基于 SARSA 强化学习的审判人力资源调度方法

吴 鹏[1,2],魏上清[1],董嘉鹏[1],潘 理[1,2]
(1. 上海交通大学 电子信息与电气工程学院,上海 200240;
2. 信息内容分析技术国家工程实验室,上海 200240)

摘 要:为对法官员额资源进行调度优化,平衡司法资源有限和现实司法需求之间的矛盾,该文建立审判人力资源调度优化模型,提出基于强化学习的审判团队调度优化策略。基于对审判人员调度问题和场景的分析,建立以案件的平均处理时间最小化为优化目标的审判人员调度优化数学模型以及相应的约束条件。在此基础上建立宏观的司法系统排队模型,定义审判人力资源调度马尔可夫决策过程,并基于状态/动作/奖励/状态/动作(Sate-Action-Reward-State-Action,SARSA)算法提出动态自适应的审判人员调度强化学习算法。该算法以案件的平均处理时间为奖励,通过贪婪行为策略选择调度策略,采用时序差分更新方法在与司法系统交互的过程中学习最优调度策略。相比于传统分案方法及其他基于规则的简单启发式算法,该算法能够提高案件审判效率、优化人力资源配置。

关键词:强化学习;资源调度;决策优化;贪婪策略;马尔可夫决策过程
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2022)09-0082-07
doi:10.3969/j.issn.1673-629X.2022.09.013

Trial Human Resources Scheduling Method Based on SARSA Reinforcement Learning

WU Peng[1,2],WEI Shang-qing[1],DONG Jia-peng[1],PAN Li[1,2]
(1. School of Electronic Information and Electrical Engineering,Shanghai Jiao Tong University,
Shanghai,200240 China;
2. National Engineering Laboratory for Information Content Analysis Technology,Shanghai 200240,China)

Abstract:In order to optimize the scheduling of legal officials and balance the contradiction between the limited judicial resources and the actual judicial needs,a trial human resource scheduling optimization model and the trial team scheduling optimization strategy based on reinforcement learning are proposed. On the basis of analysis of the judiciary scheduling problems and scenarios,a mathematical model of judiciary scheduling optimization with the optimization goal of minimizing the average processing time of the case is established. On this basis,a macroscopic judicial system queuing model is established,the Markov decision-making process of trial human resource scheduling is defined,and a dynamic adaptive reinforcement learning algorithm for judicial personnel scheduling based on SARSA (Sate-Action-Reward-State-Action) is proposed. The algorithm uses the average processing time of the case as a reward,selects the scheduling strategy through the greedy behavior strategy,and uses the time-series differential update method to learn the optimal scheduling strategy in the process of interacting with the judicial system. Compared with the traditional division method and other simple rule-based heuristic algorithms,the proposed algorithm can improve the efficiency of case trials and optimize the allocation of human resources.

Key words:reinforcement learning;resource scheduling;decision optimization;greedy strategy;Markov decision process

0 引 言

随着国内人民法律意识的普及和司法体系的完善与发展,各级人民法院每年接收的、审理的案件数量迅速增长。然而,目前国内的司法资源是有限且不足的[1]。特别是司法资源中的审判人力资源,各级法院的审判工作人员数量年增长速率远小于案件数量增长。审判工作人员包括法官、法官助理、书记员三类。如何对审判人力资源进行调度优化,平衡司法资源有限和现实司法需求之间的矛盾,成为了各级人民法院亟待解决的问题[2]。

收稿日期:2021-10-09 **修回日期**:2022-02-10
基金项目:国家自然科学基金(62002219);上海市扬帆计划项目(19YF1424700)
作者简介:吴 鹏(1989-),男,博士,助理研究员,通信作者,研究方向为司法智能化、图神经网络等。

