

神经网络二次集成预测算法研究

顾海艳¹, 曹 林², 朱 涛¹, 袁 明¹

(1. 江苏警官学院, 江苏 南京 210031)

(2. 中国人民解放军第 66072 部队, 北京 100144)

[摘要] 针对常用预测算法不同程度地存在泛化能力不足的缺陷, 提出了基于神经网络二次集成的优化算法 (NNE2-QQ). 该算法在第一次集成时采用量子粒子群算法进行个体网络的选择优化, 在第二次集成时采用量子免疫算法进行集成结论生成优化, 并通过多次迭代自适应寻求个体和权值的最佳组合, 实现神经网络二次集成模型的性能最优, 最后实验验证了 NNE2-QQ 算法的有效性和实用性. NNE2-QQ 可从海量数据中发现各种因素之间的联系及其规律, 为预测判断提供支持.

[关键词] 神经网络二次集成, 量子粒子群算法, 量子免疫算法, 数据预测

[中图分类号] TP18 **[文献标志码]** A **[文章编号]** 1001-4616(2022)02-0136-06

Research on Quadratic Integration Prediction Algorithm of Neural Network

Gu Haiyan¹, Cao Lin², Zhu Tao¹, Yuan Ming¹

(1. Jiangsu Police Institute, Nanjing 210031, China)

(2. PLA 66072 Troops, Beijing 100144, China)

Abstract: The common prediction algorithms of data mining have the defect of insufficient generalization ability, so this paper proposes an optimization algorithm based on two-level neural network ensemble (NNE2-QQ). The first ensemble uses quantum particle swarm optimization algorithm to select individuals, the second ensemble uses quantum immune algorithm to optimize the result, and then the best ability of the ensemble model is obtained by finding the best selection and weights through multiple iterations. Finally, simulation results demonstrate the practicality and efficiency of NNE2-QQ algorithm. NNE2-QQ can find the relationship between various factors and their laws from massive data, and provide support for prediction and judgment.

Key words: two-level neural network ensemble, quantum particle swarm optimization algorithm, quantum immune algorithm, data forecasting

大数据时代分分秒秒都在产生大量信息, 这些信息具有类型多样、价值密度低等明显特征. 从海量数据信息中挖掘数据间具有的潜在相关关系, 构建预测模型、做出预测分析是数据处理的重要内容, 已在各行各业得到普遍重视和应用. 随着当前计算能力的不断提升, 在各种建模方法中, 具有自学习功能的神经网络模型受到了越来越多的关注. 此类模型能较好地拟合复杂非线性函数, 十分适合用于复杂过程模拟, 但存在鲁棒性差和泛化能力弱等缺陷^[1].

近年来, 各种优化智能算法的相继出现, 使得越来越多的研究将群智能算法与神经网络相结合, 通过神经网络的集成, 提高神经网络模型的泛化性能、提升预测精度和适用性已经较为普遍^[2]. 李晓峰等^[3]提出基于卷积神经网络与特征选择的医疗图像误差预测算法, 预测结果拟合度和预测耗时均优于传统算法; 张兴辉等^[4]提出了一种基于反向改进的灰狼算法, 来优化预测模型以提高交通流预测的精度与鲁棒性; Choi 等^[5]则提出 Gabor DCNN 的集成方法来提升人脸识别性能; 王洁等^[6]提出采用自动摘要-卷积神经网络的集成模型进行文档情感分析.

随着神经网络集成在识别^[7-8]、诊断^[9-10]、预测^[11-12]、评判^[13]等很多领域的广泛应用, 对集成性能的

收稿日期: 2021-11-24.

基金项目: 国家自然科学基金资助项目(61802155)、江苏省教改课题(2019JSJC595).

通讯作者: 顾海艳, 副教授, 研究方向: 信息安全、大数据技术和智能优化. E-mail: ghy7388@126.com

要求也越来越高. 提高集成性能主要有两个改进方向^[14], 一是运用选择方法构建高格群体, 即个体网络选择优化; 二是改进集成结论生成方法, 即集成结论生成优化. 个体网络选择优化能够提高集成模型的预测精度, 缺点是模型的泛化性能下降; 集成结论生成优化能够提升模型的泛化性能, 但其预测精度有待提高. 通过建立多层结构, 可整合上述两种方法, 有效且相对容易地实现神经网络集成的改进^[15]. 实际问题中一般建立两层结构, 即神经网络二次集成 (Two-level Neural Network Ensemble, 简称 NNE2). 因此, 二次集成的算法选择和二次集成后对泛化性能的影响是本文需要研究的问题.

