DOI 10.37882/2223–2966.2021.02.23

ОПТИМИЗАЦИЯ ПРИЗНАКОВ В ЗАДАЧЕ РАСПОЗНАВАНИЯ ПОЛА ЧЕЛОВЕКА ПО ЕГО ПОЧЕРКУ

OPTIMIZATION OF FEATURES IN THE PROBLEM OF RECOGNIZING A PERSON'S GENDER BY HIS HANDWRITING

R. Myasoutov

Summary. In the task of recognizing the gender of a person by his handwriting the characteristics contained in any studied sample play an extremely important role. At the moment 3193 features have been collected that describe a person's handwriting. It is rather difficult to work with such a large number of features since their search and analysis takes a lot of time. The goal of this work is to optimize the number of features so as to leave the smallest number of features at which the accuracy of sex determination remains acceptable.

Keywords: feature optimization, dimensionality reduction, classification problem.

Мясоутов Рамиль Хамзяевич
Аспирант, Волгоградский Государственный Университет
ramilmyasoutov@yandex.ru

Аннотация. В задаче распознавания пола человека по его почерку крайне важную роль играют признаки, содержащиеся в каком — либо изучаемом образце. На данные момент собрано 3193 признаков описывающих почерк человека. Работать с таким большим количеством признаков довольно сложно, так как их поиск и анализ занимает большое количество времени. Цель данной работы заключается в том, чтобы оптимизировать количество признаков так, чтобы оставить наименьшее количество признаков, при котором точность определения пола остается приемлемой.

Ключевые слова: оптимизация признаков, уменьшение размерности, задача классификации.

Данная задача[1] сводится к задаче уменьшения размерности[2], которая подразумевает уменьшение количества признаков набора данных. Уменьшение набора признаков может быть осуществлено с помощью методов выбора признаков или выделения признаков. Методы выделения признаков работают так, что они из уже существующего набора признаков создают свой набор признаков. Такой формат противоречит цели данной работы, поэтому они не использовались при решении данной задачи. Методы выбора признаков можно разбить на несколько разных типов. В данной работе были рассмотрены фильтры, оберточные методы и встроенные методы.

Фильтры

Фильтры измеряют релевантность признаков на основе некой функции *f*, а затем по правилу K выбираются признаки, которые необходимо оставить в результирующем подмножестве. Для функции f были выбраны методы:

Хи-квадрат[3] — $X^2 = \sum \frac{(observed - expected)^2}{expected}$

F-критерий[4] — $F = \frac{\sigma_X^2}{\sigma_Y^2} \sim F(m-1, n-1)$

Для правил выбора K были использованы методы:

1. KBest[5] — K лучших значений
2. VarianceThreshold[6] — метод оценки дисперсии признака. Общий принцип звучит так, что признаки с почти нулевой дисперсией не являются значимыми, поэтому их можно удалить

Оберточные методы

Методы из данной категории работают на основе следующего принципа. Классификатор запускается на разном подмножестве признаков исходного тренировочного набора данных, после выбирается подмножество признаков с наилучшими параметрами на обучающей выборке, а затем выполняются проверки на тестовом наборе данных. Одним из преимуществ оберточных методов перед фильтрами является учитывание зависимостей между признаками. Одним из примеров таких методов является recursive feature elimination (RFE) [7]. В данной работе RFE был использован в паре с классификатором дерево решений (DecissionTreeClassifier)[8].

