План-проспект курсовой работы

студента ИМИ группы ПМИ-18-2 Федорова Никиты на тему:

«Построение рекомендательной системы с помощью методов машинного обучения»

Оглавление

Введение

Глава 1.

1.1.

1.2.

Глава 2.

2.1.

2.2.

Заключение

Список использованных источников

Введение

В настоящее время объем информации, доступной человеку стал настолько велик, что даже используемые сервисы для поиска информации не всегда справляются с задачами людей. Для осуществления поиска объектов необходимо знать об их существовании. С ростом объема доступной информации знать обо всем становится затруднительно. Данную задачу решают рекомендательные системы, которые связывают между собой объекты и пользователей и ищут для пользователей объекты, вероятно, их интересующие.

 Алгоритмы коллаборативной фильтрации обладают рядом проблем, наиболее существенными из которых являются проблема холодного старта, первого рейтинга, избирательности внимания и долгого времени работы при пересчете данных. Алгоритмы, основанные на анализе данных о поведении, всегда привязаны к специфике области, к которой они применяются.

Большинство крупных веб-сайтов рекомендует своим пользователям различные предложения, например товары для дальнейшего изучения или людей, к которым целесообразно обратиться. Рекомендательные механизмы сортируют огромные объемы данных (целесообразно говорить о «Больших данных») для выявления потенциальных предпочтений пользователей.

Amazon.com утверждает, что 40% продаж генерируются через механизмы рекомендаций.

Рекомендательные системы изменили способы взаимодействия веб-сайтов со своими пользователями. Вместо предоставления статической информации, когда пользователи ищут и, возможно, покупают продукты, рекомендательные системы увеличивают степень интерактивности для расширения предоставляемых пользователю возможностей.

Рекомендательные системы формируют рекомендации независимо для каждого конкретного пользователя на основе его прошлых действий, а также на основе поведения других пользователей.

Актуальность работы:

Цель моей работы состоит в исследовании и разработке эффективных алгоритмов для рекомендательных систем, позволяющих выбирать рекомендации с приемлемым уровнем релевантности в условиях большого числа пользователей при неполной или отсутствующей информации об их предпочтениях, а также разработка архитектуры системы, использующей такие алгоритмы.

Для решения поставленной цели исследования были определены следующие задачи:

1. Выбор и построение модели данных для выбора рекомендаций.
2. Разработка алгоритма выбора рекомендаций в условиях «холодного старта».

Объект исследования: рекомендательные системы человеко-машинного взаимодействия.

Предмет исследования: структуры данных и алгоритмы выбора релевантных рекомендаций.

Методы исследования: анализ, моделирование.

Глава 1. Алгоритмы построения рекомендательных систем

1.1. Что такое рекомендательные системы

Рекомендательная система - комплекс алгоритмов, программ и сервисов, задача которого предсказать, что может заинтересовать того или иного пользователя. В основе работы лежит информация о профиле человека и иные данные.

Такая система включает в себя весь процесс — от получения информации до её представления пользователю. Важен каждый этап. От информации, которая будет обрабатываться, зависит, какие алгоритмы окажутся более подходящими. Хорошие алгоритмы дают хорошие, полезные рекомендации. Критерии оценки результата позволяют выбрать наиболее подходящие алгоритмы. Первоначально задача кажется простой, но создать хорошую систему сложно. Нужно очень бережно подходить к построению системы. Даже лучшие алгоритмы не всегда дают подходящие результаты.

1.2. Как работают рекомендательные системы

Существуют два уровня:

* Особенности и предпочтения, не меняющиеся месяцами или годами; глобальные оценки; зависимость от характерных пользовательских черт: пол, место проживания; интересные страницы и т.п.
* Тренды и быстрые изменения интересов.

Данные собираются явным и/или неявным способами. В первом случае определяется предпочтения пользователя анкетами, опросами и т.п. Метод эффективный, только пользователи не всегда соглашаются.

Во втором методе фиксируется поведения потребителя на сайте или в приложении: просмотр страниц/разделов, добавление в корзину, комментарии, отзывы и т.п. При правильном сборе данных и аналитике метод дает хорошие результаты.

1.3. Типы рекомендательных систем

Есть четыре разных типа рекомендательных систем:

1. Коллаборативные (collaborative filtering)
2. Основанные на контенте (content-based)
3. Основанные на знаниях (knowledge-based)
4. Гибридные (hybrid)

Рассмотрим эти типы рекомендательных систем.

Коллаборативные рекомендательные системы:

Это системы, в которых рекомендации пользователю рассчитывается на основе оценок других пользователей. Здесь существует множество алгоритмов, но наиболее популярные - User/User(поиск соседей по оценкам), Item/Item(определение схожести предметов по оценкам пользователей) и SVD(самообучающийся алгоритм).

