DOI:10.20033/j.1003-7241.(2025)03-0021-05.

基于深度强化学习的铁路线路方案生成方法研究

祖家伟[1]，王明生[1]，吕希奎[2]

(1.石家庄铁道大学 土木工程学院,河北 石家庄 050043;
2.石家庄铁道大学 交通运输学院,河北 石家庄 050043)

摘 要: 为解决铁路线路方案初步设计阶段的短时需求,结合深度强化学习理论提出智能生成线路方案方法。通过简化地形建立强化学习环境模型,以选线设计工作经验设置智能体探索状态与动作,关联铁路选线任务设立奖惩反馈,搭建PPO框架寻优并输出线路走向,在平纵规范的约束下进行路径分段与拟合线形,最终得到线路中心线方案。以某铁路验证了该方法的有效性,且智能线路方案较原始设计方案节省费用20.55%。基于深度强化学习的铁路智能选线方法,大幅度减少了线路方案初步设计时间,节约了工程费用,为后期定线提供优质的参考方案。

关键词: 智能选线;深度强化学习;PPO;线路走向;线形拟合
中图分类号:TP18 文献标识码:A 文章编号:1003−7241(2025)03−0021−05

Research on the Generation Method of Railway Line Scheme Based on Deep Reinforcement Learning

ZU Jia-wei[1], WANG Ming-sheng[1], LV Xi-kui[2]

(1. Shijiazhuang Tiedao University, School of Civil Engineering, Shijiazhuang 050043 China;
2. Shijiazhuang Tiedao University, School of Transportation, Shijiazhuang 050043 China)

Abstract: In order to solve the short-term requirements in the preliminary design stage of railway line scheme, a method for intelligent generation of line scheme is proposed based on deep reinforcement learning theory. By simplifying the terrain, a reinforcement learning environment model is established, the agent exploration state and action are set with the work experience of line selection design, the reward and punishment feedback is set up in association with the railway line selection task, the PPO framework is built to seek optimization and output the line direction, and the path segmentation and line shape are carried out under the constraints of the flat and longitudinal specification, and finally the line centerline scheme is obtained. The effectiveness of the method is verified by a railway, and the intelligent line scheme saves 20.55% compared with the original design scheme. The railway intelligent line selection method based on deep reinforcement learning greatly reduces the preliminary design time of the line scheme, saves the project cost, and provides a high-quality reference scheme for the later line fixing.

Keywords: smart line selection; deep reinforcement learning; PPO; line direction; fit the line

0 引言

选线工作作为铁路建设的龙头,是铁路设计中相关领域综合性较强的工作,对整体项目的工程费用多少、施工难度大小和后期运营与管理起决定性作用[1]。因受到地理环境等客观因素和专家经验与地方政策等主观因素的影响,选线设计工作难度大、决策风险高[2]。随着技术不断进步,计算机与铁路选线工作相结合的智能选线领域逐渐开展,生成满足规范且目标最优的方案供设计人员参考,以此减轻工作量加快设计速度[3]。1960年以来,大量的研究人员涌入并深入研究铁路智能选线方向。变分法提出两种数值积分的方法优化线路平面;网络优化法把起终点之间区域离散成若干网格,将其连接成路径以供计算出最优路径;动态规划法把起终点划分为若干区段,由后向前反向计算出最优解;遗传算法[4]将生物进化规律应用于线路平面优化,随机生成的线形经过运算得到最优方案;易思蓉[5]提出类规则知识表达模式与面向对象技术结合,使铁路选线领域的决策类知识运用更灵活;蒲浩等[6]针对于三维空间线形整体优化,以最小综合代价为目标进行全局搜索和多目标寻优。近期,深度强化学习(Deep Reinforcement Learning,DRL)在各大领域进展迅速,且算法频繁更新迭代,在游戏、商业和控制等领域已经实现大面积应用[7]。

