УДК 519.237.5

Е.Б. Грибанова, Р.С. Герасимов

Гибридный алгоритм построения разреженной регрессии

Предложен гибридный алгоритм для построения разреженной регрессии. Выполнено тестирование алгоритма с использованием реальных и синтетических данных. Результаты проведенных экспериментов свидетельствуют о возможности применения алгоритма к рассматриваемым задачами и демонстрируют его эффективность в сравнении с известными методами.

Ключевые слова: разреженная регрессия, Lasso, отбор признаков, обратная задача.
DOI: 10.21293/1818-0442-2025-28-1-86-92

В условиях стремительного роста объемов данных и сложности современных аналитических задач регрессия как один из ключевых методов машинного обучения остается востребованным инструментом для выявления количественных зависимостей и прогнозирования [1–2]. Разреженная регрессия [3] – это метод регрессионного анализа, который используется для моделирования зависимостей в ситуациях, когда имеется большое количество признаков (независимых переменных), но лишь немногие из них являются значимыми для предсказания целевой переменной.

Основная идея разреженной регрессии заключается в том, чтобы отобрать наиболее информативные переменные и игнорировать остальные, уменьшая тем самым размерность модели и повышая её интерпретируемость. При наличии большого количества признаков, многие из которых могут быть не связаны с целевой переменной, обычная линейная регрессия может страдать от переобучения, неустойчивости, а также низкой точности из-за необходимости подстраиваться под «шум».

Ещё одна проблема, которая часто возникает при большой размерности, – мультиколлинеарность [4], которая может приводить к искажениям оценок коэффициентов регрессионной модели и увеличивать их стандартные ошибки. Таким образом, процесс отбора признаков играет ключевую роль в формировании качественной модели, что, в свою очередь, напрямую сказывается на точности и интерпретируемости результатов анализа и принимаемых на основе этих результатов решениях и рекомендациях в различных областях.

Разреженная регрессия является ценным инструментом в тех сферах, где необходимо рассматривать большое число потенциальных переменных, подходящих для включения в модель. В области экономики существует множество макро- и микроэкономических показателей [5, 6], таких как выручка, капитал и активы, которые могут быть проанализированы с целью их потенциального использования в модели. В сфере маркетинга возникает необходимость в анализе потребительских данных, что позволяет выявить значимые признаки, влияющие на покупательское поведение и оптимизировать маркетинговые стратегии [7]. В медицине возникает необходимость отбора диагностических признаков для повышения точности диагностики и прогноза заболеваний [8].

Выделяют две группы методов, осуществляющих отбор в контексте модели: методы-обёртки и встроенные методы [9]. Обёртки функционируют как отдельные процедуры, которые применяются к уже заданным моделям, позволяя провести многоэтапный процесс выбора переменных на основе различных критериев эффективности, таких как минимизация ошибки прогноза или максимизация объясненной дисперсии. Примером обёртки является пошаговая регрессия, которая последовательно добавляет или убирает независимые переменные из модели, основываясь на статистических критериях.

Встроенные методы, напротив, интегрированы непосредственно в процесс построения модели, автоматически идентифицируя наиболее значимые переменные в ходе обучения. Одним из наиболее распространенных встроенных методов является Lasso (Least absolute shrinkage and selection operator), который включает L1-регуляризацию, что приводит к обнулению некоторых коэффициентов и осуществлению отбора значимых признаков [10].

Использование обёрток является ресурсоемким способом, что может привести к проблемам при наличии большого числа признаков. С другой стороны, из-за того, что не рассматриваются все возможные комбинации, решение может оказаться неоптимальным. Применение встроенных методов также сопряжено с рядом недостатков, таких как чувствительность к выбору параметра регуляризации и сложность для интерпретации, поскольку механизмы отбора переменных не всегда очевидны. Кроме того, встроенные методы не гарантируют, что модель не будет включать высоко коррелированные между собой и неинформативные переменные. Эти недостатки обусловливают актуальность разработки и исследования методов отбора признаков.

В данной работе предложен гибридный алгоритм, который сочетает в себе Lasso и пошаговую регрессию для выбора признаков.

Пошаговая регрессия

Алгоритм пошаговой регрессии может быть реализован как в прямом, так и в обратном направлении. В работе рассмотрен прямой выбор, при котором на каждом шаге происходит добавление переменной

в модель. В этом случае алгоритм включает два основанных шага:

Шаг 1. Оценить модель с добавлением каждой из независимых переменных, которые ещё не включены в модель.

Шаг 2. Выбрать переменную, которая оптимизирует выбранный критерий, и добавить её в модель.

Шаги 1 и 2 повторяются до тех пор, пока не будет достигнуто состояние, когда ни одна из переменных не может быть добавлена без ухудшения модели по выбранному критерию.

