控制与决策
Control and Decision

基于时变马尔可夫链的在线医疗服务医生排班研究

马颢洲, 刘冉

引用本文:
马颢洲, 刘冉. 基于时变马尔可夫链的在线医疗服务医生排班研究[J]. 控制与决策, 2025, 40(4): 1172-1180.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2024.0327

基于时变马尔可夫链的在线医疗服务医生排班研究

马颢洲，刘 冉†

(上海交通大学 工业工程与管理系，上海 200240)

摘 要: 伴随线上医疗不断发展, 医院在线上线下联合医疗服务的模式中面临着对线上服务医生进行排班优化决策的问题, 其主要挑战在于时变的患者需求和线上医疗特殊的服务模式. 针对此决策问题, 首先将线上医疗服务系统建模为资源共享队列, 采用时变马尔可夫链和均匀化方法对患者逗留时间、队列长度和医生加班时间进行建模和分析评估; 然后基于以上系统评估方法, 提出变邻域搜索的启发式算法对医生排班问题进行求解; 最后基于合作医院的实际数据开展数值实验分析, 以验证基于时变马尔可夫链建模的准确性. 结果表明, 所提出算法可以得到相对医院实际方案更好的排班结果, 从而可以更加合理地安排医生工作时间, 减少患者逗留时间, 控制系统中的患者数量, 并具有优良的鲁棒性. 所做的研究对于完善我国线上医疗服务系统的运作管理具有实际意义.

关键词: 线上医疗服务；时变马尔可夫链；资源共享队列；均匀化方法；医生排班；变邻域搜索算法

中图分类号: F224 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2024.0327

引用格式: 马颢洲, 刘冉. 基于时变马尔可夫链的在线医疗服务医生排班研究 [J]. 控制与决策, 2025, 40(4): 1172-1180.

A study of physician scheduling for online medical service system based on time-varying Markov chains

MA Hao-zhou，*LIU Ran*†

(Department of Industrial Engineering & Management, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: The continuous development of online medical has confronted hospitals with the problem of how to allocate physician resources in joint online and offline medical service systems, whose main challenges lie in the time-varying demand and the different service models of online medical. To address the physician scheduling problem in the online medical service system, the online medical service system is modeled as a processor-sharing queue firstly, and the time-varying Markov chain and uniformization methods are used to analyze and evaluate the patient sojourn time, queue length, and physician overtime time. Then, based on the uniformization method, a heuristic algorithm for variable-neighborhood search is proposed. Finally, numerical experimental analysis is carried out based on the actual data of the cooperative hospital, which verifies the accuracy of the time-varying Markov chain based modeling, and proves that the proposed algorithm can obtain better scheduling results relative to the actual scheduling of the hospital, so that it can more reasonably arrange the working time of the physicians, reduce the sojourn time of the patients, and control the number of patients in the system, and has excellent robustness. The study has practical significance for improving the operation and management of online medical service systems in China.

Keywords: online medical service； time-varying Markov chain； processor-sharing queue； uniformization methods；physician scheduling；variable neighborhood search algorithms

0 引 言

目前我国在线医疗服务正快速发展并逐渐成熟, 形成了与传统线下问诊并行的医疗服务模式. 2020 ～ 2022 年, 我国在线问诊需求快速增长, 截至 2022 年 6 月, 线上医疗用户规模达 3 亿[1]. 目前全国大部分三甲医院都在传统线下医疗服务基础上, 引入线上医疗服务, 并取得了良好的效果. 对于患者而言, 线上医疗服务可以避免往返的时间和费用, 减少在医院的等待时间, 降低流行病爆发期间被院内感染的风险. 对医院而言, 在线医疗服务可以覆盖更多的患

收稿日期: 2024-03-29；**录用日期:** 2024-08-29.
基金项目: 国家自然科学基金面上项目 (72371161).
责任编委: 李登峰.
†通信作者. E-mail: liuran2009@sjtu.edu.cn.
本文附带电子附录文件,可登录本刊官网该文“资源附件”区自行下载阅览.

者人群, 减轻线下医疗就诊的压力, 缓解院内拥挤, 改善患者医疗体验.

医生是医院重要医疗资源, 其工作排班是否合理直接影响医疗服务的质量. 但是从医院角度, 在同时提供线下和线上服务的模式下, 需要安排医生为线下和线上的患者同时提供服务, 如何在线上与线下服务之间科学合理分配有限的医生资源, 是医院面临的客观挑战, 其难点总结为以下几个方面.

首先, 线上医疗服务处于不确定的环境, 多维不确定因素给医生排班产生很大困难. 目前我国很多医院的线上医疗服务推行了预约制度, 但是患者迟到、提前到达、爽约等现象, 造成患者线上到达的不确定性, 使线上医疗服务系统面临随机到达的患者需求. 同时, 线上医疗服务中, 医生的服务速率也是不确定的, 对每个患者的服务时长也是不确定的. 并且, 根据对合作医院的调研发现, 一名医生可以同时为多名患者提供服务, 伴随同时服务患者数量的变化, 医生的服务速率也有所不同. 这些复杂的多维随机因素对相关医生排班带来了很大的困难.

进一步, 线上和线下医疗服务在运作方式等方面存在差异, 这也对相关医生排班产生了困难. 经典线下医疗的医患之间一般采用“一对一”的服务方式, 一名患者在诊室内就诊时, 其他患者在诊室外进行等待; 而典型的线上医疗服务中, 医患之间一般采用“一对多”的服务方式, 即医生在等待一名患者准备其相关材料时 (例如等待患者上传病例、等待患者回复等), 医生还与其他患者交流以提供服务, 从而同时与多名患者进行沟通并加以诊断. 因此, 对线上医疗服务的医生排班管理必须和其“一对多”等重要特点相适应, 才能实现更好的服务效果. 但是如何在线上新服务模式下实现科学合理的医生排班, 满足有效缩短患者等待时间、控制患者在线排队队长等目标, 目前对医院而言具有挑战.