对于审判人力资源的调度主要集中在案件分配环节进行。目前法院主要采用均衡分案,案件立案后交给相应业务审判庭,由庭长(或其指定人员)按照一定的简单分案原则和庭室的具体情况为案件分配承办法官等审判人力资源。除了传统的人工分案方式之外,随着信息技术的高速发展,部分法院提出了电脑随机分案机制[3],分配法官时主要参考的是法官的办案数及存案数量的多少,将案件按序分给案件积存数量最少的法官。这些分案方法是一种细粒度静态资源分配,规则明确,容易操作。但是没有宏观考虑法院整体人员资源利用效率,难以根据宏观状况及时动态调整资源配置方案,也没有综合考虑法官助理、书记员等司法辅助人员的情况。因此为了提升审判人力资源使用效率,需要结合审判动态分析和司法统计智能分析建立宏观的审判人力资源调度优化模型。现实司法场景中,审判人力资源调度建模存在以下难题:一是案件审判不仅需要法官,还需要司法辅助人员,且不同类型的案件对审判人员配置的具体要求也不相同;二是案件审理请求、司法系统的状态是动态的,并具有一定随机性;三是人力资源不同于普通的物质资源,难以对其或者说对劳动力进行数学化定义。

为了解决以上问题,提升整体审判人力资源利用效率,该文第一次从宏观角度建立审判人力资源调度优化模型,基于案件平均处理时间最小化定义优化目标函数,根据资源限制和案件审理要求定义了优化的约束条件。为了在目标优化过程中实时感知审判人力资源的宏观状态和积存案件的情况,引入审判团队配置数组,将审判资源优化问题转换为使案件平均处理时间最小化的调度决策问题,进而提出基于 SARSA 强化学习的审判人力资源优化算法。强化学习算法利用智能体与环境的交互学习最优策略[4-5],相对于传统的资源调度方法,具有动态、自适应的优势[6-9]。近年来,将强化学习应用到资源分配与调度方向的研究在国内外已取得了诸多进展与成果[10-14],证实了强化学习算法在资源调度方面的有效性。

1　审判人力资源调度问题

考虑一个拥有一定量的司法人员并接受各类案件审理请求的司法系统,在本研究中,司法人员包括法官和审判辅助人员,其中审判辅助人员又分为法官助理和书记员,辅助法官处理案件。案件审理请求进入司法系统后等待司法系统调度司法人员进行审理。该系统具有以下合理假设:

(1)案件审理采用类型化裁判机制。基于较为通用的案件类型划分标准[15],将案件按照案件性质、适用程序分为五个类型:适用于简易程序的民事案件、适用于普通程序的民事案件、适用于简易程序的刑事案件、适用于普通程序的刑事案件以及适用于速裁程序的刑事案件。

(2)不同类型的案件在审理时对于审判人力资源的配置需求不同。民事诉讼法及刑事诉讼法中,分别对国内民事案件、刑事案件审判所需的成员做出了初步规定[15]。综合考虑了各个类型案件的要求和法官办案饱和工作量等因素,本研究对假设(1)中五种不同类型案件所对应的审判团队组成进行了更加详细具体的设置,审判团队的配置设定可以调整修改,且不影响本研究算法的执行。

(3)法官分为刑事法官和民事法官,刑事法官只审理刑事案件,民事法官只审理民事案件。此外,每位法官或司法辅助人员可以同时审理多个案件,但同时审理的案件数量是有限的。

(4)后文中所称的案件的处理时间包含案件等待分配审判团队的时间和案件开始审理到结案所需的预期审理时间,案件的预期审理时间假设为一随机变量,其估计为同类型案件历史审理时间的平均。该文提出的调度策略在分配案件时就充分考虑法官的案件工作量,因此假设出现由于案件堆积法官精力不足而使案件审理拖沓、最终审理时长远大于预期审理时间估计情况的概率较小。任意新案件的预期审理时间服从其估计上下一定范围内的均匀分布。

在具有以上假设前提的司法系统中,以一年 365 天为一个调度回合,一天进行一次调度决策与执行。每次调度时,根据司法系统的当前状态信息,例如系统中等待审理的案件类型、各类案件等待审理的案件数量和正在审理的案件数量等,选择配置哪些审判团队并按序分配给需要的案件。审判人力资源的调度策略具体地说就是按照不同类别案件的配置需求,组成相应的一定数量的审判团队完成系统中案件的审理。由此,审判人力资源调度优化问题可以定义为:给定司法系统的审判人力资源状态、案件等待状态和案件审理状态等,在资源限制条件、案件审理请求限制条件下进行审判团队的实时调度,来实现案件平均处理时间期望最小化的优化目标。

