



DOI: <https://doi.org/10.15688/mpcm.jvolsu.2026.1.3>

УДК 519.7
ББК 32.813

Дата поступления статьи: 03.03.2026
Дата принятия статьи: 09.04.2026

МОДЕЛЬ ОПТИМИЗАЦИИ ПРИЗНАКОВОГО ПРОСТРАНСТВА В ЗАДАЧАХ РЕГРЕССИИ НА ОСНОВЕ МЕДИЦИНСКИХ ДАННЫХ¹

Кирилл Сергеевич Демин

Аспирант кафедры математического анализа и теории функций,
Волгоградский государственный университет
diominkirill@volsu.ru
<https://orcid.org/0009-0002-4571-3437>
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Илья Васильевич Гермашев

Доктор технических наук, профессор кафедры математического анализа
и теории функций,
Волгоградский государственный университет
germashev@volsu.ru
<https://orcid.org/0000-0001-5507-8508>
просп. Университетский, 100, 400062 г. Волгоград, Российская Федерация

Цитирование: Демин К. С., Гермашев И. В. Модель оптимизации признакового пространства в задачах регрессии на основе медицинских данных // Математическая физика и компьютерное моделирование. 2026. Т. 29, № 1. С. 26–37.
DOI: <https://doi.org/10.15688/mpcm.jvolsu.2026.1.3>

Citation: Diomin K.S., Germashev I.V. A Model for Optimizing Feature Space in Regression Problems Based on Medical Data. *Mathematical Physics and Computer Simulation*, 2026, vol. 29, no. 1, pp. 26-37.
DOI: <https://doi.org/10.15688/mpcm.jvolsu.2026.1.3>

Аннотация. В работе представлена математическая модель оптимизации признакового пространства в задачах с использованием моделей машинного обучения на основе медицинских данных. Была предложена гибридная модель, учитывающая три основных вида отбора признаков, сочетающая как оценку значимости признаков и их избыточности, так и качество работы моделей машинного обучения. Методом нахождения оптимального подмножества признаков было предложено использование генетического алгоритма. В вычислительном эксперименте, демонстрирующем эффективность предложенной модели, использовался открытый набор данных о заболеваемости раком в Соединенных Штатах Америки. Исходный набор данных содержал 26 признаков и целевую переменную, количество записей составляло 3 047. Были обучены и протестированы разные модели на полном наборе признаков, наилучшей из которых оказалась модель случайного леса со значением коэффициента детерминации 0,936 и средней абсолютной ошибки 18,92. В результате работы предложенной модели удалось оптимизировать исходный набор признаков до 5, при этом значение коэффициента детерминации составило 0,936, значение средней абсолютной ошибки 20,11. Полученная модель может использоваться как инструмент для отбора оптимального количества признаков, так и для выбора лучшей модели машинного обучения.

Ключевые слова: математические модели, машинное обучение, интеллектуальный анализ данных, оптимизация признакового пространства, регрессия.

Введение

Медицинские наборы данных в последние десятилетия приобретают значение для создания прогностических математических моделей, используемых в диагностике, оценке рисков и планировании терапии. На сегодняшний день существует множество математических моделей, которые помогают врачам в решении ряда задач [1; 3; 5].

Переход к цифровому здравоохранению привел к накоплению массивов данных, где каждая запись пациента может содержать сотни и тысячи клинических, лабораторных, геномных и инструментальных параметров [13; 15]. При этом, большая часть таких данных неструктурирована. Они могут быть в виде свободного текста, изображения без какой-либо дополнительной связанной информации или сигналов с медицинских устройств. Из-за этого возникают проблемы для построения математических моделей, которые могли бы автоматизировать часть задач в медицинской сфере. Выделяют три основные проблемы медицинских данных [2; 6; 10].