面对传统的线下医疗服务系统, 医生以及护士排班问题 (简称“医护排班问题”) 和预约调度是两个重要研究方面. 其中, 医护排班问题涉及设计医生护士排班以满足系统要求, 同时优化各种指标, 如患者等待时间和服务水平. 目前, 大量模型和算法已被学者开发并用于研究医生或护士人员配备和排班的各个方面[2-3]. 其中, 很多文献使用精确算法求解排班问题[4-5]. 此外, 各类启发式算法也在求解医护排班问题中被使用. 其中, 诸多文献采用禁忌搜索算法求解排班问题. Wang 等[6] 针对带回流的急诊服务系统建立了离散时间的解析模型, 用于近似患者的等待时间, 并采用禁忌搜索算法求解相关医生排班问题. Yang 等[7] 基于马尔可夫链对急诊服务系统解析建模, 提出计算患者等待时间的解析方法, 并基于禁忌搜索算法对急诊的医生排班问题求解. 变邻域搜索算法 (variable neighborhood search, VNS) 在解决排班问题中的应用也十分广泛. Liu 等[8] 在线下医生排班中考虑到急诊患者到达的随机性和时变的到达率, 利用稳态排队模型和马尔可夫链等技术计算了时变患者队列长度和患者平均等待时间, 建立了医生调度模型并设计了变邻域搜索算法以有效求解问题. Lan 等[9] 提出了一种带有自适应启发式的变邻域搜索算法, 嵌入动态规划算法, 解决了综合医生计划和排班问题. 需要提及的是, 排班问题不仅在医疗服务中有较多研究, 在其他相关场景下, 例如呼叫中心等, 也有很多类似研究[10].

虽然传统线下医疗服务运作管理问题, 例如医护排班以及预约调度[11] 等已经得到重视, 但是针对线上医疗服务中的医护排班问题研究才刚刚起步, 仅有 Ji 等[12] 针对远程医疗中医生和患者可能临时离开的情况, 建立了以最小化医生工作成本和患者得不到服务的惩罚成本为目标的两阶段医生排班和调度模型, 并使用求解器求解该模型. Yu 等[13] 结合线上线下医疗服务, 以满足患者护理需求的同时最大限度地降低运营成本为目标, 解决护士分配问题. 该研究设计了一种基于列生成的启发式算法, 以确定诊所选择、患者分配和巡诊护士路径规划问题的联合决策规则. 相对而言, 针对线上医疗服务中其他方面的研究较为充分. 例如, 在线医疗预约调度问题得到一定关注[14-15], 也有一些研究关注在线医疗服务策略和定价策略等方面[16-18].

值得注意的是, 如上文中所述, 线上医疗服务与传统的线下服务不同, 医生可以同时为多位患者提供服务, 而不是采用一对一的诊断和治疗方法. 这是研究线上医疗服务调度系统时需要考虑的一个关键因素. 在这方面, 缺乏专门研究医疗系统的相关研究文献, 仅在针对商业平台的线上客户服务系统, 部分文献考虑了一个客服代表同时服务多个客户的模式, 研究了系统建模等问题[19].

从以上文献分析可以看出, 虽然线下医疗服务的医护排班问题已经得到了充分研究, 线上医疗服务的服务策略和定价策略方面的研究也得到了重视, 但是面向线上服务的医护排班问题研究目前几乎尚未开始. 本文以线上医疗服务系统的医生排班问题为研究对象, 首先基于时变马尔可夫链的均匀化方法对动态的线上医疗服务患者排队系统进行建模. 在此基础上, 由于 VNS 在医疗服务系统运作优化中

已经得到成功应用[8-9], 本文设计一种 VNS, 将均匀化方法应用到解评估中, 对系统里重要的医生资源进行排班优化, 希望通过科学的排班方案适应时变随机到达的患者需求, 有效控制患者排队队长, 减少患者在系统中的逗留时间, 并使得医生的工作强度得到合理地控制. 研究对于提升线上医疗的服务水平, 缓解医生的工作负荷, 具有积极现实意义.

1 问题描述

本文合作医院的线上医疗服务过程描述如下. 设定医院共有 N 名医生, 每天医生的工作 (问诊) 时间可以分为 T 个时间段, 每个时间段的时间长度相等, 记为Δ. 首先, 就线上医疗服务而言, 患者在每个时段 t 内随机到达线上问诊系统, 假设 t 时段内患者按照速率为λ_t的泊松过程到达. 在进入线上医疗服务系统后, 患者按照先到达先服务的原则接受服务, 形成患者的线上排队队列. 如前文所述, 由于线上问诊的特点, 每个医生可以同时为多个患者提供服务, 但每个医生“同时”服务患者数量具有上限 K. 假定所有医生的服务能力相同, 每个医生服务一个患者的服务速率服从独立的指数分布, 平均服务时间为 $1/\mu(k)$(其中 k 代表同时服务的患者数量). 与线下医疗服务不同, 医生线上服务一名患者的服务时间会随着同时服务的患者数量增加而增加, 即当 k 增大时, $\mu(k)$随之降低. 这是因为医生同时服务的患者数量更多, 意味着医生将与多名患者进行线上沟通与诊疗, 服务一名患者的时间随之增加. 另一方面, 当医生同时服务更多的患者时, 医生的总服务能力 $(k \times \mu(k))$ 将会增加, 这是由于医生同时服务多名患者时, 可以在等待一名患者回复信息时与其他患者进行沟通与诊疗, 更充分地分配自己的时间并减少空闲时间.

处于线上排队的患者, 将被分配给服务患者人数最少的医生. 如果所有正在上班的医生都到达服务患者数目的上限, 则排队患者必须继续等待. 服务结束后, 患者将离开系统. 本文设定患者不会在等待中途离开系统, 一定在接受医疗服务后才离开.

各个医院的医生工作都要遵循一定排班规则. 根据对合作医院的调研, 归纳其医生排班需要遵循的假设和约束如下: 1) 任何时段都必须至少有一名医生在线上工作. 2) 每位医生每天最多有两个工作班次, 其中线下工作班次每天最多一个, 即每位医生可以有两个线上班次, 或一个线上班次和一个线下班次. 3) 医生的一个线上班次时段长度不大于医院规定的最长工作时长 UBD(upper bound of working duration), 不小于规定的最短工作时长 LBD (lower bound of working duration). 4) 每名医生两次班次之间的间隔不少于 R 个时段, 以保证医生有足够的休息. 5) 医生在每个时段的开始交接班, 即医生的上班和下班均发生在每个时段的开始, 每个时段内正在上班的医生的数量固定.