1 基于神经网络二次集成的优化算法

1.1 神经网络二次集成的基本结构

神经网络二次集成的基本框架如图 1 所示. 第一次集成将输入数据随机分组构成 J 个数据集 $D_j (j=1, 2, \dots, J)$, 采用不同网络模型或对网络参数设置不同的初始值, 来独立训练优选出 J 组、每组 K 个神经网络 $NN = \{NN_{jk} \} (j=1, 2, \dots, J, k=1, 2, \dots, K)$, 于是第一次集成的个体网络总数 $T=K \times J$. 然后每组个体网络中随机选取 $K_j (j=1, 2, \dots, J)$ 个作为第一次集成的个体, 记为 SN . 第二次集成采用集成结论优化方法, 将第一次集成得到的 J 个神经网络 $NNE_j (j=1, 2, \dots, J)$, 集成为一个网络模型, 记为 $NNE2$.

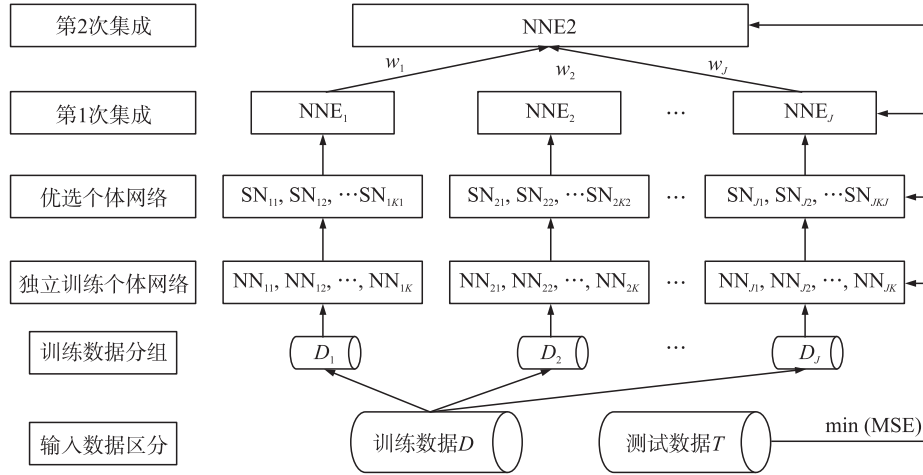


图 1 神经网络二次集成框架图

Fig. 1 The framework diagram of neural network secondary integration

神经网络二次集成的难点在于寻找恰当的组合 λ 和权值 w , 合理调控每次集成的优化程度, 从而提升二次集成的泛化能力. 第一次集成优化的目标是选择合适的个体, 第二次集成优化的目标是为每个个体分配合理的权重, 神经网络二次集成的终极目标是平衡这两次集成优化, 实现模型泛化误差最小化.

1.2 神经网络二次集成的目标函数

泛化性能是基于神经网络的数据预测模型的重要性能指标. 由于集成优化的泛化误差可以用集成模型在测试集的标准差表示, 因此依据泛化误差最小化原则, 神经网络集成的目标函数可表示为:

$$\min f(\lambda, w) = \sum_{i=1}^M \left[y_i - \sum_{j=1}^J w_j NNE_j(x_i) \right]^2 / M, \quad (1)$$

式中, λ 是 1 个 J 位的二进制数, 表示第一次集成时个体组合; $w = [w_1, w_2, \dots, w_J]$ 为第二次集成的个体 NNE_j 的权值组合; M 为测试集 $T(x_i, y_i)$ 的样本数量; J 为第一次集成得到的集成模型数量; $NNE_j(x_i)$ 可由各组选中的神经网络个体加权平均计算, 算式如下:

$$NNE_j(x_i) = \sum_{k=1}^{K_j} SN_{jk}(x_i) / K_j, \quad (2)$$

式中, $SN_{jk}(x_i)$ 为集成时选择的个体网络 SN_{jk} 的输出; K_j 为第一次集成时第 j 组个体网络入选集成的数量. 令向量 $KJ = \{K_1, K_2, \dots, K_J\}$, 公式(1)中的 λ 与向量 KJ 存在一一对应关系, KJ 相当于 λ 的十进制表示.

