基于 GA-XGBoost 的宁波港物流需求预测

李 顺[1]，李 君[1]，吴 鑫[1]，梅碧舟[2]
（1.浙江万里学院，浙江 宁波 315100；2.浙江易锻精密机械有限公司，浙江 象山 315700）

摘 要：通过灰色关联法，分析宁波港物流需求同腹地经济的相关度，建立宁波港物流需求预测指标集，提出一种基于遗传算法优化的极限梯度提升树模型。实验结果表明：这种港口物流需求预测模型取得的平均绝对误差和平均绝对百分比误差分别为 21.62 和 1.05%，优于近年来的港口物流需求预测模型。

关 键 词：港口物流；需求预测；极限梯度提升树；遗传算法
中图分类号：F552.7 **文献标识码**：A **文章编号**：1671-2250(2021)02-0071-07

0 引言

随着经济全球化的迅猛发展，港口作为与世界贸易交流的窗口也变得日趋重要。因此，通过历史数据对港口物流需求量进行准确预测，为港口的建设、规划提供一个数据支持，成为近年来国内外研究的一个热点。随着大数据信息时代的发展，与港口相关的统计数据也越来越详细丰富，这对港口物流需求预测建模提供了极大的便利。港口是一个复杂的大规模非线性系统，受到各种因素的影响与制约，如国家政策、港口腹地经济、地理位置和集疏运系统等。港口物流需求量往往以港口货物吞吐量或者集装箱吞吐量作为衡量指标，并且已被证实同腹地经济呈联动发展的趋势[1]。

王景敏等[2]利用三次指数平滑法(Cubic Exponential Smoothing，CES)预测广西北部湾港口物流需求量，把港口吞吐量当作港口物流需求目标，使用 2000 至 2009 年的港口吞吐量数据建立模型，预测 2010 至 2012 年的港口吞吐量。但是，港口是一个受到众多因素制约的复杂系统，而指数平滑法仅能进行单序列预测，没有考虑各种因素对模型预测的影响，所以这种港口物流需求预测方法存在一定的缺陷。王炳丹等[3]采用支持向量机预测广州港集装箱吞吐量，取得 4.8%到 8.0%的相对误差。支持向量机模型的预测结果虽然比较稳定，也考虑到影响港口物流需求量的因素，但容易出现过拟合问题，会降低在测试样本上的预测精度。李洪磊等[4]使用灰色模型 GM(1,1)预测大连港物流需求规模，使用 2006 至 2013 年的吞吐量数据建立模型，对 2014 至 2018 年的吞吐量进行预测，取得平均误差为 1.14%的预测结果。然而灰色模型 GM(1,1)和指数平滑法一样仅能进行单序列预测，忽略了港口物流需求的各种影响因素，存在一定的弊端。魏辉等[5]以港口吞吐量作为港口物流需求指标，并以大连港为例，采用 2009 至 2016 年的指标数据作为训练样本，以 2017 至 2018 年的数据作为测试样本，并考虑到港口是一个受众多因素影响的非线性系统，利用具有非线性预测特点的 BP 神经网络建立港口物流需求预测模型。然而 BP 神经网络在训练过程中会随机更新权重参数，容易陷入局部最优解以及出现欠拟合、过拟合的问题，从而会导致模型每次预测的结果存在一定的偏差，降低模型的预测精度。

收稿日期：2020-04-30
基金项目：宁波市科技厅惠民项目(2017C50028)；国家级大学生创新创业训练计划项目(201910876035)。
作者简介：李顺(1992-)，男，河南商丘人，浙江万里学院信息与智能工程学院 2018 级研究生，研究方向：数据挖掘与机器学习。