Lasso

Уравнение линейной функции регрессии представляется следующим образом:

$$y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \eta,$$

где y – зависимая переменная; $\boldsymbol{\beta}$ – вектор коэффициентов регрессии размера p+1; x – объясняющая переменная; η – случайный остаток; p – число признаков.

Тогда задача определения параметров регрессии с помощью метода наименьших квадратов заключается в минимизации суммы квадратов остатков и представляется следующим образом:

$$f(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \to \min,$$

где x_{ij} – значение j-го признака в i-м наблюдении; y_i – i-е значение зависимой переменной; n – число наблюдений.

Метод Lasso представляет собой L1-регуляризацию линейной регрессии. Таким образом, в процессе настройки параметров модели осуществляется одновременная минимизация не только стандартной функции потерь, которая выражается как сумма квадратов ошибок, но и суммы абсолютных значений коэффициентов. В математическом виде задача оптимизации имеет вид

$$\sum_{i=1}^{n} \left(y_i - \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \to \min, \qquad (1)$$

где λ – параметр регуляризации.

Классическим способом настройки параметра λ является применение кросс-валидации [11]. Данный способ является трудоемким, что побуждает исследователей к поиску способов решения задачи без применения данного параметра. С целью исключения необходимости настройки параметра в данной работе задача (1) была преобразована в задачу условной оптимизации. В частности, было выполнено реформулирование (1) в виде одноточечной обратной задачи. В отличие от подхода условной оптимизации, описанного в работе [12], где минимизируется сумма квадратов ошибок с одновременным ограничением на сумму абсолютных значений параметров, в представленном варианте отсутствует необходимость устанавливать ограничение на значения параметров, что позволяет более гибко подходить к решению задач. Кроме того, полученная задача является более простой для численного решения.

Гибридный алгоритм

Предложенный алгоритм основан на реформулировании Lasso (1) в виде обратной задачи при минимизации суммы абсолютных значений коэффициентов [13–15]:

$$\begin{aligned} g(\beta) &= \sum_{j=0}^{p} |\beta_j| \to \min, \\ f(\beta) &= \sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2 = y^*, \end{aligned} \qquad (2)$$

где $\bar{y}_i$ – предсказанное i-е значение; y^* – целевое значение.

Задача (2) в представленных ранее алгоритмах решения обратной задачи [13, 14] решается итерационно до выполнения условия останова, которым в том числе служит факт достижения функцией f значения y^*. В данной работе считается, что целевое значение y^* является некоторым заранее неизвестным малым числом, которое не будет достигнуто, поэтому данное условие останова было исключено из алгоритма. Кроме того, в исходный алгоритм [13, 14] были внесены изменения для интеграции пошаговой регрессии.

Исходные данные, используемые в алгоритме, включают параметр α шага изменения аргумента, точность ε, а также значение индикатора **u**, который характеризует возможность дальнейшего изменения аргумента и изначально принимается равным 1. Вместо определения точности ε можно задать максимальное количество итераций $r_{\max}$, которое может быть выполнено в процессе вычислений. Начальные значения элементов вектора параметров $\boldsymbol{\beta}$ регрессионной модели устанавливаются равными 0.

Перед применением алгоритма необходимо выполнить стандартизацию значений независимых переменных.

Алгоритм представлен на рис. 1.

Первый этап работы алгоритма включает в себя расчет необходимых величин в четырех начальных блоках. После этого, если выбранный для изменения параметр β_k имеет значение, равное нулю, производится оценка его значимости для модели с использованием установленного критерия отбора (IC). Это осуществляется путём построения модели как без исследуемого k-го признака, так и с его включением, после чего рассчитываются значения критериев для обоих случаев.

Если включение нового признака приводит к улучшению значения критерия модели по сравнению с моделью, в которую этот признак не входит, то данный признак добавляется в модель, и значение соответствующего параметра корректируется. В противном случае, если улучшения не происходит, признак исключается из дальнейших расчетов.

Таким образом, в процессе обучения модели проходит этап пошаговой регрессии по сравнению обученных моделей, при этом исследуются не все возможные для включения переменные, а только та, которая выбрана для изменения на текущей итерации.

