基于知识迁移的 Ant-Q 算法

王雪松,潘 杰,程玉虎
(中国矿业大学信息与电气工程学院,江苏徐州 221116)

摘 要: 常规 Ant-Q 算法计算复杂度随问题的规模呈现出阶乘级的增长,极大地抑制了算法的收敛速度,同时其仅关注单一任务本身,使得求出的解不具有可重用性,在处理一系列相关联任务时效率较低.为此,提出一种基于知识迁移的 Ant-Q 算法,通过贝叶斯理论分析源任务与目标任务的相似率,并以此为权值确定各源任务的迁移样本数,然后将各源任务样本按迁移价值降序排列,筛选出有效迁移样本,指导 Agent 快速做出合理决策.在 att532 旅行商问题上的仿真结果表明,知识迁移能够有效降低目标任务的学习难度,从而快速找到问题的最优解.

关键词: 知识迁移; Ant-Q 算法; 贝叶斯理论; 样本筛选; 旅行商问题
中图分类号: TP18 **文献标识码:** A **文章编号:** 0372-2112 (2011) 10-2359-07

Ant-Q Algorithm Based on Knowledge Transfer

WANG Xue-song, PAN Jie, CHENG Yu-hu
(*School of Information and Electrical Engineering, China University of Mining and Technology, Xuzhou, Jiangsu* 221116)

Abstract: The computational complexity of traditional Ant-Q algorithm shows factorial growth with the scale of the studied problem, which greatly reduces the convergence speed. Moreover, the traditional Ant-Q algorithm only focuses on a single task, therefore, the solution for the task cannot be reusable and the algorithm will handle a series of related tasks with low efficiency. In order to improve the convergence speed, a kind of Ant-Q algorithm based on knowledge transfer is proposed. At first, the similarity between each source task and a target task is computed according to the Bayesian theory. Then the obtained similarities are viewed as the weights to determine the number of samples transferred from every source task. In the third step, the samples from source tasks are listed in a descending order according to its transfer values and some valid samples are selected. In this way, the selected samples can guide an Agent to make a rational decision quickly. Simulation results involving a traveling salesman problem att532 illustrate that the knowledge transfer technology can effectively reduce the difficulty of learning a new task and quickly find an optimal solution.

Key words: knowledge transfer; ant-Q algorithm; bayesian theory; sample selection; traveling salesman problem

1 引言

Ant-Q 算法是将利用仿生学技术的蚁群算法与基于值迭代的 Q 学习相融合的一类优化算法,由 Gambardella 与 Dorigo 提出[1].该算法不但继承了蚁群算法分布式计算、信息正反馈以及启发式搜索等优点,还兼具强化学习的试错法与延时回报等行为心理学机制[2],在处理旅行商问题(Traveling Salesman Problem, TSP)[3]、移动机器人路径规划[4]、核燃料重载[5]、配水灌溉网络[6]以及流程规划[7]等方面有着极其重要的应用.然而,当待处理问题规模较大时,Ant-Q 算法的时间复杂度将随问题规模呈现出阶乘级的增长,极大地抑制了算法的收敛速度.另外,Ant-Q 算法在本质上属于蚂蚁系统的延伸,因而蚁群算法难以解决的局部最优问题,也同时制约着 Ant-Q 算法的进一步发展.

近年来,针对 Ant-Q 算法在处理大规模问题时收敛速度较慢的缺陷,许多研究人员对其进行了改进.Wang[8]提出一种 Ant(λ)算法,将资格迹机制引入到信息素的局部更新中,并融合时间差分(Temporal Difference, TD)与蒙特卡罗方法,使算法能够及时获得反向传播的延时回报.同样地,为解决时间信用分配问题,Lee[9]把 TD 误差融合到 Ant-Q 算法中,建立了多蚁群交互式强化学习模型(Multi Colony Interaction Ant Reinforcement Learning Model, MCIARLM),其在每个学习步均能预

收稿日期:2010-10-21;修回日期:2011-03-16
基金项目:国家自然科学基金(No.60804022, No.60974050, No.61072094);教育部新世纪优秀人才支持计划(No.NCET-08-0836, No.NCET-10-0765);霍英东教育基金会青年教师基金(No.121066);江苏省自然科学基金(No.BK2008126).

测当前及下一时刻的状态,并根据预测结果实时更新 Q 值函数逼近器,同时在不同群体间应用精英策略,通过正负作用机制以实现算法在探索与利用方面的平衡. Ant(λ)和 MCIARLM 能够在不同程度上提高 Ant-Q 的收敛速度,但是它们仅关注单一任务本身,使得最终结果均不具有再利用性,即无法将其应用到相似的任务中以降低学习难度.

知识迁移(Knowledge Transfer,KT)技术能够通过以往所学任务的知识来提高当前相关任务的学习性能[10],具有模拟人类迁移思维能力的显著特点.与行为迁移中单纯的迁移策略或最优动作不同[11],KT 技术旨在发现任务间的内在联系,即更重视相似性规律的总结,并将其提炼为知识的形式予以迁移.KT 技术本身不具有独立性,通常与其他机器学习方法相结合,以改善相应的算法性能.本文将 KT 技术应用到 Ant-Q 学习中,通过源任务与目标任务间的相似性衡量,确定最有价值的迁移样本,以指导 Agent 做出高效的决策.采用数据集 TSPLIB 中的 att532 旅行商问题进行仿真研究,以验证算法的合理性和有效性.

