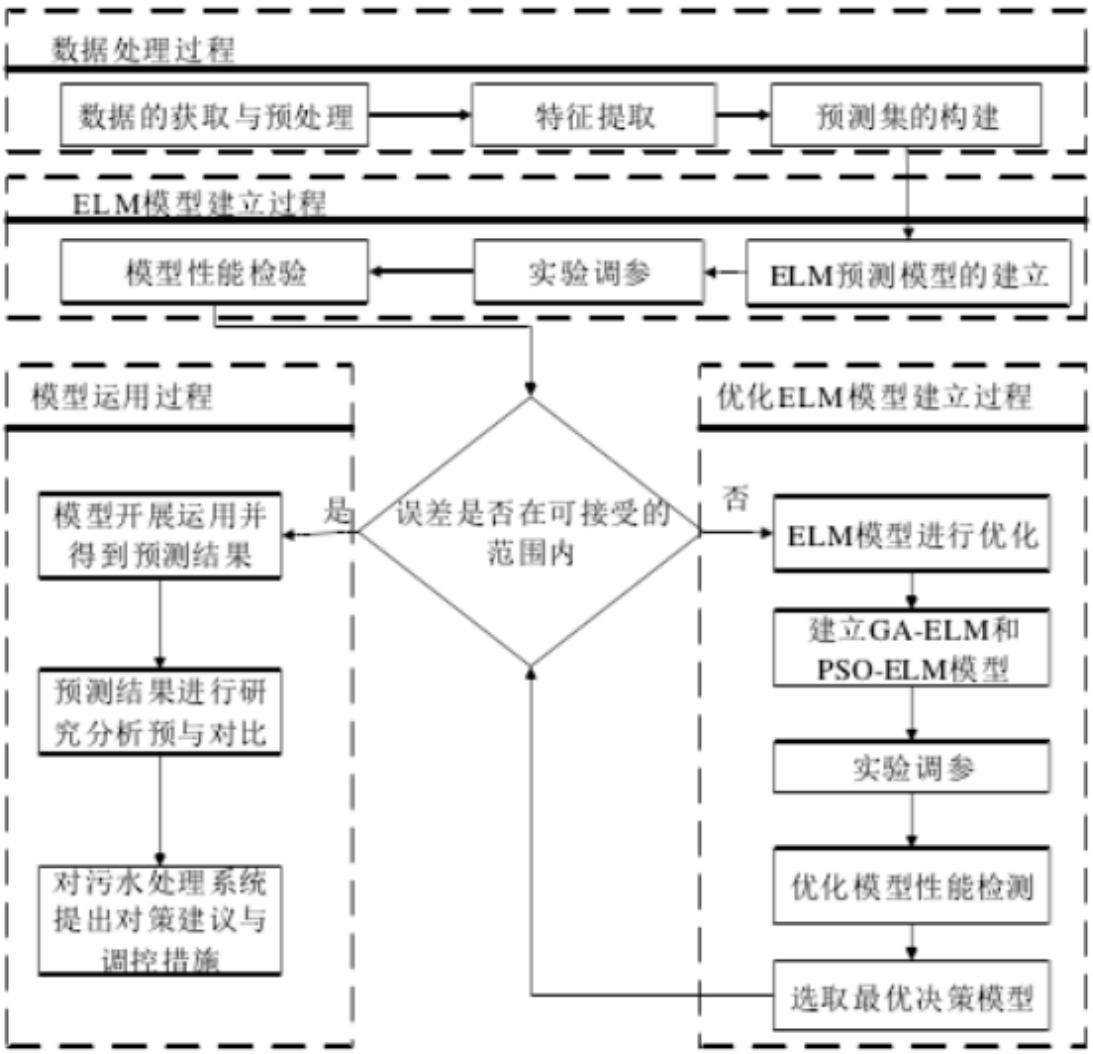

图 1.2 本文的技术路线

Fig.1.2 The technical route of this paper

3.1 数据预处理

根据得到的实际数据，有着一些实际情况可能产生的一些不可避免的数据，如数据不完整、有噪声、不一致等情况，可能有如下情况产生：

- 采集数据的设备存在一定问题；
- 录入数据过程中由于计算机或是人为的错误；

- 在数据传输过程中产生的错误。

针对于可能发生的情况，我们可以认为的提前进行规避，或者针对性的差错，提高数据的完整性和校验效率。

1. **噪声数据的处理**

常用的方法：

- 剔除噪声点；
- 向上、向下填充；（相邻的数据可能存在相似性）
- 平滑处理；（局部均值处理等）
- 根据实际应用实际情况也可以设计出不一样的处理方式。

2. **数据的归一化**

由于我们的实际数据会比较庞大，并且特征项比较多，导致了我们不同的特征项会有不一样的量纲，那么我们进行归一化处理，让数据都在[0,1]之间，使得每个数据项在网络训练的时候，具有同等低位，并且最后的数据结果更加具有可信度。

本文选用的归一化公式如下：**每个值的范围是[-1, 1]**

$$p_n = \frac{2(p_n - p_{min})}{p_{max} - p_{min}} - 1$$

3. **数据的降维**

- （主成分分析）PCA：针对于多个变量的相关性的一种统计方法。利用正交变换将可能存在相关性的一组变量转换成线性不相关，相当于将多个相关性的变量进行合并，实现从高维投射到低维的一种降维方式，并且降维的同时，新的数据也会尽量多的保留原始的数据特征。
- （独立成分分析）ICA：作用于将混合的特征集分成单个的特征集，适合于特征提取。
