文章编号:1671-9352(2024)01-0017-10 DOI:10.6040/j.issn.1671-9352.4.2022.8254

多服务器串联排队系统中平均排队时间的预测

李绎冉[1,2],赵宁[1,2],张志坚[1*]
(1.昆明理工大学理学院, 云南 昆明 650500; 2.昆明理工大学数据科学研究中心, 云南 昆明 650500)

摘要:研究了具有 2 个服务站且缓冲区无限的多服务器串联排队系统,利用机器学习的线性回归模型和非线性回归模型对 2 个站的平均排队时间进行预测,并对各种机器学习方法的预测结果进行误差分析。数值实验结果显示,非线性回归模型优于线性回归模型,RF、XGBoost、GBDT 方法可以作为分析多服务器串联排队网络的有效手段。

关键词:串联排队系统;多服务器;机器学习;平均排队时间;模拟
中图分类号:O226 **文献标志码**:A
引用格式:李绎冉,赵宁,张志坚.多服务器串联排队系统中平均排队时间的预测[J]. 山东大学学报(理学版),2024,59(1):17-26.

Prediction of average queue time in multi-server tandem queueing systems

LI Yiran[1,2], ZHAO Ning[1,2], ZHANG Zhijian[1*]
(1. Faculty of Science, Kunming University of Science and Technology, Kunming 650500, Yunnan, China; 2. Data Science Research Center, Kunming University of Science and Technology, Kunming 650500, Yunnan, China)

Abstract: This paper studies a multi-server tandem queueing system with two stations and infinite buffers before each station. The average queueing time of the two stations is predicted by linear regression models and nonlinear methods of machine learning, and the error in the prediction results of various machine learning methods is analyzed. Numerical experiments show that the nonlinear method exhibits better performance than the linear regression model. Moreover, the RF, XGBoost and GBDT methods are effective to predict the average waiting time of multi-server tandem queueing networks.

Key words: tandem queueing system; multi-server; machine learning; average waiting time; simulation

0 引言

串联排队系统高效、准确的性能评价对于有效设计和控制复杂的排队系统至关重要。串联排队系统由多个服务站串联而成,是排队网络的基本结构,系统中由于站与站之间相互依赖,导致系统的性能分析比较困难。

对于到达过程是更新过程、服务时间具有马尔可夫性且每个服务站只有一个服务器的串联排队系统,已经有多位学者给出了分析方法。Zhu[1]研究了具有批量到达,且站与站之间不存在缓冲区的串联排队系统,其中顾客的到达过程是泊松过程,服务时间均服从一般分布,得到了顾客逗留时间的拉普拉斯变换。Gómez-Corral[2]利用马尔可夫更新过程分析了具有马尔可夫到达过程(markovian arrival processes, MAPs)的 2 个站的串联排队系统,得到了系统的稳态概率分布、平均队长等系统性能指标。Van Houd 等[3]基于第一个站的首位顾客在系统内消耗的时间,提出了一种求解带阻塞的离散时间串联排队系统响应时间的方法。

收稿日期:2022-08-12; **网络出版时间**:2023-09-07 16:03:08
网络出版地址:https://link.cnki.net/urlid/37.1389.N.20230906.1639.016
基金项目:2021 年度工业控制技术国家重点实验室开放课题(ICT2021B51)
第一作者简介:李绎冉(1996—),女,硕士研究生,研究方向为排队论. E-mail:1639061882@ qq.com
*通信作者简介**:张志坚(1980—),男,讲师,博士,研究方向为机器学习. E-mail:zhijian@ kust.edu.cn

同时,Lian 等[4]针对具有 MAPs 输入的单服务器串联排队系统,提出了一种能够有效地分析服务站间的时间依赖和稳态队列长度分布的方法,同时还得到了顾客在排队网络中的逗留时间。Wu 等[5]通过模拟实验分析了串联排队系统中各服务站之间的关系,探究了具有一般服务时间的串联排队系统中上游服务站对下游服务站的排队时间的影响。吴登磊等[6]引入指标比的概念来刻画串联排队系统中上游服务站对下游服务站的排队时间的影响,并运用指标比的性质近似得出了到达过程为泊松过程、服务时间服从一般分布、具有 2 个服务站的串联排队系统的平均排队时间。侯佳辰等[7]研究了一个到达过程是更新过程、服务时间服从一般分布、具有 2 个服务站的单服务器串联排队系统,根据到达时间间隔和服务时间的三阶矩,构建相应的马尔可夫过程,同时利用矩阵几何解的方法对系统的平均排队时间进行近似求解。

多服务器串联排队系统指的是由多个相互依赖的服务站串联而成,并且每个服务站至少有一个服务器的排队系统。与单服务器串联排队系统相比,由于服务站内服务器个数的增加,使多服务器串联排队系统的理论分析更加困难,因此多服务器串联排队系统的性能分析仍有待进一步突破。目前,多服务器串联排队系统的研究主要关注于近似分析或相关应用。Kim 等[8]研究了具有无限缓冲区、第一个服务站为单服务器而第二个站为多服务器的串联排队系统,并通过数值实验计算出系统的平稳分布和性能指标。Dudin 等[9]利用具有马尔可夫到达过程的串联排队系统作为远程技术支持模型,提出了计算该系统平稳分布的数值算法,并推导出逗留时间的拉普拉斯变换。Kim 等[10]根据顾客类型对多服务器串联排队系统的缓冲区进行划分,分析了系统状态的遍历性条件和稳态分布,并推导出顾客逗留时间分布的拉普拉斯变换。Sinu Lal 等[11]以医院为应用背景,研究了第一个服务站具有多个相同且平行服务器、第二个站只有一个服务器的串联排队系统,并利用矩阵解析方法得到系统的平稳分布。Banu 等[12]通过 Burke 定理,研究了具有 4 个平行队列且缓冲区无限的开放排队网络,每个队列均是 $M/M/1$ 模型,构建了分析系统内平均顾客数、平均等待时间等性能的公式。Sagir 等[13]研究了到达过程为泊松过程,服务时间服从指数分布的串联排队系统,并且证明了 2 个站的客户数和等待时间在稳态下是相互独立的。Kumar 等[14]分析了具有多处理器的两阶段串联呼叫重审排队网络,在稳态条件下,利用矩阵分析法求解系统的平稳分布等指标。

