Advances in Applied Mathematics 应用数学进展, 2026, 15(3), 23-33
Published Online March 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.153084

Hans汉斯

基于强化学习的距离约束的最大加权目标覆盖问题研究

梁泳劲，何伟骅*

广东工业大学数学与统计学院，广东 广州

收稿日期：2026年2月1日；录用日期：2026年2月25日；发布日期：2026年3月3日

摘　要

距离约束的最大加权目标覆盖问题(MaxWTCDL)是一个NP-难组合优化问题。传统的启发式和近似算法在处理此问题时，往往容易陷入局部最优，难以求解。为了解决这些挑战，我们提出了一种基于强化学习的方法来求解MaxWTCDL问题。首先，我们将该问题重新表述为最大预算分组集合覆盖问题(MaxBGSC)的一个特例。我们将该问题建模为马尔可夫决策过程，并采用Q-learning算法来学习有效的传感器移动策略。实验结果表明，所提出的强化学习方法相比传统算法取得了更优越的性能，突显了其在高效解决带有距离约束的复杂覆盖问题方面的潜力。

关键词

组合优化，强化学习，马尔科夫决策过程，Q-Learning算法

A Reinforcement Learning Framework for Solving the Maximum Weighted Target Cover Problem with Distance Limitations

Yongjing Liang, Weihua He*

School of Mathematics and Statistics, Guangdong University of Technology, Guangzhou Guangdong

Received: February 1, 2026; accepted: February 25, 2026; published: March 3, 2026

Abstract

The Maximum Weighted Target Cover Problem with Distance Limitations (MaxWTCDL) is an NP-hard combinatorial optimization problem. Traditional heuristic and approximation algorithms often struggle with this problem due to a tendency to converge to local optima. To address these challenges,

*通讯作者。

文章引用: 梁泳劲, 何伟骅. 基于强化学习的距离约束的最大加权目标覆盖问题研究[J]. 应用数学进展, 2026, 15(3): 23-33. DOI: 10.12677/aam.2026.153084

we propose a reinforcement learning-based approach for solving the MaxWTCDL problem. Specifically, we reformulate the problem as a special case of the Maximum Budgeted Group Set Cover Problem (MaxBGSC). We model the problem as a Markov Decision Process and employ the Q -learning algorithm to learn effective sensor movement strategies. Experimental results demonstrate that the proposed reinforcement learning method achieves superior performance compared to traditional algorithms, highlighting its potential in efficiently solving complex coverage problems with distance constraints.

Keywords

Combinatorial Optimization, Reinforcement Learning, Markov Decision Process, Q-Learning Algorithm

1. 引言

目标覆盖问题作为无线传感器网络中的一个基本问题，具有广泛的应用，例如农业监测[1]、环境应用[2]以及工业检查[3]。目标覆盖问题这一主题吸引了许多研究人员的关注，目前已经有许多综述和专著[4]-[7]。已有诸多研究表明，传感器移动所消耗的能量远大于监测所消耗的能量[8] [9]。基于这一观察，许多研究人员针对在满足一定覆盖要求的前提下最小化移动距离的问题展开了研究[10]-[12]，还有一些研究对单个传感器加入了移动距离限制[13]。

在实际场景中，传感器数量有限、传感器能量受限以及环境因素等约束条件常常可能使我们无法实现完全覆盖。在这种情况下，一个很自然的问题就是如何有效地安排有限的资源，以覆盖最关键的目标。基于这样的考虑，Jin 等人提出了距离约束的最大加权目标覆盖问题(MaxWTCDL) [14]。给定平面上的一组目标，每个目标都有一个表示其重要程度的权重，MaxWTCDL 的目标是在单个移动距离约束和总距离约束下移动移动传感器，使得被覆盖目标的权重最大化。距离约束的最大加权目标覆盖问题(MaxWTCDL)是一个经典的组合优化问题，同时，在 2024 年被 Jin 等人证明是 NP 难问题，目前仅有两种多项式时间的近似算法求解这个问题，分别是贪心算法和随机线性规划算法。

组合优化问题作为计算机科学和运筹学的基本问题在过去几十年中受到了理论和算法设计社区的广泛关注[15]-[18]。解决 NP 难问题的传统方法包括精确算法、近似算法和启发式算法。多项式时间近似算法通常可以带来质量保证的解，但其优化保证不如精确算法强。特别是对于不适用于多项式近似算法的问题，可能根本不存在优化保证。此外，许多解决组合优化问题的传统算法都涉及到使用手工设计的启发式方法，这些方法会顺序地构建解决方案。这些启发式方法是由领域专家设计的，可能由于问题的复杂性而常常不是最优的[19]。

对此，强化学习(RL)提出了一个很好的替代方案，通过以监督或自监督的方式训练智能体来自动化这些启发式的搜索[20]-[22]。此外，一些研究将强化学习与传统启发式算法结合，通过智能体自动学习启发式规则，显著提升了算法效率和适用性[23]-[25]。

相关工作

强化学习近年来在组合优化领域得到了广泛且有效的应用，这为目标覆盖等类似问题的求解提供了

重要的方法论参考。

Khalil 等人(2017)率先提出将图组合优化问题建模为序列决策过程，其框架结合了拟合 Q 学习与图嵌入技术，通过贪婪策略逐步构造解，并在最小顶点覆盖、最大割及旅行商等经典问题上验证了所习得启发式方法的优越性与泛化能力[26]。

Hu 等人(2020)针对多旅行商问题的复杂性，设计了基于共享图神经网络与分布式策略网络的架构，将原问题分解为多个并行的子问题进行求解，其采用的 S 样本批量训练方法有效提升了学习稳定性，在求解质量与速度上均超越传统方法[27]。

Wang 等人(2023)则聚焦于无容量限制的 P-中值问题，利用图注意力网络与多对话头机制学习节点表示，并在 REINFORCE 算法框架下进行训练，取得了质量与效率的较好平衡[28]。

