DOI: 10.19797/j.cnki.1000-0852.20200025

耦合 LASSO 回归的 HHO-LSVR 中长期径流预报模型

韩晓育[1,2]，郭颍奎[3]

(1. 黄河水利职业技术学院水利工程学院，河南 开封 475004；2. 开封市水生态修复工程技术研究中心，河南 开封 475004；3. 华北水利水电大学 土木与交通学院，河南 郑州 450046)

摘 要：为验证 LASSO 回归在剔除冗余预报因子中的高效性，并探讨基于 LASSO 回归的哈里斯鹰群算法(HHO)与支持向量回归(SVR)耦合的 HHO-LSVR 模型的年径流预报效果，利用 LASSO 回归分别求得各气象因子的拟合系数并以此作为优选预报因子的依据，采用 HHO 算法来优化选择 SVR 模型的控制参数进行年径流预报并设置对照模型，利用 Friedman 检验量化上述两种方法对模型性能的贡献程度。结果显示 HHO-LSVR 拟合和检验效果均为最优，对照检验结果显示 LASSO 回归在提升模型性能上占主要地位。与其他预报模型相比较，HHO-LSVR 具有较高的预报精度和稳定性，可为实际预报提供支撑。

关键词：LASSO 回归；哈里斯鹰群优化算法；支持向量回归；年径流预报

中图分类号：P338+.2 **文献标识码**：A **文章编号**：1000-0852(2021)03-0069-06

0 引言

中长期径流预报研究一直是水文水利领域的热点，其预见期一般在 3d 以上一年以内[1]。常见的预报时间尺度多为旬尺度、月尺度和年尺度等，在指导工农业用水、水电站调度计划等领域发挥着重要作用[2-3]。目前径流预报方法主要关注在优选预报因子和学习模型两个方面。近期有关优选预报因子的方法很多，比如互信息法[4]和核主成分分析法[5]等；研究较多的学习模型有 BP 神经网络[6]、随机森林模型[7]和投影寻踪回归[8]等。以上这些方法已被验证能针对某些实例预报提供较为可靠的精度。雷莉等[9]基于主成分分析法筛选后的预报因子，构造了 BP 神经网络、Elman 神经网络和 PSO-SVR 三种预报模型，结果表明三种模型的预报合格率分别为 75%、83.33%和 91.67%。由此可以看出在相同输入的预报因子条件下，除了 BP 神经网络外，学习模型之间的效果差距并不明显。卢迪等[10]以互信息量来衡量因子的复合相关关系从而达到优选目的，虽然该方法针对碧流河的汛期预报验证了其相对于相关系数法的优越性，但是这种筛选方法需要在预报前期进行试选因子预报来分析因子的可靠性，过程繁杂且存在较大偶然性。谢帅等[11]利用 LASSO 回归来量化因子的回归系数来优选预报因子，再将其作为支持向量回归(SVR)的输入形成耦合的 LSVR 模型，检验结果显示该模型能有效地避免多余干扰的因子项。但是该模型的性能依赖于控制参数的选择，交叉验证法增加了人为调参的随机性。因此，许多先进的智能优化算法例如多组群教学优化算法[12]等，常被用来进行预报模型的调参。此外，许多学者分别从序列处理[13-14]、结果校正[15]和组合预报[16]等其他角度，提供了许多预报改进技术。

年径流预报作为中长期径流预报的重要组成部分，因为有关年尺度的径流及相关资料系列较短，以及水文过程不可避免的非线性、随机性，导致其预报精度往往较低。在这种背景下，支持向量回归发挥了其针对小样本拟合的优势。王文川等[17]基于伊犁河雅马渡站的年径流和相应的 4 个前期气象因子的资料，利用粒子群算法(PSO)寻优支持向量回归的三个控制参数来训练和预报年径流量，但这种 PSO-SVR 模型未考虑输入因子的影响，而且 PSO 算法寻优能力有

收稿日期：2020-02-21

基金项目：水利部堤防安全与病害防治工程技术研究中心开放课题基金项目(2019005)；2019 开封市社科联基金项目(基于智慧水利的开封市水生态健康评价体系研究 ZXSKGH-2019-A010)