综上所述,当服务时间不具有马尔可夫性,上游站离去过程是非更新过程时,串联排队模型的排队时间没有解析表达式,只能通过近似方法来分析。机器学习的方法作为一种近似分析方法,在排队系统的性能分析预测方面取得了较好的效果。Efrosinin 等[15]通过人工神经网络方法研究了具有异构服务器排队系统中的经典优化问题,提供了最优阈值的估计。Tan 等[16]基于高斯过程回归算法提出了一种分析单服务器开放排队网络模型的有监督学习方法(supervised learning based queueing network analyzer, SLQNA),用来分析由延迟、批量、合并和拆分块组成的制造系统开放排队网络模型。Khayyati 等[17]根据 SLQNA 来预测不同服务规则下排队网络中队列延迟、拆分、批量和合并等情况的性能指标,采用高斯过程回归方法预测各队列的离去时间和周期的分布特征,并得出了对应队列的均值、变异系数和周期。对于评估排队系统以及排队网络的性能指标,SLQNA 是一种通用、准确和有效的方法。

机器学习在排队网络的性能分析方面取得了一定的研究成果,但目前的成果主要关注于单站的排队系统或是特定生产制造系统的分布特征研究,并未将研究成果推广至生产生活中应用更广泛的串联排队系统,例如 $M/G/1\rightarrow/G/c$、$M/G/c\rightarrow/G/1$ 和 $M/G/c\rightarrow/G/c$ 等串联排队系统。

本文的创新点是:(1) 研究了到达过程为泊松过程、服务时间为一般分布、具有 2 个服务站、每个站内有多个独立且相同服务器的串联排队系统,该模型更具有一般性,更符合实际生产生活需求;(2) 利用仿真模拟,获得系统内各站点的性能参数,数据更具有可靠性;(3) 采用机器学习方法中的线性回归模型和非线性回归模型对多服务器串联排队系统的平均排队时间进行预测。

1 模型描述

本文研究具有 2 个服务站的 $M/G_1/N_1\rightarrow/G_2/N_2$ 串联排队系统,假设第 i 个站有 N_i 个性能相同的服务器($i=1,2$),如图 1 所示。2 个服务站的缓冲区均为无限大,顾客到达后在 2 个服务站中依次完成服务,满足先到先服务(first come first served, FCFS)的服务规则。顾客到达系统后,如果发现第一个站有空闲的服务器,则任选一个空闲服务器接受服务,否则需要等待;当完成第一个站的服务后,顾客进入第二个站接受服

务。顾客到达过程为泊松过程,相邻 2 个顾客到达时间间隔 X 服从指数分布,每个服务器的服务时间均服从一般分布,第一个站的服务时间记为 S_1,第二个站的服务时间记为 S_2。X、S_1 和 S_2 的平方变异系数分别记为 C_a^2、$C_{S_1}^2$和 $C_{S_2}^2$,且

$$C_a^2=\frac{D(X)}{E^2(X)},\tag{1}$$

$$C_{S_1}^2=\frac{D(S_1)}{E^2(S_1)},\tag{2}$$

$$C_{S_2}^2=\frac{D(S_2)}{E^2(S_2)}。\tag{3}$$

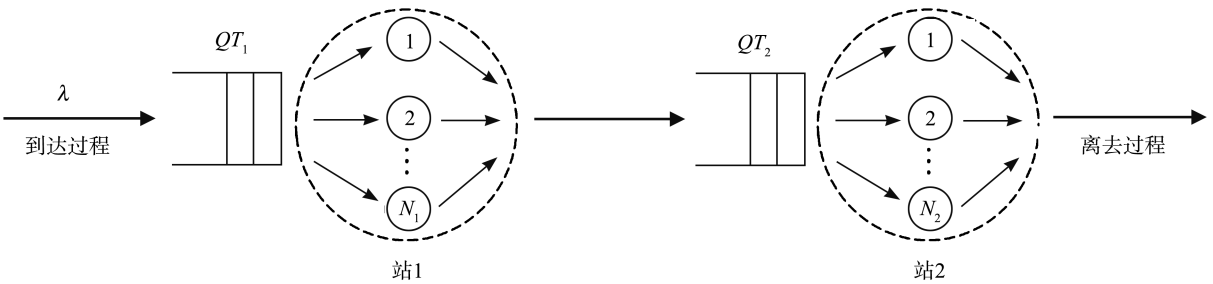


图 1　2 个站的多服务器串联排队系统

Fig.1　Multi-server tandem queueing system with two stations

假设系统的到达率为 λ,$\lambda=1/E(X)$,第 i 个服务站的服务速率为 μ_i,$\mu_i=N_i/E(S_i)$,$i=1,2$,系统中第 i 个服务站的服务器利用率为 $\rho_i=\lambda/\mu_i$,令 $\rho=\max(\rho_1,\rho_2)$,服务强度高的工作站被称为瓶颈站。系统中第 i 个服务站的平均排队时间记为 $E(QT_i)$,$i=1,2$。

在具有 2 个服务站的串联排队系统中,顾客在第一个站完成服务后进入第二个站,第一个站的输出过程是第二个站的到达过程,一般情况下,第二个站的到达过程不同于系统的初始到达过程,因此,第二个站的平均排队时间受到第一个站的影响。

本文研究的 $M/G_1/N_1\rightarrow/G_2/N_2$ 串联排队系统中,因为第一个站的离去过程不是更新过程,所以第二个站的平均排队时间无法利用理论方法进行精确求解。本文利用 MATLAB 进行模拟,在模拟中,假设相邻顾客的到达时间间隔、服务时间分布均服从 Gamma 分布,与平均排队时间相关的参数记为(C_a^2,N_1,N_2,μ_1,μ_2,$C_{S_1}^2$,$C_{S_2}^2$,ρ),将参数任意组合,均可构成一个多服务器串联排队系统。对于每个系统,模拟 30 个样本,每个样本中串联排队系统第 400 001 个至第 600 000 个作业为 2 个站的平均排队时间 $E(QT_i)$,$i=1,2$,平均排队时间模拟值的置信水平大于 95%,保证了模拟数据的可靠性。