Guo 等人(2023)进一步提出了基于交换操作的深度强化学习方法，将设施重定位建模为马尔可夫决策过程，采用近端策略优化进行训练，并结合模仿学习与密度初始化策略，显著提升了求解性能与收敛速度[29]。

受先前工作的启发，我们采用强化学习框架来解决 MaxWTCDL 问题。我们提出的方法可以自适应地调整传感器的移动策略，以最大化被覆盖的总权重，同时遵守单个和总移动距离约束。

2. 研究问题

2.1. MaxWTCDL 问题定义

给定平面上的 n 个目标 $T=\{t_1,t_2,\cdots,t_n\}$，其中 $n\geq 1$ 为正整数，每个目标 $t_j\ (j=1,\cdots,n)$ 具有权重 $w_j\ (j=1,\cdots,n)$，m 个移动传感器 $S=\{s_1,s_2,\cdots,s_m\}$ 随机部署在平面上，其中 $m\geq 1$ 是正整数，感知半径 $r>0$，每个传感器的移动距离约束 $b>0$ 以及所有传感器的总移动距离约束 B，其中 $B>b$，MaxWTCDL 问题的目标是在距离约束 b 和 B 内移动移动传感器，使得被覆盖目标的总权重最大化。

为了便于定量分析和计算传感器与目标之间的空间关系，我们将平面建模为二维坐标网格。

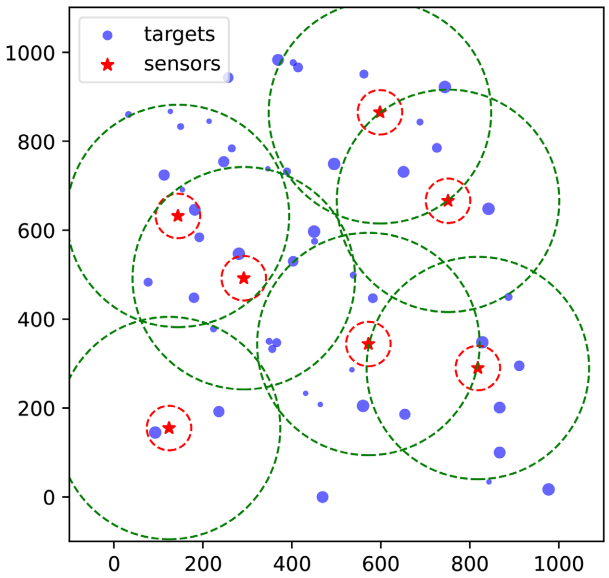


Figure 1. Distribution of sensors and targets on a discrete grid
图 1. 离散网格上的传感器和目标的分布

在图 1 中，传感器和目标分布在 1000 × 1000 的二维离散网格上。红色五角星表示传感器

$s_i\left(i=1,2,\cdots,7\right)$，蓝色圆点表示目标 $t_j\left(j=1,2,\cdots,50\right)$。蓝色圆点的大小对应于目标 $t_j\left(j=1,2,\cdots,50\right)$ 的权重 w_j。任何网格点 $location_{xy}\left(x=0,1,\cdots,1000;y=0,1,\cdots,1000\right)$ 都是传感器放置的潜在候选点。传感器 $s_i\left(i=1,2,\cdots,7\right)$ 周围半径为 r 的红色虚线圆圈表示其覆盖区域；如果目标 $t_j\left(j=1,2,\cdots,50\right)$ 到 s_i 的欧几里得距离小于 r，则认为该目标被覆盖。设 $d\left(s_i,location_{xy}\right)$ 表示传感器 $s_i\left(i=1,2,\cdots,7\right)$ 重新定位到网格点 $location_{xy}\left(x=0,1,\cdots,1000;y=0,1,\cdots,1000\right)$ 所需的移动距离。在单个移动距离约束 b 下，每个传感器 $s_i\left(i=1,2,\cdots,7\right)$ 都有一个相应的可移动区域 $E_i\left(i=1,2,\cdots,7\right)$，由图 1 中的绿色虚线圆圈表示。因为传感器的移动收到了总移动距离 B 的限制，所以问题的解空间是 $E_1\cup E_2\cup\cdots\cup E_7$ 的一个有限离散的子集合。

MaxWTCDL 问题可以近似表示为 0-1 整数规划模型。在这个 0-1 整数规划模型中，$y_k\left(k=1,2,\cdots,n\right)$ 表示目标 $t_k\left(k=1,2,\cdots,n\right)$ 是否被覆盖，其中 $y_k\in\{0,1\}$。$x_{ijz}\left(i=1,2,\cdots,m;j=1,2,\cdots,p;z=1,2,\cdots,q;p,q\in Z_{\geq 1}\right)$ 表示传感器 $s_i\left(i=1,2,\cdots,m\right)$ 是否移动到位置坐标 $location_{jz}\left(j=1,2,\cdots,p;z=1,2,\cdots,q;p,q\in Z_{\geq 1}\right)$。符号 d_{ijz} 和 g_i 的定义可以在第 2.3 节中找到。

$$\max\sum_{k=1}^{n}w_k y_k$$

$$\text{s.t. }\ y_k\leq\sum_{i=1}^{m}\sum_{jz:t_k\in A_{ijz}}x_{ijz}\quad\forall k=1,2,\cdots,n$$

