Эффективность фильтрационных алгоритмов отбора признаков

А. Д. Черемухин

Нижегородский государственный инженерно-экономический университет

ngieu.cheremuhin@yandex.ru

Аннотация. **В работе исследуется эффективность фильтрационных методов отбора признаков в задачах регрессии с учётом свойств данных. Предложен подход, в рамках которого эффективность отбора рассматривается как функция характеристик выборки, используемой меры статистической связи, алгоритма модели и метрик качества. На основе синтетического бенчмарка, включающего различные семейства распределений и параметры структуры данных, проведён вычислительный эксперимент для комбинаций метрик зависимости и правил формирования подмножеств признаков. Показано, что эффективность методов носит контекстно-зависимый характер и определяется как параметрами распределения, так и структурными характеристиками задачи. Установлено отсутствие универсального метода отбора признаков, оптимального для всех типов данных.**

Ключевые слова: отбор признаков; фильтрационные методы; регрессия; меры статистической связи; эффективность моделей; синтетические данные; бенчмарк; расстояние Танимото; анализ данных; машинное обучение

I. Введение

Задачи регрессии занимают центральное место в современном анализе данных и находят широкое применение в экономике, медицине, инженерных и социальных исследованиях. В рамках таких задач требуется установить зависимость между набором независимых переменных и числовой зависимой переменной. Их целесообразно рассматривать как систему, включающую структурные компоненты (функция связи, свойства зависимой и независимых переменных) и процессуальные компоненты (метрика качества, метод оптимизации и данные).

Одной из ключевых проблем при решении задач регрессии является высокая размерность пространства признаков, наличие избыточных и слабоинформативных переменных, а также коррелированность факторов [1]. Эти особенности могут приводить к ухудшению качества моделей, снижению их интерпретируемости и увеличению вычислительной сложности. В этой связи важную роль играет задача отбора признаков (feature selection, FS), направленная на формирование подмножества переменных, наиболее релевантных для построения модели.

II. Постановка задачи

A. Фильтрационные методы отбора признаков

Существует несколько основных классов методов отбора признаков: фильтрационные, обёрточные, встроенные, гибридные и методы, основанные на обучении представлений. Среди них фильтрационные методы занимают особое место благодаря своей универсальности, вычислительной эффективности и независимости от конкретного алгоритма построения модели, что предопределило их выбор в качестве предмета исследования

Как правило, они основываются на оценке статистической связи между каждой независимой переменной и зависимой переменной. В этом смысле задача отбора признаков сводится к ранжированию признаков по некоторой мере зависимости и последующему выбору подмножества признаков на основе заданного правила.

Исторически в качестве таких мер использовались классические коэффициенты корреляции, прежде всего коэффициент Пирсона, а также ранговые коэффициенты Спирмена и Кендалла. Однако данные меры обладают рядом ограничений: они ориентированы на выявление линейных или монотонных зависимостей и не способны обнаруживать более сложные, нелинейные или нефункциональные зависимости.

Современные исследования в области статистических мер зависимости направлены на преодоление этих ограничений – в последнее время появился широкий класс обобщённых мер зависимости, включающий дистанционную корреляцию, статистику Хёффдинга, статистику Бергсмы–Дассиоса, взаимную информацию, максимальный информационный коэффициент и ряд других подходов, позволяющий выявлять существенно более широкий класс зависимостей, включая нелинейные и немонотонные структуры.

При этом очевидно, что использование различных мер зависимости в фильтрационных методах отбора признаков может приводить к выбору различных подмножеств признаков, что в конечном итоге влияет на результат решения задачи регрессии.