针对以上研究的不足，本文引入正则化项的极限梯度提升树模型，进行建模并预测宁波港物流需求量，以防止过拟合问题，并摒弃单序列模型忽略港口物流需求影响因素的缺点；然后采用遗传算法(Genetic Algorithm，GA)进行极限梯度提升树(eXtreme Gradient Boosting，XGBoost)模型的参数寻优，进一步提高模型的预测精度；最后采用平均绝对误差(Mean Absolute Error，MAE)和平均绝对百分比误差(Mean Absolute Percentage Error，MAPE)作为衡量模型精度的评估指标，并将构建的预测模型与其他港口物流需求预测模型进行比较。

1　模型介绍

XGBoost 算法是一种并行集成式机器学习算法，在梯度提升决策树的基础上增加了对目标函数的二阶泰勒展开和引入正则化项。本文将采用 XGBoost 算法建立宁波港物流需求预测模型。下面是对 XGBoost 算法原理的介绍。

(1) XGBoost 的学习目标函数。XGBoost 的目标函数由两部分组成，分别是训练损失函数与正则化项，如公式(1)所示。

$$Obj=\sum_{i=1}^{n}\iota(y_i,\hat{y}_i^t)+\sum_{i=1}^{t}\Omega(f_i)=\sum_{i=1}^{n}\iota(y_i,\hat{y}_i^{(t-1)}+f_t(x_i))+\Omega(f_t)+C \tag{1}$$

公式(1)中第一个等号后面第一项是训练样本的损失函数，第二项是正则化项。其中 y_i 是输入样本 x_i 的预测值，ι 是损失函数。回归问题常见的损失函数有均方误差 MSE、平均绝对误差 MAE 等。f_t 表示第 t 轮训练的树模型。$\Omega(f_i)$表示树模型的复杂度，对所有树模型的复杂度求和作为正则化项。前 $t-1$ 棵树的结构已经确定，故 $t-1$ 棵树的复杂度 C 可以视为常量。那么第 t 棵树要学习的目标函数则为上式第二个等号后面的前两项。

(2) 利用二阶泰勒展开式展开损失函数，如公式(2)所示。

$$\iota(y_i,y_i^{(t-1)}+f_t(x_i))\approx\iota(y_i,y_i^{(t-1)})+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i) \tag{2}$$

公式(2)中 g_i、h_i 分别是 $\iota(y_i,y_i^{(t-1)}+f_t(x_i))$的一阶导数和二阶导数。

(3) 将(2)代入到公式(1)中得到目标函数 Obj 的近似值：

$$Obj=\left[\sum_{i=1}^{n}\iota(y_i,\hat{y}_i^{(t-1)}+g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right]+\Omega(f_t)+C \tag{3}$$

公式(3)中 g_i 和 h_i 分别是 $\iota(y_i,y_i^{(t-1)}+f_t(x_i))$的一阶导数和二阶导数。$\iota(y_i,y_i^{(t-1)})$是之前 $t-1$ 轮的训练损失，不影响本轮的训练损失值，可以视其为常量。由于常量不影响本轮的训练结果，所以可以将所有常量去掉，最终目标函数值 Obj 为：

$$Obj\approx\sum_{i=1}^{n}\left[g_if_t(x_i)+\frac{1}{2}h_if_t^2(x_i)\right]+\Omega(f_t) \tag{4}$$

(4) 树的定义为叶子节点权重向量 ω 和叶子节点的映射关系 q，q 表达的是树的分支结构，如公式(5)所示。

$$f_t(x)=\omega_{q(x)},\omega\in R^T,q:R^D\rightarrow\{1,2,\cdots,T\} \tag{5}$$

公式(5)中 ω 是长度为 T 的一维向量，它的值是叶子节点的权重；q 代表一棵树的结构，它可以将输入映射到某个叶子节点，这里假设这棵树有 T 个叶子节点。

(5) 树的复杂度，包括叶子节点的数量 T 和叶子节点权重向量 ω_j 的 L2 范数。

$$\Omega(f_t)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{6}$$

(6) 由公式(5)和公式(6)可以得到新的目标函数，如公式(7)所示。

$$\begin{aligned} Obj &\approx \sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\Omega(f_t) \\ &=\sum_{i=1}^{n}\left[g_i\omega_{q(x_i)}+\frac{1}{2}h_i\omega_{q(x_i)}^2\right]+\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \\ &=\sum_{j=1}^{T}\left[\left(\sum_{i=I_j}g_i\right)\omega_{q(x_i)}+\frac{1}{2}\left(\sum_{i=I_j}h_i+\lambda\right)\omega_j^2\right]+\gamma T \end{aligned} \tag{7}$$