- （线性判别算法）LDA：一种监督学习的特征降维方式，会使得高维相关性高的特征项，投射到低维的时候，彼此尽量靠近，而不相关的特征之间会尽量的远离。
- （局部线性嵌入）LLE：非线性降维算法，特征是降维后的数据，依然能保留较多的原数据的流行结构[64]，结构的保留使得基于流形学习算法进行降维。
- （核主成分分析）KPCA*：基于PCA，利用核函数，对于非线性、高阶的数据提取都有着良好的效果，并且对原始数据的空间分布没有要求，对于一些位置的数据，完全可以基于该方法进行数据处理，使得数据具有特征性，并在空间分布上具有一定的特征，然后针对于这些数据再利用其他更优秀的算法进行二次处理。（本文则将KPCA运用到水质检测模型中。

3.2 评价标准

本文的评价标准用到了：RMSE（均方根误差）、MAE（平均绝对误差）、MAPE（平均绝对百分比误差）、R^2 （决定系数）。

- RMSE：计算预测值与观测值之差的平方的算数平方根。

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y - \hat{y}_i)^2}$$

- MAE：计算观测值于预测值之差的绝对值的平均值。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y - \hat{y}_i|$$

- MAPE：计算预测值和观测值的百分比的平均值。

$$MAPE = \frac{100}{N}\sum_{i}^{N}|\frac{y_i - \hat{y}_i}{y_i}|$$

- R^2 ： 计算预测值和观测值的拟合程度。

$$R^2 = \sqrt{1 - \sum_{i=1}^{N}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{N}(y_i - \overline{y})^2}$$