Суть этого типа – нахождение ближайших соседей. Близость двух пользователей или предметов определяется метриками схожести.

Преимущество – высокая теоретическая точность.

Минус – невозможно порекомендовать новым пользователям, т. к. отсутствует информация об этих пользователях (холодный старт).

Пример: Саша и Вова любят покушать суши и позаниматься спортом. А еще они оба любят машины марки Mercedes и ездят на них. Еще есть Никита, который тоже любит покушать суши после тяжелого дня и позаниматься спортом вечером. Но у него нет машины. Исходя из одинаковых интересов, Никите можно порекомендовать машины марки Mercedes.

Рекомендательные системы, основанные на контенте:

Этот тип лежит в основе многих рекомендательных систем. В отличие от коллаборативной фильтрации, этап знакомства с пользователем опускается. Система работает на основе: проанализировать контент предметов и составить набор его критериев (жанры, тэги, слова), узнать какие критерии нравится пользователю, сопоставить эти данные и получить рекомендации. Критерии составляют из пользователей и предметов точки в системе координат, и если точка пользователя и предмета рядом, то с высокой вероятностью предмет понравится пользователю.

Рекомендательная система, основанная на знаниях:

Этот тип работает на основе знаний о какой-то предметной области: о пользователях, товарах и других, которые могут помочь в ранжировании. Существуют разновидности: case-based, demographic-based, utility-based, critique-based, whatever-you-want-based и т. д. Пользователи указывают предмету другие схожие предметы. На основе этих данных создаются рекомендации. Рекомендательная система, основанная на знаниях, сможет расширить географию выбора пользователя за счет рекомендаций, которые оставили предыдущие пользователи.

Очевидное преимущество системы — высокая точность.

Минус — для разработки этой системы требуется много времени и ресурсов.

Гибридные рекомендательные системы:

Эти системы объединяют несколько выше представленных алгоритмов в один. У всех описанных ранее типов есть определенные недостатки. Комбинирование нескольких алгоритмов в рамках одной платформы позволяет если не устранить их полностью, то хотя бы минимизировать.

Существуют несколько распространенных типов комбинирования:

* реализация по отдельности коллаборативных и контентных алгоритмов и объединение их предположений;
* включение некоторых контентных правил в коллаборативную методику;
* включение некоторых коллаборативных правил в контентную методику;
* построение общей модели, включающей в себя правила обеих методик.

Пример: рекомендательная система Netflix построена на 27 алгоритмах.

Основной недостаток гибридных систем – сложность разработки этих рекомендательных систем.

1.4. Постановка задачи

Дана обучающая выборка

Для задачи регрессии – числовой признак.



Глава 2. Методы прогнозирования

2.2. Метод k-ближайших соседей

Метод k-ближайших соседей – алгоритм классификации или регрессии объектов. В нашем случае используется для регрессии, где объекту присваивается среднее значение по k ближайшим к нему объектам.

1. Вычислить расстояние до каждого объекта обучающей выборки
2. Выбрать k самых близких объектов
3. Предсказать класс наиболее часто встречающихся во множестве

Метрики:

1. Эвклидово 
2. Манхэттенское расстояние 
3. Расстояние Чебышева 
4. Расстояние Минковского 

При k=1 алгоритм неустойчив к «выбросам», в результате получим ошибочные значения для объектов, близких к «объектам-выбросам» (не встречавшиеся ранее в алгоритме данные).

При k=N метод вырождается в константу, т.к. с какого-то элемента существенных изменений в значениях не будет.



2.1. Деревья решений

Дерево решений – представляет собой иерархическую структуру правил, которая состоит из «узлов», соединенных с помощью ребер. Начальный узел называется «корнем» и не имеет входного ребра. Атрибут корневого узла определяется из того, на сколько хорошо он в одиночку может разделять на классы имеющийся набор данных. Узел с исходящим ребром называется «внутренним узлом», где для атрибута в соответствии с его значением создается ребро. Процесс выполняется рекурсивно по каждому атрибуту. Все остальные узлы, которые имеют только по одному входящему ребру, называются «листьями» и являются конечным узлом дерева, содержащим целевую переменную определяющую класс объекта [7]. Однако нерелевантные параметры набора данных могут негативно сказаться на построенном дереве решений. В свою очередь, немного измененные или дополненные данные могут привести к весьма серьезному перестроению дерева решений. Основным преимуществом деревьев решений является простота интерпретации и понимания полученного результата даже для неспециалиста [8]. Заложенные принципы деревьев решений были использованы во множестве алгоритмов классификации: C4.5, ID3, CART, MARS, а также целый ряд их модификаций [9].