本文结合深度强化学习理论,在复杂地形环境情况下,基于寻优速度及寻优效果建立铁路智能选线模型,通

*基金项目:河北省自然基金(E2021210027)
收稿日期:2024-04-23

过奖惩函数实现降低工程开销。针对智能选线模型输出的线路走向，提出一种自动分段方法，调整拟合公式以适应线形规范的需求，最终输出线路设计的初步方案。由实际方案与智能方案对比，线路大致走向趋于一致，验证了该方法的可行性，并展示该方法的时效性以及节省工程费用方面的优越性。

1 铁路智能选线模型

1.1 深度强化学习原理

强化学习按学习类型划分为基于价值(value based)和基于概率(policy based)两种，基于价值输出状态动作的值，该值的大小决定动作选择，其代表性算法为Q-Learning；基于概率输出各动作概率，依据各动作概率选择动作，其代表性算法为Policy Gradient。

基于Q-Learning框架之上通过神经网络近似值函数的方法提出DQN(Deep Q-Network)算法。DQN解决了状态和动作高维连续时计算量增长的难题，其中评估网络以状态s_i作为输入预测出价值函数$Q(s_i)$，目标网络以状态s_{i+1}作为输入预测出价值函数$Q(s_{i+1})$，利用迭代公式$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s'+a')-Q(s,a)]$，算出价值函数与目标函数之间的误差，通过反向传播优化网络参数，收敛后得出各状态下最优动作策略[8]。

由蒙特卡洛发展出的策略梯度(Policy Gradient)算法，不通过误差反向传播，而是通过观测信息选出动作反向传播，利用奖惩增加优质动作的选中概率，减少劣质动作的选中概率，这一过程没有误差产生[9]。但每组训练数据只能更新一次参数，为解决耗时过长等问题，OPENAI提出更简洁优化的算法框架PPO(Proximal Policy Optimization)[10]。PPO使用Actor-Critic框架，结合了基于价值和基于概率的特点，实现了在多个训练步骤中小批次更新，其中PPO迭代函数见文献[11]。

1.2 马尔科夫五元组

强化学习最基本的理论模型是马尔科夫决策过程(Markov Decision Process，MDP)。一般情况下，MDP模型的基本要素可以由马尔科夫五元组{S，A，R，P，γ}表示：S为环境所有的状态集合，指的是能够指导智能体做出决策的信息；A为探索的动作集合，指的是智能体于当前状态下可选择的动作集；R为回报函数，指的是智能体在状态s_i做出动作a_i获得的反馈，用于帮助智能体学习到最优的决策；P为状态转移概率，指的是在当前状态下转移到下一状态的概率，这跟探索策略的设置相关；γ为折扣因子，指的是在计算累计奖励时，考虑长远奖励影响的权重。针对本文研究的铁路选线决策问题，以马尔科夫五元组构建相关模型。

1.3 环境模型

用栅格法[12]简化地形建立深度强化学习所需的环境模型，供智能体交互并探索路径。本文以GDAL库读取数字高程模型(DEM)栅格数据，根据栅格个数和尺寸对应生成方格，并储存高程数据矩阵。自然保护区和地质不良区等铁路修建时不可通过的区域称为线路禁修区，将预先处理标注好的shp文件统一赋值于NoData处，供程序识别线路禁修区坐标，实现绕行。起终点以经纬度为单位，需手动输入，其中经纬度转换模型坐标伪代码如公式(1)所示。

$$X=\mathrm{int}((|x-start|)/unit)*unit+unit/2 \tag{1}$$

式中，x表示读取文件的边界坐标；start表示起点坐标；$unit$表示单元格单位长度；X表示转化后的方格中间点坐标。

通过读取以上数据文件生成多层网格环境模型。(X，Y)为网格中心坐标，H为网格高程数据。线路禁修区赋予属性Else，非线路禁修区赋予属性Danger。

1.4 探索状态-动作空间

智能体动作A设为：上，下，左，右，左上，左下，右上，右下；智能体所处状态S由网格中心坐标(X，Y)表示。状态变化由状态列表记录为Statelist=$\{s_1,s_2,\cdots,s_n\}$；动作变化由动作列表记录为Actionlist=$\{a_1,a_2,\cdots,a_n\}$。连续两次动作产生如图1所示的五种转角，其中135°和180°的动作序列在线路走向探索过程中输出的线路方案很难符合规范，增加不必要的探索时间[13]。加入矢量判断，使得前后动作关联，去除135°与180°的探索情况。