2　机器学习方法对比

本文利用 8 种有监督学习的机器学习方法[18]预测多服务器串联排队系统 2 个站的平均排队时间。8 种机器学习方法分为线性回归模型和非线性回归模型。线性回归模型包括多元线性回归(multiple linear regression, MLR)模型、岭回归(ridge regression, RR)模型和 LASSO 回归(least absolute shrinkage and selection operator regression, LASSO)模型。非线性回归模型包括决策树(decision tree, DT)算法、随机森林(random forest, RF)算法、梯度提升树(gradient boosting decision tree, GBDT)算法和 XGBoost(eXtreme gradient boosting)算法,以及与决策树算法功能类似的 K 近邻(K nearest neighbor, KNN)算法。

MLR 模型的原理是根据已知的自变量 X 的值预测未知的因变量 y 的值。模型的目标函数为

$$y=X\beta+\varepsilon,\tag{4}$$

其中:β 表示多元线性回归模型的回归系数;ε 表示模型拟合后每一个样本的误差项。

RR 模型是一种用于共线性数据分析的有偏估计回归模型,通过在目标函数上添加惩罚项来缩减线性回归模型的偏回归系数。RR 模型的目标函数为

$$J(\beta)=\sum(y-X\beta)^2+\lambda\|\beta\|_2^2, \tag{5}$$

其中:λ 为非负数,表示惩罚因子;$\|\beta\|_2^2$ 为 l_2 正则,表示回归系数 β 的平方和;$\lambda\|\beta\|_2^2$ 为惩罚项。当 $\lambda=0$ 时,$J(\beta)$ 为线性回归模型的目标函数;当 $\lambda\to+\infty$ 时,通过缩减回归系数 β 来使目标函数达到最小。

LASSO 回归是一种压缩估计模型,与岭回归模型类似,LASSO 回归同样属于缩减性估计。在回归系数的缩减过程中,将不重要的回归系数直接缩减为 0,达到变量筛选的功能,并且降低复杂度。LASSO 回归模型的目标函数为

$$J(\beta)=\sum(y-X\beta)^2+\lambda\|\beta\|_1, \tag{6}$$

其中:λ 为惩罚项系数;$\|\beta\|_1$ 为回归系数 β 的 l_1 正则,表示所有回归系数绝对值的和;$\lambda\|\beta\|_1$ 为目标函数的惩罚项。

DT 算法的原理是在已知各种情况发生概率的基础上,通过构成决策树来求取净现值的期望值大于等于 0 的概率。决策树作为一个树结构,代表的是对象属性与对象值之间的一种映射关系。在机器学习中,DT 算法简单直观,不需要复杂的数学推理,具有很强的解释性,可以用来预测数值型因变量和分类离散型变量,也可以作为其他算法的弱分类器。

RF 算法属于集成算法,算法的核心思想是通过多棵决策树的投票机制来实现预测或分类。“森林”是由多棵经过充分生长的分类回归决策树(classification and regression tree, CART)组成的集合。“随机”是指构成多棵决策树的数据是随机生成的。该算法的优势是运行速度快和预测准确率高。

GBDT 算法的原理为利用损失函数的负梯度值近似残差,简化目标函数的求解。该算法的优势主要体现在提升(boosting)、梯度(gradient)和决策树(decision tree) 3 个方面。如果因变量为连续型的数值型变量,则 GBDT 算法中有多种具有一阶导函数的损失函数可供选择。

XGBoost 算法的根本思想是将多个弱决策树串联起来形成一个较强的决策树,从而得到更准确的结果。相比 GBDT,XGBoost 的优势主要体现在以下几个方面:(1)支持线性分类器;(2)可以自定义损失函数;(3)防止过拟合;(4)计算量小;(5)对于特征值有缺失的样本,可以自动学习出它的分裂方向;(6)支持并行计算,运行效率高。

KNN 算法与决策树算法类似,是一种惰性学习算法,即模型的构建与未知数据的预测同时进行。算法原理为搜索最近的 K 个已知类别样本用于未知类别样本的预测。

综合以上 8 种机器学习的方法可以发现:MLR、RR 和 LASSO 回归模型的区别在于目标函数的不同;RR 和 LASSO 回归可以缩减目标函数中的回归系数,降低模型的复杂度;RF、GBDT 和 XGboost 算法是以决策树为弱分类器的算法,通过分类器组合、损失函数负梯度近似残差等方式提高算法的预测精度。

由于多服务器串联排队系统中上、下游服务站之间存在依赖性,通过理论方法很难求解平均排队时间,因此本文利用以上 8 种方法对多服务器串联排队系统 2 个站的平均排队时间进行预测。

3 数值实验

本章利用 8 种机器的方法分别对 $M/G_1/N_1\to/G_2/N_2$ 串联排队系统的第一个站和第二个站的平均排队时间进行预测。模拟系统参数的取值如下:$\mu_1\in\{1/10,1/20,1/25,1/29,1/30\}$,$\mu_2\in\{1/10,1/20,1/25,1/30\}$,$C_a^2=1$,$C_{S_1}^2\in\{0.1,0.5,0.9,2,5,10\}$,$C_{S_2}^2\in\{0.1,0.5,0.9,2,5,10\}$,$N_1\in\{2,5,10,100\}$,$N_2\in\{2,5,10\}$,$\rho\in\{0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.95\}$,以上的 8 个参数任意组合,都能构成一个多服务器串联排队系统。