2.3. Случайный лес (Random forest)

Алгоритм Random Forest основывается на использовании большого количества деревьев решений, которые по отдельности не являются «сильными» алгоритмами, но объединенный ансамбль из таких деревьев будет показывать результаты выше, нежели просто один «сильный» алгоритм. Обучение алгоритма происходит путем построения деревьев решений на случайно выбранных параметрах обучающей выборки. Целевой класс для нового объекта определяется большинством голосов построенных деревьев решений [12]. Random Forest считается весьма эффективным с точки зрения точности прогнозирования при работе с большим объемом данных, а также предлагает интуитивно понятный подход к оценке важности используемых параметров [11]. При работе с большим объемом данных показывает высокие результаты точности, также хорошо работает с малым количеством объектов и большим количеством переменных. Из-за построения большого количества деревьев интерпретируемость результатов низкая, а время на построение всех деревьев высокое. В зависимости от количества построенных деревьев может быть требователен к объему памяти.

2.4. Сравнение

Многообразие существующих алгоритмов машинного обучения не позволяет говорить об однозначно лучшем, так как каждый из алгоритмов имеет свои слабые и свои сильные стороны. На основе проведенного анализа методов классификации [54-56, 59, 60] составлена таблица сравнения основных методов классификации по нескольким параметрам (табл. 1.1), где для каждого из них была приведена оценка по шкале от 0 до 2, где 2 – максимальный балл, а также приведено сравнение и составлена таблица (см.табл.1.2) с указанием достоинств и недостатков основных методов классификации.

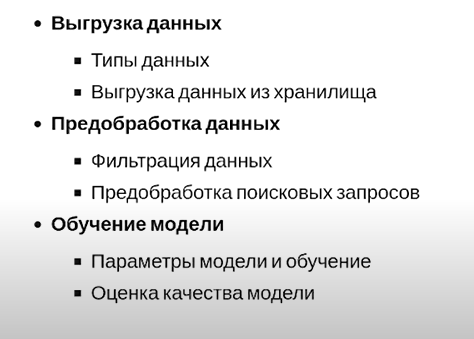


Представленные методы имеют разные области применимости, разные требования к набору входных данных и типу параметров, разные вычислительные характеристики и уровень потребления памяти. Использование данных методов классификации в рекомендательных системах будет также давать различный результат. Следовательно, для каждой задачи и каждого набора данных необходимо использовать свой метод классификации, который может основываться на ансамблевом подходе и включать в себя множество методов с последующим объединением результатов при принятие итогового решения [6].

Помимо этого, построение ансамблей из различных алгоритмов позволяет

устранить один из недостатков машинного обучения, когда структура небольшой обучающей выборки в силу своего размера мало соответствует структуре всей рассматриваемой совокупности данных, особенно на больших данных. В этом случае обучение каждого алгоритма может проводиться на разных обучающих выборках.

Глава 3. Реализация методов и построение моделей



Список использованных источников

1. Федоровский, А.Н. Архитектура рекомендательной системы, работающей на основе неявных пользовательских оценок / А.Н. Федоровский, В.К. Логачева // Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2011, Воронеж, Россия, 2011.
2. Правиков, А.А. Разработка и применение метода формализации проектирования рекомендательных систем с естественно-языковым интерфейсом: дис.... канд. технических наук, Москва, 2011. С. 160
3. “How Computers Know What We Want — Before We Do” [электронный ресурс] URL: <http://content.time.com/time/magazine/article/0,9171,1992403,00.html> [дата обращения: 16 мая 2021]
4. Decision Trees, Random Forests, and Nearest-Neighbor classifiers: <https://pages.mtu.edu/~shanem/psy5220/daily/Day13/treesforestsKNN.html> [дата обращения: 17 мая 2021]
5. Классификация, регрессия и другие алгоритмы Data Mining с использованием R: <https://ranalytics.github.io/data-mining/> [дата обращения: 17 мая 2021]
6. Что такое дерево решений и где его используют?: <https://habr.com/ru/company/productstar/blog/523044/> [дата обращения: 17 мая 2021]
7. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. Classification and regression trees. Monterey, CA: Wadsworth & Brooks // Cole Advanced Books & Software, 1984.
8. Patra A., Singh D. A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms // International Journal of Computer Applications. 2013. № 7 (75). Р. 14–18.
9. Satyanarayana N., Ramalingaswamy C., Ramadevi Y. Survey of classification techniques in data mining // IJISET - International Journal of Innovative Science, Engineering & Technology. 2014. № 9 (1). Р. 268–278.
10. Altman N. S. An introduction to kernel and nearest-neighbor nonparametric regression // The American Statistician, 1992. 46 (3). Р. 175–185.
11. Crisci C., Ghattas B., Perera G. A review of supervised machine learning algorithms and their applications to ecological data // Ecological Modelling, 2012. (240). Р. 113–122.
12. Breiman, Leo. Random Forests //Machine Learning, 2001. 45 (1), pp. 5–32.
13. Understanding Random Forest: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>