基于 SHO-SA 算法的案例推理预测模型特征权重优化

严爱军[1,2,3], 丁 凯[1,2]
(1. 北京工业大学信息学部, 北京 100124; 2. 数字社区教育部工程研究中心, 北京 100124;
3. 城市轨道交通北京实验室, 北京 100124)

摘 要: 针对案例推理(case-based reasoning,CBR)检索过程中特征权重的分配结果直接影响 CBR 预测模型性能的问题,提出了一种基于自私牧群优化-模拟退火(selfish herd optimizer-simulated annealing,SHO-SA)算法的特征权重优化分配方法. 首先,将 CBR 预测模型的均方根误差定义为 SHO 算法和 SA 算法中权重寻优的适应度;然后,通过 SHO 算法的牧群运动、捕食及恢复等步骤得到种群内最小均方根误差所对应的权重;最后,采用 SA 算法对上述权重进行随机搜索,从而获得特征权重的近似最优解. 采用加州大学欧文分校(University of California Irvine,UCI)数据集中的 5 个标准回归数据集进行实验,结果表明该方法与一些典型的优化方法相比可以显著提高 CBR 预测模型的精度,说明 SA 算法能够改善 SHO 算法陷入局部最优的问题.

关键词: 案例推理; 案例检索; 特征权重; 自私牧群优化; 模拟退火; 分配权重
中图分类号: U 461; TP 308 **文献标志码**: A **文章编号**: 0254 - 0037(2022)04 - 0355 - 12
doi: 10.11936/bjutxb2020090007

Feature Weights Optimization Based on SHO-SA Algorithm for Case-based Reasoning Prediction Model

YAN Aijun[1, 2, 3], DING Kai[1, 2]
(1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China;
2. Engineering Research Center of Digital Community, Ministry of Education, Beijing 100124, China;
3. Beijing Laboratory for Urban Mass Transit, Beijing 100124, China)

Abstract: Performance of the case-based reasoning (CBR) prediction model is directly affected by feature weight allocation in the retrieval process. A method based on selfish herd optimizer-simulated annealing (SHO-SA) algorithm was proposed to calculate the feature weights. The root mean square error (RMSE) of the CBR prediction model was first defined as the fitness function in the SHO algorithm and the SA algorithm to evaluate the rationality of the weight distribution. Then, the weights distribution with the minimum RMSE in the population were obtained through the herd movement, predation and recovery steps of the SHO. Finally, SA algorithm was employed to search randomly based on the above weight, and an approximate optimal solution of the feature weights was obtained. Performance evaluation was carried out by five benchmark regression datasets from University of California Irvine (UCI) datasets. Results show that compared with other typical optimization methods, the proposed method can improve the

收稿日期: 2020-09-10; 修回日期: 2020-11-13
基金项目: 国家自然科学基金资助项目(61873009); 北京市自然科学基金资助项目(4192009)
作者简介: 严爱军(1970—), 男, 教授, 主要从事过程建模与控制、人工智能及应用方面的研究, E-mail: yanaijun@ bjut.edu.cn

accuracy of the CBR prediction model significantly. Meanwhile, it illustrates that the matter of SHO suffering from local minima can be mended by SA algorithm.

Key words: case-based reasoning (CBR); case retrieval; feature weight; selfish herd optimizer (SHO); simulated annealing; weight allocation

案例推理(case-based reasoning, CBR)是一种基于知识的解决问题方法[1],它在解决问题的过程中主要包括 4 个步骤,即案例检索、案例重用、案例修正和案例存储. CBR 的工作机制是通过检索相似的问题并利用其解决方案来解决当前问题[2],该方法在分类、诊断、预测等领域得到了广泛应用[3-5]. 值得关注的是,案例检索性能的好坏对 CBR 求解性能的高低有着至关重要的影响[6],为了提高案例检索的性能,大量针对案例检索的方法被提出[7-9]. 研究结果表明,在案例检索过程中特征权重的分配结果直接影响案例检索的性能,因而,特征权重的分配方法具有重要的研究价值.

目前,特征权重的分配方法分为主观法、客观法和主客观结合法[10]. 主观法是指根据领域专家对相关领域的认知和经验完成特征权重的分配[11],如专家打分法、调查统计法、层次分析(analytic hierarchy process, AHP)法[12]、网络层次分析(analytic network process, ANP)法[13]以及相关改进方法[14-16]等. 其中,专家打分法和调查统计法主要依据专家经验和数理统计等方法确定特征权重,主观性较强;AHP 法能对特征变量进行定性定量的分析;ANP 法可以解决多个评价指标间存在依赖和反馈的问题;相关改进方法在信息确定的前提下降低了不确定性,但依旧存在主观不确定性以及适用范围的局限性. 客观法则是依据数据以及数据之间的关系确定特征权重,可以避免决策人主观判断带来的影响. 文献[17]采用神经网络等方法确定特征权重,提高了分类性能,但网络的结构不易确定;熵权法利用熵表达信息量的特点分配权重,然而所有熵值都趋近于 1 时,会导致权重分配不符合实际[18];文献[19]采用粗糙集等方法通过评价条件属性和决策属性实现了权重的分配,然而有时对事物的评价只根据影响它的各因素做出综合评价, 导致决策属性不可知;文献[20]提出的注水法分配特征权重提高了分类精度,但对于案例库的变化难以适应;文献[21]采用遗传算法通过随机搜索来分配权重,在一定程度上解决了上述问题,但遗传算法容易陷入局部最优,案例样本较少时,容易导致权重与实际不符;文献[22]提出了一种基于模拟退火粒子群的优化算法,其中还引入了高斯变异和遗传算法的杂交概念,取得了较好的效果,但在一定程度上增加了算法的复杂度;文献[23]提出了一种遗传模拟退火算法,在一定程度上缓解了陷入局部最优的问题,但计算参数具有一定的依赖性. 主客观结合法的主要思想是利用主观法和客观法各自的优势来确定特征权重,但是计算比较复杂,并且存在着较大的随机性[24]. 因此,对于特征权重的分配方法需要进一步研究.

针对上述特征权重分配方法存在的问题,本文提出一种基于自私牧群优化-模拟退火(selfish herd optimizer-simulated annealing, SHO-SA)算法的特征权重优化方法. 相比传统优化方法,SHO 算法将整个群体分为猎物和捕食者 2 组,并考虑个体的独立行为,这使得 SHO 算法在一定程度上缓解了陷入局部最优、过早收敛等问题[25]. 但 SHO 算法在利用搜索空间上存在一些不足,可能陷入局部最优[26],而 SA 算法局部搜索能力较好,能以一定概率接受初始权重附近新的权重,因此,将 SHO 算法与 SA 算法融合可以有效缓解陷入局部最优的问题[23]. 其中,SHO 算法通过不同的运动阶段增加特征权重的多样性;通过捕食阶段去除适应度较差的特征权重组;通过恢复阶段保证每次迭代的权重规模[25]. 本文根据 SHO 和 SA 原理,首先,设计了权重寻优的适应度函数;其次,通过寻优计算的运动阶段、捕食阶段和恢复阶段选择最佳适应度所对应的权重;然后,采用 SA 算法进一步优化权重对象;最后,通过对比实验验证了本文方法的有效性.