作者简介：韩晓育(1984—)，女，河南洛阳人，硕士，讲师，主要研究方向为水利工程、水工结构工程、水利优化建模。E-mail：hxy1507022@163.com

限。总的来看,优选预报因子和学习模型这两种预报模型的改进角度对最终的径流预报精度的影响并不是平等的。由以上分析可以看出,以往的研究多是从优选预报因子或学习模型单方面设置对照,没有分析这两个角度对径流预报精度的贡献程度。综上所述,鉴于前述的 LSVR 模型在优选预报因子的优势和人工调参上的缺点,本文将采用一种新型的哈里斯鹰群算法(HHO)来进行模型调参,提出一种 HHO-LSVR 径流预报模型。同时就 LASSO 回归法、HHO-LSVR 模型分别设置相关系数法和 PSO-LSVR 作为对照,最后据此模型进行雅马渡站的年径流预报,根据结果利用 Friedman 检验来量化改进的贡献程度,验证提出模型在中长期径流预报中的有效性和优越性。

1 基本原理

1.1 LASSO 回归

LASSO(Least absolute shrinkage and selection operator)回归集合了岭回归和子集选择法的优点,能通过将某些相关系数设置为 0 达到有效减少系数数量、保留最好特征的目的,已被证明其强竞争性[18]。LASSO 回归从最小二乘估计的两个缺陷出发:最小二乘估计能保持较小误差但往往方差较大;针对大数量级数据,没有经过筛选而利用了全部的特征,包括冗余的特征。因此,它在岭回归求解最小误差平方和的基础上,增加了对回归系数 R 的约束来牺牲少量的误差保证对序列方差的平衡,其目标函数如下:

$$(b,R)=\text{argmin}\left(\sum_{i=1}^{n}(y_i-b_i-X_iR)^2+\lambda\|R\|_1\right) \tag{1}$$

式中:y_i 为因变量样本,本文中指 n 年中某一年份的年平均径流量;X_i 为自变量样本,本文中指对应该年份的待选预测因子实测值组成的行向量;b 为偏置项;R 为 LASSO 回归系数;λ 为控制参数。

式(1)即选取最优向量 b 和 R 使得误差最小。因为 LASSO 回归是对回归系数的 1 范数进行约束,可能导致个别对应样本间的回归系数为 0。这些对应 0 系数的样本,其对应的径流预报因子就可以被剔除。在本文中,将首先利用 LASSO 回归计算预报因子的回归系数并进行排列,根据结果进行筛选。

1.2 哈里斯鹰群算法(HHO)

哈里斯鹰群优化算法(Harris hawks optimization, HHO)是 Heidari 等在 2019 年提出的一种新型群智能优化算法,基于标准测试函数的结果好于萤火虫算法、灰狼优化算法等[19]。该算法受哈里斯鹰群协作捕捉猎物(兔子)的行为启发,鹰群中的每个个体代表搜索空间中的一个位置,将鹰群当前最优位置当作下一次迭代过程中的待捕捉猎物,并通过随机触发鹰群的软围攻、硬围攻、快速俯冲软围攻和快速俯冲硬围攻 4 种不同机制来达到寻优目的。HHO 算法的一个显著优点就是用户不需要定义任何参数,主要过程如下:

(1)探索阶段:主要发生在算法寻优的前期,此时鹰群往往比较分散,个体将通过式(2)中的两种方式随机移动位置来寻找猎物;

$$\begin{cases} X_m(t)=\dfrac{1}{N}\sum_{i=1}^{N}X_i(t) \\ X(t+1)=\begin{cases} X_{rand}(t)-r_1\left|X_{rand}(t)-2r_2X(t)\right|, q\geq 0.5 \\ (X_{rabbit}(t)-X_m(t))-r_3(LB+r_4(UB-LB)), q<0.5 \end{cases} \end{cases} \tag{2}$$

式中:t 为迭代次数;X 为当前鹰的位置向量;X_{rabbit} 为猎物的位置向量,即上一次迭代后鹰群最优位置;X_m 为鹰群在各个维度上的平均位置;X_{rand} 为随机选出的一只鹰;r_1,r_2,r_3,r_4,q 指(0,1)内产生的均匀分布随机数;N 为鹰的总数。