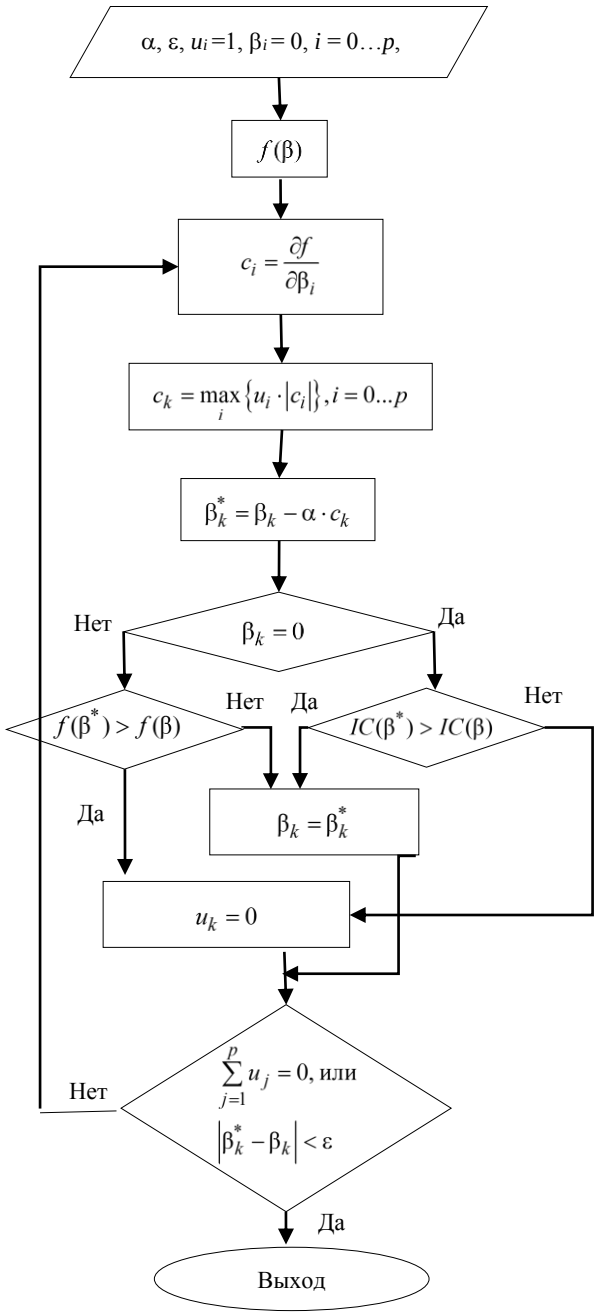

Рис. 1. Гибридный алгоритм построения разреженной регрессии

Результаты экспериментов

Построение регрессии выполнено с использованием разработанного алгоритма, а также пошаговой регрессии (Stepwise regression). Кроме того, в данной работе рассматриваются классическая линейная регрессия (LinearRegression), метод Lasso с применением кросс-валидации (LassoCV) и метод наименьших углов (Least-angle regression, Lars), который использует информационный критерий Акаике (AIC) для оценки модели (LassoLarsIC). LinearRegression, LassoCV и LassoLarsIC реализованы в библиотеке sklearn.linear_model, что обеспечивает удобство их применения для решения задач отбора признаков.

Реализация всех алгоритмов выполнена на языке Python, для методов LassoLarsIC, LinearRegression использованы параметры по умолчанию. Для сравнения были использованы показатели: число выбранных признаков (p^*), фактор инфляции дисперсии (vif), критерий Маллоуса (C_p), AIC, байесовский критерий (BIC), показатель устойчивости (γ). Критерий Маллоуса, AIC и BIC отражают баланс между точностью и сложностью модели, при этом меньшее значение показателей соответствует лучшей модели.

Критерий Маллоуса был вычислен по формуле

$$C_p = \frac{r_a}{r} - n + 2p,$$

где r_a – сумма квадратов остатков модели с p признаками; r – сумма квадратов остатков для модели со всеми признаками.

Коэффициент vif характеризует мультиколлинеарность и определен как максимальный коэффициент по всем признакам

$$\text{vif} = \max_j \frac{1}{1 - R_j^2},$$

где R_j^2 – коэффициент детерминации j-го признака относительно всех остальных признаков, используемых в модели.

Считается, что значения vif, превышающие 10, указывают на значительные проявления мультиколлинеарности, которые могут повлечь за собой негативные последствия и привести к некорректным оценкам коэффициентов [16].

Показатель устойчивости был вычислен на основе числа обусловленности матрицы $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ ($\mathbf{x}$ – матрица объясняющих переменных размера $n \times (p+1)$, первый столбец которой состоит из единиц). Более высокие значения этого показателя указывают на большую устойчивость модели к изменениям входных данных

$$\gamma = \ln \frac{\lambda_{\min}}{\lambda_{\max}},$$

где $\lambda_{\max}$ и $\lambda_{\min}$ – наибольшее и наименьшее собственное значение матрицы $\mathbf{x}^{\mathrm{T}}\mathbf{x}$.

