车联网中基于迁移强化学习的跨域充电站推荐算法

林海[1]，赵家仪[1]，曹越[1]，苏航宇[1]，王丽园[2]
（1. 武汉大学国家网络安全学院，湖北 武汉 430072；2. 中交第二公路勘察设计研究院有限公司，湖北 武汉 430056）

摘　要：深度强化学习已广泛应用于车联网充电站推荐，但传统方法通常需要为每个区域单独训练神经网络，增加了计算负担和数据需求。迁移学习通过利用先前任务的知识加速新任务学习，减少重复训练。为此，提出了基于迁移强化学习的跨域充电站推荐算法。该算法引入嵌入编码器对齐源域和目标域中系统状态空间和动作空间的维度，有效解决了维度差异问题。同时，该算法基于互信息构造变分分布，最大化对齐前后目标域状态相似度，确保迁移有效性。相较 3 个典型的充电站推荐算法，在低维向高维迁移中，该算法平均总充电时间分别减少 57.6%、59.3% 和 7.1%；在高维向低维迁移中，分别减少 12.3%、40.8% 和 4.7%。仿真结果证明该算法具备较强的迁移性，显著提升了跨域充电站推荐系统的性能。

关键词：深度强化学习；迁移学习；互信息；充电站推荐
中图分类号：TN915.08
文献标志码：A
doi：10.11959/j.issn.2096−3750.2025.00462

A transfer reinforcement learning-based approach for cross-domain charging station recommendation in the Internet of vehicles

LIN Hai[1], ZHAO Jiayi[1], CAO Yue[1], SU Hangyu[1], WANG Liyuan[2]
1. School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
2. China Communications Construction Company Second Highway Consultants Co., Ltd., Wuhan 430056, China

Abstract: Deep reinforcement learning has been widely applied in charging station recommendations in the internet of vehicles, but training separate neural networks for each region are often required by traditional methods, leading to increased computational load and data demands. Transfer learning accelerates the learning process for new tasks by leveraging knowledge from previous tasks, thus reducing redundant training. Therefore, a transfer reinforcement learning-based cross-domain charging station recommendation algorithm was proposed. An embedding encoder was introduced by this algorithm to align the system state and action space dimensions between the source and target domains, effectively solving the dimensionality discrepancy problem. Additionally, variational distributions were constructed based on mutual information to maximize the similarity between pre-aligned and post-aligned target domain states to ensure effective transfer. Compared to three typical charging station recommendation algorithms, in the low-dimensional to high-dimensional transfer, the average total charging time of the proposed algorithm was reduced by 57.6%, 59.3%, and 7.1%. In the high-dimensional to low-dimensional transfer, the reductions were 12.3%, 40.8%, and 4.7%, respectively. Simulation results demonstrate that the proposed algorithm exhibits strong transferability and significantly enhances the performance of cross-domain charging station recommendation systems.

Key words: deep reinforcement learning, transfer learning, mutual information, charging station recommendation

收稿日期：2024−10−13；**修回日期：**2025−05−30
通信作者：赵家仪，zhaojiayi@whu.edu.cn
基金项目：国家重点研发计划项目（No. 2023YFB3907105）；湖北省重点研发计划项目（No. 2023BAB022）
Foundation Items: The National Key Research and Development Program of China (No. 2023YFB3907105), Hubei Province Key Research and Development Program (No. 2023BAB022)

0 引言

随着新能源汽车的快速普及，充电资源供需矛盾日益突出。充电站推荐系统作为缓解这一矛盾的重要手段，其性能直接影响用户充电体验和充电资源利用效率。深度强化学习凭借其出色的决策能力和自适应学习特性，已广泛应用于车联网中[1-2]。

车联网中的充电站推荐通常被定义为顺序决策问题，并且可以建模为马尔可夫决策过程（MDP, Markov decision process）。通过深度强化学习算法[3]，充电站推荐系统能够根据系统状态（如充电站排队情况、用户位置等）自适应地做出推荐决策。然而，深度强化学习模型的有效性与环境联系紧密，当环境发生变化时，已训练模型往往难以直接应用于新环境，需要重新设计或训练模型，造成了大量计算资源的浪费[4]。

一种可行方法是利用源域的已训练模型，在新区域（目标域）上进行模型的再训练，从而加速目标区域的模型优化过程。然而，不同区域在充电站数量、分布位置等方面的显著差异[5]导致MDP模型在状态空间和动作空间维度上不一致。因此，源域训练好的网络模型既无法直接应用于目标域，也难以在目标域继续优化。这使得迁移学习成为解决这一问题的必要手段。

目前，学术界已进行了一系列使用迁移学习加速强化学习的研究工作。例如，示例深度Q学习[6]算法通过使用源任务样本数据进行预训练，结合经验回放加速任务的学习。考虑样本相似度，文献[7]基于目标任务和源任务样本的相似度重塑奖励，加速学习过程。此外，为了克服源域与目标域之间的显著差异，文献[8]构建中间域增强源域和目标域之间的关联性，并结合主要和非主要置信类别进行有效的知识迁移。针对负迁移的风险，文献[9]选择性地过滤与目标域无关的源域数据，并设计了基于样本重构误差的奖励机制，有效降低了负迁移的发生。

在车联网中进行充电站推荐任务的跨域迁移时，系统状态维度的差异性是一个亟待解决的关键挑战。图神经网络[10]被认为是解决跨域迁移强化学习中维度不匹配的有效方法。例如，文献[11]构建智能体形态图，并通过图变分自编码器引入结构归纳偏差。图神经网络充分利用跨域任务间的共性和差异，实现了知识的有效迁移。另一种常见解决方案是通过学习状态空间的映射函数来匹配源域和目标域的任务关系。文献[12]自主学习映射函数，通过将源域和目标域的状态映射至共同的特征空间，指导目标域的奖励函数，实现知识迁移。

上述基于图神经网络或映射的方法虽然在特定任务中展现良好性能，但仍存在明显局限性。首先，它们主要聚焦于机器人肢体控制问题，通用性有限。其次，这些方法在对齐状态维度时往往缺乏有效性保证机制，难以确保对齐后的特征能够完整保留原始状态信息。这些局限性使得现有方法难以直接应用于跨域充电站推荐。

针对上述问题，本文提出了一种基于迁移强化学习的跨域充电站推荐算法。该算法引入嵌入编码器，将不同维度的状态和动作空间映射到统一的维度，解决了不同充电区域间充电站数量和分布差异导致的状态空间和动作空间维度不一致问题。同时，该算法设计了基于互信息最大化的优化机制，有效保持了对齐前后目标域状态的相关性，显著增强了知识迁移的有效性和模型的域适应能力。此外，该算法设计了分层自适应知识迁移机制，逐层迁移源策略网络的知识，并通过动态调整源域和目标域特征的融合权重，有效降低了负迁移的风险。与现有方法相比，本文提出的算法更好地保留了原始特征的语义信息，确保了知识迁移的有效性。本文主要贡献如下。

（1）通过嵌入编码器将不同区域的状态空间和动作空间映射到统一的维度，解决了跨域充电站推荐系统中的维度差异问题，实现了源域知识向目标域的有效迁移，显著提升了模型的泛化能力和训练效率。