由于本文聚焦医生线上工作的排班问题, 设定每位医生的线下班次是已知的. 需要注意的是, 不同医院对医生交接班时, 如何处理下班医生手头的患者规定各不相同. 本文设定当一名线上工作医生下班时, 如果有一名新医生接班, 则接手其正在服务的患者; 反之, 如果没有新的上班医生接手, 则该下班医生正在服务的患者将返回等待队列.

医生排班除了需要遵守以上排班规则约束外, 出于提升对患者服务水平的出发点, 医院希望控制线上服务系统中的患者人数, 以免患者进行过多地等待, 即对系统的“患者排队队长”有一定的约束限制. 已有研究大部分在控制医疗服务系统患者队长方面, 均只考虑队长的期望值, 即一般设定为患者队长低于一个设定的阈值. 但由于排队队长是一个随机量, 仅控制队长的期望值并不严谨. 本文问题定义提出以下重要约束, 即对系统中患者的人数提出上限阈值θ, 并设定队长的约束为每个时段末排队队长低于设定阈值的概率应大于 95 %, 通过如此约束实现对系统服务水平、患者等待时间的有效控制.

对线上医生排班的优化目标设定为 3 部分: 医生工作成本、医生加班成本和患者等待服务的成本. 相关指标也在医疗服务管理的文献中被使用[20-21]. 以 c_t表示 t 时段上班的医生数量, 则全部医生工作时间为$\sum_{t=1}^{T}(\Delta \times c_t)$(其中$\Delta$表示一个时间段的时间长度). 在一天的最后一个时段结束时, 如果还有患者没有被医生服务完成, 则此医生加班直到系统中的所有患者完成服务, 本文以 OT 表示医生的加班时间. 对于患者等待服务的成本, 线上服务与经典线下服务模式不同: 线上服务系统中一个医生可以同时服务多个患者, 因此一名患者在服务开始后并不完全占用医生资源, 即其被服务的过程中可能也存在等待, 所以仅考虑经典客户初次被服务之前的等待时间并不能完全体现服务水平. 因此, 本文关注患者在线上医疗服务系统中的完整逗留时间, 此逗留时间定义为: 一名患者从进入线上医疗服务系统至离开系统的时间.

综上, 本文医生排班的优化目标定义如下:

$$\min\Big(\sum_{t=1}^{T} W_t + \alpha \times \mathrm{OT} + \beta \sum_{t=1}^{T}(\Delta \times c_t)\Big). \quad (1)$$

其中: W_t为t时段患者总逗留时间; OT 为医生加班时间; α和β为参数, 分别为医生加班时间和工作时段数的权重.

2　系统建模

马尔可夫链建模方法在类似不确定条件下的随机过程建模中应用广泛[22], 本文将线上排队系统建模为连续时间马尔可夫链, 并采用均匀化方法解析计算系统性能指标.

2.1　系统时变马尔可夫链模型

对线上医疗服务系统, 定义如下状态变量以基于马尔可夫链对系统进行建模. $q_0(t)$: 表示在t时段的开始, 系统中等待被服务的患者数量; $q_i(t)$: 表示在t时段初, 医生i正在同时服务的患者数量, $i = 1,2,\ldots,N$. 值得注意的是, 若医生i在t时段不工作, 则$q_i(t)=0$, 并且需要满足约束$q_i(t) \leqslant K, \forall i = 1,2,\ldots,N$. 基于以上定义, 系统状态变量$s(t)$定义如下: $s(t)=\{q_0(t), q_1(t), \ldots, q_N(t), t \geqslant 0\}$. 对于每个时段$t$, 基于以上问题描述, 患者到达率和医生最大总服务速率是不变的. 假设$v_{s,s'}$表示系统从状态s转移到s'的转移速率, 系统状态转移的情况可以描述如下.

1) 患者到达. 若正在上班的所有医生正在服务的患者人数都为上限K, 则该患者将在等待队列进行等待, 那么转移后的系统状态为$s'_0 = (q_0+1, q_1, \ldots, q_N)$. 否则, 在正在上班的医生中选择当前正在服务患者数量最少的医生j对该患者进行服务, 则转移后的系统状态为$s'_j = (q_0, q_1, \ldots, q_j+1, \ldots, q_N)$. 在这种情况下, 由于患者到达系统的速率为$\lambda_t$, 系统状态的转移速率为$v_{s,s'_j} = v_{s,s'_0} = \lambda_t$.

2) 患者服务完成. 当医生j服务完一名患者, 该患者将会直接离开系统. 此时, 如果$q_0 > 0$, 即系统中有患者等待被服务, 则将队列中的第 1 个患者分配给医生j. 因为在有患者等待的情况下, 所有正在上班的医生服务的患者数量必须达到上限K, 所以在医生j服务完一名患者后, 只能将等待的患者分配给医生j. 转移后的系统状态为$s''_0 = (q_0-1, q_1, \ldots, q_N)$. 否则, 如果$q_0 = 0$, 则系统中没有患者等待, 转移后的系统状态为$s''_j = (q_0, q_1, \ldots, q_j-1, \ldots, q_N)$. 在这种情况下, 由于医生$j$的正在服务的患者数目为$q_j$, 服务单个患者的服务速率为$\mu(q_j)$, 该医生的总服务速率为$q_j \times \mu(q_j)$. 基于此, 对于医生$j$服务完一名患者的情况, 系统状态的转移速率为$v_{s,s''_j} = v_{s,s''_0} = q_j \times \mu(q_j)$. 由于系统中共有$N$名医生, 故总转移速率为$v_{s,s'} = \sum_{j=1}^{N} q_j \times \mu(q_j)$. 综上, 系统转移速率矩阵可表示为

$$v_{s,s'} = \begin{cases} \lambda_t, \ s' = s'_0 \text{ or } s' = s'_j; \\ q_j \times \mu_{q_j}, \ s' = s''_0 \text{ or } s' = s''_j; \\ 0, \ \text{otherwise.} \end{cases} \quad (2)$$