В табл. 1 представлены результаты решения задачи с использованием набора данных Auto-mpg, включающего характеристики автомобилей [17]. Набор содержит 7 признаков и 398 наблюдений. В качестве критерия отбора в методах был использован AIC. В модуле statsmodels данный показатель вычисляется по следующей формуле:

$$\text{AIC} = -2L + 2p,$$

$$L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}{n}\right) - \frac{n}{2}.$$

Для полученного уравнения регрессии также представлены значения MSE, скорректированного индекса детерминации R^2, F-статистики (табл. 2).

Таблица 1

Решение задачи с использованием тестового набора Auto-mpg

Метод решения	p^*	vif	C_p	BIC	AIC	γ
LinearRegression	7	21,9	–383	2 128	2 096	–4,9
LassoCV, λ изменяется от 0 до 5 с шагом 0,01	7	21,9	–383	2 128	2 096	–4,9
LarsIC	7	21,9	–383	2 128	2 096	–4,9
Stepwise regression	3	1,6	–389	2 112	2 097	–1,5
Гибридный алгоритм, α=0,0001, r_{max}=10000	3	1,6	–389	2 112	2 097	–1,5

Таблица 2

Решение задачи с использованием тестового набора Auto-mpg: MSE, R^2, F-статистика

Метод решения	MSE	F	R^2
LinearRegression	10,9	256	0,817
LassoCV, λ изменяется от 0 до 5 с шагом 0,01	10,9	256	0,817
LarsIC	10,9	256	0,817
Stepwise regression	11,1	588	0,816
Гибридный алгоритм, α = 0,0001, r_{max} = 10000	11,1	588	0,816

Согласно полученным значениям Stepwise regression и гибридный алгоритм обеспечили одинаковый результат, выполнив отбор 3 признаков (масса, год модели, происхождение) и обеспечив лучшие значения показателей vif, C_p, BIC и γ по сравнению со встроенными методами. Все значения скорректированного индекса детерминации являются значимыми на уровне $p < 0{,}001$.

На рис. 2 проиллюстрирована зависимость критерия AIC от параметра α гибридного алгоритма, максимальное число итераций не изменялось и равно 10000. Оптимальные значения AIC наблюдаются при $\alpha = 10^{-5}...10^{-3}$. При больших значениях α метод не сходится к решению из-за того, что изменение параметра при большом шаге оказывается худшим решением по сравнению с его нулевым значением. Если α меньше 10^{-5}, то оптимальное решение не достигается за заданное число итераций из-за малого шага.

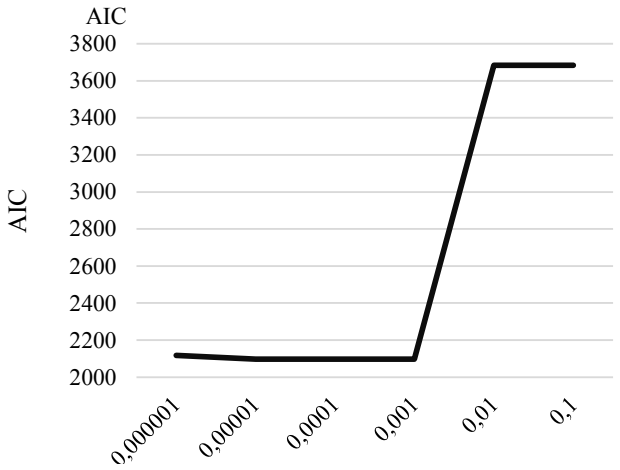

Рис. 2. Зависимость AIC от параметра α

В табл. 3 содержатся результаты, когда в качестве исходных данных использован случайно сгенерированный набор данных из 1 000 наблюдений и 299 признаков. Сгенерированные данные имеют равномерный закон распределения на интервале от 0 до 1. Гибридный алгоритм обеспечил более низкое значение критерия AIC, а также меньшее количество отобранных признаков по сравнению с методом Stepwise regression. С помощью LassoCV получить результат за установленное по умолчанию число итераций не удалось из-за отсутствия сходимости метода оптимизации.

