基于 GA-GWO 混合优化算法的 ANN-KNN 融合模型的股票价格区间预测

孙怡曈，宫雨竹，徐祥博，李甘露，莫慧杰，杨凯 *
长春工业大学数学与统计学院，吉林　长春 130012
DOI:10.61369/ASDS.12182

摘　　要： 本文提出一种基于混合优化算法的多模型融合预测框架，即为一种基于人工神经网络（ANN）和 K 近邻（KNN）融合的预测方法。该方法在训练期间，先运用遗传算法（GA）进行若干代的优化，初步定位潜在的最优区域，再运用灰狼优化算法（GWO）仔细搜索最优结果，进而得到最优的模型参数。同时构建两个双序列模型，通过仿真实验对比融合模型与单一模型的预测效果。结果显示，与单模型的预测效果相比，融合模型在捕捉股票价格趋势和波动性方面表现更为卓越。

关 键 词： 区间值时间序列；神经网络；KNN；融合模型；混合优化算法

Stock Price Interval Prediction Using a GA-GWO Hybrid Optimization Algorithm-Based ANN-KNN Fusion Model

Sun Yitong, Gong Yuzhu, Xu Xiangbo, Li Ganlu, Mo Huijie, Yang Kai*
School of Mathematics and Statistics, Changchun University of Technology, Changchun, Jilin 130012

Abstract: This paper proposes a hybrid optimization algorithm–based multi–model fusion prediction framework, which integrates an Artificial Neural Network (ANN) and K–Nearest Neighbors (KNN). During the training phase, the Genetic Algorithm (GA) is first employed to optimize potential solution spaces over multiple generations, followed by the Grey Wolf Optimizer (GWO) to refine the search for optimal parameters. Additionally, dual–sequence models are constructed to compare the prediction performance between the fusion model and individual models through simulation experiments. The results demonstrate that the fusion model outperforms single models in capturing stock price trends and volatility.

Keywords: interval–valued time series; neural networks; KNN; fusion model; hybrid optimization algorith

引言

在众多领域中，区间值数据广泛存在，它能够表示数据的不确定性、可变性等特征。在金融市场，股票收益 [1] 等时间序列随着时间推移呈现出复杂的波动性变化。吴玉霞等人利用 ARIMA 模型对创业板市场股票价格变动的规律和趋势进行了预测 [2]，未能充分考虑区间值时间序列的非线性特征和波动性，难以准确捕捉市场波动。杨星等人利用非对称非线性平滑转换的广义自回归条件异方差算法的碳价格均值自回归检验 [3]，虽能刻画波动聚集性，但其依赖方差假设

为结合不同算法的优势，融合模型逐渐成为研究热点 [5]。田瑞杰等（2019）尝试将 BP 与 ANN 形成融合模型 [6]，丰富了融合模型的领域。温泉等人提出 TSO-XGBoost 并行融合模型 [7]，但未针对区间值数据的双序列特性进行适配。本文将以改进上述模型的问题为目的，针对现有研究的数据表示单一化、模型协同不足和动态适应性弱等的局限性，提出一种基于双序列分解的 ANN-KNN 的融合模型，通过分阶段协同预测与残差优化，为提升区间值时间序列预测的准确性和效率提供坚实的理论基础与技术支持。

项目信息：国家自然科学基金项目 (12471249) 长春工业大学大学生创新创业训练计划项目 (202410190088)。
作者简介：孙怡曈（2004-），女，吉林省长春市人，长春工业大学学生，研究方向：时间序列分析；
通讯作者：杨凯（1984-），男，辽宁省沈阳市，副教授，博士生导师，研究方向：时间序列分析 .Email: yangkai@ccut.edu.cn

一、模型构建

在传统数据分析中，现实存在的一些信息无法表示，这促使了区间数据分析方法（SDA）的发展。区间数据指的是变量取值范围在某个区间内。为了解析这类数据，我们通过区间的中点和半径来进行研究。即当区间值数据按照时间顺序进行收集时，我们得到了区间值时间序列数据[8]在相同的时间里。（t=1,2,⋯,n 其中 n为实际序列观察到的区间个数）。因此我们得到 X_{U_t}和 X_{L_t}并 $X_{L_t} \leqslant X_{U_t}$且，此时 X_{U_t}和 X_{L_t}分别为时刻的区间上界和下界，则原时间序列数据即可表示为：

$$\left[X_{L_1}, X_{U_1}\right], \left[X_{L_2}, X_{U_2}\right], \cdots, \left[X_{L_n}, X_{U_n}\right] \tag{1}$$