2　问题建模

该文提出一个司法系统排队模型,该系统主要由案件等待队列、审判人力资源调度决策器两部分组成,案件立案后进入系统等待调度相应的审判团队开始审理,调度决策器将利用强化学习方法对审判人力资源进行调度,配置所需审判团队并分配。对审判人力资源进行调度,配置所需审判团队并分配。系统拥有的审判人力资源包括 N_{jcivil} 名民事类型法官、$N_{\text{jcriminal}}$ 名刑

事类型法官、N_a 名法官助理和 N_c 名书记员，由于待审理的案件类型不同，系统需要将这些人力资源组成一定的审判团队，案件类型与审判团队类型是一一对应的。为避免案件积压，假设法官最多能够同时审理 m_1 个案件，包括法官助理和书记员在内的司法辅助人员最多能够同时处理 m_2 个案件。为了将审判人力资源调度问题中法官和案件的关系由一对多简化为一对一，以便于后续调度策略的选择等，将 1 位司法人员同时处理 m 个案件等效为 m 个法官同时各自处理 1 个案件。等效后，系统拥有的审判人力资源实际为 $m_1 * N_{\text{jcivil}}$ 名民事类型法官、$m_1 * N_{\text{jcriminal}}$ 名刑事类型法官、$m_2 * N_a$ 名法官助理和 $m_2 * N_c$ 名书记员。

案件立案后明确其类型，进入系统等待分配审判团队审理。案件的类型决定了审判团队配置，参数化审判团队配置，用 R_{ik} 表示第 i 类案件审判团队配置中的第 k 类审判人力资源的数量。司法系统模型如图 1 所示。

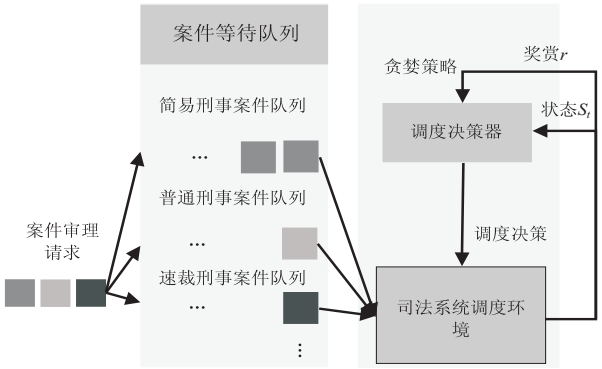

图 1 司法系统模型架构

案件等待队列存储等待分配审判团队审理的案件请求，并按照其案件类型分别放入 5 类案件队列中。调度决策器观测司法系统的状态，决定调度策略。调度环境代表司法系统除了等待队列、调度决策器以外的其他部分，包括审判人力资源、决策执行器等，用来为调度决策器提供所需的系统状态信息、反馈奖励值、执行调度策略等。一个新的案件审理请求产生后进入司法系统，按照类型进入不同的案件等待队列等待为其调度分配审判团队。在每个时隙的开始时刻，系统的调度决策器观测司法系统调度环境的状态，得到可用的审判人力资源、新收案件、各类型案件待审理请求等信息。结合以上状态信息，调度决策器通过强化学习方法学习调度审判人力资源的最优策略，而调度环境根据调度策略完成这一天对于审判人力资源的调度，计算奖赏值来帮助调度决策器后续优化。

假设每天会产生一批新的案件审理请求，请求已知案件的类型并给出案件的预期审理时间。在时隙 $[t, t+1]$ 的开始时刻 t，系统调度决策器决定该时隙如何利用空闲的审判人员组成审判团队并分配给待审理的案件。令 $Q_i(t)$ 表示时刻 t 开始时系统等待队列中等待分配审判团队审理的第 i 类案件的数量，$N_i^N(t)$ 表示在该时隙开始审理案件的第 i 类审判团队的数量，$N_i^P(t)$ 表示在上一时隙已经在审理案件且该时隙还未结案需要继续审理的第 i 类审判团队的数量，$N_i(t)$ 为该时隙内在审理案件的第 i 类审判团队的总数量，则在时隙 $[t, t+1]$ 内，所调度用于审理案件的第 i 类审判团队的总数量即为该时隙开始调度的审判团队数量和上一时隙已经调度且该时隙仍需调度的审判团队数量之和：