Встроенные методы

В данном подходе для начального множества признаков создается несколько подмножеств признаков, а затем эти группы пересекают так, чтобы получить набор самых релевантных признаков. В данной работе был использован случайный лес (RandomForestClassifier)[9]

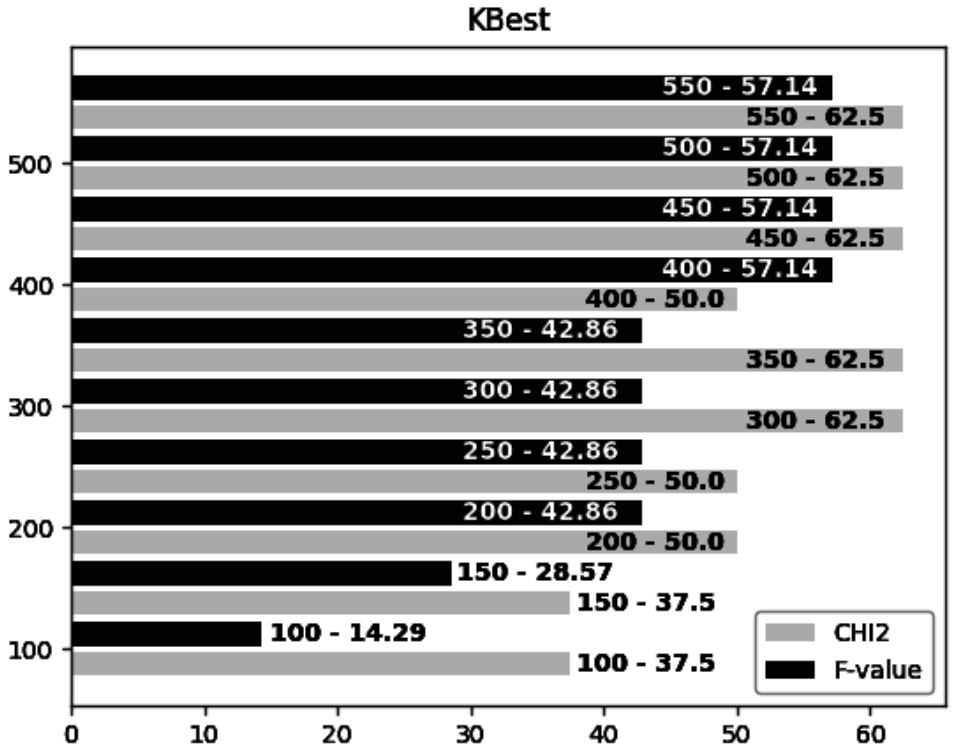
KBest
550 - 57.14
550 - 62.5
500 - 57.14
500 - 62.5
450 - 57.14
450 - 62.5
400 - 57.14
400 - 50.0
350 - 42.86
350 - 62.5
300 - 42.86
300 - 62.5
250 - 42.86
250 - 50.0
200 - 42.86
200 - 50.0
150 - 28.57
150 - 37.5
100 - 14.29
100 - 37.5
CHI2
F-value
500
400
300
200
100
0
10
20
30
40
50
60

