СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ ОТБОРА ПРИЗНАКОВ НА ДЕСЯТИ НАБОРАХ ДАННЫХ ДЛЯ КЛАССИФИКАЦИИ

Г.Х. Ташматов

Ташкентский университет информационных технологий имени Мухаммада ал-Хоразмий, Ташкент, Узбекистан

(E-mail: ganijon.tashmatov@gmail.com)

Аннотация

Отбор признаков применяется для снижения размерности данных, повышения интерпретируемости моделей и уменьшения вычислительных затрат. В работе сравниваются пять алгоритмов отбора признаков и базовый вариант без отбора на десяти наборах данных классификации и трех классификаторах. Всего оценено 180 конфигураций «набор данных — селектор — модель». Наилучший средний результат показал RF_Importance_SFM со средней сбалансированной точностью 0.8296. Лучшим классификатором в среднем стала логистическая регрессия со значением Balanced Accuracy 0.8378. Применение отбора позволило снизить среднее число признаков с 37.1 до 13.3, то есть примерно на 64.2%, при сохранении конкурентного качества классификации.

Ключевые слова: отбор признаков; машинное обучение; ANOVA F-test; взаимная информация; mRMR; Random Forest; Extra Trees; классификация.

1. Введение

Современные задачи классификации часто содержат нерелевантные, избыточные или шумовые переменные. Отбор признаков решает эту

проблему путем выбора подмножества входных характеристик перед обучением модели. Основные цели такого подхода — повышение качества прогноза, сокращение времени обучения и улучшение интерпретируемости результата. Цель данной работы состоит в сравнении нескольких алгоритмов отбора признаков по качеству классификации, степени сокращения признакового пространства и вычислительным затратам.

2. Материалы и методы

Исследование включало 10 наборов данных, 6 условий отбора признаков с учетом базового варианта и 3 классификатора, что дало 180 экспериментальных конфигураций. Для воспроизводимости использовалось значение random seed = 42. Каждый набор данных разделялся на обучающую и тестовую части в соотношении 70/30 с сохранением распределения классов. Масштабирование и селектор признаков обучались только на обучающей части, после чего тестовая выборка преобразовывалась уже обученными компонентами.

Сравнивались следующие варианты: NoSelection, ANOVA_F_SelectKBest, MutualInfo_SelectKBest, MRMR_MI_Corr, RF_Importance_SFM и ExtraTrees_Importance_SFM. В качестве моделей использовались Logistic Regression, Random Forest и Gaussian Naive Bayes. Целевое число отобранных признаков определялось по формуле:

$$k = \min(30, \max(2, \text{ceil}(0.35 \cdot p))).$$

Для наборов с пятью и менее признаками использовалось значение k = floor(p/2), но не менее одного признака. Основной метрикой выступала Balanced Accuracy, дополнительно учитывались Accuracy, macro F1 и время выполнения.

Таблица 1. Характеристика наборов данных

Набор данных	Тип	N	p	Классы	Мин. доля	k
BreastCancerWisconsin	реальный	569	30	2	0.373	11
Wine	реальный	178	13	3	0.270	5
Digits	реальный	600	64	10	0.097	23
Iris	реальный	150	4	3	0.333	2
SynthLinear20	синтетический	320	20	2	0.497	7
SynthRedundant60	синтетический	340	45	3	0.332	16
SynthNoisy100	синтетический	260	50	2	0.481	18
SynthImbalanced40	синтетический	340	40	2	0.171	14
SynthMulticlass80	синтетический	260	45	4	0.242	16
SynthSmallNHighP120	синтетический	160	60	2	0.500	21

3. Результаты

Среди всех наборов данных и классификаторов метод RF_Importance_SFM достиг наибольшей средней Balanced Accuracy — 0.8296. Второе место занял ANOVA_F_SelectKBest со значением 0.8240. Базовый вариант без отбора признаков остался конкурентоспособным, что показывает: польза отбора зависит от структуры конкретного набора данных.

Таблица 2. Общая эффективность методов отбора признаков

Селектор	Acc	BalAcc	F1	Признаки	Время
RF_Importance_SFM	0.8479	0.8296	0.8319	13.3000	0.1125
ANOVA_F_SelectKBest	0.8430	0.8240	0.8243	13.3000	0.0595
NoSelection	0.8396	0.8205	0.8207	37.1000	0.0599
MutualInfo_SelectKBest	0.8360	0.8181	0.8186	13.3000	0.2918
ExtraTrees_Importance_SFM	0.8329	0.8125	0.8132	13.3000	0.1162
MRMR_MI_Corr	0.8259	0.8034	0.8039	13.3000	0.3027