3.1 第一个站平均排队时间的预测

在 $M/G_1/N_1\to/G_2/N_2$ 串联排队系统中,系统缓冲区无穷大,由于下游站点对上游站点没有影响,因此第二个站对第一个站的排队时间没有影响,第一个站平均排队时间 $E(QT_1)$ 只与参数 C_a^2、N_1、$C_{S_1}^2$、S_1、ρ 有关。当串联排队系统中第一个站的平均排队时间数据较小时,会造成机器学习的模型无法识别数据,因此首先对数据进行筛选,去除 $E(QT_1)<0.1$ 的数据,有效数据为 15 092 组。

在机器学习中,将同一个数据集按照不同的比例划分为测试集和训练集。数据集划分比例的不同会导

致预测误差的不同,因此,针对第 2 章中的 8 种机器学习方法,在本章中,每种机器学习方法都在同一组参数下根据不同的训练集、测试集比例进行了 8 次实验,训练集的比例分别为 0.95、0.9、0.85、0.8、0.75、0.7、0.6、0.5,对应的测试集比例为 0.05、0.1、0.15、0.2、0.25、0.3、0.4、0.5。根据测试集比例的变化,将 3 种线性回归模型和 5 种非线性回归模型的数值实验结果进行整理、对比,如表 1、2 所示。

表 1　线性回归模型预测 $E(QT_1)$ 的平均相对误差

Table 1　The mean relative error of predicting $E(QT_1)$ by linear regression model　　%

比例	MLR	RR	LASSO
0.05	300.99	301.11	301.61
0.10	267.60	267.67	268.10
0.15	288.79	288.91	289.34
0.20	280.00	280.06	280.48
0.25	300.53	300.61	301.05
0.30	286.05	286.14	286.51
0.40	287.56	287.71	288.01
0.50	284.37	284.47	284.78

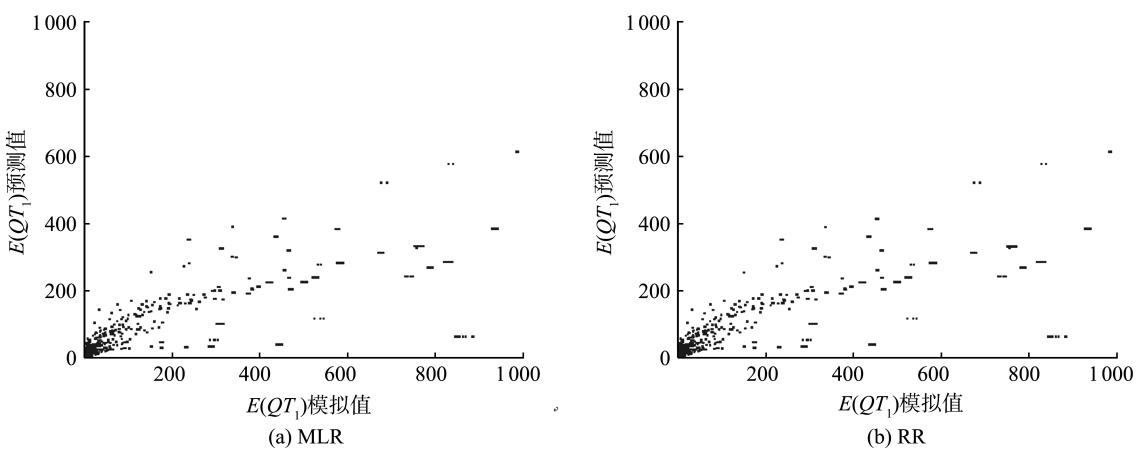
表 2　非线性回归模型预测 $E(QT_1)$ 的平均相对误差

Table 2　The mean relative error of predicting $E(QT_1)$ by nonlinear method　　%

比例	DT	RF	GBDT	XGBoost	KNN
0.05	2.29	0.45	0.78	0.56	0.46
0.10	3.12	0.42	0.79	0.66	0.42
0.15	5.03	0.48	0.85	0.65	0.48
0.20	7.44	0.46	0.81	0.63	0.48
0.25	8.81	0.46	0.80	0.66	0.46
0.30	9.21	0.47	0.80	0.62	0.46
0.40	9.93	0.49	0.79	0.66	0.46
0.50	11.34	0.56	0.84	0.74	0.50

由表 1 可知,线性回归模型 MLR、RR 和 LASSO 的预测效果较差,3 种模型的误差范围均在 267%~300%左右。表 2 中,非线性回归模型的预测误差随着测试集比例增大也逐渐增大;DT 算法的误差范围为 2.29%~11.34%;RF、GBDT、KNN、XGBoost 算法的预测误差均小于 1%,当测试集比例低于 0.3 时,RF 算法的预测误差为最小,预测效果较好。与非线性回归模型相比,线性回归模型的预测效果较差,因此,本文所考虑的 3 种线性回归模型不适合用于预测串联排队系统排队时间。

当测试集比例过大或过小时,会出现样本比例失衡、预测结果过拟合等情况。本章中,8 种不同的测试集比例、预测误差都相对稳定。结合数值实验中的数据集样本量,测试集比例为 0.25 是较为合适的比例,也是机器学习预测时常用的比例,因此,选取测试集比例为 0.25 时,分别将每种机器学习方法求出的第一个站平均排队时间的预测值和模拟值进行线性相关性对比,结果如图 2 所示。



