– на каждом устройстве выполняется поиск локальных для данного фрагмента разладок с использованием критериев, основанных на статистиках от вейвлет коэффициентов [1].

Стадия обобщения состоит из следующих основных шагов:

– аппроксимирующие коэффициенты на уровне разрешения J копируются на выделенное устройство, из этих коэффициентов составляется временной ряд в порядке следования фрагментов исходного временного ряда;

– выполняется поиск неоднородностей в построенном временном ряду, который представляет собой результат масштабирования исходного ряда, а содержащаяся в его отсчетах информация относится к глобальным свойствам исходного ряда;

– на основании оценок моментов возникновения неоднородностей в построенном временном ряду определяем номер соответствующего фрагмента временного ряда и приблизительное место возникновения неоднородности;

– для получения точных оценок моментов возникновения неоднородностей проводится дополнительный анализ соответствующих фрагментов исходного ряда.

Библиографические ссылки

1. Абрамович М.С., Мицкевич М.Н. Обнаружение скачкообразных изменений среднего с использованием вейвлет-преобразования Хаара // Информатика. – 2008. – №4. – С. 59-66.

ПРОГНОЗИРОВАНИЕ ПОПУЛЯРНОСТИ СТАТЕЙ В ИНТЕРНЕТЕ НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Артеменко В. А., Кузьмин К. Г.

Белорусский государственный университет, Минск,
e-mail: artemenko.vladimi@gmail.com, kuzminkg@bsu.by

В данной работе рассмотрена задача определения паттерна поведения пользователей для предсказания популярности статьи на известном ресурсе habrahabr по содержанию и времени публикации. В качестве метрики популярности статьи использована функция, зависящая от доли статей за последний месяц, которые имеют рейтинг меньше, чем у текущей статьи. Тем самым, доле рейтинга поставлены в соответствие квантили стандартного распределения [1]. Отметим, что подобные задачи относятся к широко известному [1, 2] классу задач классификации с категориальными признаками.

Как инструмент для решения задачи была использована библиотека Vowpal Wabbit – одна из наиболее широко используемых библиотек в индустрии. Ее отличает высокая скорость работы и поддержка большого количества различных режимов обучения. В частности, онлайн обучение позволяет работать с большими и высокоразмерными данными. В библиотеке реализовано хэширование признаков, а также Vowpal Wabbit отлично подходит для работы с текстовыми данными.

В данной работе был изучен вопрос применения нейронных сетей для задачи прогнозирования временных рядов. Для рассматриваемой задачи построен эффективный алгоритм прогнозирования и дана оценка его точности.

Библиографические ссылки

1. Bishop, C. M. Pattern Recognition and Machine Learning / C. M. Bishop. Springer. – 2006. – 738 p.

2. Дьяконов, А. Г. Методы решения задач классификации с категориальными признаками / А. Г. Дьяконов // Прикладная математика и информатика. Труды факультета Вычислительной математики и кибернетики МГУ имени М.В. Ломоносова. – 2014. – № 46. – С. 103–127.

ПРОГНОЗИРОВАНИЕ ВЫБОРА КЛИЕНТОМ ОТЕЛЯ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Голубева Л. Л., Мурашко А. С.

Белорусский государственный университет, Минск, Беларусь,
e-mail: goloubeva@bsu.by, aliaksandra.murashka@gmail.com

В работе рассматриваются вопросы выявления внутренних закономерностей больших объемов данных методами машинного обучения в применении к задаче мультиклассификации на выборке данных, которая представляет собой информацию о поисковых запросах пользователя относительно конкретного отеля при бронировании на сайте Expedia. Выборка данных содержит 40 198 536 образцов, из них 2 528 243 тестовых образца, для каждого образца указаны 24 признака. Согласно условию имеется 100 кластеров отелей, к одному из них и требуется отнести отель.

Процесс моделирования включал в себя настройку виртуальной машины для выполнения вычислений на облаке, подготовку данных (создание и отбор признаков с помощью статистических процедур [1], обработку пропущенных значений, приведение данных к подходящему формату). Были построены модели на основании различных алгоритмов машинного обучения (Decision Tree, Random Forest, Gradient Boosted Trees, Stacking). Проведено исследование поведения моделей на различных наборах признаков, выбран наиболее информативный набор. Точность прогноза различными алгоритмами оценивалась с помощью метрики MAP@5. Специфика данной метрики заключается в том, что для нее необходимо, чтобы предсказание состояло не из одного значения предполагаемого кластера, а из пяти наиболее вероятных. Для моделей, показавших более высокую точность на начальном этапе, подобраны параметры (число классификаторов, входящих в ансамбль, максимальная глубина и др.) [2], которые позволили улучшить первоначальный результат.

В приведенной сводной таблице (см. Табл. 1) указаны характеристики точности и времени обучения методов, показавших лучшие результаты в своей категории.

Табл. 1. Сравнение точности и времени обучения моделей

Модель	Точность на тестовом наборе	Map@5	Время обучения
Решающее дерево	0.185124	0.235172	39.4 с
Случайный лес	0.369408	0.338504	9 ч 34 мин
Градиентный бустинг	0.363490	0.385587	11 ч 37 мин 2 с
Стекинг (градиентный бустинг + случайный лес + решающее дерево)	0.186890	0.241722	1ч 58 мин 3 с

Библиографические ссылки

1. Hastie T. The elements of statistical learning. Data mining, Inference and Prediction second edition/ T. Hastie, R. Tibshirani, J. Friedman. -Springer, 2008, 764 p.

2. Jain A. Complete Guide to Parameter Tuning in XGBoost (with codes in Python) [Electronic source]. Mode of access: https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/. Date of access: 10.03.2017