1 CBR 预测模型及问题分析

1.1 CBR 预测模型

CBR 经过近 40 年的发展得到了广泛应用,它的主要思想是通过评估当前问题与存储在案例库中的源案例之间的相似性,继而在案例库中检索出相同或相似的案例,并利用这个(或这些)源案例的解决方案来解决当前问题[27]. CBR 预测模型的求解过程主要包括案例检索、案例重用、案例修

正和案例存储 4 个步骤，如图 1 所示. 求解之前，需要以一定的方式表示源案例，即案例表示. 下面对案例表示和上述 4 个步骤分别进行介绍.

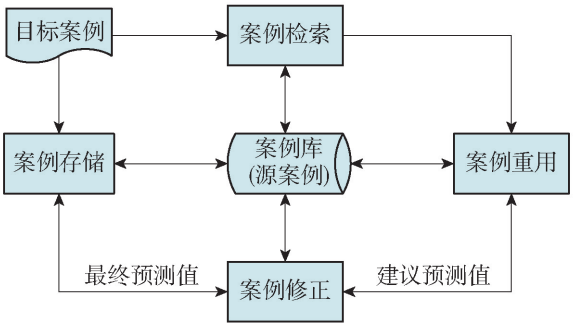

图 1　CBR 预测模型结构原理

Fig. 1　Structure principle of CBR prediction model

1）案例表示

以具有数值类型的案例为例，案例库中的源案例可以表示为

$$C_t = (X(t); y(t)), t = 1, 2, \cdots, A \tag{1}$$

式中：A 为源案例总数；$y(t)$ 为第 t 个案例的输出值；$X(t)$ 为第 t 个案例的输入数据，可以表示为

$$X(t) = (x_{1,t}, \cdots, x_{j,t}, \cdots, x_{J,t}) \tag{2}$$

式中：$x_{j,t}$ 为第 j 个输入特征的归一化特征值；J 为特征总数.

2）案例检索

设目标案例（即待求解案例）为 $X = (x_1, x_2, \cdots, x_j, \cdots, x_J)$，待求解的输出值记为 y. 采用基于欧氏距离的相似度评估策略计算 X 与 $X(t)$ 的相似度[28]，公式为

$$S_t = 1 - \sqrt{\sum_{j=1}^{J} \omega_j (x_j - x_{j,t})^2}, t = 1, 2, \cdots, A \tag{3}$$

式中 ω_j 为第 j 个特征的权重，所有的特征权重需满足

$$\sum_{j=1}^{J} \omega_j = 1, \omega_j \geqslant 0 \tag{4}$$

将式(3)计算出的 A 个相似度从大到小排列，依次选取前 M 个案例，至此完成案例检索.

3）案例重用与案例修正

若检索到的源案例与目标案例的相似度较高，可以不对源案例的输出值进行调整，直接将其作为目标案例的建议输出；否则，需要对源案例的输出值进行改编. 经过重用，输出结果如果不正确，此时须添加案例修正步骤，以修正错误的输出，从而得到目标案例的正确输出结果.

4）案例存储

获得目标案例的正确输出结果后，形成一个新案例，将其以式(1)所示的源案例表示形式存储于案例库中，从而丰富了案例库的记录.

1.2　问题分析

式(3)中的 ω_j 的大小表示每个特征的重要程度，值越大表示该特征越重要. 满足式(4)的特征权重分配组合对相似度的计算结果会产生影响，准确合理地分配权重可以提高 CBR 模型的求解性能[29].

针对权重分配问题，许多研究人员从最初的主观法到客观法以及二者组合法做了大量的方法研究. 主观法包括专家打分法、AHP 法等；客观法包括膜计算法[30]、注水法[31]、遗传算法[32]、神经网络法[33]等；主客观分配权重法是将主观法与客观法结合起来使用. 因为主观法存在主观不确定性和适用范围的局限性，而组合法[24]存在较大的随机性等不足之处，所以目前多采用客观法分配权重. 表 1 列出了各种权重分配方法的特点.

表 1　权重分配方法的优缺点

Table 1　Advantages and disadvantages of weight distribution methods

方法类型	具体方法	优点	缺点
主观法	专家打分法	方法简单、成熟	主观性较强，有局限性
	AHP 法	将复杂问题层次化，将定量问题定性化	
客观法	膜计算法	分布式计算和并行式计算	基本膜个数不易确定
	注水法	直观、简洁	难以适应案例库变化
	遗传算法	运行简单，鲁棒性强	易陷入局部最优
	神经网络法	通过网络内部自适应算法不断修改连接权值分布	网络结构不易确定，收敛速度慢
组合法	组合法	同时体现主、客观信息	随机性较大

由表 1 可知,相比之下,客观法的优势比较明显,但这些方法自身还存在一些问题. 遗传算法虽然能在一定程度上解决其他客观方法的问题,但它容易陷入局部最优,因而,对于权重的分配方法还需进一步研究. Fausto 等[25]提出的 SHO 方法是对猎物群和捕食者 2 种类型的搜索算子之间相互作用的模拟,通过运动阶段、捕食阶段和恢复阶段进行寻优,在群寻优算法中具有一定的竞争力. 虽然 SHO 算法将种群划分为 2 种操作算子,并考虑个体的独立操作,但在利用搜索空间上存在不足,导致 SHO 算法也有可能会陷入局部最优[34],而 SA 算法将搜索过程中的时间变化特性与趋于零的概率跳变特性相结合,在局部最优解附近随机产生新解,采用 Metropolis 准则接受产生的最优解,可以有效地避免局部收敛[35-36]. 因此,本文拟将 SHO 和 SA 算法结合起来对特征权重进行寻优.

2 权重的 SHO-SA 分配算法

基于上述分析,本节介绍 SHO 与 SA 算法结合起来形成的 SHO-SA 算法,并给出了算法伪代码.

2.1 算法实现

根据 SHO 和 SA 原理,SHO-SA 算法结构图如图 2 所示. 采用 SHO-SA 算法优化分配权重前,需要将随机生成的 N 组权重对象分成猎物和捕食者 2 组,猎物组的规模大于捕食者的规模. 猎物组的成员集合用 H 表示,数量为 N_h;另一部分为捕食者,成员集合用 P 表示,数量为 N_p. 由于每个个体的生存能力与最安全和最危险的个体相关,因此,为每个个体分配一个生存值 S(如图 3 所示)来代表它们的生存能力[25]. 根据生存能力和适应度,图中的 SHO 算法主要通过运动阶段、捕食阶段和恢复阶段实现权重的初步寻优过程,得到当前迭代的最优权重,当未达到迭代次数时,采用 SA 算法继续寻找当前最佳权重至温度最低,从而实现完整的单次权重寻优.

下面首先介绍适应度函数的定义,然后介绍采用 SHO 算法的权重寻优,最后介绍采用 SA 算法进行特征权重的进一步优化,从而得到最优特征权重的近似解.