$$\sum_{jz:A_{ijz}\in g_i}x_{ijz}\leq 1\quad\forall i=1,2,\cdots,m$$

$$\sum_{i=1}^{m}\sum_{jz:t_k\in A_{ijz}}x_{ijz}\cdot d_{ijz}\leq B$$

$$x_{ijz}\in\{0,1\}\quad\forall i=1,2,\cdots,m\text{ and }A_{ijz}\in G_i$$

$$y_k\in\{0,1\}\quad\forall k=1,2,\cdots,n$$

2.2. 工作流程

本文在强化学习框架内解决 MaxWTCD 问题，利用 Q-learning 算法。所提出方法的具体工作流程如图 2 所示。

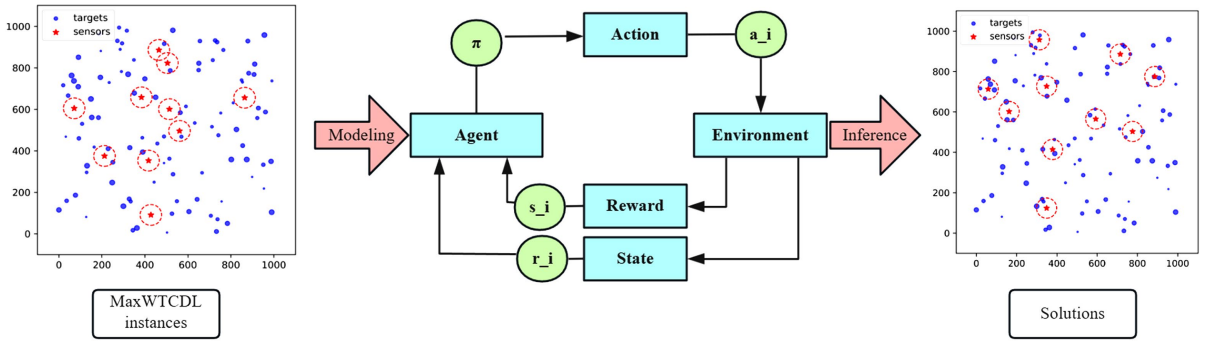


Figure 2. Flowchart of the reinforcement learning algorithm framework
图 2. 强化学习算法框架流程图

解决 MaxWTCDL 问题取决于找到最优的传感器移动策略 π_0。最优移动策略结果是一个传感器移动动作集 $g'=\{a_1',a_2',\cdots,a_n'\}$，其中 $a_i'=(1,2,\cdots,n)$ 表示传感器 $s_i=(1,2,\cdots,n)$ 在策略 π_0 下的最优动作结果。每个传感器 s_i 在执行动作 a_i' 后，获得一个新的分布位置。在第 2.3 节中，我们将 MaxWTCDL 问题重新表述为 MaxBGSC 的一个特例。在第 2.4 节中，基于强化学习框架，将研究问题建模为马尔可夫决策过程，并使用 Q-learning 算法进行求解。

2.3. 最大预算分组集合覆盖(MaxBGSC)问题

假设 $T=\{t_1,\cdots,t_n\}$ 是要覆盖的元素集合，其中 $n\geq 1$ 是正整数。每个目标 $t_j\ (j=1,2,\cdots,50)$ 有一个权重 $w_j\ (j=1,\cdots,n)$。设 g 是 T 的子集集合，分为 m 组 $g=g_1\cup g_2\cup\cdots\cup g_m$，其中 $m\geq 1$ 是正整数。每个子集 $A\in g$ 都有一个成本 $c(A)$。设 $B>0$ 为总预算，$n_1,n_2,\cdots,n_m$ 为 m 个正整数。MaxBGSC 的目标是选择一个子集集合 $g'\subseteq g$，使得：

- $c(g)=\sum_{A\in g'}c(A)\leq B$；
- 每个组 g_i 有 $|g_i\cap g|\leq n_i\ (i=1,2,\cdots,m)$，且；
- $w(C(g))=\sum_{t_j\in C(g')}w_j$ 最大化。

其中 $C(g)=\bigcup_{A\in g'}A$ 是被 g' 覆盖的元素集合。

MaxWTCDL 问题可以看作是 MaxBGSC 问题的一个特例。每个传感器 $s_i\ (i=1,2,\cdots,m)$ 对应于一组子集 $g_i=\{A_{i00},A_{i01},\cdots,A_{ipq}\}$，其中 $A_{ixy}\ (i=1,2,\cdots,m;x=0,1,\cdots,p;y=0,1,\cdots,q;p,q\in Z_{\geq 1})$ 表示在移动距离约束条件 $b>0$ 下，传感器 $s_i\ (i=1,2,\cdots,m)$ 可以移动到位置 $location_{ij}\ (i=1,2,\cdots,p;j=1,2,\cdots,q;p,q\in Z_{\geq 1})$ 然后覆盖所有距离小于 r 的目标。定义 $A_{ijz}\ (i=1,2,\cdots,m;j=1,2,\cdots,p;z=1,2,\cdots,q;p,q\in Z_{\geq 1})$ 对应的成本为：

$$c(A_{ixy})=\begin{cases}d_{ijz}, & \text{if } d_{ijz}\leq b\\ +\infty, & \text{if } d_{ijz}>b\end{cases}$$

这里，$d_{ijz}\ (i=1,2,\cdots,m;j=1,2,\cdots,p;z=1,2,\cdots,q;p,q\in Z_{\geq 1})$ 表示传感器 $s_i\ (i=1,2,\cdots,m)$ 初始位置到 $location_{xy}\ (x=1,2,\cdots,p;y=1,2,\cdots,q;p,q\in Z_{\geq 1})$ 的欧几里得距离。如果 $d_{ixy}\leq b$，这意味着 s_i 可以移动到 $location_{xy}$，因此 $c(A_{ixy})$ 等于 d_{ixy}。如果 $d_{ixy}>b$，这意味着 s_i 无法移动到 $location_{xy}$，因此 $c(A_{ixy})$ 等于 $+\infty$。设 $n_1=n_2=\cdots=n_m=1$，且 $g=g_1\cup g_2\cup\cdots\cup g_m$。