2 基于知识迁移的 Ant-Q 算法

Ant-Q 算法的基本思想是用 AQ 值来代替原蚁群算法中的信息素,用 ΔAQ 表示立即回报,从而可以采用 Q 学习中值迭代的方法来更新 AQ 值,以找到最优策略 π^*,其迭代公式为[1]:

$$AQ(s,s') = (1-\alpha)AQ(s,s') + \alpha[r_{ss'} + \gamma \max_{z\in J_l(s')} AQ(s',z)] \tag{1}$$

其中,$0<\alpha<1$ 为学习因子,表征学习速度,其值越大,学习速度越快,然而过大的 α 值会产生振荡,导致系统的不稳定;$0<\gamma<1$ 为折扣因子,表征 Agent 在学习过程中的远视程度,其值越小,越重视眼前利益;s 和 s' 分别表示当前状态与下一状态,对于 TSP 问题而言,专指节点城市;$J_l(s')$为第 l 只蚂蚁位于状态 s' 时的下一步可选状态集合;$r_{ss'}$为立即回报,定义为[1]:

$$r_{ss'} = \Delta AQ(s,s') = \begin{cases} R/L_l, & \text{若第 } l \text{ 只蚂蚁经过状态}(s,s') \\ 0, & \text{否则} \end{cases} \tag{2}$$

其中,L_l 为第 l 只蚂蚁爬行路线的总长度,R 为回报强度系数.由式(2)可以知道,$r_{ss'}$虽称为立即回报,事实上并非立即获得,只有在进行一次完整的寻径后方可更新,且同一条路径上的不同状态组(s,s')具有相同的立即回报值.

设任务空间 $\Gamma=(S_1,S_2,\cdots,S_n,T)$,包含 n 个已学习的源任务 $S_k(k=1,2,\cdots,n)$与 1 个待求的相关目标任务 T,S_k 与 T 中分别含有 m 与 t 个样本.通常情况下,对 T 采集样本的代价较高,因而有 $t\ll m$,且 t 的数目常不足以训练 T 使其快速收敛.另一方面,对于 T 而言,能够利用已学习的 n 个 S_k,并从中提取所需样本,比起从当前环境中重新采样将付出更少的代价.因而,考虑采用 KT 技术,辅助 Ant-Q 算法进行学习,以求提高目标任务 T 的学习速度.KT 的映射关系描述为:

$$\hat{S}\times\hat{T}\to H \tag{3}$$

其中,$\hat{S}$ 与 $\hat{T}$ 分别对应源任务与目标任务样本集,H 为最优迁移样本集.

KT 技术的要点是避免负迁移的发生.事实上,并非任意形式的迁移都是有效的,亦并非所提炼出作为知识的数据或规则越多越好,无价值或无关联的知识只会使目标任务的性能变得更差,即产生负迁移.如何避免负迁移的发生是一项关键技术.针对这个问题,本文将基于知识迁移的 Ant-Q 算法分为 3 个阶段:迁移阶段Ⅰ、迁移阶段Ⅱ与 Ant-Q 学习阶段.算法的结构框图如图 1 所示.

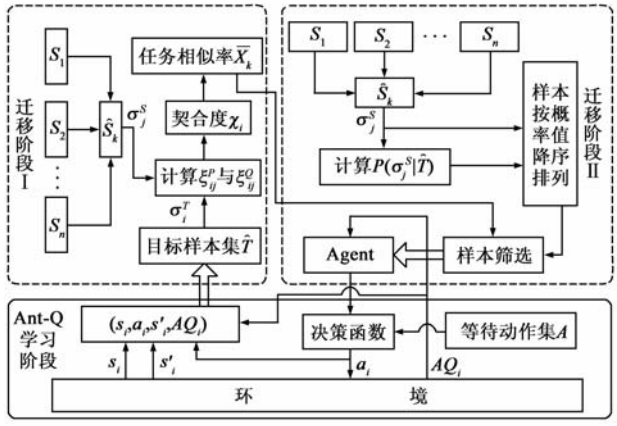

图1 基于知识迁移的Ant-Q算法

迁移阶段Ⅰ用以实现对源任务的任务空间筛选,对各源任务与目标任务间的相似率 $\bar{X}_k$ 进行评价.在该阶段,首先给出相似度评价指标 ξ_{ij}^P与ξ_{ij}^Q,以衡量 $\hat{S}$ 与 $\hat{T}$ 中样本间一对一的相似程度;其次,根据贝叶斯概率分析理论求出样本 $\sigma_i^T\in\hat{T}$ 相对于 S 的契合度 χ_i 以及 $\hat{S}_k$ 相对于 $\hat{T}$ 的似然度 X_k;最后,将 X_k 归一化后即得相似率 $\bar{X}_k$,并按比例从中迁移出所需的样本.阶段Ⅱ的工作是对源任务的样本空间进行筛选.与阶段Ⅰ相同,应首先给出样本迁移价值评价依据.本文通过条件概率 $P(\sigma_j^S|\hat{T})$来衡量 $\sigma_j^S\in\hat{S}_k$ 对 $\hat{T}$ 的匹配程度,并将结果按降序排列,前 $\bar{X}_k(m-t)$个样本即为最有价值的迁移样本.Ant-Q 学习阶段的核心部分是决策函数,其目标是从待选动作集 $A=\{a_v\in R^2\mid v=1,2,\cdots,q\}$中选择针对当前状态的合适动作,其中 $a_v=s'-s$ 表示待选动作,q 为待选动作的个数.由于 Agent 已从迁移阶段Ⅱ获得了充足的样本,如何利用这些样本使其发挥出最大的效用,即选