2.1.1 定义 SHO-SA 算法的适应度

本文将 CBR 预测模型的均方根误差(root mean

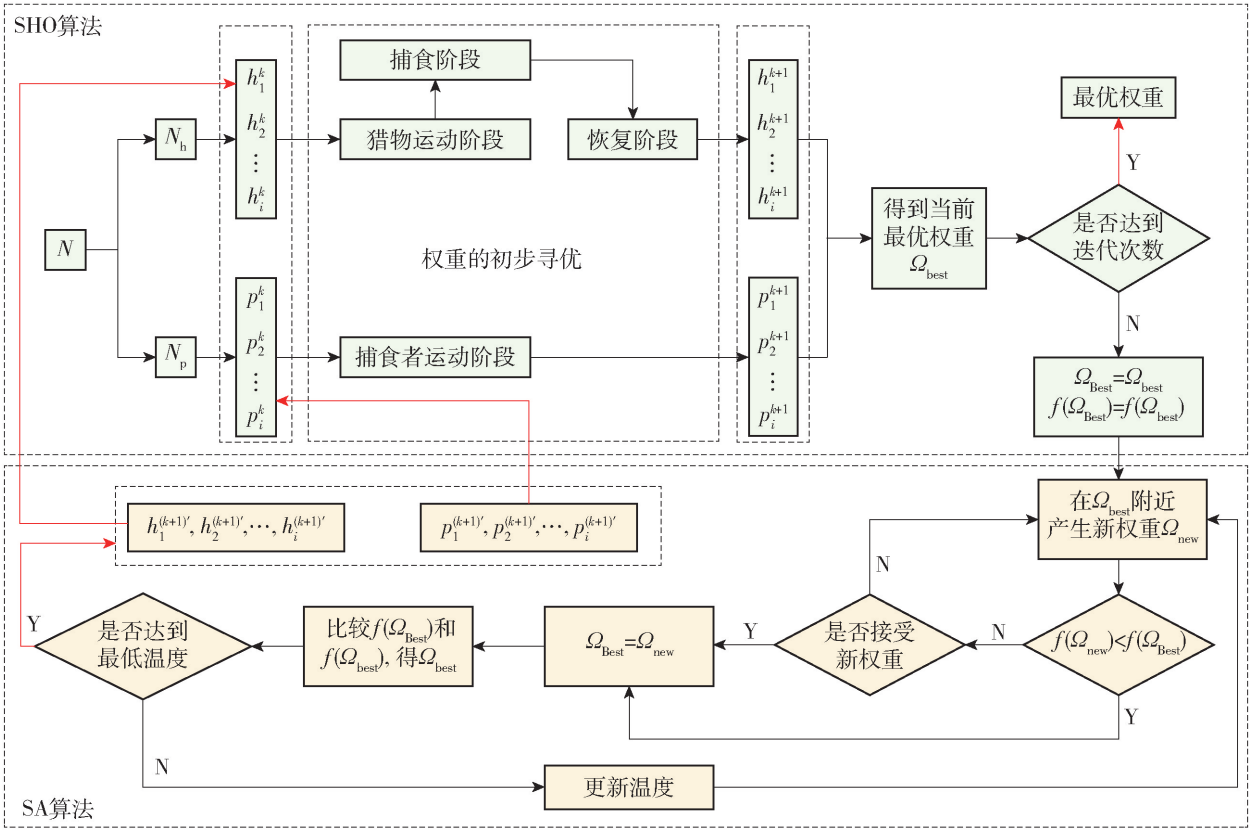

图 2 SHO-SA 算法结构

Fig. 2 Structure of SHO-SA algorithm

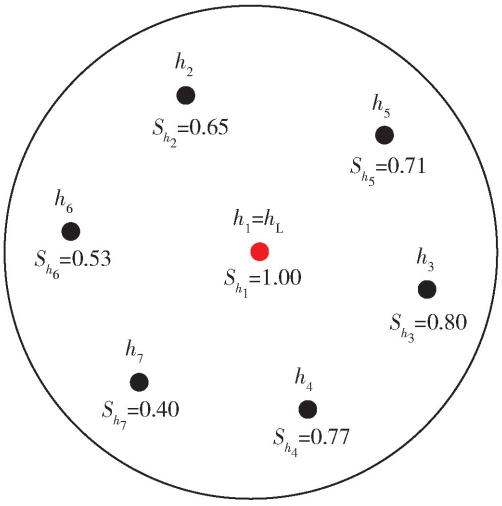

图 3　生存值[25]

Fig. 3　Survival value[25]

square error，RMSE）定义为 SHO-SA 算法中权重寻优的适应度函数 f，即

$$f=\sqrt{\frac{\sum_{l=1}^{m}(y_l-\hat{y}_l)^2}{m}} \tag{5}$$

式中：m 为测试集中案例数量；y_l 为第 l 个测试案例的实际输出值；$\hat{y}_l$ 为第 l 个测试案例的预测值.

2.1.2　基于 SHO 算法的权重寻优

根据 SHO 算法原理[25]，在求解前将随机生成的 N 组权重分为猎物 H 和捕食者 P 两部分，猎物 H 为权重总数的 70% ~90%，用 N_h 表示，捕食者 P 则为 30% ~10%，用 N_p 表示，并通过式（5）及生存值的计算公式

$$S_{\Omega(i)}=\frac{f(\Omega(i))-f_{\text{worst}}}{f_{\text{best}}-f_{\text{worst}}},\ i=1,2,\cdots,N \tag{6}$$

分别计算这两部分中的每组权重所对应的适应度和生存值. 式中：$f(\Omega(i))$ 为 N 组权重中第 i 组所对应的适应度，$\Omega(i)=(\omega_{i,1},\omega_{i,2},\cdots,\omega_{i,j},\cdots,\omega_{i,J})$，$\omega_{i,j}$ 为 $\Omega(i)$ 中的第 j 个特征权重；f_{best} 和 f_{worst} 分别为 SHO 算法进化过程中发现的最佳适应度和最差适应度.

SHO 算法通过运动阶段、捕食阶段和恢复阶段进行寻优. 运动阶段主要是产生新的权重，从而提高权重的多样性；捕食阶段主要是淘汰适应度较差的权重，其目的是增强收敛性；恢复阶段主要是恢复权重规模以保证每次迭代规模相同.

1）运动阶段

根据猎物的结构，将猎物组的权重划分为 3 个角色——猎物的首领、追随者以及独立于猎物群移动的逃亡者，并对运动阶段进行分析. 首领 $\Omega(h_L)$ 代表当前迭代猎物组中生存值最佳的权重，当 $\Omega(h_i^k)\neq\Omega(h_L^k)$ 时，通过比较生存值与一个（0，1）的随机数得出当前迭代过程中的追随者和逃亡者，生存值大于等于随机数的猎物为追随者 $\Omega(H_f^k)$，其余的猎物为独立于猎物群移动的逃亡者 $\Omega(H_d^k)$，其中：$\Omega(h_i)$ 表示每一个猎物组的权重；k 表示当前的迭代次数.

基于以上描述，3 种角色的运动方式表示如下.

首领运动为

$$\Omega(\boldsymbol{h}_L^{k+1})=\begin{cases}\Omega(\boldsymbol{h}_L^k)+\boldsymbol{c}^k, & S_{\Omega(\boldsymbol{h}_L^k)}=1\\ \Omega(\boldsymbol{h}_L^k)+\boldsymbol{s}^k, & S_{\Omega(\boldsymbol{h}_L^k)}<1\end{cases} \tag{7}$$

式中：$\Omega(\boldsymbol{h}_L^{k+1})$ 为下一次迭代的最佳适应度对应的权重；$\boldsymbol{c}^k$ 为牧群首领决定驱赶捕食者时的运动矢量；$\boldsymbol{s}^k$ 为牧群首领决定转移到安全位置时的运动矢量.