在这种情况下，选择满足约束条件 $c(g')\leq B$ 的子集 $g'\subseteq g$ 等价于所有传感器 $s_i\ (i=1,2,\cdots,m)$ 在其可移动区域 $E_i\ (i=1,2,\cdots,m)$ 内执行动作 $A_{ixy}\ (i=1,2,\cdots,m;x=0,1,\cdots,p;y=0,1,\cdots,q;p,q\in Z_{\geq 1})$，对应于产生所有距离成本 $\sum_{i=1}^{m}c(A_{ixy})\leq B$。

2.4. $\mathcal{Q}$-Learning 强化学习算法

$\mathcal{Q}$-Learning 是一种无模型强化学习算法，用于在没有先验环境信息的情况下在给定环境中找到最优策略。它允许智能体在环境中不断尝试、探索和学习，根据收到的奖励反馈调整对动作价值的评估，从而逐渐找到最优动作策略。

2.4.1. 算法框架

我们在强化学习框架内将我们的研究问题制定为马尔可夫决策过程(MDP)。MDP 的关键组成部分定义如下：

- **状态空间 $\mathcal{S}$**：状态空间 $\mathcal{S}$ 包含系统的所有可能状态。在本研究中，时间步 t 的状态 $s_t\in S$ 由所有目标和传感器的位置定义，因为目标是不可移动的，状态 s_t 主要是由传感器来决定的。因为连续的空间是太难以建模，本文一般对空间采用离散化处理的方式，例子说明如章节 2.1 所示。传感器移动受到移动距离 b 和 B 的限制，所以传感器的移动空间是 $E_1\cup E_2\cup\cdots\cup E_m$ 的一个有限离散的子集合。
- **动作空间 $\mathcal{A}$**：动作空间 $\mathcal{A}$ 由给定状态下任何传感器的所有可行移动组成。动作 $a_t\in\mathcal{A}(s_t)$ 对应于特定传感器的移动决策，受单个移动距离限制 b 和总移动距离预算 B 的约束。
- **奖励函数 $\mathcal{R}$**：奖励函数 $R(s,a)$ 产生在状态 s 采取动作 a 后的即时奖励。在我们的设置中，奖励定义为

状态从 s 转移到 s' 后，被覆盖目标总权重的增量收益：$R(s,a)=\sum_{k\in c(s')}w_k-\sum_{k\in c(s)}w_k$，其中 $c(s)$ 表示状态 s 下被覆盖目标的集合。

- $\mathcal{Q}$ 值：$\mathcal{Q}$ 值 $\mathcal{Q}(s,a)$ 表示在状态 s 采取动作 a 并此后遵循策略 π 的预期累积折扣奖励：

$$\mathcal{Q}(s,a)=E_\pi\left[\sum_{l=0}^{\infty}\gamma^k R_{t+k+1}\mid S_t=s,A_t=a\right]$$

其中 R_t 是在时间步 i 收到的奖励，$\gamma\in[0,1)$ 是平衡即时和未来奖励重要性的折扣因子。

- **策略 π**：策略 $\pi(a\mid s)$ 定义了在状态 s 采取动作 a 的概率。在这项工作中，我们采用源自 $\mathcal{Q}$-learning 算法的 ε-贪婪策略进行动作选择。

2.4.2. 算法步骤

$\mathcal{Q}$-Learning 算法在马尔可夫决策过程框架内运行。其核心机制涉及智能体迭代更新 $\mathcal{Q}$ 值以逼近最优状态－动作价值函数，进而指导最优策略的推导。更新公式如下：

$$\mathcal{Q}(s,a)\leftarrow(1-\alpha)\cdot\mathcal{Q}(s,a)+\alpha\left[r+\gamma\cdot\max_{a'}\mathcal{Q}(s',a')\right]$$

其中 α 是学习率 $(0<\alpha\le1)$，控制新信息覆盖旧 $\mathcal{Q}$ 值的程度。较大的 α 赋予最近的经验更多权重，而较小的 α 则保留过去的估计。

在我们的实验设置中，状态空间 $\mathcal{S}$ 和动作空间 $\mathcal{A}$ 如第 2.4.1 节所定义实例化，结合了距离约束 b 和 B。

在初始化期间，建立一个 $\mathcal{Q}$ 表，行表示 S 中的状态，列表示 A 中的动作。所有 $\mathcal{Q}$ 值初始化为 0，反映了智能体对环境奖励的初始估计，这将通过学习得到细化。学习率 α 和折扣因子 γ 也在 此阶段设置。

在每一回合中，智能体从初始状态 s_0 开始，并根据 ε-贪婪策略选择动作 a。以概率 ε，它选择随机动作进行探索；否则，它选择具有最高 $\mathcal{Q}$-值的动作，$a=\arg\max_a\mathcal{Q}(s,a)$，以利用当前知识。智能体在环境中执行动作 a，这会返回即时奖励 r 和新状态 s'。与我们的奖励函数定义一致，奖励 r 等于由状态转换导致的被覆盖目标总权重的变化。然后使用方程 $\mathcal{Q}(s,a)$ 更新状态-动作对 (s,a) 的 $\mathcal{Q}$ 值。最后，当前状态设置为 s' 以进行下一个决策步骤。

经过充分的探索和经验，$\mathcal{Q}$ 值保证收敛到其最优值。最优 $\mathcal{Q}$ 函数 $\mathcal{Q}^*$ 满足贝尔曼最优方程：

$$\mathcal{Q}^*(s,a)=E_\pi\left[R(s,a)+\gamma\cdot\max_{a'}\mathcal{Q}(s',a')\right]$$

3. 实验与结果分析

3.1. 实验数据与设定

在本研究中，为了评估算法性能，在二维平面 $[0,1000]\times[0,1000]$ 上随机生成 m 个传感器和 n 个目标。传感器和目标均服从均匀分布，目标的权重系数在区间 $[1,10]$ 内均匀分布。传感器覆盖半径设置为 $r_s=50$，每个传感器的最大移动距离约束为 $b=250$，所有传感器的总移动距离约束为 $B=2000$。