1.3 神经网络二次集成的优化算法选择

神经网络集成方法除了常用的线性组合方法, 还有贪心算法、遗传算法、粒子群算法等优化算法, 学者

们在这些算法的基础上又研究提出了量子粒子群算法和量子免疫算法^[15-18].

量子粒子群算法(quantum particle swarm optimization, QPSO)是一种基于群体智能的优化计算技术^[19],该算法将粒子群算法与量子计算相结合,对粒子采用量子比特进行编码,为并行计算提高搜索效率提供可能;另外,该算法还采用量子旋转门更新种群,大大丰富了种群的多样性,为提升算法的全局寻优能力提供了可能^[16]. 由于个体网络选择优化是离散问题优化求解,该算法具有较好的适应性. QPSO 算法的核心思想:在粒子更新时根据当前粒子与最优粒子的相位差,自适应调整旋转角,提升算法的寻优效率;并在变异操作时加大旋转角,改善算法的全局寻优性能. 其基本流程如图 2 所示.

量子免疫算法(quantum immune algorithm, QIA)是另一种群体智能优化计算技术,在免疫算法中融入量子遗传算法思想,该算法因具有较优的收敛性能而受到广泛关注. 由于集成结论优化是组合权值优化问题,组合权值优化是连续问题的优化求解,量子免疫算法具有较好的适应性^[15]. QIA 算法的基本思想是在量子免疫算法中引入精英策略、反转策略,利用精英策略保留优秀抗体,提高寻优效率;利用反转策略增加多样性,避免早熟问题,增强全局发现能力. 该算法的基本流程如图 3 所示.

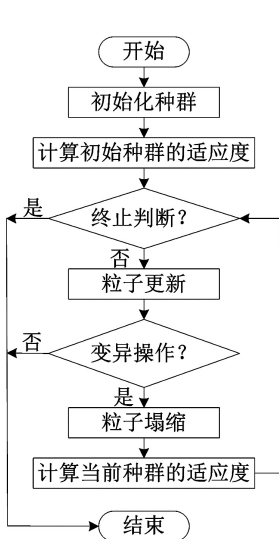


图 2 QPSO 算法的基本流程

Fig. 2 The basic process of QPSO algorithm

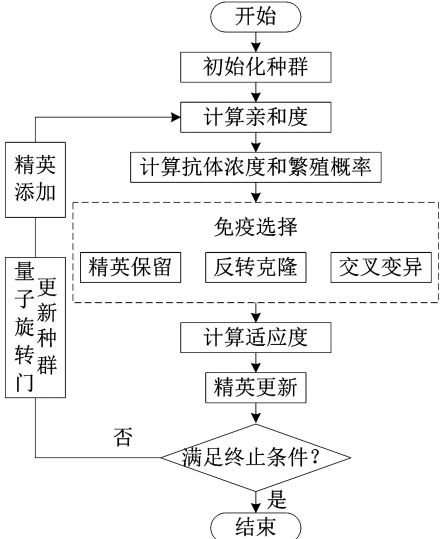


图 3 量子免疫算法的基本流程

Fig. 3 The basic process of QIA algorithm

鉴于上述算法的特性,本文提出基于量子粒子群算法和量子免疫算法的神经网络二次集成(Two-level neural network ensemble based on QPSO and QIA, NNE2-QQ),在第一次集成时采用量子粒子群算法实现个体网络选择优化,在第二次集成时采用量子免疫算法进行集成结论生成优化,然后通过分步优化,使预测模型具有较好泛化性能.

2 NNE2-QQ 算法流程

2.1 NNE2-QQ 算法的基本流程

NNE2-QQ 算法的基本流程如图 4 所示,具体步骤如下:

Step1: 数据处理. 将输入数据按照一定比例,随机划分成训练集 D 和测试集 T ,再将训练集 D 随机分成 J 组.

Step2: 第一次集成. 首先依据 J 组训练数据独立训练优化出 J 组每组 K 个个体网络;然后采用量子粒子群算法优化个体网络选择,得到最优个体组合 λ .

Step3: 第二次集成. 以第一次集成得到的 J 个集成模型为个体,采用量子免疫算法优化集成结论,得到最优权值组合 w .

Step4: 迭代优化. 当达到最大迭代次数时,停止迭代、输出结果,否则返回 Step2.