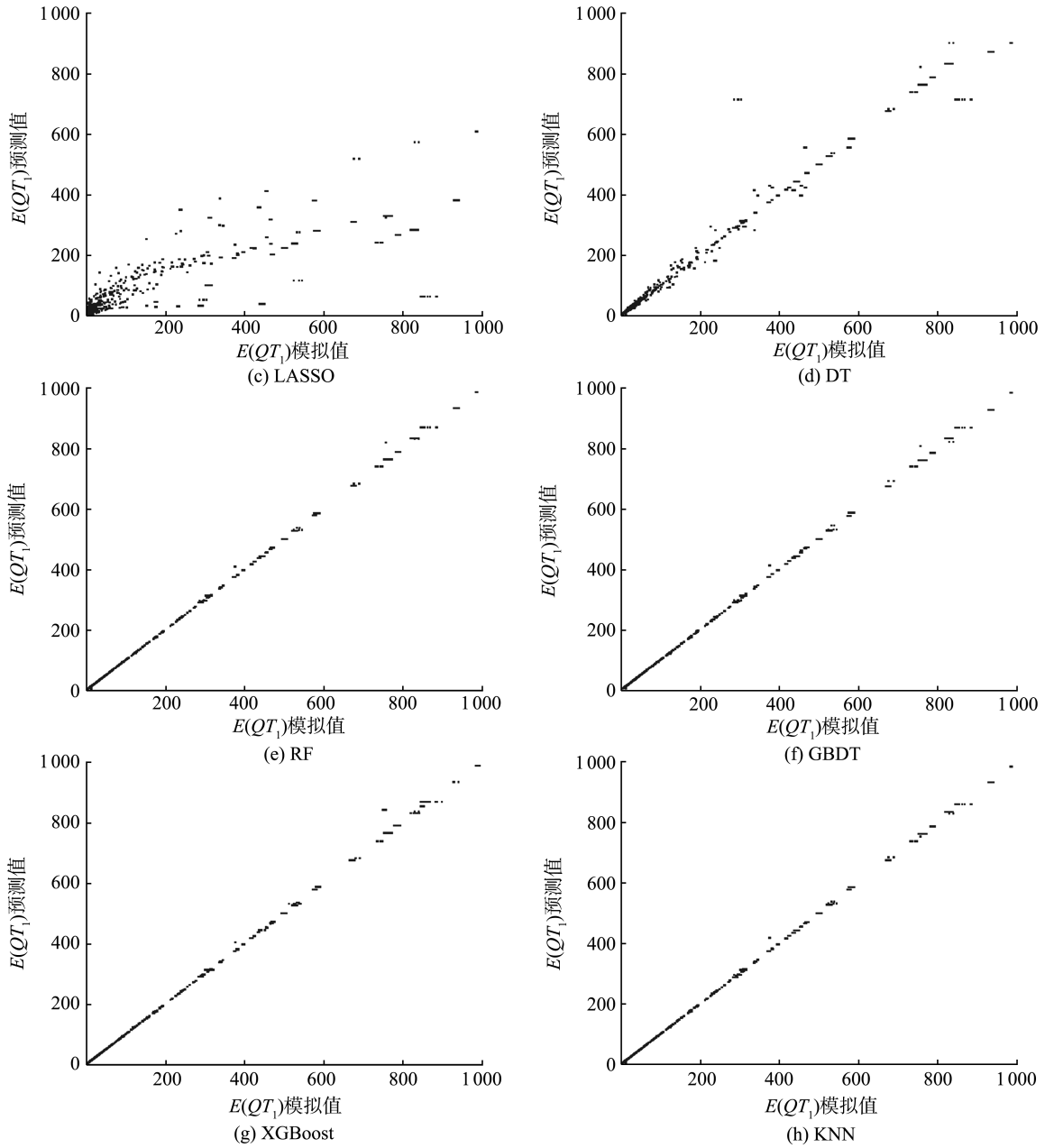


图 2 第一个站平均排队时间预测效果图

Fig.2 Prediction graph of the average queueing time of the first station

图 2 中,非线性回归模型 DT、RF、GBDT、XGBoost、KNN 的预测值与模拟值呈线性相关,预测效果较好。MLR、RR、LASSO 方法的预测结果呈现偏态,大部分情况下,预测值和模拟值误差较大,预测结果不可靠,不具有参考价值。

本文研究的 $M/G_1/N_1 \rightarrow /G_2/N_2$ 串联排队系统中,由于第二个站对第一个站的排队时间没有影响,因此第一个站可以看作是独立的 $M/G/N$ 排队系统。针对 $M/G/N$ 排队系统,郭亚亚等[19]对 David[20]、Hokstad[21]、Kimura[22]和 Sakasegawa[23]这 4 位学者的理论方法进行对比分析,发现 4 种理论方法的近似效果与 $C_{S_1}^2$ 和 ρ 有关。对于不同的 $C_{S_1}^2$ 和 ρ,本章选取测试集比例为 0.25,将 RF、GBDT、XGBoost 和 KNN 这 4 种机器学习方法的预测结果与 4 种理论方法的近似结果进行对比。当参数为 $E(S)=30$, $N=2$, $C_{S_1}^2=2$ 时,8 种方法的相对误差比较结果如表 3 所示。当参数为 $E(S)=30$, $N=2$, $C_{S_1}^2=0.9$ 时,8 种方法的相对误差比较结果如表 4 所示。

比较表 3 和表 4 的数值结果,可以发现平均排队时间是 ρ 和 $C_{S_1}^2$ 的增函数,4 种理论方法的误差随着 ρ 的增大逐渐减小,随着 $C_{S_1}^2$ 增大而增大,误差波动范围比较大,最小误差为 0.01%,最大误差为 106.99%;而 4 种机器学习方法的预测误差随着参数 ρ 和 $C_{S_1}^2$ 的变化波动较小,误差均小于 0.7%,整体预测效果较好。表 3

和表 4 的预测结果均验证了机器学习方法预测排队系统平均排队时间的可行性。

表 3　第一个站平均排队时间相对误差比较($C_{S_1}^2=2$)

Table 3　The comparison of relative error of the mean queuing time at the first station ($C_{S_1}^2=2$)　%

ρ	模拟值	David	Hokstad	Kimura	Sakasegawa	RF	GBDT	XGBoost	KNN
0.10	0.85	10.03	5.94	3.95	106.99	0.23	0.23	0.18	0.25
0.20	9.58	7.72	4.66	4.65	52.31	0.11	0.11	0.16	0.12
0.30	8.58	6.10	3.79	0.55	30.90	0.08	0.07	0.09	0.10
0.40	16.62	4.82	3.14	1.74	19.57	0.13	0.13	0.16	0.13
0.50	29.40		2.03	1.12	12.08	0.11	0.10	0.13	0.12
0.60	49.78	2.37	1.70	3.11	7.78	0.10	0.10	0.14	0.10
0.70	85.40	1.57	1.25	2.99	4.74	0.08	0.09	0.12	0.09
0.80	158.60	0.96	0.88	2.41	2.66	0.14	0.13	0.16	0.17
0.90	381.07	0.65	0.68	1.21	1.36	0.32	0.33	0.46	0.35
0.95	837.03	0.51	0.47	1.48	0.18	0.53	0.52	0.61	0.49