Рис. 1

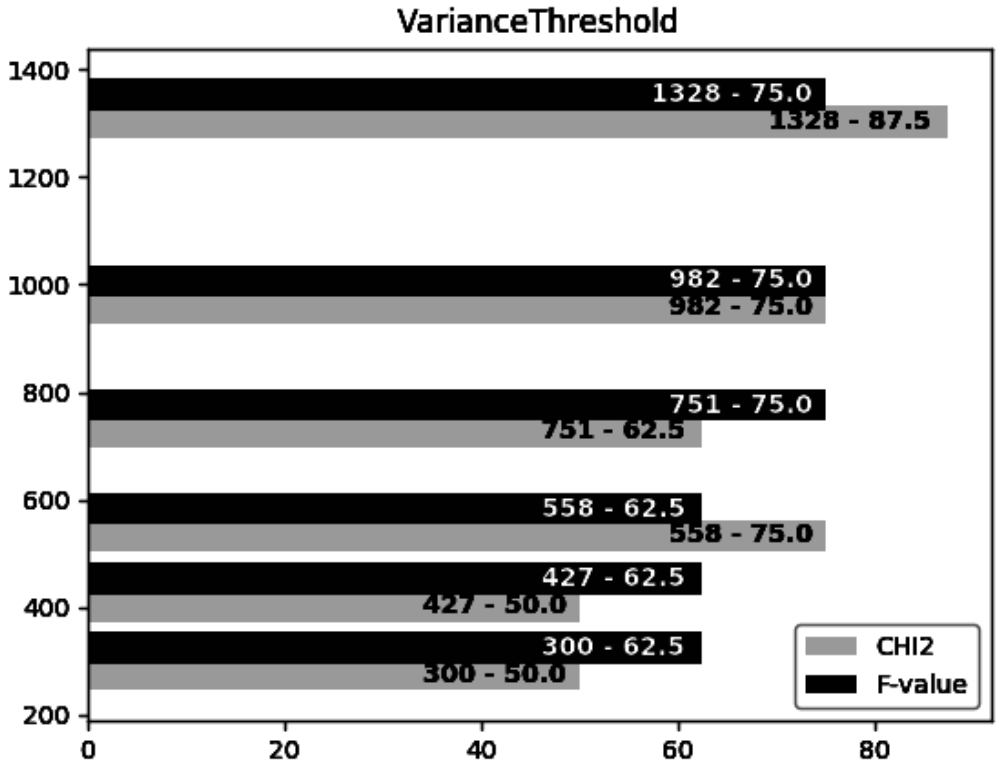
VarianceThreshold
1328 - 75.0
1328 - 87.5
982 - 75.0
982 - 75.0
751 - 75.0
751 - 62.5
558 - 62.5
558 - 75.0
427 - 62.5
427 - 50.0
300 - 62.5
300 - 50.0
CHI2
F-value
1400
1200
1000
800
600
400
200
0
20
40
60
80

Рис. 2

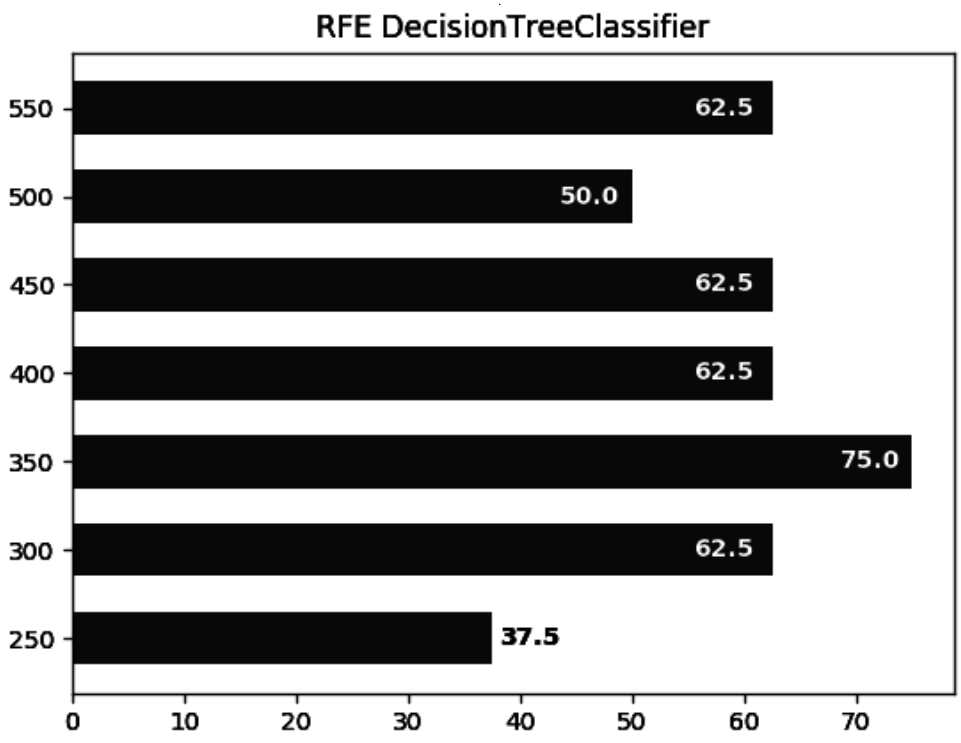

Рис. 3

Критерий выбора

Для сбора статистических данных для всех перечисленных методов использовался следующий подход:

- для всех признаков от 3193 до 0 с шагом 50 делать:
- инициализировать метод оптимизации с указанным количеством признаков
- выполнить оптимизацию признаков
- проверить работу классификатора для указанного набора признаков
- сохранить точность и признаки
- На текущий момент для определения пола используется следующая последовательность действий:
- представить изучаемый образец в виде вектора признаков $p = (0, 1, 0, 1, \ldots)$, где позиция 0 или 1 — наличие i-го признака,
- вычислить косинусное расстояние[10] между изучаемым образцом и векторами всех эталонных (обучающих) образцов,
- выбрать наиболее ближайший эталонный вектор,
- присвоить пол ближайшего эталонного образца изучаемому образцу.

Поэтому при решении поставленной задачи выбор признаков осуществлялся на базе обучающих образцов, а при проверке во всех векторах тестовых и обучающих образцах выбирались только те признаки, которые были получены на этапе оптимизации признаков. Сравнение производилось на базе уменьшенного вектора признаков. Так как основная цель работы заключается в подборе наименьшего количества признаков, при котором точность определения остается приемлемой. Тогда критерием выбора будет: минимальное количество признаков с максимальной точностью определения.

Рассмотрим результаты работы методов фильтрации признаков.

На рис. 1 отображено сравнение точности определения пола с количеством выбранных признаков функциями хи-квадрат (CHI2) и F-критерий (F-value), на оси абсцисс лежит точность распознавания, а на оси ординат — количество признаков. На рис. 1 представлено сравнение для метода выбора KBest, а на рис. 2 — для метода VarianceThreshold.

Судя по графикам, можно сделать вывод, что ограничение фильтрами не дает особо большую точность распознавания пола, так как с увеличением количества признаков повышается и точность, за счет покрытия большего множества признаков.

Рассмотрим работу оберточного метода RFE на базе DecissionTreeClassifier

На рис. 3 отображено сравнение точности определения пола с количеством выбранных признаков, на оси

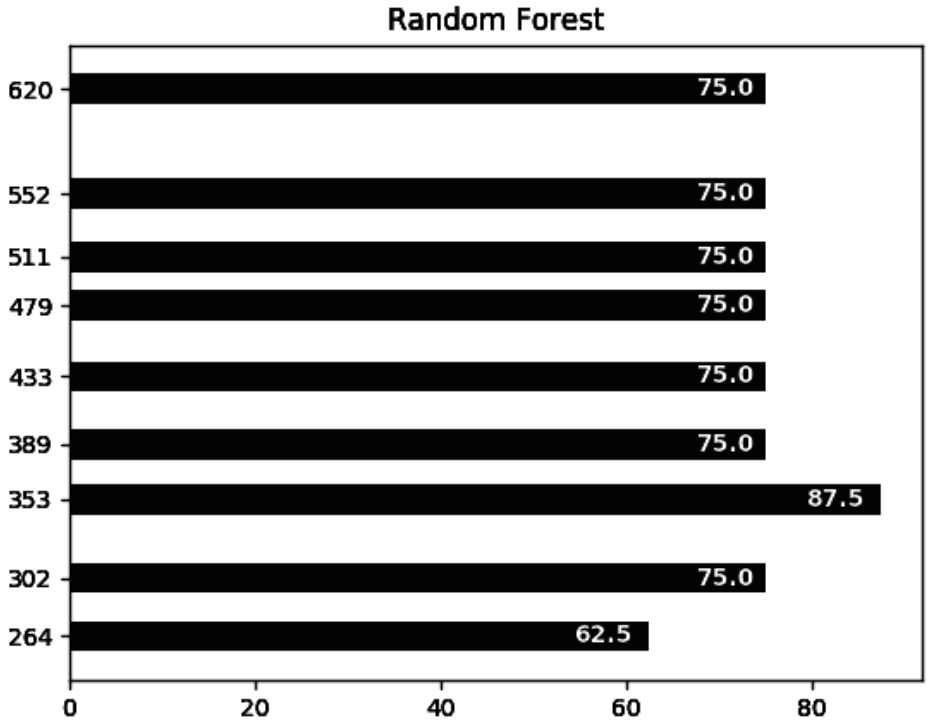

Рис. 4

Рис. 5.