（2）基于互信息理论，通过变分分布拟合目标域状态分布并最大化相似度，使得源策略网络提取的特征能够充分表达目标系统状态信息，从而确保了编码前后目标域状态的相关性，保证知识迁移的有效性。

1 相关研究

1.1 充电站推荐

一种典型的充电站推荐系统包括电力系统控制中心、智能交通系统中心、充电站和电动汽车终端4个模块[13]。大部分充电站推荐系统采用先进先服

务策略[14]，但这种简单的策略可能带来电量紧张难以支撑等待时间，影响用户体验。针对这一问题，研究者提出了多种改进方案。文献[15]设计了基于紧急程度的优先充电策略，通过综合考虑电动汽车的能源需求和停车时间确定充电优先级。为了进一步提高用户满意度，文献[16]考虑了用户的充电意图，通过分析历史充电数据预测用户需求，并结合用户当前位置和时间进行实时充电站推荐。

上述确定性推荐方法往往忽视了交通环境的动态性和排队时间的不确定性，这些因素会导致推荐结果产生不可预见的偏差，进而影响充电站推荐的有效性[17]。为了应对这一挑战，研究者将充电站推荐问题建模为马尔可夫决策过程，并引入强化学习算法求解。文献[18]将电动汽车的充/放电调度问题定义为具有未知转移概率的MDP模型，将充电价格、车辆位置和充电状态作为系统状态，以充电成本作为奖励。文献[19]进一步考虑了充电站推荐中的环境不确定性，也将其建模为MDP，并采用强化学习算法优化推荐性能。

随着充电站推荐问题规模的增大，传统强化学习方法难以应对高维状态空间。深度强化学习的引入为解决这一问题提供了有效途径[20-24]。文献[20]提出了一种基于图的深度强化学习充电导航系统，有效降低了电动汽车的行驶时间和充电成本。文献[21]设计了多智能体演员–评论家框架。每辆电动汽车作为单独的智能体做出决策，而集中评论家则评估每个智能体的决策，以实现更优的推荐效果。尽管深度强化学习显著提升了充电站推荐系统的性能，但现有研究主要局限于单一区域，跨域推荐系统的性能仍有待提高。迁移学习作为一种有效的知识迁移方法，尽管在充电站推荐系统中尚未得到充分应用，在相关领域已有一些研究工作。文献[24]将迁移学习应用于电动汽车充电策略的跨域迁移，实现了充电策略的快速部署。此外，迁移学习被普遍认为能够解决（新环境）数据量不足的问题。当新电动汽车用户的充电行为数据不足时，研究者通过迁移已训练好的模型（基于老用户的数据）来预测新用户的充电行为[25]。类似地，文献[26]利用数据充足充电站的模型来预测数据匮乏充电站的充电负载。这些研究表明，迁移学习在应对车联网环境差异性方面具有重要价值，也在跨域充电站推荐领域发挥关键作用。

1.2　迁移学习

迁移学习主要分为同域迁移和跨域迁移两类。在同域迁移中，源任务与目标任务的域相同，可直接迁移知识。文献[27]提出了一种基于策略重用的学习算法。该算法包括探索策略和新旧策略的相似度函数，并在Q学习框架下不断更新复用概率。

跨域充电站推荐属于跨域迁移范畴，因源域与目标域之间存在差异，需要解决任务之间的域映射问题，以确保迁移模型的有效性和泛化能力[28]。文献[29]基于状态映射函数X_S和动作映射函数X_A推导出Q值的映射函数。随后，研究者提出了通过自动学习和推测源任务与目标任务之间的映射实现迁移的算法[30]。

为了对齐任务间的空间维度，文献[31]通过神经网络参数的迁移和微调加速混合动力车辆能源管理策略的学习。文献[32]提出了将各种状态空间统一为固定大小输入的框架，实现了深度强化学习策略的多场景复用以及多智能体的迁移学习。类似地，文献[33]提出了一个能够从多个系统的共享模型中选择适当训练模型的范式，计算输入数据与模型适配数据的相似度，并使用相似度高的模型。

2　问题定义与建模

跨域充电站推荐系统旨在解决不同充电区域间学习经验迁移的问题，通过有效利用已有区域的学习经验来加速新区域模型的训练过程，从而克服新区域训练数据不足和训练周期过长等问题。本文将充电站推荐的跨域迁移问题建模为MDP。MDP模型通常由五元组$(\boldsymbol{K},\boldsymbol{S},\boldsymbol{A},\boldsymbol{P},\boldsymbol{R})$构成，其中，$\boldsymbol{K}$表示决策周期，$\boldsymbol{S}$表示系统状态空间，$\boldsymbol{A}$表示动作空间，$\boldsymbol{P}$表示状态转移概率，$\boldsymbol{R}$表示奖励。

在充电站推荐系统中，系统状态$\boldsymbol{s}_t \in \boldsymbol{S}$由充电站信息、电动汽车信息和交通负载水平组成。在$t$时刻，系统状态定义为元组$\boldsymbol{s}_t=(\boldsymbol{\xi}_t,\boldsymbol{v}_t,\boldsymbol{k}_t)$。其中，$\boldsymbol{\xi}_t=(\mathrm{CS}_{1,t},\mathrm{CS}_{2,t},\cdots,\mathrm{CS}_{n,t})$表示$n$个充电站在$t$时刻的系统状态。对于每个充电站$\mathrm{CS}_i$，其状态$\mathrm{CS}_{i,t}=[\boldsymbol{c}_{i,t},\mathrm{type}_{i,t},\mathrm{qu}_{i,t},\mathrm{ow}_{i,t}]$包含以下属性：充电桩数量$c_i \in [1,\max_i{}^{\mathrm{c}}]$（$\max_i{}^{\mathrm{c}}$表示最大充电桩数量）、充电站类型$\mathrm{type}_i$（如快充站）、当前排队车辆数量$\mathrm{qu}_i$和已接受推荐正在途中的车辆数量$\mathrm{ow}_i$。请求充电的电动汽车信息$\boldsymbol{v}_t$表示为二元组$(\boldsymbol{l}_t,\boldsymbol{e}_t)$。其中，

$l_t \in \{l_1, l_2, l_3, \cdots, l_m\}$表示电动汽车请求充电时的位置，$e_t \in \{1, 2, 3, \cdots, e_{\max}\}$表示剩余电量能级。交通负载水平$k_t \in \{1, 2, 3\}$反映当前路网状况，分别对应轻度、中度和重度负载。

充电站推荐系统为处于状态$\boldsymbol{s}_t$的请求充电的电动汽车推荐适当的充电站，记该动作为a。系统的动作空间定义为$\boldsymbol{A} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n\}$。其中，$a_i$表示向请求车辆推荐充电站$\mathrm{CS}_i$的决策。充电站推荐系统根据当前状态$\boldsymbol{s}_t = (\boldsymbol{\xi}_t, \boldsymbol{v}_t, \boldsymbol{k}_t)$和采取的动作，按照转移概率$P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, a_t)$演化至下一状态$\boldsymbol{s}_{t+1} = (\boldsymbol{\xi}_{t+1}, \boldsymbol{v}_{t+1}, \boldsymbol{k}_{t+1})$，并获得相应的奖励。