- Количество признаков сильно больше количества наблюдений. Это связано с особенностью всех медицинских данных, так как здоровых пациентов гораздо больше больных. Это вызывает проблемы в построении математических моделей, что ведет к низкой обобщающей способности и повышению риска переобучения.
- Взаимозависимые признаки. Так, например, при возникновении различных заболеваний, при анализе крови взаимно растут такие показатели, как скорость оседания эритроцитов и С-реактивный белок, что указывает на воспаление в организме. Хотя эти данные могут быть полезны для врача, принимающего решения, для математических моделей это накладывает проблемы определения «весов» признаков, а также уменьшает интерпретирующую способность моделей.

- Зашумленные признаки. Медицинские данные особенно подвержены шумам из-за ошибок измерений, опечаткам при внесении данных. Также сам пациент может вносить шум из-за особенностей организма на момент измерения каких-либо показателей, например, циркадные ритмы или фаза менструального цикла. Это приводит к тому, что математическая модель обучается на частично ложной информации, и на практике будет иметь более низкую точность.

Решение вышеуказанных проблем возможно за счет оптимизации входного набора признаков. Под оптимизацией в данном контексте понимается целенаправленный отбор переменных, которые обеспечивают сохранение клинически значимого сигнала при одновременном подавлении шума и избыточности. Такой подход позволяет значительно сократить проблемы переизбытка признаков к числу наблюдений, стабилизировать оценки за счет устранения корреляции признаков и повысить отношение сигнал/шум. Более того, оптимизированный набор позволяет повысить интерпретируемость моделей, что, в свою очередь, повышает доверие врачей к ним.

Несмотря на наличие большого количества математических методов для отбора признаков, в исследованиях редко применяется системный подход к формированию входных наборов данных. Часто выбор переменных остается либо исключительно алгоритмическим, либо субъективно экспертным, без учета особенностей используемых моделей машинного обучения и набора данных. В данной работе предлагается использовать комбинированный метод отбора признаков в задачах регрессии с использованием медицинских данных на основе интеллектуального анализа данных и моделей машинного обучения.

Цель данной статьи заключается в разработке математической модели оптимизации признакового пространства, которая позволит значительно сократить используемый набор признаков, повысить интерпретируемость модели и незначительно уменьшить точность по сравнению с моделью, использующей полный набор признаков.

1. Построение концептуальной модели оптимизации признакового пространства

Перед тем как перейти к построению математической модели оптимизации признакового пространства, необходимо проанализировать имеющиеся методы отбора признаков, выделить ключевые особенности и недостатки. В современной литературе методы разделяют на три основные категории [7; 8; 14]: фильтрующие, встроенные и оберточные. Каждый из них характеризуется специфическим механизмом взаимодействия с данными, вычислительной сложностью и степенью учета контекста конечной модели.

Фильтрующие методы отбора признаков основаны на использовании статистической или информационной меры связи между независимым признаком и целевой зависимой (от этого признака) переменной без учета используемой модели машинного обучения. Примерами таких методов являются корреляционный анализ, взаимная информация и оценки дисперсионного анализа. При этом вычислительная сложность линейна или квадратична относительно числа признаков, что обеспечивает масштабируемость при большом числе признаков. Недостатком таких методов является то, что в условиях зашумленных данных приоритет может отдаваться стабильным, но малозначимым признакам, исключая слабые нелинейные, но более значимые признаки.

Во встроенных методах процесс отбора признаков интегрирован непосредственно в процедуру обучения модели машинного обучения. Примером такого метода является

Лассо-регрессия, которая добавляет штраф за большое количество признаков и может обнулять веса при их низкой значимости. Преимуществом данного подхода является баланс между точностью и вычислительной эффективностью, так как отбор происходит в момент обучения и не требует дополнительных вычислений. Однако результаты сильно зависят от выбора параметров регуляризации или архитектуры модели, на которой происходит процесс обучения.

В оберточных методах процесс отбора основан на оценке подмножеств признаков, которая производится через производительность используемой модели на основе итеративных стратегий поиска, таких как последовательные методы отбора или популяционные алгоритмы (генетический алгоритм, муравьиный алгоритм). Как правило, такие методы обеспечивают максимальную эффективность на конкретном наборе данных и позволяют напрямую оптимизировать целевые метрики эффективности. Главным недостатком данного типа методов является вычислительная сложность из-за многократного обучения модели на разных наборах признаков.