（一）中值 – 对数半径序列 (M–LR)

为进行后续的模型预测，本段将聚焦构建两个时间序列：区间中值序列 X_t^M 与对数半径序列 X_t^{lR}。结合文中 (1) 所给出的区间序列形式，我们能够分别运用以下方式来构建这两个序列：

$$X_t^M = \frac{X_{U_t} + X_{L_t}}{2} \text{和} X_t^{lR} = \log_2\left(\frac{X_{U_t} - X_{L_t}}{2}\right) \tag{2}$$

通过模型得到这两个序列拟合的值并将其用于预测区间的上界和下界的值，因此对区间的上界和下界的预测分别为：$\hat{U}_t = \hat{X}_t^M + 2^{\hat{X}_t^{lR}}$ 和 $\hat{L}_t = \hat{X}_t^M - 2^{\hat{X}_t^{lR}}$。其中 $\hat{X}_t^M$ 和 $\hat{X}_t^{lR}$ 分别表示中值和对数半径值序列的模型预测结果。

我们定义预测半径为 $\hat{R} = 2^{\hat{x}_t^{lR}}$，将区间中值序列的残差和区间半径序列的残差分别表示为：$\varepsilon_{X_t} = X_t^M - \hat{X}_t^M$ 和 $\varepsilon_{X_t} = X_t^R - \hat{X}_t^R$。其中，$\hat{x}_t^M$ 是区间中值序列通过模型的预测结果，$\hat{X}_t^R$ 是将模型得到的结果去对数化后得到的预测结果。

基于上述内容，区间时间序列的最终预测表示如下：

$$\hat{X}_t^H = \hat{U}_t + \hat{N}_{U_t} \text{和} \hat{X}_t^L = \hat{L}_t + \hat{N}_{L_t} \tag{3}$$

其中，$\hat{N}_{U_t}$ 是其他模型对区间上界的预测误差，$\hat{N}_{L_t}$ 是其他模型对区间下界的预测误差。这些误差由以下式子得出：$\hat{N}_{U_t} = \hat{\varepsilon}_{X^M} + \hat{\varepsilon}_{X^R}$ 和 $\hat{N}_{L_t} = \hat{\varepsilon}_{X^M} - \hat{\varepsilon}_{X^R}$，这里的 $\hat{\varepsilon}_{x^M}$ 是其他模型对时间时区间中值序列误差的预测结果，$\hat{\varepsilon}_{x^R}$ 是其他模型对时间时区间半径序列误差的预测结果。

综上，利用区间中值序列 $\hat{x}_t^M$ 和区间对数半径序列 $\hat{x}_t^{lR}$ 构建的双序列模型，通过对区间上界和下界进行预测，最终得到的预测区间为 $\left[\hat{X}_t^L, \hat{X}_t^H\right]$。

（二）中值 – 半径序列 (M–R)

在本部分中，考虑两个时间序列：区间中值序列 X_t^M 和区间半径序列 X_t^R。考虑到给出的 (1) 的区间序列形式，我们可以分别用：

$$X_t^M = \frac{X_{U_t} + X_{L_t}}{2} \text{和} X_t^R = \frac{X_{U_t} - X_{L_t}}{2} \tag{4}$$

在进行模型训练和数据预测中，我们运用区间中值序列 X_t^M 和区间半径序列 X_t^R。因此对区间的上界和下界的预测值分别为：$\hat{U}_t = \hat{X}_t^M + \hat{X}_t^R$ 和 $\hat{L}_t = \hat{X}_t^M - \hat{X}_t^R$ 其中 $\hat{X}_t^M$ 和 $\hat{X}_t^R$ 分别表示中值和区间半径序列的模型预测结果。

根据上述预测结果，我们定义中值残差序列即为 $\varepsilon_{X^M} = X_t^M - \hat{X}_t^M$，半径残差序列为：$\varepsilon_{X^R} = X_t^R - \hat{X}_t^R$。其中 $\hat{X}_t^M$ 和 $\hat{X}_t^R$ 分别表示中值和区间半径序列的模型预测结果。因此，区间时间序列的最终预测为：

$$\hat{X}_t^H = \hat{U}_t + \hat{N}_{U_t} \text{和} \hat{X}_t^L = \hat{L}_t + \hat{N}_{L_t} \tag{5}$$