考虑用户体验，本文设计的奖励函数r_t和总充电时间成反比，即总充电时间越少，奖励越多，从而鼓励系统尽量做出总充电时间少的推荐。总充电时间指从用户请求充电到完成充电的全过程时间，包括行驶时间τ_{tr}、排队等待时间τ_{qu}和充电时间τ_{ch}，表示为

$$r_t = \frac{1}{\tau_{\mathrm{tr}} + \tau_{\mathrm{qu}} + \tau_{\mathrm{ch}}} \tag{1}$$

对于跨域充电站推荐系统，源域（已训练完成区域）的MDP模型状态记为$\boldsymbol{s}_{\mathrm{src}} = (\boldsymbol{\xi}_{\mathrm{src}}, \boldsymbol{v}_{\mathrm{src}}, \boldsymbol{k}_{\mathrm{src}})$。目标域（新区域）的MDP模型状态记为$\boldsymbol{s}_{\mathrm{tg}} = (\boldsymbol{\xi}_{\mathrm{tg}}, \boldsymbol{v}_{\mathrm{tg}}, \boldsymbol{k}_{\mathrm{tg}})$。由于不同区域的充电站数量存在差异，$\boldsymbol{\xi}_{\mathrm{src}}$和$\boldsymbol{\xi}_{\mathrm{tg}}$的维度不一致，相应的动作空间维度也不同。这种维度不匹配是充电站推荐系统跨域迁移的主要挑战。

3　基于迁移强化学习的跨域充电站推荐算法

针对充电站推荐系统的跨域迁移问题，本文提出了一种基于迁移强化学习的跨域充电站推荐算法（TRCSA, transfer reinforcement learning-based cross-domain charging station recommendation algorithm）。该算法通过结合互信息理论和迁移学习技术，实现了不同区域间知识的有效迁移。TRCSA整体框架如图1所示。

TRCSA的核心思路主要包括3个关键步骤。首先，将目标策略网络的中间层初始化为与已在源域完成训练的源策略网络相同的结构。其次，目标域系统状态$\boldsymbol{s}_{\mathrm{tg}}$分别输入编码器$\phi$进行维度对齐和目标策略网络进行特征提取。编码器基于互信息通过深度神经网络将$\boldsymbol{s}_{\mathrm{tg}}$映射为与源域状态维度一致的向量表示$\phi(\boldsymbol{s}_{\mathrm{tg}})$，实现了不同区域状态空间的对齐。最后，将源策略网络基于$\phi(\boldsymbol{s}_{\mathrm{tg}})$提取的中间层特征$\boldsymbol{z}_{\theta'}$与目标策略网络的对应特征$\boldsymbol{z}_{\theta}$进行自适应融合，融合后的特征经过深层网络处理，输出充电站推荐决策。

3.1　状态空间对齐

为了实现跨域充电站推荐系统中源域和目标域状态空间维度的有效对齐，本文引入了嵌入编码器ϕ，将目标域系统状态$\boldsymbol{s}_{\mathrm{tg}}$编码为特征向量$\boldsymbol{s}_{\mathrm{emb}} = \{\phi(\boldsymbol{s})|\boldsymbol{s} \in \boldsymbol{s}_{\mathrm{tg}}\}$，嵌入编码器$\phi$如图2所示。

为了确保$\phi(\boldsymbol{s}_{\mathrm{tg}})$从源策略网络中提取有效知识

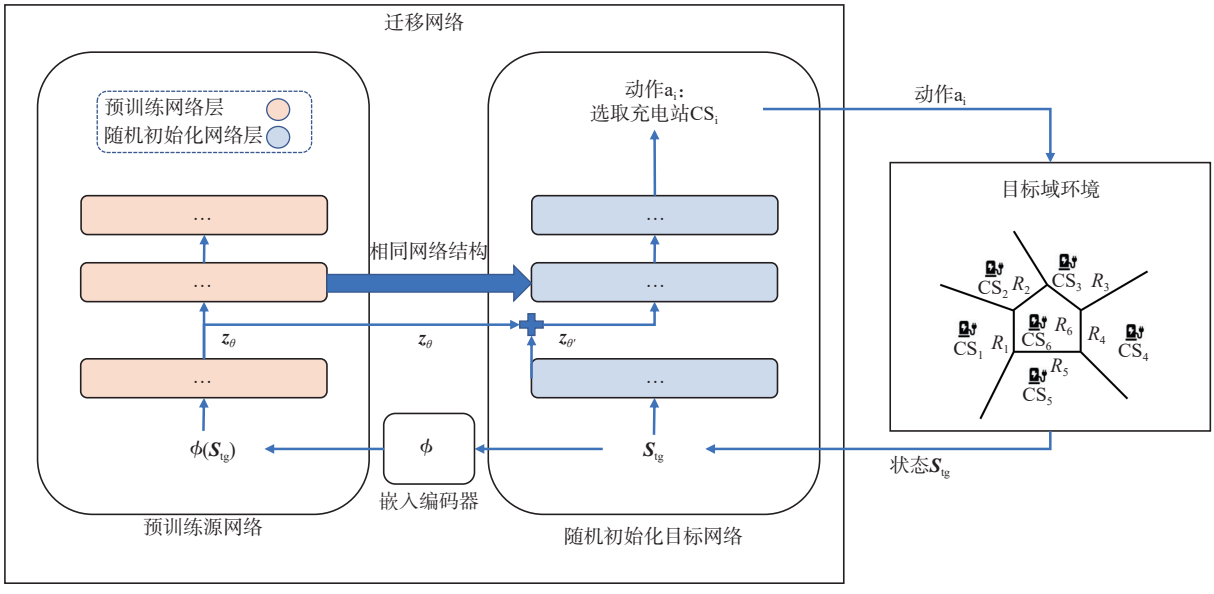


图1　TRCSA整体框架

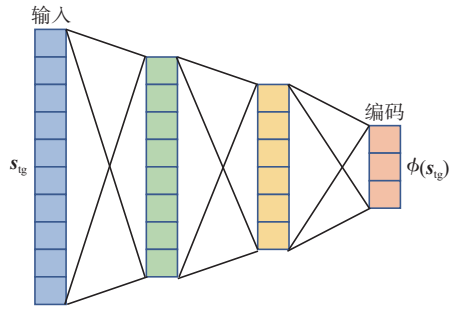


图2　嵌入编码器ϕ

以实现迁移，编码后的特征向量维度需要与源系统状态$\boldsymbol{s}_{src}$保持一致，即$|\boldsymbol{s}_{emb}|=|\boldsymbol{s}_{src}|$。

3.2　分层自适应知识迁移机制

分层自适应知识迁移机制主要包括3个关键步骤：源域特征提取、目标域特征学习和自适应特征融合。

首先，将对齐后的特征向量$\boldsymbol{s}_{emb}$输入源策略网络，提取线性中间层特征$\boldsymbol{z}'$，并将其迁移至目标策略网络。线性中间特征计算如下

$$\boldsymbol{z}'=\boldsymbol{W}'\phi\left(\boldsymbol{s}_{tg}\right)+\boldsymbol{b}' \tag{2}$$