表 4　第一个站平均排队时间相对误差比较($C_{S_1}^2=0.9$)

Table 4　The comparison of relative error of the mean queuing time at the first station ($C_{S_1}^2=0.9$)　%

ρ	模拟值	David	Hokstad	Kimura	Sakasegawa	RF	GBDT	XGBoost	KNN
0.10	0.58	10.76	0.97	0.78	93.48	0.15	0.15	0.12	0.15
0.20	2.39	8.04	0.72	3.31	44.48	0.09	0.09	0.10	0.09
0.30	5.68	5.89	0.67	0.23	25.27	0.09	0.09	0.17	0.09
0.40	10.91	4.22	0.46	0.03	15.38	0.07	0.06	0.19	0.06
0.50	19.07		0.36	0.16	15.38	0.09	0.09	0.15	0.09
0.60	32.16	1.51	0.30	0.20	5.66	0.05	0.05	0.06	0.06
0.70	54.83	0.74	0.12	0.32	3.32	0.10	0.11	0.08	0.11
0.80	101.34	0.23	0.01	0.33	1.76	0.08	0.08	0.10	0.08
0.90	243.36	0.19	0.15	0.04	0.52	0.26	0.26	0.36	0.24
0.95	527.32	0.01	0.06	0.16	0.35	0.51	0.51	0.61	0.47

3.2　第二个站平均排队时间的预测

与第一个站的平均排队时间的预测相比,第二个站平均排队时间的预测需要考虑第一个站的参数和第二个站自身的参数,参数的增多会对预测结果产生影响。为了提高预测的准确率,先对第二个站平均排队时间进行归一化处理,再利用机器学习方法进行预测,最后得出第二个站平均等待时间的预测值。

将前面所涉及的数值模拟的参数 C_a^2、N_1、N_2、μ_1、μ_2、$C_{S_1}^2$、$C_{S_2}^2$、ρ 任意组合,一共得到 17 280 组数据,对数据进行筛选,去除 $E(QT_2)<0.1$ 的数据,获得有效的数据 13 847 组。

与第一个站平均排队时间的预测类似,针对每一种方法,都在同一组参数下根据不同的训练集、测试集比例进行了 8 次实验,训练集的比例分别为 0.95、0.9、0.85、0.8、0.75、0.7、0.6、0.5,对应的测试集比例为 0.05、0.1、0.15、0.2、0.25、0.3、0.4、0.5。将实验结果进行整理、对比,结果如表 5、6 所示。

表 5 中,MLR、RR、LASSO 方法的误差均在 30%左右,预测效果较差。表 6 中,RF、GBDT、XGBoost 算法的误差较为接近,误差范围为 1.4%~2.7%,其中,GBDT 算法的平均相对误差最小,误差范围为 1.42%~1.87%,误差波动范围小,表现出良好的预测效果。KNN 算法的误差范围为 3.99%~5.57%,预测效果次于 RF、GBDT、XGBoost 算法。5 种非线性回归模型中,DT 算法的误差最大。第二个站与第一个站的预测结果一致,本文选取的非线性回归模型适用于平均排队时间的预测。

表 5　线性回归模型预测 $E(QT_2)$ 的平均相对误差

Table 5　The mean relative error of predicting $E(QT_2)$ by linear regression model　%

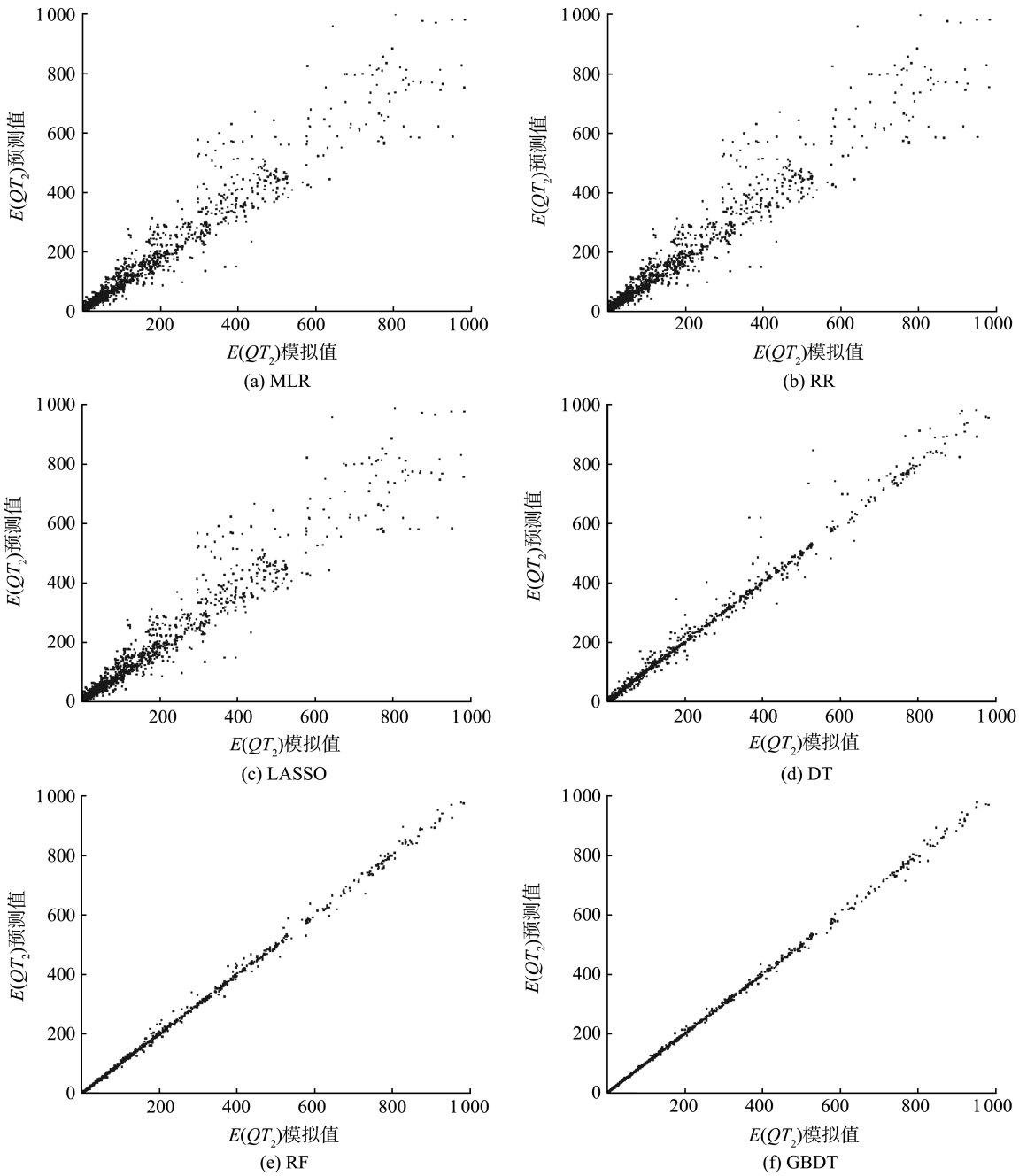
比例	MLR	RR	LASSO
0.05	32.52	32.49	32.23
0.10	28.97	28.94	28.71
0.15	28.53	28.49	28.28
0.20	28.54	28.51	28.30
0.25	28.37	28.34	28.13
0.30	29.33	29.26	29.08
0.40	29.61	29.56	29.38
0.50	30.28	30.21	30.05