三、学习模型

1. ELM

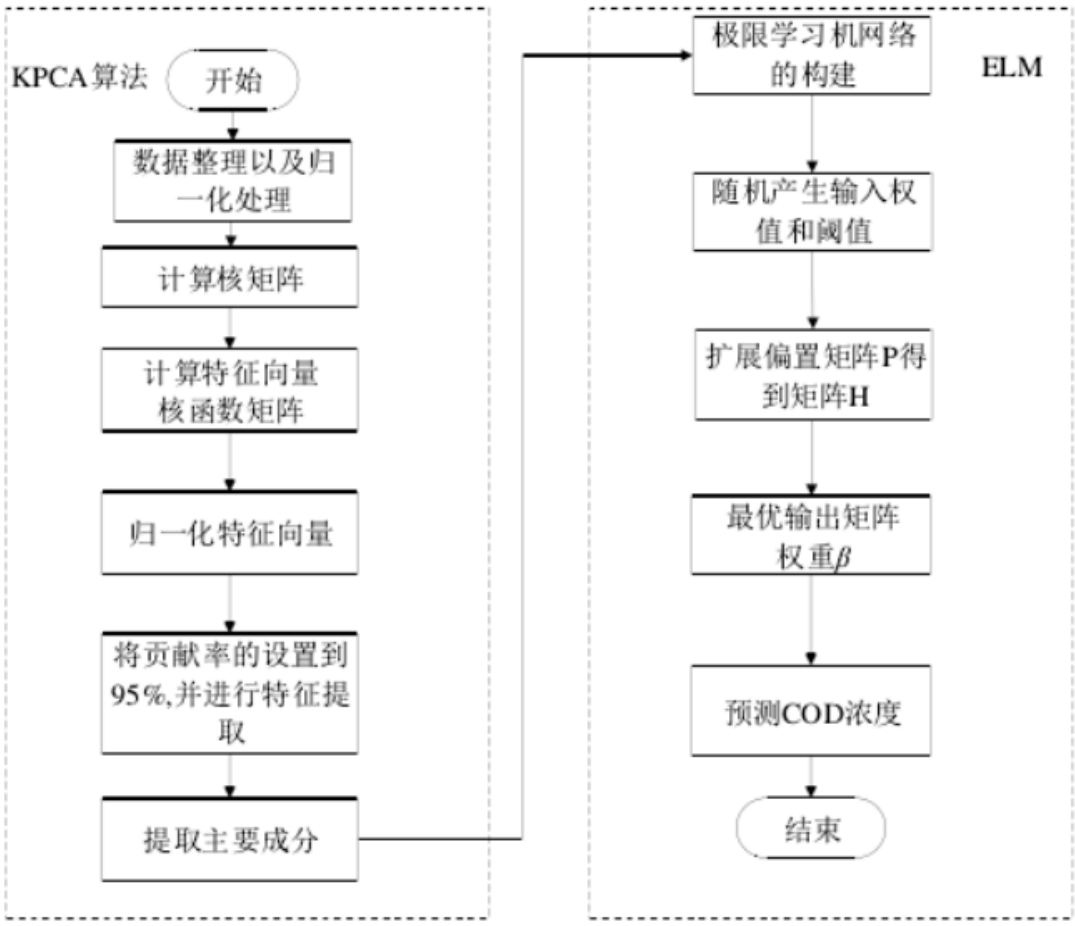

图 3.2 极限学习机模型的建立过程

Fig.3.2 Establishment of extreme learning machine model

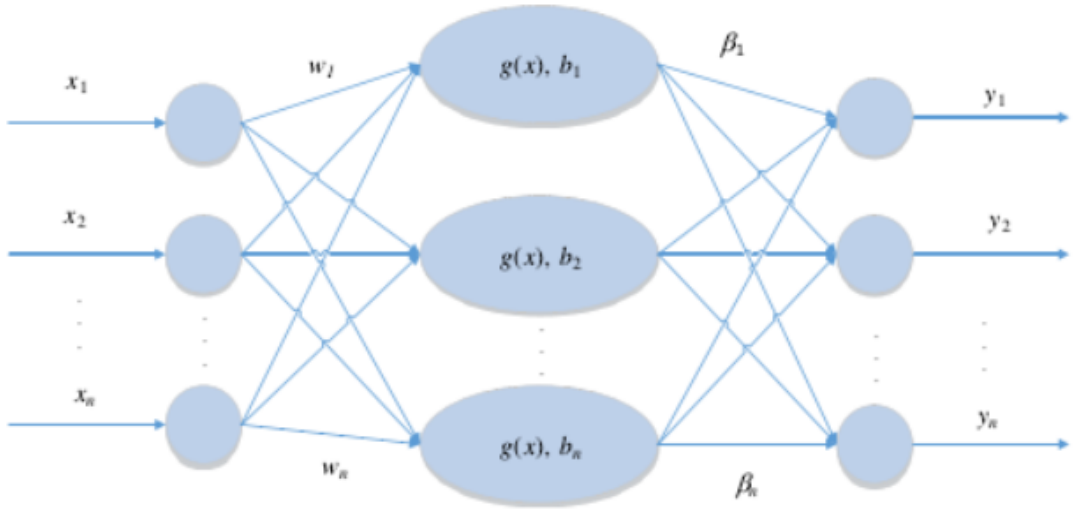

图 3.1 极限学习机网络结构图

Fig. 3.1 Extreme learning machine network architecture

ELM隐含节点个数的选取对于COD的预测有着影响，四个评价指标的结果也大不相同。本文希望R^2 尽量高，所以选取的进出水的时候，ELM的隐含节点个数也不一样。（在实际应用中，根据使用的评价指标，选取不一样的值）

2. 比较

将ELM和BPNN以及LSSVM进行比较。

- BPNN是BP的一种应用，主要是通过隐含层进行层层筛选优化实现的一种算法。
- LSSVM是对于SVM的一种优化应用，将不等式约束改为等式约束，二次规划问题转化为线性方程组问题，简化问题的计算过程，提高求解速度收敛精度。

比较结果：

表 3.5 对比模型的评价指标统计

Tab.3.5 Statistics of evaluation indicators for comparative models

模型	进水预测结果				出水预测结果			
	RMSE	MAE(mg/L)	MAPE(%)	R^2	RMSE	MAE(mg/L)	MAPE(%)	R^2
BPNN	6.521	4.673	2.289	0.849	3.047	2.201	20.450	0.314
LSSVM	5.492	4.369	2.282	0.891	2.203	1.882	15.491	0.579
ELM	3.318	2.437	1.621	0.950	1.310	1.064	8.692	0.661

在整体比较伤，ELM的评分都是最高的。

3. 优化ELM

无论是使用GA还是PSO对ELM进行优化，都是将KPCA的数据提取后的数据传给GA或者PSO，然后由GA和PSO对数据集进行优化处理，数据集的每一条数据作为一个群体的个体，利用各自的运算机制，不断的迭代产生优质个体，将权值和阈值作为适应度值函数的评价标准，不断得到更优的权值和阈值。