其中，$\boldsymbol{W}'$和$\boldsymbol{b}'$分别表示源策略网络的权重矩阵和偏置。

其次，将目标域系统状态$\boldsymbol{s}_{tg}$输入目标策略网络，计算目标域线性中间特征为

$$\boldsymbol{z}=\boldsymbol{W}\boldsymbol{s}_{tg}+\boldsymbol{b} \tag{3}$$

其中，$\boldsymbol{W}$和$\boldsymbol{b}$分别表示目标策略网络中与源策略网络对应层的权重矩阵和偏置。

最后，为了实现平滑的知识迁移，设计动态特征融合机制，动态加权$\boldsymbol{z}$和$\boldsymbol{z}'$得到特征向量$\boldsymbol{h}$为

$$\boldsymbol{h}=(1-\varepsilon)\boldsymbol{z}+\varepsilon\boldsymbol{z}' \tag{4}$$

其中，$\varepsilon\in[0,1]$为自适应权重参数。在训练初期，ε接近1，特征向量$\boldsymbol{h}$主要由源网络特征$\boldsymbol{z}'$构成，充分利用源域知识。随着训练的进行，ε逐渐降低至0，$\boldsymbol{h}$逐步过渡为由目标网络特征$\boldsymbol{z}$主导，实现知识的渐进式迁移。融合特征$\boldsymbol{h}$经过激活函数$f(\cdot)$后传递至网络深层，该迁移过程在策略网络的每一层执行。单层网络的迁移计算过程如图3所示，图3中蓝色节点代表目标策略网络，黄色节点代表源策略网络。

TRCSA采用神经网络实现嵌入编码器ϕ。为了提升ϕ的性能，从两个方面进行优化。首先，TRCSA通过最大化目标MDP的累计奖励来确定嵌入参数，以增强编码效果。其次，引入互信息（MI, mutual information）保证$\boldsymbol{s}_{tg}$与$\phi(\boldsymbol{s}_{tg})$之间的高度相关性，

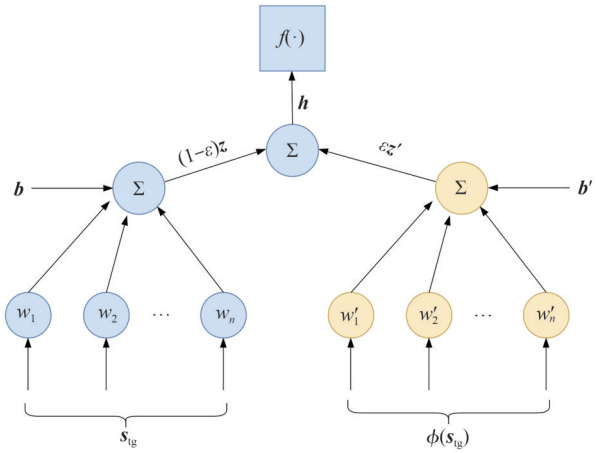


图3　单层网络的迁移计算过程

使每个目标策略的系统状态都能获得源策略的有效指导。这种双重优化机制显著提升了状态维度对齐和知识迁移的效果。

3.3　互信息最大化

本文引入互信息度量目标系统状态$\boldsymbol{s}_{tg}$与其编码表示$\phi(\boldsymbol{s}_{tg})$之间的相关性，保证源策略网络知识迁移的有效性。互信息可量化一个随机变量包含的另一个随机变量的信息量，是衡量两个变量统计相关性的重要指标。将目标域状态$\boldsymbol{s}_{tg}$作为随机输入$\boldsymbol{s}$，嵌入编码器ϕ的输出作为随机变量$\boldsymbol{e}$，两者的互信息定义如下

$$I(\boldsymbol{s};\boldsymbol{e})=H(\boldsymbol{s})-H(\boldsymbol{s}|\boldsymbol{e}) \tag{5}$$

其中，H表示差分熵。由于随机变量的条件密度未知，无法直接最大化互信息。为了解决这一问题，本文使用变分分布$q_w(\boldsymbol{s}|\boldsymbol{e})$获得互信息的下界，用于近似真实条件分布$p(\boldsymbol{s}|\boldsymbol{e})$，式(5)改写为

$$\begin{aligned}&I(\boldsymbol{s};\boldsymbol{e})=H(\boldsymbol{s})-H(\boldsymbol{s}|\boldsymbol{e})=H(\boldsymbol{s})+E_{s,e}\left[\log p(\boldsymbol{s}|\boldsymbol{e})\right]=\\&H(\boldsymbol{s})+E_{s,e}\left[\log q_w(\boldsymbol{s}|\boldsymbol{e})\right]+E_e\left[D_{\mathrm{KL}}\left(p(\boldsymbol{s}|\boldsymbol{e})\|q_w(\boldsymbol{s}|\boldsymbol{e})\right)\right]\geqslant\\&H(\boldsymbol{s})+E_{s,e}\left[\log q_w(\boldsymbol{s}|\boldsymbol{e})\right]\end{aligned} \tag{6}$$

其中，不等式的成立基于KL（Kullback-Leibler）散度$D_{\mathrm{KL}}(\cdot)$的非负性，这一方法被称为变分互信息最大化算法。此外，根据目标MDP状态和编码器参数，将优化目标形式化为变分参数和编码器参数的联合优化问题：

$$\max_{\boldsymbol{W},\phi}E_{s_{tg}}\left[\log q_w\left(\boldsymbol{s}_{tg}|\phi\left(\boldsymbol{s}_{tg}\right)\right)\right] \tag{7}$$

其中，由于$H(\boldsymbol{s})$是与参数无关的常数项，可在优化过程中忽略。最终，损失函数L_{MI}形式化为

$$L_{\mathrm{MI}}(\boldsymbol{W},\phi)=-E_{s\sim\rho_{\pi_\theta}}\left[\log q_w(\boldsymbol{s}|\phi(\boldsymbol{s}))\right] \tag{8}$$

其中，ρ_{π_θ}是目标MDP中目标策略的系统状态分布。通过最小化该损失函数，可以有效提升状态编码的质量，确保知识迁移的准确性和有效性。

3.4　TRCSA框架及工作机制

TRCSA的核心组件包括嵌入编码器、源策略网络和目标策略网络，这3个关键模块协同实现知识的有效迁移，TRCSA核心组件及工作机制如图4所示。

（1）嵌入编码器ϕ：是实现充电站推荐系统跨域迁移的关键组件。其主要功能是将目标域系统状态维度与源域对齐。编码器（蓝色）接受目标域状态输入，通过非线性变换生成维度对齐的向量表示$\phi(\boldsymbol{s}_{\mathrm{tg}})$。为了确保编码质量，编码器的参数优化同时考虑两个损失函数：互信息损失L_{MI}和值函数损失L_{PPO}。其中，L_{MI}用于保证$\phi(\boldsymbol{s}_{\mathrm{tg}})$和$\boldsymbol{s}_{\mathrm{tg}}$之间的高度相关性。而$L_{\mathrm{PPO}}$结合了环境反馈的奖励信息，指导编码器生成更有利于决策的特征表示。