公式(7)中 I_j 的为每个叶子节点上的样本集合。

给定 C_i 和 H_j 的计算方法，如公式(8)所示。

$$C_i=\sum_{i\subset I_j}g_i, H_i=\sum_{i\subset I_j}h_i \tag{8}$$

(7) 将公式(8)代入到公式(7)中，再次得到新的目标函数，如公式(9)所示。

$$Obj^{(t)}=\sum_{j=1}^{T}\left[C_j\omega_j+\frac{1}{2}(H_j+\lambda)\omega_j^2\right]+\gamma T \tag{9}$$

对公式(9)求导得最值点和最优目标函数，分别如等式(10)和(11)所示。

$$\omega_j^*=-\frac{C_j}{H_j+\lambda} \tag{10}$$

$$Obj^*=-\frac{1}{2}\sum_{j=1}^{T}\frac{C_j^2}{H_j+\lambda}+\lambda T \tag{11}$$

(8) 找出最佳分裂节点

在训练过程中，当建立第 t 棵树时，XGBoost 采用贪心算法进行树结点的分裂。每一个节点分裂后需要计算目标函数的增益。如果增益大于 0，则目标函数就会下降。可以对目标函数设置一个阈值，当达到该阈值时模型就会停止训练。增益的计算方法如公式(12)所示。

$$\begin{aligned} Gain&=Obj_{L+R}-(Obj_L+Obj_R) \\ &=\left[-\frac{1}{2}\frac{G_L+G_R}{H_L+H_R+\lambda}+\lambda\right]-\left[-\frac{1}{2}\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}+2\lambda\right] \\ &=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)2}{H_L+H_R+\lambda}\right]-\lambda \end{aligned} \tag{12}$$

2　基于遗传算法的 XGBoost 模型参数寻优

遗传算法是一种受生物有机体的遗传过程启发而被提出的启发式智能进化算法。遗传算法可以采用多种编码方式进行编码，比较常用的是二进制编码。二进制遗传算法中的每个染色体都包含多个具有二进制值的基因。用 0 和 1 来确定每个个体的属性。一个种群由一组染色体组成。使用适应度函数评估每个染色体的优点来选择适合的染色体，从而产生新的染色体。在该过程中，选择两个合适的染色体，并通过交叉步骤进行合并，以产生新的后代。最后，对种群进行变异操作，来增加个体的随机性，从而减小陷入局部最优的可能性[6]。

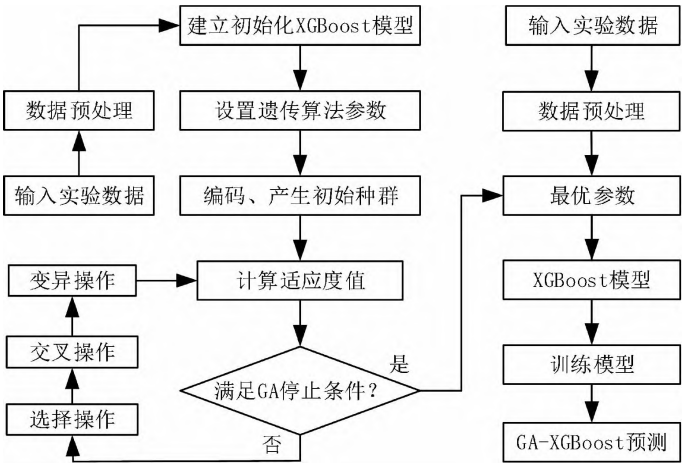

图 1　遗传算法优化 XGBoost 参数的流程

标准遗传算法的数学模型可以表示为：

$$SGA=f(C,E,P_0,N,\Phi,\Psi,\Gamma,T) \tag{13}$$

上式中，C 表示个体编码方法；EE 表示个体适应度评价函数；P_0 为初始种群；N 种群大小；Φ 为选择算子；Γ 为交叉算子；Ψ 为变异算子；Γ 为遗传算子终止条件。遗传算法优化 XGBoost 参数的流程如图 1 所示。遗传算法优化 XGBoost 模型参数的主要步骤如下：

