УДК 004.8

Д.С. Ткаченко

студент, Северо-Кавказский федеральный университет, г. Ставрополь, Россия

З.М. Альбекова

канд. пед. наук, доцент Департамента цифровых, робототехнических систем и электроники Института перспективной инженерии, Северо-Кавказский федеральный университет, г. Ставрополь, Россия

Ф.В. Самойлов

канд. техн. наук, доцент Департамента цифровых, робототехнических систем и электроники Института перспективной инженерии, Северо-Кавказский федеральный университет, г. Ставрополь, Россия

МНОГОКРИТЕРИАЛЬНАЯ МЕТОДИКА ОЦЕНКИ МЕТОДОВ ОТБОРА ПРИЗНАКОВ ДЛЯ МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ

Аннотация. Предложена многокритериальная методика сравнения методов отбора признаков для задач многоклассовой классификации. Рассмотрены три семейства методов — фильтрующие (взаимная информация), обёрточные (рекурсивное исключение признаков) и встроенные (важность признаков случайного леса) — на шести разнородных наборах данных. Эксперименты выполнены в рамках повторяемой вложенной кросс-валидации, исключающей утечку данных. Для интегральной оценки предложен обобщённый показатель, объединяющий точность, F1-меру, устойчивость отбора, степень редукции признакового пространства и время вычислений, с тремя сценариями весов. Статистический анализ проводился критерием Фридмана и попарными тестами Уилкоксона с поправкой Холма. Парето-анализ показал, что встроенный метод оптимален на всех рассмотренных наборах данных: он достигает статистически неразличимого качества с обёрточным методом при кратном

выигрыше по скорости. Сформулированы практические рекомендации по выбору метода в зависимости от приоритетов предметной области.

Annotation. A multicriteria methodology for comparing feature selection methods in multiclass classification is proposed. Three method families are considered — filter (mutual information), wrapper (recursive feature elimination) and embedded (random forest importance) — across six heterogeneous datasets. Experiments follow a repeated nested cross-validation protocol preventing data leakage. An aggregate index integrating accuracy, F1-score, selection stability, dimensionality reduction and computation time is introduced with three weighting scenarios. Statistical analysis relies on the Friedman test and pairwise Wilcoxon tests with Holm correction. Pareto analysis shows that the embedded method is optimal on all datasets: it achieves quality statistically indistinguishable from the wrapper method while being several times faster. Practical guidelines on method choice depending on application priorities are provided.

Ключевые слова: отбор признаков; многокритериальная оптимизация; вложенная кросс-валидация; Парето-фронт; устойчивость отбора; взаимная информация; случайный лес

Keywords: feature selection; multicriteria optimization; nested cross-validation; Pareto front; selection stability; mutual information; random forest

Введение

Высокая размерность признакового пространства — одна из ключевых проблем анализа данных. При числе признаков, сопоставимом с числом наблюдений, классические алгоритмы теряют обобщающую способность и переобучаются на неинформативных переменных. Отбор признаков решает эту проблему, сохраняя физический смысл исходных переменных, что критически важно для интерпретируемых решений в медицине, биоинформатике и финансах.

Большинство исследований оценивает методы отбора по единственному критерию — точности классификации, игнорируя противоречивые показатели: время вычислений, устойчивость отобранного подмножества и степень редукции размерности. Многие работы не используют вложенную кросс-валидацию, из-за чего оценки точности оказываются оптимистично смещёнными вследствие утечки данных. Цель работы — разработать многокритериальную методику, учитывающую все перечисленные показатели в едином интегральном критерии, и апробировать её на разнородных наборах данных.

Методика

Принята классическая таксономия методов отбора. Фильтрующие методы оценивают релевантность признаков независимо от модели; в работе использована взаимная информация. Обёрточные методы переобучают модель на каждом шаге; реализовано рекурсивное исключение признаков RFE на основе случайного леса. Встроенные методы получают важности признаков в процессе обучения; использована impurity-based важность случайного леса. Для каждого селектора обучено три классификатора — логистическая регрессия, случайный лес и SVM, всего девять комбинаций на каждый набор данных.

Интегральный показатель Q_m — взвешенная сумма пяти нормированных критериев: точности, макро-усреднённой F1-меры, устойчивости отбора (индекс Жаккара), степени редукции и времени вычислений с обратным нормированием. Рассмотрены три сценария весов: сбалансированный, с приоритетом точности и с приоритетом скорости (таблица 1). Дополнительно выполнен Парето-анализ в двумерном и трёхмерном пространствах критериев, не требующий задания весов и устраняющий субъективность выбора.

Таблица 1 — Сценарии весовых коэффициентов показателя Q_m

Сценарий	Accuracy	F1	Устойчивость	Редукция	Время
Сбалансированный	0,20	0,20	0,20	0,20	0,20
Приоритет точности	0,35	0,25	0,15	0,15	0,10

Приоритет скорости	0,25	0,20	0,15	0,20	0,20

Эксперименты выполнены на шести открытых наборах данных: Mobile Price Classification (2000 образцов, 4 класса), Cardiotocography (2126, дисбаланс классов), Image Segmentation (210, 7 классов — проверка устойчивости при дефиците данных), Wine Quality (6497, бинарная задача), Vehicle Silhouettes (845) и Dry Bean (13 611). Признаки стандартизировались по обучающей части каждого фолда, что исключает утечку информации.

