ВВЕДЕНИЕ

Большая часть существующих систем построения онлайн очередей построена на жестких признаках, которые позволяют довольно простым образом сортировать очередь, но в любом случае это упирается в проблему времени прихода в очередь и чаще выходит, что электронная очередь не соотносится с контекстом живой очереди.

Классическим решением для такой системы является подход firs-in-firs-out (fifo), что в переводе с английского языка означает «первый пришел, первый ушел», в таких системах нет никакой адаптивности.

Целью научной исследовательской работы является предварительная реализация телеграм бота по постановке людей в очередь и их простейшей сортировки.

Для достижения цели, были поставлены следующие задачи:

- подбор методов, которые лягут в основу будущей системы организации адптивных онлайн очередей;
- рассмотреть существующие подходы в машинном обучение и определить подходящий для дальнейшей реализации продукта;
- выбрать алгоритм для программной реализации продукта в будущем и обосновать этот выбор путем сравнения его с другими алгоритмами;
- выполнить тестирование созданной системы.

ГЛАВА 1. ОБЗОР ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ И АЛГОРИТМОВ

1.1. Типы искусственных нейронных сетей

Варианты подходов при использовании машинного обучения: обучение с учителем; обучение без учителя; обучение с частичным привлечением учителя; обучение с подкреплением.

1.1.1. Обучение с учителем

Обучение с учителем рассматривается в контексте поставленной задачи, но существуют проблемы из-за того, что нужна информация о разметке очереди, а таких данных нет, ведь очередь строится относительно времени прихода и не понятно, на чём обучаться. Так как системы сильно отличаются друг от друга, нет какого-то унифицированного метода научить на какой-то очереди, т.е. идеальной очереди не существует, а значит должна быть адаптивная система. Также нет лейблирования, т.е. что идёт за чем, кто имеет больший приоритет и почему, то мы продолжим рассмотрение различных видов машинного обучения дальше.

1.1.2. Обучение без учителя

Обучение без учителя обучение основано на поиске закономерностей или структур в данных без какой-либо помеченной информации. В контексте онлайн-системы массового обслуживания отсутствие помеченных данных об оптимальной конфигурации очереди затрудняет определение наилучшей организации очереди на основе одних только шаблонов. Кроме того, неконтролируемое обучение, возможно, не сможет полностью охватить сложность и динамику проблемы, поскольку оно не учитывает вознаграждения или штрафы, связанные с различными действиями, выполняемыми в рамках системы массового обслуживания.

1.1.3. Обучение с частичным подкреплением

В обучении с частичным привлечением учителя используется комбинация помеченных и немаркированных данных. Хотя такой подход может быть полезен в некоторых случаях, онлайн-система массового обслуживания страдает от отсутствия согласованной и надежной информации с маркировкой, что затрудняет предоставление подходящего набора данных для обучения. Как и в случае обучения с учителем, рассматриваемое обучение требует определенных знаний об идеальной организации очередей, которых может не существовать или которые могут сильно различаться в разных системах.

Исключив эти подходы, мы можем подчеркнуть пригодность алгоритмов обучения с подкреплением (RL) для онлайн-системы массового обслуживания. Такие алгоритмы обучаются методом проб и ошибок, оптимизируя действия на основе полученных вознаграждений или штрафных санкций. Это позволяет модели адаптироваться и находить оптимальную конфигурацию очереди, не полагаясь на помеченные данные. Адаптивный характер алгоритмов RL делает их хорошо подходящими для универсальной онлайн-системы массового обслуживания, поскольку они могут научиться справляться с различными ситуациями и динамикой системы.

1.2. Подходы rl алгоритмов

RL содержит в себе такие подходы, как: dynamic programming policy iteration; q-learning; monte-carlo; policy gradient; sarsa.

1.2.1 dynamic programming policy iteration

Динамическое программирование, т.е. мы должны понимать что-то о системе, в нашем случае есть изначальные условия (появление в очереди; соц. дем факторы; смещение, которое организуется за счет того, что система некоторым образом вознаграждает людей за добровольную потерю места.

1.2.2 Q-learning

Q-learning - это алгоритм обучения с подкреплением без моделей, основанный на ценности, который направлен на изучение оптимальной функции "действие-ценность". Он использует таблицу для хранения ожидаемых вознаграждений для каждой пары состояние-действие и итеративно обновляет эти значения, чтобы приблизиться к оптимальной политике. В контексте онлайн-системы массового обслуживания Q-learning может использоваться для оценки ожидаемого вознаграждения за различные действия (например, изменение порядка очереди, предоставление приоритета определенным лицам) в различных состояниях (например, текущая конфигурация очереди, социально-демографические факторы).