Таблица 3

Решение задачи с использованием случайно сгенерированных значений

Метод решения	p^*	vif	C_p	BIC	AIC	γ
LinearRegression	299	1,658	–401	2121	554,26	–11
LassoCV, λ изменяется от 0 до 5 с шагом 0,01	–	–	–	–	–	–
LarsIC	8	1,02	–983	136	332,97	–4,8
Stepwise regression	71	1,107	–855	561	231,30	–8,5
Гибридный алгоритм, α=0,0001, r_{max}=100000	66	1,096	–865	527	**231,28**	–8,4

Кроме того, было выполнено исследование случая наличия мультиколлинеарности. Для этого был сгенерирован синтетический набор данных в соответствии с процедурой, представленной в статье [18]: 100 переменных, высоко коррелированных с целевым вектором, 100 переменных, ортогональных целевому вектору, и 100 случайных переменных. В этой ситуации наблюдается высокая мультиколлинеарность, к устранению которой не приводит использование рассматриваемых методов (табл. 4). В частности, использование методов Stepwise regression и LinearRegression привело к тому, что индекс детерминации между выбранными признаками оказался равным единице, что, в свою очередь, вызвало бесконечное значение vif. Кроме того, в этом случае матрица $\mathbf{x}^{\mathrm{T}}\mathbf{x}$ не является хорошо обусловленной, что приводит к невозможности определения показателя устойчивости, который стремится к минус бесконечности. При использовании разработанного гибридного алгоритма для решения данной задачи также возникла проблема мультиколлинеарности.

Таблица 4

Решение задачи при использовании синтетического набора данных с мультиколлинеарностью

Метод решения	p^*	vif	C_p	BIC	AIC	γ	t, с
LinearRegression	300	$+\infty$	–399	2073	–4 410	$-\infty$	35
LassoCV, λ изменяется от 0 до 5 с шагом 0,01	–	–	–	–	–	–	–
LarsIC	200	98,8	–599	1382	–4 610	–15	11
Stepwise regression	119	$+\infty$	–758,9	829	–4 713	$-\infty$	397
Гибридный алгоритм, α=0,0001, r_{max}=10000, u = **z**	99	**2,1**	–795,4	693	–3 285	–7,8	51

В связи с этим была предложена модификация алгоритма: вместо использования начальных единичных значений признака u, начиная со второго эле-

мента, был применен вектор вероятностей **z**, полученный в результате решения задачи квадратичного программирования, представленной в [19–20]:

$$\mathbf{z}^{\mathrm{T}}\mathbf{Q}\mathbf{z} - \mathbf{b}^{\mathrm{T}}\mathbf{z} \to \min,$$
$$\|z\|_1 \le 1,$$

где **z** – искомый вектор ненормализованной вероятности для выбора признака; **Q** – корреляционная матрица признаков; **b** – вектор корреляции признаков и результирующего показателя.

В процессе решения данной задачи проводится уменьшение количества взаимозависимых переменных и увеличение числа релевантных переменных. Таким образом, для переменных, которые демонстрируют высокую корреляцию с целевым вектором и низкую корреляцию с остальными переменными, соответствующее значение элемента вектора **z** будет максимальным. В результате был осуществлен отбор 99 переменных, коллинеарных целевому вектору, при этом значение коэффициента vif находится в допустимых пределах.

Также в табл. 4 представлено время (t), затраченное на решение задачи: встроенный метод LarsIC демонстрирует наивысшую скорость, в то время как метод Stepwise regression является наиболее ресурсоемким. Кроме того, был рассмотрен набор данных Communities and Crime, описание которого приводится по ссылке [21]. Данный набор включает 100 признаков и 1 994 наблюдения. В данном эксперименте было выполнено разделение выборки на обучающую и тестовую в пропорции 0,7 и 0,3 соответственно.

Согласно полученным результатам (табл. 5–7), гибридный алгоритм обеспечил наименьшее значение AIC на обучающей выборке, а гибридный алгоритм с использованием вектора вероятностей **z** – наименьшее значение vif. На тестовой выборке наилучшие значения информационных критериев показал гибридный алгоритм с использованием вектора вероятностей **z**.

Таблица 5

Решение задачи с использованием набора Communities and Crime, обучающая выборка

Метод решения	p^*	vif	C_p	BIC	AIC	γ
LinearRegression	100	964	–1 192	754	–1 604	–11
LassoCV, λ изменяется от 0 до 5 с шагом 0.01	100	964	–1 192	754	–1 604	–11
LarsIC	64	102	–1 266	486	–1 636	–8
Stepwise regression	25	28	–1 342	212	–1 656	–6
Гибридный алгоритм, $\alpha = 0{,}0001$, $r_{\max} = 10000$	40	44	–1 314	313	–1 671	–7
Гибридный алгоритм, α=0,0001, $r_{\max} = 10000$, **u** = **z**	14	13	–1 364	135	–1 568	–5

Необходимость отбора признаков может возникать не только в контексте линейных моделей; в рамках исследования было рассмотрено применение алгоритма для построения логистической регрессии.