Таким образом, ни один из трех типов методов не решает в полной мере совокупность проблем, характерных для медицинских данных. Фильтрующие подходы игнорируют контекст модели и взаимодействия переменных, оберточные не масштабируются, встроенные зависят от параметров и архитектуры рассматриваемой модели. Разрабатываемая модель должна объединять статистическую значимость признаков и их избыточность, а также предсказательную способность моделей машинного обучения на оптимизированном множестве признаков. На рисунке 1 показана концептуальная схема разрабатываемой модели.

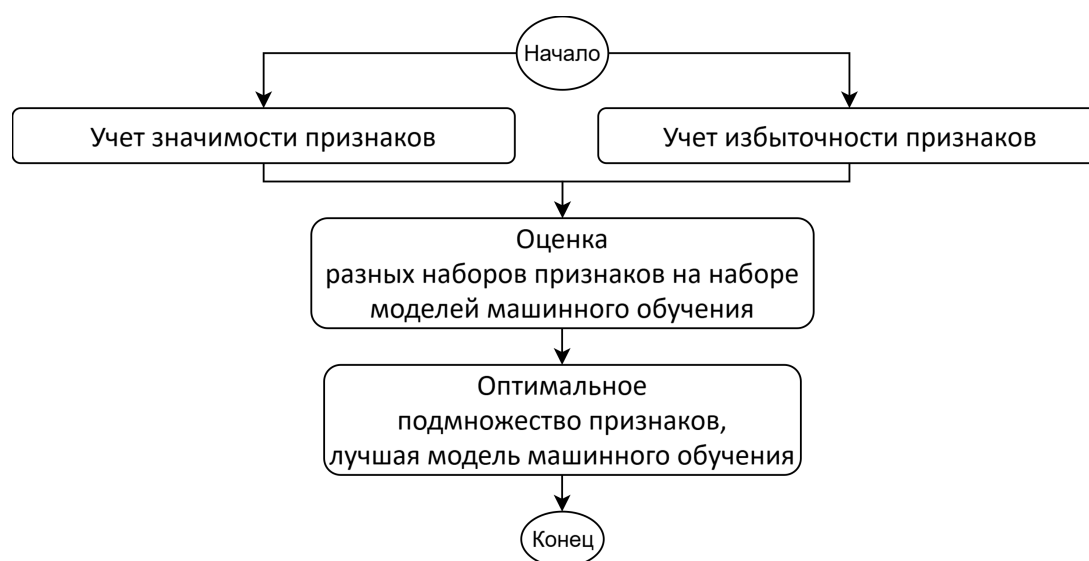


Рис. 1. Концептуальная схема модели оптимизации признакового пространства

Предлагаемая концептуальная гибридная модель структурирована в три связанных этапа, каждый из которых рассматривает аспект проблемы высокой размерности, коррелированности и зашумленности медицинских данных.

На этапе учета значимости признаков производится статистическая оценка вклада каждого признака в прогнозирование значения целевой переменной. Результатом этапа является множество всех признаков с соответствующим числовым значением значимости.

На этапе учета избыточности анализируется попарная зависимость рассматриваемых признаков с целью уменьшения корреляции и стабилизации последующих оценок. Результатом этапа является матрица попарной избыточности с соответствующим числовым значением корреляции.

На этапе оценки реализуется оберточный метод с учетом результатов, полученных на предыдущих этапах. Генерируется множество кандидатов подмножеств признаков, при этом каждый кандидат оценивается на наборе различных моделей машинного обучения.

В результате работы данной модели оптимизации на выходе получается оптимальное подмножество признаков и лучшая модель машинного обучения из заранее определенного набора. Данная концептуальная модель оптимизации служит основой для формальной математической постановки задачи оптимизации, которая изложена в следующем разделе.