2.2 NNE2-QQ 的自适应迭代优化

由于存在两组需要优化的变量 λ 和 w ,可采用自适应迭代寻优的方法,如图 5 所示. 首先将第二次集

成的权值组合设为 $w_0 = [1/J, 1/J, \dots, 1/J]$, 即第二次集成结论生成时初始化采用简单平均方法, 然后用量子粒子群算法搜索第一次集成的最优组合 λ_1 , 以 λ_1 确定的组合群体为基础, 再用量子免疫算法优化组合权值, 得到 w_1 , 如此循环, 直到满足终止条件, 最终得到第一次集成的最优个体组合 λ^* 和第二次集成的最优权值组合 w^* .

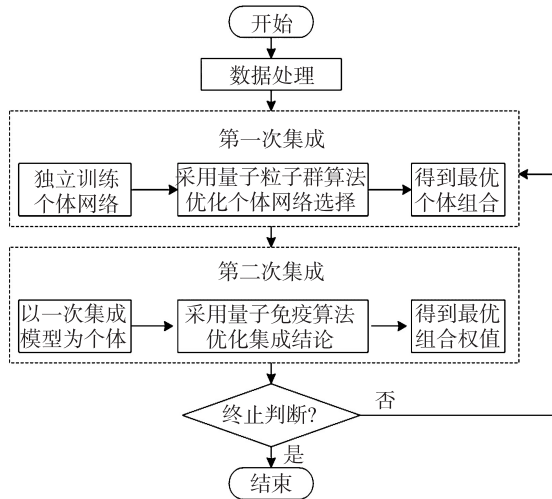


图4 NNE2-QQ 算法的基本流程

Fig. 4 The basic process of NNE2-QQ algorithm

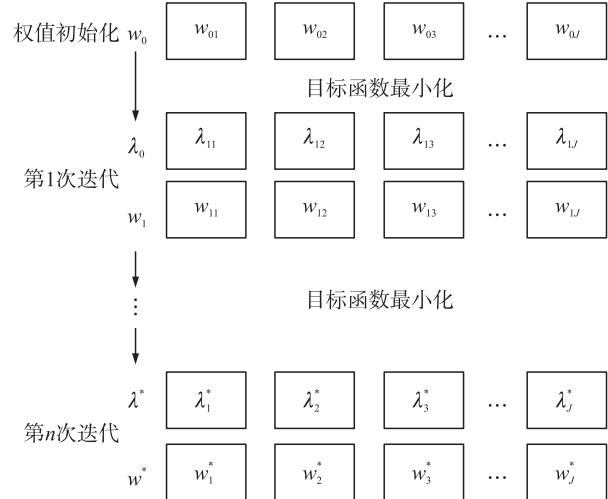


图5 迭代优化示意图

Fig. 5 Iterative optimization diagram

3 实验验证

3.1 仿真实验

参考文献[14]的做法, 使用2个函数 Friedman#1 和 Friedman#3, 且第一次集成时独立训练优选出5组、每组20个个体神经网络, 研究 NNE2-QQ 算法的适用性和有效性, 并与简单集成方法(BEM)、基于量子粒子群的个体选择优化方法(QPSO)、基于量子免疫算法的集成结论优化方法(QIA)相比较。

3.1.1 使用 Friedman#1 函数预测

Friedman#1 函数:

$$y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon$$

式中, $x_i (i=1, 2, \dots, 5)$ 在区间 $[0, 1]$ 均匀分布, 令 $\varepsilon=0$, 即不考虑噪声的影响。

首先利用 Friedman#1 函数生成1200个数据, 之后随机选择其中1000个组成训练集, 剩余200个组成测试集, 用于估计集成的泛化误差。10次仿真实验的平均误差见表1, 其中平均误差是10次仿真实验泛化误差的平均值, 泛化误差用测试集上的均方误差表示。

仿真实验结果表明, 相比 BEM 而言, QPSO、QIA 和 NNE2-QQ 得到的集成模型泛化能力均有显著提高, 且 NNE2-QQ 的泛化性能明显最优。

3.1.2 使用 Friedman#3 函数预测

Friedman#3 函数:

$$y = \arctan[x_3 - 1/x_1 x_2 x_4] + \varepsilon.$$

其中约束条件为:

$$0 \leq x_1 \leq 100, 20 \leq x_2/2\pi \leq 280, 0 \leq x_3 \leq 1, 1 \leq x_4 \leq 11.$$