追随者和逃亡者的运动为

$$\Omega(\boldsymbol{h}_i^{k+1})=\begin{cases}\Omega(\boldsymbol{h}_i^k)+\boldsymbol{f}_i^k, & \Omega(\boldsymbol{h}_i^k)\in\Omega(\boldsymbol{H}_f^k)\\ \Omega(\boldsymbol{h}_i^k)+\boldsymbol{d}_i^k, & \Omega(\boldsymbol{h}_i^k)\in\Omega(\boldsymbol{H}_d^k)\end{cases} \tag{8}$$

式中 $\Omega(\boldsymbol{H}_f^k)$ 和 $\Omega(\boldsymbol{H}_d^k)$ 分别为追随者和逃亡者的集合. 当 $\Omega(\boldsymbol{h}_i)$ 被标识为主要或从属追随者时，$\boldsymbol{f}_i^k$ 为追随者的运动矢量；当 $\Omega(\boldsymbol{h}_i)$ 被标识为逃亡者时，$\boldsymbol{d}_t^k$ 为逃亡者的运动矢量. 运动矢量 $\boldsymbol{c}^k$、$\boldsymbol{s}^k$ 以及运动矢量 $\boldsymbol{f}_i^k$ 和 $\boldsymbol{d}_t^k$ 的定义主要依赖于羊群的位置和生存值，其公式的详细推导、位置图和生存值参见文献[25].

捕食者的运动只依赖于与猎物个体的距离，并且捕食者通常攻击距离较近的个体，其运动公式可以表示为

$$\Omega(\boldsymbol{p}_I^{k+1})=\Omega(\boldsymbol{p}_I^k)+2\rho(\Omega(\boldsymbol{h}_r^k)-\Omega(\boldsymbol{p}_I^k)) \tag{9}$$

式中：$\Omega(\boldsymbol{p}_I^k)$ 为当前的捕食者组的权重，$I\in(1,N_p)$；ρ 为一个[0,1]的随机数；$\Omega(\boldsymbol{h}_r^k)\in\Omega(\boldsymbol{H}^k)$ 为采用轮盘赌的方式在猎物组中选出的权重；$r\in(1,2,\cdots,N_h)$，N_h 为猎物组权重的数量.

2）捕食阶段

在捕食阶段，将淘汰的权重定义为

$$\Omega(\boldsymbol{K})=\{\Omega(\boldsymbol{k}_n)=(\Omega(\boldsymbol{h}_w)\in\Omega(\boldsymbol{H}))\},\quad n=1,2,\cdots,N_k,w\in\{1,2,\cdots,N_h\} \tag{10}$$

式中 N_k 为在捕食阶段被淘汰的权重总数.

3）恢复阶段

为了保证每次迭代权重的规模相同，采用轮盘赌方式恢复权重规模，恢复权重定义为

$$\Omega(\boldsymbol{h}_w)=\Omega(\boldsymbol{h}_{\text{new}})=\text{mix}([\Omega(\boldsymbol{h}_{r_1,1}),\Omega(\boldsymbol{h}_{r_2,2}),\cdots,\Omega(\boldsymbol{h}_{r_s,s}),\cdots,\Omega(\boldsymbol{h}_{r_J,J})]) \tag{11}$$

式中 $\Omega(\boldsymbol{h}_{r_s,s})$ $(s=1,2,\cdots,J)$ 为随机候选权重 $\Omega(\boldsymbol{h}_{r_s})$ 的特征元素.

恢复阶段后,种群规模恢复,得到了新的迭代权重.

基于上述描述,SHO 算法通过运动阶段改变了每组权重的属性值,从而提高了特征权重的多样性;通过捕食阶段去除适应度较差的权重,选择适应度较好的权重进入下次迭代;通过恢复阶段保证了每次迭代权重组数相同. 通过 SHO 算法得到种群内部 RMSE 最小所对应的权重,并将上述权重和此权重的适应度分别记为 Ω_{best} 和 $f(\Omega_{\text{best}})$,同时存储在 Ω_{Best} 和 $f(\Omega_{\text{Best}})$ 中,即

$$\Omega_{\text{Best}}=\Omega_{\text{best}} \tag{12}$$

$$f(\Omega_{\text{Best}})=f(\Omega_{\text{best}}) \tag{13}$$

因为 SHO 算法在利用搜索空间上存在一些不足,影响了搜索性能,所以每次进入下次迭代的当前最佳权重并不准确,而 SA 算法局部搜索能力较强,因此,采用 SA 算法继续寻找当前最佳权重.

2.1.3 基于 SA 算法的权重再优化

将 Ω_{best} 作为 SA 算法的初始值,在权重 Ω_{best} 附近进行随机搜索,从而产生新的权重 Ω_{new},并计算其适应度 $f(\Omega_{\text{new}})$,再采用 Metropolis 准则[36-37] 接受新权重. 当新权重的适应度优于 Ω_{Best} 的适应度时直接接受,否则以一定概率 ζ 接受新权重,即 $\zeta<\xi$ 时接受,ζ 为[0,1]的随机数,ξ 为概率;否则不接受,然后在权重 Ω_{best} 附近重新进行搜索,并进行如上的判断. 如果接受新的权重,则将新解 Ω_{new} 和其适应度 $f(\Omega_{\text{new}})$ 赋予 Ω_{Best} 和 $f(\Omega_{\text{Best}})$;判断 $f(\Omega_{\text{Best}})$ 与 $f(\Omega_{\text{best}})$ 的大小,将值较小的权重赋予 Ω_{best};赋值结束后,判断当前温度否到最低,当 SA 算法中温度没有达到最低时,则在更新温度后在权重 Ω_{best} 附近继续进行随机搜索,当温度达到最低时,表明 SHO-SA 算法一次迭代寻优结束,选择本次迭代得到的最优权重进入下一次迭代寻优,直到达到迭代次数,从而获得最优权重的近似解. 更新温度等式、Metropolis 准则以及权重 Ω_{best} 的赋值公式表示为

$$T_{n+1}=\tau\cdot T_n \tag{14}$$

$$\xi=\begin{cases}1, & f(\Omega_{\text{new}})<f(\Omega_{\text{Best}})\\ \exp\left(-\dfrac{f(\Omega_{\text{new}})-f(\Omega_{\text{Best}})}{T_{\text{SA}}}\right), & f(\Omega_{\text{new}})\geqslant f(\Omega_{\text{Best}})\end{cases} \tag{15}$$

$$\Omega_{\text{best}}=\begin{cases}\Omega_{\text{best}}, & f(\Omega_{\text{best}})\leqslant f(\Omega_{\text{Best}})\\ \Omega_{\text{Best}}, & f(\Omega_{\text{best}})>f(\Omega_{\text{Best}})\end{cases} \tag{16}$$

式中:T_n 为第 n 次退火温度,并且设初始温度为 T_1;τ 表示温度衰减率;Ω_{Best} 为当前最佳适应度所对应的权重;T_{SA} 为温度控制参数.

2.2 算法伪代码

综上所述,SHO-SA 算法的伪代码如下.

输入:特征权重组数 N、特征数 J、特征权重上下限 ω_{u} 和 ω_{d}、适应度函数 f、总迭代次数 itern、当前迭代次数 k、初始温度 T_1、最终温度 $T_{\min}$、当前温度 T、温度控制参数 T_{SA} 和温度衰减系数 τ,以及特征权重的初始化.