择何种学习策略,便显得尤为重要.一般来说,Q 学习的策略均可用于 Ant-Q 算法,而 ε-greedy 策略由于引入了随机因素使得算法在探索与利用间可以有效地进行调节,因而本文更倾向于采用该种策略. ε-greedy 策略与 Ant-Q 算法相结合的动作选择表达式为[1]:

$$s' = \begin{cases} \arg\max\limits_{z\in J_i(s)} \{[AQ(s,z)]^{\theta}\cdot[HE(s,z)]^{\beta}\}, & \varepsilon \leqslant \varepsilon_0 \\ \text{rand}, & \text{其他} \end{cases} \tag{4}$$

其中,$0\leqslant\varepsilon_0\leqslant 1$ 为随机动作控制常数,$0\leqslant\varepsilon\leqslant 1$ 是随机变量,$HE(s,z)$表征启发式信息,一般取为城市 s,z 间距离的倒数,θ 与 β 分别为相应的权重因子.可以看出,Agent 将以 ε_0 的概率选择使$[AQ(s,z)]^{\theta}[HE(s,z)]^{\beta}$ 最大的动作,而以$(1-\varepsilon_0)$的概率随机选择动作,从而保证了探索与利用的平衡.

3　任务空间与样本空间筛选

为提高迁移效率,首先需从任务空间 Γ 中选择出最适合迁移的源任务.这可以看作是已知含有少量样本 $\sigma_i^T(i=1,2,\cdots,t)$的目标样本集 $\hat{T}$,求各源任务样本集 $\hat{S}_k$ 的似然度问题.根据贝叶斯理论,有[12]:

$$P(S_k\mid\hat{T}) = \frac{P(S_k)P(\hat{T}\mid S_k)}{\sum\limits_{v=1}^{n}P(S_v)P(\hat{T}\mid S_v)} \propto P(S_k)P(\hat{T}\mid S_k) \tag{5}$$

考虑到 Ant-Q 算法样本的具体表达形式,有:

$$\begin{aligned} P(\hat{T}\mid S_k) &= \prod_{u=1}^{t} P(s_u,a_u,s_u',AQ_u\mid S_k) \\ &= \prod_{u=1}^{t} P(s_u'\mid s_u,a_u)P(AQ_u\mid s_u,a_u)P(S_k) \end{aligned} \tag{6}$$

综合式(5)与式(6),得:

$$P(S_k\mid\hat{T}) \propto \prod_{u=1}^{t} P(s_u'\mid s_u,a_u)P(AQ_u\mid s_u,a_u)P(S_k) \tag{7}$$

由式(7)可知,要求得各源任务 S_k 对目标样本集 $\hat{T}$ 的似然度,仅需已知状态转移模型 $P(s_u'\mid s_u,a_u)$与回报模型 $P(r_u\mid s_u,a_u)$即可.下面引入参数 ξ 与 χ 来描述这两个模型.

定义 1[12]　设 $\hat{S}$ 与 $\hat{T}$ 分别为源任务与目标任务的样本集,且有 $\sigma_j^S=(s_j,a_j,s_j',AQ_j)\in\hat{S}$, $\sigma_i^T=(s_i,a_i,s_i',AQ_i)\in\hat{T}$,则样本 σ_i^T 与 σ_j^S 具有状态相似度 ξ_{ij}^P与回报相似度 ξ_{ij}^Q:

$$\xi_{ij}^P = w_{ij}\cdot\exp\left(-\left(\frac{\|s_i'-s_i+a_j\|}{\delta_s}\right)^2\right) \tag{8}$$

$$\xi_{ij}^Q = w_{ij}\cdot\exp\left(-\left(\frac{|AQ_i-AQ_j|}{\delta_q}\right)^2\right) \tag{9}$$

其中,$w_{ij}=\dfrac{\exp(-\|(s_j,a_j),(s_i,a_i)\|^2/\delta_{sa}^2)}{\sum\limits_{k=1}^{m}\exp(-\|(s_k,a_k),(s_i,a_i)\|^2/\delta_{sa}^2)}$ 为相似权值.

分析式(8)可以得知,等号右侧第一项表征 σ_i^T 与 σ_j^S 的状态相似度,第二项表征输出相似度,衡量的是 σ_i^T 与 σ_j^S 中动作的作用效果.显然,两样本的相似度越高,ξ_{ij}^P 的值越大.同样地,式(9)中两样本 AQ 值愈接近,ξ_{ij}^Q值愈大.

接下来求出 $\hat{S}$ 中所有样本 $\sigma_j^S(j=1,2,\cdots,m)$对 $\sigma_i^T\in\hat{T}$ 的状态相似度 ξ_{ij}^P与回报相似度 ξ_{ij}^Q.容易得知,$\sum\xi_{ij}^P$与$\sum\xi_{ij}^Q$的值越大,说明 $\hat{S}$ 中与 σ_i^T 相似的样本越多,亦即对于模型 S 而言,样本 σ_i^T 具有较大的触发概率,因而有[12]:

$$P_{\hat{S}}(s_i'\mid s_i,a_i) \propto \sum_{j=1}^{m}\xi_{ij}^P \tag{10}$$

$$P_{\hat{S}}(AQ_i\mid s_i,a_i) \propto \sum_{j=1}^{m}\xi_{ij}^Q \tag{11}$$

基于以上原理,有如下定义.

