Отчет о проекте «Идентификация интернет-пользователей»

Сысак Михаил Алексеевич

22 декабря 2019 г.

1. Описание проекта

Идентификация пользователей по их поведению — сложная и интересная задача, у которой много применений. В этом проекте решалась задача идентификации пользователей по последовательности сайтов, посещенных ими. В исходном датасете для каждого пользователя имеется csv-файл, в котором последовательно записаны сайты, посещенные пользователем, и время их посещения.

2. Первичная обработка и анализ признаков

Основной идеей, которая использовалась на протяжении всего проекта, являлось использование мешка слов для хранения обучающей выборки. Недостатком такого подхода является большая размерность данных; достоинством — их разреженность, позволившая воспользоваться разреженными матрицами для хранения.

Обучающие примеры строились следующим образом: из всей последовательности сайтом скользящим окном выделялись последовательности заданной длины, и для каждой такой последовательности создавался объект обучающей выборки. Исследования проводились для размера сессии в 10 сайтов.

На рисунке 1 изображена гистограмма числа различных сайтов в сессии. Критерий Шапиро-Уилка уверенно отвергает гипотезу о нормальном распределении этой величины; ясно, что она имеет сложное распределение с двумя пиками.

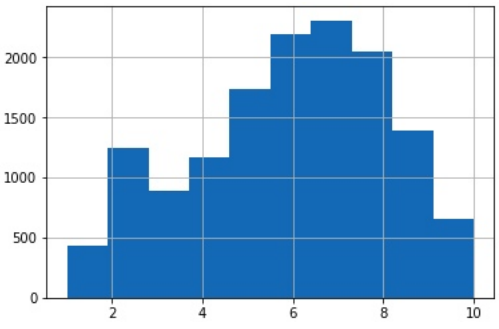

Рис. 1: Число различных сайтов

Для исследования данных были визуально проанализированы различные признаки для выделенных 10 пользователей. В процессе анализа было замечено, что пользователи принципиально отличаются между собой закономерностями, по которым они посещают различные сайты. Например, кто-то не выходит в интернет в выходные, в то время как другие, наоборот, меньше времени проводят в интернете по будням. У всех пользователей сессии начинаются в разное время дня, по разному распределено и количество сайтов в сессии. Это говорит о том, что по этим данным действительно можно строить идентифицирующую модель.

3. Дальнейшее построение признаков

По исходным данным был построен ряд признаков, анализ которых был описан в предыдущем разделе. Для каждого пользователя извлекались данные из соответствующей csv-таблицы, и вносились в новый датасет. Были построены следующие признаки:

- Продолжительность сессии в секундах
- Число уникальных сайтов в сессии
- Час начала сессии
- День недели
- Дневное/ночное время
- Месяц
- Число сайтов в сессии, входящих в топ-30 по посещениям

4. Сравнение алгоритмов классификации

На выборке из 10 пользователей с длиной сессии 10 были опробованы различные алгоритмы классификации. Результаты приведены в таблице 1.

Алгоритм	Accuracy на кросс-валидации	Accuracy на отложенной выборке
KNN	0.565	0.584
Random Forest	0.723	0.735
Логистическая регрессия	0.759	0.775
SVM	0.766	0.782

Таблица 1: Качество различных алгоритмов

Был проведен анализ зависимости качества классификации от длины сессии и ширины окна. Выяснилось, что с ростом длины сессии и уменьшением ширины окна возрастает точность классификации.

Кроме того, была решена задача классификации одного пользователя против всех остальных, и построена кривая обучения для этой задачи. Она приведена на рисунке 2. Видно, что с увеличением размера выборки увеличивается обобщающая способность алгоритма.

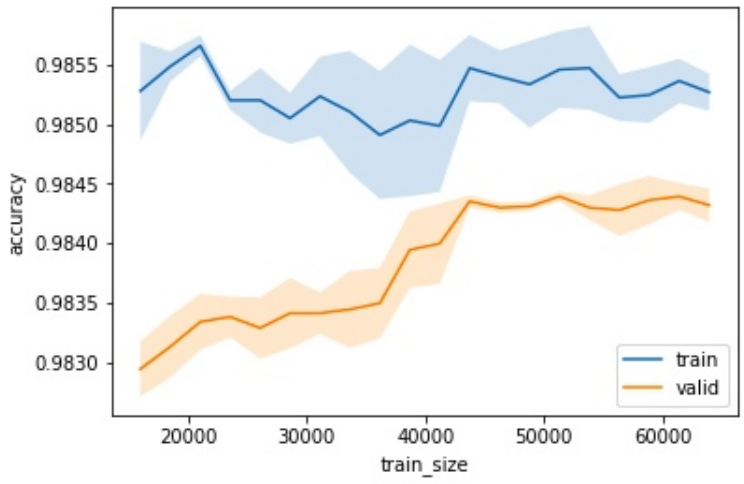

Рис. 2: Кривые обучения в задаче One-vs-All

5. Дальнейший выбор модели и метрики

Была проведена работа с данными соревнования на Kaggle «Catch me if you can». Снова решалась задача идентификации One-vs-All, на этот раз на сильно большей выборке. По этой причине был использован алгоритм SGDClassifier, который хорошо работает с большими объемами данных. В качестве метрики использовался ROC-AUC Score, поскольку он устойчив к несбалансированным выборкам. С помощью GridSearchCV были подобраны параметры алгоритма, побивающие первый бенчмарк. Второй бенчмарк побить не удалось — возможно, с этим помогли бы дополнительные признаки.
Для задачи идентификации 400 пользователей была применена библиотека Vowpal Wabbit. Для сравнения анализировалась также работа SGD и логистической регрессии для этой же задачи. Результаты приведены в таблице 2.

Алгоритм	Точность на отложенной выборке	Время обучения
Vowpal Wabbit	0.345	24.5 s
SGDClassifier	0.294	24.9 s
Логистическая регрессия	0.363	3min 14s

Таблица 2: Сравнение алгоритмов для классификации 400 пользователей

Из таблицы видно, что VW совмещает в себе лучшее от остальных: он быстро обучается, при этом мало отставая от долгой логистической регрессии.

6. Выводы

В данном проекте было проведено исследование алгоритмов, метрик и методов для работы с объемными разреженными выборками. Основной вывод состоит в том, что для такого типа задач очень хорого подходит Vowpal Wabbit в силу эффективного распараллеливания и качества, превышающего обычный SGDClassifier. Кроме того, для несбалансированных задач бинарной классификации хорошо подходит метрика ROC-AUC, устойчивая к различным по размеру классам.

Проведенная работа позволяет с уверенностью сказать, что исследованная задача может быть эффективно решена с помощью алгоритмов машинного обучения. Рассмотренные алгоритмы могут применяться для кластеризации пользователей по их предпочтениям в интернете, для ИИ, отличающего владельца аккаунта от взломщика, для идентификации заблокированных пользователей, зашедших с другого IP-адреса, а также для множества других подобных задач.