论文题目: 基于信用分配的决策优化——以中国象棋为例

报告日期: 2023-08-28

报告人: xxx

1. 背景与目标

信用分配（credit assignment）即奖励分配，是强化学习领域的一个重要问题。对于单智能体强化学习，智能体可能在中间的时间步没有收到任何奖励信号，而在最终的时间步才收到奖励信号，信用分配就是要解决这种奖励延迟的问题，合理量化智能体每个时间步应得的奖励，从而鼓励智能体执行价值高的动作，避免无价值的动作。所以、本论文旨在深入研究单智能体强化学习领域中的信用分配问题，以中国象棋这一典型的完全信息博弈为切入点，详细探讨如何有效地对中国象棋中的每个对弈步进行奖励分配，以优化智能体的学习效果。通过对比分析多种信用分配方法的特点，寻找一种切实可行且高效的奖励机制。

本文将提出一种基于深度神经网络评估的奖励机制。该机制将充分考虑当前实际对弈结果以及涉及的各个棋子的棋力，从而为每一步着法赋予恰当的价值。这一奖励机制的引入将有助于激励智能体选择那些具有更高价值的动作，从而加速智能体的学习过程。

1. 项目进展情况

当前进行奖励分配的主要方法是对奖励函数进行重塑（reward shaping）。Jose A. Arjona-Medina等【1】提出RUDDER方法进行奖励重塑，该方法指出可以使用LSTM进行关于奖励的模式识别，从而对奖励分解后再重分配。如图1所示，得益于LSTM的网络架构：遗忘门、记忆门、输出门，如果信息没变化，网络不会学到新的模式，当有新的信息，相应的模式就会被学习到。通过分析LSTM的“记忆”，可以重构出关键事件这些信息，并根据最后预测的奖励的贡献度分配给每个状态-动作配对（图2所示）。

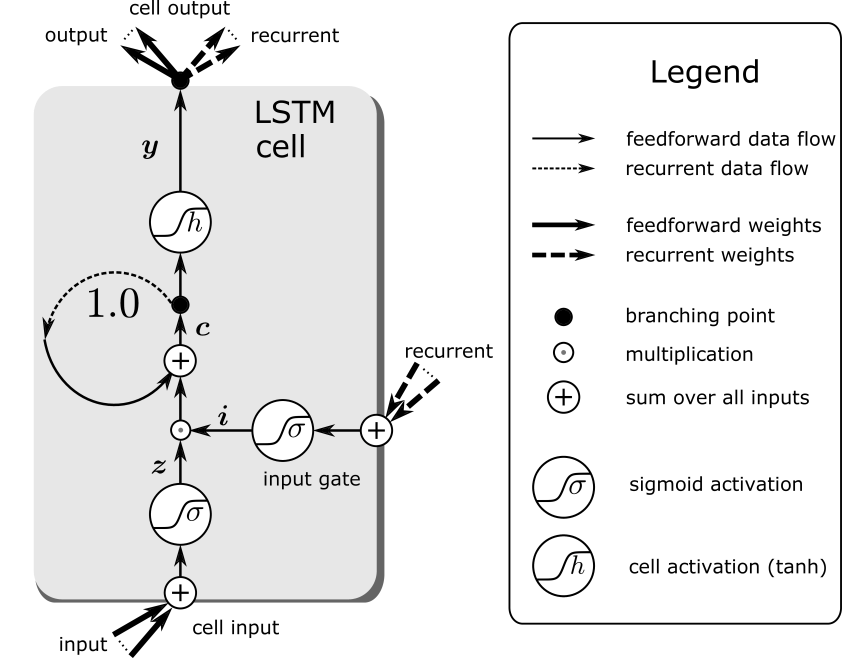


图1 LSTM单元结构[[1]](#footnote-0)

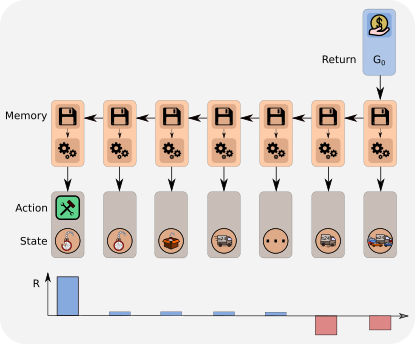


图2 使用LSTM检测关键事件[[2]](#footnote-1)

多触点归因（MTA: Multi Touch Attribution）是一种确定每个触点对最终转化效果的价值贡献进行分析的方法，常用于数字营销场景，如顾客从开始对某种产品产生初始印象到最终购买该产品期间，可能会涉及多个渠道，如果把“功劳”公平归因于各个渠道，是多触点归因要解决的问题。 Ning li等【参考2】提出了一种使用带注意力机制的深度神经网络模型DNAMTA，使用LSTM来拟合顾客所触达的广告序列路径，学习每个广告渠道的权重，并且将顾客个人静态信息通过全连接神经网络编码融入到整体的训练模型中。

受此启发，中国象棋的博弈过程可以类比于数字营销中顾客所触达的广告渠道路径：下棋者落子的每部棋可类比于每个广告渠道，下棋的最终输赢结果则类比于整体广告效果是否让顾客最终购买了该产品，从而中国象棋的信用分配问题可以借鉴广告的多触点归因分析来解决。本文将使用LSTM为主要神经网络架构来实现多触点归因分析，识别每一步落子的贡献度大小。