$$N_i(t) = N_i^N(t) + N_i^P(t)$$

以案件的平均处理时间期望最小为优化目标，而案件的处理时间包括案件在等待队列中等待分配审判团队的时间以及案件的预期审理时间两部分，后者服从某个小范围内的均匀分布。令 T_i^j 表示第 j 个 i 类案件的总处理时间，W_i^j 表示第 j 个 i 类案件的等待时间，S_i^j 表示第 j 个 i 类案件的预期审理时间，则有 $T_i^j = W_i^j + S_i^j$。第 i 类案件的平均处理时间即为：

$$\mathbb{E}(T_i) = \lim_{t\to\infty} \frac{\sum_{\tau=0}^{t-1} \sum_{j=1}^{N_f(\tau)} T_i^j}{\sum_{\tau=0}^{t-1} N_f(\tau)}$$

其中，$N_f(\tau)$ 为时隙 $[t, t+1]$ 内结束审理的案件数量。优化目标是最小化案件的总体平均处理时间，案件的总体平均处理时间定义如下：

$$\mathbb{E}(T) = \frac{\sum_{i=1}^{5} \mathbb{E}(T_i)}{5}$$

进行资源调度策略的优化时，还要考虑司法系统的资源有限和案件的审理需求条件，所调度的审判人力资源不能超出系统拥有的资源数量；调度时新的开始审理案件的审判团队数量显然不能多于待审理的案件数量，也不可能多于该时隙内调度的总数量。因此，审判人力资源优化问题即为在以下限制条件下通过最小化平均处理时间 $\mathbb{E}(T)$ 优化审判团队配置 $N_i(t)$，$t = 0, 1, \cdots, \infty$：

$$\begin{cases} \sum_{i=1}^{5} N_i(t) \cdot R_{ik} \leqslant C_k, \forall k \in K \\ N_i^N(t) \leqslant Q_i(t), i = 1, 2, \cdots, 5 \\ 0 \leqslant N_i^N(t) \leqslant N_i(t), i = 1, 2, \cdots, 5 \end{cases}$$

其中，C_k 表示司法系统所拥有的第 k 类审判人力资源数量，审判人力资源类别包括法官、法官助理和书记员三类。

在上一节中，假设将案件分为 5 大类型，对应 5 类审判团队。新的案件审理请求进入系统后，审判人力

资源调度是指根据案件审理请求确定不同类型审判团队的数量,并按照不同类型审判团队的司法人员配置组成审判团队分配给指定待审理案件。系统开始运行后,每个时隙的开始时刻,调度决策器需要根据司法系统的实时状态决定该时隙内调度审判团队的策略,即该时隙所要调度的不同类型审判团队的数量。使用每类审判团队的调度总数量组成的数组作为审判团队调度策略。一条审判团队调度策略定义为 $P_1=[a_1,a_2,a_3,a_4,a_5]$,其中 a_1、a_2、a_3、a_4、a_5 分别表示在时隙 $[t,t+1]$ 内所有正在审理案件的 5 类审判团队的调度数量,包括上一时隙已经在审理且该时隙仍需审理的团队以及该时隙新增的调度团队。调度的审判团队数量也需满足一定的限制条件。有效的审判团队调度策略调度的所有审判团队消耗的各类人力资源总量应不超过当时的可用数量。其次,当前时隙下开始调用的每一类审判团队数量不能超过该时隙系统等待队列中的对应类型案件数量。因此调度策略需满足下列约束条件:

$$\begin{cases}\sum_{i=1}^{5} a_i \cdot R_{ik} \leqslant C_k, \forall k \in K \\ N_i^P(t) \leqslant a_i, i=1,2,\cdots,5 \\ a_i - N_i^P(t) \leqslant Q_i(t), i=1,2,\cdots,5\end{cases}$$

显然,每个时隙内有效的调度策略通常不止一条,所以再将所有有效的调度策略合并为一个二维数组,称为审判团队调度策略数组 $P_{A_t\times 5}$,A_t 表示 t 时刻的调度策略总数。数组的每个行向量即为一条审判团队调度策略。调度决策器的任务相应地转换为在审判团队调度策略数组中尽可能找到最优的调度策略,以达到案件平均处理时间期望最小的目的。综上,审判人力资源的调度优化问题转换为在审判团队调度策略数组中找到满足上述约束条件的调度策略 P 使 $\mathbb{E}(T)$ 最小。