（2）源策略网络和目标策略网络：将编码后的目标域状态表示$\phi(\boldsymbol{s}_{\mathrm{tg}})$输入源和目标策略网络，在网络的每一层分别计算源网络的特征表示$\boldsymbol{z}_{\theta'}$（θ'为源网络参数）和目标网络的特征表示$\boldsymbol{z}_{\theta}$（θ为目标网络参数）。自适应权重机制融合两个网络的特征表示（如式(4)所示），经过激活函数处理，传递至目标网络的下一层，实现源策略网络的逐层知识迁移。TRCSA工作流程如下所示。

算法　基于迁移强化学习的跨域充电站推荐算法

输入　源策略网络参数θ'，目标策略网络参数θ，目标策略网络的目标Q网络参数$\bar{\theta}$，嵌入编码器参数ϕ，变分分布参数w

输出　选择的充电站CS_i

for 每个请求 do

　　计算嵌入后的系统状态$\phi(\boldsymbol{s}_{\mathrm{tg}})$；

　　$\phi(\boldsymbol{s}_{\mathrm{tg}})$输入源策略网络得到可迁移知识$\boldsymbol{z}_{\theta'}$；

　　$\phi(\boldsymbol{s}_{\mathrm{tg}})$输入互信息网络得到$L_{\mathrm{MI}}(\theta,w)$；

　　$\boldsymbol{s}_{\mathrm{tg}}$输入目标策略网络得到线性中间特征$\boldsymbol{z}_{\theta}$和$L_{\mathrm{PPO}}(\theta,\phi,\theta')$；

　　将$\boldsymbol{z}_{\theta}$与$\boldsymbol{z}_{\theta'}$相加后通过激活函数；

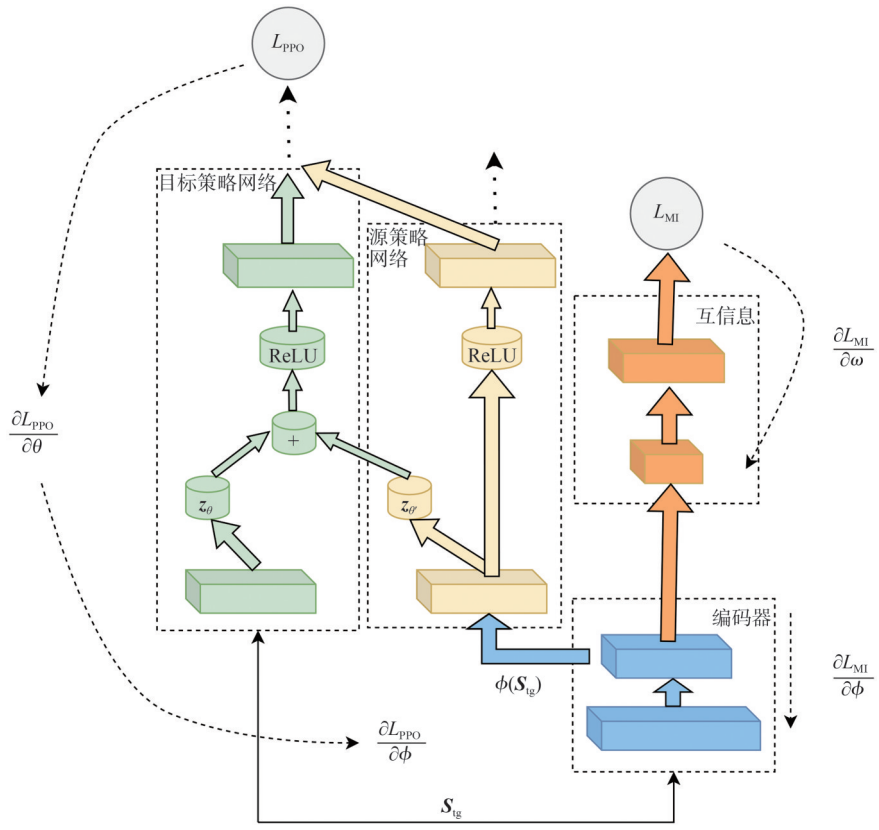


图4　TRCSA核心组件及工作机制

用$\nabla_{\theta}\left(L_{\mathrm{PPO}}(\theta,\phi,\theta')\right)$更新参数$\theta$；

用$\nabla_{\phi}\left(L_{\mathrm{MI}}(\phi,w)+L_{\mathrm{PPO}}(\theta,\phi,\theta')\right)$更新参数$\phi$；

用$\nabla_{w}L_{\mathrm{MI}}(\phi,w)$更新参数$w$；

end for

if 请求次数%5 == 0 then

更新目标策略网络的目标网络参数$\bar{\theta}=\theta$。

end if

为了有效训练TRCSA，本文使用深度Q网络（DQN, deep q-network）作为强化学习算法。DQN包含Q网络和目标网络两个核心组件，其中目标网络的动作值函数y计算如下

$$y=r_t+\eta\max_{a'}Q(\boldsymbol{s}',a') \tag{9}$$

其中，$\boldsymbol{s}'$和a'分别表示下一个系统状态和对应的动作，η表示学习折扣因子。然后，训练过程通过均方误差损失函数（式(10)）执行梯度下降

$$L_{\mathrm{PPO}}=E\left[\left(Q(\boldsymbol{s},a)-y\right)^2\right] \tag{10}$$

4　仿真结果与分析

为了全面评估提出的TRCSA性能，本节将其与3种典型的基准算法进行如下对比分析。

（1）DQN算法[34]：该算法无迁移学习机制，在目标域中从零开始训练充电站推荐策略。

（2）预训练DQN算法（Pre-DQN）[35]：该方法通过直接迁移源策略网络的中间层参数实现知识迁移，同时将源策略网络的输入层和输出层替换为适应目标策略网络的输入输出层。

（3）策略蒸馏算法[36]：如同TRCSA算法，该算法采用逐步迁移策略实现知识迁移，可视为无互信息的TRCSA算法。

4.1　实验环境设置

本文采用Voronoi图对充电站推荐区域Z进行空间划分。充电站推荐区域Z_n包括n个互不重叠的子区域，每个子区域内设置一个充电站。为了验证算法的跨域迁移能力，选取了具有不同充电站数量的区域Z_4和Z_6作为实验场景，区域Z_4与区域Z_6的迁移如图5所示。

仿真实验评估了Z_4和Z_6间的跨域迁移性能，其交通负载水平见表1，仿真详细参数设置见表2。表2中，μ_{a}表示区域内车辆请求充电的期望时间间隔（假设充电请求服从泊松分布）；μ_{d}表示车辆到达相同子区域内充电站的期望行驶时间；μ_{c}表示单位电量的期望充电时间。

4.2　跨域迁移的收敛性分析

当充电站推荐系统从区域Z_4向区域Z_6迁移时，系统的状态维度和动作数量增加。Z_4迁移到Z_6的奖励变化如图6所示，所有算法的奖励值随着训练次数的增加逐渐增加，但DQN算法的奖励值始终最小，这是因为DQN算法的初始参数是随机生成的，缺乏先验知识，需要从零开始学习策略网络的参数，进而奖励值的提升较为缓慢。即在无迁移机制的情况下，DQN无法在有限的训练步数内收敛至较好的性能。