定义 2[12]　设 $\hat{S}$ 与 $\hat{T}$ 分别为源任务与目标任务的样本集,且 $\sigma_i^T\in\hat{T}$,$\hat{S}$ 中各样本 σ_j^S 与 σ_i^T 的状态相似度与回报相似度分别为 ξ_{ij}^P与 ξ_{ij}^Q,则 σ_i^T 与模型 S 的状态契合度 χ_i^P 以及回报契合度 χ_i^Q 为:

$$\chi_i^P = P_{\hat{S}}(s_i'\mid s_i,a_i) = \frac{1}{Z_P}\sum_{j=1}^{m}\xi_{ij}^P \tag{12}$$

$$\chi_i^Q = P_{\hat{S}}(AQ_i\mid s_i,a_i) = \frac{1}{Z_Q}\sum_{j=1}^{m}\xi_{ij}^Q \tag{13}$$

其中,Z_P 与 Z_Q 为概率修正参数.

易知,将式(12)与(13)相乘,即得样本 $\sigma_i^T=(s_i,a_i,s_i',AQ_i)$与模型 S 的总契合度 $\chi_i=P(s_i,a_i,s_i',AQ_i\mid S)=P_{\hat{S}}(s_i'\mid s_i,a_i)\cdot P_{\hat{S}}(AQ_i\mid s_i,a_i)=\dfrac{1}{Z_PZ_Q}\left(\sum\limits_{j=1}^{m}\xi_{ij}^P\right)\cdot\left(\sum\limits_{j=1}^{m}\xi_{ij}^Q\right)$.这样,只需将 χ_i 的表达式代入式(7),即求得各源任务对目标任务的似然度 X_k[12]:

$$X_k = P(S_k\mid\hat{T}_\sigma) = \frac{1}{Z_X}\prod_{u=1}^{t}\chi_u\cdot P(S_k) \tag{14}$$

其中,$P(S_k)$为模型 S_k 的先验概率,Z_X 为似然度修正项.

将各源任务的似然度 X_k 归一化后,得到 S_k 与 T 的相似率 $\overline{X}_k$.由于 $t<<m$,因而需从各源任务样本集中迁移出共$(m-t)$个样本补足目标任务的样本数.容易想到,迁移的样本数应与各源任务的相似率成正比,这样,从各 $\hat{S}_k$ 中分别迁移$\overline{X}_k(m-t)$个样本是一种合理的选择.下面将说明如何从 $\hat{S}_k$ 的 m 个样本中筛选出 $\overline{X}_k(m-t)$个进行迁移.

与之前的分析类似,首先应求出样本 $\sigma_j^S\in\hat{S}_k$ 属于

目标任务 T 的概率[12]:

$$P(\sigma_j \mid \hat{T}) = P_{\hat{T}}(s_j' \mid s_j, a_j) \cdot P_{\hat{T}}(AQ_j \mid s_j, a_j)$$

$$= \frac{1}{Z_P Z_Q}\left(\sum_{i=1}^{t} \xi_{ji}^{P}\right) \cdot \left(\sum_{j=1}^{t} \xi_{ji}^{Q}\right) \tag{15}$$

其中,ξ_{ji}^{P}与ξ_{ji}^{Q}分别表示σ_j^S与σ_i^T的相似度,其定义可参考式(8)与(9)推出.然后,将求出的 $P(\sigma_j \mid \hat{T})$按照降序排列,选出其中前 $\overline{X}_k(m-t)$个即为从源任务集 $\hat{S}_k$ 中迁移的样本.

4 仿真研究

为探讨所提算法的有效性,针对 TSP 问题进行仿真研究.所谓 TSP 问题,即对于给定数量 p 的城市,为旅行商规划出一条最短的路径,要求该路径能访问到所有城市,并且对于每个城市仅访问一次,如图 2 所示.该问题属于 NP - 难问题,其计算复杂度为 $O((p-1)!)$,当城市规模 p 较大时,无法用传统算法予以求解,一般的 Ant-Q 算法虽能用于解决该问题,然而学习时间较长.本文所提算法旨在通过知识迁移的方法提高其学习速度,下面对其进行仿真验证.

4.1 单源迁移 TSP 问题

采用国际通用测试集 TSPLIB 中的 att532 问题进行研究,该问题为美国 532 个主要城市的布局.为保证任务间的相似性,源任务与目标任务均随机选择其中的 400 个城市,如图 2 所示,其中(a)为 att532 问题的城市布局,(b)为目标任务 T,(c) - (f)为已经学习完的源任务 $S_k(k=1,2,3,4)$,图中所示路径为其学习结果.首先考察单源迁移的效果,即选出与目标任务匹配程度最高的源任务,并仅从该任务中选取样本进行迁移.

