УДК 004.021

ОТБОР ПРИЗНАКОВ ОСНОВАННЫЙ НА MRMR И ГЕНЕТИЧЕСКОМ АЛГОРИТМЕ ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ ДАННЫХ

Тисленко М.Д. (Самарский национальный исследовательский университет им. академика С. П. Королева)

Научный руководитель – кандидат технических наук, Парингер Р. А.
(Самарский национальный исследовательский университет им. академика С. П. Королева)

Введение. Отбор признаков давно показал эффективность при предобработке данных для построения модели в машинном обучении, он позволяет ускорить классификацию данных, выделить подмножество признаков, в котором связь между свойствами и целевым значением легче проследить, убрать ненужные признаки. В рамках задачи классификации отбор признаков на этапе предварительной обработки данных заключается в выборе такого подмножества признаков, что точность классификации на нём будет максимальна среди всех подмножеств той же мощности.

Цель данной работы создать собственный алгоритм отбора признаков для задач классификации данных, использование которого при предобработке данных позволит с большей точностью по сравнению с существующими алгоритмами случайного леса и выбора k наилучших признаков классифицировать объекты.

Основная часть. В качестве решения предлагается использовать алгоритм отбора признаков, основанный на генетическом алгоритме, в котором вероятность добавления признака при скрещивании оценивается с помощью mRMR (maximum relevance minimum redundacy) алгоритма, суть которого заключается в том, что мы стремимся выбрать подмножество признаков, которое имеет максимальную корреляцию (релевантность) с целевым значением, а признаки внутри подмножества имеют минимальную корреляцию (избыточность) друг с другом.

Основная идея генетического алгоритма при отборе признаков состоит в создании популяции, подмножеств признаков, выборе родителей, пар подмножеств, кроссовере, создании подмножества-ребенка, состоящего из признаков, содержащихся в родителях, и мутации, изменении подмножества-ребенка путем внесения признаков, не содержащихся в родителях, которые напоминают биологический процесс образования хромосом.

Далее представлен пошагово алгоритм отбора.

1. На первом шаге создается популяция из различных наборов признаков. Размер популяции равен количеству признаков в них. В [1] показано, что это количество является оптимальным. Количество отобранных признаков в каждом наборе на порядок меньше, чем в исходном наборе данных, на котором производится классификация. Признаки для индивидуумов в начальную популяцию выбираются случайным образом.
2. Кроме того, предварительно вычисляется качество признаков на основе алгоритма выбора k наилучших признаков и метрики, выбранной в качестве гиперпараметра, оценка качества признаков хранится в виде отсортированного массива.
3. Далее в цикле пока качество лучшего индивида из популяции не начнет стагнировать, производится скрещивание индивидов. Под стагнацией понимается отсутствие увеличения максимального показателя взвешенной F1–меры на протяжении трех итераций.

3.1. Для этого оценивается качество каждого индивида в популяции с помощью взвешенной F1-меры.

3.2. После выбираются пары индивидов, которые будут скрещиваться, количество потомков равно количеству предков в популяции. Предки с лучшим показателем взвешенной F1-меры с большей вероятностью будут скрещиваться.

3.3. Перед скрещиванием происходит мутация: из набора тех признаков, которых нет в предках, выбираются несколько признаков, которые будут добавлены в потомка, вероятность добавления признака зависит от качества в наборе признаков, вычисленном вначале.

Количество добавляемых признаков зависит от количества одинаковых признаков в каждом из двух скрещиваемых предков, а также от значения взвешенной F1-меры для каждого из предков. Чем таких признаков больше и чем меньше значение взвешенной F1-меры, тем больше признаков, не содержащихся в двух предках, будет добавлено потомку.
3.4. Далее идет добавление признаков из предков, вероятность добавления вычисляется на основе mRMR алгоритма на потомке.
4. После выхода из цикла, при возникновении стагнации, берется индивид из популяции, который показывает наилучший результат при классификации, это подмножество признаков и будет возвращено в качестве результирующего.
Результаты работы алгоритма сравнивались с результатами работы алгоритмов выбора k наилучших признаков и случайного леса из библиотеки Scikit-learn на наборе признаков The broken machine[2]. Было получено, что при использовании классификатора на основе случайного леса для вышеописанного алгоритма значение взвешенной F1-меры при классификации больше, чем при использовании алгоритмов выбора k наилучших признаков и случайного леса из библиотеки Scikit-learn.

Выводы. Описанный алгоритм можно использовать при предобработке наборов данных из различных отраслей в сочетании с грамотным выбором классификаторов для достижения более высокой, по сравнению с алгоритмами выбора k наилучших признаков и случайного леса из библиотеки Scikit-learn, точности классификации данных. Учитывая, что разработанный алгоритм был реализован в виде программного пакета и использует программный интерфейс, соответствующий библиотеке Scikit-learn, его можно с легкостью использовать для предобработки данных.

Список использованных источников:

1. Alander, J. On optimal population size of genetic algorithm [Text] / J. Alander // CompEuro 1992 Proceedings Computer Systems and Software Engineering (04-08 May 1992, The Hague, Netherlands)
2. The broken machine [Электронный ресурс]. — Режим доступа: https://www.kaggle.com/ivanloginov/the-broken-machine (25.12.2022)
3. Панченко, Т. Генетические алгоритмы [Текст]: учебно-методическое пособие / под ред. Ю. Ю. Тарасевича. – Астрахань: Издательский дом «Астраханский университет», 2007. – С. 5-6
4. Vora, S. A comprehensive study of Eleven Feature Selection Algorithms and their Impact on Text Classification [Text] / S. Vora, H. Yang // Computing Conference 2017 (18-20 July 2017, London, United Kingdom). - 2017. – P. 440 - 449.

Тисленко М. Д. (автор) Подпись

Паринger Р. А. (научный руководитель) Подпись