令 $\varepsilon=0$, 即不考虑噪声的影响, 产生2200个数据, 随机选取2000个组成训练集, 其余200个组成测试集。实验重复10次, 取泛化误差的平均值, 结果如表2所示。由此表可知, 与简单集成 BEM 相

表1 优化算法的比较之一

Table 1 The first comparison of optimization algorithms				
方法	BEM	QIA	QPSO	NNE2-QQ
平均误差	0.118	0.023	0.016	0.001

表2 优化算法的比较之二 (10^{-3})

Table 2 The second comparison of optimization algorithms (10^{-3})				
方法	BEM	IQIA	IQPSO	NNE2-QQ
平均误差	5.124	4.211	4.127	3.325

比,QPSO 和 QIA 的泛化误差明显降低,而融合了 QPSO 和 QIA 两种优化算法的 NNE2-QQ 则泛化误差更低,泛化性能显著提升.

3.2 实例检验

为进一步验证 NNE2-QQ 的数据预测性能,本文以 Boston-housing 实例数据^[14]进行验证分析. Boston-housing 是 UCI 数据库中的一个实际问题,包含 506 个数据样本,输入数据是 13 维的预测房价的基本指标,输出数据是 1 维的该地区的房价,目的是根据基本指标来预测房价.

首先,随机选择 400 个样本组成训练集,并将训练集随机选择均分为 5 组、每组 80 个样本,剩余 106 个样本组成测试集. 之后,针对每组数据,采用 BP(Back Propagation)神经网络模型并通过设置不同参数,独立训练优选出 20 个个体神经网络,再通过分步优化使得二次集成的泛化误差最小. 由公式(1)和(2)可得目标函数如下:

min f(λ,w)=∑_{i=1}¹⁰⁶[y_i−∑_{j=1}^Jw_jNNE_j(x_i)]²/106,

其中:

NNE_j(x_i)=∑_{k=1}^{K_j}SN_{jk}(x_i)/K_j.

分别采用 BEM、QPSO、QIA 和 NNE2-QQ 等 4 种算法对目标函数进行优化,仿真实验 10 次结果如图 6、表 3 所示.

表 3 不同算法优化结果的比较

算法	最优结果	最差结果	平均时间/s
BEM	8.53	9.26	0.02
QPSO	4.20	6.45	5.32
QIA	3.24	4.60	6.12
NNE2-QQ	1.07	2.70	9.45

图 6 结果表明,NNE2-QQ 的泛化误差最低且比较稳定,BEM 虽然较为稳定但泛化误差最高,QPSO、QIA 的泛化误差较低但稳定性较差,这表明 NNE2-QQ 算法具有非常好的泛化性能. 表 3 结果表明,NNE2-QQ 的最优泛化误差值仅为 1.07,最差泛化误差值也只有 2.70. 虽然该算法泛化性能非常优异,但其运行时间较长,运算速度比其他 3 种算法逊色很多,如图 7 所示. 这也表明该算法适用于对准确性要求高、但对运算速度要求不高的数据预测应用场景.

由上述仿真实验验证和实例对比两方面分析可见,NNE2-QQ 算法的泛化性能明显优于简单集成(BEM)、量子粒子群算法(QPSO)、量子免疫算法(QIA),提高了预测的准确性. 因此,NNE2-QQ 算法能够较好地平衡两次集成的优化程度,使得集成模型的泛化误差更小,从而得到更优的预测性能.

4 结论

本文针对数据预测问题,提出了一种神经网络二次集成的优化算法 NNE2-QQ. 通过仿真实验验证和实例对比分析,发现相比于简单集成方法、基于量子粒子群算法、基于量子免疫算法,NNE2-QQ 算法具有准确度高、泛化性能好的显著优势,但其运行速度较慢,适用于对速度要求不高但对准确度有较高要求的数据预测场景. 后续将就如何提高该算法运行速度开展进一步研究,从而为安全预警、运营决策和预测评估等相关工作提供更好的技术支持.