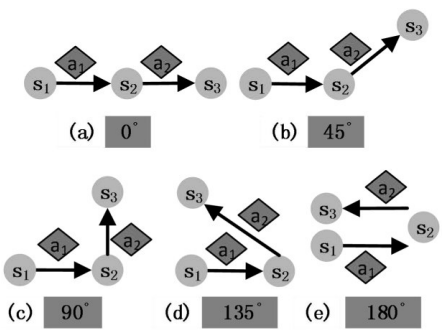

图1 矢量动作图

方格尺寸由地图的比例尺大小而定，比例尺上毫米标的实际长度为单个方格的尺寸。例如：比例尺1∶50 000，方格代表实际尺寸为50 m，地形图分辨率取50 m。

步长N为智能体单步移动n个方格。以往栅格法寻优大多采用固定步长思路，即每次选择动作后移动固定间距n个方格，众多可行的探索方案被限制产生。因此本文采用以最小步长的新思路，即智能体移动间距为单个方格，单方向移动满足最小步长n个方格后，智能体可以选择改变或不改变动作方向，探索方案群相比于固定步长的探索方案群更加全面。如图2所示，例如最小步长设

为3，智能体单方向移动4、5方格处均发生转向，若为固定步长，则只能在3、6、9……处(3的倍数)发生转向。

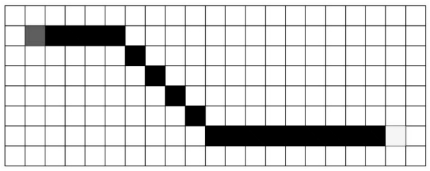

图2 路径探索轨迹

1.5 奖惩反馈

智能体与环境交互后反馈的信号分为两种：是否继续本轮探索信号、奖励或惩罚数值信号。如式(2)所示，位于状态s_i时是否继续本轮探索由done反馈，done=False继续本轮探索，done=True终止本轮探索；奖惩值由R反馈并记录为reward=$\{r_1, r_2, \cdots, r_n\}$。

$$\text{state}=\begin{cases} Danger, done = True \\ Else, done = False \\ Cost, done = False \\ Terminal, done = True \end{cases} \tag{2}$$

式中：线路禁修区(Danger)，reward=$-C_1$(常数)。正常探索状态(Else)，reward=$\phi_1-\phi_2$，此项为起点与终点之间的势能差作为奖励势能函数[14]，赋予智能体更明确的探索方向，加快探索速度。智能体移动所对应的工程费用(Cost)，reward=$-(C_T+C_B+C_E+C_L)$[15]，其中：C_T为隧道建设费用；C_B为桥梁建设费用；C_E为填挖土方费用；C_L为铺轨费用。终点(Terminal)，reward=$+C_2$(常数)。

1.6 算法对比

以范围1 km×1 km区段进行样本实验。DQN框架参数设置：learning rate取0.001，γ取0.9，epsilon取0.5，epsilon min取0.05，batch size取32；PPO框架参数设置：learning rate取0.001，γ取0.9，lambda取0.95，epoch取5，loss clipping取0.2。

探索策略：DQN采用ε贪心策略，即每个动作a被选择的概率是ε，有1−ε的概率选择当前价值最大的动作；PPO采用随机策略，即依靠输出的每个动作的概率，对所有动作按权重随机进行选取。

训练方法采用类似“课程学习”的方法，即探索初期于终点附近随机生成起点，随探索轮数的增加随机生成位置范围逐渐扩大，探索后期起点位置为初始设置值不再变化。先学习简单任务再逐渐增加难度的策略不仅加快了学习速度，还增加了方案探索度。

根据DQN框架与PPO框架实验结果(图3)分析：PPO奖励曲线迭代1.4万次趋于收敛，DQN奖励曲线迭代2.2万次趋于收敛，故PPO收敛速度较快，更加节省寻优时间；PPO奖励值收敛于25.6，DQN奖励值收敛于22.5，故PPO收敛值较高，寻优效果更好。通过以上对比实验，本文选取PPO算法框架构建铁路智能选线模型。