1.2.3 Monte-carlo

Monte-Carlo - это класс алгоритмов обучения с подкреплением, которые основаны на усреднении результатов нескольких выборочных эпизодов для оценки ценности состояний и действий. Эти методы могут быть использованы как для задач прогнозирования, так и для задач управления. Для системы онлайн-массового обслуживания методы Монте-Карло могут быть применены для оценки ценности различных действий в различных состояниях, помогая системе определить оптимальную политику путем многократной выборки и обновления оценок ценности на основе результатов предпринятых действий.

1.2.4 Policy gradient

Policy gradient - это класс алгоритмов обучения с подкреплением, основанных на политике, без использования моделей, которые оптимизируют политику напрямую, а не изучают функцию значения. Эти методы используют градиентное восхождение для обновления параметров политики, стремясь максимизировать ожидаемое совокупное вознаграждение. В контексте онлайн-системы массового обслуживания методы градиента

политики могут использоваться для прямой оптимизации политики массового обслуживания, позволяя системе узнавать, какие действия следует предпринять в различных состояниях для достижения наилучшей организации очереди и минимизации времени ожидания или других соответствующих показателей.

1.2.5 SARSA

State-Action-Reward-State-Action - это еще один свободный от моделей алгоритм обучения с подкреплением на основе ценностей, который можно рассматривать наряду с ранее обсуждавшимися алгоритмами RL. SARSA расшифровывается как Состояние-действие-Награда-Состояние-действие, представляя последовательность событий, которые происходят во время взаимодействия агента с окружающей средой. Алгоритм является методом на основе политики, что означает, что он изучает функцию значения для текущей политики, следуя этой политике. В контексте онлайн-системы массового обслуживания модель SARSA может быть применена следующим образом:

- Состояние: текущая конфигурация очереди, включая положение отдельных лиц и любые соответствующие социально-демографические факторы;

- Действие: решение, принятое системой управления очередью, такое как изменение порядка очереди, предоставление приоритета конкретным пользователям или другие возможные действия;

- Вознаграждение. Числовое значение, представляющее непосредственный результат предпринятого действия, например, сокращение времени ожидания, повышение справедливости или другие соответствующие показатели;

- Следующее состояние. Обновленная конфигурация очереди, полученная в результате выполненного действия;

- Следующее действие: Последующее решение, принимаемое системой управления очередью на основе обновленного состояния. Алгоритм SARSA обновляет функцию значения действия (Q-функция) на основе немедленно полученного вознаграждения и расчетного значения следующей пары состояние-действие, следуя текущей политике. Этот итеративный процесс позволяет системе выработать оптимальную политику для управления онлайн-очередью, принимая во внимание динамический характер проблемы и различные факторы, влияющие на организацию очереди.

По сравнению с другими алгоритмами RL, модель SARSA имеет преимущество в том, что она является методом на основе политики, который может привести к более стабильному обучению и лучшему решению проблемы компромисса между разведкой и эксплуатацией. Это может быть полезно в контексте онлайн-системы массового обслуживания, где принятие наилучших решений на основе текущих знаний при одновременном изучении новых возможностей имеет решающее значение для оптимизации организации очереди.

Каждый из этих подходов к обучению с подкреплением обладает уникальными преимуществами и может быть рассмотрен для универсальной онлайн-системы массового обслуживания в зависимости от конкретных требований и ограничений задачи.

ГЛАВА 2. ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ СИСТЕМЫ

ЗАКЛЮЧЕНИЕ

Было рассмотрено применение алгоритма SARSA для оптимизации онлайновых систем очередей. Адаптивный характер обучения с подкреплением делает его хорошо подходящим для таких приложений, учитывая отсутствие маркированных данных и различные условия в разных системах. Также были подчеркнуты трудности и ограничения алгоритма SARSA, а также важность разработки признаков, настройки гиперпараметров и оценки производительности.

Потенциальные реальные приложения систем онлайн-очередей с использованием SARSA или других алгоритмов RL охватывают различные области, включая здравоохранение, транспорт и управление событиями. Будущие направления исследований включают разработку новых алгоритмов, передаточное обучение, многоагентное обучение с усилением, ограничения справедливости и обучение с усилением с участием человека.

В дальнейшем будет реализована финальная версия приложения, которая будет содержать рассмотренные алгоритмы и методики из этой работы.

СПИСОК СОКРАЩЕНИЙ И УСЛОВНЫХ ОБОЗНАЧЕНИЙ

SARSA - State-Action-Reward-State-Action;

RL - Reinforcement Learning;

UCB - Upper Confidence Bound;

СУБД – система управления базами данных