基于单源迁移的 Ant-Q 算法(Single Source Knowledge Transfer-Based Ant-Q, SSKT-AQ),其参数设定分为 Ant-Q 学习阶段与迁移阶段两个部分.其中学习阶段待设定的参数有 AQ 值权重因子θ、启发式信息权重因子β、学习因子 α、最大迭代次数 $N_{\max}$、折扣因子 η、随机动作控制常数 ε_0 以及蚂蚁个数 d;迁移阶段需要设定的参数为高斯核宽度 δ_s、δ_q、δ_{sa},契合度概率修正项 Z_P、Z_Q 以及似然度修正项 Z_X.无迁移的 Ant-Q 算法(No Transfer Ant-Q, NTAQ)仅须设定学习阶段的参数.表 1 给出了算法的参数设置情况.

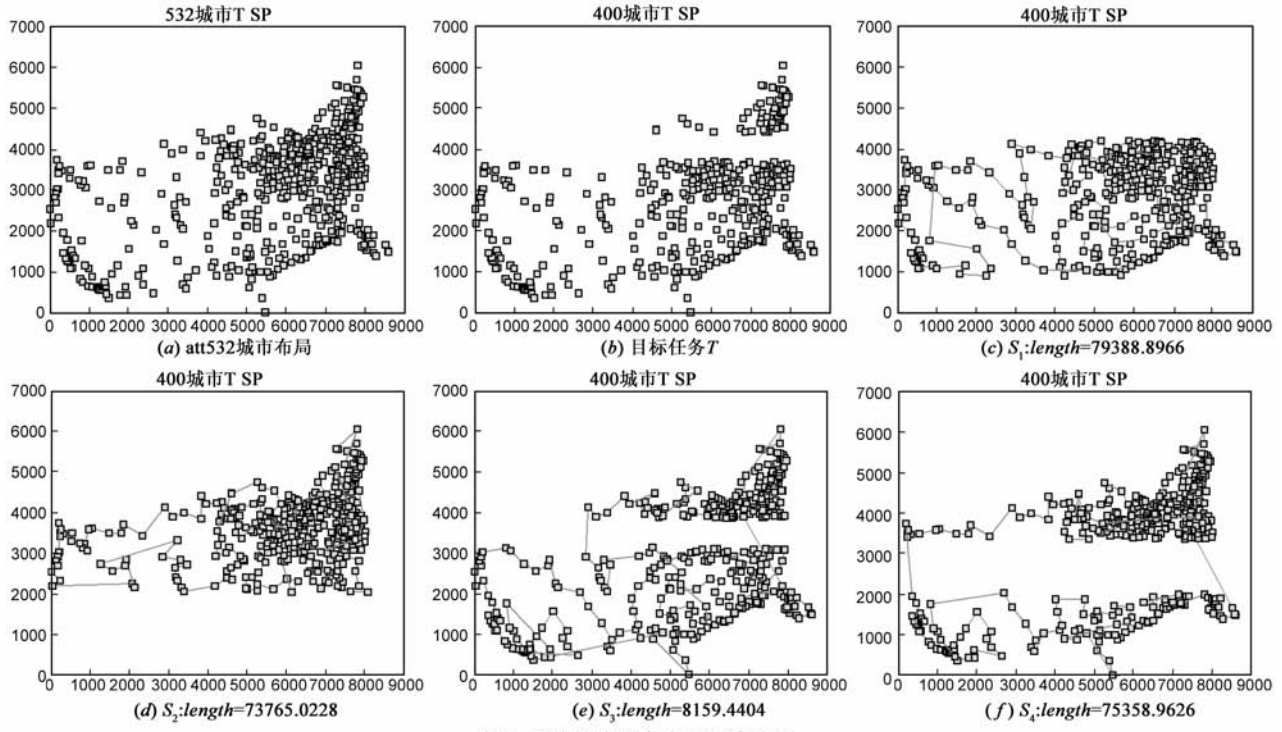

图2 源任务与目标任务城市布局

表 1 参数设置

Ant-Q 算法类型	Ant-Q 学习阶段							迁移阶段					
	θ	β	α	γ	ε_0	d	N_{max}	δ_s	δ_{sa}	δ_q	Z_P	Z_Q	Z_X
SSKT-AQ	1	2	0.5	0.95	0.1	4	100	10^3	10^3	10^{-8}	1	1	1
NTAQ	1	2	0.5	0.95	0.1	4	100	/	/	/	/	/	/

表 1 中,Ant-Q 学习阶段参数设置与常规 Ant-Q 算法相同,不再赘言,需要说明的是迁移阶段的设置.由于 SSKT-AQ 算法对高斯核宽度 δ_s、δ_q 和 δ_{sa}的取值较为敏感,一旦设置不当,会对迁移阶段产生较大影响,导致部分迁移样本失效.考察式(8)与(9),对于等号右侧的第二项而言,当 $\| s_i', s_i + a_j \| >> \delta_s$ 以及 $| AQ_i - AQ_j |$

$>> \delta_{sa}$时,有 $\xi_{ij}^{P} \to 0, \xi_{ij}^{Q} \to 0$;当 $\| s_i', s_i + a_j \| << \delta_s$ 与 $|AQ_i - AQ_j| << \delta_{sa}$时,有 $\xi_{ij}^{P} \to 1, \xi_{ij}^{Q} \to 1$.这两种情形均给数据的后续处理带来极大麻烦,为避免这个问题,应使 δ_s、δ_q 和 δ_{sa}与其相应的分子项拥有相同或相近的数量级,故此处设置 $\delta_s = 10^3, \delta_{sa} = 10^3, \delta_q = 10^{-8}$.