表 6 非线性回归模型预测 $E(QT_2)$ 的平均相对误差

Table 6 The mean relative error of predicting $E(QT_2)$ by nonlinear method %

比例	DT	RF	GBDT	XGBoost	KNN
0.05	5.80	1.82	1.42	1.61	3.99
0.10	6.48	1.88	1.53	1.64	4.06
0.15	6.52	2.00	1.46	1.62	4.35
0.20	6.07	2.06	1.54	1.73	4.37
0.25	6.19	2.09	1.52	1.71	4.59
0.30	6.34	2.17	1.50	1.89	4.78
0.40	6.91	2.43	1.67	2.22	5.07
0.50	7.78	2.70	1.87	2.37	5.57

选取测试集比例为 0.25 时,将 8 种机器学习方法求出的第 2 个站平均排队时间的预测值和模拟值对比,结果如图 3 所示。



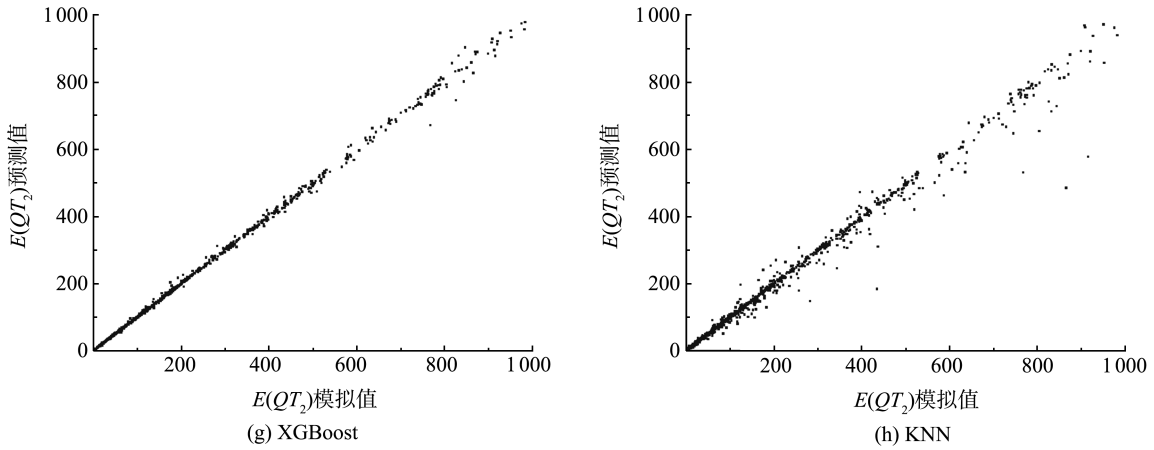


图 3　第二个站平均排队时间预测效果

Fig.3　Prediction graph of the average queueing time of the second station

由图 3 可以明显看出,RF、XGBoost、GBDT 算法的预测值与模拟值呈线性相关,预测效果较好。KNN、DT 算法的预测值与模拟值具有线性趋势,但仍有部分预测值与模拟值偏差较大。MLR、RR、LASSO 方法的预测结果未呈现出线性关系,大部分情况下,预测值误差较大,预测结果不具有可靠性,不具有参考价值。

综上所述,针对本文所考虑的数据集,MLR、RR、LASSO 这 3 种模型不适用于多服务器串联排队系统平均排队时间的预测。非线性回归模型预测效果较好,从数值结果分析,RF、GBDT、XGBoost、KNN 算法的预测误差较为接近,预测效果较好;相较于 RF、GBDT、XGBoost 和 KNN 算法,DT 算法的误差偏大。从线性关系分析,RF、XGBoost、GBDT 算法的线性相关性比 DT、KNN 算法的线性关系明显。从算法的性质分析,RF 算法本身具有随机性,运行速度快和预测准确率高;XGBoost 算法自身具有良好的性质,预测结果误差低,预测结果可靠性高;GBDT 算法本身包容性很强,但是损失函数的选择不如 XGBoost 算法灵活,计算效率稍逊于 XGBoost 算法;KNN、DT 算法预测结果较好,但是算法本身具有局限性。综上,RF、XGBoost、GBDT 这 3 种算法的预测结果较好,并且对数据类型包容性强,具有一定的参考价值,可用于串联排队系统的预测,也可用于排队网络的研究。

