МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СУМСЬКИЙ ДЕРЖАВНИЙ УНІВЕРСИТЕТ

ІНФОРМАТИКА, МАТЕМАТИКА, АВТОМАТИКА

IMA :: 2017

МАТЕРІАЛИ
та програма

НАУКОВО-ТЕХНІЧНОЇ КОНФЕРЕНЦІЇ

(Суми, 17–21 квітня 2017 року)

Суми
Сумський державний університет
2017

Сравнение алгоритмов адаптивного и градиентного бустинга в задаче классификации текстов

Ломотин К.Е., *студент*
Национальный исследовательский университет
«Высшая Школа Экономики», г. Москва, Россия

Алгоритмы машинного обучения позволяют более эффективно рубрицировать тексты, выделять из них знания, а также решать множество других задач, связанных с обработкой естественного языка. Бустинг – это один из ансамблевых подходов к улучшению моделей машинного обучения, суть которого состоит в том, что базовые модели обучаются последовательно: каждая следующая обучается на ошибках предыдущей. Работа посвящена сравнению двух наиболее популярных алгоритмов бустинга: AdaBoost и градиентного бустинга в задаче классификации научных статей по рубрикам первого уровня УДК. Главное различие этих алгоритмов заключается в методе коррекции весовых коэффициентов и параметров базовых моделей, входящих в их состав.

Обучающая выборка состояла из 39 тысяч статей с кодами УДК, загруженных с онлайн-ресурса cyberleninka.ru. Тексты прошли базовую предобработку и лемматизацию. Векторное представление текстов было сформировано из ключевых слов, выделенных с помощью меры важности слова TF-IDF [1]. В ходе обучения для каждого алгоритма бустинга были подобраны гиперпараметры, соответствующие лучшей метрике качества F1. Тестирование проводилось на статьях, не участвовавших в обучении. Алгоритм градиентного бустинга показал качество классификации 0.77 по метрике F1. Для популярного алгоритма AdaBoost качество по метрике F1 составило 0.67.

Таким образом, построение ансамбля классификаторов по алгоритму градиентного бустинга для решения задачи рубрикации текста является на 10% более эффективным, чем применение адаптивного бустинга.

Руководитель: Романов А.Ю., *доцент*

1. A. Romanov, K. Lomotin, et. al., *2016 SIBCON Proc.* **543fu4t** (2016).