Таблица 6

Решение задачи с использованием набора Communities and Crime, обучающая выборка: MSE, *F*-статистика, R^2

Метод решения	MSE	F	R^2
LinearRegression	0,016	31	0,68
LassoCV, λ изменяется от 0 до 5 с шагом 0,01	0,016	31	0,68
LarsIC	0,0165	48,75	0,69
Stepwise regression	0,0172	118,4	0,68
Гибридный алгоритм, $\alpha = 0{,}0001$, $r_{\max} = 10000$	0,0167	78,76	0,69
Гибридный алгоритм, $\alpha = 0{,}0001$, $r_{\max} = 10000$, **u** = **z**	0,018	190	0,65

Таблица 7

Решение задачи с использованием набора Communities and Crime, тестовая выборка

Метод решения	C_p	BIC	AIC	MSE	R^2
LinearRegression	–396	657	–470	0,0191	0,6
LassoCV, λ изменяется от 0 до 5 с шагом 0,01	–396	657	–470	0,0191	0,6
LarsIC	–470	421	–545	0,019	0,6
Stepwise regression	–546	200	–619	0,0196	0,62
Гибридный алгоритм, $\alpha = 0{,}0001$, $r_{\max} = 10000$	–518	301,2	–584	0,0193	0,61
Гибридный алгоритм, $\alpha = 0{,}0001$, $r_{\max} = 10000$, **u** = **z**	–568	120	–624	0,0196	0,63

В качестве встроенных методов решения был рассмотрен метод LogisticRegression с L1-регуляризацией, также входящий в библиотеку sklearn.linear_model и использующий решатели saga и liblinear. В наборе данных Auto-mpg зависимая переменная была преобразована в бинарный формат на основе медианного значения. В результате гибридный алгоритм выполнил отбор 4 признаков (табл. 8), обеспечив лучшие значения для всех критериев, включая среднеквадратичную ошибку (MSE). При этом значение критерия AIC меньше, чем при использовании пошаговой регрессии, в 6 раз.

Таблица 8

Решение задачи при логистической регрессии

Метод решения	p^*	vif	C_p	BIC	AIC	γ	MSE
LogisticRegression, saga	6	21,7	–383	76	164	–4,8	0,085
LogisticRegression, l iblinear	6	21,7	–385	71	174	–3,5	0,088
Stepwise regression	5	8,7	–385	73	196	–4	0,093
Гибридный алгоритм, $\alpha = 0{,}0001$, $r_{\max} = 100000$	4	5,3	–387	55	33	–3,2	0,062

Заключение

Выполнена разработка гибридного алгоритма для построения разреженной регрессии, сочетающего в себе элементы встроенных методов и пошаговой регрессии. Алгоритм основан на реформулировании Lasso в виде обратной задачи и сравнения регрессий с разным числом признаком в ходе обучения модели.

Алгоритм был протестирован с применением как реальных, так и синтетических данных, и полученные результаты свидетельствуют о возможности его применения для решения задач отбора признаков. Результаты исследований могут быть использованы для улучшения регрессионных моделей, что, в свою очередь, может способствовать более эффективному решению практических проблем в различных областях.

Дальнейшие исследования будут сосредоточены на применении методов решения некорректных задач в таких областях машинного обучения, как обучение нейронных сетей и кластеризация данных.

Исследование выполнено при финансовой поддержке Российского научного фонда (проект № 25-21-00123).

Литература

1. Economic modeling of mechanized and semi-mechanized rainfed wheat production systems using multiple linear regression model / A.K. Mobin, R. Reza, E.-T. Mahdi, K.-M. Armaghan // Information Processing in Agriculture. – 2020. – Vol. 7, No. 1. – P. 30–40.

2. Using a linear regression approach to sequential inter-industry model for time-lagged economic impact analysis / H. Kehan, M. Zhifu, C. D'Maris, G. Dabo // Structural Change and Economic Dynamics. – 2022. – Vol. 62. – P. 399–406.

3. On sparse regression, Lp-regularization, and automated model discovery / J.A. McCulloch, S.R. St. Pierre, K. Linka, E. Kuhl // International Journal for Numerical Methods in Engineering. – 2024. – Vol. 125, No. 14. – P. e7481.

4. Upendra S. Multicollinearity in Multiple Linear Regression: Detection, Consequences, and Remedies / S. Upendra, R. Abbaiah, P. Balasiddamuni // International Journal for Research in Applied Science and Engineering Technology. – 2023. – Vol. 11, No. 9. – IJRASET55786.

5. Ojo O.O. Bayesian analysis of macroeconomic variables on national savings / O.O. Ojo, A.A. Adepoju // Communications in Statistics: Case Studies, Data Analysis and Applications. – 2021. – Vol. 7, No. 3. – P. 432–441.

6. Vukovic D.B. Predicting the Performance of Retail Market Firms: Regression and Machine Learning Methods / D.B. Vukovic, L. Spitsina, E. Gribanova, V. Spitsin, I. Lyzin // Mathematics. – 2023. – Vol. 11, No. 8. – P. 1916.