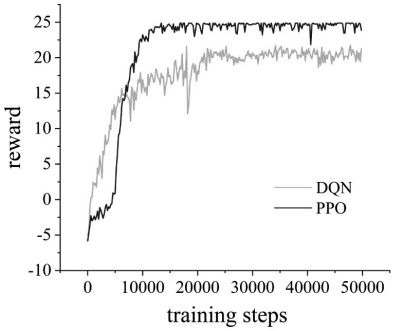

图3 奖惩曲线对比图

2 线路中心线方案生成

通过铁路智能选线深度强化学习模型生成线路走向线，这是在带状区域的地形中进行探索的结果，并不涉及平纵参数和线元构成。因此，想呈现出直观的线路中心线方案，为选线后续阶段提供参考，需进一步对线路走向线进行线形拟合。

2.1 自适应线容器

本文采取以直线−曲线−直线的要素形式绘制铁路线形，采取先拟合直线后配置曲线的原则，其中首先需要解决线形分段问题，因此本文提出一种基于线路走向点的分段方法：自适应线容器。自适应线容器由两条无限长的平行线构成，平行线中间的空隙为容器，容器内部可以包含的线路走向线控制点判定为趋于同一直线的分段，当出现某一点无法被包含于容器内部时，容器从此处断开，自动适应下一段线形生成容器，继续判定后续控制点。

其中两根平行线的间距称之为限宽(W)，判定公式为$W=b_1+b_2$，即上限宽b_1与下限宽b_2之和为自适应线容器的限宽。限宽设置大小为$W=\left(\frac{N}{2\sqrt{2}}\right)\times Rr$，其中$N$为智能体的探索步长，Rr为栅格精度/m。自适应线容器实际模型如图4所示。

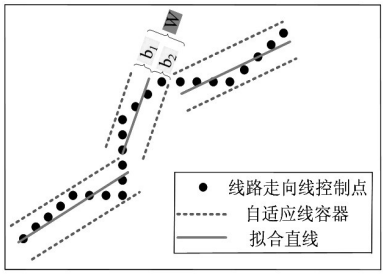

图4 自适应线容器示意图

2.2 拟合直线段

控制点分段完成后，选取正交最小二乘法对各个分段拟合直线方程。正交最小二乘法既考虑了横坐标X的误差还考虑了纵坐标Y的误差，传统最小二乘法仅考虑

纵坐标Y的误差，二者中正交最小二乘法拟合效果更优。直线方程中所需斜率k和截距b的计算方法由拟合公式(3)所示。

$$\begin{aligned}&\begin{bmatrix}k\\b\end{bmatrix}=\begin{bmatrix}\delta k\\\delta b\end{bmatrix}+\begin{bmatrix}k^0\\b^0\end{bmatrix}\\&\begin{bmatrix}\delta k\\\delta b\end{bmatrix}=(B^TB)^{-1}(B^TL)\\&B=\begin{bmatrix}S_1,T_1\\\cdots\\S_i,T_i\\\cdots\\S_n,T_n\end{bmatrix}\quad L=\begin{bmatrix}P_1\\\cdots\\P_i\\\cdots\\P_n\end{bmatrix}\end{aligned}\tag{3}$$

其中：

$$S_i=\frac{2X_i(k^0X_i-Y_i+b^0)}{1+(k^0)^2}-\frac{2k^0(k^0X_i-Y_i+b^0)^2}{[1+(k^0)^2]^2}$$

$$T_i=\frac{2(k^0X_i-Y_i+b^0)}{1+(k^0)^2}$$

$$P_i=-\frac{(k^0X_i-Y_i+b^0)^2}{1+(k^0)^2}$$

式中：k为拟合斜率，b为拟合截距，k^0为斜率的近似值，b^0为截距的近似值；δk为斜率的修正值，δb为截距的修正值；$(X_i, Y_i, i=1\cdots n)$为该分段内各数据点的横纵坐标。

