多臂赌博机两阶段框架下的策略研究

江木力

2025 年 4 月 26 日

1 算法策略

1.1 实验阶段策略

我们实现了以下探索-利用策略：

1.1.1 固定探索策略

- **ϵ-贪婪 (Epsilon-Greedy)**: 以 ϵ 概率随机探索，$1-\epsilon$ 概率选择当前最优臂。
- **UCB(Upper Confidence Bound)**: 选择上置信界最高的臂，平衡探索与利用。
- **Thompson 采样 (Thompson Sampling)**: 基于 Beta 分布后验采样。
- **Softmax**: 按指数加权概率选择臂，温度参数控制探索强度。

1.1.2 动态探索策略

- **基于 Hoeffding 不等式的动态探索**:
  - 根据后续承诺阶段长度 N 自适应调整置信水平 ($\delta = 1/N$)。
  - 当某臂的置信下界高于所有其他臂的置信上界时停止探索。
- **基于贝叶斯后验的动态探索**:
  - 使用蒙特卡洛方法评估每个臂是最优臂的后验概率。
  - 当某臂的后验概率超过阈值 $(1-1/N)$ 时停止探索。

1.2 承诺阶段策略

- **最佳经验值 (BestEmpirical)**: 选择估计平均奖励最高的臂。
- **最常拉动 (MostPulled)**: 选择被拉动次数最多的臂。
- **置信下界 (LCB)**: 选择置信下界最高的臂，更保守的选择方法。

2 理论分析

2.1 T/N 比率的理论意义

在总预算 $B = T + N$ 固定的情况下，如何分配探索轮数 T 和利用轮数 N 是一个关键问题。理论上：

- 当 $B \to \infty$ 时，最优探索轮数应该是 $O(\log B)$，使得 $T/B \to 0$。
- 当 B 较小时，存在一个取决于问题难度的最优 T/N 比率。

对于差距为 Δ 的情况 (最优臂与次优臂均值差)，UCB 策略的理论最优 T 约为:

$$T_{opt} \approx \frac{c \cdot \log N}{\Delta^2}$$

其中 c 是一个常数。这表明随着 N 的增加，最优 T 应该对数级增长，而随着问题难度 (Δ 减小) 的增加，最优 T 应该二次级增长。

2.2 动态探索策略分析

2.2.1 基于 Hoeffding 的动态策略

当使用 $\delta = 1/N$ 时，该策略确保:

- 错误识别概率 P(选择次优臂) $\leq 1/N$。
- 期望后悔值 $E[R] \leq \Delta \cdot N \cdot (1/N) + T \cdot \Delta = \Delta(1 + T)$。
- 要满足终止条件，需要至少 $\Omega(\log(kN)/\Delta^2)$ 次拉动。

2.2.2 贝叶斯策略分析

当停止阈值为 $1 - 1/N$ 时:

- 错误率随 N 增大而减小，确保后悔值期望上界为常数。
- 探索轮数理论上与 $\log(N)/\Delta^2$ 成正比。
- 在简单问题 (臂间差距明显) 上通常能更快停止探索。

3 实验结果与分析

3.1 单次实验示例

3.2 T/N 关系分析

3.3 动态探索策略评估

3.4 N 值对探索轮数的影响

4 讨论

4.1 主要发现

1. **最优 T/N 比率**: 实验表明，对于固定预算 $B = T + N$，存在最优的资源分配比例。这一比例与问题难度和所选策略密切相关。
2. **动态策略优势**: 动态探索策略能够根据统计证据自适应决定何时停止探索，在很多情况下优于预设固定 T 值的策略，特别是当 N 值变化大时。
3. **策略组合性能**: 实验表明 Thompson 采样与 BestEmpirical 的组合在多数场景下表现最佳，而简单的 ϵ-贪婪策略在参数选择得当时也能取得不错的效果。

4.2 实际应用指导

基于我们的研究，可以提出以下实际应用指导：

1. **当 N 值已知且较大时**：应采用动态探索策略，特别是基于贝叶斯后验的方法，可以在保证选择质量的同时减少不必要的探索。
2. **当问题难度未知时**：UCB 策略是稳健的选择，它能在各种难度条件下都取得不错的表现。
3. **当计算资源受限时**：使用 ϵ-贪婪策略，设定 T/N 比率约为 0.5 左右，是简单而有效的做法。

5 结论与未来工作

5.1 结论

本研究系统分析了多臂赌博机两阶段框架中的决策策略，重点关注了 T/N 比率对总后悔值的影响。我们提出的动态探索策略能够根据承诺阶段长度 N 自适应确定所需的探索量，在多种场景下表现优异。实验结果验证了我们的理论分析：最优探索轮数与 $\log(N)$ 成正比，与问题难度的倒数平方 $(1/\Delta^2)$ 成正比。

5.2 未来工作

1. **非平稳环境扩展**：将研究扩展到奖励分布随时间变化的非平稳环境。
2. **多臂相关性研究**：考虑臂之间存在相关性的情况，例如线性上下文赌博机模型。
3. **实际应用验证**：在实际业务场景如临床试验或广告投放中验证本研究的发现。
4. **理论边界改进**：提供动态策略在不同问题特征下的更精确后悔值上界。