其中 $\hat{N}_{U_t}$ 和 $\hat{N}_{L_t}$ 分别为对区间上界和下界的预测误差。通过 $\hat{N}_{U_t} = \hat{\varepsilon}_{X_t^M} + \hat{\varepsilon}_{X_t^R}$ 和 $\hat{N}_{L_t} = \hat{\varepsilon}_{X_t^M} - \hat{\varepsilon}_{X_t^R}$ 即可得到区间上界误差 $\hat{N}_{U_t}$ 和区间下界误差 $\hat{N}_{L_t}$。其中 $\hat{\varepsilon}_{X_t^M}$ 和 $\hat{\varepsilon}_{X_t^R}$ 分别为用模型对 t 时间时的区间中值序列的误差和区间半径序列的误差预测结果。

综上运用区间中值序列 $\hat{x}_t^M$ 和区间半径序列 $\hat{x}_t^R$ 的双序列模型得到对于区间上界和下界的预测，即可得到最终的预测区间为 $\left[\hat{X}_t^L, \hat{X}_t^H\right]$。

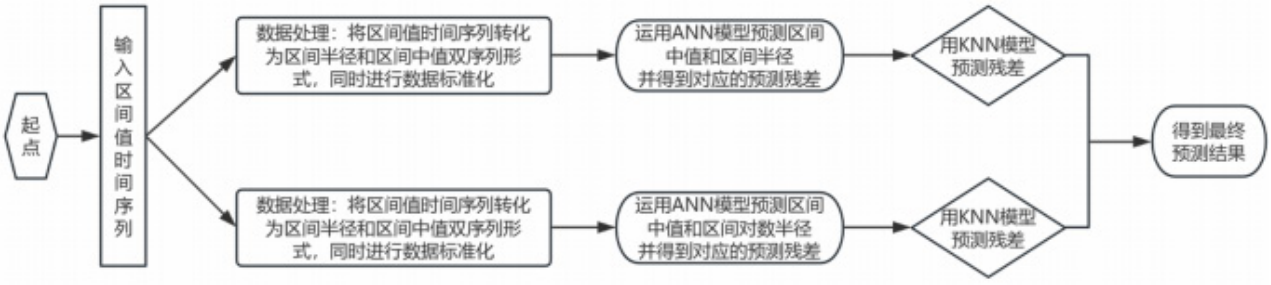
（三）ANN–KNN 融合模型

单一模型通常难以充分提取时间序列数据中的信息，为实现更高精度的预测，提出了融合模型[9]的预测方法。图 1.1 展示了融合模型结构，其核心思想是通过 ANN 模型进行区间中值、区间半径和区间对数半径的预测，再运用 KNN 模型[10]对上述区间的残差进行预测。将预测结果结合得到最终的预测结果。

为实现更高效地提取有用信息，可针对时间序列的不同特征开展建模工作。我们可以合理的认为区间值时间序列 $X_t = \left[X_t^L, X_t^H\right]^T (t = 1, 2, \cdots, n)$ 是由两部分组成。即

$$X_t = L_t + N_t \tag{6}$$

其中 X_t 表示原始时间序列数据，L_t 表示线性分量，N_t 表示非线性分量。首先，假设残差仅包含非线性关系，利用 ANN 对线性分量进行建模并预测残差；再通过 KNN 模型对非线性分量进行建模和预测。最终将线性和非线性的结果模型进行组合，得到整体预测结果。

