УДК 004.41

Д.А. Радиванович, магистрант; И.Г. Сухорукова ст. преп.
(БГТУ, г. Минск)

МНОГОКРИТЕРИАЛЬНАЯ ФИЛЬТРАЦИЯ ЭЛЕКТРОННОЙ КОРРЕСПОНДЕНЦИИ

Наиболее популярным и эффективным подходом для фильтрации электронной корреспонденции на данный момент является использование машинного обучения с учителем с различными признаками, основанными как на содержании сообщений, так и на свойствах отдельных профилей пользователей.

Машинное обучение (machine learning) – это область научного знания, имеющая дело с алгоритмами, «способными обучаться». Необходимость использования методов машинного обучения объясняется тем, что для многих сложных – «интеллектуальных» – задач (например, распознавание рукописного текста, речи и т. п.) очень сложно (или даже невозможно) разработать «явный» алгоритм их решения, однако часто можно научить компьютер обучиться решению этих задач. Одним из первых, кто использовал термин «машинное обучение», был изобретатель первой самообучающейся компьютерной программы игры в шашки А. Л. Самуэль в 1959 г. Под обучением он понимал процесс, в результате которого компьютер способен показать поведение, которое в нее не было заложено «явно». Это определение не выдерживает критики, так как не понятно, что означает наречие «явно». Более точное определение дал намного позже Т. М. Митчелл: говорят, что компьютерная программа обучается на основе опыта E по отношению к некоторому классу задач T и меры качества P, если качество решения задач из T, измеренное на основе P, улучшается с приобретением опыта E.

На этапе классификации и оценки были протестированы 5 алгоритмов: наивный байесовский классификатор (NaiveBayes classifier), метод k ближайших соседей (k-nearest neighbors algorithm, k-NN)), метод опорных векторов (SVM), дерево принятия решений (Decision tree), случайные леса (Random forest).

В результате проведенного исследования данных алгоритмов машинного обучения лучший результат продемонстрировал алгоритм Random Forest - 93%. Наиболее высокий известный результат в данной задаче составляет 98%, однако в нем отсутствуют какие-либо ограничения на историчность признаков.