Pre-DQN、TRCSA和蒸馏3种迁移算法的初始奖励值较高，且收敛速度较快。Pre-DQN直接迁移源策略网络的中间层，对于知识的利用有限。对于TRCSA和蒸馏算法，其嵌入编码器函数最初也是随机初始化的，因此初始奖励值并不比Pre-DQN高很多，但是它们收敛更快。特别是TRCSA，因采用了互信息对嵌入编码器进行更新，获得了最好的性能表现。

当充电站推荐系统从区域Z_6向区域Z_4迁移时，系统的状态维度和动作数量减少。Z_6迁移到Z_4的奖励变化如图7所示，随着训练次数的增加，4种算

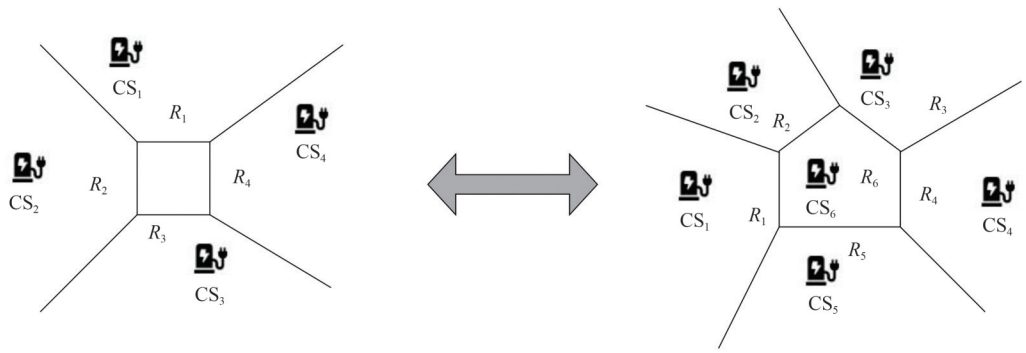


图5　区域Z_4与区域Z_6的迁移

表 1　　交通负载水平

时间段	Z_4 负载值	Z_6 负载值
6:00-7:00	2	1
7:00-9:00	3	2
9:00-16:00	2	1
16:00-19:00	3	3
19:00-21:00	2	1
21:00-次日 6:00	1	1

表 2　　仿真详细参数设置

参数	定义	值
η	学习折扣因子	0.1
γ	学习率	0.01
μ_a	充电请求间隔	360~600 s
μ_d	单位距离行驶时间	180 s
μ_c	单位电量充电时间	360 s

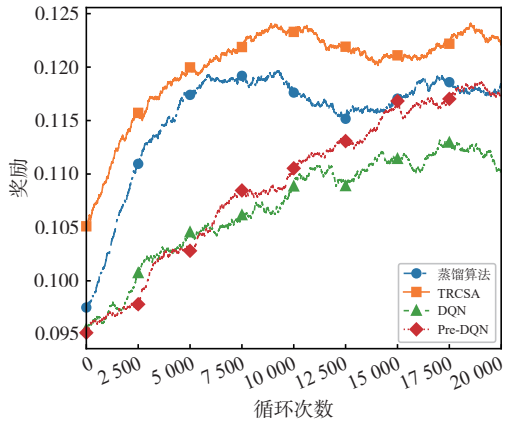


图 6　Z_4 迁移到 Z_6 的奖励变化

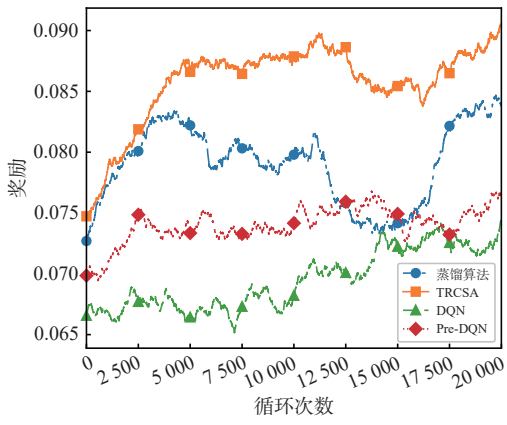


图 7　Z_6 迁移到 Z_4 的奖励变化

法在充电站推荐任务中获得的奖励逐步提升。同样，因为 DQN 算法没有迁移学习机制，其奖励值低于其他 3 种机制。TRCSA 和蒸馏算法在训练初期的奖励值更高，并具有更快的收敛速度。这是因为 TRCSA 和蒸馏算法利用嵌入编码器迁移源策略网络的中间特征，这种方式在迁移中更稳定且有效。进一步地，TRCSA 利用互信息避免了系统状态嵌入后的失真问题，利用互信息优化嵌入效果，更好地利用可迁移知识。

4.3　跨域迁移性能分析

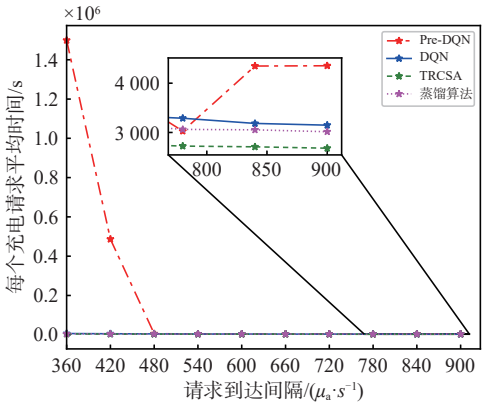
通过改变参数 μ_a 和 μ_d，分析从低维域 Z_4 向高维域 Z_6 迁移时充电站推荐模型的性能。低维域向高维域迁移的平均总充电时间如图 8 所示。首先，设定 μ_d 为 300 s，通过改变充电请求时间间隔 μ_a 分析性能，结果如图 8(a)所示。观察到，随着充电请求时间间隔的增加，4 种算法的总充电时间减少。μ_a 的增加意味着充电请求的频率降低，充电站的充电负载减少。所以，一方面所有算法都会尽量推荐就近充电站，使得行驶时间变小；另一方面，充电排队时间减少。从而，整体充电时间变少。

相对于其他算法，DQN 和 Pre-DQN 算法的平均总充电时间较长。因为 DQN 没有迁移机制，对目标区域的训练完全是从零开始，而没有利用源域的知识，在训练次数有限的情况下，并不能实现网络模型最优化。而 Pre-DQN 将源域的模型完整地迁移到目标域，当目标域和源域相差较大时，相比于模型随机初始化（DQN），这种迁移并不能达到更好的效果，例如，在 μ_a 值较小时，Pre-DQN 的性能比 DQN 还差。

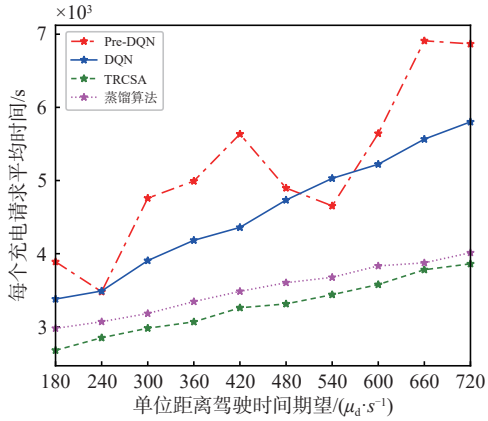
TRCSA 和蒸馏算法性能相对较好，也说明了迁移机制能够在有限的训练数据下，更快地优化模型。同时，观察发现，TRCSA 比蒸馏算法能够带来更好的性能。这两种算法都使用嵌入编码器对齐目标域的状态空间维度与源区域，并使用源策略网络获取中间特征。而 TRCSA 相比于蒸馏算法多了互信息机制，用于更新嵌入编码器参数。这表明互信息机制在处理不同维度的输入和输出时，能够加速迁移过程中的网络收敛。