Step1：输入实验数据并划分训练样本集和测试样本集；

Step2：归一化处理以消除变量之间的量纲差异；

Step3：建立初始化 XGBoost 模型；

Step4：设置遗传算法参数编码方式、初始种群、变异概率等；

Step5：对染色体进行解码，计算种群内的个体适应度值；

Step6：对适应度值进行判断，若满足停止条件（达到理想的适应度值），则将输出的最佳参数带入 XGBoost 模型进行训练，否则执行第 7 步；

Step7：进行染色体的选择、交叉和变异操作；

Step8：再次对适应度值进行判断，若满足停止条件，则将输出的最佳参数带入 XGBoost 模型进行训练，否则执行第 7 步；

Step9：利用训练好的 GA-XGBoost 模型进行回归预测。

3　实验数据

3.1　数据的收集

对于港口物流需求指标，以往大部分研究都是选择集装箱吞吐量或者货物吞吐量作为衡量指标。本文继承以往的研究，选择集装箱吞吐量作为宁波港物流需求指标。对于港口物流需求影响因素指标，需要结合腹地经济进行选择；而国家政策、港口地理环境等影响因素短期内不易变化，所以不予考虑。宁波港的直接经济腹地包括浙江省的 11 个地级市[7]，如图 2 所示。宁波港以宁波市为依托，进而辐射整个浙江省。因为港城联动发展的紧密关系[8]，所以在选取宁波港物流需求影响因素时，除了考虑宁波港的直接腹地（浙江省）经济指标，同时也重点考虑了宁波港依托城市（宁波市）经济指标。为了科学地选取物流需求的影响指标，需要遵循高度相关性、全面性以及可获取性的原则。指标须涵盖经济腹地 GDP、腹地经济产业结构、对外贸易水平、交通运输状况、居民消费水平等，最终在选取指标上按照港口所在省份和港口所在城市划分。港口所在省份的输入指标有：$X1$ 生产总值（亿元）、$X2$ 第一产业（亿元）、$X3$ 第二产业（亿元）、$X4$ 第三产业（亿元）、$X5$ 进出口贸易总额（亿美元）、$X6$ 社会消费品零售总额（亿元）、$X7$ 公路通车里程（万公里）、$X8$ 铁路营业长度（万公里）、$X9$ 货运周转量（亿吨公里）。港口所在城市的输入指标有：X10 生产总值（亿元）、$X11$ 口岸进出口总额（亿美元）、$X12$ 社会消费品零售总额（亿元）、$X13$ 城市人均可支配收入（亿元）、X14 城市人均消费支出（亿元）、$X15$ 农村人均可支配收入（亿元）、$X16$ 农村人均消费支出（亿元）、$X17$ 固定资产投资总额（亿元）、$X18$ 货运量（万吨）。以上指标均来自浙江省统计年鉴和宁波市统计年鉴，时间范围在 1990 年到 2018 年。

图 2　宁波港的直接经济腹地

3.2　灰色关联法筛选指标

初选的输入变量有 18 个，将利用灰色关联法计算这些输入指标同输出指标（宁波港集装箱吞吐量）之间的相关度。

(1) 计算关联系数。给定参考数列 $S=\{x_0,x_1,x_2,\cdots,x_i,\cdots,x_m\}$，$x_0$ 为母序列，x_i 是被比较和母序列相关性的子序列。关联系数如公式(14)所示。

$$\gamma(x_0,x_i)=\frac{\min\min\Delta_{0i}+\zeta\max\max\Delta_{0i}}{\Delta_{0i}+\zeta\max\max\Delta_{0i}} \tag{14}$$