输出:特征权重的近似最优解 Ω_{best} 及其适应度 $f(\Omega_{\text{best}})$.

1. 将 N 随机生成两部分:猎物群 H 和捕食者 P,数量分别为 N_{h} 和 N_{p};
2. 通过式(5)(6)分别计算每组权重的适应度和生存值;
3. for $k=1$: itern
4. 运动阶段:通过式(7)~(9)更新权重,并通过式(5)(6)重新计算适应度和生存值;
5. 捕食阶段:通过式(10)淘汰部分权重;
6. 恢复阶段:采用轮盘赌的方式恢复种群规模,恢复个体如式(11)所示,并得到当前迭代过程中最小适应度对应的权重 Ω_{best},通过式(12)(13)进行存储;

//SA 算法进一步寻优

7. $T=T_0$;
8. While $T>T_{\min}$
9. 设置 Ω_{best} 为 SA 算法的初始值,在 Ω_{best} 附近随机搜索产生新权重 Ω_{new},通过式(4)计算新权重的适应度 $f(\Omega_{\text{new}})$;
10. 通过式(12)(13)存储 Ω_{best} 和 $f(\Omega_{\text{best}})$;
11. if $f(\Omega_{\text{new}})<f(\Omega_{\text{Best}})$
12. $\Omega_{\text{Best}}=\Omega_{\text{new}}$, $f(\Omega_{\text{Best}})=f(\Omega_{\text{new}})$,并通过式(14)更新温度;
13. else
14. if $\zeta<\xi$
15. $\Omega_{\text{Best}}=\Omega_{\text{new}}$, $f(\Omega_{\text{Best}})=f(\Omega_{\text{new}})$,并通过式(14)更新温度;
16. end if
17. end if
18. 通过式(16)给 Ω_{best} 赋值;
19. end while
20. end for

3　实验研究

为了验证 SHO-SA 算法的有效性，采用加州大学欧文分校（University of California Irvine，UCI）数据集中的 5 个标准回归数据集进行实验，数据集分别为 Concrete Compressive Strength、Concrete Slump Test、QSAR aquatic toxicity、Airfoil Self-Noise 和 Energy Efficiency，为了方便查阅分别记为 D1 ~ D5，基本信息如表 2 所示. 将使用均权重的 CBR 算法记为 CBRMA，使用注水法分配权重的 CBR 算法记为 CBRWFA，使用遗传算法分配权重的 CBR 算法记为 CBRGA，使用 SHO 分配权重的 CBR 算法记为 CBRSHO，采用本文方法分配权重的 CBR 算法记为 CBRSHO-SA. 其中，由于变异概率会在一定程度上影响遗传算法分配权重的结果，因此，本文分别采用 0.03、0.06 和 0.09 共 3 种变异概率的遗传算法进行实验.

表 2　数据集信息

Table 2　Information of data sets

简称	数据集名称	案例数	特征数
D1	Concrete Compressive Strength	1 030	7
D2	Concrete Slump Test	103	7
D3	QSAR aquatic toxicity	546	8
D4	Airfoil Self-Noise	1 503	5
D5	Energy Efficiency	768	8

3.1　实验设计

本文方法可以通过 2 个实验进行验证：

1）针对表 2 中的 5 个数据集，采用上述 5 种权重分配方法的 CBR 预测模型进行实验，并比较本文方法与其他 4 种方法的预测结果.

2）通过计算 10 组 RMSE 的标准差（standard deviation，SD）衡量预测实验的稳定性，并与 CBRMA、CBRWFA、CBRGA 以及 CBRSHO 等方法的实验结果做比较.

评价指标除了 RMSE 外，还使用了平均绝对误差（mean absolute error，MAE）和 SD 作为评价实验结果的指标. 为了保证实验的可靠性和全面性，本文采用五折交叉验证的方法，记录每次实验的预测值、RMSE，并计算 5 次实验的平均值和 RMSE 的标准差.

参数设置为：5 种方法的近邻个数 M 全部取 3，初始权重规模全部取 100，迭代次数全部取 50；遗传算法采用 4 位二进制编码的方式，交叉概率 p_c 取 0.6，变异概率 p_m 分别取 0.03、0.06 和 0.09；SHO 算法中每个特征值的下限 ω_d 取 0，每个特征值的上限 ω_u 取 1；在 SHO-SA 算法中，初始温度 T_1 取 100 ℃，最低温度 T_{min} 取 0.001 ℃，温度衰减率 τ 取 0.95，温度控制参数 T_{SA} 取 T_n，其余参数与 SHO 算法和 SA 算法中的一致.

3.2　实验步骤

针对上述的 2 个实验内容，实验步骤设计如下.

步骤 1　参数初始化.

步骤 2　构建案例库. 将数据集进行归一化处理，每个特征值和预测值都映射到[0，1]，将特征值和相应的预测值表示成式（1）的形式，并储存在案例库中.

步骤 3　构造训练样本和测试样本.

步骤 4　权重分配. 分别采用均权重法、注水法、遗传算法、SHO 算法和 SHO-SA 算法分配权重.

步骤 5　案例检索. 通过式（3）计算相似度，将其从大到小进行排序并取出前 M 个案例.

步骤 6　案例重用. 计算前 M 个案例输出值的平均值作为建议的预测值.

步骤 7　计算 RMSE 和 MAE. 重复步骤 5、6，得到所有测试数据的预测值，并计算 RMSE 和 MAE.

步骤 8　案例修正. 对案例重用给出的数据进行实际验证或专家评估，若相差较大则需要对案例进行修正.

步骤 9　案例存储. 将目标案例和预测值形成一个新的案例并存储于案例库中.

步骤 10　重复步骤 1 ~ 9，记录 2 次五折实验的结果，并计算实验结果的平均值以及 10 组 RMSE 的 SD.

3.3　比较分析

3.3.1　有效性分析

为了测试本文方法分配特征权重的有效性，本文分别采用均权重法、注水法、遗传算法、SHO 算法和 SHO-SA 算法分别实现表 2 中 5 个数据集的特征权重分配，并绘制 5 个回归数据集的拟合效果（见图 4 ~ 8），由于遗传算法在变异概率 $p_m = 0.06$ 时实验效果最好，因此，绘图时选择遗传算法（$p_m = 0.06$）进行绘制，从中可以直观看出本文方法的拟合效果较好. 另外，在 RMSE 和 MAE 方面，将各个方法的 RMSE 和 MAE 的平均值进行对比，结果如表 3 所示. 对于表 2 中的 5 个数据集，基于 5 种权重分配方法的 CBR 预测精度由高到低依次是：

CBRSHO-SA、CBRSHO、CBRGA、CBRMA、CBRWFA. 可见,具有不同操作算子和独立个体行为[25]的 SHO 算法能在一定程度上缓解局部最优问题,并且 SHO-SA 算法通过退火策略进行全局搜索得到最优权重,进一步改善了陷入局部最优的缺点,表明了采用 SA 算法优化 SHO 算法的有效性.

