УДК 004.8

К ВОПРОСУ ОБОБЩЕННОГО АЛГОРИТМА РЕАЛИЗАЦИИ ФИЛЬТРАЦИОННЫХ МЕТОДОВ ОТБОРА ПРИЗНАКОВ В ЗАДАЧАХ РЕГРЕССИИ

Черемухин А. Д. (НГИЭУ), **Колбанев М. О.** (СПбГУ ЛЭТИ)
Научный руководитель – доктор технических наук, профессор Колбанев М.О. (СПбГУ ЛЭТИ)

Введение.

Отбор признаков в задачах регрессии является ключевым этапом построения корректных и устойчивых моделей, поскольку позволяет уменьшить размерность пространства факторов, снизить влияние нерелевантных и избыточных переменных, повысить интерпретируемость результатов и качество обобщения при ограниченном объёме наблюдений. При этом описано большое число алгоритмов и модификаций отбора признаков; однако на практике это разнообразие слабо поддержано программными реализациями. Соответственно, возникает необходимость разработки таксономии методов отбора признаков, ориентированной на формализацию их общих структурных компонентов и на построение обобщённого алгоритма реализации. В настоящей работе в качестве объекта исследования рассматриваются фильтрационные методы отбора признаков как наиболее универсальный и вычислительно эффективный класс подходов, допускающий систематизацию по используемым мерам качества и правилам формирования поднабора признаков без привязки к конкретному регрессионному алгоритму.

Основная часть.

Под фильтрационным методом отбора признаков в задачах регрессии будем понимать алгоритм, реализующий две последовательно выполняемые процедуры:

- вычисление меры качества (релевантности) каждого признака по отношению к зависимой переменной на основе исходных данных;
- формирование поднабора признаков на основе полученного вектора оценок по заданному правилу.

Формально фильтрационный метод может быть представлен в виде пары: отображение, ставящее в соответствие каждому признаку числовую оценку качества; и правило выбора подмножества признаков на основе этих оценок.

Принципиальными особенностями фильтрационного подхода являются:

- отсутствие итеративного пересчёта качества поднабора на основе обучения регрессионной модели (в отличие от обёрточных методов);
- отсутствие встроенной процедуры (в отличие от встроенных методов) штрафа и других механизмов, интегрированных в процесс обучения модели (в отличие от встроенных методов);
- независимость процедуры отбора от конкретного алгоритма регрессии.

В задачах регрессии ключевым элементом фильтрационного отбора является вычисление числовой меры связи между независимой и зависимой переменными - функционал, который по выборке наблюдений возвращает одно числовое значение, характеризующее степень статистической зависимости между переменными

Используемые современные меры связи можно сгруппировать по нескольким концептуальным направлениям.

1. Корреляционные и ранговые меры.[1]
2. Метрические и геометрические меры. Интерпретируют зависимость как согласованность геометрической структуры выборок.[2]
3. Энтропийные и информационные меры. Трактуют зависимость как уменьшение неопределённости одной переменной при знании другой. [2]
4. Операторные и функциональные меры. Определяют зависимость через

операторы условного ожидания и ковариационные структуры.[3,4,5]

5. Унифицирующие функционалы зависимости. Предлагают общий формализм, включающий многие известные меры как частные случаи.[6]

6. Критерии независимости.[7]

После вычисления вектора характеристик для каждого признака задача сводится к формированию подмножества; можно выделить следующие подходы к нему:

- Кардинальный отбор: выбираются k признаков с наибольшими значениями выбранной характеристики. Размер множества задаётся заранее или в виде доли от общего числа признаков.

- Пороговый отбор: включаются признаки, значение характеристики которых превышает заданный порог. Порог может быть фиксированным, статистическим или адаптивным. В частности, в ряде работ используется относительный порог в виде верхних $p\%$ признаков или квантиля распределения оценок [8], в других подходах порог калибруется относительно уровня «шума», определяемого через искусственные или перестановочные признаки, что позволяет эмпирически задать границу значимости [9]. Такой механизм делает правило отбора зависимым от структуры данных и снижает произвольность выбора параметра.

- Метод «локтя»: размер подмножества определяется по точке излома в графике убывания оценок. Отбор производится до момента резкого снижения вклада признаков.

Предложенная структуризация фильтрационных методов через раздельное описание метрик связи и правил формирования подмножества позволяет задать обобщённый алгоритм их реализации.

Выводы.

Также формализация фильтрационного подхода в виде обобщённого алгоритма открывает возможность систематического исследования его поведения в нестандартных регрессионных постановках. Это позволяет тестировать фильтрационные методы на моделях со специфическими свойствами зависимой переменной, включая бета-регрессию, регрессию с избыточными нулями и другие нетривиальные обобщения классической регрессии.

Список использованных источников:

1. Chatterjee S. A survey of some recent developments in measures of association //Probability and stochastic processes: a volume in Honour of Rajeeva L. Karandikar. – 2024. – С. 109-128.
2. de Siqueira Santos S. et al. A comparative study of statistical methods used to identify dependencies between gene expression signals //Briefings in bioinformatics. – 2014. – Т. 15. – №. 6. – С. 906-918.
3. Székely G. J., Rizzo M. L. Energy statistics: A class of statistics based on distances //Journal of statistical planning and inference. – 2013. – Т. 143. – №. 8. – С. 1249-1272.
4. Ren Y. et al. Learning adaptive kernels for statistical independence tests //International Conference on Artificial Intelligence and Statistics. – PMLR, 2024. – С. 2494-2502.
5. Tjøstheim D., Otneim H., Støve B. Statistical dependence: Beyond Pearson's ρ //Statistical science. – 2022. – Т. 37. – №. 1. – С. 90-109.
6. Azadkia M., Roudaki P. A new measure of dependence: Integrated R^2 //arXiv preprint arXiv:2505.18146. – 2025.
7. Gao Z. et al. Studentized tests of independence: Random-lifter approach //The Annals of Statistics. – 2025. – Т. 53. – №. 2. – С. 703-723.
8. Zouhri H., Idri A., Ratnani A. Evaluating the impact of filter-based feature selection in intrusion detection systems //International Journal of Information Security. – 2024. – Т. 23. – №. 2. – С. 759-785.
9. Wang Q. G., Li X., Qin Q. Feature selection for time series modeling //Journal of Intelligent Learning Systems and Applications. – 2013. – Т. 5. – №. 03. – С. 152-164.