因为我们的ELM模型对于权值和阈值的选取不同，就会得到不同的输出结果，在最终得到最优的权值和阈值传输给ELM，由ELM对初始数据集进行训练，得到水质预测模型。

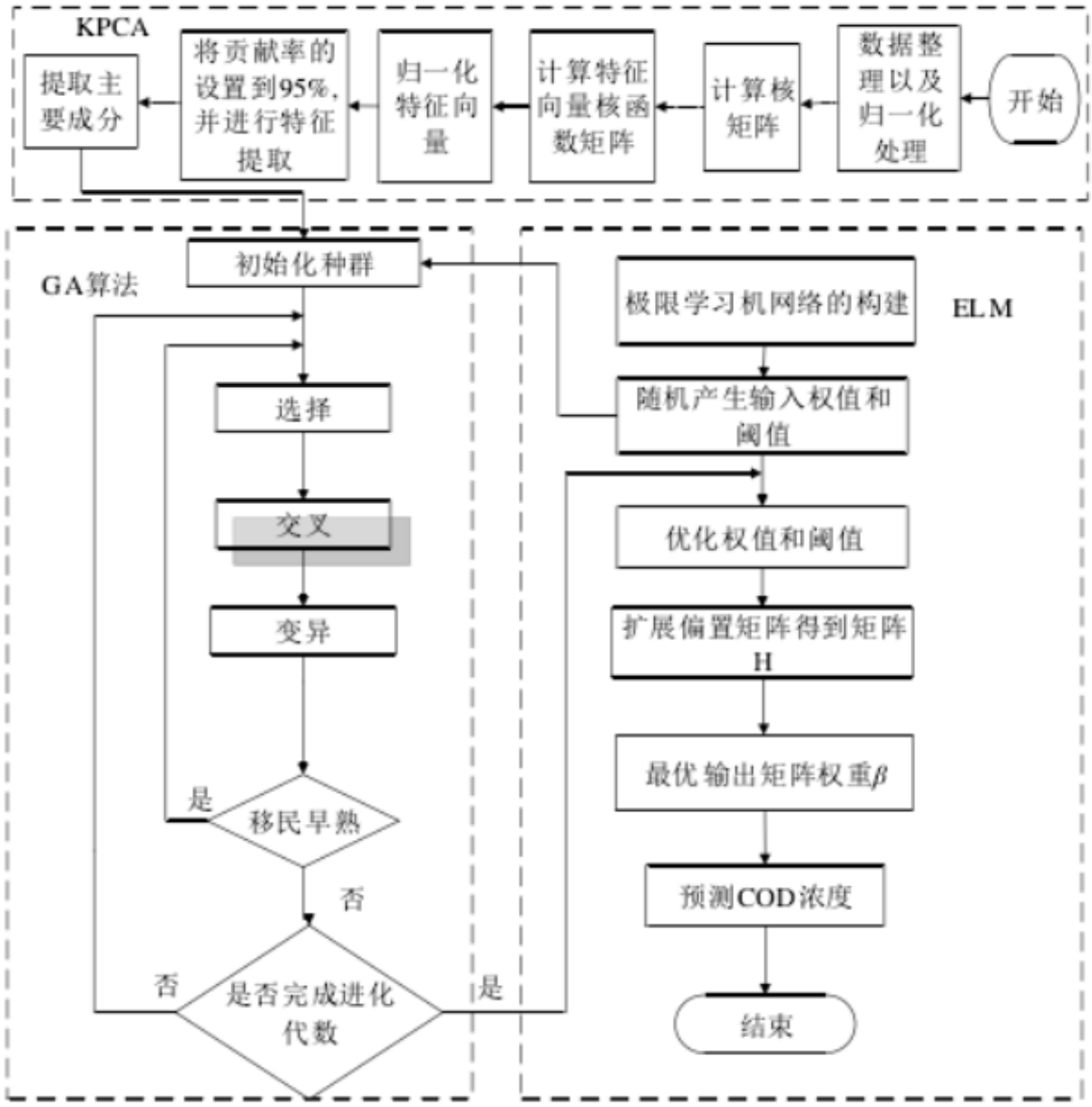

图4.1 GA-ELM模型的实验流程图

Fig.4.1 Experimental flow chart of GA-ELM model

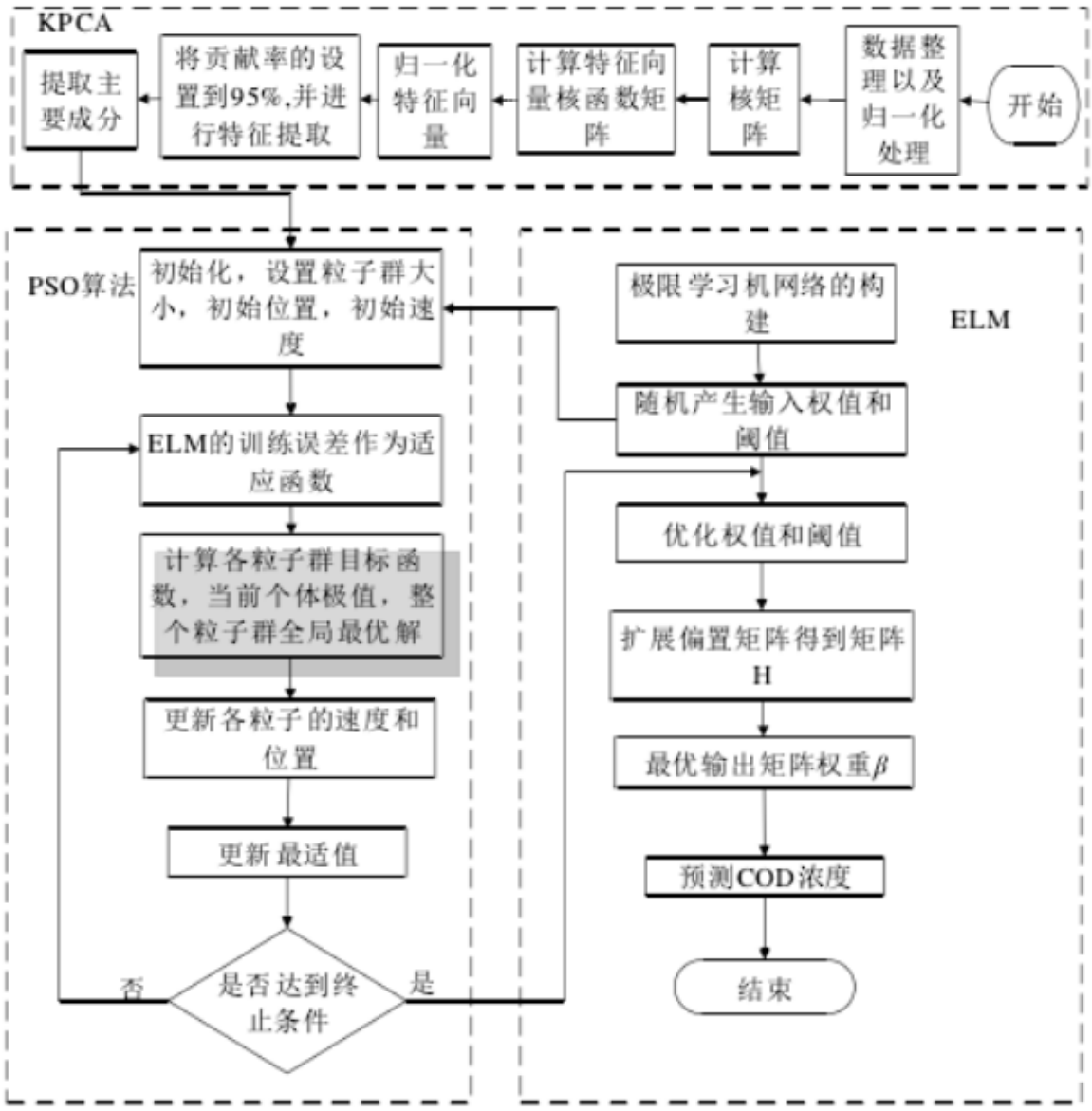

图 4.2 PSO-ELM 模型的实验流程图

Fig.4.2 Experimental flow chart of PSO-ELM model

在整个优化过程中，PSO对于ELM的优化更好，且效果更好，拟合程度更高。

表4.3 对比模型的评价指标统计

Tab.4.3 Statistics of evaluation indexes of contrast model

模型	评价指标			
	RMSE	MAE（mg/L）	MAPE（%）	R^2
GA-BPNN	1.457	1.107	9.167	0.833
PSO-BPNN	1.345	1.081	8.957	0.867
GA-LSSVM	0.944	0.722	7.954	0.901
PSO-LSSVM	0.895	0.622	5.330	0.913
GA-ELM	0.700	0.474	4.073	0.916
PSO-ELM	0.421	0.340	3.072	0.975