1. 绪论
   1. 课题背景
   2. 本文内容安排

第二章 相关理论及算法

深度强化学习是将深度学习的特征提取功能和强化学习的动态学习策略结合起来的算法理论，可以综合性解决用户偏好动态变化，奖励反馈，未来回报率等问题，也可以很好地忽略推荐特征之间的联动影响。本文基于深度强化学习理论构建基于DQN网络的推荐系统。

2.1 DQN（Deep Q Network）

DQN是一种常见的Value Based强化学习算法，是将深度学习模型与强化学习结合在一起从而成功地直接从高维的输入学习控制策略，适用于在复杂场景中大规模数据中学习，音乐推荐是一个较为复杂的场景，因此，本文使用DQN来构建模型是适合的。

2.2 LSTM（Long Short-Term Memory）

LSTM是一种时间循环神经网络，适合于处理和预测时间序列中间隔和延迟非常长的重要事件。由于CNN并不能对时序数据进行处理，而LSTM在时序数据的处理上表现出较好的效果，因此本文使用LSTM来处理歌曲播放歌曲记录和点赞收藏记录这类时序性数据。

第三章 算法设计

音乐推荐是一个互动过程，音乐推荐系统需要反复在不同的音乐中进行选择来推荐，因此需要分析用户与推荐系统之间交互的细节。用户进入系统中，得到系统与推荐的音乐列表，可以选择听歌或者不听歌，也可以进行收藏，下载，加入歌单，评价等操作，每次推荐后，用户浏览记录，歌曲歌单，评分信息会更新，推荐系统也会更新自身参数，并更新推荐列表，因此将音乐推荐过程作为强化学习任务是合适的。

本文对音乐推荐过程进行强化学习建模，将用户模拟为强化学习中的环境，将推荐系统模拟成智能体对象，系统将用户的属性特征，历史浏览记录，歌曲歌单，点赞记录，听歌次数，评分信息等作为输入数据，用户进入系统，向推荐系统发出请求推荐歌曲的请求，系统收到请求后将输入数据和带推荐音乐池p输入模型，根据模型算法选出一组top-k歌曲列表，生成推荐歌单列表，用户对歌单中推荐的歌曲进行收听，收藏，下载，加入歌单，评价等操作，这些操作是用户对推荐系统的反馈和奖励，同时，强化学习探索环境并采取行动来最大化积累奖励，目标是最大化用户满意度，即预期的总奖励。

3.1用户的状态特征设计

因为用户对系统的反馈是正负向的双向反馈，如果用户收听，点赞，收藏歌曲，就意味着用户对该歌曲有正向的反馈，而如果用户略过该歌曲，则意味着用户对该歌曲有负向的反馈。负反馈能够在一定程度上帮助过滤用户不喜欢的歌曲。

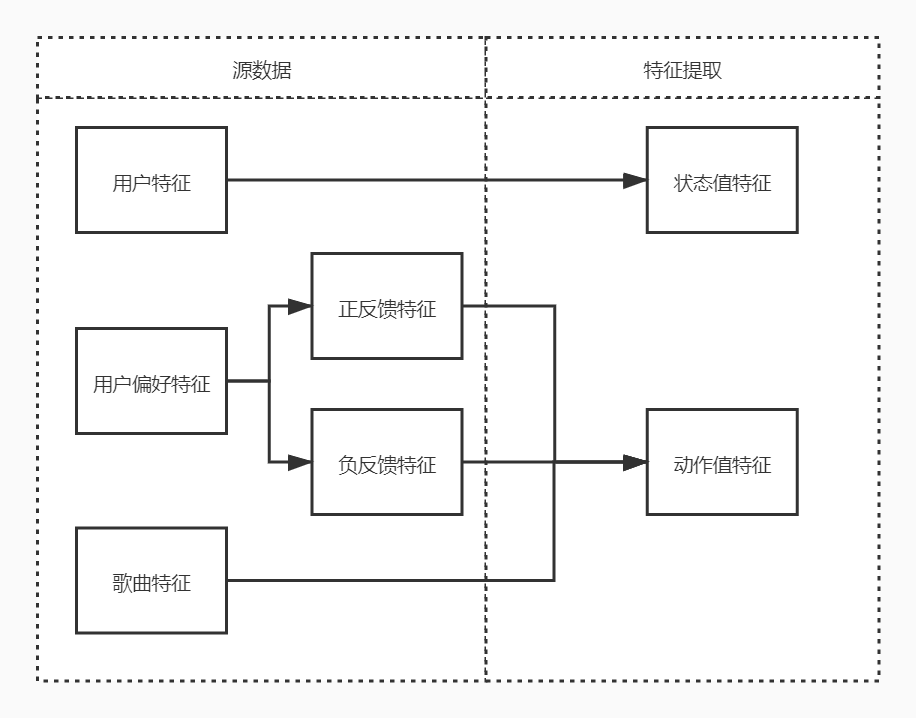
将目前的用户偏好表示为状态s，s+={i1,i2,…,in}表示用户最近收听，点赞收藏的n首歌曲的特征集合，即用户正反馈信息集合。s-={j1,j2,…,jm}表示系统推荐给用户的歌曲中被略过的m个歌曲集合，即用户负反馈集合。s=(s+,s-)，其中s+和s-中的歌曲集合是按照时间顺序排列的。

当推荐系统将歌曲a在状态s=(s+,s-)的状态下推荐给用户时，如果用户略过歌曲a，则正反馈保持不变s‘+=s+，负反馈更新为s’-={j2,j3,…,jm,a}，如果用户点击收听，点赞，收藏歌曲，则负反馈不变s‘-=s-，正反馈更新为s‘+={i2,i3,…,in,a}，状态更新为s’=(s’+,s‘-)。

3.2特征提取

本文所使用的特征可以分为两大类：用户特征包括用户性别，年龄，职业，受教育水平，所处地域位置等用户信息，描述了用户本身的一些特征；用户偏好特征，包括用户对歌曲的收听评分收藏点赞等记录等偏好信息，表达了用户对音乐的潜在长期偏好，在本文设计中分为正反馈特征和负反馈特征；用户特征和用户偏好特征描述了音乐推荐系统的用户兴趣画像；音乐特征包括歌词，歌曲作者，歌唱者，评论，类别等信息，描述了音乐的特征信息。

由于复杂的特性之间的关联影响关系，将特征提取分为两部分：第一部分只与用户特征单独用卷积提取特性，这部分特性跟用户的状态更相关，其值函数对应状态价值函数，第二部分则同时依赖用户特征，歌曲特征和歌曲偏好，相当于与用户状态和动作相关，其值函数对应动作价值函数，将这两部分的提取出的特征集合起来而已更快更准确提供推荐策略。



特征关系图

3.3值函数定义

Q(s，a；θ，α，β ) = V (s；θ，β ) + A(s，a；θ，α)

3.4 神经网络结构

3.5 奖励函数定义

3.6 探索策略设计

3.7 推荐系统框架设计

第四章 总结与展望