当种群移动完毕后,鹰群最优位置即为猎物的位置,HHO 将根据目标函数 $f(X)$ 计算适应度值并将其赋予猎物的自身能量 E_{rabbit}。

(2)猎物逃脱阶段:猎物要摆脱鹰群的追击,尝试向其周围逃脱。HHO 算法针对猎物逃脱设定了逃脱能量 E,可以表示为:

$$E=2E_0(1-t/T) \tag{3}$$

式中:T 为最大迭代次数;E_0 为每次迭代过程中猎物的初始能量,对于每只鹰来说是不同的,从(-1,1)中随机产生。HHO 将根据 E 值来触发探索阶段或开发阶段来选择不同的移动机制。

(3)开发阶段:主要发生在算法寻优的后期,HHO 将判别猎物的逃脱能量 E 来随机触发下列 4 种移动机制中的任意一种:

a. 软围攻:

$$X(t+1)=X_{rabbit}(t)-X(t)-E\left|JX_{rabbit}(t)-X(t)\right| \tag{4}$$

式中:J 为(0,2)之间的随机数。

b. 硬围攻:

$$X(t+1)=X_{rabbit}(t)-E\left|X_{rabbit}(t)-X(t)\right| \tag{5}$$

c. 俯冲式软围攻:

$$\begin{cases} Y=X_{rabbit}(t)-E\left|JX_{rabbit}(t)-X(t)\right| \\ Z=Y+S\times LF(D) \end{cases} \tag{6}$$

$$X(t+1)=\begin{cases} Y, & if \quad f(Y)<f(X(t)) \\ Z, & if \quad f(Z)<f(X(t)) \end{cases} \tag{7}$$

式中:S 为(0,1)均匀分布上的 D 维随机向量;LF为由莱维飞行产生的 D 维随机向量。

d. 俯冲式硬围攻:移动选择同俯冲式软围攻相同,只是在 Y、Z 的确定上略有改变,见式(8):

$$\begin{cases} Y=X_{rabbit}(t)-E\left|JX_{rabbit}(t)-X_m(t)\right| \\ Z=Y+S\times LF(D) \end{cases} \tag{8}$$

详细的开发阶段选择移动方式的判别条件见 HHO 算法伪代码(见图 1)。

Algorithm: HHO 算法

Inputs: 已知目标函数 $f(x)$,其搜索空间上、下限 UB 和 LB,设置鹰群个体数量 N 和最大迭代次数 T;

1: 随机初始化种群位置 $X_i(i=1,2,\cdots,N)$。其中,$X_i=(x_{i1},x_{i2},\cdots,x_{iD})$,$D$ 代表维数,初始化兔子的能量 E_{rabbit} 为无限大和兔子的位置 X_{rabbit};

2: While $t<T$ **do**

3: 针对鹰群中的每个个体,轮流计算个体适应度值并与 E_{rabbit} 相比,E_{rabbit} 和 X_{rabbit} 更新为较优适应度值及其相应的位置;

4: **for** 鹰群中的每个个体 **do** 根据式(3)更新相应的兔子逃跑能量 E;

5: **if** $|E|\geq 1$ **then** 根据式(2)更新个体 **end**;

6: **if** $|E|<1$ **then**

7: **if** $rand\geq 0.5\&\&|E|\geq 0.5$ **then** 根据式(4)更新个体位置;

8: **elseif** $rand\geq 0.5\&\&|E|<0.5$ **then** 根据式(5)更新个体位置;

9: **elseif** $rand<0.5\&\&|E|\geq 0.5$ **then** 根据式(6)~(7)更新个体位置;

10: **elseif** $rand<0.5\&\&|E|<0.5$ **then** 根据式(7)~(8)更新个体位置;

11: **end**

12: **end**

13: **end for**

14: $t=t+1$;

15: endwhile

Outputs: 兔子的位置 X_{rabbit} 和能量 E_{rabbit},即为最优点和最优值.