3　审判人力资源调度优化算法

审判人力资源是一个长期调度问题,没有“回合”概念,且人力资源调度状态空间和动作空间比较复杂,难以寻找最优策略。针对这些特点,该文选择 SARSA 算法作为调度优化算法的基础。引入审判团队配置数组后,审判资源优化问题就转换为使案件平均处理时间最小化的调度决策问题。决策器会在当前状态下选择行为,从而改变系统状态,其与司法系统交互的过程是一个马尔可夫决策过程(Markov Decision Process, MDP)。马尔可夫决策过程定义如下[16]:

$$M=(S,A,T,d_0,r,\gamma)$$

其中,S 为有限的状态集;A 为有限的动作集;T 为条件概率分布;d_0 定义初始状态分布;r 为奖励函数;γ 为标量折扣因子。在审判人力资源调度问题中,主要元素定义如下:

(1)状态空间(State)。

在每个时刻 t,可以从司法系统调度环境观测到一个全局状态向量 $\boldsymbol{s}_t \in \boldsymbol{S}$,为了更好地描述司法系统的可用资源和案件审理情况,状态向量由上一时隙已经调度且当前时隙仍需继续审理的审判团队数量和系统等待队列中的案件数量组成,即 $s_t=(N^P(t),Q(t)),t=0,1,2\cdots$。系统从开始到结束所有时刻下的状态向量就组成了状态空间。

(2)动作空间(Action)。

每次调度时,可以认为一条有效的审判团队调度策略就是一个可执行的调度动作,而所有的调度动作组成了动作空间 $A=P=(P(p_t))_{1\times 5}$,其中 $p_t \in \{1,2,\cdots,A_t\}, t=0,1,\cdots,\infty$。

(3)奖励函数(Reward)。

每次进行调度决策并执行后,计算奖励值。审判人力资源调度系统中以最小化案件平均处理时间为优化目标,故对于时隙 $[t,t+1]$ 内完成的调度,当存在结束审理的案件时,比较该案件的处理时间与设定阈值,该案件的处理时间相对更小时,可以认为此时达到了使案件平均处理时间尽可能小的目的,奖励为 1,否则为-1;另外当系统等待队列为空时,也认为调度策略效果较好,给予奖励值 1;其他情况下奖励均为 0。其中所设定阈值是案件平均处理时间估计的 β 倍,β 是一个大于 0 的系数。则奖励函数定义为:

$$r=\begin{cases}1,\text{if}\sum_{i=1}^{5} N_i^F(t)>0 \wedge \forall i, T_i(t) \leqslant \beta \bar{D}_i \\ -1,\text{if}\sum_{i=1}^{5} N_i^F(t)>0 \wedge \exists i, T_i(t)>\beta \bar{D}_i \\ 1,\text{if}\sum_{i=1}^{5} N_i^F(t)=0 \wedge \sum_{i=1}^{5} Q_i=0 \\ 0,\text{others}\end{cases} \tag{1}$$

其中,$T_i(t)$ 是 t 时刻第 i 类案件的处理时间,$\bar{D}_i$ 是第 i 类案件的平均处理时间估计,$N_i^F(t)$ 是 t 时刻第 i 类案件的结案数量。

基于强化学习的调度优化算法:

在以上定义与数学描述的基础上,该文基于 SARSA 算法提出审判人力资源调度优化的强化学习算法。初始时刻定义动作价值函数为 0,随后在训练过程中逐步更新,用来求解最优策略。每一个调度时隙开始时,系统感知环境的状态,计算状态向量,若该状态已知即存在于状态空间中,则借助 $q(s,a)$ 得到局部最优策略;否则将该状态加入,更新状态空间。在当

前状态下根据 ϵ -贪婪行为策略[17-18]决定调度策略，采取行为 a_t。最后按照对应决策执行审判团队调度分配，计算奖励值、预期回报等，更新价值估计。

$$q_\pi(s,a) = \mathbb{E}_\pi[r_{t+1} + \gamma q_\pi(S_{t+1},A_{t+1}) \mid S_t = s, A_t = a] \quad (2)$$