对于 SSKT-AQ 算法,其基本思路是通过已知的 t 个少量目标样本(此处设 $t = 0.2p^2$),来衡量各源任务 S_k 的迁移价值,即相似率 $\bar{X}_k$,仅将相似率最大的任务作为迁移任务,并从中筛选出$(m - t)$个样本迁移给目标任务 T.由于具有了充足的样本,处于 Ant-Q 学习阶段的 Agent 能够充分利用样本做出正确的决策,以实现高效学习的目的.NTAQ 算法则是从 t 个样本出发,在 Agent 与环境的交互中不断探索学习,以期寻到最优路径.图 3给出了 NTAQ 算法以及从各 S_k 迁移的 SSKT-AQ 算法所寻最优路径的收敛曲线,表 2 与图 4 给出了从不同源任务迁移的 SSKT-AQ 算法迁移指标对比.事实上,仅须找出相似率 $\bar{X}_k$ 最大的源任务即可,本实验考察所有源任务的迁移性能,旨在分析各迁移指标的有效性.

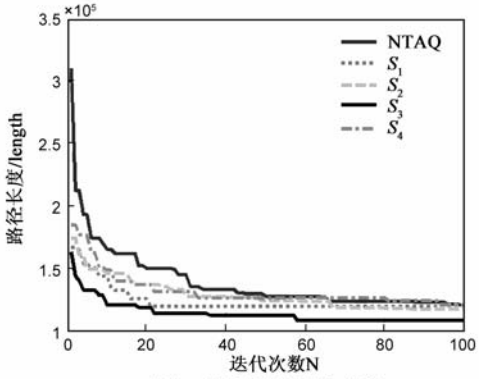

图3　最短路径收敛曲线

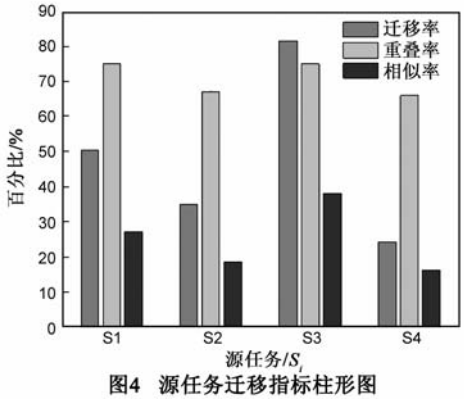

图4　源任务迁移指标柱形图

由图 3 的路径收敛曲线可以看出,NTAQ 算法由于缺乏知识,不得不在与环境的交互中不断积累样本,致使其无论是在算法的起始阶段还是迭代收敛阶段,性能均劣于其余 4 种不同类型的 SSKT-AQ 算法.另一方面也可知道,知识迁移的意义不仅在于提高了算法的初始性能,在新路径的选择与决策方面也起到至关重要的指导作用,从而能够加快算法迭代收敛的速度.虽然在图中能较为直观地看到各源任务的迁移性能,仍然有必要对迁移的效果给出定量的指标予以衡量.定义任务迁移率:

$$\Psi_k = 1 - \frac{A_k - A_0}{A_{NT} - A_0} \tag{16}$$

其中,A_k 为从源任务 S_k 迁移的 SSKT-AQ 算法收敛曲线与 x 轴所围图形的面积,A_{NT}为 NTAQ 算法收敛曲线与 x 轴所围面积,A_0 为 $y = length_b$ 与 x 轴所夹面积,$length_b$ 是目标任务 T 在各类算法运行下的最优路径.易知,Ψ_k 的取值范围是$(-\infty, 1]$.当 $\Psi_k > 0$ 时表示正迁移,其值越接近 1,迁移效果越好;当 $\Psi_k < 0$ 时为负迁移,其值越小,迁移效果越差;当 $\Psi_k = 0$ 时,有 $A_k = A_{NT}$,即迁移前后算法性能基本不变.各任务迁移率 Ψ_k 的计算结果见表 2,与该指标同列的还有城市重叠率 Ω 与任务相似率$\bar{X}$.城市重叠率的计算公式为 $\Omega = p_k/p$,其中 p_k 为源任务 S_k 与目标任务 T 重叠城市的数目.

表 2　源任务迁移指标对比

源任务	S_1	S_2	S_3	S_4
任务迁移率 Ψ	0.5016	0.3463	0.8163	0.2415
城市重叠率 Ω	0.7500	0.6700	0.7500	0.6600
任务相似率 $\bar{X}$	0.2713	0.1854	0.3809	0.1624

图 4 是与表 2 相对应的迁移指标柱形图,其较为直观地呈现了各任务各迁移指标间的区别与联系.观察图 4 与表 2 可以知道,由于各源任务与目标任务拥有较高的城市重叠率,少则 66%,多则 75%,使得从各 S_k 迁移的 SSKT-AQ 算法均呈现出良好的正迁移特性,分别为 $\Psi_1 = 0.5016, \Psi_2 = 0.3463, \Psi_3 = 0.8163, \Psi_4 = 0.2415$.然而要说明的是,城市重叠率 Ω 这个指标仅表明了可供迁移的可能性,即较大的 Ω 有可能产生较好的 Ψ,而非必然产生.观察图 4 可知,源任务 S_1 与 S_3 的城市重叠率高于 S_2 与 S_4,同样地,其迁移率也高于 S_2 与 S_4,而 $\Omega_2 > \Omega_4$,亦有 $\Psi_2 > \Psi_4$;但是要看到,尽管 $\Omega_1 = \Omega_3 = 75\%$,$S_3$ 的迁移率 $\Psi_3 = 0.8163$ 却高出 $\Psi_1 = 0.5016$ 许多.这也说明了城市重叠率与最终的迁移效果并不具有密切相关性,仅可作为一项指标予以参考,只有通过样本匹配程度计算得出的任务相似率 $\bar{X}$,才是衡量迁移效果优劣的根本依据,这一点从图表中亦可以得到印证.