接着，将 μ_a 设定为 600 s，通过改变 μ_d 分析迁移性能，如图 8(b)所示。μ_d 的增加意味着行驶时间增加，因此电动汽车平均总充电时间增加。

从图 8 中观察到，Pre-DQN 的波动较大，这是因为 Pre-DQN 机械地迁移源策略网络的中间网络层参数，在源域环境和目标域环境较相似的情况下得到较好的结果，而当环境相差较大时，迁移效果大打折扣。这也是 Pre-DQN 的充电时间具有较大的波动（图 8(b)）的原因。相反，通过调整 ε，进行渐进迁移（TRCSA 和蒸馏算法）的方法保证了迁移的稳定性和有效性。



(a) 平均总充电时间vs. μ_a



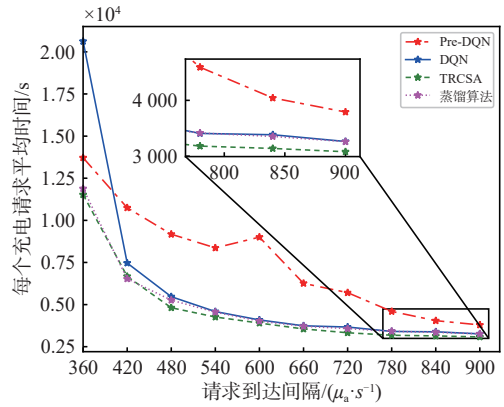
(b) 平均总充电时间vs. μ_d

图8　低维域向高维域迁移的平均总充电时间

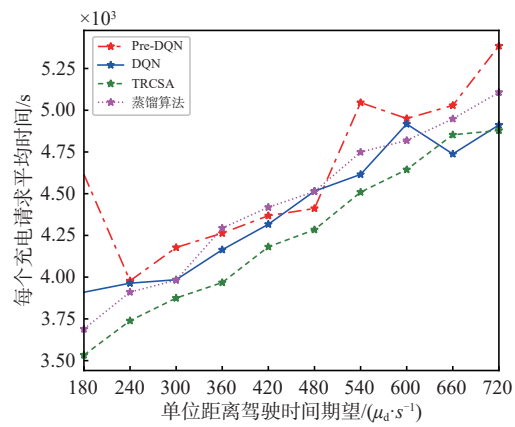
当充电站推荐系统从充电站较多的区域向较少的区域迁移时（从Z_6向Z_4迁移），系统的状态维度和动作数量减少。高维域向低维域迁移的平均总充电时间如图9所示。图9(a)所示为μ_a对各个算法性能的影响。随着μ_a的增大，所有算法的平均总充电时间减少，原因同上面分析。

对比图9(a)和图8(a)，可以发现各个算法的表现相似。一方面，TRCSA和蒸馏算法在总体上要好于Pre-DQN和DQN算法；另一方面，在高负载下（$\mu_a < 540$ s），TRCSA和蒸馏算法表现更出色。这里观察到，Pre-DQN在很高的负载下（$\mu_a < 420$ s），其获得了最高的奖励。仅表明此时目标域的环境和源域环境较相似，使得Pre-DQN表现出色。

图9(b)所示为μ_d对算法的影响。随着μ_d的增加，车辆到充电站的行驶时间增加，所有算法的平均总充电时间增加。Pre-DQN表现出较大波动性，这主要源于其仅通过简单的参数复制进行网络初始化，而未充分考虑源域环境与目标域环境之间的特征关联性和分布差异。当源域环境与目标域环境存在较大差异时，这种简单的参数迁移策略无法有效适应目标域的特征分布，导致算法性能呈现不稳定性。TRCSA因采用了渐进迁移和互信息机制，总体上取得了最优的性能。但同时观察到，高维向低维的迁移过程中，DQN也表现出较好的性能，特别是在轻负载的环境中DQN取得了不错的效果。这是因为在低维度下，网络更容易训练（因为状态空间状态量和动作空间动作量小），所以即使对一个网络重新进行训练，在有限的训练次数下，DQN也能收敛至较好的性能。值得注意的是，当μ_d为660 s时，DQN表现优于TRCSA，这主要是较大的行驶时间参数导致算法倾向于推荐距离较近的充电站。在这种情况下，充电请求位置的随机性造成了性能评估的波动。



(a) 平均总充电时间vs. μ_a



(b) 平均总充电时间vs. μ_d

图9　高维域向低维域迁移的平均总充电时间

5　结束语

本文针对车联网中充电站推荐系统的跨域迁移挑战，提出了基于迁移强化学习的跨域充电站推荐算法TRCSA。该算法引入编码器对齐不同区域的维度，有效解决了源域与目标域在状态空间和动作空间维度

不一致的问题。同时，TRCSA采用互信息保证知识迁移的有效性。实验结果表明，所提TRCSA展现出优异的迁移性能。特别是在从低维域向高维域的迁移过程中，TRCSA能够充分利用低维网络中的先验知识，显著提升训练效率，有效缓解了高维域中的维度灾难问题。在未来研究中，将探讨跨应用迁移问题，例如，将充电站推荐系统迁移至网约车系统。

参考文献：

[1] 熊凯, 冷甦鹏, 张可, 等. 车联雾计算中的异构接入与资源分配算法研究[J]. 物联网学报, 2019, 3(2): 20-27.
XIONG K, LENG S P, ZHANG K, et al. Research on heterogeneous radio access and resource allocation algorithm in vehicular fog computing[J]. Chinese Journal on Internet of Things, 2019, 3(2): 20-27.

[2] 罗梓珲, 江呈羚, 刘亮, 等. 基于深度强化学习的智能车间调度方法研究[J]. 物联网学报, 2022, 6(1): 53-64.
LUO Z H, JIANG C L, LIU L, et al. Research on deep reinforcement learning based intelligent shop scheduling method[J]. Chinese Journal on Internet of Things, 2022, 6(1): 53-64.

[3] WANG X, WANG S, LIANG X X, et al. Deep reinforcement learning: a survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(4): 5064-5078.

[4] 张启阳, 陈希亮, 曹雷, 等. 深度强化学习中的知识迁移方法研究综述[J]. 计算机科学, 2023, 50(5): 201-216.
ZHANG Q Y, CHEN X L, CAO L, et al. Survey on knowledge transfer method in deep reinforcement learning[J]. Computer Science, 2023, 50(5): 201-216.

[5] HARIZAJ M, BISHA I, BASHOLLI F. IOT integration of electric vehicle charging infrastructure[J]. Advanced Engineering Days (AED), 2023, 6: 152-155.

[6] HESTER T, VECERIK M, PIETQUIN O, et al. Deep Q-learning from demonstrations[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2018: 3223-3230.

[7] ALKAABNEH F, DIABAT A, GAO H O. Benders decomposition for the inventory vehicle routing problem with perishable products and environmental costs[J]. Computers & Operations Research, 2020, 113: 104751.