2.2　均匀化方法

2.2.1　状态分布和转移矩阵

均匀化方法用于将有限状态的连续时间马尔科夫链转化为离散时间马尔科夫链. 对于每个时段t, 定义均匀化转移速率为$\gamma_t = \lambda_t + c_t \times K \times \mu(K)$. 为了保证每个状态的转出速率$\gamma_t$相同, 需要引入虚拟的状态转移, 其速率为 $v_{s,s} = \gamma_t - \lambda_t - \sum_{j=1}^{N} q_j \times \mu(q_j)$. 定义$s(t)$为$t$时段初系统状态, 令$\pi_t$为向量$[\pi_{s,t}]_{s\in S}$, 其中$\pi_{s,t} = P(s(t)=s)$. 假设$t$时段的状态转移次数为$N_t$, 根据问题定义, 在每个时段初存在医生上班和医生下班, 本文将这一过程定义为医生交接班. 令t^+和t^-表示t时段初医生交接班后和交接班前的瞬时时刻. 因此, t时段结束时系统状态的概率分布表示为

$$\pi_{(t+1)^-} = \sum_{n=0}^{\infty} \pi_t P^n \times P(N_t = n) = \sum_{n=0}^{\infty} \pi_t P^n \times \frac{(\gamma_t \times \Delta)^n \mathrm{e}^{-\gamma_t \times \Delta}}{n!}, \quad (3)$$

其中$P = [p_{s,s'}]_{(s,s')\in S^2}$是每个均匀化事件的转移概率矩阵, 定义如下:

$$p_{s,s'} = \begin{cases} \lambda_t/\gamma_t, \ s' = s'_0 \text{ or } s' = s'_j; \\ q_j \times \mu_{q_j}/\gamma_t, \ s' = s''_0 \text{ or } s' = s''_j; \\ v_{s,s}/\gamma_t, \ s' = s; \\ 0, \ \text{otherwise.} \end{cases} \quad (4)$$

本文设定医生交接班时, 正在被下一时段下班的医生服务的患者, 会被退回系统等待队列. 这造成系统状态会出现瞬时转换, 即$\pi_{t^+} = \pi_{t^-} \times P'$, 其中$P' = \left[p'_{s,s'}\right]_{(s,s')\in S^2}$表示医生交接班时的瞬时转移矩阵. 对于$t$时段末的每个系统状态, 正在被将要下班医生服务的患者将返回等待队列. 医生交接班后, 将根据医生的空闲状态分配候诊队列中的患者, 直到候诊队列为空或所有正在上班的医生的服务能力达到上限. 分配规则与患者到达时的分配规则一致. 以$s(t^+)$表示瞬时转换后的系统状态, $s(t^-)$表示瞬时转换前的系统状态, 则矩阵P'定义如下:

$$p'_{s,s'}=\begin{cases}1,\ s=s(t^-)\ \text{and}\ s'=s(t^+);\\0,\ \text{otherwise.}\end{cases}\tag{5}$$

2.2.2 患者逗留时间计算

本节提出患者在系统中的逗留时间的计算方法. t 时段患者在系统中总逗留时间期望计算公式为

$$W_t=\sum_{n=0}^{\infty}(\mathrm{E}[\mathrm{WT}|N_t=n]\times P(N_t=n))=\sum_{n=0}^{\infty}\Big(\mathrm{E}[\mathrm{WT}|N_t=n]\times\frac{(\gamma_t\times\Delta)^n\mathrm{e}^{-\gamma_t\times\Delta}}{n!}\Big).\tag{6}$$

其中: W_t为患者在 t 时段的随机总停留时间, Δ为一个时间段的时间长度. 且有

$$\mathrm{E}[\mathrm{WT}|N_t=n]=\sum_{k=0}^{n}\sum_{s\in S}\pi(k,s)\times q_{to}\times\frac{\Delta}{n+1}.\tag{7}$$

其中: $q_{to}=q_0+q_1+\ldots+q_N$, $\pi(k,s)$表示在发生 k次均匀化事件后系统处于状态 s 的概率, 向量 $\pi(k)$服从$\pi(0)=\pi_{t^+}$, $\pi(k+1)=\pi(k)P$. 因此可以得到

$$W_t=\sum_{n=0}^{\infty}\Big(\sum_{k=0}^{n}\sum_{s\in S}\pi(k,s)\times q_{to}\times\frac{\Delta}{n+1}\times\psi\Big)=\sum_{k=0}^{n}\sum_{s\in S}\pi(k,s)\times q_{to}\times B_k,\tag{8}$$

其中$\psi=\dfrac{(\gamma_t\times\Delta)^n\mathrm{e}^{-\gamma_t\times\Delta}}{n!}$. 并且$B_k=\displaystyle\sum_{n=k}^{\infty}\frac{\Delta}{n+1}\times\frac{(\gamma_t\times\Delta)^n\mathrm{e}^{-\gamma_t\times\Delta}}{n!}$, 满足 $B_{k+1}=B_k-{\gamma_t}^{-1}\times\dfrac{(\gamma_t\times\Delta)^{k+1}\mathrm{e}^{-\gamma_t\times\Delta}}{(k+1)!}$, $B_0=1-{\gamma_t}^{-1}\times\mathrm{e}^{-\gamma_t\times\Delta}$. 综上可计算得到患者的逗留时间.

2.2.3 医生加班时间计算

在线医疗服务系统的最后一个时段结束后, 系统关闭 (停止挂号), 医生一般服务完滞留的患者 (加班) 再下班. 本节将计算医生加班时间的期望值. 医生加班时间的计算过程可分为两个阶段. 在第 1 阶段, 对于每个系统状态, 根据分配原则将仍在系统中的患者分配给正在工作的医生, 然后计算每位医生的加班时间总和. 第 2 阶段, 计算T时段末每个系统状态的概率与加班时间的乘积并求和, 得出加班时间期望. 总加班时间期望计算如下:

$$\mathrm{OT}=\sum_{s\in S}\mathrm{E}(\mathrm{OT}|s(T+1)=s)\times P(s(T+1)=s),\tag{9}$$

其中$\mathrm{E}(\mathrm{OT}\,|s(T+1)=s)$为系统处于状态 s 时的期望加班时间, 且有

$$\mathrm{E}(\mathrm{OT}|s(T+1)=s)=\sum_{j=1}^{N}\mathrm{E}(n_j|s(T+1)=s),\tag{10}$$

$\mathrm{E}(n_j\,|s(T+1)=s)$为系统处于状态 s 时, 医生 j 的加班时间期望. 需要注意的是, 由于所有医生的能力相同, 对每个患者的服务持续时间是互相独立的指数分布, 服务速率为$\mu(k)$. 根据指数分布的特性, 一名医生为n名患者服务的期望时间为

$$\mathrm{E}(n)=\begin{cases}1/\mu(1)+1/\mu(2)+\ldots+1/\mu(n),\ n\leqslant K;\\1/\mu(1)+1/\mu(2)+\ldots+1/\mu(K-1)+\\1/\mu(K)\times(n-K+1),\ n>K.\end{cases}\tag{11}$$

其中 n 为整数且$k>1$.