B. Эффективность отбора признаков

Несмотря на широкое распространение методов отбора признаков, вопрос оценки их эффективности остаётся недостаточно формализованным [2]. В большинстве работ эффективность методов FS рассматривается изолированно, без учёта их влияния на итоговое качество модели; однако, по мнению авторов, такой подход представляется недостаточным, поскольку отбор признаков является лишь промежуточным этапом решения задачи регрессии, и эффективность задачи отбора признаков включает в себя все метрики эффективности решения исходной задачи регрессии

С этой точки зрения эффективность метода отбора признаков следует рассматривать в контексте всей системы построения модели. В частности, результат

применения FS определяется не только выбранной мерой зависимости, но и последующим алгоритмом регрессии, используемой метрикой качества и свойствами данных. Таким образом, эффективность FS может быть интерпретирована как функция нескольких компонентов:

- характеристик данных (включая распределения переменных, наличие шума и выбросов);
- используемой меры зависимости;
- алгоритма построения модели;
- метрики качества.

Следовательно, корректная оценка эффективности методов отбора признаков должна учитывать их влияние на итоговое качество решения задачи регрессии, а не только свойства самих отобранных признаков.

III. Материалы и методы эксперимента

A. Гипотеза исследования

В настоящей работе выдвигается гипотеза о том, что эффективность фильтрационных методов отбора признаков существенно зависит от свойств выборки, прежде всего от распределений независимых и зависимой переменных, а также от структуры зависимости между ними.

Данная гипотеза основана на следующих соображениях. Во-первых, различные меры статистической связи по-разному чувствительны к типам зависимостей (линейным, нелинейным, монотонным, локальным), а также к особенностям данных, таким как выбросы, тяжёлые хвосты распределений и гетероскедастичность. Во-вторых, свойства данных являются ключевым элементом в системном описании задачи регрессии и определяют поведение как алгоритмов, так и метрик качества .

Таким образом, можно ожидать, что один и тот же метод отбора признаков будет демонстрировать различную эффективность на выборках с различными распределениями и структурой зависимостей. Это приводит к выводу о невозможности существования универсального метода отбора признаков, оптимального для всех типов данных.

B. Методика генерации данных для проведения эксперимента

Проверка данной гипотезы требует проведения систематического экспериментального исследования, в рамках которого оценивается влияние свойств данных на эффективность различных фильтрационных методов отбора признаков в задачах регрессии.

Проверка выдвинутой гипотезы проводилась на специально сформированном синтетическом бенчмарке для задач регрессии, в котором целенаправленно варьировались свойства распределений и структура данных. Общая логика построения такого бенчмарка ранее была описана в работе [3]; в соответствии с ней в настоящем исследовании использовалась синтетическая часть бенчмарка, включающая семь семейств распределений: бета-распределение, симметричное бета-распределение, гамма-распределение, логнормальное распределение, обратное гамма-распределение, бета-распределение второго рода и t-распределение Стьюдента; для каждого семейства было сформировано по десять вариантов параметров, что обеспечило в совокупности 70 различных датафреймов – однако на этом этапе их размер ограничен (не более 100 наблюдений и 100 переменных).

Генерация параметров распределений и структурных характеристик датафреймов осуществлялась в два этапа. На первом этапе с помощью генетического алгоритма подбирались параметры распределений, максимизирующие различие между формами плотностей. В программной реализации для этого использовалась функция Йенсена–Шеннона, вычисляемая попарно по матрице плотностей, а сам алгоритм включал инициализацию популяции, турнирный отбор, кроссовер и мутацию параметров (весь код и результаты его выполнения доступны по ссылке [4]). На втором этапе независимо формировались метапараметры самих датафреймов: число строк, количество значимых и незначимых признаков, а также целевой уровень связи между значимыми факторами и откликом. Для этого использовалось латинское гиперкубическое планирование, после чего применялся второй генетический алгоритм, максимизирующий суммарное попарное расстояние между параметрическими векторами, то есть обеспечивающий максимальное разнообразие получаемых сценариев.