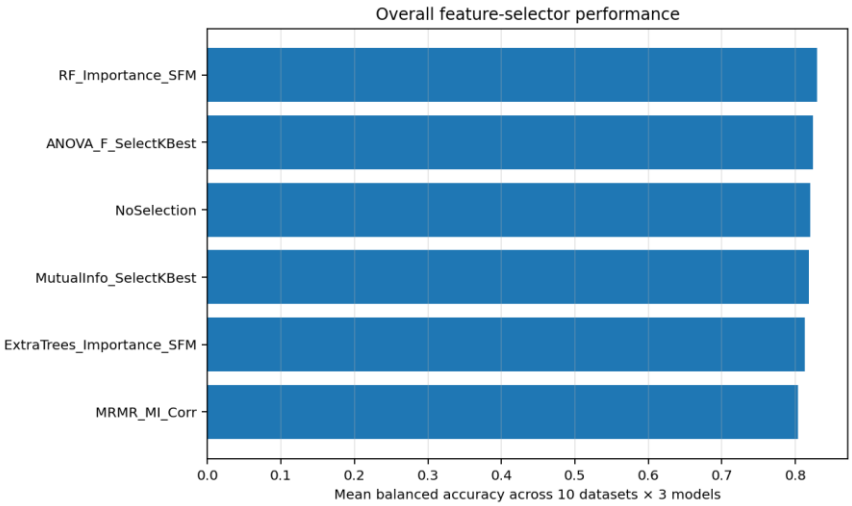


Рис. 1. Средняя Balanced Accuracy методов отбора признаков.

Логистическая регрессия оказалась лучшим классификатором в среднем по всем конфигурациям: Balanced Accuracy = 0.8378. Далее следуют Gaussian Naive Bayes и Random Forest.

Таблица 3. Общая эффективность классификаторов

Модель	Acc	BalAcc	F1	Время
LogisticRegression	0.8544	0.8378	0.8418	0.1366
GaussianNB	0.8297	0.8172	0.8171	0.1315
RandomForest	0.8286	0.7990	0.7974	0.2032

Таблица 4. Лучшая конфигурация для каждого набора данных

Набор данных	Селектор	Модель	BalAcc	F1	Acc	Призн.	Время
BreastCancerWisconsin	NoSelection	LogReg	0.9875	0.9875	0.9883	30	0.0234
Digits	NoSelection	LogReg	0.9556	0.9551	0.9556	64	0.1047
Iris	ANOVA_F	RandomForest	0.9556	0.9554	0.9556	2	0.0995
SynthImbalanced40	NoSelection	GaussianNB	0.7706	0.7908	0.8922	40	0.0094
SynthLinear20	ANOVA_F	RandomForest	0.9062	0.9062	0.9062	7	0.1130
SynthMulticlass80	RF_SFM	LogReg	0.7053	0.7045	0.7051	16	0.1259

SynthNoisy100	ExtraTrees_SFM	LogReg	0.7697	0.7692	0.7692	18	0.0626
SynthRedundant60	ANOVA_F	LogReg	0.9314	0.9315	0.9314	16	0.0258
SynthSmallNHighP120	NoSelection	GaussianNB	0.7917	0.7913	0.7917	60	0.0081
Wine	NoSelection	RandomForest	1.0000	1.0000	1.0000	13	0.0900

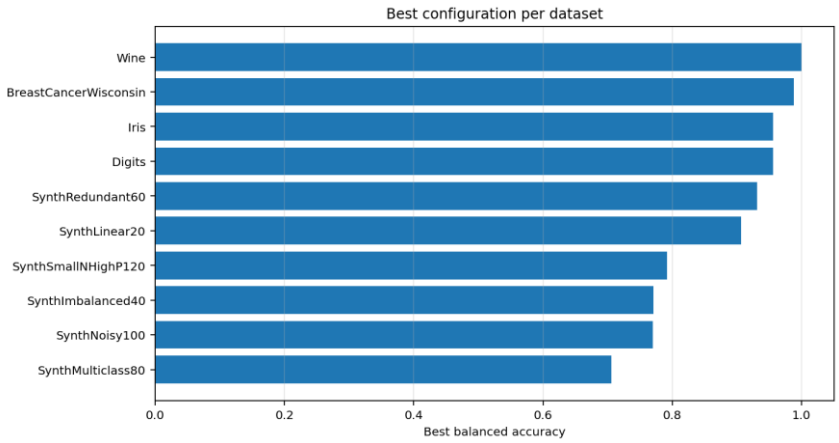


Рис. 2. Лучшие конфигурации по наборам данных.

В среднем базовый вариант сохранял 37.1 признака. Пять алгоритмов отбора оставляли примерно 13.3 признака, что соответствует сокращению размерности примерно на 64.2%.