3 算法设计

为了解决在线医疗服务的医生排班问题, 基于第 2 节中提出的时变马尔可夫链系统建模方法以及均匀化计算技术, 设计一种 VNS 算法.

3.1 算法基本框架

VNS 算法首先生成一个初始解$\mathcal{S}^0$, 并定义一组邻域集合$H_l(l=1,2,\ldots,l_{\max})$, 设定算法迭代的当前解$\mathcal{S}^c\leftarrow\mathcal{S}^0$. 算法运行由多次迭代组成, 其中每一次迭代描述如下. 首先, 设置$l=1$, 通过从当前解$\mathcal{S}^c$的第 1 个邻域集H_1执行 shaking 操作生成解$\mathcal{S}^d$; 然后应用局部搜索优化获得最好邻域解$\mathcal{S}^n$. 如果该最好邻域解优于当前解, 则算法设置$\mathcal{S}^c\leftarrow\mathcal{S}^n$并设定$l=1$, 重新开始对解$\mathcal{S}^c$进行搜索. 否则, 算法将切换到$l=l+1$, 以下一个邻域继续执行 shaking 操作, 并继续优化当前解$\mathcal{S}^c$. 算法重复上述过程, 直到 $l>l_{\max}$, 完成一次迭代. 算法结束一次迭代后, 开始下一次迭代. 当迭代次数达到预定次数时, VNS 终止, 输出全部迭代搜索过程中产生的最高质量的解.

3.2 初 始 解

VNS 初始解生成过程如下. 首先, 选择一名医生从第 1 时段开始排班, 其班次长度是满足工作时间和排班约束条件的最大长度; 然后, 从上一个班次最后一个时段的下一个时段, 选择另一位医生开始排班, 其班次长度同样是满足工作时间和排班约束条件的最大长度. 如此反复, 直到满足所有时段至少有一名医生上班的约束条件为止; 最后, 在剩余可排班医生中按医生编号顺序安排班次. 班次设置如下: 班次时长先设置为最长时长, 班次开始时段从第 1 时段开始逐时段搜索, 若满足排班约束条件则安排该班次给这名医生, 否则将班次时长减少一个时段, 重

新从第 1 时段开始搜索, 直至班次时长为最短时长, 若该医生没有符合条件的班次可排班, 则进行下一名医生的排班.

3.3 shaking 操作

在 VNS 中, shaking 操作可以实现邻域间的跳跃, 避免算法搜索陷入局部最优. 本文算法定义了 $l_{\max}$ 个不同的邻域, 第 l 个邻域称为 $H_l(l=1,2,\ldots,l_{\max})$. 对于当前解, 邻域 H_l 将从解中随机选择 l 个医生班次进行随机生成 (随机选择班次开始时段和班次时长) 直到满足约束. 例如, H_1 表示随机选择解中的一个排班重新生成, 使新解满足约束. 本文设置 $l_{\max}=\lceil N/2\rceil$. 需要注意的是, 在 shaking 操作中可能会出现违反医生排班工作约束的不可行解, 对这种情况算法继续重复 shaking 操作, 直到 shaking 产生可行解.

3.4 局部搜索

对解进行 shaking 操作后, 继续通过局部搜索对解进一步优化. 本文考虑的局部搜索方法如下:

1) 将一名医生一个班次的工作结束时间提前一个时段 (缩短此班次时间长度一个时段), 该邻域的规模为 $O(N)$; 2) 将一名医生一个班次的工作结束时间推迟一个时段 (延长此班次时间长度一个时段), 该邻域的规模为 $O(N)$; 3) 将一名医生一个班次的开始时间提前一个时段 (延长此班次时间长度一个时段), 该邻域的规模为 $O(N)$; 4) 将一名医生一个班次的开始时间推迟一个时段 (缩短此班次时间长度一个时段), 该邻域的规模为 $O(N)$; 5) 将一名医生一个班次整体向后平移一个时段 (班次长度不变), 该邻域的规模为 $O(N)$; 6) 将一名医生一个班次整体向前平移一个时段 (班次长度不变), 该邻域的规模为 $O(N)$; 7) 从目前的排班方案中取消一个现有的医生班次, 该邻域的规模为 $O(N)$; 8) 为一名医生增加一个可行的新班次 (增加方式与初始解中生成医生排班的方式相同), 该邻域的规模为 $O(N\times T)$.

4 数值实验

本节利用 10 组实例数据进行数值实验, 包含小规模和大规模实例各 5 组. 数据来源于上海某三甲医院的呼吸内科就诊记录. 其中大规模实例数据来源于 2022 年 3 月上海病情爆发期的数据, 小规模数据源于 2022 年 7 月病情平稳时期数据. 下文将两类实例称为“病情爆发场景下实例”(编号 1-1～1-5) 与“病情平稳场景下实例”(编号 2-1～ 2-5). 两组实例在患者到达数目方面具有明显的区别, 病情爆发阶段患者的线上服务需求量更大. 在病情平稳场景下, 线上医疗服务的工作时间为 8 : 00 ～ 17 : 30, 每半小时为一个时段, 共 19 个时段; 有 6 名医生可供调度. 患者到达率在一天中有两个明显的峰值, 分别是第 3 ～ 5 个时段 (9 : 00 ～ 10:30) 和第 12 ～ 14 个时段 (14 : 30 ～ 16 : 00). 病情爆发场景下, 线上医疗服务的工作时间延长至 7 : 30 ～ 22 : 30; 有 7 名医生可以调度. 同样, 患者到达率在一天中有两个明显的峰值, 分别是第 3 ～ 5 时段 (8 : 30 ～ 10 : 00) 和第 19 ～ 22 时段 (16 : 30 ～ 18 : 00). 所有数值实验运行于 Windows 10 系统下 3.7 GHz Xeon CPU.