7. Terui N. Measuring large-scale market responses from aggregated sales – Regression model for high-dimensional sparse data / N. Terui, Y. Li // Research Papers in Economics. – 2019. – Vol. 38, No. 5. – P. 440–458.

8. Islamiyati A. Studi longitudinal pada analisis data gula darah pasien diabetes melalui principal component analysis / A. Islamiyati, S. Sahriman, S. Oktoni // Jambura Journal of Mathematics. – 2022. – Vol. 4, No. 1. – P. 41–49.

9. Phan T.-T.-H. Enhancing rice seed purity recognition accuracy based on optimal feature selection / T.-T.-H. Phan, L.H.B. Nguyen // Ecological Informatics. – 2025. – Vol. 86. – P. 103044.

10. Regularization methods for high-dimensional data as a tool for seafood traceability / C. Yokochi, R. Bispo, F. Ricardo, R. Calado // Journal of Statistical Theory and Practice. – 2023. – Vol. 17, No. 3. – P. 44.

11. Roozbeh M. Generalized cross-validation for simultaneous optimization of tuning parameters in ridge regression / M. Roozbeh, M. Arashi, N.A. Hamzah // Iranian Journal of Science and Technology, Transactions A: Science. – 2020. – Vol. 44. – P. 473–485.

12. Tibshirani R. Regression shrinkage and selection via the lasso // Journal of the Royal Statistical Society, Series B. – 1996. – Vol. 58. – P. 267–288.

13. Gribanova E. Elaboration of an algorithm for solving hierarchical inverse problems in applied economics / E. Gribanova // Mathematics. – 2022. – Vol. 10, No. 15. – P. 2779.

14. Gribanova E. Algorithm for solving the inverse problems of economic analysis in the presence of limitations // EUREKA: Physics and Engineering. – 2020. – No. 1. – P. 70–78.

15. Грибанова Е.Б. Методы решения обратных задач экономического анализа с помощью минимизации приращений аргументов // Доклады ТУСУР. – 2018. – Т. 21, № 2. – С. 95–99.

16. Belsley D.A. Conditioning diagnostics: Collinearity and weak data in regression. – New York: John Wiley & Sons, 1991. – 396 p.

17. UC Irvine Machine Learning Repository. Auto MPG [Электронный ресурс]. – Режим доступа: https://archive.ics.uci.edu/dataset/9/auto+mpg, свободный (дата обращения: 24.02.2025).

18. Katrutsa A.M. Stress test procedure for feature selection algorithms / A.M. Katrutsa, V.V. Strijov // Chemometrics and Intelligent Laboratory Systems. – 2015. – Vol. 142. – P. 172–183.

19. Katrutsa A. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria / A. Katrutsa, V. Strijov // Expert Systems with Applications. – 2017. – Vol. 76. – P. 1–11.

20. Грибанова Е.Б. Алгоритм выбора признаков линейной регрессии для решения проблемы мультиколлинеарности // Искусственный интеллект и принятие решений. – 2025. – № 1. – С. 95–104.

21. UC Irvine Machine Learning Repository. Community and crime [Электронный ресурс]. – Режим доступа: https://archive.ics.uci.edu/dataset/183/communities+and+crime, свободный (дата обращения: 8.04.2025).

Грибанова Екатерина Борисовна
Д-р техн. наук, проф. каф. автоматизированных систем управления (АСУ) Томского государственного ун-та систем управления и радиоэлектроники (ТУСУР)
Ленина пр-т, 40, г. Томск, Россия, 634050
ORCID: 0000-0001-6499-5893
Тел.: +7 (382-2) 70-15-36
Эл. почта: ekaterina.b.gribanova@tusur.ru

Герасимов Роман Сергеевич
Магистрант каф. АСУ ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: +7 (382-2) 70-15-36
Эл. почта: roman.s.gerasimov@tusur.ru

Поступила в редакцию: 25.02.2025.
Принята к публикации: 18.04.2025.

Gribanova E.B., Gerasimov R.S.
Hybrid sparse regression algorithm
A hybrid algorithm to construct sparse regression is proposed. The proposed algorithm was tested using real and synthetic data. The experimental results demonstrate the applicability of

the proposed algorithm to the tasks under consideration and show its efficiency compared to known methods.

Keywords: sparse regression, Lasso, feature selection, inverse problem.
DOI: 10.21293/1818-0442-2025-28-1-86-92

References

1. Mobin A.-K., Reza R., Mahdi E.-T., Armaghan K.-M. Economic modeling of mechanized and semi-mechanized rain-fed wheat production systems using multiple linear regression model. *Information Processing in Agriculture*, 2020, vol. 7, no. 1, pp. 30–40.

