Сравнение алгоритмов отбора признаков для задач классификации данных

М.Д. Тисленко
Самарский национальный исследовательский университет им. академика С.П. Королева
Самара, Россия
makstislenko@gmail.com

А.В. Гайдель
Институт систем обработки изображений - филиал ФНИЦ «Кристаллография и фотоника» РАН
Самара, Россия
andrey.gaidel@gmail.com

***Аннотация*—В данной статье рассматриваются различные алгоритмы отбора признаков, производится сравнение качества классификации с использованием алгоритмов отбора и без их использования на различных наборах данных.**

Ключевые слова— отбор признаков, алгоритм, классификация.

1. Введение

Отбор признаков, как один из способов предварительной обработки данных, доказал свою эффективность и действенность при подготовке информации (особенно данных с большим числом признаков) для различных задач интеллектуального анализа данных и машинного обучения. В отбор признаков входит построение более простых и понятных моделей, улучшение качества интеллектуального анализа данных, производительности и подготовка более понятного для восприятия набора признаков [1].

Существует огромное количество различных методов определить лучшее подмножество признаков. Это связано с тем, что данная задача является NP-трудной, гарантировано оптимальное решение может быть найдено только путем полного перебора, который может занимать много времени при большом числе атрибутов [2].

Задачей данного исследования является сравнение различных способов отбора признаков для совершенствования выбора алгоритма для решения задач классификации данных.

2. Сравнение алгоритмов отбора признаков

Для сравнения различных способов отбора признаков было взято 4 набора данных [3]-[6]. В качестве классификатора было взято две модели: логистическая регрессия, которая работает быстро, по сравнению с другими моделями и случайный лес, который работает сравнительно точно относительно других моделей, но гораздо медленнее.

В исследовании сравниваются алгоритм SelectKBest с тремя критериями отбора: information gain, Хи-квадрат, F-статистика дисперсионного анализа и алгоритм отбора на основе случайного леса RandomForest. Алгоритм SelectKBest оценивает по отдельности каждый признак по одному из критериев, указанных выше, далее выбирается k лучших признаков. Данный алгоритм считается одним из самых простых и быстрых среди алгоритмов отбора признаков. RandomForest основан на построении так называемых деревьев решений, обучающихся на основе выбранного случайного подмножества признаков, таким образом, он относится к большой группе ансамблиевых методов классификации. После обучения результирующее значение будет браться как среднее от результата работы всех решающих деревьев. Данный метод позволяет производить достаточно точную классификацию, кроме того, можно явно указывать количество признаков, которое необходимо взять для классификации. Таким образом, с помощью данного алгоритма можно не просто классифицировать, но и успешно отбирать признаки. В данном исследовании этот алгоритм используется как в качестве классификатора, так и в качестве алгоритма отбора признаков.

Число выбранных свойств не превышает 10% от изначального количества. Оценка эффективности каждого из алгоритмов проводилась с помощью сравнения взвешенного среднего F-меры по каждому из возможных результирующих значений.

F-мера достигает наилучшего значения при 1 и наихудшего при 0, вычисляется по следующей формуле:

$$F=2\cdot(precision\cdot recall)/(precision+recall) \quad (1)$$

С помощью использования взвешенного среднего при оценке критерия или алгоритма можно принимать во внимание количество элементов, принадлежащих каждому из классов.

$$\bar{x} = \frac{\sum_k w_k x_k}{\sum_k w_k} \quad (2)$$

Где w_k - веса, а x_k - соответствующие этим весам значения. Весом в нашем случае является количество отнесенных к этому классу объектов, так как классов в каждом из наборов данных два, можно переписать формулу в следующем виде:

$$\bar{x} = (w_0x_0 + w_1x_1)/(w_0 + w_1) \quad (3)$$

Ниже в таблицах представлены результаты работы каждого алгоритма отбора признаков на наборах данных [3]-[6] и соответствующих классификаторах. На диаграммах по оси ординат представлены значения взвешенного среднего F-меры результатов классификации для каждого из наборов данных при использовании различных метод отбора и без них, по оси абсцисс указаны наборы данных, на которых проводится классификация.