4.1 建模方法有效性验证

本节验证所提出的基于时变马尔可夫链和均匀化方法的建模方法, 检验不同实例采用医院实际排班的情况下解析计算得到的患者队列长度、逗留时间和医生加班时间的精度. 为了验证计算结果, 构建系统仿真模型, 仿真 (10^5 次) 得到患者队列长度、逗留时间等性能指标, 将仿真和解析计算结果对比, 验证所提出解析建模的计算精度. 医院实际线上医生排班为: 病情平稳时期, 医院在前 4 个时段安排一名医生线上服务, 后 15 个时段均安排两名医生线上服务; 病情爆发时期, 在患者到达高峰期 (5 ～ 12 时段、18 ～ 25 时段) 均安排 3 名医生线上服务, 其余各时段均安排两名医生线上服务.

数值实验结果表明, 在病情平稳和病情爆发场景下, 均匀化方法与仿真在患者排队队长的平均偏差分别为 0.06 % 与 0.04 %, 目标函数值的平均偏差分别为 0.03 % 与 0.01 %. 此外, 均匀化方法的运行时间相对于仿真大大缩短. 因此, 基于时变马尔可夫链和均匀化方法的建模方法可以精确计算线上医疗服务系统的各项指标, 具有很高的精确性, 可以作为评估一个医生线上工作排班效果 (例如评估一个排班方案解所对应目标值) 的快速精准方法.

4.2 医生排班求解结果与分析

4.2.1 VNS 收敛性能分析

本节测试 VNS 的收敛性能. 在小规模实例 (实例 1-4) 下, 算法目标值在前 10 次迭代中快速降低, 在经历 80 次迭代后趋于平稳. 在大规模实例 (实例 2-5) 下, 算法目标值在前 30 次迭代中快速降低, 在经历 100 次迭代后趋于平稳. 进一步, 本文对更高的迭代次数进行了测试, 例如, 迭代 1 000 次等. 根据测试得出结论, 在迭代次数超过 200 次以后, 算法很难再发现新的最好解. 基于以上对算法收敛性能的分析, 本文设定迭代次数为 200, 这可以有效地保证算法已经充分收敛, 并控制算法的运行时间.

4.2.2 病情平稳时期数据的计算结果对比

本节采用医院病情平稳时期的 5 组实例进行数值实验. 使用 VNS 对每个实例求解, 得到医生排班方案, 再对排班方案通过计算机仿真得到系统性能结果, 包括每位患者平均逗留时间、医生总工作时段数、医生加班时间、目标函数值等. 为了验证算法解的质量, 其求解结果与目前文献常用的启发式算法(模拟退火算法[23]) 以及医院实际排班方案对比, 即通过计算机仿真得到每种结果的性能参数并进行对比. 对于每个不同实例, VNS、模拟退火算法均运行 5 次, 取 5 次中目标函数最优的解作为最终得到的排班结果并评估各指标, 算法运行时间为 5 次运行时间的平均值.

表 1 显示了 5 个实例的数值实验结果信息. 其中: ST 表示每位患者的平均逗留时间, WP 表示医生的总工作时段数, OT 表示医生的加班时间, Obj 表示目标函数值, RT 表示算法运行时间, Gap 表示算法得到排班方案的目标函数值与实际排班的差距. 从表中数据可以得出, 所提出的 VNS 得到的医生排班方案相对于实际排班方案有一定程度的优化, 平均差距约为 4.53 %, 其中对于实例 1-4 的优化程度最高, 达到了 7.61 %, 对于实例 1-3 的优化程度最低, 但也有 3.38 %. 与实际排班相比, 在 VNS 得到的排班方案下, 患者在系统中逗留时间平均缩短了 3 min, 表明 VNS 得到的排班缓解了患者在系统内的等待. 但是, 医生的平均工作时段数除了实例 1-2 外均有些许增多, 同时医生的加班时间在所有实例中均略微增加, 这表明在病情平稳场景下的实际排班并不能完全满足患者的实际医疗需求, 偶尔会出现医疗资源短缺的情况, 而 VNS 得到的排班则为解决这一问题提供了一种可行的方法.

表1 病情平稳场景下 VNS、模拟退火排班和实际排班结果对比

实例	VNS					模拟退火		实际排班				Gap/%
	ST/min	WP	OT/h	Obj	RT/s	Obj	RT/s	ST/min	WP	OT/h	Obj	
1-1	22.69	35	0.60	67.46	596	67.47	3 803	25.63	34	0.50	70.32	4.07
1-2	22.23	34	0.49	63.68	393	63.68	915	24.27	34	0.35	66.02	3.55
1-3	22.09	35	0.47	64.88	401	64.88	3 090	24.69	34	0.40	67.15	3.38
1-4	22.62	37	0.53	64.85	521	65.04	2 167	26.79	34	0.46	70.19	7.61
1-5	22.85	36	0.69	70.77	516	70.77	3 130	26.38	34	0.61	73.76	4.04
均值	22.50	35.40	0.56	66.33	485	66.37	2 621	25.55	34.00	0.46	69.49	4.53

对比 VNS 与模拟退火算法, 在病情平稳场景下, VNS 的性能优于模拟退火算法. 从目标函数值的角度而言, 实例 1-1、1-4 下 VNS 所得排班目标值优于模拟退火, 其余 3 组实例目标值相同. 从运行时间的角度看, 模拟退火算法的平均运行时间为 2 621 s, 本文提出的 VNS 耗时 485 s, 后者具有运算时间上的显著优势.

进一步, 图 1 显示了实例 1-4 在 VNS 求解结果与实际排班下, 每个时段末患者在线排队队长的 95 分位值, 以及所希望控制的队长阈值. 可以看出, VNS 优化后的排班使队长得到了有效控制, 所有时段的队长都控制在阈值之下. 对比医院使用的实际排班结果可以发现, 部分时段队长超出队长阈值, 即形成了较长的患者排队现象. 此外, 排班结果显示, 两个不同的医生排班方案下, 医生的总工作时段数没有明显增加, 即医生工作负荷基本相同, 但是 VNS 求解得到的医生排班方案中, 医生的数目变化更加复合患者到达的波动, 即 VNS 优化后的排班方案使得医生的配置和患者需求更加匹配.