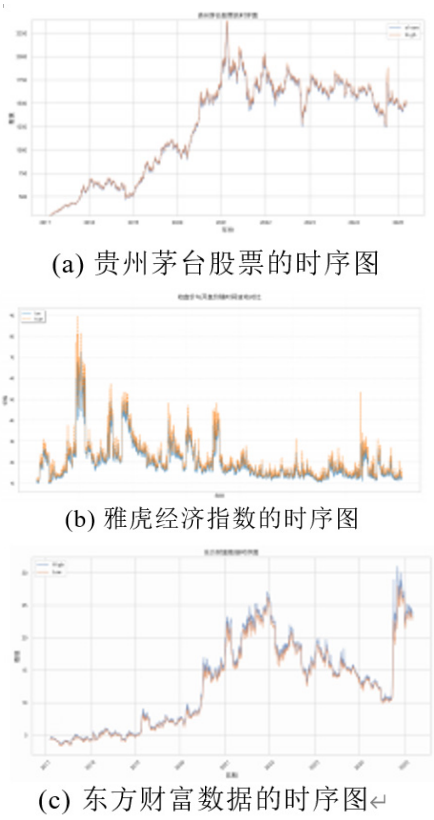


> 图 1.1: 融合模型流程图

二、实证分析

（一）数据获取

本次研究选用了贵州茅台股票数据为分析对象，时间为2017年2月7日至2025年3月7日共1964观测值。而后又选择从2007年1月3日到2016年12月30日的雅虎经济指数和2017年2月17到2025年3月7日的东方财富数据等的波动性数据。其中雅虎经济指数数据共包含2518条数据，东方财富数据共包含1964条数据。



(a) 贵州茅台股票的时序图

(b) 雅虎经济指数的时序图

(c) 东方财富数据的时序图

> 图2.1: 三种不同数据的时序图

图2.1(a)、图2.1(b)、图2.1(c)分别展示了三个不同数据情况的波动性数据集的走势情况。图2.1(a)为贵州茅台股票数据，从2017年-2021年，数据总体呈上升趋势，且2022年-2025年间整体在一定区间内上下震荡。图2.1(b)为雅虎经济指数的时序图，自2007年开始雅虎经济指数随着时间一路上涨，在2009年达到峰值。图2.1(c)为东方财富指数的数据，从2017年-2021年数据呈现上升趋势，随后在2022年-2024年经历了较大幅度的波动下降随后达到又一高峰。

（二）单模型拟合

为了衡量不同模型的预测效果，本研究通过区间平均相对方差（ARV）、区间覆盖率(PICP)和区间平均宽度(PINAW)这三种经典的精度评估指标。具体来说，这些精度的定义如下：

$$ARV=\frac{\sum_{t=1}^{n}\left(X_{t+1}^{H}-\hat{X}_{t+1}^{H}\right)^{2}+\sum_{t=1}^{n}\left(X_{t+1}^{L}-\hat{X}_{t+1}^{L}\right)^{2}}{\sum_{t=1}^{n}\left(X_{t+1}^{H}-\overline{X^{H}}\right)^{2}+\sum_{t=1}^{n}\left(X_{t+1}^{L}-\overline{X^{L}}\right)^{2}} \tag{7}$$

$$PICP=\frac{\min\left\{X_{t}^{H},\hat{X}_{t}^{H}\right\}-\max\left\{X_{t}^{L},\hat{X}_{t}^{L}\right\}}{X_{t}^{H}-X_{t}^{L}} \tag{8}$$

$$PINAW=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{X}_{t}^{H}-\hat{X}_{t}^{L}\right) \tag{9}$$

其中n为样本数量，X_{t+1}^{H}、X_{t+1}^{L} X_{t}^{H} X_{t}^{L}为第$t+1$个和第t个真实区间，$\hat{X}_{t+1}^{H}$、$\hat{X}_{t+1}^{L}$ $\hat{X}_{t}^{H}$ $\hat{X}_{t}^{L}$为第$t+1$个和第t个拟合区间，$\overline{X^{H}}$和$\overline{X^{L}}$分别为真实区间的最高值和最低值的平均值。

ARV值表示数据点相对于区间平均值的离散程度。ARV值越小，说明模型精度较高。PICP值指的是真实值落在预测区间内部的比例，当PICP处于合理范围时，PINAW越窄，代表预测区间越精确。当多个模型的PINAW相近时，PICP越高，说明模型精度更高。

1.模型预测

为了评估模型的泛化能力，我们以80%划为数据集的训练集，20%划为数据集的测试集。想要得到最优模型，我们在所有单模型中引入遗传算法[11]和灰狼优化算法[12]形成混合优化模型GA-GWO进行每一个单模型的最优模型参数的选择。其基本思想是：利用遗传算法全局搜索能力强和灰狼优化算法局部效率高的特点形成优于两种单一优化模型的优化算法。即先用GA进行全局搜索，生成多样化的初始种群，再将GA的结果作为GWO的初始种群，利用GWO的包围机制进行精细搜索，进而得到更优的模型参数。

2.单模型拟合结果

表1展示了单模型的预测情况分析，分别采用单一模型拟合三种不同数据的中值-半径序列和中值-对数半径序列的双序列预测效果。我们可以看到ANN模型表现最佳，在PINAW值相差不大的情况下具有最高的PICP值和最低的ARV值，表明该模型在拟合和预测区间值时间序列上非常有效。