图 1　HHO 算法伪代码

Fig.1 The pseudo-code of HHO algorithm

1.3　支持向量回归(SVR)

支持向量机(Support vector machine,SVM)是一种经典的机器学习模型,可用来解决分类和回归问题,其中支持向量回归(Support vector regression,SVR)特指用来解决回归问题[20]。SVR 的一个区别于传统的回归方法的特点是该模型认为回归值与实际值存在一定误差是可以接受的。SVR 的基本思想是通过核函数把低维空间中的非线性问题变换到高维特征空间中进行线性或非线性回归,对应地在这个高维特征空间中寻求一个最优超平面或最优超曲面来使所有样本点离最优面的间隔最大不超过一定误差。其基本原理如下:

假设某训练样本的集合为$\{(x_i,y_i),i=1,2,\cdots,N;x_i,y_i\in R^M\}$,其中 x 为输入向量,y 为对应 x 的输出向量,N 为样本个数,M 为维数。设经过变换后的非线性回归函数为 $f(x)=<w,\Phi(x)>+b$,其中 w 和 b 为系数向量;$<w,\Phi(x)>$为 w 与 $\Phi(x)$的内积。根据结构风险化最小化原则,通过引入相关损失函数和松弛变量,SVR 的拟合过程将转化为求解以下凸二次优化问题,即:

$$\begin{cases} \min\dfrac{1}{2}w^2+C\sum\limits_{i=1}^{N}(\xi_i+\xi_i^*) \\ s.t.\begin{cases} y_i-\langle w,\Phi(x)\rangle-b\leq\varepsilon+\xi_i \\ -y_i+\langle w,\Phi(x)\rangle+b\leq\varepsilon+\xi_i^* \\ \xi_i,\xi_i^*\geq 0 \end{cases} \end{cases} \tag{9}$$

式中:$\Phi(x)$为非线性映射函数;ε 为不敏感损失函数;ξ_i 和 ξ_i^* 为松弛变量;C 为惩罚因子。

为了求解这个凸规划问题,引入拉格朗日函数和 KKT 条件,回归方程最终确定为:

$$f(x)=\sum_{i=1}^{N}(a_i-a_i^*)K(x,x_i)+b \tag{10}$$

式中:a_i^*、a_i 为二次规划中拉格朗日乘子;$K(x,x_i)$为核函数,应满足 Mercer 条件 $K(x,x_i)=<\Phi(x),\Phi(x_i)>$,本文将采用径向基核函数其表达式为:

$$K(x,x_i)=\exp(-g\|(x-x_i\|^2) \tag{11}$$

式中:$g>0$。因此,SVR 的性能受(C,g,ε)控制。

2　HHO-LSVR 径流预报模型

按照前述的 LASSO 回归、HHO 算法以及 SVR 模型的原理,本文需要首先计算前期气象因子与年径流的 LASSO 拟合系数,再将筛选过的预报因子和年径流值进行归一化,分别作为 SVR 的输入和输出,同时利用 HHO 算法搜寻一组最优的向量(C,g,ε)使得模型检验样本(归一化值)均方误差最小。HHO-LSVR 预报模型实现步骤如下:

Step1:LASSO 回归筛选前期气象因子,选择回归系数显著较大的作为预报因子;

Step2:HHO 算法初始化。定义鹰群规模、最大迭代次数、SVR 参数设置范围,随机生成鹰群个体的位置;

Step3:HHO 中每只鹰的位置即一个向量,对应(C,g,ε),并将该向量作为 SVR 模型的参数选择,模型学习得到均方误差作为适应度值;

Step4:根据 HHO 算法流程进行迭代寻优;

Step5:判断是否达到最大迭代次数,若达到则保留当前最优解并输出,若未达到则重复循行Step4~Step5;