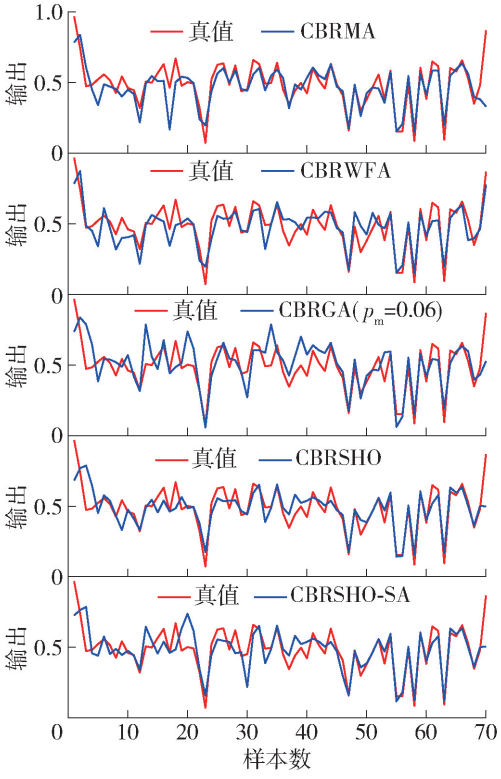

图 4 各方法对数据集 D1 的拟合效果

Fig. 4 Fitting effect of D1 for each method

3.3.2 稳定性分析

为了验证本文方法的稳定性,在每次实验后记录其 RMSE. 因为 MA 和 WFA 每次分配权重固定,所以同一数据集每次所得 RMSE 相同,使得 MA 和 WFA 方法所得 RMSE 的 SD 为 0,并且遗传算法在 $p_m=0.06$ 时实验效果最优,因此,对每个数据集和遗传算法($p_m=0.06$)、SHO 算法以及 SHO-SA 算法进行对比,均计算 10 组 RMSE 的 SD 并绘制柱状图,如图 9 所示. 由图可以看出,本文方法的 SD 在数据集 D1 ~ D5 中均低于其他方法,结果表明此方法的稳定性较好. 另外,为了减小误差,本文计算了各方法针对 5 个回归数据集 SD 的平均值,如表 4 所示. 表中的 DZ:SD 表示第 Z 个数据集的SD,采用遗传算法($p_m=0.06$)、SHO 算法、SHO-SA 算法分配特征权重得出 CBR 预测模型的平均 SD 分别为 0.007 7、0.006 3、0.004 1. 由结果分析可知,