本文建立了多个实验数据集，以评估每种算法模型在不同场景下的性能，从而对其鲁棒性进行比较分析。对于传感器和目标的数量关系，分别按照 $m:n=1:5$ 和 $m:n=1:10$ 的比例配置了六组实验数据：

- 比例 $m:n=10:50,20:100$，和 30:150 下的三组数据集；
- 比例 $m:n=10:100,20:200$，和 30:300 下的三组数据集。

对此，我们可以明确得到各个传感器的坐标，根据传感器的最大移动距离约束为 $b=250$，可以得到各个传感器 $s_i\,(i=1,2,\cdots,m)$ 都有一个相应的可移动区域 $E_i\,(i=1,2,\cdots,m)$，从而可以得到一个传感器移动的“候选位置”集合，即解空间，为 $E_1\cup E_2\cup\cdots\cup E_m$ 的一个有限离散的子集合。

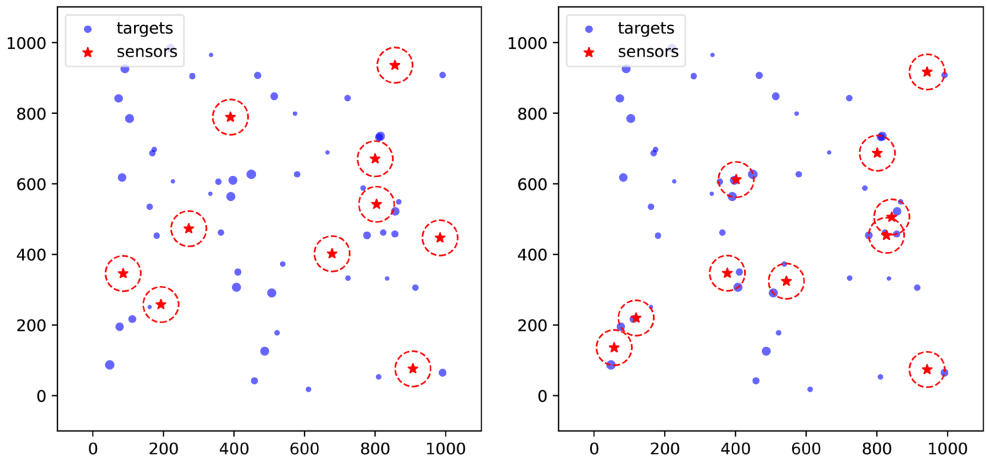
本文的实验模型算法主要包括六种类型：Gurobi 求解器、$\mathcal{Q}$-Learning 算法、贪婪算法(Greedy)、遗传算法(GA)、模拟退火算法(SA)和粒子群优化算法(PSO)。

Obj.表示优化问题中最优解的目标函数值。对于 MaxWTCDL 问题，它是被覆盖目标的权重之和。*Gap*.定义为使用特定方法获得的目标函数值与最优目标函数值之间的差值。在本研究中，特定方法获得的差距定义为：

$$Gap = \frac{Obj_{\text{gurobi}} - Obj}{Obj_{\text{gurobi}}}$$

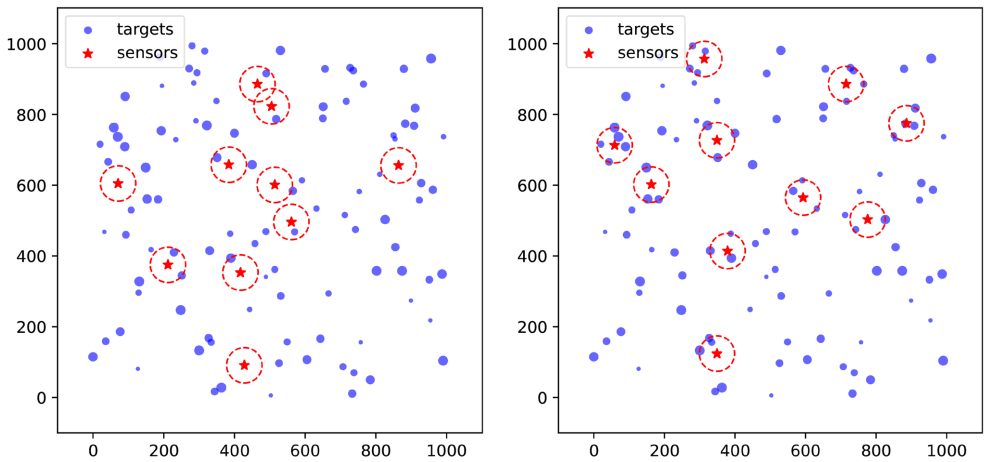
其中 Obj_{gurobi} 表示 Gurobi 求解器获得的最优解的目标函数值，Obj 表示使用特定算法获得的最优解的目标函数值。

3.2. 实验结果



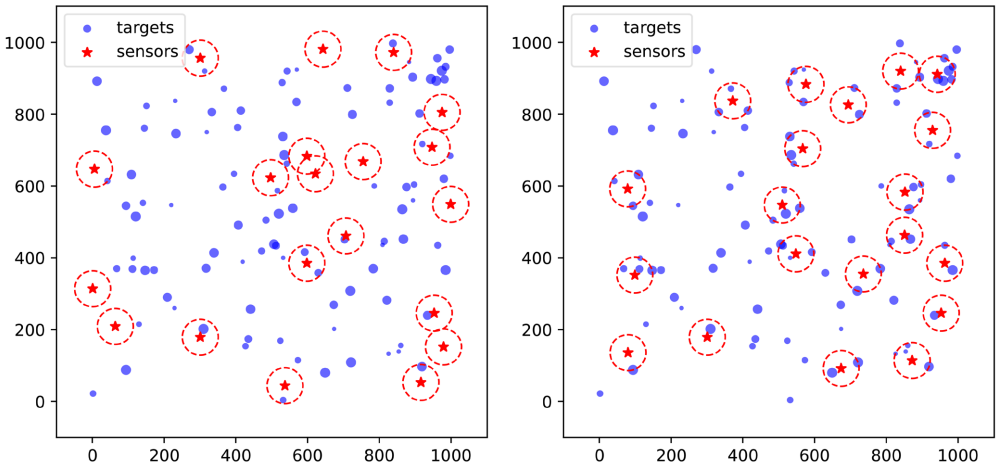
(a) 初始随机分布　　(b) Q-learning 后的覆盖结果

Figure 3. $\mathcal{Q}$-Learning algorithm in sensor deployment optimization ($m=10, n=50$)
图 3. 传感器部署优化中的 $\mathcal{Q}$-Learning 算法($m=10, n=50$)