Step6:输出猎物的位置,即(C,g,ε)的值。回代 SVR 模型中作为最佳学习参数进行预报。

3 模型验证

3.1 前期准备

伊犁河位于我国西北部，是一条边界内陆河流，由特克斯河、喀什河和巩乃斯河在新疆维吾尔自治区伊宁市汇合而成。该河全长1 439km，流域面积141 600km^2，其中位于中国国界内河长441km，流域面积61 640km^2，多年平均径流量达117×10^{8}m^3[21]。雅马渡水文站位于该河上，是其上游与中游的分界线，经该站实测年最大径流量156×10^{8}m^3，年最小径流量88×10^{8}m^3，年径流深238.1mm。现有雅马渡水文站23a实测径流资料和其相应的4个前期气象因子数据，4个前期气象因子分别是对应年份的上一年11月至当年3月伊犁气象站的总降雨量x_1、上一年8月欧亚地区月平均纬向环流指数x_2、上一年5月欧亚地区月平均经向环流指数x_3和上一年6月2 800MHz太阳射电流量x_4，数据来源于文献[17]。为了分别考察HHO-LSVR模型中HHO算法和LASSO回归对预报模型效果的提升程度，基于PSO算法和皮尔逊相关系数法设置对照模型PSO-LSVR模型和HHO-SVR模型。模型评价指标采用平均绝对误差*MAE*和平均相对误差*MRE*，相关公式见式(12)~(13)。

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|y_i^*-y_i| \tag{12}$$

$$MRE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i^*-y_i}{y_i}\right|\times 100\% \tag{13}$$

式中：y_i为实测值；y_i^*为预报值。

样本进行归一化处理，采用前20a资料作为训练样本，后3a作为检验样本。为了公平性，HHO算法和PSO算法的种群规模和最大迭代次数同设为30和500，PSO算法的控制参数设置同文献[17]。为避免随机误差，LASSO回归系数的计算采用独立计算10次的十折交叉检验；SVR模型中的学习参数惩罚因子C、核函数参数g和不敏感参数ε的上下限设为[10^{-6} 10^{-6} 0]和[2000 500 1]。本文中所有建模计算均基于MATLAB R2019b环境，SVR利用libsvm 3.24工具箱编程，模型采用10次独立运算并记录结果。

3.2 模型预报及结果分析

LASSO回归计算的结果显示4个x_1、x_2、x_3和x_4相对于年径流的拟合系数分别是0.207、-19.469、42.97和0.025；皮尔逊相关系数计算的结果为0.757、-0.491、0.542和0.448。两者计算结果差异显著，LASSO回归更能区分因子相关性。针对模型HHO-LSVR和PSO-LSVR，采用x_2和x_3作为输入；针对HHO-SVR模型，采用x_1和x_3作为输入。选取10次模型独立运行中的最优预报，拟合检验结果及其评价见图2和表1。

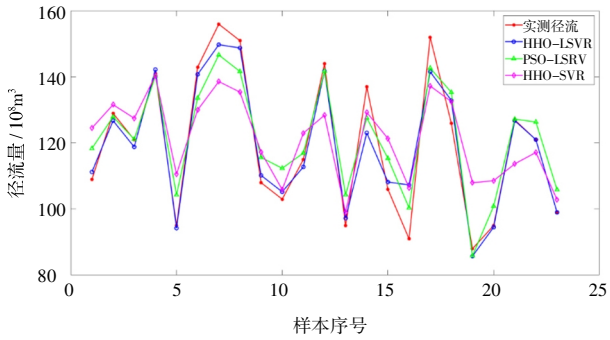

图2 模型拟合检验结果

Fig.2 The fitting and test results of models

表1 最优预报下各模型拟合与检验结果评价

Table1 The evaluation of the fitting and test results of models based on the optimal prediction

评价指标	拟合结果			检验结果		
	HHO-LSVR	PSO-LSVR	HHO-SVR	HHO-LSVR	PSO-LSVR	HHO-SVR
MAE	4.14	6.69	10.99	0.09	4.16	7.00
MRE/%	3.49	5.79	9.56	0.07	3.85	5.85

从拟合检验的结果来看，HHO-LSVR模型的拟合检验径流曲线最贴近实测径流曲线，虽然其中第14、16和17年的训练结果与实测值误差较大，但是从侧面保证了模型的容错性，降低了“过拟合”的可能。表1的评价结果进一步验证了拟合检验曲线的贴合程度，可以得到以下结论：(1)本文提出的HHO-LSVR模型在拟合期和检验期的平均相对误差均小于《水文情报预报规范》[22]中有关中长期径流预报规定的20%误差的标准；(2)HHO-LSVR模型无论是在模型拟合期还是检验期，2个评价指标结果均好于两个对照模型；(3)定性地将HHO-LSVR模型与PSO-LSVR模型、HHO-SVR模型对比可以得出LASSO回归筛选预报因子和利用HHO优化算法调参这两种改进手段均是有效的。