2. Kehan H., Zhifu M., D'Maris C., Dabo G. Using a linear regression approach to sequential interindustry model for time-lagged economic impact analysis. *Structural Change and Economic Dynamics*, 2022, vol. 62, pp. 399–406.

3. McCulloch J.A., St. Pierre S.R., Linka K., Kuhl E. On sparse regression, Lp-regularization, and automated model discovery. *International Journal for Numerical Methods in Engineering*, 2024, vol. 125, no. 14, p. e7481.

4. Upendra S., Abbaiah R., Balasiddamuni P. Multicollinearity in Multiple Linear Regression: Detection, Consequences, and Remedies. *International Journal for Research in Applied Science and Engineering Technology*, 2023, vol. 11, no. 9, IJRASET55786.

5. Ojo O.O., Adepoju A.A. Bayesian analysis of macroeconomic variables on national savings. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 2021, vol. 7, no. 3, pp. 432–441.

6. Vukovic D.B., Spitsina L., Gribanova E., Spitsin V., Lyzin I. Predicting the performance of retail market firms: regression and machine learning methods. *Mathematics*, 2023, vol. 11, no. 8, p. 1916.

7. Terui N., Li Y. Measuring large-scale market responses from aggregated sales: regression model for high-dimensional sparse data. *Research Papers in Economics*, 2019, vol. 38, no. 5, pp. 440–458.

8. Islamiyati A., Sahriman S., Oktoni S. Studi longitudinal pada analisis data gula darah pasien diabetes melalui principal component analysis. *Jambura Journal of Mathematics*, 2022, vol. 4, no. 1, pp. 41–49.

9. Phan T.-T.-H., Nguyen L.H.B. Enhancing rice seed purity recognition accuracy based on optimal feature selection. *Ecological Informatics*, 2025, vol. 86, p. 103044.

10. Yokochi C., Bispo R., Ricardo F., Calado R. Regularization methods for high-dimensional data as a tool for seafood traceability. *Journal of Statistical Theory and Practice*, 2023, vol. 17, no. 3, p. 44.

11. Roozbeh M., Arashi M., Hamzah N.A. Generalized cross-validation for simultaneous optimization of tuning parameters in ridge regression. *Iranian Journal of Science and Technology, Transactions A: Scienc*e, 2020, vol. 44, pp. 473–485.

12. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 1996, vol. 58, pp. 267–288.

13. Gribanova E. Elaboration of an algorithm for solving hierarchical inverse problems in applied economics. *Mathematics*, 2022, vol. 10, no. 15, p. 2779.

14. Gribanova E. Algorithm for solving the inverse problems of economic analysis in the presence of limitations. *EUREKA: Physics and Engineering*, 2020, no. 1, pp. 70–78.

15. Gribanova E.B. [Methods for solving inverse problems of economic analysis by minimizing argument increments]. *Proceedings of TUSUR University*, 2018, vol. 21, no. 2, pp. 95–99 (in Russ.).

16. Belsley D.A. Conditioning diagnostics: Collinearity and weak data in regression. New York, John Wiley & Sons, 1991, 396 p.

17. UC Irvine Machine Learning Repository. Auto MPG. Available at: https://archive.ics.uci.edu/dataset/9/auto+mpg (Accessed: February 24, 2025).

18. Katrutsa A.M., Strijov V.V. Stress test procedure for feature selection algorithms. *Chemometrics and Intelligent Laboratory Systems*, 2015, vol. 142, pp. 172–183.

19. Katrutsa A., Strijov V. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications*, 2017, vol. 76, pp. 1–11.

20. Gribanova E.B. [An algorithm for selecting linear regression features to solve the multicollinearity problem]. *Artificial Intelligence and Decision Making*, 2025, no. 1, pp. 95–104 (in Russ.).

21. UC Irvine Machine Learning Repository. Community and crime. Available at: https://archive.ics.uci.edu/dataset/183/communities+and+crime (Accessed: April 08, 2025).

Ekaterina B. Gribanova
Doctor of Science in Engineering, Professor,
Department of Automated Control System (ACS),
Tomsk State University of Control Systems
and Radioelectronics (TUSUR)
40, Lenin pr., Tomsk, Russia, 634050
ORCID: 0000-0001-6499-5893
Phone: +7 (382-2) 70-15-36
Email: ekaterina.b.gribanova@tusur.ru

Roman S. Gerasimov
Master student, Department of ACS TUSUR
40, Lenin pr., Tomsk, Russia, 634050
Phone: +7 (382-2) 70-15-36
Email: roman.s.gerasimov@tusur.ru

Received: 25.02.2025.
Accepted: 18.04.2025.