上式中 $\Delta_{0i}=|x_0-x_i|$，$\zeta\in[0,1]$，称为分辨系数。ζ 越小，分辨力越大，通常取 $\zeta=0.5$ 。

(2) 计算关联度。关联系数是描述子序列与母序列在同时刻的相关程度。由于各个时刻都有一个关联系数，信息比较分散，不利于比较。因此，将各个时刻的关联系数求平均值。关联度 γ_i 如公式(15)所示。

$$\gamma_i=\frac{1}{n}\sum_{i=1}^{n}\gamma(x_0,x_i) \tag{15}$$

先将数据进行归一化到[0,1]区间，消除变量之间的量纲差异后，再通过以上方法计算集装箱吞吐量与各影响指标的关联度。以上步骤的实际操作采用 Python 编程实现，结果如图 3 所示。其中 $X2$ 第一产业、$X7$ 公路通车里程，因关联强度较弱，选择舍弃。

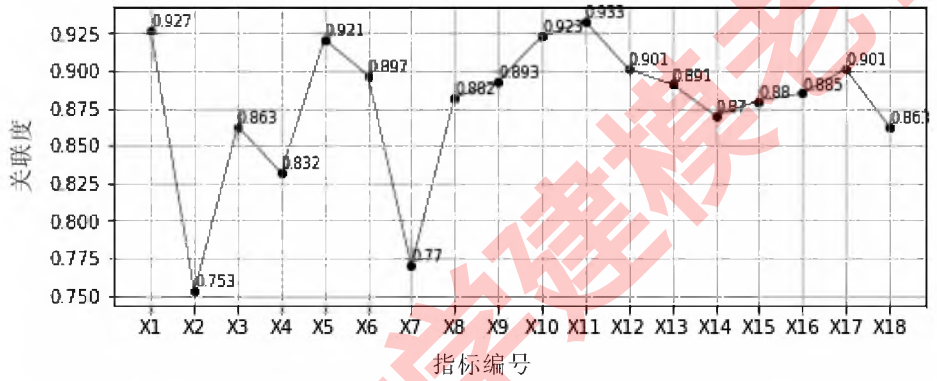

图 3　集装箱吞吐量与其各影响指标关联度

4　实验与结果分析

4.1　GA-XGBoost 预测集装箱吞吐量

本文基于 Python 遗传算法工具箱 Geatpy 来实现优化 XGBoost 模型的参数。该工具箱操作简单、方便，非常适合遗传算法的优化问题。实现遗传算法优化 XGBoost 参数用到的 Python 库和函数如表 1 所示。

表 1　实现 GA-XGBoost 相关的 Python 库和函数

Python 库和函数	实现功能
import numpy,pandas	读取实验数据
from sklearn.preprocessing import MinMaxScaler,scale	数据归一化
from sklearn import xgboost	XGBoost 模块
from sklearn.metrics import ,mean_absolute_error	MAE 函数
import geatpy	遗传算法模块

首先进行数据预处理即归一化，消除量纲差异。然后将 1990 年至 2013 年的数据作为训练集，将 2014 年至 2018 年的数据作为测试集。实现遗传算法优化 XGBoost 模型的具体步骤如下：

(1) 设置 XGBoost 初始参数。XGBoost 参数有很多，主要优化三个较为重要的参数，它们分别是树的最大深度 max_depth、学习率 learning_rate 和迭代器数量 n_estimators。结合样本的数量，设置参数的初始范围。max_depth 的初始范围为[1,10]，类型为整型，它可以有效防止过拟合；learning_rate 的初始范围为(0,0.3)，类型为浮点型，它可以减少每一步的权重，提高模型的鲁棒性；n_estimators 的初始范围为[1,50]，类型为整型，该参数是最大迭代次数。其他初始参数都设为默认值。