进一步排除随机性对模型的影响，表2记录了10次独立运行结果的最小值、平均值、最劣值和标准差。Friedman检验可以用来定量分析针对最优化问题的数个方案综合性能优劣并进行排序[23]，利用该方法对三个模型的性能进行定量计算，各秩均值记录如表3所示。

表2　各模型10次独立检验预报误差对比

Table2 The comparison of forecast errors of 10 independent tests of different models

指标	检验结果		
	HHO-LSVR	PSO-LSVR	HHO-SVR
最小值	4.28E-06	5.49E-03	1.49E-02
平均值	2.86E-03	7.63E-03	1.82E-02
最劣值	5.51E-03	8.16E-03	1.97E-02
标准差	2.80E-03	1.13E-03	2.24E-03

表3　基于Friedman检验的模型秩均值

Table3 The mean ranks of models based on Friedman test

模型	秩均值
HHO-LSVR	1.50
PSO-LSVR	1.75
HHO-SVR	2.75

HHO-LSVR 模型除了在标准差上大于对照模型,其他项目上均保持相当优势。但是两个对照模型的平均值均较差,标准差可参考价值不大。HHO-LSVR 模型的最好运行结果明显好于其他模型,最劣值接近于 PSO-LSVR 模型的最优值,展现了提出模型的优越性。HHO-LSVR 模型与 PSO-LSVR 模型、HHO-SVR 模型的秩均值之差达到 0.25、1.25,分别代表了 HHO 算法和 LASSO 回归的改进效果。因此,LASSO 回归筛选因子对改善预报精度贡献更大,这也反映了预报因子的选择在中长期径流预报问题中占主导地位,验证了中长期径流预报的首要问题是预报因子的选择,在实际预报中应首先考虑通过 LASSO 回归优选预报因子来保障预报精度。此外,HHO 算法在 SVR 模型参数寻优问题上效果好于 PSO 算法。

4　结论

基于 LASSO 回归在筛选预报因子中的强竞争性,利用 HHO 算法良好的寻优性能来调参支持向量回归,构建了 HHO-LSVR 径流预报模型,针对一个实例的年径流预报结果验证了 HHO-LSVR 模型的有效性和优越性。此外,非参数检验的结果显示 LASSO 回归对于提升模型性能的贡献程度占主导地位,HHO 算法优化 SVR 模型居其次。本文提出的 HHO-LSVR 中长期径流预报模型是合理可行的,LASSO 回归筛选预报因子的方法值得推广。未来的研究工作将利用其他实例进一步研究验证该模型在旬尺度、月尺度上的预报效果,同时进一步试验 LASSO 回归与其他学习模型耦合的效果。

参考文献:

[1] 石继海, 宋松柏, 李航. 中长期径流预报模型优选研究[J]. 西北农林科技大学学报(自然科学版), 2019,47(7):147-154. (SHI Jihai, SONG Songbai, LI Hang. Optimization of mid-long term runoff forecasting models [J]. Journal of Northwest A&F University (Natural Science Edition), 2019,47(7):147-154. (in Chinese))

[2] 何振奇, 乔光建. 基于多项式回归模型的枯季径流预报与分析[J]. 南水北调与水利科技, 2010,8(5):85-88. (HE Zhenqi, QIAO Guang jian. Polynomial regression model based on the low flow forecasting and analysis [J]. South-to-North Water Transfers and Water Science & Technology, 2010,8(5):85-88. (in Chinese))

[3] 李继伟, 纪昌明, 张新明, 等. 基于支持向量机的水电站中长期径流组合预报[J]. 水电能源科学, 2013,31(11):13-16. (LI Jiwei, JI Changming, ZHANG Xinming, et al. Medium and long-term runoff combinational forecast based on support vector machine [J]. Water Resources and Power, 2013,31(11):13-16. (in Chinese))