表1 单模型预测结果

模型	雅虎经济指数			贵州茅台股票			东方财经经济指数		
	PICP	PINAW	ARV	PICP	PINAW	ARV	PICP	PINAW	ARV
SVRM-R	0.8811	1.7993	0.0329	0.8468	32.9006	0.0075	0.5468	0.5551	0.0272
KNNM-R	0.917	1.6558	0.0096	0.8966	36.108	0.0095	0.4465	0.5851	0.0106
ANNM-R	0.9999	1.8171	0.0000	0.9304	33.3034	0.0071	0.9973	0.5888	0.0000
SVRM-LR	0.9932	1.9185	0.0400	0.9805	30.6800	0.0002	0.7683	0.6028	0.0056
KNNM-LR	0.4177	1.7400	0.0480	0.9027	36.4862	0.0096	0.4126	0.5748	0.0105
ANNM-LR	0.7923	1.8000	0.027	0.4054	22.2836	0.413	0.8034	0.6333	0.0029

图2.2(a)、图2.2(b)、图2.2(c)展示了单模型分别对三种不同数据的两个双序列模型的的预测 曲线。从图中可以观察到，在三个不同类型的数据集中，ANN模型拟合效果最好，说明ANN模型可以有效区间值时间序列的趋势和波动。KNN在中值－半径序列上表现较好，但整体波动较大，未能充分学习到区间值时间序列的特征。SVR的预测值整体上和实际值较为接近，但预测精度不如ANN模型。

(a) 贵州茅台股票单模型预测曲线

(b) 雅虎经济指数单模型预测曲线

(c) 东方财富数据单模型预测曲线

> 图2.2：单模型预测曲线

（三）融合模型拟合

通过分析以上单模型的拟合效果，发现ANN模型在区间值时间序列的预测上表现最佳。针对模型在某些预测点上存在较大相对误差的问题，在此考虑运用两个模型[6]分别对线性分量L_t和非线性分量N_t进行预测，并将两部分的预测结果以加法的形式组合，得到交易价格的最终预测结果。

表2分别是采用融合模型[4]拟合三种不同数据的中值－半径序列和中值－对数半径序列的双序列预测效果。将融合模型的预测效果与单模型的预测效果进行对比，发现融合模型在区间值时间序列预测中表现更好。其中提升效果比较明显的是东方财经指数中的ANN-KNN和KNN-SVR融合模型，这说明融合模型能从不同模型的优点中获益，并减少了单一模型的缺点。

表2 融合模型预测结果

模型	雅虎经济指数			贵州茅台股票			东方财经经济指数		
	PICP	PINAW	ARV	PICP	PINAW	ARV	PICP	PINAW	ARV
KNN-SVRM-R	0.8394	1.7539	0.0125	0.7245	32.7467	0.0073	0.8046	0.5478	0.0048
ANN-KNNM-R	1.0000	1.8171	0.0000	0.9918	29.9399	0.0001	0.9919	0.5894	0.0003
ANN-SVRM-R	0.9999	1.8174	0.0000	0.9727	30.5175	0.0039	0.9991	0.587	0.0000
KNN-SVRM-LR	0.8964	1.7395	0.0122	0.9064	34.3805	0.0056	0.6913	0.6214	0.0056
ANN-KNNM-LR	0.9787	1.8613	0.0436	0.9856	40.0000	0.4106	0.6781	0.5788	0.0046
ANN-SVRM-LR	0.9724	1.8885	0.0186	0.9608	38.4330	0.4150	0.7012	0.5275	0.0033

在雅虎经济指数中，ANN-KNN模型的预测区间相对实际值的比例较为一致，模型精度较高。在贵州茅台股票上，ANN-KNN模型在三个融合模型中表现最佳，PICP值达到最高。虽然PINAW 也相对较高但处于合理范围，同时ARV最低。在东方财经经济指数，PICP值达到了0.99以上，PINAW也相对较高但处于，且ARV最低近乎为0。综合四个指标来看，发现ANN-KNN效果最好。

图2.3(a)、图2.3(b)、图2.3(c)分别绘制了融合模型在三种不同数据的两个双序列模型的的预 测曲线。综合来看各融合模型预测结果与实际值高度吻合。其中ANN-KNN在中值－半径序列上表现优异，能够更准确地捕捉到数据的非线性特征和时序关系。