算法 1：基于强化学习的审判人力资源调度优化算法。

(1)初始化：$q(s,a) = 0$，$S = \{0\}$；
(2)for 每个时间步长 do
(3)感知系统状态 s_t
(4)if $s_t \notin S$ then
(5)将 s_t 加入 S
(6)end if
(7)计算审判团队配置数组 $P_{A_t \times 5}$
(8)基于 ϵ -贪婪策略选择行为 a_t
(9)执行审判团队调度，根据公式(1)计算奖励值，并由公式(2)更新 $q(s,a)$
(10)end for

4 实验分析

本研究所有仿真实验在一台运行 Windows 系统的笔记本电脑(Intel Core i5 3317U，16G RAM)上完成，实验代码使用 Python 编写。

4.1 实验设置

根据对中国多家不同层级的法院的实际调研以及相关法律文献对各个类型案件的审判团队配置要求的研究[15-16]，对假设(1)中五种不同类型案件所对应的审判团队组成进行了合理的设置，即表 1 中所示。需要注意的是，审判团队的设置并不影响本研究提出的算法的使用，对其他可能的设置该算法同样适用。

实验中以 1 时隙代表 1 天，在每个时隙生成一定数量的各类案件请求进入系统，设新收案件的产生速率为 λ 件/天。考虑到真实场景中新收案件类别具有不均衡的特点，尤其是民事案件和刑事案件之间的数量差距往往较大，将新收案件为民事案件的概率设为 p_c，则新收案件是刑事案件的概率为 $1 - p_c$。民事案件和刑事案件大类下适用于不同程序的案件概率则相等。通过 p_c 的变化模拟真实场景中民事案件和刑事案件数量不均的情形，并进行实验。

表 1 5 类审判团队的人员配置

案件类型	法官数量/人	法官助理数量/人	书记员数量/人
适用于简易程序的民事案件	1	2	1
适用于普通程序的民事案件	2	2	1
适用于简易程序的刑事案件	3	1	1
适用于普通程序的刑事案件	3	2	2
适用于速裁程序的刑事案件	3	2	0

每个案件的预期审理时间在各类案件预期审理时间估计的上下一定范围内随机产生。各类案件预期审理时间估计则为该类型案件之前的平均值。根据法院调研，5 类案件，即适用于简易程序的民事案件，适用于普通程序的民事案件，适用于简易程序的刑事案件，适用于普通程序的刑事案件以及适用于速裁程序的刑事案件的预期审理时间估计分别设为 10 天，20 天，12 天，25 天和 7 天，新的案件预期审理时间在对应类型估计的上下 2 天内均匀分布。

借助过去司法领域的相关调研报告，将每名法官最多同时审理的案件数量设为 8，每名司法辅助人员最多同时审理的案件数量设为 20，以保证法官等审判人员的精力足以适应案件审判工作量。最后，实验使用 365 天的调度数据来评价，通过计算案件的平均处理时间并比较进行调度效果的评估和检验。

4.2 对比算法

目前缺少审判人力资源调度优化的相关研究，本研究根据合理的启发式规则设计以下调度策略，并将它们与提出的基于强化学习的审判人力资源调度优化算法(RL)进行对比。

(1)完全随机策略(Random)：调度器在满足调度策略有效可行的前提下，每天为待审理的案件随机选择分配审判团队的调度。

(2)最短时间案件优先策略(Shortest Case First，SCF)：类似最短作业优先调度算法，调度器在分配调度时，为预期审理时间最小的案件分配审判团队进行审理优先。

(3)总调度团队数量最大策略(Maximum Total First，MTF)：选择使所调度的审判团队总数量最大的策略，使当前时隙能够开始审理的案件数量最多。

(4)传统电脑随机分案机制(Computer Random Division，CRD)：新的案件请求进入系统后，为其分配当前积存案件数量最少的法官并配备司法辅助人员。

4.3 实验结果

鉴于真实场景中民事案件占大多数，设民事案件的产生概率 p_c 等于 0.8 作为通用情况。在以上仿真条件设置下，将基于强化学习的审判人力资源调度优化算法多次训练，在不同新收案件产生速率下实验得到

的结果对比见表 2。可以看到,本研究提出的基于强化学习的审判人力资源调度优化算法得到的案件平均处理时间总是最小的,在 $\lambda = 1$ 时与次优的电脑随机分案机制相比降低了 29.2%。相比其他基于简单规则的调度策略算法,案件的平均处理时间均有明显的减少。说明在动态的、具有随机性的现实司法场景下,基于强化学习的审判人力资源调度优化算法能够通过实时地调度审判人力资源,极大地提升司法审判效率。