* 1. 数据收集与整理

通过与象棋程序进行自动博弈来收集博弈数据。数据经过去除重复值、去除空值和异常值处理后，还基于原始数据组合新的数据特征。 数据字段说明如表1所示。

表1 博弈数据字段说明

|  |  |
| --- | --- |
| 字段名 | 说明 |
| match\_id | 博弈场次id，标识每一场博弈 |
| round\_id | 轮次id，标识每一场博弈中每一轮对战 |
| state | 当前棋局状态 |
| action | 落子动作 |
| next\_state | 下一个棋局状态 |
| done | 当前博弈局是否结束：0表示未终局，1表示已终局 |
| chapture\_reward | 吃子奖励：正数为红方吃子，负数表示黑方吃子 |
| win | 胜/负/和局标识：1表示胜，-1表示负，0表示和局 |
| converted | 胜或负为1，和局为0 |
| jid | 每次落子的唯一标识 |

详细数据样例可参考图3。

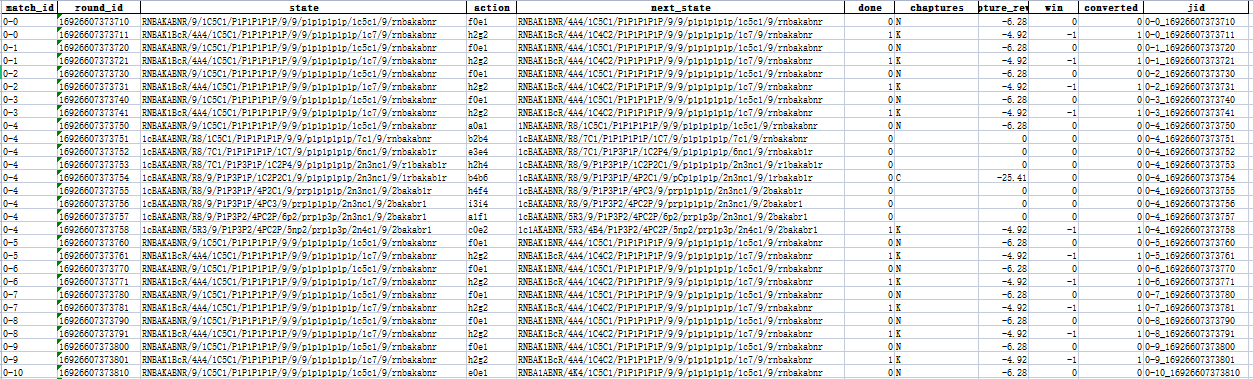


图3 博弈数据样例

1. 原始数据经过清洗后，由于state棋局状态特征为字符串型，现使用sentence-bert【参考3】的方法将编码为数值型向量，而action落子动作特征则编码为one-hot向量。最终将每个样本特征拼接为2871维的向量，作为训练数据。当前已收集了10777条博弈记录（仅红方），根据match\_id进行分组，形成形如[batch, step, feature]的训练样本2462条，测试样本308条，验证样本为308条。
   1. 模型开发与训练

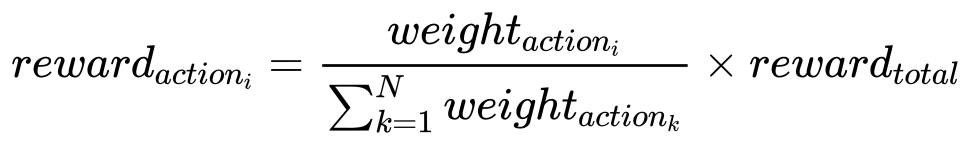
本论文借鉴DNAMTA【参考2】的神经网络归因模型，主要使用LSTM作为模型的主要架构（图4），并通过注意力层学习奖励分配权重。训练过程如下：

1. 使用6层LSTM对博弈数据特征进行编码，分别保留每一步的输出结果；
2. 每一步的编码结果输入到注意力网络层，输出注意力权重;
3. 注意力权重与LSTM的每一步编码结果进行加权求和，最后通过一个全连接层输出预测结果（converted=1或converted=0）。

【补充图4神经网络架构】

* 1. 系统集成与测试

模型训练完后，可提取注意力层的权重，基于历史博弈数据可统计出每个落子动作对应的贡献权重。目前基于算式（1）进行奖励分配。

 （1）

1. 当前成果

（1）已完成了神经网络模型训练并已验证训练收敛，如图5所示。

1. 已输出落子动作权重，如图6所示。
2. 已输出AlphaGo Zero【参考4】的奖励分配方式（所有动作都得到相同的奖励分配，即平均分配方式）和本论文的奖励分配方式对比结果，如图7所示，作为后续分析使用。
3. 遗留问题与解决方案
4. 基于公式（1）的奖励分配方式缺乏验证。解决方案：需增加多种奖励分配方式进行验证。
5. 本论文所提出的奖励分配方式的所表达优势阐述不充分。解决方案：除了加入平均分配方式，还可以考虑加入基于最后触点分配、基于时间衰减分配等方式进行对比验证。
6. 下阶段计划

在本节中，概述项目接下来的计划和目标。这可能包括：

优化与改进： 对已有模型和系统的进一步优化和改进计划。

扩展应用： 考虑将该人工智能预算分配技术应用到其他领域或问题上的计划。

持续监测： 计划如何对项目中的技术和系统进行持续监测和维护。

1. 总结与展望

在本节中，总结项目目前的阶段性成果，强调取得的进展和重要性，并展望项目的未来发展。

1. 附件

如果有必要，可以在本节添加附录，包括详细的数据分析结果、技术细节、代码示例等。

以上是关于基于人工智能的预算分配项目的进度成果报告模板，你可以根据实际情况进行调整和填写。

【附件1】

参考文献

【1】RUDDER: Return Decomposition for Delayed Rewards

【2】Deep Neural Net with Attention for Multi-channel Multi-touch Attribution

【3】Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks

1. 图片来自https://ml-jku.github.io/rudder/ [↑](#footnote-ref-0)
2. 图片来自https://ml-jku.github.io/rudder/ [↑](#footnote-ref-1)