2. Математическое построение модели оптимизации признакового пространства

Пусть имеется некоторый набор данных $X = (x_{ij})_{i=1, j=1}^{N,d}$, где $X \in \mathbb{R}^{N \times d}$, N представляет собой количество примеров, d — количество признаков. Каждому $x_i = (x_{ij})_{j=1}^d$ приписана целевая переменная $y \in \mathbb{R}$.

В рамках данной работы задачей является выделение такого набора признаков, который одновременно позволяет улучшить качество модели машинного обучения при сокращении вычислительной нагрузки и повышении интерпретируемости модели.

Пусть $D = \{1, 2, \dots, d\}$ — множество номеров признаков, множество отобранных номеров признаков обозначим через S , то есть $S \subseteq D$. Определим оператор проекции как $\pi_S(x) = (x_j)_{j \in S}$, где $x_j = (x_{ij})_{i=1}^N$.

Рассматривается набор моделей $\mathcal{H} = \{h_1, h_2, \dots, h_K\}$ машинного обучения, где K — количество рассматриваемых моделей. Обученная модель из набора $\mathcal{H}$ на множестве отобранных признаков S задается как

$$\hat{h}(S) = h^*(\pi_S(x)) = \arg \min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^N l(h(\pi_S(x)_i), y_i),$$

где l — функция потерь модели машинного обучения.

Определим качество конкретной модели как

$$Q(S, h) = \sum_{z=1}^Z \eta_z G_z(S, h),$$

где G_z — функция оценки качества модели машинного обучения h на наборе признаков S , Z — количество учитываемых оценок качества, $h \in \mathcal{H}$, $\eta_z \in \mathbb{R}^+$ — уровень значимости функции G_z .

В итоге задача оптимизации признакового пространства сводится к максимизации качества моделей

$$\max_{S \subseteq D} Q(S, h) \quad \text{при ограничении} \quad 1 \leq |S| \leq ST, \quad (1)$$

где ST — заданное количество признаков. Далее будем считать $ST = d$.

Для нахождения подмножества признаков S введем вектор отбора признаков $w = (w_1, w_2, \dots, w_d)$, где $w_j \in [0, 1]$.

На основе концептуальной модели (рис. 1), описанной в разделе 1, построим целевую функцию оценки качества оптимизированного набора признаков:

$$Q(S, h) = \frac{1}{K} \sum_{k=1}^K l(h_k(S)) - \eta_1 |S| + \eta_2 \sum_{j \in S} \sigma_j - \eta_3 P_{\text{corr}}(S), \quad (2)$$

где

- $\frac{1}{K} \sum_{k=1}^K l(h_k(S))$ — средняя оценка качества рассматриваемого подмножества признаков S на наборе моделей;
- $-|S|$ — штраф за количество используемых переменных;
- $\sum_{j \in S} \sigma_j$ — мера значимости признаков из рассматриваемого подмножества S ;
- $-P_{\text{corr}}(S)$ — мера избыточности признаков из рассматриваемого подмножества S .

Таким образом, первое слагаемое в (2) обеспечивает оценку качества признаков для заданного набора моделей, второе и четвертое — штраф за избыточность признаков, третье — поощрение за использование признаков с высокой информативностью. При этом предлагаемая модель может использоваться как в задачах классификации, так и регрессии. В данной работе будет проведен вычислительный эксперимент только для задачи регрессии.

Так как значимость признаков зависит от рассматриваемой задачи, опишем подробнее, что будет иметься в виду под членами функции (2).

Под функцией потерь будем понимать R^2 (коэффициент детерминации) для оценки качества набора моделей, которая задается формулой:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}},$$

где $SS_{\text{res}} = \sum_{i=1}^N (y_i - \hat{y}_i)^2$, $SS_{\text{tot}} = \sum_{i=1}^N (y_i - \bar{y})^2$, $\hat{y}_i$ — предсказанное значение y моделью h , $\bar{y}$ — среднее значение.