(2) 设置遗传算法参数。在 Geatpy 工具箱中遗传算法的染色体有三种最基础的编码方式，分别是 BG(二进制/格雷编码)、RI(实数整数混合编码)以及 P(排列编码)；但是一条染色体只能是这三种编码方式的一种。这里采用编码方式 BG，初始种群 NIND 设为 20，最大迭代次数 MAXGEN 为 100，进化停滞判断阈值 TrappedValue 设为 1e-6，进化停滞计数器最大上限值 maxTrappedCount 设为 10。选择算子并设置为 dup(基于适应度排序的直接复制选择)，交叉概率 PC 设为 0.95，变异概率 PM 设为 0.05。

(3) 适应度函数选择平均绝对误差 MAE，即以此最小值来作为遗传算法求解的最优解。

经过步骤(1)、(2)、(3)之后开始训练模型：model=xgboost.XGBRegressor(max_depth=m,n_estimators=n,learning_rate=lr).fit(x_train,y_train)，输出拟合值 predict_value=model.predict(x_train)，然后计算适应度 obj =mean_absolute_error(y_train, predict_value)。得到最优的适应度 MAE 为 0.0035；得到最优的控制变量值为：树的最大深度 max_depth=5、学习率 learning_rate=0.1941、迭代器数量 n_estimators=30。有效进化代数100，最优的一代是第 74 代，适应度曲线如图 3 所示，拟合效果如图 4 所示。由此可见 GA-XGBoost 模型具有很好的拟合效果。

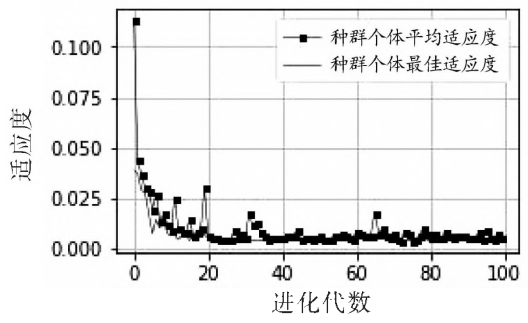

图 3　遗传算法参数寻优的适应度曲线

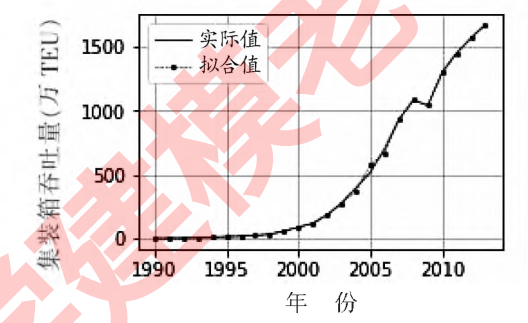

图 4　GA-XGBoost 模型的拟合结果

4.2 模型预测结果分析

采用平均绝对误差 MAE 和平均绝对百分比误差 MAPE 作为回归预测的误差指标。具体公式分别如(16)和(17)所示。其中 N 为要预测的样本个数，y_{True} 和 $y_{Predict}$ 分别是真实值和预测值。

$$MAE=\frac{1}{N}\sum_{i=1}^{N}|y_{Predict}-y_{True}| \tag{16}$$

$$MAPE=\sum_{i=1}^{N}\left|\frac{y_{True}-y_{Predict}}{y_{True}}\right|\times\frac{100\%}{N} \tag{17}$$

表 2 模型的预测结果

年份	真实值(万TEU)	GA-XGBoost(万TEU)	CES(万 TEU)	SVM(万 TEU)	GM(1,1)(万TEU)	BP-NN(万TEU)
2014	1870	1870.58	1863.24	1885.27	1834.04	1882.27
2015	1982.4	1940.88	2025.84	2027.53	1992.01	2013.33
2016	2069.6	2124.54	2193.54	2155.01	2163.58	2102.1
2017	2356.6	2349.54	2366.34	2308.51	2349.93	2352.21
2018	2510	2506.01	2544.25	2427.32	2552.33	2461.68