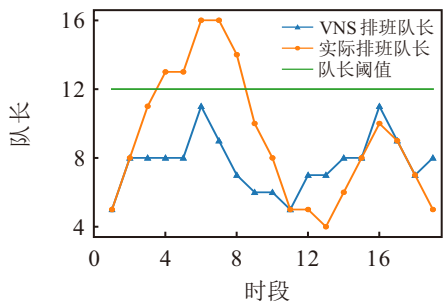

图1 病情平稳期算法与实际排班队长对比 (实例 1-4)

4.2.3 病情爆发时期数据的计算结果对比

本节采用病情爆发时期的 5 组实例进行数值实验. 同样, 使用 VNS 对每个实例求解, 通过仿真得到系统各项指标, 并与模拟退火算法结果以及医院实际排班方案进行对比. 表 2 显示了对病情爆发时期 5 个实例的求解结果. 从目标函数的角度看, VNS 得到的医生排班方案相对于实际排班方案有明显的优化, 平均差距约为 4.74 %, 其中实例 2-4 的优化程度达到了 7.98 %. 虽然对于实例 2-2 的优化程度仅有 1.39 %, 但算法求得的排班减少了医生的工作时段数, 对医疗资源进行了更为合理的分配. 从患者逗留时间的角度看, VNS 得到的医生排班均减少了患者在系统中的逗留时间, 平均缩短了约 3 min, 这满足

表2　病情爆发场景下 VNS 排班、模拟退火排班和实际排班结果对比

实例	VNS					模拟退火		实际排班				Gap/%
	ST/min	WP	OT/h	Obj	RT/h	Obj	RT/h	ST/min	WP	OT/h	Obj	
2-1	21.83	79	0.74	150.50	0.78	150.50	26.22	25.48	76	0.73	159.15	5.44
2-2	22.56	74	0.82	145.08	0.72	145.06	25.61	22.67	76	0.67	147.12	1.39
2-3	22.53	74	0.77	144.78	0.80	144.84	25.48	23.59	76	0.73	149.97	3.46
2-4	21.39	76	0.90	147.68	0.72	147.68	28.91	24.58	76	0.94	160.49	7.98
2-5	22.33	76	0.98	148.98	0.86	148.98	27.20	24.98	76	1.04	157.52	5.42
均值	22.13	75.80	0.84	147.40	0.77	147.41	26.68	24.26	76.00	0.82	154.85	4.74

了病情爆发时期患者对医疗资源的需求. 从医生工作时段数的角度看, VNS 得到的医生排班与实际排班差距较小, 因此, VNS 得到的医生排班在不增加医生工作负担的情况下优化了其他各项指标, 表明了所提出算法的有效性. 从医生加班时间的角度看, VNS 求得的排班在部分情况下 (实例 2-4、2-5) 可以减少医生加班现象, 减少医生的工作负担, 而在医疗资源不足 (实例 2-1、2-2、2-3) 的情况下, 则会增加一些加班时间使患者的需求得到保障.

对比 VNS 与模拟退火算法, 从目标函数值来看, 模拟退火和 VNS 各在一个实例上占优, 其余 3 个实例两者目标值相同. 从运行时间的角度来看, 模拟退火和 VNS 平均运算时间分别为 26 h 和 0.77 h. 相比模拟退火算法, VNS 可以在更短时间内得到质量更高解, 具有更好算法性能.

进一步, 图 2 以实例 2-5 为例, 显示了 VNS 求解结果与实际排班每个时段末队长的 95 分位值以及队长阈值. 可以看出, 经过 VNS 优化后的排班对队长阈值进行了有效控制, 每个时段的队长都被有效控制在期望的阈值之下, 这也可以表明基于 VNS 求得的排班方案, 患者的排队等待时间得到了控制. 相比之下, 医院实际排班则无法有效控制患者的在线队长, 出现了部分时段队长超出阈值的情况. 此外, 结果显示, VNS 求得的排班方案各时段医生数目随患者到达率波动变化得更加明显, 例如在患者到达高峰期有 4 名医生线上服务, 而在低谷期仅有 1 名医生线上服务. 这样更能满足病情爆发期间患者对医疗资源的高需求, 减少患者在系统内的逗留时间, 改善实际排班在患者到达率高峰期医疗资源不足的情况.

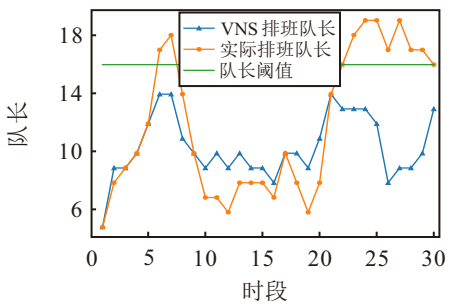

图2　病情爆发期算法与实际排班队长对比 (实例 2-5)

4.3　求解方案的鲁棒性数值实验

以上研究均是基于在线服务时间长度服从指数分布的假设. 虽然这是排队系统相关研究中的常见假设, 但是某些情况下此假设不能精确刻画现实情况. 为了验证本文所提出方法的鲁棒性, 本节突破此假设, 对服务时间采用 G 分布加以描述, 以验证建模和排班方法有效性. 通过离散事件仿真, 在 G 分布条件下, 对前文各实例在指数分布假设下通过 VNS 所得的排班方案加以仿真, 得到优化目标函数值等系统性能指标. 同时, 通过离散事件仿真在 G 分布条件下, 对实际排班方案的目标函数等指标进行评估, 并对每个实例的两种排班方案进行对比. 结果表明, 针对两类数据, 本文提出的 VNS 求解出的医生排班方案都均显优于实际排班方案, 尤其是从患者逗留时间和目标函数的角度看, 患者逗留时间明显减少, 目标函数值也得到了优化: 病情平稳期患者平均逗留时间从 32.20 min 缩短至 29.15 min, 目标函数值优化 3.89 %; 病情爆发期患者平均逗留时间从 30.96 min 减少到 29.07 min, 目标函数值优化 3.43 %, 且医生加班时间并未增加. 这些结果证实了本文所提出算法的优越性. 可以得出结论, 虽然本文的算法和模型是基于服务时间呈指数分布的假设, 但得到的排班方案足够鲁棒稳健, 能够在更加实际的场景中获得比目前实际排班更好的运行效果, 求解结果具有鲁棒性.

