УДК 004.8

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДОВ СОКРАЩЕНИЯ РАЗМЕРНОСТИ ВЕКТОРОВ ИСХОДНЫХ ДАННЫХ ПРИ РЕШЕНИИ ЗАДАЧИ КЛАССИФИКАЦИИ АНСАМБЛЯМИ СЛУЧАЙНЫХ ДЕРЕВЬЕВ

© Шибаева А.О.

e-mail: aoshibaeva@yandex.ru

Самарский национальный исследовательский университет имени академика С.П. Королёва, г. Самара, Российская Федерация

В данной работе рассматриваются несколько упаковочных методов сокращения размерности векторов исходных данных, так как методы этой группы в общем случае дают наилучший результат. В данной работе рассматриваются алгоритм оценки признаков последовательным исключением признаков и алгоритм Boruta.

Алгоритм работы метода оценки признаков последовательным исключением признаков состоит из двух главных шагов [1]:

1) построить модели на начальной выборке данных и оценить важность всех признаков;

2) исключить наименее важных признаков из текущей выборки.

Эта процедура рекурсивно повторяется в сокращенном наборе до тех пор, пока в конечном итоге не будет достигнуто требуемое количество объектов для выбора.

Ниже приведена пошаговая работа алгоритма Boruta [2]:

1) добавить случайность к исходной выборке данных, путём добавления перемешанных копий всех признаков (которые называются теневыми признаками).

2) построить классификатор ансамбля случайных деревьев для расширенного набора данных и к каждому признаку применить меру важности признака (по умолчанию – уменьшение средней точности) для оценки важности каждого признака, где более высокий результат означает большую важность.

3) для каждого признака проверить, имеет ли реальный признак лучшее значение, чем лучший из его теневых признаков. Если нет, такие признаки считаются неважными и удаляются.

4) остановить алгоритм, если все признаки являются важными или неважными, либо если достигается заданный предел запусков ансамбля.

В качестве данных для обучения было решено взять два набора данных: ирисы Фишера [3] и типы стекла [4]. Набор ирисов Фишера включает в себя данные о 150 экземплярах ирисов, по 50 экземпляров для трёх видов – Ирис щетинистый (Iris setosa), Ирис виргинский (Iris virginica) и Ирис разноцветный (Iris versicolor). Для каждого экземпляра предоставлены четыре характеристики (в сантиметрах):

− длина чашелистика;

− ширина чашелистика;

− длина лепестка;

− ширина лепестка.

В качестве второго набора данных были использованы открытые данные о типах стекла, в зависимости от химического состава (содержания определённых химических элементов) и показателя преломления. Количество атрибутов, включая класс – 10, всего типов стекла – 7, количество векторов – 214.

В таблицах 1 и 4 приведены результирующие метки выбора признаков (1 – значимый признак, 0 – незначимый). В таблицах 2 и 5 приведены доли правильных ответов для исходных данных, а также для данных, которые были сформированы из исходных с помощью исключения незначимых признаков по таблице 1. В таблицах 3 и 6 приведено время работы алгоритмов.

Таблица 1. Значимость признаков на выборке ирисов Фишера

Номер признака	1	2	3	4
Рекурсивное исключение признаков	0	0	1	0
Boruta	0	0	0	1

Таблица 2. Доли правильных ответов на выборке ирисов Фишера

Доля правильных ответов		
На исходной выборке	На выборке с рекурсивным исключением признаков	На выборке Boruta
0,86	0,89	0,93

Таблица 3. Время работы на выборке ирисов Фишера

Время работы метода, с	
Рекурсивного исключения признаков	Boruta
72	15

Таблица 4. Значимость признаков на выборке типов стекла

Номер признака	1	2	3	4	5	6	7	8	9
Рекурсивное исключение признаков	1	0	1	0	0	0	1	1	0
Boruta	1	0	1	1	0	1	1	0	0

Таблица 5. Доли правильных ответов на выборке типов стекла

Доля правильных ответов		
На исходной выборке	На выборке с рекурсивным исключением признаков	На выборке Boruta
0,6	0,69	0,74

Таблица 6. Время работы на выборке типов стекла

Время работы метода, с	
Рекурсивного исключения признаков	Boruta
323	20

По полученным данным видно, что, несмотря на то, что алгоритмы выбирали значимыми неодинаковые наборы признаков, точность классификации после обработки на обработанных исходных данных выше, чем на исходной выборке. Однако, алгоритм Boruta показывает лучшие результаты, чем алгоритм рекурсивного исключения признаков, так как у него выше доля правильных ответов и меньше время работы.

Сравнение показало, что оба алгоритма позволяют увеличить точность задачи классификации, а алгоритм Boruta позволяет получить лучший результат по точности классификации при меньшем времени работы, чем алгоритм рекурсивного исключения признаков.

Библиографический список

1. Featureselection [Электронный ресурс] // scikit-learn.org. – https://scikit-learn.org/stable/modules/feature_selection (дата обращения: 17.11.2018).

2. How to perform feature selection (i.e. pick important variables) using Boruta Package in R? [Электронный ресурс] // Analytics Vidhya. – https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/ (дата обращения: 03.12.2018).

3. Boruta 0.1.5 [Электронный ресурс] // PyPI. – https://pypi.org/project/Boruta/ (дата обращения: 05.11.2018).

4. Iris Data Set [Электронныйресурс] // UC Irvine Machine Learning Repository – URL: http://archive.ics.uci.edu/ml/datasets/Iris (датаобращения: 3.12.2018)