Секция «Вычислительная математика и кибернетика»

Линейная комбинация случайных лесов в задаче предсказания релевантности документов

Фигурнов Михаил Викторович
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: michael@figurnov.ru

В работе представлен метод ранжирования документов по пользовательскому поведению, который был разработан для конкурса Интернет-математика «Relevance Prediction Challenge» [3] и занял там второе место из 85 участников (лучший результат среди российских команд). Поисковые системы выдают по запросу пользователя документы, ранжированные по убыванию релевантности (степени соответствия) документа запросу. Ранжирование производится на основе данных о документе. Для настройки параметров алгоритма ранжирования используется обучающая выборка, содержащая оценки релевантности для части документов. Особенности задачи соревнования:

1. Данные представляют собой обезличенные поисковые логи. Строки лога соответствуют действиям пользователей в сессиях.

2. Данные имеют достаточно большой объем (размер 16 ГБ, 340,8 млн. действий) и зашумлены.

3. Используется нестандартный функционал качества (средний AUC [3]).

Для решения задачи было использовано множество из 43 признаков документа, основанных на показах, истории кликов и времени. Примеры признаков: средняя позиция в выдаче поисковой системы, вероятность последнего в сессии клика, среднее время, потраченное на просмотр документа. Поскольку требуется упорядочить документы по релевантности, применён попарный (pair-wise) подход: рассматриваются признаки пар документов из одного запроса и ставится задача предсказания разности оценок релевантностей этих документов (т.е. какой из документов лучше подходит к запросу). В этой задаче естественно оптимизировать среднеквадратичную ошибку. Для этого мы используем случайные леса [1], которые хорошо зарекомендовали себя при решении других задач с зашумлёнными данными, например, кредитного скоринга. Исходная обучающая выборка разбивается на обучающую и валидационную выборку. На подгруппах признаков обучающей выборки создаются 28 случайных лесов. Попарные оценки релевантности каждого леса для валидационной и тестовой выборки переводятся в оценки релевантности отдельных документов. Затем применяется идея метода «LENKOR» [2]: оценки различных лесов объединяются в линейную комбинацию с неотрицательными коэффициентами. Её параметры настраиваются методом покоординатного спуска с убывающим шагом для достижения максимального функционала качества исходной задачи на валидационной выборке. Ответы на тестовой выборке усредняются с полученными коэффициентами. Предложенный метод позволил решить задачу с нестандартным функционалом качества и зашумлёнными данными. Оценки релевантности,

полученные предложенным методом, можно использовать в поисковых системах для улучшения ранжирования документов.

Литература

1. Breiman L. Random Forests // Machine Learning. 2001. Vol. 45. Pp. 5-32.
2. D'yakonov A. Two Recommendation Algorithms Based on Deformed Linear Combinations // ECML-PKDD 2011 Discovery Challenge Workshop. 2011. Pp. 21-27.
3. Интернет-математика «Relevance Prediction Challenge»: http://imat-relpred.yandex.ru/

Слова благодарности

Работа выполнена при финансовой поддержке РФФИ, проект 12-07-00187-а.