Таблица 5. Эффективность сокращения признакового пространства

Селектор	Acc	BalAcc	F1	Призн.	Доля	Сокр., %	Время

RF_Importance_SFM	0.8479	0.8296	0.8319	13.3000	0.3585	64.1509	0.1125
ANOVA_F_SelectKBest	0.8430	0.8240	0.8243	13.3000	0.3585	64.1509	0.0595
NoSelection	0.8396	0.8205	0.8207	37.1000	1.0000	0.0000	0.0599
MutualInfo_SelectKBest	0.8360	0.8181	0.8186	13.3000	0.3585	64.1509	0.2918
ExtraTrees_Importance_SFM	0.8329	0.8125	0.8132	13.3000	0.3585	64.1509	0.1162
MRMR_MI_Corr	0.8259	0.8034	0.8039	13.3000	0.3585	64.1509	0.3027

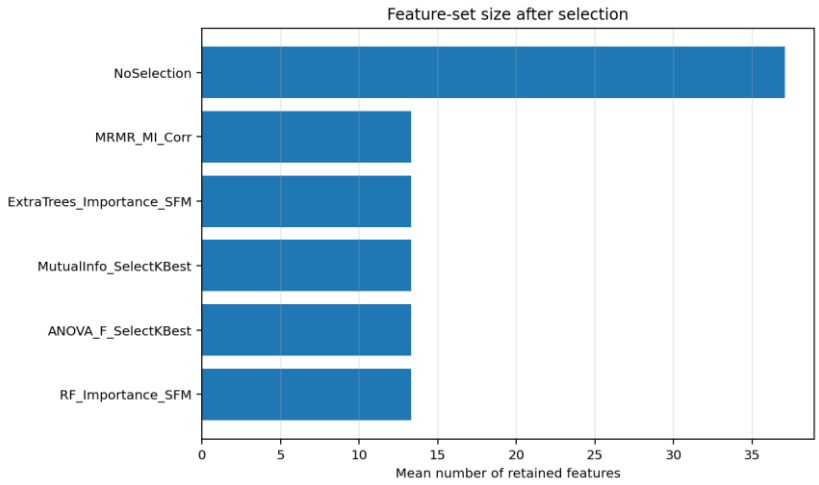


Рис. 3. Среднее число сохраненных признаков.

Матрица взаимодействия «модель — селектор» показывает, что ни один алгоритм не доминирует во всех сочетаниях. Для Logistic Regression наиболее сильным оказался RF_Importance_SFM, тогда как для GaussianNB конкурентным остался вариант без отбора признаков.

Таблица 6. Матрица Balanced Accuracy для пар «модель — селектор»

Модель	ANOVA_F	ExtraTrees	MRMR	MutualInfo	NoSelection	RF_SFM
GaussianNB	0.8128	0.8081	0.8095	0.8211	0.8301	0.8218
LogisticRegression	0.8477	0.8345	0.8226	0.8392	0.8291	0.8537
RandomForest	0.8115	0.7948	0.7779	0.7939	0.8023	0.8134

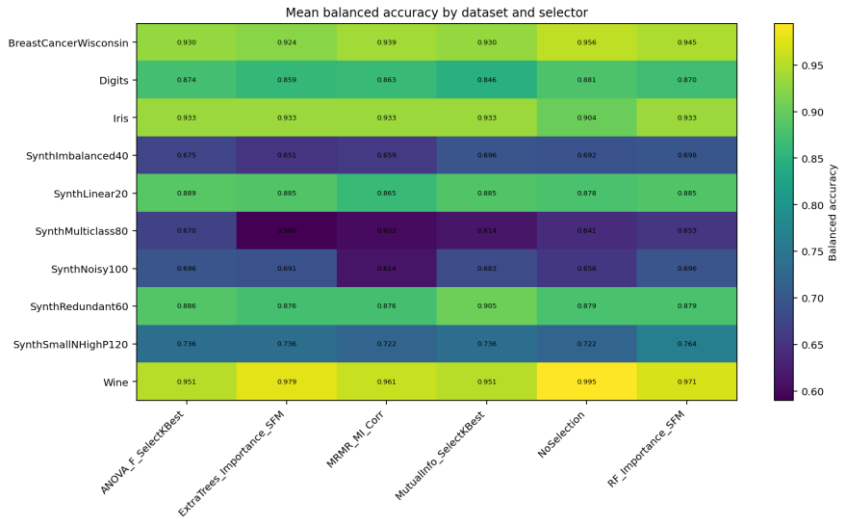


Рис. 4. Тепловая карта Balanced Accuracy по наборам данных и селекторам.