На диаграмме видно, что наиболее качественная классификация с помощью логистической регрессии происходит при отборе признаков с помощью случайного леса, доля верно классифицированных объектов примерно на 2% выше, чем при классификации без отбора

признаков, а на наборе данных The broken machine взвешенное среднее F-меры выше более чем на 5%.

ТАБЛИЦА1. РЕЗУЛЬТАТЫ РАБОТЫ ПРИ ИСПОЛЬЗОВАНИИ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ

	Email spam	**Parkinson dicease**	**The broken machine**	**Financial indicators**
Без отбора	0.5751	0.5206	0.7617	0.9562
Хи-квадрат	0.5469	0.5078	0.7931	0.9474
F-статистика ANOVA	0.5837	0.5202	0.7961	0.9403
RandomForest	0.5934	0.5199	0.8146	0.9595
Взаимная информация	0.5	0.5262	0.7816	0.952

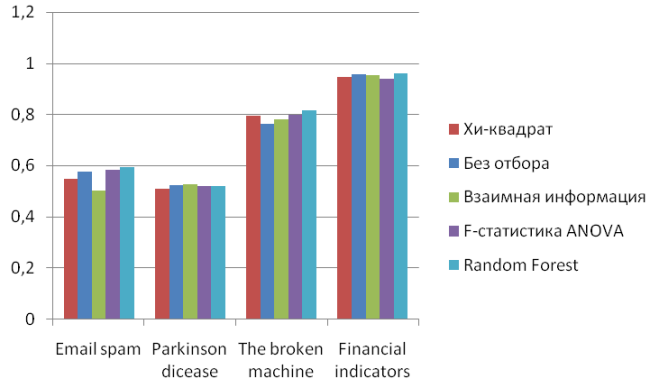

Рис. 1. Диаграмма работы различных методов с использованием логистической регрессии

На остальных наборах данных преимущество в доле правильно классифицированных объектов у RandomForest не такое большое. Видно, что использование критерия взаимной информации на наборе данных Email spam ухудшает качество классификации по сравнению с классификацией без отбора признаков. В целом результаты при использовании отбора и без него практически не отличаются, однако можно отметить, что стабильно качественная классификация также возможна при использовании критериев Хи-квадрат и F-статистики дисперсионного анализа.

На диаграмме видно, что наиболее качественная классификация с помощью случайного леса происходит при отборе признаков с помощью критерия взаимной информации, доля верно классифицированных объектов примерно на 2,5% выше, чем при классификации без отбора признаков.

ТАБЛИЦА 2. РЕЗУЛЬТАТЫ РАБОТЫ ПРИ ИСПОЛЬЗОВАНИИ RANDOM FOREST

	Email spam	**Parkinson dicease**	**The broken machine**	**Financial indicators**
Без отбора	0,61295	0,570025	0,830825	0,9614
Хи-квадрат	0,6217	0,5748	0,8704	0,9676
Взаимная информация	0,6304	0,6163	0,8667	0,964
F-статистика ANOVA	0,6012	0,5877	0,8507	0,9675
Random Forest	0,6199	0,5698	0,83	0,9604

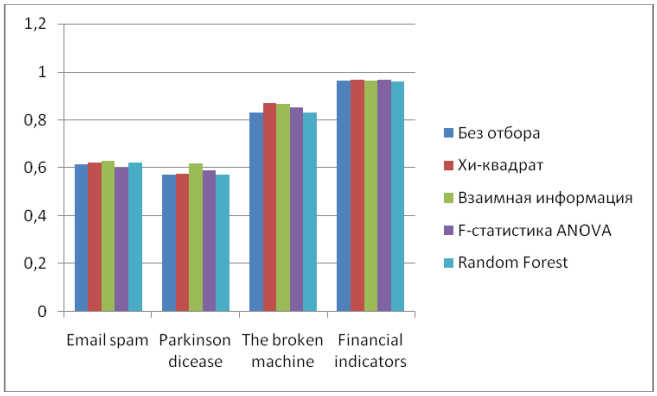

Рис. 2. Диаграмма работы различных методов с использованием классификатора Random Forest