C. Методика оценки эффективности отбора признаков

Далее на каждом из сгенерированных датафреймов проводился единообразный вычислительный эксперимент. В качестве зависимой переменной использовалась числовая переменная target, а в качестве кандидатов на включение в итоговое подмножество — все числовые предикторы, кроме целевой переменной. Для каждой пары «признак — отклик» рассчитывались одиннадцать мер статистической зависимости, рассмотренных ранее. На этой основе последовательно применялись пять правил формирования подмножества признаков:

- Первое правило соответствовало кардинальному отбору top-k признаков; в используемой реализации выбирались три наиболее сильных признака.
- Второе правило представляло собой фиксированный пороговый отбор по значению меры связи; в коде использовался порог 0,2.
- Третье правило было относительным и соответствовало выбору верхних 10% признаков по рангу.
- Четвёртое правило представляло собой адаптивный порог на основе «теневых» признаков: зависимая переменная случайно переставлялась, для каждой генерации вычислялось максимальное значение выбранной метрики по всем признакам, после чего в качестве порога брался 95%-квантиль распределения этих максимумов. В настоящем эксперименте число таких генераций было зафиксировано равным пяти.
- Пятое правило соответствовало автоматическому определению точки «локтя» по упорядоченному вектору значений метрики.

Таким образом, для каждого датафрейма проводился полный перебор всех сочетаний из 11 метрик зависимости и 5 правил формирования подмножества, то есть рассматривалось 55 экспериментальных конфигураций.

Поскольку использовались синтетические данные с заранее известной структурой, качество отбора признаков можно было оценивать не косвенно через последующую регрессионную модель, а напрямую через сравнение найденного и истинного подмножеств. Для каждой экспериментальной конфигурации вычислялись:

- расстояние Танимото между найденным и истинным множествами значимых признаков;
- процент правильно найденных значимых признаков;
- процент правильно исключённых незначимых признаков;
- процент пропущенных значимых признаков;
- процент ошибочно включённых незначимых признаков.

Дополнительно фиксировались вычислительные характеристики процедуры: время выполнения, изменение объёма занятой памяти и размер объекта итоговой модели; также сохранялось число отобранных признаков.

D. Анализ показателей оценки эффективности методов отборов признаков

На завершающем этапе для отдельных семейств распределений результаты агрегировались по десяти вариантам параметров. Из результатов извлекался показатель расстояния Танимото, после чего он сопоставлялся с параметрами соответствующих распределений и структурными характеристиками датафреймов, после чего рассчитывались коэффициенты корреляции между значением итогового показателя качества отбора и параметрами распределения или структуры данных. Это позволило перейти от простого попарного сравнения методов к анализу того, какие именно свойства выборки связаны с улучшением или ухудшением качества фильтрационного отбора признаков.

В итоге реализованная методика позволила исследовать не только различия между конкретными метриками связи и правилами формирования подмножеств, но и зависимость их эффективности от характеристик распределения и структуры данных.

IV. Результаты

Для семейства распределений Пирсона I типа (бета-распределение) были получены следующие результаты (табл. 1):

ТАБЛИЦА I. ФАКТОРЫ, ВЛИЯЮЩИЕ НА ЭФФЕКТИВНОСТЬ РАЗНЫХ МЕТОДОВ ОТБОРА ПРИЗНАКОВ

Мера статистической зависимости	**Правило Отбора**	**Показатель, с которым выявлена значимая корреляция у индекса Танимото**	**Величина корреляции**
Корреляция Пирсона	top-3	Число значимых признаков	0.845
Показатель Hoeffding's D	top-3	Число значимых признаков	0.879
Коэффициент Бергсмы-Дассиоса	Адаптивный порог	Число значимых признаков	0.852
Показатель дистанционной корреляции	Пороговый отбор по значению меры связи	Число незначимых признаков	0.845
HHG-статистика	top-3	Число значимых признаков	0.945
HHG-статистика	Выбор 10% лучших признаков	Число значимых признаков	0.902
MIC-коэффициент	top-3	Число значимых признаков	0.857

По данным таблицы 1 можно констатировать, что для большинства выявленных случаев значимой корреляции эффективность соответствующих методов отбора признаков прямо пропорциональна числу значимых признаков.