为消除随机性因素对算法的影响,共进行 10 次独立实验,表 3 给出了从各 S_k 迁移的 SSKT-AQ 算法与 NTAQ 算法的对比统计数据,其中 SSKT-AQ(S_1)表示基于源任务 S_1 迁移的 SSKT-AQ 算法.

观察表 3 中数据可以知道,除了给出任务相似率 $\bar{X}_k$ 之外,表中信息还包括最优路径长度 L_b 与收敛迭代

次数 N,这两项指标分别从精度与速度方面衡量了算法的优劣程度.源任务 S_3 拥有与目标任务最高的相似率 $\bar{X}_3 = 0.3809$,相应地,其获得的最优路径,无论是最小值 1.0324×10^5,最大值 1.0973×10^5 还是平均值 1.0603×10^5 均优于从其他 S_k 迁移的 SSKT-AQ 算法.另外,在收敛迭代次数方面,从 S_3 迁移的 SSKT-AQ 算法亦呈现出良好的性能,仅次于源自 S_1 的迭代效果.同时综合两方面指标来看,源自 S_1 的 SSKT-AQ 算法其最优路径平均值高达 1.1670×10^5,尽管收敛速度很快,但事实上,其过早地陷入了局部最优.

表 3 目标 TSP 任务的学习结果统计

Ant-Q 算法类型	任务相似率 $\bar{X}_k$	最优路径长度 $L_b/10^5$			收敛迭代次数 N		
		最小值	最大值	平均值	最小值	最大值	平均值
NTAQ	/	1.1652	1.2112	1.1801	63	97	71.3
SSKT-AQ(S_1)	0.2713	1.1478	1.1987	1.1670	18	41	34.8
SSKT-AQ(S_2)	0.1854	1.1245	1.1789	1.1466	49	77	65.2
SSKT-AQ(S_3)	0.3809	1.0324	1.0973	1.0603	31	62	49.3
SSKT-AQ(S_4)	0.1624	1.1542	1.2016	1.1745	47	84	64.5

4.2 多源迁移 TSP 问题

所谓多源迁移 Ant-Q 算法(Multi-Source Knowledge Transfer Ant-Q,MSKT-AQ),其基本思想是在求得各源任务对于目标任务的似然度 X_k 并将其归一化后,按相似程度的比例从各个源任务迁移样本,即从每个 S_k 迁移 $\bar{X}_k(m-t)$个样本,以补足目标任务 T 的训练样本数.为方便与 SSKT-AQ 算法进行对比,MSKT-AQ 算法的仿真仍采用图 2 所示地图,目标任务与源任务均不变,参数设定也与 SSKT-AQ 算法相同,具体可参见 4.1 节.

图 5 给出了 NTAQ、SSKT-AQ 与 MSKT-AQ 三类算法的仿真效果对比.可以看出,从算法寻路的精度来说,MSKT-AQ 算法所得路径长度 $length = 8.7976 \times 10^4$,远优于 NTAQ 算法的 1.1713×10^5 以及 SSKT-AQ 算法的 1.0579×10^5.事实上这是可以预见的,由于 MSKT-AQ 算法扩大了待迁移的任务空间与样本空间,使得从源任务筛选出的样本具有更高的迁移价值,因而能够更有效地指导 Agent 在 Ant-Q 学习阶段进行决策.表 4 是 NTAQ、SSKT-AQ 与 MSKT-AQ 三类算法 10 次实验下的统计数据.由表可知,MSKT-AQ 算法在寻找最优路径的速度方面亦具有优良的性能.

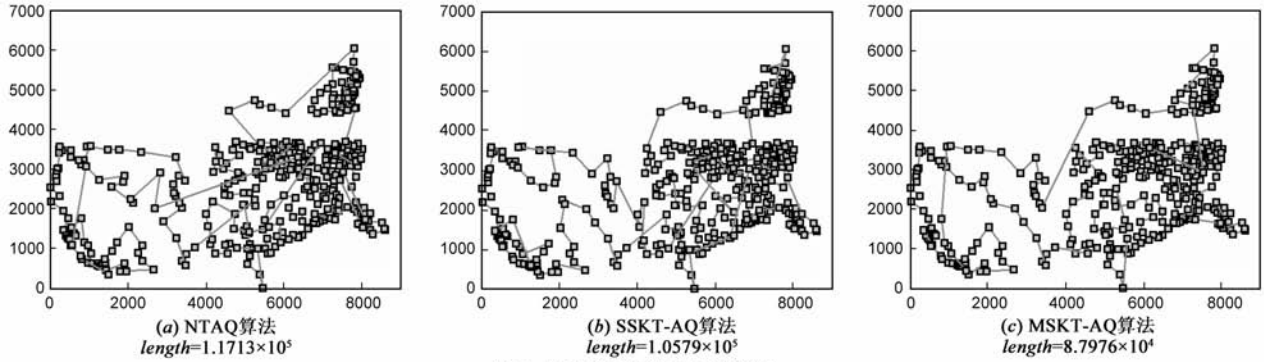

图5 目标TSP任务的寻径结果

表 4 目标 TSP 任务的学习结果统计

Ant-Q 算法类型	最优路径长度 $L_b/10^5$			收敛迭代次数 N		
	最小值	最大值	平均值	最小值	最大值	平均值
NTAQ	1.1652	1.2112	1.1801	63	97	71.3
SSKT-AQ	1.0324	1.0882	1.0603	31	62	49.3
MSKT-AQ	0.8534	0.9174	0.8847	15	33	26.7