(a) 初始随机分布　　(b) Q-learning 后的覆盖结果

Figure 4. $\mathcal{Q}$-Learning algorithm in sensor deployment optimization ($m=20, n=100$)
图 4. 传感器部署优化中的 $\mathcal{Q}$-Learning 算法($m=20, n=100$)



(a) 初始随机分布　　(b) Q-learning 后的覆盖结果

Figure 5. $\mathcal{Q}$ -Learning algorithm in sensor deployment optimization ($m=30, n=150$)
图 5. 传感器部署优化中的 $\mathcal{Q}$ -Learning 算法($m=30, n=150$)

Table 1. Performance comparison of optimization algorithms for sensor deployment ($m:n=1:5$)
表 1. 传感器部署优化算法的性能比较($m:n=1:5$)

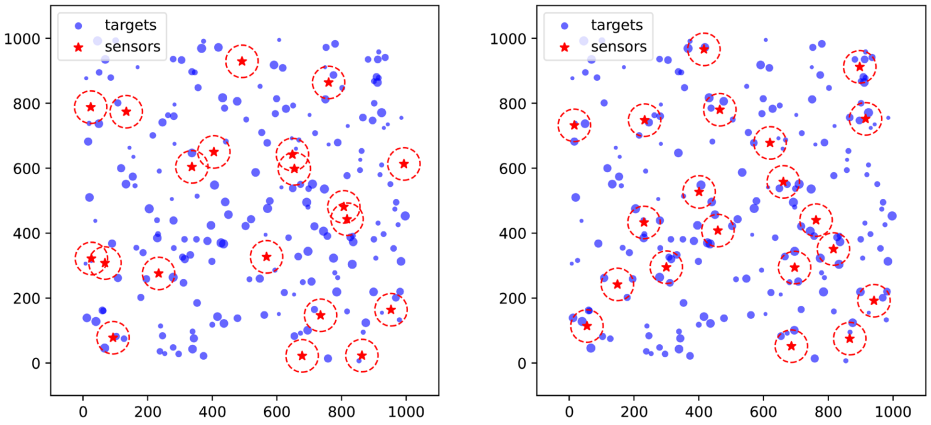
算法	$m=10, n=50$			$m=20, n=100$			$m=30, n=150$		
	Obj.	*Gap* (%)	*Time* (h)	*Obj.*	*Gap* (%)	*Time* (h)	*Obj.*	*Gap* (%)	*Time* (h)
Gurobi	104	0.00	0.11	326	0.00	2.31	566	0.00	5.21
Q-learning	104	0.00	0.40	324	0.61	1.12	551	2.65	2.88
Greedy	104	0.00	0.30	296	9.26	0.66	486	14.13	1.23
GA	104	0.00	0.49	296	9.26	2.55	532	6.01	8.91
SA	104	0.00	0.48	300	7.22	2.87	518	8.48	12.80
PSO	89	14.42	0.41	283	15.19	1.84	515	9.01	10.15

图 3、图 4 和图 5 展示了三种不同规模下优化前后传感器网络的比较。可以直观地观察到，经 Q-learning 算法优化后，传感器网络的分布得到了显著改善，从而获得了更好的覆盖性能。

根据上述的表 1，可以得到不同的算法模型在传感器与目标的数量关系比例为 1:5 的三种场景下的实验结果。在求解最优目标覆盖结果上，Q-learning 算法很接近 Gurobi 求解器，并且优于四种元启发式算法。相较于元启发式算法，Q-learning 算法也有明显优势。

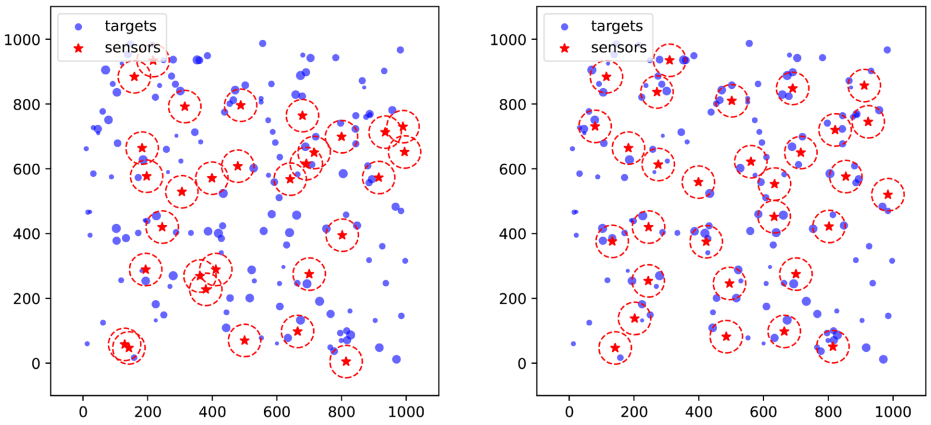
Table 2. Performance comparison of optimization algorithms for sensor deployment ($m:n=1:10$)
表 2. 传感器部署优化算法的性能比较($m:n=1:10$)