表 2　各算法的案件平均处理时间

Methods	$\lambda = 0.8$	$\lambda = 1$	$\lambda = 1.2$
RL	41.996±1.38%	39.362±2.03%	43.692±1.83%
Random	52.949±1.26%	76.823±1.69%	90.396±1.11%
SCF	55.036±1.81%	71.063±1.40%	86.763±1.16%
MTF	55.338±1.53%	57.127±1.41%	60.720±1.67%
CRD	47.672±4.25%	55.610±2.15%	57.672±1.75%

当新收案件的产生速率 λ 由 0.8 件/天逐渐增大到 1.2 件/天时,审判人力资源调度优化的强化学习算法得到的案件平均处理时间始终最低且较为稳定,其他的调度算法的结果则均有不同程度的增长。也就是说,在该司法系统这一年审判人力资源数量一定但接收的案件数量越来越多的情况下,强化学习算法可以在较大的审判工作量压力下保持对审判人力资源的优化配置和较高的审判效率;然而,包括传统电脑随机分案机制在内的其余算法会因为案件审判量的增大,其调度效果越来越差,无法很好地处理“案多人少”的问题。

本研究提出的算法经过 2 000 轮迭代训练平均需要 142 分钟,考虑到在实验中是对 365 天的收案情况进行调度,如果只对法院当天的收案情况进行审判人力资源调度平均所需时间不超过 1 分钟,因此该算法适合真实场景的应用。

4.4　司法系统敏感性实验

为了进一步分析案件类别不均时上述各种算法的表现,在实验中对民事案件产生概率 p_c 设置不同的参数值,表现民事案件与刑事案件之间的不均衡程度,对比结果见图 2。

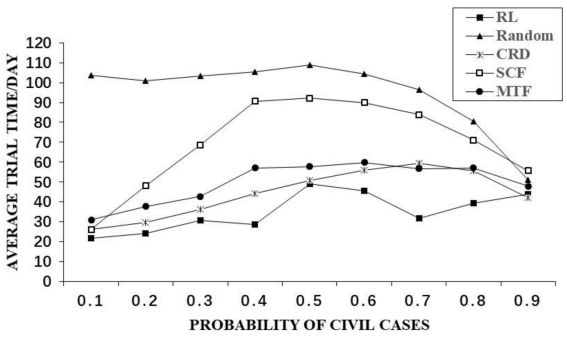

图 2　不同民事案件产生概率下各算法的实验结果

随机策略在刑事案件较多时对于案件的均衡程度并不敏感,案件平均处理时间均处在较高水平,但在民事案件较多时有所下降。传统的电脑分案机制和总调度团队数量最大算法则随着民事案件的增多,案件平均处理时间呈现缓慢上升趋势。最短时间案件优先算法的案件平均处理时间则先升后降,在民事案件与刑事案件产生概率相等时效果最好。基于强化学习的调度优化算法在民事案件与刑事案件数量大致相同时案件平均处理时间最长,当案件类别的不均衡程度增加时平均处理时间减少,同时在各种情况下其案件平均处理时间几乎都是最小。由此可得,审判人力资源调度的强化学习算法可以对案件的均衡程度变化自适应地调整策略,尽可能地进行优化,且调度效果均是最好。

4.5　参数敏感性实验

在强化学习方法中,奖励函数的定义是极其重要的部分,很大程度上决定了强化学习训练效果的好坏。本研究中,针对案件平均处理时间期望值最小化的优化目标设定了相应的奖励函数,主要是将当前策略得到的案件平均处理时间与设定阈值比较,较小就给予正奖励,否则给予负奖励。其中设定阈值为 5 类案件的平均预期审理时间估计的 β 倍。当 β 值过小时,难以获得正奖励,因此无法有效引导强化学习,当 β 值过大时,太容易获得正奖励,算法学习收敛效率较低。β 的设置需要平衡正奖励策略和负奖励策略的数量。为了分析审判人力资源调度的强化学习算法对于这个 β 值的敏感性,分别在不同的 β 取值下进行了实验,其中 $\beta = 3$ 是正式实验时所采用的参数值。实验结果见图 3。