(a)贵州茅台股票融合模型预测曲线

(b) 雅虎经济指数融合模型预测曲线

(c)东方财富数据融合模型预测曲线

> 图2.3：融合模型预测曲线

三、研究结论

本文利用 ANN、KNN、SVR 单一模型，同时在单一模型的基础上选取两个模型形成融合模型，同时分别建立中值 – 半径序列和中值 – 对数半径序列的双序列结构，运用单模型和融合模型分别对上述两个双序列进行数据预测。为了衡量模型的预测效果，选择 PICP、PINAW、ARV 作为评估指标，主要有以下发现：

（一）从单一模型的预测效果来看，ANN 模型在三种不同数据情况中表现最佳，其中 ANN 模型能够有效捕捉价格变化的趋势和波动，可能在个别点上存在较大的相对误差。

（二）相较于单一模型，融合模型显著提高了预测效果，表现出更高的 PICP 和更好的 PINAW、ARV 值，其中 ANN 模型和 KNN 模型在区间值时间序列的预测中整体表现最佳。

（三）针对中值 – 半径序列和中值 – 对数半径序列的双序列预测结果，我们可以看到对于不同的数据情况和不同的模型方法，中值 – 半径序列更适合进行本次预测分析，模型效果更优。

本研究提出的多模型融合预测框架，为区间时间序列 [13] 预测提供了创新思路，期望能为金融市场参与者在风险管理 [14]、投资决策等方面提供有力的支持，并为后续相关研究提供有益的参考与借鉴，推动金融时间序列分析领域的持续进步。

参考文献

[1] 王浩 . 中国证券市场股票价格预测模型综述 [J]. 四川教育学院学报 ,2009,25(07):58-60.

[2] 吴玉霞 , 温欣 . 基于 ARIMA 模型的短期股票价格预测 [J]. 统计与决策 ,2016,(23):83-86.

[3] 杨星 , 李斌 , 曾悦 , 等 . 非对称非线性平滑转换的广义自回归条件异方差算法的碳价格均值回归检验 [J]. 控制理论与应用 ,2019,36(04):622-628.

[4] 李昭毅 , 孙虎元 , 蔡振宇 , 等 . 基于 Sine-SSA-BP 人工神经网络的腐蚀速率预测研究 [J]. 海洋科学 ,2024,48(08):17-28.

[5] 赵山 , 仝新朵 , 王晓波 , 等 . 多指标定量、化学模式识别结合加权 TOPSIS 与灰色关联度融合模型的木芙蓉叶质量评价 [J]. 中国中医药信息杂志 ,2025,32(03):129-135.

[6] 田瑞杰 , 张维石 , 翟华伟 . 基于时间序列与 BP-ANN 的短时交通流速度预测模型研究 [J]. 计算机应用研究 ,2019,36(11):3262-3265

[7] 温泉 , 余玉欢 , 庄尚德 , 等 . 融合 SHAP 和 TSO-XGBoost 模型的水路货运量预测 [J]. 水利水运工程学报 ,2024,(06):86-96.

[8] 陶志富 , 刘金培 , 朱家明 , 等 . 区间值时间序列预测效果测度研究 [J]. 模糊系统与数学 ,2018,32(04):135-144.

[9] 韩自奋 , 陈宁 , 范义 , 等 . 基于多种机器学习算法的风电功率预测融合模型 [J]. 干旱气象 ,2024,42(05):710-718.

[10] 吴胜义 , 王义贵 , 王飞 , 等 . 基于多距离度量 kNN 模型的森林蓄积量反演 [J]. 中南林业科技大学学报 ,2023,43(02):10-18.

[11] 刘哲 , 李凤军 . 遗传算法与蚂蚁算法的融合在神经网络中的研究 [J]. 科技广场 ,2012,(10):6-9.

[12] 杨红光 , 刘建生 . 一种结合灰狼优化和 K- 均值的混合聚类算法 [J]. 江西理工大学学报 ,2015,36(05):85-89.

[13] 万昆 , 柳瑞禹 . 区间时间序列向量自回归模型在短期电力负荷预测中的应用 [J]. 电网技术 ,2012,36(11):77-81.

[14] 章凡 . 论股票投资风险控制及应对策略 [J]. 商展经济 ,2023,(13):84-87.