算法	$m=10, n=100$			$m=20, n=200$			$m=30, n=300$		
	Obj.	*Gap* (%)	*Time* (h)	*Obj.*	*Gap* (%)	*Time* (h)	*Obj.*	*Gap* (%)	*Time* (h)
Gurobi	195	0.00	0.11	479	0.00	3.98	826	0.00	8.31
Q-learning	192	1.54	0.67	469	2.09	2.14	815	1.33	3.27
Greedy	186	4.62	0.33	431	10.02	0.80	690	16.46	1.22
GA	186	4.62	0.59	456	4.80	5.15	760	7.99	15.13
SA	188	3.59	0.66	458	4.38	6.46	768	7.02	17.78
PSO	188	3.59	0.57	454	5.22	5.57	748	9.44	14.12



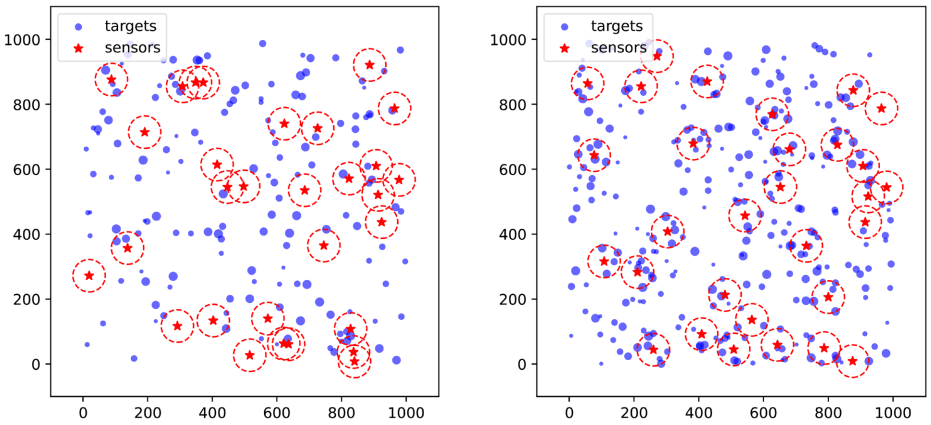
(a) 初始随机分布　　(b) Q-learning 后的覆盖结果

Figure 6. Q -Learning algorithm in sensor deployment optimization ($m=10, n=100$)
图 6. 传感器部署优化中的 Q -Learning 算法($m=10, n=100$)



(a) 初始随机分布　　(b) Q-learning 后的覆盖结果

Figure 7. Q -Learning algorithm in sensor deployment optimization ($m=20, n=200$)
图 7. 传感器部署优化中的 Q -Learning 算法($m=20, n=200$)



(a) 初始随机分布　　(b) Q-learning 后的覆盖结果

Figure 8. Q -Learning algorithm in sensor deployment optimization ($m=30, n=300$)
图 8. 传感器部署优化中的 Q -Learning 算法($m=30, n=300$)

图 6、图 7 和图 8 展示了三种不同规模 $(m,n=10,100;20,200;30,300)$ 下优化前后传感器网络的比较。显然，Q-learning 算法显著提高了传感器网络的覆盖性能。

根据上述的表 2，可以得到不同的算法模型在传感器与目标的数量关系比例为 1:10 的三种场景下的实验结果。在求解最优目标覆盖结果上，Q-learning 算法很接近 Gurobi 求解器，并且优于四种元启发式算法。相较于元启发式算法，Q-learning 算法也有明显优势。

4. 结论

基于强化学习算法框架，我们创新性地使用 Q-Learning 算法来解决 MaxWTCDL 问题。通过实验证明，该算法能够自适应地调整传感器移动策略，以在满足单个和总移动距离约束的同时，最大化目标覆盖的权重之和。与传统算法相比，本文提出的 Q-Learning 强化学习算法在解决本文研究问题方面具有更好的性能。

参考文献

[1] Mois, G., Folea, S. and Sanislav, T. (2017) Analysis of Three IoT-Based Wireless Sensors for Environmental Monitoring. *IEEE Transactions on Instrumentation and Measurement*, **66**, 2056-2064. https://doi.org/10.1109/tim.2017.2677619

[2] Lombardo, L., Corbellini, S., Parvis, M., Elsayed, A., Angelini, E. and Grassini, S. (2018) Wireless Sensor Network for Distributed Environmental Monitoring. *IEEE Transactions on Instrumentation and Measurement*, **67**, 1214-1222. https://doi.org/10.1109/tim.2017.2771979

[3] Li, X., Li, D., Wan, J., Vasilakos, A.V., Lai, C. and Wang, S. (2015) A Review of Industrial Wireless Networks in the Context of Industry 4.0. *Wireless Networks*, **23**, 23-41. https://doi.org/10.1007/s11276-015-1133-7

[4] Liang, J., Liu, M. and Kui, X. (2014) A Survey of Coverage Problems in Wireless Sensor Networks. *Sensors & Transducers*, **16**, 240-248.

[5] Chaturvedi, P. and Daniel, A.K. (2021) A Comprehensive Review on Scheduling Based Approaches for Target Coverage in Wsn. *Wireless Personal Communications*, **123**, 3147-3199. https://doi.org/10.1007/s11277-021-09281-7

[6] Wang, B. (2011) Coverage Problems in Sensor Networks. *ACM Computing Surveys*, **43**, 1-53. https://doi.org/10.1145/1978802.1978811

[7] Wu, W., Zhang, Z., Lee, W. and Du, D.-Z. (2020) Optimal Coverage in Wireless Sensor Networks, Volume 162 of Springer Optimization and Its Applications. Springer International Publishing.