5 结论

结合仿生学技术与行为心理学机制的 Ant-Q 算法,在处理路径规划、流程设计、路由管理等领域具有先天性优势,然而其随问题规模阶乘级递增的时间复杂度极大地降低了算法收敛速度,限制了其发展与应用,其解的不可重用性使算法在处理相关联任务时效率较低.为此,将知识迁移技术引入到 Ant-Q 算法,从源任务空间中筛选出较高相似率的源任务,通过贝叶斯概率分析理论,将样本按照迁移价值的高低降序排列,从而挑选出最适合迁移的样本补充给目标任务学习器,使 Agent 能够快速做出合理地决策,减少了学习时间,提高了算法收敛速度.其模拟人类迁移思维能力的智能性与处理相似问题的高效性是基于 KT 技术的 Ant-Q 算法所特有的优势.SSKT-AQ、MSKT-AQ 和 NTAQ 的仿真对比,说明了所提算法的合理性、快速性与高效性.

参考文献

[1] L M Gambardella, M Dorigo. Ant-Q: A reinforcement learning approach to the traveling salesman problem[A]. Proceedings of 12th International Conference on Machine Learning[C]. New York: ACM Press, 1995. 252 – 260.

[2] Y H Cheng, H T Feng, X S Wang. Actor-Critic learning based on adaptive importance sampling[J]. Chinese Journal of Electronics, 2010, 19(4): 583 – 588.

[3] H M Rais, Z A Othman, A R Hamdan. Improved dynamic ant colony system (DACS) on symmetric traveling salesman problem[A]. Proceedings of International Conference on Intelligent and Advanced Systems[C]. Piscataway: IEEE Inc, 2008. 43 – 48.

[4] N A Vien, N H Viet, S G Lee, T. H. Chung. Obstacle avoidance path planning for mobile robot based on Ant-Q reinforcement learning algorithm[J]. Lecture Notes in Computer Science, 2007, 4491: 704 – 713.

[5] L Machado, R Schirru. The Ant-Q algorithm applied to the nuclear reload problem[J]. International Journal of Annals of Nuclear Energy, 2002, 29(12): 1455 – 1470.

[6] C E Mariano, E Morelos. A multiple objective Ant-Q algorithm for the design of water distribution irrigation[A]. Proceedings of the Genetic and Evolutionary Computation Conference[C]. San Francisco: Morgan Kaufmann, 1999. 894 – 901.

[7] X J Liu, Z H Ni. Ant-Q algorithm based optimization approach for process planning[A]. Proceedings of the 8th IEEE International Conference on Control and Automation[C]. Piscataway: IEEE Inc., 2010. 620 – 623.

[8] X R Wang, T J Wu. The Ant(λ) ant colony optimization algorithm based on eligibility trace[A]. Proceedings of the IEEE International Conference on Systems, Man and Cybernetics [C]. Piscataway: IEEE Inc., 2003. 4065 – 4070.

[9] S G Lee, T C Chung. A reinforcement learning algorithm using temporal di? erence error in ant model[J]. Lecture Notes in Computer Science, 2005, 3512: 217 – 224.

[10] S J Pan, Q Yang. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22 (10): 1345 – 1359.

[11] 王皓,高阳,陈兴国. 强化学习中的迁移:方法和进展[J]. 电子学报, 2008, 36(12): 39 – 43.
Wang Hao, Gao Yang, Chen Xing-Guo. Transfer of reinforcement learning: the state of the art[J]. Acta Electronica Sinica, 2008, 36(12): 39 – 43. (in Chinese)

[12] A Lazaric, M Restelli, A Bonarini. Transfer of samples in batch reinforcement learning[A]. Proceedings of the 25th International Conference on Machine Learning[C]. New York: ACM Press, 2008. 544 – 551.

作者简介

王雪松　女,1974 年生于安徽泗县,2002 年获中国矿业大学控制理论与控制工程专业博士学位,现为中国矿业大学信息与电气工程学院教授,博士生导师. 主要研究方向为机器学习、复杂系统优化与控制、生物信息学等.
E-mail: wangxuesongcumt@163.com

潘　杰　男,1986 年生于江苏徐州,2009 年获中国矿业大学学士学位,现为中国矿业大学控制理论与控制工程专业博士研究生,研究方向为知识迁移学习.
E-mail: panjie1616@126.com

程玉虎　男,1973 年生于安徽淮南,2005 年获中国科学院自动化研究所控制理论与控制工程专业博士学位,现为中国矿业大学信息与电气工程学院教授,博士生导师. 主要研究方向为机器学习和智能系统等.
E-mail: chengyuhu@163.com