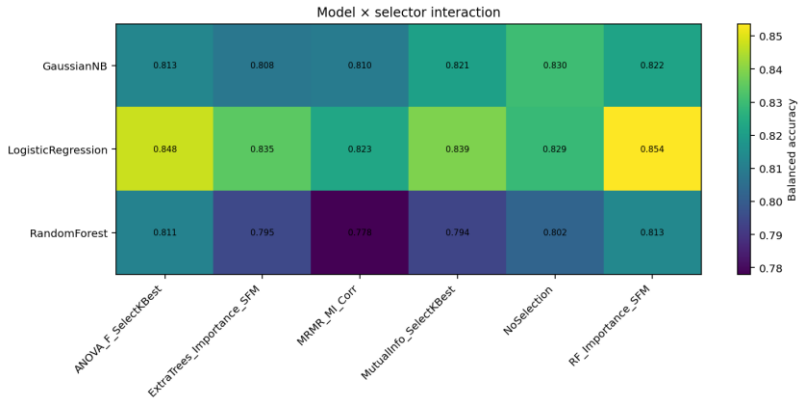


Рис. 5. Взаимодействие модели и метода отбора признаков.

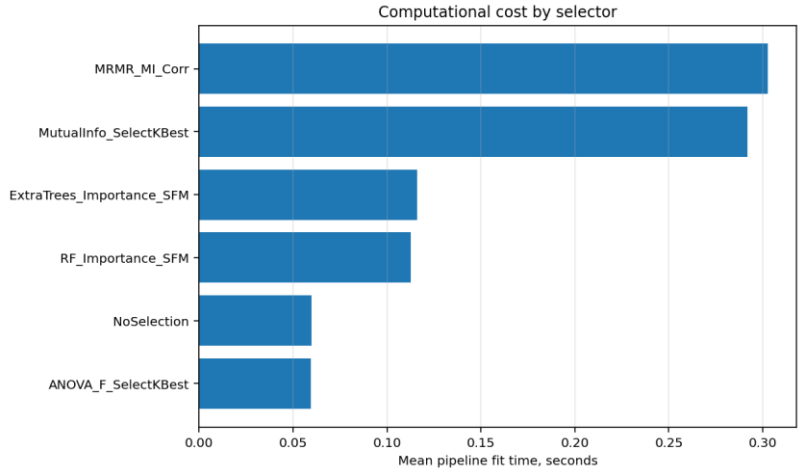


Рис. 6. Вычислительные затраты методов отбора признаков.

4. Обсуждение

Полученные результаты отражают практический компромисс, а не наличие универсально лучшего метода. Random-Forest importance selection обеспечил максимальную среднюю Balanced Accuracy, ANOVA F-score оказался сильным и быстрым фильтровым базовым методом, а NoSelection сохранил конкурентность на компактных или уже информативных признаковых пространствах. Отбор был особенно полезен в условиях шумовых и избыточных измерений.

Метод MRMR_MI_Corr не стал лучшим в среднем. Возможная причина состоит в том, что для ускорения использовалась взаимная информация с

целевой переменной и корреляционная оценка избыточности, а не полная попарная взаимная информация между признаками. Поэтому точные варианты mRMR могут показать иное поведение на отдельных задачах.

5. Ограничения исследования

- использован один стратифицированный holdout split, а не повторная кросс-валидация;
- шесть из десяти наборов являются синтетическими, хотя они воспроизводимы и отражают разные условия отбора;
- набор Digits был уменьшен до 600 наблюдений для ускорения эксперимента;
- гиперпараметры моделей фиксировались и не проходили расширенную оптимизацию;
- реализация mRMR применяла корреляционную оценку избыточности для ускорения вычислений.

6. Заключение

В работе проведено сравнение пяти алгоритмов отбора признаков и базового варианта без отбора на десяти наборах данных и трех классификаторах. Лучшим селектором по средней Balanced Accuracy стал RF_Importance_SFM, а лучшей моделью — Logistic Regression. В среднем методы отбора сократили размерность примерно на 64.2%, сохранив конкурентное качество классификации. Практический рабочий процесс должен включать базовый вариант без отбора, быстрый одномерный фильтр и встроенный метод на основе важности признаков деревьев решений.

Список литературы

1. Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. 2003. Vol. 3. P. 1157–1182. DOI: 10.1162/153244303322753616.
2. Peng H., Long F., Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005. Vol. 27, No. 8. P. 1226–1238.
3. Chandrashekar G., Sahin F. A survey on feature selection methods // Computers & Electrical Engineering. 2014. Vol. 40, No. 1. P. 16–28.
4. Pedregosa F. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. 2011. Vol. 12. P. 2825–2830.
5. scikit-learn developers. Feature selection documentation: SelectKBest, SelectFromModel and related selectors. 2026.
6. scikit-learn developers. Dataset loading utilities and bundled toy datasets. 2026.