

РАЗРАБОТКА МОДЕЛИ ДЛЯ МЭТЧИНГА ДАТАСЕТОВ

А. С. Буримский

*Белорусский государственный университет,
пр. Независимости, 4, 220030, г. Минск, Беларусь, leha.burimskiy.05@gmail.com
Научный руководитель – И. М. Галкин, кандидат физико-математических наук, доцент*

В работе представлена разработка модели мэтчинга датасетов. Основным результатом является модель, которая сопоставляет одинаковые сущности в различных данных.

Ключевые слова: метод KNN, косинусное сходство, линейная регрессия, мэтчинг датасетов.

Введение

В современном мире, где данные играют ключевую роль в принятии стратегических решений, задача сопоставления (мэтчинг) датасетов становится все более актуальной и востребованной. Мэтчинг данных представляет собой процесс сопоставления информации из различных источников. Это позволяет компаниям, исследователям и аналитикам извлекать ценные инсайты, делать точные прогнозы и принимать обоснованные решения на основе полной и точной информации. В данной работе была разработана модель, которая сопоставляет два датасета с информацией о продаже автомобилей.

Цель работы

Основной целью данной работы является разработка модели для мэтчинга двух датасетов, которая будет способна точно определять соответствие между записями из двух наборов данных.

Задачи работы

1. Изучение методов мэтчинга данных: ознакомление с основными подходами и алгоритмами, используемыми для сопоставления записей в различных датасетах.
2. Разработка модели подбора кандидата на основе метода ближайших соседей (KNN): создание алгоритма, способного находить ближайших соседей для каждой записи из одного датасета в другом на основе косинусного сходства строчковых данных.
3. Реализация линейной регрессии для определения соответствия записей: разработка модели, которая на основе косинусного сходства между двумя ближайшими соседями принимает решение о сопоставлении записей из двух датасетов.
4. Анализ результатов и выводы.

Изучение мэтчинга данных

В начале работы проводился обзор различных источников по методам мэтчинга данных. Основной задачей было понять различные методы решения задачи сопоставления данных. Были рассмотрены такие методы, как метод ближайших соседей в качестве подбора кандидатов, а также линейная регрессия в качестве решающей модели для сопоставления.

Разработка модели подбора кандидата на основе метода ближайших соседей (KNN)

Первым шагом в реализации модели мэтчинга стало написание модели подбора кандидата. В качестве модели был выбран метод KNN, так как он прост в реализации и эффективен на небольших наборах данных. В качестве метрики расстояния было использовано косинусное сходство, так как оно хорошо подходит для работы с текстовыми данными. Так как для мэтчинга важен лишь ближайший кандидат, в реализации KNN проводился поиск лишь одного ближайшего соседа, который становился кандидатом на матч. Отдельно стоит отметить, что в

реализации KNN использовались не все поля для расчета расстояния между записями, а лишь небольшое количество, которые содержали в себе самую важную информацию о машине.

Реализация линейной регрессии для определения соответствия записей

В качестве модели для определения, являются ли две записи матчем, была выбрана линейная регрессия. Она работает по простому принципу: если приходящее значение расстояния выше некоторого порога, то считается, что это две разных записи. Если же значение расстояния меньше порога, то считается, что это одна и та же запись. Основная проблема в реализации линейной регрессии – это подбор порога, который был выбран исходя из точности модели на обучающем наборе данных.

Анализ результатов и выводы

Процесс оценки точности модели заключался в запуске модели на наборе тестовых данных. Модель показала высокую точность работы, однако была достаточно медленна, несмотря на то что объем данных был небольшим. В будущем хотелось бы улучшить скорость модели.

Результаты работы

1. Изучены различные методы мэтчинга данных.
2. Разработана модель подбора кандидатов на основе метода KNN.
3. Реализована линейная регрессия для определения соответствия записей.
4. Результаты работы модели были проанализированы и сделаны соответствующие выводы.

Перспективы развития

Основной перспективой развития модели является ускорение ее работы. В качестве одного из вариантов ускорения рассматривается замена метода KNN на метод ANN. Метод ANN позволяет ускорить поиск соседей за счет потери точности. Другими словами, ANN найдет не обязательно ближайшего соседа, однако сделает это быстрее. В зависимости от обстоятельств с помощью ANN можно найти баланс между скоростью работы модели и ее точностью.

Библиографические ссылки

1. Raschka S. Python Machine Learning. Birmingham, 2017. 595 с.