[4] 丁公博, 农振学, 王超, 等. 基于 MI-PCA 与 BP 神经网络的石羊河流域中长期径流预报[J]. 中国农村水利水电, 2019(10):66-69. (DING Gongbo, NONG Zhenxue, WANG Chao, et al. Long-term runoff forecasting model based on MI-PCA and BP neural network in Shiyang River basin [J]. China Rural Water and Hydropower, 2019(10):66-69. (in Chinese))

[5] 杨易华, 罗伟伟. 基于 KPCA-PSO-SVM 的径流预测研究[J]. 人民长江, 2017,48(3):44-47. (YANG Yihua, LUO Weiwei. Research on runoff forecast based on KPCA-PSO-SVM [J]. Yangtze River, 2017,48(3):44-47. (in Chinese))

[6] 刘勇, 陈元芳, 王银堂, 等. 基于 OSR-BP 神经网络的丹江口秋汛期径流长期预报研究[J]. 水文, 2010,30(6):32-36. (LIU Yong, CHEN Yuanfang, WANG Yintang, et al. Long-term forecasting for autumn flood season in Danjiangkou Reservoir basin based on OSR-BP neural network [J]. Journal of China Hydrology, 2010,30(6):32-36. (in Chinese))

[7] 赵铜铁钢, 杨大文, 蔡喜明, 等. 基于随机森林模型的长江上游枯水期径流预报研究[J]. 水力发电学报, 2012,31(3):18-24+38. (ZHAO Tongtiegang, YANG Dawen, CAI Ximing, et al. Predict seasonal low flows in the upper Yangtze River using random forests model [J]. Journal of Hydroelectric Engineering, 2012,31(3):18-24+38. (in Chinese))

[8] 徐刚, 余冲. 基于遗传算法的参数投影寻踪回归径流预报模型及应用[J]. 水电能源科学, 2013,31(9):20-23. (XU Gang, YU Chong. Application of parametric projection pursuit regression based on genetic algorithm in runoff forecasting [J]. Water Resources and Power, 2013,31(9):20-23. (in Chinese))

[9] 雷莉, 王超. 基于 BP、Elman、PSO-SVR 三种预报模型在石羊河流域的应用比较[J]. 中国农村水利水电, 2019(9):28-32. (LEI Li, WANG Chao. Comparison and application of three prediction mod-

els based on BP, Elman and PSO-SVR in Shiyang River basin [J]. China Rural Water and Hydropower, 2019(9):28-32. (in Chinese))

[10] 卢迪, 周惠成. 基于互信息量与 BP 神经网络的中长期径流预报方法研究[J]. 水文, 2014,34(4):8-14+67. (LU Di, ZHOU Huicheng. Medium and long-term runoff forecasting based on mutual information and BP neural network [J]. Journal of China Hydrology, 2014,34(4):8-14+67. (in Chinese))

[11] 谢帅, 黄跃飞, 李铁键, 等. LASSO 回归和支持向量回归耦合的中长期径流预报[J]. 应用基础与工程科学学报, 2018,26(4):709-722. (XIE Shuai, HUANG Yuefei, LI Tiejian, et al. Mid-long term runoff prediction based on a Lasso and SVR hybrid method [J]. Journal of Basic Science and Engineering, 2018,26(4):709-722. (in Chinese))

[12] 崔东文. 多组群教学优化算法-神经网络-支持向量机组合模型在径流预测中的应用[J]. 水利水电科技进展, 2019,39(4):41-48+84. (CUI Dongwen. Combined model of multi-group teaching optimization algorithm-neural network-support vector machine in runoff prediction application [J]. Advances in Science and Technology of Water Resources, 2019,39(4):41-48+84. (in Chinese))

[13] 赵雪花, 桑宇婷, 祝雪萍. 基于 CEEMD-GRNN 组合模型的月径流预测方法[J]. 人民长江, 2019,50(4):117-123+141. (ZHAO Xuehua, SANG Yuting, ZHU Xueping. Monthly runoff prediction based on CEEMD and GRNN hybrid model [J]. Yangtze River, 2019,50(4):117-123+141. (in Chinese))