абсцисс лежит точность распознавания, а на оси ординат — количество признаков. На данном графике видно, что при увеличении количества признаков не всегда увеличивается точность распознавания, а при выборе 350 определенных признаков точность равна 75%, а это говорит о том, что при определенной группе наиболее важных признаков, можно добиться большей точности.

Рассмотрим работу
встроенного метода случайный лес

На рис. 4 отображено сравнение точности определения пола с количеством выбранных признаков, на оси абсцисс лежит точность распознавания, а на оси ординат — количество признаков. Так же как и в случае оберточного метода, видно, что случайный лес позволил найти более удачные группы признаков, которые не теряют точности с увеличением количества признаков. При 353 признаках точность возросла до 87.5%.

Сравнение результатов

Сравним все лучшие варианты оптимизации признаков с учетом описанного выше критерия.

На рис. 5 отображено сравнение лучших методов определения пола, на оси абсцисс лежит количество признаков, а на оси ординат — методы выбора признаков, на каждой из шкал указана точность определения с указанным количеством признаков. Судя по графику видно, что со всеми текущими признаками точность определения составляет 87.5%. Если

сократить количество признаков до 1328 с помощью фильтра VarianceThreshold с функцией хи-квадрат, то точность также останется на уровне 87.5%. В то же время можно пойти дальше и сократить количество признаков до 353 с помощью случайного леса, не потеряв точность определения. Исходя из всего вышеперечисленного видно, что для задачи определения стоит рассматривать методы оптимизации, которые учитывают связи между признаками. Следовательно, фильтры не подходят для решения текущей задачи, поэтому необходимо рассматривать встроенные и оберточные методы. У оберточных методов высок риск переобучения, а встроенные методы уменьшают данный риск, но так как полученный набор признаков отбирается на основе знаний о классификаторе, то при смене классификатора выбранное множество признаков может не быть релевантным. На текущий момент было принято решение остановиться на множестве признаков, которое было выбрано с помощью случайного леса, так как с учетом определения пола на основе косинусного расстояния между векторами признаков точность на данном множестве признаков остается высокой. В дальнейшем, при увеличении обучающей выборки с полным множеством признаков, можно повторить выбор признаков, чтобы подобрать более удачное множество признаков.

ЛИТЕРАТУРА

1. Мясоутов Р. Х. Создание системы распознавания пола человека по его почерку. // Современная наука: актуальные проблемы теории и практики. Серия «Естественные и технические науки» -№ 07,—2020, с. 86–90
2. Уменьшение размерности. URL // Университет ИТМО. Конспекты: https://neerc.ifmo.ru/wiki/index.php?title=Уменьшение_размерности (дата обращения — 10.01.2021)
3. Хи-квадрат // Sklearn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html (дата обращения — 10.01.2021)
4. F-критерий // Sklearn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html (дата обращения — 10.01.2021)
5. KBest // Sklearn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html (дата обращения — 10.01.2021)
6. VarianceThreshold // Sklearn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html (дата обращения — 10.01.2021)
7. Jason Brownlee. Recursive Feature Elimination (RFE) for Feature Selection in Python // Machine Learning Mastery. 2020. URL: https://machinelearningmastery.com/rfe-feature-selection-in-python/ (дата обращения — 10.01.2021)
8. DecissionTreeClassifier // Sklearn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (дата обращения — 10.01.2021)
9. RandomForestClassifier // Sklearn. URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (дата обращения — 10.01.2021)
10. Косинусное сходство // Wikipedia. URL: https://ru.wikipedia.org/wiki/Векторная_модель#Косинусное_сходство (дата обращения — 10.01.2021)

Журнал «Современная наука: актуальные проблемы теории и практики»