

УДК 004.048

МЕТОД ВЫБОРА ПРИЗНАКОВ С ИСПОЛЬЗОВАНИЕМ АНАЛИЗА ЧУВСТВИТЕЛЬНОСТИ НА ОСНОВЕ ДИСПЕРСИИ

Рунушкина О.Р. (Университет ИТМО), Тирский Б.Д. (Университет ИТМО)
Научный руководитель – доцент, кандидат технических наук, Муромцев Д.И.
(Университет ИТМО)

Введение. Выбор признаков является важной проблемой во многих задачах классификации и регрессии, особенно когда задача обучения включает в себя многомерный набор данных. Конечная цель состоит в том, чтобы выбрать оптимальное подмножество из большего набора признаков, исключив признаки с малой прогностической информацией или без нее. Это также может привести к повышению производительности модели машинного обучения (ML) [1].

Существующие подходы к выбору признаков в целом можно разделить на три группы: фильтр, обертка и встроенные подходы [2].

В этой работе рассматривается метод определения важности признаков в задачах ML, на примере FNN модели, в которой ее выходные данные изучаются с учетом входных факторов, которые являются входными признаками.

Основная часть. Шаги предлагаемого метода выбора признаков можно резюмировать следующим образом:

- 1) Инициализация параметров и обучение модели FNN до критериев ранней остановки.
- 2) Определение модели и ее выходных и входных факторов, которые вы хотите включить в свой анализ.
- 3) Назначение функции плотности вероятности с диапазонами изменения каждому входному фактору.
- 4) Генерация образцов и оценка модели.
- 5) Оценка влияния относительной важности каждой функции и вычисление нормализованного процента вклада для всех функций.
- 6) Удаление входных объектов с низким процентом вклада на основе определенного порога.

Выводы. В этой статье моделирование проводится с использованием нескольких задач классификации с намерением охватить примеры небольших и средних наборов данных высокой размерности. Предлагаемый в статье метод реализован на языке программирования python с использованием библиотек с открытым исходным кодом в стеке Scipy [3]. Эффективность предложенного метода подтверждается анализом результатов и визуализацией данных для серии экспериментов с несколькими известными наборами данных, взятыми из репозитория машинного обучения UCI.

Список использованных источников:

1. Дэш, М. и Х. Лю, Выбор признаков для классификации. Интеллектуальный анализ данных, 1997. 1(1): с. 131-156.
2. Хуан Л. и Ю. Лей. На пути к интеграции алгоритмов выбора признаков для классификации и кластеризации. IEEE Transactions on Knowledge and Data Engineering, 2005. 17(4): с. 491-502.
3. Oliver, M., *Introduction to the Scipy Stack—Scientific Computing Tools for Python*. Google Scholar, 2013.

Рунушкина О.Р. (автор)

Муромцев Д.И. (научный руководитель)