5　结　论

我国很多大型医院都开始提供线上线下联合医疗服务. 由于医生资源有限, 医院需要根据患者到达率等多种因素合理安排医生, 对医生的线上工作进行科学的排班和调度. 本文探讨了线上医疗服务系统的医生排班问题, 提出了基于时变马尔可夫链和均匀化方法的系统建模方法, 解析计算时变在线医疗服务系统中患者的逗留时间、队长等性能指标. 在

此基础上, 提出了 VNS 对医生排班问题加以求解. 基于合作医院病情平稳时期和病情爆发时期的数据, 数值实验验证了所提出方法可以帮助医院做出更合理的决策, 求解得到的医生排班优于医院实际采用的排班方案, 可以有效控制患者在线排队队长, 提升在线医疗服务水平. 在本文研究基础上, 后续将进一步对目标函数进行拓展, 例如尽量使得不同医生的工作负荷更加均衡, 研究不同目标函数对在线医疗服务系统中医生排班的影响.

参考文献 (References)

[1] 中国互联网络信息中心. 中国互联网络发展状况统计报告 [Z]. 2022.

[2] Erhard M, Schoenfelder J, Fügener A, et al. State of the art in physician scheduling[J]. European Journal of Operational Research, 2018, 265(1): 1-18.

[3] Xu D, Liu H W, Qi E S. Comparative study of nurse scheduling problem with different overtime strategies[J]. Journal of Systems Engineering, 2018, 33(2): 279-288.

[4] Zaerpour F, Bijvank M, Ouyang H Y, et al. Scheduling of physicians with time-varying productivity levels in emergency departments[J]. Production and Operations Management, 2022, 31(2): 645-667.

[5] 范雯娟, 兰绍雯, 裴军, 等. 多院区门诊医生排班方法及系统 [Z]. 2018.

[6] Wang Z X, Liu R, Sun Z K. Physician scheduling for emergency departments under time-varying demand and patient return[J]. IEEE Transactions on Automation Science and Engineering, 2023, 20(1): 553-570.

[7] Yang K, Liu Y X, Yang Z T, et al. A heuristic approach for emergency department weekly staffing and scheduling problem for time-varying demands[J]. Industrial Engineering and Management, 2020, 25(3): 171-178.

[8] Liu R, Xie X L. Physician staffing for emergency departments with time-varying demand[J]. Informs Journal on Computing, 2018, 30(3): 588-607.

[9] Lan S W, Fan W J, Yang S L, et al. A variable neighborhood search algorithm for an integrated physician planning and scheduling problem[J]. Computers & Operations Research, 2022, 147: 105969.

[10] Hu X W, Xu Y, Wang X L. Shift scheduling and rostering problem of call centers considering off-day fairness and same shift constraints[J]. Journal of Systems Engineering, 2021, 36(1): 88-101.

[11] Ahmadi-Javid A, Jalali Z, Klassen K J. Outpatient appointment systems in healthcare: A review of optimization studies[J]. European Journal of Operational Research, 2017, 258(1): 3-34.

[12] Ji M L, Li J L, Peng C. Two-stage chance-constrained telemedicine assignment model with No-show behavior and uncertain service duration[C]. Springer Proceedings in Business and Economics. Cham: Springer International Publishing, 2022: 431-442.

[13] Yu T, Guan Y P, Zhong X. Visiting nurses assignment and routing for decentralized telehealth service networks[J]. Annals of Operations Research, 2024: 1-31.

[14] Bayram A, Deo S, Iravani S, et al. Managing virtual appointments in chronic care[J]. IISE Transactions on Healthcare Systems Engineering, 2020, 10(1): 1-17.

[15] Guo H N, Xie Y, Jiang B W, et al. When outpatient appointment meets online consultation: A joint scheduling optimization framework[J]. Omega, 2024, 127: 103101.

[16] Rajan B, Tezcan T, Seidmann A. Service systems with heterogeneous customers: Investigating the effect of telemedicine on chronic care[J]. Management Science, 2019, 65(3): 1236-1267.

[17] 王娜, 李亚飞, 王洪峰. 基于鲁棒优化的云医疗资源配置问题[J]. 控制与决策, 2021, 36(2): 469-474.
(Wang N, Li Y F, Wang H F. Robust optimization based medical resource allocation problem in cloud healthcare system[J]. Control and Decision, 2021, 36(2): 469-474.)

[18] 路薇, 赵杰, 翟运开. 混合决策下考虑第三方偏好的远程医疗服务匹配方法[J]. 控制与决策, 2021, 36(11): 2803-2811.
(Lu W, Zhao J, Zhai Y K. Matching method for telemedicine service considering third-party preferences in context of mixed decision-making[J]. Control and Decision, 2021, 36(11): 2803-2811.)

[19] Luo J, Zhang J H. Staffing and control of instant messaging contact centers[J]. Operations Research, 2013, 61(2): 328-343.

[20] Fügener A, Brunner J O. Planning for overtime: The value of shift extensions in physician scheduling[J]. Informs Journal on Computing, 2019, 31(4): 732-744.

[21] Tang J F, Yan C J, Fung R Y K. Optimal appointment scheduling with no-shows and exponential service time considering overtime work[J]. Journal of Management Analytics, 2014, 1(2): 99-129.

[22] 代亮, 张金龙, 秦雯. 面向交通能源融合的路侧单元传输控制优化策略[J]. 控制与决策, 2023, 38(12): 3354-3362.
(Dai L, Zhang J L, Qin W. Optimization strategy of roadside units' transmission control for transportation-energy integration[J]. Control and Decision, 2023, 38(12): 3354-3362.)

[23] Wang Z X, Liu R. Managing appointments of outpatients considering the presence of emergency patients: The combination of the analytical and data-driven approach[J]. International Journal of Production Research, 2022, 60(13): 4214-4228.

作者简介

马颢洲 (2000−), 男, 博士生, 主要研究方向为服务系统的运作管理, E-mail: mahaozhou@sjtu.edu.cn;

刘冉 (1981−), 男, 副教授, 博士生导师, 主要研究方向为生产与服务系统的运作管理, E-mail: liuran2009@sjtu.edu.cn.