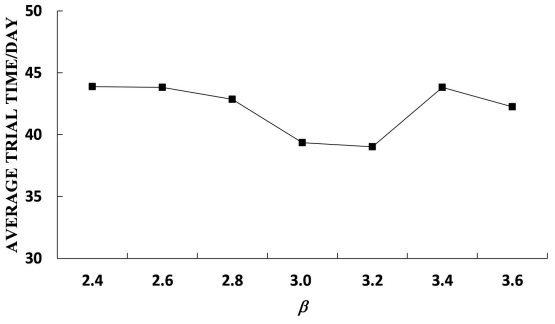

图 3　不同 β 取值下强化学习算法的实验结果

可以看到,案件的平均处理时间基本维持在 41 天上下,浮动不超过 4.8%,也就是说奖励函数设置中算法参数 β 在合理取值范围内对算法的实现效果影响有限,在可接受的范围内。

5 结束语

对司法场景下的审判人力资源问题进行了分析,并针对司法系统进行了数学建模,将审判人力资源调度问题转换为可计算的数学问题。在此基础上,为了提高司法审判效率、优化司法资源配置,以尽可能减小案件的平均处理时间为优化目标,提出了基于强化学习的审判人力资源调度算法,并在仿真条件下通过实验验证了算法的有效性和优越性。

参考文献:

[1] 北京市海淀区人民法院课题组,鲁为. 优化基层法院审判资源配置的调研报告——以审判人力资源配置为中心[J]. 人民司法,2011(19):59-62.

[2] 万 涛. 员额制背景下审判资源有效配置研究[J]. 福建法学,2019(1):67-74.

[3] 韦西妮. 法院分案制度研究[J]. 法制博览,2020(8):93-94.

[4] 刘虹庆,王世民. 基于强化学习的车辆路径规划问题研究[J]. 计算机应用与软件,2021,38(8):303-308.

[5] 王 欣,王 芳. 基于强化学习的动态定价策略研究综述[J]. 计算机应用与软件,2019,36(12):1-6.

[6] 亢中苗,汪 莹,张珮明,等. 云计算环境下基于强化学习的虚拟机资源调度[J]. 自动化与仪器仪表,2020(10):68-72.

[7] ASGHARI A,SOHRABI M K,YAGHMAEE F. A cloud resource management framework for multiple online scientific workflows using cooperative reinforcement learning agents [J]. Computer Networks,2020,179(10):107340.

[8] CUNHA B,MADUREIRA A,FONSECA B,et al. Intelligent scheduling with reinforcement learning [J]. Applied Sciences,2021,11(8):3710-3731.

[9] WANG L,HU X,WANG Y,et al. Dynamic job-shop scheduling in smart manufacturing using deep reinforcement learning[J]. Computer Networks,2021,190(2):107969.

[10] MASON K,GRIJALVA S. A review of reinforcement learning for autonomous building energy management[J]. Computers & Electrical Engineering,2019,78:300-312.

[11] JIANG B,FEI Y. Smart home in smart microgrid:a cost-effective energy ecosystem with intelligent hierarchical agents [J]. IEEE Transactions on Smart Grid,2015,6(1):3-13.

[12] LIU Y,ZHANG D,GOOI H B. Optimization strategy based on deep reinforcement learning for home energy management [J]. CSEE Journal of Power and Energy Systems,2020,6(3):572-582.

[13] PINTO G,PISCITELLI M S,VÁZQUEZ-CANTELI J R,et al. Coordinated energy management for a cluster of buildings through deep reinforcement learning[J]. Energy,2021,229(8):120725-120737.

[14] 刘冠男,曲金铭,李小琳,等. 基于深度强化学习的救护车动态重定位调度研究[J]. 管理科学学报,2020,23(2):38-52.

[15] 于 猛. 人民法院审判团队制度建设与模式选择——以基层人民法院审判团队的构建为例[J]. 法律适用,2018(11):66-71.

[16] 上海市第二中级人民法院课题组,郭伟清,阮忠良,等. 司法辅助人员配置问题研究[J]. 人民司法,2019(19):54-59.

[17] LEVINE S,KUMAR A,TUCKER G,et al. Offline reinforcement learning: tutorial, review, and perspectives on open problems[J]. arXiv:2005.01643,2020.

[18] SUTTON R S,BARTO A G. Reinforcement learning:an introduction[J]. Robotica,1999,17(2):229-235.