[8] WANG H, TAO C, QI J, et al. Avoiding negative transfer for semantic segmentation of remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 4413215.

[9] CHEN Z H, CHEN C, CHENG Z W, et al. Selective transfer with reinforced transfer network for partial domain adaptation[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 12703-12711.

[10] WU Z H, PAN S R, CHEN F W, et al. A comprehensive survey on graph neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 32(1): 4-24.

[11] 贺晓, 王文学. 基于图嵌入编码形态信息的非均匀多任务强化学习方法[J]. 计算机应用研究, 2024, 41(4): 1022-1028.
HE X, WANG W X. Method for inhomogeneous multi-task reinforcement learning based on morphological information encoding by graph embedding[J]. Application Research of Computers, 2024, 41(4): 1022-1028.

[12] GUPTA A, DEVIN C, LIU Y X, et al. Learning invariant feature spaces to transfer skills with reinforcement learning[C]//Proceedings of the International Conference on Learning Representations. Toulon: ICLR Press, 2017: 1456-1469.

[13] GUO Q L, XIN S J, SUN H B, et al. Rapid-charging navigation of electric vehicles based on real-time power systems and traffic data[J]. IEEE Transactions on Smart Grid, 2014, 5(4): 1969-1979.

[14] WU H, PANG G K H, CHOY K L, et al. A scheduling and control system for electric vehicle charging at parking lot[C]//Proceedings of the 2017 11th Asian Control Conference (ASCC). Piscataway: IEEE Press, 2017: 13-18.

[15] LIU S H, XIA X, CAO Y, et al. Reservation-based EV charging recommendation concerning charging urgency policy[J]. Sustainable Cities and Society, 2021, 74: 103150.

[16] TIAN Z Y, JUNG T, WANG Y, et al. Real-time charging station recommendation system for electric-vehicle taxis[J]. IEEE Transactions on Intelligent Transportation Systems, 2016, 17(11): 3098-3109.

[17] PAREEK S, SUJIL A, RATRA S, et al. Electric vehicle charging station challenges and opportunities: a future perspective[C]//Proceedings of the 2020 International Conference on Emerging Trends in Communication, Control and Computing (ICONC3). Piscataway: IEEE Press, 2020: 1-6.

[18] BI X W, WANG R H, JIA Q. On the speed-varying range of electric vehicles in time-windowed routing problems with en-route partial re-charging[J]. IEEE Transactions on Consumer Electronics, 2024, 70(1): 3650-3657.

[19] WANG S Y, BI S Z, ZHANG Y A. Reinforcement learning for real-time pricing and scheduling control in EV charging stations[J]. IEEE Transactions on Industrial Informatics, 2021, 17(2): 849-859.

[20] XING Q, XU Y, CHEN Z, et al. A graph reinforcement learning-based decision-making platform for real-time charging navigation of urban electric vehicles[J]. IEEE Transactions on Industrial Informatics, 2023, 19(3): 3284-3295.

[21] ZHANG W J, LIU H, WANG F, et al. Intelligent electric vehicle charging recommendation based on multi-agent reinforcement learning[C]//Proceedings of the Web Conference 2021. New York: ACM Press, 2021: 1856-1867.

[22] CUI F F, LIN X X, ZHANG R N, et al. Multi-objective optimal scheduling of charging stations based on deep reinforcement learning[J]. Frontiers in Energy Research, 2023, 10: 1042882.

[23] 丁瑞金, 高飞飞, 邢玲. 基于深度强化学习的物联网智能路由策

略[J]. 物联网学报, 2019, 3(2): 56-63.
DING R J, GAO F F, XING L. Intelligent routing strategy in the Internet of things based on deep reinforcement learning[J]. Chinese Journal on Internet of Things, 2019, 3(2): 56-63.

[24] WANG K, WANG H X, YANG Z H, et al. A transfer learning method for electric vehicles charging strategy based on deep reinforcement learning[J]. Applied Energy, 2023, 343: 121186.

[25] FOROOTANI A, RASTEGAR M, ZAREIPOUR H. Transfer learning-based framework enhanced by deep generative model for cold-start forecasting of residential EV charging behavior[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(1): 190-198.

[26] CAI Y H, LI M X, CHENG Y F. Small sample load forecasting of electric vehicle charging stations based on transfer learning[C]// Proceedings of the 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT). Piscataway: IEEE Press, 2024: 2041-2045.

[27] FERNÁNDEZ F, VELOSO M. Probabilistic policy reuse in a reinforcement learning agent[C]//Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems. New York: ACM Press, 2006: 720-727.

[28] ZHUANG F Z, QI Z Y, DUAN K Y, et al. A comprehensive survey on transfer learning[EB]. 2020.

[29] TAYLOR M E, STONE P, LIU Y X. Transfer learning via intertask mappings for temporal difference learning[J]. Journal of Machine Learning Research, 2007, 8: 2125-2167.

[30] TAYLOR M E, STONE P. Cross-domain transfer for reinforcement learning[C]//Proceedings of the 24th International Conference on Machine Learning. New York: ACM Press, 2007: 879-886.

[31] LIAN R Z, TAN H C, PENG J K, et al. Cross-type transfer for deep reinforcement learning based hybrid electric vehicle energy management[J]. IEEE Transactions on Vehicular Technology, 2020, 69(8): 8367-8380.

[32] NIPU A S, LIU S M, HARRIS A. Enabling multi-agent transfer reinforcement learning via scenario independent representation[C]// Proceedings of the 2023 IEEE Conference on Games (CoG). Piscataway: IEEE Press, 2023: 1-8.

[33] NOGUCHI H, ISODA T, ARAI S. Shared trained models selection and management for transfer reinforcement learning in open IoT[C]//Proceedings of the 2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC). Piscataway: IEEE Press, 2021: 2170-2176.

[34] KHALID M, WANG L, WANG K Z, et al. Deep reinforcement learning-based long-range autonomous valet parking for smart cities[J]. Sustainable Cities and Society, 2023, 89: 104311.

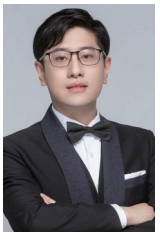
[35] FANG X, GONG G C, LI G N, et al. Cross temporal-spatial transferability investigation of deep reinforcement learning control strategy in the building HVAC system level[J]. Energy, 2023, 263: 125679.

[36] ZHU Z D, LIN K X, JAIN A K, et al. Transfer learning in deep reinforcement learning: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11): 13344-13362.

[作者简介]

林海（1976– ），男，博士，武汉大学国家网络安全学院副教授，主要研究方向为计算机网络。

赵家仪（2000– ），女，武汉大学国家网络安全学院硕士生，主要研究方向为车联网、强化学习、数据融合。



曹越（1984– ），男，博士，武汉大学国家网络安全学院教授、系主任，主要研究方向为网络安全。

苏航宇（2002– ），男，武汉大学国家网络安全学院硕士生，主要研究方向为车联网、迁移学习、人工智能。

王丽园（1980– ），女，中交第二公路勘察设计研究院有限公司正高级工程师、首席研究员，主要研究方向为公路智慧交通技术。