4　结论

本文利用机器学习方法,预测了多服务器串联排队系统的平均排队时间,并对机器学习方法进行对比,从数值结果、线性关系和算法自身性质进行分析,选取出可靠的机器学习预测方法。通过数值实验结果,验证了 RF、XGBoost、GBDT 方法可以作为分析多服务器串联排队网络的有效手段,该结果可以推广到复杂排队网络的研究中。

本文主要利用机器学习方法对多服务器串联排队系统平均排队时间进行预测,但并未分析系统各参数对预测结果的影响程度,以及各参数和平均排队时间的相关性。未来研究可以进一步利用机器学习的方法深入分析参数对平均排队时间的影响,同时,也可以将目前可行的机器学习方法进行改进,与神经网络方法相结合,降低排队系统性能指标的预测误差,提升预测精度。

参考文献:

[1] ZHU Yixin. Tandem queue with group arrivals and no intermediate buffer[J]. Queueing Systems, 1994, 17(3/4):403-412.

[2] GÓMEZ-CORRAL A. A tandem queue with blocking and markovian arrival process[J]. Queueing Systems, 2002, 41(4): 343-370.

[3] VAN HOUD T B, ALFA A S. Response time in a tandem queue with blocking, Markovian arrivals and phase-type services [J]. Operations Research Letters, 2005, 33(4):373-381.

[4] LIAN Zhaotong, LIU Liming. A tandem network with MAP inputs[J]. Operations Research Letters: A Journal of the Operations Research Society of America, 2008, 36(2):189-195.

[5] WU Kan, ZHAO Ning. Dependence among single stations in series and its applications in productivity improvement[J]. Euro-

pean Journal of Operational Research, 2015, 247(1):245-258.

[6] 吴登磊,赵宁,刘文奇. 基于指标比对串联排队系统平均排队时间的近似方法[J]. 南京航空航天大学学报, 2020, 52(4):644-649.
WU Denglei, ZHAO Ning, LIU Wenqi. Approximation method for average queuing time of tandem queuing system based on index comparison[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2020, 52(4):644-649.

[7] 侯佳辰,赵宁,刘文奇,等. 串联排队系统平均等待时间的近似分析[J]. 山西大学学报(自然科学版), 2022, 45(1):41-49.
HOU Jiachen, ZHAO Ning, LIU Wenqi, et al. Approximate analysis of average waiting time of tandem queuing system[J]. Journal of Shanxi University (Natural Science Edition), 2022, 45(1):41-49.

[8] KIM C S, KLIMENOK V, TARAMIN O. A tandem retrial queueing system with two Markovian flows and reservation of channels[J]. Computers & Operations Research, 2010, 37:1238-1246.

[9] DUDIN A, DUDIN S, DUDINA O. Tandem queueing system MAP| M| N| K-N→●| M| R|∞ with impatient customers as a model of remote technical support[C]//2012 2nd Baltic Congress on Future Internet Communications. Vilnius: IEEE, 2012: 134-139.

[10] KIM C, DUDIN A, DUDINA O, et al. Tandem queueing system with infinite and finite intermediate buffers and generalized phase-type service time distribution[J]. European Journal of Operational Research, 2014, 235(1):170-179.

[11] LAL T S S, KRISHNAMOORTHY A, JOSHUA V C. A multiserver tandem queue with a specialist server operating with a vacation strategy[J]. Automation and Remote Control, 2019, 81(4):760-773.

[12] BANU P, RAJENDRAN P. Performance measures of parallel tandem open queueing network[J]. International Journal of Pervasive Computing and Communications, 2021, 17(1):37-48.

[13] SAGIR M, SAGLAM V. Optimization and analysis of a tandem queueing system with parallel channel at second station[J]. Communication in Statistics-Theory and Methods, 2021, 51(2):1-14.

[14] KUMAR B K, SANKAR R, KRISHNAN R N, et al. Performance analysis of multi-processor two-stage tandem call center retrial queues with non-reliable processors[J]. Methodology and Computing in Applied Probability, 2022, 24(1):95-142.

[15] EFROSININ D, STEPANOVA N. Estimation of the optimal threshold policy in a queue with heterogeneous servers using a heuristic solution and artificial neural networks[J]. Mathematics, 2021, 9(11):1267.

[16] TAN B, KHAYYATI S. Supervised learning based approximation method for single-server open queuing networks with correlated interarrival and service times[J]. International Journal of Production Research, 2022, 60(22):6822-6847.

[17] KHAYYATI S, TAN B. Supervised-learning-based approximation method for multi-server queueing networks under different service disciplines with correlated interarrival and service times[J]. International Journal of Production Research, 2022, 60(17):5176-5200.

[18] 刘顺祥. 从零开始学 Python 数据分析与挖掘[M]. 北京:清华大学出版社,2018.
LIU Shunxiang. Learning Python data analysis and mining from scratch[M]. Beijing: Tsinghua University Press, 2018.

[19] 郭亚亚, 赵宁, 戴琳, 等. 对 M/G/m 排队系统平均等待时间估计方法的数值比较[J]. 江苏科技大学学报(自然科学版), 2017, 31(2):252-258.
GUO Yaya, ZHAO Ning, DAI Lin, et al. Numerical comparison of methods for estimating average waiting time in M/G/m queuing system[J]. Journal of Jiangsu University of Science and Technology (Natural Science Edition), 2017, 31(2):252-258.

[20] DAVID D Y. Refining the diffusion approximation for the M/G/m queue[J]. Operations Research, 1985, 33(6):1266-1277.

[21] HOKSTAD P. Approximations for the M/G/m queue[J]. Operations Research, 1978, 26(3):510-523.

[22] KIMURA T. Approximations for the delay probability in the M/G/s queue[J]. Mathematical and Computer Modelling, 1995, 22(10/11/12):157-165.

[23] SAKASEGAWA H. An approximation formula $L_q=\alpha.\rho^3/(1-\rho)$[J]. Annals of the Institute of Statistical Mathematics, 1977, 29(1):67-75.

(编辑:于善清)