将测试样本输入到已经获取最优参数的 GA-XGBoost 模型中进行预测：predict_y=model.predict(x_test)，然后将预测结果进行反归一化处理 scaler.inverse_transform(predict_y)，得到 2014 年至 2018 年的集装箱吞吐量。同时，将提出的 GA-XGBoost 模型与近年来的港口物流需求预测模型做对比，实验结果如表 2 所示。采用公式(16)和(17)计算预测误差，结果如表 3 所示。从预测结果看，本文所建立模型的预测效果要优于三次指数平滑法、灰色模型 GM(1,1)、支持向量机 SVM 和 BP 神经网络模型。出现这种结

果的主要原因在于：XGBoost 模型在目标函数上采用了二阶泰勒展开，使训练过程挖掘到更深的信息；引入了正则化项，有效地防止了过拟合问题；基于遗传算法的参数寻优进一步提高了 XGBoost 模型的预测精度。

表 3　模型的预测误差

模型	MAE 误差	MAPE 误差
GA-XGBoost	21.62	1.05%
CES	43.62	2.06%
SVM	54.15	2.45%
GM(1,1)	37.71	1.78%
BP-NN	25.68	1.18%

5　结语

通过灰色关联法分析港口物流需求和其影响指标之间的关联度，发现对外贸易经济水平、腹地产业结构、交通运输、居民经济消费水平等对宁波港物流需求量有很大影响；建立的 GA-XGBoost 港口物流需求预测模型取得的 MAE 和 MAPE 误差分别为 21.62 和 1.05%，要优于三次指数平滑法、支持向量机、灰色模型和 BP 神经网络。下一步的工作应探究不同的优化算法，进一步提高港口物流需求预测模型的精度。

参考文献：

[1]吴桥，曹非燕. 宁波舟山港与浙江经济腹地联动发展实证研究[J]. 港口经济，2016(11):5-9.
[2]王景敏，朱芳阳，等. 广西北部湾港口物流需求预测及发展模式研究[J]. 物流科技，2010(12):26-28.
[3]王炳丹. 基于 SVM 的集装箱吞吐量预测研究[D]. 北京市：北京交通大学，2011.
[4]李洪磊，王德闯. 基于灰色系统理论的大连港口物流需求预测[J]. 物流科技，2016(001):17-20.
[5]魏辉. 基于 BP 神经网络的港口物流需求预测[J]. 决策探索(中)，2019(09):86-88.
[6]Ghamisi P，Benediktsson J A. Feature Selection Based on Hybridization of Genetic Algorithm and Particle Swarm Optimization[J]. Geoscience & Remote Sensing Letters IEEE，2015(2):309-313.
[7]吴桥. 宁波港—浙江经济腹地空间结构演变及其实证研究[J]. 港口经济，2016(01):10-13.
[8]占慧杰，冯路. 创新港城联动路径 助推海洋经济发展[J]. 宁波经济(三江论坛)，2017(07):19-20.

Forecast of Ningbo Port Logistics Demand Based on GA-XGBoost

LI Shun[1]，LI Jun[1]，WU Xin[1]，MEI Bi-zhou[2]
(1.Zhejiang Wanli University，Ningbo Zhejiang 315100;
2. Zhejiang Yiduan Precision Machinery Co.，Ltd.，Xiangshan Zhejiang 315700)

Abstract: Through grey correlation method，this paper analyzed the correlation between Ningbo port logistics demand and hinterland economy，established Ningbo port logistics demand forecasting index set，and put forward the eXtreme Gradient Boosting model based on genetic algorithm optimization. The results showed that the mean absolute error and the mean absolute percentage error obtained by this port logistics demand prediction model were 21.62 and 1.05% respectively，which were better than those of the port logistics demand prediction models in recent years.

Key Words: port logistics; demand forecasting; eXtreme Gradient Boosting; genetic algorithm

(责任编辑：顾亿天)