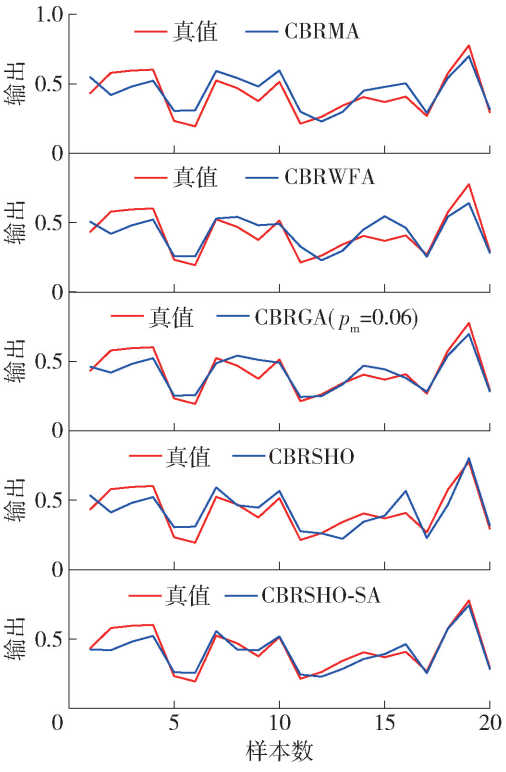

图 5 各方法对数据集 D2 的拟合效果

Fig. 5 Fitting effect of D2 for each method

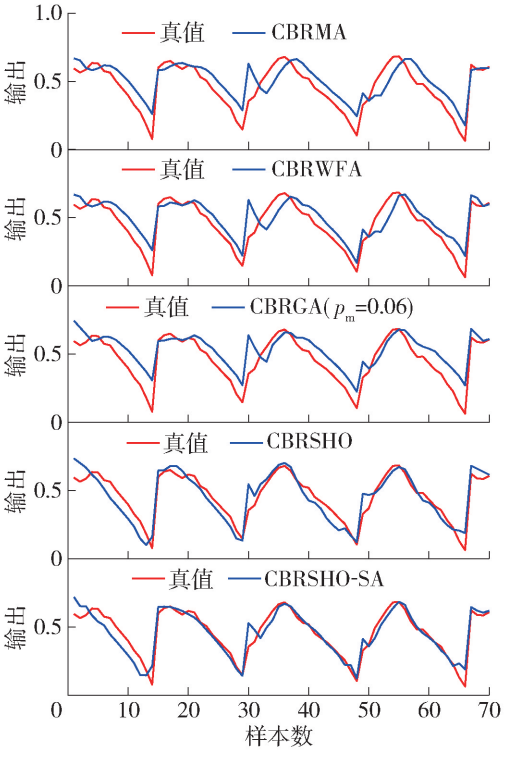

图 6 各方法对数据集 D3 的拟合效果

Fig. 6 Fitting effect of D3 for each method

SHO-SA 算法通过运动阶段、捕食阶段、恢复阶段以及退火策略不断对权重迭代寻优,使得 CBR 预测模

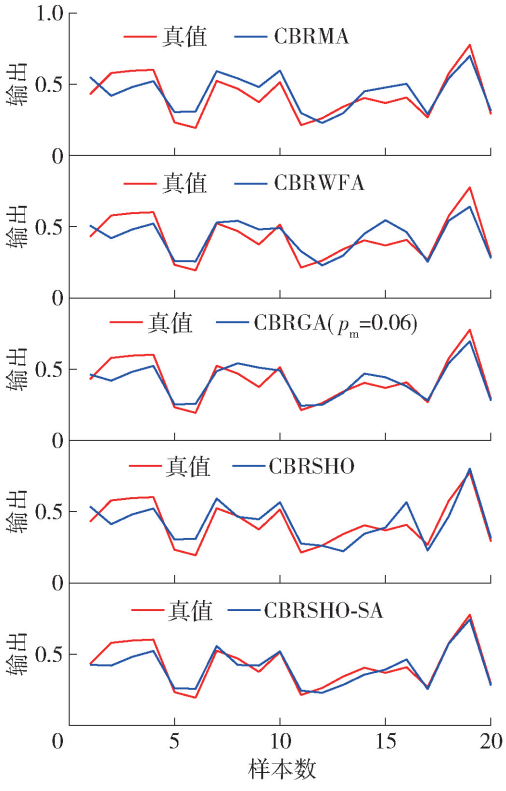

图 7　各方法对数据集 D4 的拟合效果

Fig. 7　Fitting effect of D4 for each method

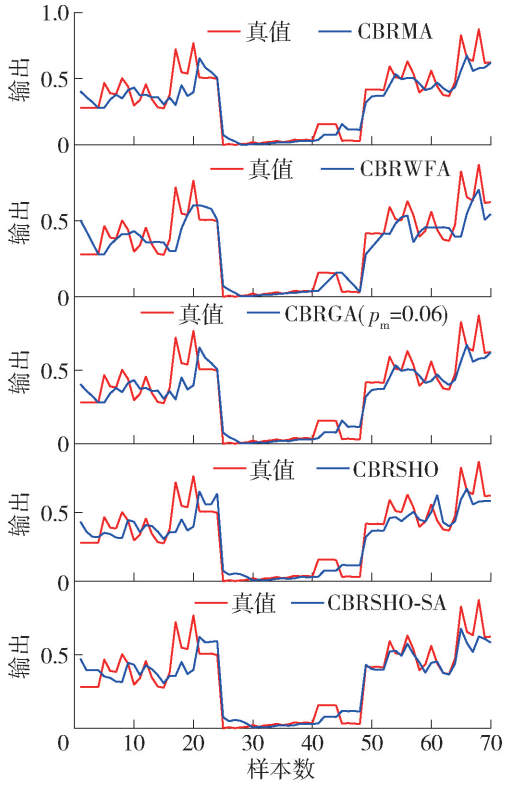

图 8　各方法对数据集 D5 的拟合效果

Fig. 8　Fitting effect of D5 for each method

表 3　不同方法对各数据集 RMSE 和 MAE 的对比

Table 3　Comparison of RMSE and MAE of different methods for each data set

评价指标	方法	D1	D2	D3	D4	D5	平均值
RMSE	CBRMA	0.122 8	0.097 5	0.123 1	0.102 6	0.087 3	0.106 7
	CBRWFA	0.131 6	0.099 6	0.127 1	0.139 0	0.008 0	0.101 1
	CBRGA(p_m=0.03)	0.127 1	0.086 9	0.129 8	0.095 2	0.068 7	0.101 5
	CBRGA(p_m=0.06)	0.125 1	0.083 3	0.126 3	0.094 1	0.064 4	0.098 6
	CBRGA(p_m=0.09)	0.127 3	0.087 8	0.128 2	0.095 2	0.067 2	0.101 1
	CBRSHO	0.113 6	0.080 6	0.125 7	0.093 3	0.058 5	0.094 3
	CBRSHO-SA	0.101 5	0.074 4	0.120 5	0.089 3	0.056 6	0.088 5
MAE	CBRMA	0.094 0	0.078 3	0.087 6	0.088 8	0.057 7	0.081 3
	CBRWFA	0.101 4	0.078 8	0.093 1	0.113 9	0.052 8	0.088 0
	CBRGA(p_m=0.03)	0.096 3	0.066 1	0.092 0	0.079 7	0.051 3	0.077 1
	CBRGA(p_m=0.06)	0.094 5	0.065 8	0.089 9	0.077 5	0.046 5	0.074 8
	CBRGA(p_m=0.09)	0.097 9	0.068 9	0.092 2	0.078 2	0.050 9	0.077 6
	CBRSHO	0.086 9	0.063 8	0.089 9	0.074 4	0.037 2	0.070 4
	CBRSHO-SA	0.074 7	0.056 2	0.087 3	0.070 5	0.036 2	0.065 0

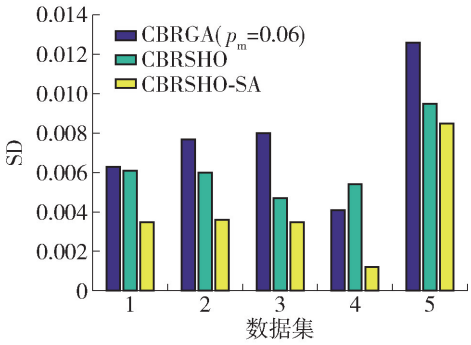

图 9 各方法的 SD 对比

Fig. 9 Comparison of SD for different methods

表 4 不同方法对各数据集 SD 的对比

Table 4 Comparison of SD of each data set for different methods

数据集	CBRGA (p_m = 0.06)	CBRSHO	CBRSHO-SA
D1:SD	0.006 3	0.006 1	0.003 5
D2:SD	0.007 7	0.006 0	0.003 6
D3:SD	0.008 0	0.004 7	0.003 5
D4:SD	0.004 1	0.005 4	0.001 2
D5:SD	0.012 6	0.009 5	0.008 5
平均值	0.007 7	0.006 3	0.004 1

型的结果更具稳定性.

综上所述,采用本文基于 SHO-SA 算法进行特征权重分配后,CBR 预测模型的精度较高,并具有良好的稳定性.

4 结论

1) 针对特征权重优化分配存在的问题,本文提出一种基于 SHO-SA 算法的特征权重优化分配方法. 首先,采用适应度函数和 SHO 算法的运动阶段、捕食阶段和恢复阶段优化特征权重,得到单次迭代 RMSE 最小所对应的权重;然后,采用 SA 算法进一步对上述权重进行随机搜索,进而对特征权重进行更合理的优化分配,并存储最优特征权重作为 SHO 算法下一次迭代的初始权重;最后,通过不断迭代得到近似最优解.

2) 对比实验的结果表明,本文分配权重的方法在 CBR 预测模型中发挥了一定优势,相对于其他方法来说,本文方法的准确性和稳定性均得以提高.

3) 值得注意的是,虽然实验结果表明了该特征权重优化分配方法的有效性,但是 SA 算法采用随机抽样的方法产生新解,导致算法收敛速度慢,并且在当前最优权重附近产生新解方面还缺乏成熟的理论依据,这样得到的结果可能是权重的近似最优解. 因此,为了实现更加科学、准确且稳定的特征权重优化分配方法,这个问题将是下一阶段的主要研究内容.

参考文献:

[1] YAN A J, SHAO H S, GUO Z. Weight optimization for case-based reasoning using membrane computing [J]. Information Sciences, 2014, 287(12): 109-120.

[2] AAMODT A, PLAZA E. Case-based reasoning: foundational issues, methodological variations, and system approaches[J]. AI Communications, 1994, 7(1): 39-59.

[3] YU H, YAN A J, WANG D H. Case-based reasoning classifier based on learning pseudo metric retrieval [J]. Expert Systems with Applications, 2017, 89(12): 91-98.

[4] SARAIVA R, PERKUSICH M, SILVA L, et al. Early diagnosis of gastrointestinal cancer by using case-based and rule-based reasoning [J]. Expert Systems with Applications, 2016, 61(11): 192-202.

[5] THIKE P H, XU Z, CHENG Y, et al. Materials failure analysis utilizing rule-case based hybrid reasoning method [J]. Engineering Failure Analysis, 2019, 95(1): 300-311.

[6] LU Y, LI Q M, XIAO W J. Case-based reasoning for automated safety risk analysis on subway operation: case representation and retrieval [J]. Safety Science, 2013, 57(8): 75-81.

[7] 严爱军, 戴香东, 邵宏赡, 等. 基于自组织膜计算的案例推理属性权重优化方法[J]. 北京工业大学学报, 2017, 43(5): 745-753.
YAN A J, DAI X D, SHAO H S, et al. Self-organization membrane computing-based attribute weights optimization for case-based reasoning model [J]. Journal of Beijing University of Technology, 2017, 43(5): 745-753. (in Chinese)

[8] GUO Y, WU K. Research on case retrieval of Bayesian network under big data [J]. Data & Knowledge Engineering, 2018, 118(11): 1-13.

[9] 张春晓, 严爱军, 王普. 案例推理分类器属性权重的内省学习调整方法[J]. 计算机应用, 2014, 34(8): 2273-2278.
ZHANG C X, YAN A J, WANG P. Introspective learning adjustment approach for attribute weights of case-based reasoning classifier[J]. Journal of Computer Applications,

2014，34(8)：2273-2278.（in Chinese）

[10] FU C，XU D L，XUE M. Determining attribute weights for multiple attribute decision analysis with discriminating power in belief distributions [J]. Knowledge-Based Systems，2018，143(3)：127-141.