[8] Somasundara, A.A., Ramamoorthy, A. and Srivastava, M.B. (2007) Mobile Element Scheduling with Dynamic Deadlines. *IEEE Transactions on Mobile Computing*, **6**, 395-410. https://doi.org/10.1109/tmc.2007.57

[9] Tan, R., Xing, G., Wang, J., et al. (2010) Exploiting Reactive Mobility for Collaborative Target Detection in Wireless Sensor Networks. *IEEE Transactions on Mobile Computing*, **9**, 317-332. https://doi.org/10.1109/tmc.2009.125

[10] Liao, Z., Wang, J., Zhang, S., Cao, J. and Min, G. (2015) Minimizing Movement for Target Coverage and Network Connectivity in Mobile Sensor Networks. *IEEE Transactions on Parallel and Distributed Systems*, **26**, 1971-1983. https://doi.org/10.1109/tpds.2014.2333011

[11] Chen, Z., Gao, X., Wu, F. and Chen, G. (2016) A PTAS to Minimize Mobile Sensor Movement for Target Coverage Problem. *IEEE INFOCOM* 2016—*The 35th Annual IEEE International Conference on Computer Communications*, San Francisco, 10-14 April 2016, 1-9. https://doi.org/10.1109/infocom.2016.7524334

[12] Wongwattanakij, N., Phetmak, N., Jaikaeo, C., *et al*. (2023) An Improved PTAS for Covering Targets with Mobile Sensors. arXiv:2305.03946.

[13] Quan, L.V., Hanh, N.T., Binh, H.T.T., Toan, V.D., Ngoc, D.T. and Lam, B.T. (2023) A Bi-Population Genetic Algorithm Based on Multi-Objective Optimization for a Relocation Scheme with Target Coverage Constraints in Mobile Wireless Sensor Networks. *Expert Systems with Applications*, **217**, Article 119486. https://doi.org/10.1016/j.eswa.2022.119486

[14] Jin, J., Ran, Y. and Zhang, Z. (2024) Approximation Algorithms for Maximum Weighted Target Cover Problem with Distance Limitations. *Journal of Combinatorial Optimization*, **47**, Article No. 60. https://doi.org/10.1007/s10878-024-01166-2

[15] Applegate, D.L., Bixby, R.E., Chvátal, V. and Cook, W.J. (2006) The Traveling Salesman Problem: A Computational Study. Princeton University Press.

[16] Perboli, G. and Rosano, M. (2019) Parcel Delivery in Urban Areas: Opportunities and Threats for the Mix of Traditional and Green Business Models. *Transportation Research Part C*: *Emerging Technologies*, **99**, 19-36. https://doi.org/10.1016/j.trc.2019.01.006

[17] Li, Y., Chu, F., Feng, C., Chu, C. and Zhou, M. (2019) Integrated Production Inventory Routing Planning for Intelligent Food Logistics Systems. *IEEE Transactions on Intelligent Transportation Systems*, **20**, 867-878. https://doi.org/10.1109/tits.2018.2835145

[18] Brouer, B.D., Alvarez, J.F., Plum, C.E.M., Pisinger, D. and Sigurd, M.M. (2014) A Base Integer Programming Model and Benchmark Suite for Liner-Shipping Network Design. *Transportation Science*, **48**, 281-312. https://doi.org/10.1287/trsc.2013.0471

[19] Golden, B., Bodin, L., Doyle, T. and Stewart, W. (1980) Approximate Traveling Salesman Algorithms. *Operations Research*, **28**, 694-711. https://doi.org/10.1287/opre.28.3.694

[20] Mazyavkina, N., Sviridov, S., Ivanov, S. and Burnaev, E. (2021) Reinforcement Learning for Combinatorial Optimization: A Survey. *Computers & Operations Research*, **134**, Article 105400. https://doi.org/10.1016/j.cor.2021.105400

[21] Bello, I., Pham, H., Le, Q.V., Norouzi, M. and Bengio, S. (2016) Neural Combinatorial Optimization with Reinforcement Learning. arXiv:1611.09940.

[22] Barrett, T., Clements, W., Foerster, J. and Lvovsky, A. (2020) Exploratory Combinatorial Optimization with Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 3243-3250. https://doi.org/10.1609/aaai.v34i04.5723

[23] Chen, M., Liu, S. and He, W. (2024) Learn to Solve Dominating Set Problem with GNN and Reinforcement Learning. *Applied Mathematics and Computation*, **474**, Article 128717. https://doi.org/10.1016/j.amc.2024.128717

[24] Wang, Q. and Tang, C. (2021) Deep Reinforcement Learning for Transportation Network Combinatorial Optimization: A Survey. *Knowledge-Based Systems*, **233**, Article 107526. https://doi.org/10.1016/j.knosys.2021.107526

[25] Wang, D. (2023) Reinforcement Learning for Combinatorial Optimization. In: Wang, J., Ed., *Encyclopedia of Data Science and Machine Learning*, IGI Global Scientific Publishing, 2857-2871. https://doi.org/10.4018/978-1-7998-9220-5.ch170

[26] Khalil, E., Dai, H., Zhang, Y., Dilkina, B. and Song, L. (2017) Learning Combinatorial Optimization Algorithms Over graphs. *Advances in Neural Information Processing Systems*, **30**, 6348-6358.

[27] Hu, Y., Yao, Y. and Lee, W.S. (2020) A Reinforcement Learning Approach for Optimizing Multiple Traveling Salesman Problems over Graphs. *Knowledge-Based Systems*, **204**, Article 106244. https://doi.org/10.1016/j.knosys.2020.106244

[28] Wang, C., Han, C., Guo, T. and Ding, M. (2022) Solving Uncapacitated P-Median Problem with Reinforcement Learning Assisted by Graph Attention Networks. *Applied Intelligence*, **53**, 2010-2025. https://doi.org/10.1007/s10489-022-03453-z

[29] Guo, W., Xu, Y. and Jin, Y. (2023) Swap-Based Deep Reinforcement Learning for Facility Location Problems in Networks. arXiv:2312.15658.