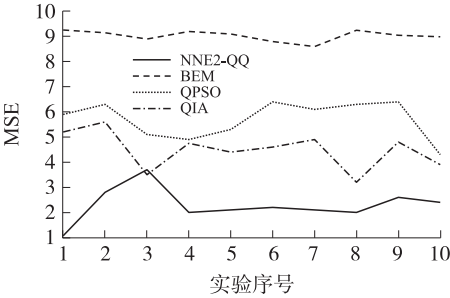


图 6 不同算法预测结果的泛化误差对比
Fig. 6 Comparison of the generalization error of different algorithms

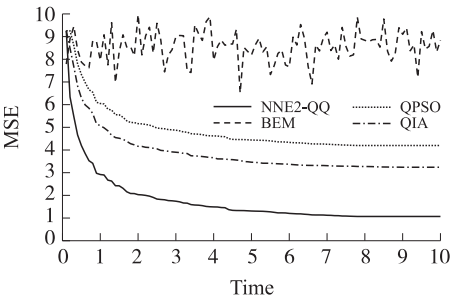


图 7 不同算法的收敛性对比
Fig. 7 Comparison of the astringency of different algorithms

[参考文献]

- [1] HANSEN L K, PETER S. Neural network ensembles[J]. Institute of Electrical and electronic Engineers transactions on pattern analysis & machine intelligence, 1990, 12(10): 993-1001.
- [2] LI H, WANG X S, DING S F. Research and development of neural network ensembles: a survey[J]. Artificial intelligence review, 2018, 49(4): 455-479.
- [3] 李晓峰, 刘刚, 卫晋, 等. 基于卷积神经网络与特征选择的医疗图像误差预测算法[J]. 湖南大学学报(自然科学版), 2021, 48(4): 90-99.
- [4] 张兴挥, 樊秀梅, 阿喜达, 等. 反向学习的灰狼算法优化及其在交通流预测中的应用[J]. 电子学报, 2021, 49(5): 879-886.
- [5] CHOI J Y, LEE B. Ensemble of deep convolution neural networks with Gabor face representations for face recognition[J]. Institute of Electrical and Electronic Engineers transactions on image processing, 2020, 29: 3270-3281.
- [6] 王洁, 乔艺璇, 彭岩, 等. 基于深度学习的美国媒体"一带一路"舆情的情感分析[J]. 计算机技术与应用, 2018, 44(11): 102-110.
- [7] LEE M, LEE J, CHANG J H. Ensemble of jointly trained deep neural network-based acoustic models for reverberant speech recognition[J]. Digital signal processing, 2019, 85: 1-9.
- [8] 韩兆宇, 周勇, 刘兵. 基于差异性神经网络集成的命名实体识别方法[J]. 计算机工程与设计, 2019, 40(4): 994-1000.
- [9] KITAMURA G, CHUNG C Y, MOORE B E. Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, De Novo Training, and Multiview Incorporation[J]. Journal of digital imaging, 2019, 32: 672-677.
- [10] 朱俊, 刘天羽, 王致杰, 等. 基于蜂群算法的选择性神经网络集成的风机齿轮箱轴承故障诊断[J]. 电机与控制应用, 2017, 44(1): 6-11.
- [11] 谢琪, 程耕国, 徐旭. 基于神经网络集成学习股票预测模型的研究[J]. 计算机工程与应用, 2019, 55(8): 238-243.
- [12] 赵辉, 周杰, 王红君, 等. 基于 CEEMDAN-PE 和 QGA-BP 的短期风速预测[J]. 电子技术应用, 2018, 44(12): 60-64.
- [13] 闫瑞姣, 尹四清. 选择性神经网络集成的微博用户信用评估模型[J]. 计算机工程与设计, 2018, 39(5): 1478-1483.
- [14] 施彦. 群体智能预测与优化[M]. 北京: 国防工业出版社, 2012.
- [15] CHAKRABORTY M, BISWAS S K, PURKAYASTHA B. A novel ensembling method to boost performance of neural net-works[J]. Journal of experimental & theoretical artificial intelligence, 2020, 32(1): 17-29.
- [16] 曹林, 王之腾, 陈亮, 等. 基于改进量子免疫算法的神经网络集成[J]. 计算机工程与应用, 2020, 56(22): 142-147.
- [17] KURNAZ T F, KAYA Y. A novel ensemble model based on GMDH-type neural network for the prediction of CPT-based soil liquefaction[J]. Environmental earth sciences, 2019, 78(11): 339.1-339.14.
- [18] RRAGAGNOLO L, SILVA R V D, GRZYBOWSKI J M V. Artificial neural network ensembles applied to the mapping of landslide susceptibility[J]. Catena, 2019, 184: 104240.
- [19] 金建海, 孙俊, 张安通, 等. 基于量子粒子群优化算法的无人艇航线规划[J]. 船舶力学, 2020, 24(3): 352-361.

[责任编辑:顾晓天]