[11] CHEN B，LI X H，LIU H W，et al. Hybrid subjective and objective evaluation method of the equipment for first class distribution network[J]. Energy Procedia，2019，158(2)：3452-3457.

[12] LEUNG L C，CAO D. On the efficacy of modeling multi-attribute decision problems using AHP and Sinarchy[J]. European Journal of Operational Research，2001，132(1)：39-49.

[13] YU R C，TZENG G H. A soft computing method for multi-criteria decision making with dependence and feedback [J]. Applied Mathematics and Computation，2006，180(1)：63-75.

[14] RAJASEKHAR M，SUDARSANA G，SREENIVASULU Y，et al. Delineation of groundwater potential zones in semi-arid region of Jilledubanderu river basin，Anantapur District，Andhra Pradesh，India using fuzzy logic，AHP and integrated fuzzy-AHP approaches [J]. HydroResearch，2019，2：97-108.

[15] FARNA S. Bimodal fuzzy analytic hierarchy process (BFAHP) for coronary heart disease risk assessment[J]. Journal of Biomedical Informatics，2018，83(7)：204-216.

[16] WANG M，NIU X D. Research on project post-evaluation of wind power based on improved ANP and fuzzy comprehensive evaluation model of trapezoid subordinate function improved by interval number [J]. Renewable Energy，2019(3)：255-265.

[17] SINGH D，SINGH B，KAUR M. Simultaneous feature weighting and parameter determination of neural networks using ant lion optimization for the classification of breast cancer[J]. Biocybernetics and Biomedical Engineering，2020，40(1)：337-351.

[18] 欧阳森，石怡理. 改进熵权法及其在电能质量评估中的应用[J]. 电力系统自动化，2013，37(21)：156-159，164.
OUYANG S，SHI Y L. A new improved entropy method and its application in power quality evaluation [J]. Automation of Electric Power Systems，2013，37(21)：156-159，164.（in Chinese）

[19] 柳炳祥，李海林. 基于模糊粗糙集的因素权重分配方法[J]. 控制与决策，2007，22(12)：1437-1440.
LIU B X，LI H L. Method of factor weights allocation based on combination of fuzzy and rough set[J]. Control and Decision，2007，22(12)：1437-1440.（in Chinese）

[20] 赵辉，严爱军，王普. 提高案例推理分类器的可靠性研究[J]. 自动化学报，2014，40(9)：2019-2036.
ZHAO H，YAN A J，WANG P. On improving reliability of case-based reasoning classifier[J]. Acta Automatica Sinica，2014，40(9)：2019-2036.（in Chinese）

[21] 严爱军，蒋伟，王普. 案例属性权重的群基数效用优化方法[J]. 北京工业大学学报，2012，38(12)：1888-1892.
YAN A J，JIANG W，WANG P. Optimization method for case attributes weights based on group cardinal utility method[J]. Journal of Beijing University of Technology，2012，38(12)：1888-1892.（in Chinese）

[22] 高鹰，谢胜利. 基于模拟退火的粒子群优化算法[J]. 计算机工程与应用，2004，41(1)：47-50.
GAO Y，XIE S L. Particle swarm optimization algorithm based on simulated annealing[J]. Computer Engineering and Applications，2004，41(1)：47-50.（in Chinese）

[23] LIU W Z，YE J H. Collapse optimization for domes under earthquake using a genetic simulated annealing algorithm [J]. Journal of Constructional Steel Research，2014，97(6)：59-68.

[24] LI C，QIN J，LI J，et al. The accident early warning system for iron and steel enterprises based on combination weighting and grey prediction model GM (1，1) [J]. Safety Science，2016，89(11)：19-27.

[25] FAUSTO F，CUEVAS E，VALDIVIA A，et al. A global optimization algorithm inspired in the behavior of selfish herds[J]. Bio Systems，2017，160(10)：39-55.

[26] YIMIT A，IIGURA K，HAGIHARA Y. Refined selfish herd optimizer for global optimization problems [J]. Expert Systems with Applications，2020，139(6)：1-15.

[27] GUO Y，CHEN W，ZHU Y X，et al. Research on the integrated system of case-based reasoning and Bayesian network[J]. ISA Transactions，2019，90(7)：213-225.

[28] 严爱军，赵辉，王普. 基于可信度阈值优化的案例推理评价分类方法[J]. 控制与决策，2016，31(7)：1253-1257.
YAN A J，ZHAO H，WANG P. Trustworthiness evaluation method with threshold optimization for case-based reasoning classification[J]. Control and Decision，2016，31(7)：1253-1257.（in Chinese）

[29] 严爱军，钱丽敏，王普. 案例推理属性权重的分配模型比较研究[J]. 自动化学报，2014，40(9)：1896-1902.
YAN A J，QIAN L M，WANG P. A comparative study of attribute weights assignment for case-based reasoning[J]. Acta Automatica Sinica，2014，40(9)：1896-1902.（in

Chinese)

[30] LIU W J, DU Y J. A novel focused crawler based on cell-like membrane computing optimization algorithm[J]. Neurocomputing, 2014, 123(1): 266-280.

[31] YAN A J, WANG W X, ZHANG C X, et al. A fault prediction method that uses improved case-based reasoning to continuously predict the status of a shaft furnace[J]. Information Sciences, 2014, 259(2): 269-281.

[32] RAY S S, MISRA S. Genetic algorithm for assigning weights to gene expressions using functional annotations [J]. Computers in Biology and Medicine, 2019, 104(1): 149-162.

[33] 李峰, 冯珊. 基于人工神经网络的案例检索与维护[J]. 系统工程与电子技术, 2004, 26(8): 1053-1056.
LI F, FENG S. ANN based approach to case retrieval and case maintenance [J]. Systems Engineering and Electronics, 2004, 26(8): 1053-1056. (in Chinese)

[34] ANAND P, ARORA S. A novel chaotic selfish herd optimizer for global optimization and feature selection[J]. Artificial Intelligence Review, 2020, 53(2): 1441-1486.

[35] BENDAOUD R, AMIRY H, BENHMIDA M, et al. New method for extracting physical parameters of PV generators combining an implemented genetic algorithm and the simulated annealing algorithm[J]. Solar Energy, 2019, 194(12): 239-247.

[36] 何庆, 吴乐意, 徐同伟. 改进遗传模拟退火算法在TSP优化中的应用[J]. 控制与决策, 2018, 33(2): 219-225.
HE Q, WU L Y, XU T W. Application of improve genetic simulated annealing algorithm in TSP optimization [J]. Coutrol and Decision, 2018, 33(2): 219-225. (in Chinese)

[37] 傅文渊, 凌朝东. 布朗运动模拟退火算法[J]. 计算机学报, 2014, 37(6): 1301-1308.
FU W Y, LING C D. Brownian motion based simulated annealing algorithm[J]. Chinese Journal of Computers, 2014, 37(6): 1301-1308. (in Chinese)

[38] ZHANG L Z, MA H, QIAN W, et al. Protein structure optimization using improved simulated annealing algorithm on a three-dimensional AB off-lattice model [J]. Computational Biology and Chemistry, 2020, 85(4): 1-8.

(责任编辑 梁 洁)