[14] 周正弘, 粟晓玲. 基于熵谱理论的月径流预报[J]. 西北农林科技大学学报(自然科学版), 2019,47(5):146-154. (ZHOU Zhenghong, SU Xiaoling. Monthly streamflow forecasting based on entropy spectral theory [J]. Journal of Northwest A&F University (Natural Science Edition), 2019,47(5):146-154. (in Chinese))

[15] 张强, 王本德, 何斌, 等. SSA 分解预测校正模型在年径流预报中的应用[J]. 武汉理工大学学报, 2010,32(7):152-155. (ZHANG Qiang, WANG Bende, HE Bin, et al. Research of dedium-long terms runoff predictor-corrector model based on SSA decomposition [J]. Journal of Wuhan University of Technology, 2010,32(7):152-155. (in Chinese))

[16] 徐炜, 张弛, 彭勇, 等. 基于多模型预报信息融合的中长期径流预报研究[J]. 水力发电学报, 2013,32(6):11-18. (XU Wei, ZHANG Chi, PENG Yong, et al. Study on medium and long-term hydrological forecasting based on data fusion [J]. Journal of Hydroelectric Engineering, 2013,32(6):11-18. (in Chinese))

[17] 王文川, 和吉, 邱林. 基于 PSO 的 SVM 年径流预报模型研究[J]. 人民黄河, 2012,34(4):17-19. (WANG Wenchuan, HE Ji, QIU Lin. SVM annual runoff forecasting model Based on PSO [J]. Yellow River, 2012,34(4):17-19. (in Chinese))

[18] Tibshirani R. Regression shrinkage and selection via the Lasso [J]. Journal of the Royal Statistical Society, 1996,58(1):267-288.

[19] Heidari A A, Mirjalili S, Faris H, et al. Harris hawks optimization: algorithm and applications [J]. Future Generation Computer Systems, 2019,97:849-872.

[20] Vapnik V. Statistical Learning Theory [M]. New York: Wiley, 1998.

[21] 朱道清. 中国水系辞典[M]. 青岛: 青岛出版社, 2007. (ZHU Daoqing. Chinese Water System Dictionary [M]. Qingdao: Qingdao Publishing House, 2007. (in Chinese))

[22] GB/T 22482-2008, 水文情报预报规范 [S]. (GB/T 22482-2008, Standard for Hydrological Information and Hydrological Forecasting [S]. (in Chinese))

[23] 赵嘉, 谢智峰, 吕莉, 等. 深度学习萤火虫算法[J]. 电子学报, 2018,46(11):2633-2641. (ZHAO Jia, XIE Zhifeng, LV Li, et al. Firefly algorithm with deep learning [J]. Acta Electronica Sinica, 2018,46(11):2633-2641. (in Chinese))

HHO-LSVR Model Combined with LASSO Regression for Medium and Long-Term Runoff Forecasting

HAN Xiaoyu[1,2], GUO Yingkui[3]

(1. *School of Hydraulic Engineering, Yellow River Conservancy Technical Institute, Kaifeng* 475004, *China*; 2. *Kaifeng Water Ecological Restoration Engineering Technology Research Center, Kaifeng* 475004, *China*; 3. *School of Civil Engineering and Communication, North China University of Water Resources and Electric Power, Zhengzhou* 450046, *China*)

Abstract: To verify the efficiency of LASSO regression in removing redundant forecasting factors and discuss the annual runoff forecast effects of HHO-LSVR model based on the Harris Hawks Optimization (HHO) algorithm and Support Vector Regression (SVR) with LASSO regression, this paper applied LASSO regression to obtain the fitting coefficients of each meteorological factor and used it as the basis for the preferred forecasting factors. The HHO algorithm was used to optimize the control parameters of SVR model for annual runoff forecasting and control models were set up. Besides, Friedman test was utilized to quantify the contribution of the above two methods to model performance. The results show that the HHO-LSVR model has the best fitting and test results, and the control test results show that LASSO regression plays a major role in improving the performance of the model. Compared with other models, the HHO-LSVR model has higher forecast accuracy and stability, which can provide support for actual forecast.

Key words: LASSO regression; Harris hawks optimization algorithm; support vector regression; annual runoff forecasting