Детально таблицы с результатами, в том числе и для других распределений, представлены по ссылке [4].

Резюмируя их, можем отметить следующее: для семейства распределений Пирсона VI типа (обобщенное бета-распределение) установлено, что:

- для комбинации метрики Хёффдинга и правила top-k наблюдается отрицательная корреляция с параметром формы распределения;
- для комбинации метрики Кендалла и квантильного правила отбора наблюдается отрицательная корреляция с параметром формы распределения;
- для комбинации метрики Хёффдинга и адаптивного порога (shadow threshold) наблюдается отрицательная корреляция с параметром формы распределения;
- для комбинации метрики Бергсмы–Дассиоса и адаптивного порога наблюдается отрицательная корреляция с параметром формы распределения;
- для комбинации метрики дистанционной корреляции и метода локтя наблюдается отрицательная корреляция с параметром формы распределения.

Для семейства распределение II типа (симметричные бета-распределения) выявлены следующие зависимости:

- для комбинации метрики Кендалла и квантильного правила отбора наблюдается отрицательная корреляция с параметром формы распределения;
- для комбинации метрики взаимной информации и расширенного правила отбора наблюдается отрицательная корреляция с числом значимых признаков;

- для комбинации метрики максимальной корреляции Реньи и соответствующего правила отбора наблюдается отрицательная корреляция с числом значимых признаков;
- для комбинации HSIC-статистики и адаптивного правила отбора наблюдается отрицательная корреляция с числом значимых признаков.

При этом для остальных рассмотренных семейств распределений (гамма-распределение, логнормальное распределение, обратное гамма-распределение и распределение Стьюдента) статистически значимых связей между качеством отбора признаков и параметрами генерации данных выявлено не было.

V. Заключение

На основании результатов проведенных исследований можно утверждать, что эффективность фильтрационных методов отбора признаков является контекстно-зависимой и определяется свойствами данных; выдвинутая гипотеза подтверждена, влияние параметров данных носит селективный характер: в одних случаях доминируют структурные характеристики задачи, в других — параметры распределения, т.е. отсутствует универсальная зависимость, справедливая для всех типов распределений.

Отдельно стоит внимание на то, что выявленные зависимости справедливы только для случаев, когда зависимая переменная цензурирована и справа, и слева – во всех остальных случаях эффективность методов отбора признаков не зависит от параметров баз данных.

Полученные результаты могут стать основой для формирования рекомендаций по применению методов отбора признаков в реальных задачах – например, при построении регрессионных моделей для решения экономических [5] или технических [6] задач.

Список литературы

[1] Bommert A. et al. Benchmark for filter methods for feature selection in high-dimensional classification data //Computational Statistics & Data Analysis. 2020. Т. 143. С. 106839.

[2] Flach P. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward //Proceedings of the AAAI conference on artificial intelligence. 2019. Т. 33. №. 01. С. 9808-9814.

[3] Черемухин А.Д., Лямин А.С. Методика построения бенчмарков для оценки эффективности методов отбора признаков при решении задач регрессии // Информационные технологии. 2026. Т. 32, № 1. С. 20-27. – DOI 10.17587/it.32.20-27. – EDN YRFUOY..

[4] FS efficiency experiment [Электронный ресурс]. Режим доступа: https://github.com/acheremuhin/FS_efficiency_experiment

[5] Завиваев Н.С., Бобер В.С. Взаимосвязь социально-экономического развития сельских аграрных территорий и сельского хозяйства // Вестник НГИЭИ. 2025. № 5(168). С. 79-87. – DOI 10.24412/2227-9407-2025-5-79-87. – EDN QBLGRQ

[6] Оценка влияния синего спектра излучения на рост и поведенческие реакции бройлеров ROSS 308 / Ю.А. Журавлева, М.В. Коренюга, О.Ю. Коваленко, А.Н. Туркин // Вестник НГИЭИ. 2026. № 2(177). С. 34-45. – DOI 10.24412/2227-9407-2026-2-34-45. – EDN SAAEUG