2.3 平面线形生成

以自适应线容器对铁路智能选线模型输出的线路走向控制点进行合理分段，对各个分段的数据点以正交最小二乘法拟合该段线形。在相邻两条直线夹角的角平分线上寻找圆心，从直线交点两侧各取五组数据点进行拟合半径，曲线半径$R\in[R_{min},R_{max}]$。其中要求拟合得出的圆曲线必须和直线段相切，因此本文在此限定条件下，对正交最小二乘法进行调整，调整后的曲线拟合方法如公式(4)所示。

$$\begin{aligned}&[R]=[\delta R]+\left[R^0\right]\\&[\delta R]=(B^TB)^{-1}(B^TL)\\&B=\begin{bmatrix}U_1\\\cdots\\U_i\\\cdots\\U_n\end{bmatrix}\quad L=\begin{bmatrix}V_1\\\cdots\\V_i\\\cdots\\V_n\end{bmatrix}\end{aligned}\tag{4}$$

其中：

$$U_i=\frac{(X_0-X_i)-k_0(Y_i-k_0X_0-b_0)}{\sqrt{(X_i-X_0)^2+(Y_i-k_0X_0-b_0)^2}}-\frac{|k_1-k_0|}{1+(k_1)^2}$$

$$V_i=\frac{|(k_1-k_0)X_0-b_0+b_1|}{1+(k_1)^2}-\sqrt{(X_i-X_0)^2+(Y_i-k_0X_0-b_0)^2}$$

式中：R为拟合半径；R^0为半径的近似值，δR为半径的修正值；X_0为圆心横坐标的近似值(以R^0和角平分线方程计算得出)；k_0为角平分线方程的斜率，b_0为角平分线方程的截距，k_1与b_1为相邻两条直线中任意一条直线的斜率与截距；$(X_i, Y_i, i=1\cdots n)$为所取集合中各个数据点的横纵坐标。

由此公式拟合得出的圆曲线，一定与相邻两直线相切。若拟合半径小于R_{min}则由R_{min}代替拟合半径配置曲线，其中最小曲线半径R_{min}根据《铁路线路设计规范》取值，R_{max}=12 000 m。若直线交点间距过短，出现无法布控圆曲线情况时，需要通过修正交点位置，以保证线形连续且满足最小曲线半径以及最小夹直线要求，如图5所示为同向曲线和反向曲线的修正方式。

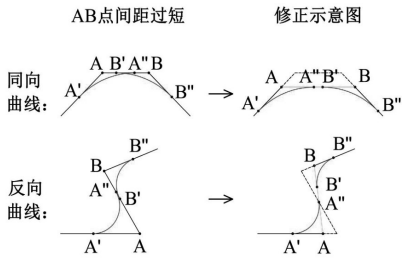

图5 交点间距过短时线形修正方式

2.4 纵断面线形生成

平面线形生成后，原始坐标位置发生改变，坐标对应的高程数据也发生变化，根据新平面坐标重新读取与之对应的高程数据，以此生成纵断面线形。考虑最小坡段长度、最大限制坡度、最大坡度代数差，采取上述同样的方法进行分段、拟合。

3 案例论证

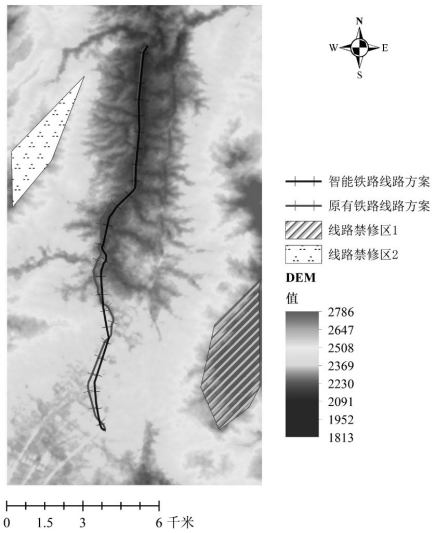

图6 选线方案对比

以云贵高原某铁路为例，验证本文提出的基于深度强化学习铁路智能选线方法。以PPO框架搭建，其模型

(下转第155页)