Протокол включал повторяемую вложенную кросс-валидацию RepeatedStratifiedKFold 3×5: во внешнем цикле оценивалась обобщающая способность, во внутреннем подбирались число признаков k из сетки {2, …, 15} и гиперпараметры классификаторов. Всего получено 810 оценок качества; детерминированные зерна генератора обеспечивают полную воспроизводимость. Статистический анализ включал критерий Фридмана и попарные тесты Уилкоксона с поправкой Холма.

Результаты

Наилучшая точность по датасетам достигала 0,84–0,99, причём встроенный метод входил в число лидеров на всех шести наборах данных. Кривые зависимости точности от числа признаков демонстрируют резкий рост до $k \approx 5$–10 с последующим плато (рисунок 1). Для крупных датасетов оптимальное k расположено у верхней границы сетки: информативность распределена по всем признакам, и агрессивная редукция размерности на таких данных не оправдана.

Рисунок 1 — Зависимость точности классификации от числа отобранных признаков

Критерий Фридмана не выявил значимых различий между методами по точности ($p = 0{,}070$), что подтверждается и попарными тестами: расхождение между RFE и важностью случайного леса пренебрежимо мало ($p = 0{,}93$, эффект Клиффа $-0{,}02$). При этом вычислительные затраты различаются драматически: время настройки RFE превышает время встроенного метода в 5,9 раза в медиане и до 13,3 раза в максимуме (таблица 2).

Таблица 2 — Результаты по датасетам

Датасет	Лучшая комбинация	Accuracy	F1	Время, с
Mobile	Embedded + LR	0,986	0,962	2,0
Cardiotocography	Embedded + RF	0,955	0,921	4,4
Image Segmentation	Filter + RF	0,971	0,956	3,8
Wine Quality	Wrapper + RF	0,839	0,826	13,7
Vehicle	Filter + SVM	0,858	0,859	0,5
Dry Bean	Wrapper + SVM	0,938	0,948	22,3

Парето-анализ идентифицировал встроенный метод rf_importance как Парето-оптимальный на всех шести датасетах в двумерном и трёхмерном пространствах: не существует метода, превосходящего его одновременно по всем

критериям, — любой выигрыш в точности оплачивается непропорционально большими вычислительными затратами. Ранжирование по Q_m устойчиво к изменению весов в сбалансированном сценарии и сценарии приоритета точности.

Обсуждение

Основной вывод: с точки зрения чистой точности выбор метода отбора признаков статистически незначим, и решающую роль приобретают вторичные критерии. Встроенный метод целесообразно рекомендовать как подход по умолчанию: он обеспечивает ту же точность, что и обёрточный, но на порядок быстрее и демонстрирует более высокую устойчивость отбора благодаря учёту взаимосвязей между признаками. Фильтрующие методы оправданы при жёстких ограничениях по времени, обёрточные — лишь в задачах с максимальными требованиями к интерпретируемости и без ограничений на вычислительные ресурсы.

Ограничения работы: наборы данных содержат не более 21 признака, что не покрывает задачи сверхвысокой размерности; для встроенного метода использована единственная мера важности — по примесям, тогда как permutation importance или SHAP могут дать иные ранжирования; сетка k не включала полное число признаков для трёх датасетов. Дальнейшие исследования предполагают расширение базы, включение дополнительных методов (LASSO, Boruta, SHAP-based) и разработку адаптивного выбора селектора по мета-признакам данных.

Заключение

Разработана и апробирована многокритериальная методика оценки методов отбора признаков, объединяющая интегральный показатель Q_m, строгий протокол вложенной кросс-валидации и Парето-анализ. Установлено, что методы трёх семейств статистически эквивалентны по точности, а встроенный метод важности случайного леса Парето-оптимален на всех шести

датасетах при кратном выигрыше по скорости. Сформулированы практические рекомендации по выбору метода в зависимости от ограничений задачи.

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.

2. Chandrashekar G., Sahin F. A Survey on Feature Selection Methods // Computers & Electrical Engineering. – 2014. – Vol. 40. – No. 1. – P. 16–28.

3. Kuncheva L.I. A Stability Index for Feature Selection // IEEE Transactions on Knowledge and Data Engineering. – 2007. – Vol. 19. – No. 10. – P. 1346–1356.

4. Cawley G.C., Talbot N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation // Journal of Machine Learning Research. – 2010. – Vol. 11. – P. 2079–2105.

5. Breiman L. Random Forests // Machine Learning. – 2001. – Vol. 45. – P. 5–32.

6. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.

References

1. Guyon I., Elisseeff A. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. – 2003. – Vol. 3. – P. 1157–1182.

2. Chandrashekar G., Sahin F. A Survey on Feature Selection Methods // Computers & Electrical Engineering. – 2014. – Vol. 40. – No. 1. – P. 16–28.

3. Kuncheva L.I. A Stability Index for Feature Selection // IEEE Transactions on Knowledge and Data Engineering. – 2007. – Vol. 19. – No. 10. – P. 1346–1356.

4. Cawley G.C., Talbot N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation // Journal of Machine Learning Research. – 2010. – Vol. 11. – P. 2079–2105.

5. Breiman L. Random Forests // Machine Learning. – 2001. – Vol. 45. – P. 5–32.

6. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine Learning in Python // Journal of Machine Learning Research. – 2011. – Vol. 12. – P. 2825–2830.