На наборе данных Parkinson Disease этот статистический критерий позволяет улучшить результат классификации на 4,5% по сравнению с классификацией без использования отбора признаков. На остальных наборах данных также есть некоторые улучшения при классификации. Важно отметить, что отбор признаков с помощью RandomForest при использовании классификатора на основе случайного леса показывает результат хуже по сравнению со статистическими критериями, чем при использовании этого же классификатора в сочетании с алгоритмом SelectKBest. В целом результаты при использовании отбора и без него практически не отличаются, однако можно отметить, что на всех наборах данных достаточно качественная классификация происходит при использовании критерия Хи-квадрат и критерия взаимной информации.

3. ЗАКЛЮЧЕНИЕ

В результате работы были использованы различные алгоритмы и критерии отбора признаков для классификации данных. В среднем отбор признаков дает небольшое улучшение классификации. Важно отметить, что использование некоторых критериев на некоторых наборах данных может дать худшее качество классификации, чем было до этого без отбора признаков. На основании результатов из таблиц можно сделать вывод, что наиболее эффективным является отбор признаков с помощью RandomForest, результаты отбора с помощью этого алгоритма почти на всех наборах данных и классификаторах лучше, чем без использования отбора. Кроме того, отбор признаков этим методом дал наибольший прирост в взвешенного среднего F-меры по сравнению с классификацией без отбора признаков. В среднем доля верно классифицированных объектов увеличилась на 1% при использовании этого алгоритма. Также можно отметить, что стабильно высокое качество классификации возможно при использовании критерия Хи-квадрат в алгоритме выбора k наилучших признаков. Остальные критерии, которые можно использовать при выборе k наилучших признаков также в среднем показывают результаты, которые незначительно лучше, чем без отбора, однако при выборе определенных наборов данных и классификаторов качество классификации может быть на несколько процентов хуже, чем без отбора признаков.

Таким образом, можно сделать вывод, что отбор признаков с помощью случайного леса в общем случае при невозможности подробного анализа набора данных

покажет наилучший результат, однако может оказаться так, что классификация происходит точнее на данном наборе данных при выборе k наилучших с помощью какого–либо статистического критерия, однако выбор критерия требует подробного изучения набора данных, что не всегда осуществимо.

Благодарности

Работа выполнена при финансовой поддержке гранта РФФИ, проект № 20-51-05008 Arm_a

Литература

[1] Li, J. Feature Selection: A Data Perspective/ J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang, H. Liu // ACM Computing Surveys. – 2017. – Vol. 50 – P. 94-139.

[2] Ходашинский, И.А. Отбор классифицирующих признаков: сравнительный анализ бинарных метаэвристик и популяционного алгоритма с адаптивной памятью / И.А. Ходашинский, К.С. Сарин // Программирование. – 2019. – Т. 45, № 5. – С. 3-9. DOI: 10.1134/S0132347419050030.

[3] Email spam classification dataset [Electronic resource]. — Access mode: https://www.kaggle.com/balaka18/email-spam-classification-dataset-csv (05.02.2022).

[4] Parkinson disease speech signal features [Electronic resource]. — Access mode: https://www.kaggle.com/dipayanbiswas/parkinsons-disease-speech-signal-features (05.02.2022).

[5] The broken machine [Electronic resource]. — Access mode: https://www.kaggle.com/ivanloginov/the-broken-machine (05.02.2022).

[6] 200 + financial indicators of US stocks (2014-2018) [Electronic resource]. — Access mode: https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018 (05.02.2022).