В качестве меры значимости признаков σ_j будем использовать комбинированную оценку, которая задается как

$$\sigma_j = \frac{1}{3}(q_{j1} + q_{j2} + q_{j3}),$$

где $q_{j1} = 1 - p_j$, а p_j — значение, получаемое в результате дисперсионного анализа (ANOVA [12]) для признака x_j относительно целевой переменной. Значение q_{j1} служит мерой статистической значимости связи признака и целевой переменной. Чем ближе q_{j1} к единице, тем более статистически значимой считается связь между ними, где q_{j2} — нормализованная оценка взаимной информации (Mutual Information [11]) между признаком и целевой переменной.

$$q_{j3} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

где q_{j3} — коэффициент корреляции Пирсона признака с целевой переменной.

Под мерой избыточности признаков $P_{\text{corr}}(S)$ понимается оценка количества парных корреляций признаков, которая задается в виде:

$$P_{\text{corr}}(S) = \frac{1}{|S|/2} \sum_{j_1, j_2} \max \left(0, \frac{R_{j_1, j_2} - \tau}{1 - \tau} \right),$$

где R_{j_1, j_2} — коэффициент корреляции Пирсона между двумя признаками x_{j_1} и x_{j_2} , $\tau = 0,8$ — значение коэффициента корреляции, больше которого корреляция полагается высокой.

3. Генетический алгоритм для поиска подмножества признаков

В силу комбинаторной природы задачи отбора признаков пространство допустимых решений задачи (1) содержит 2^d возможных подмножеств. При значительном количестве признаков полный перебор вычислительно неосуществим, поэтому в данной работе предлагается в качестве поискового механизма использовать генетический алгоритм (ГА), адаптированный под рассматриваемую задачу оптимизации. На рисунке 2 показана схема работы генетического алгоритма.

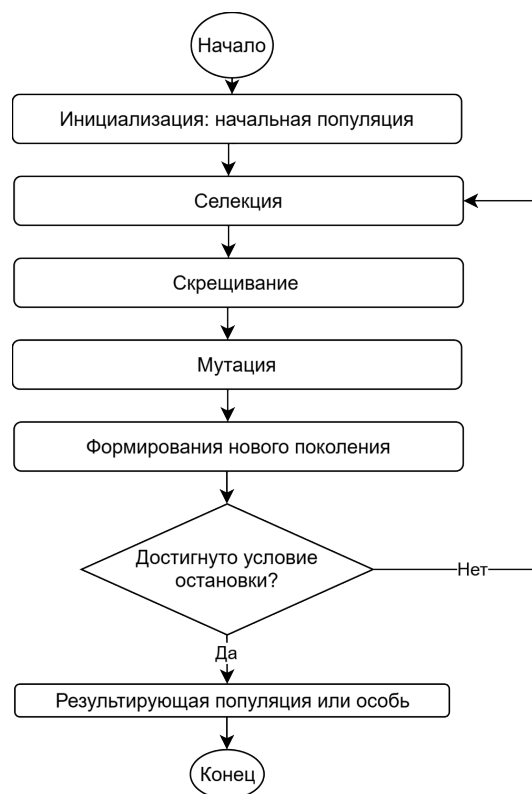


Рис. 2. Схема работы генетического алгоритма

Опишем подробнее работу алгоритма в контексте рассматриваемой задачи.

1) Шаг инициализации.

Формируется начальная популяция

$$\text{POP}_0 = \{S_1, S_2, \dots, S_{\text{size}}\},$$

задаваемая случайным образом. Здесь $size$ — размер популяции, а S_i выступает в роли особи (подмножества признаков). Каждый ген w_j особи принимает значения на отрезке $[0, 1]$. При $w_j > 0,3$ признак считается включенным и будет участвовать на этапе оценки моделей машинного обучения.

2) Шаг селекции.

Вычисляется значение целевой функции (2) для каждого набора S_i . Проводится турнирная селекция, в ходе которой отбираются особи с максимальным значением функции $\max F(w, h)$. При этом 5 % лучших особей исключаются из текущего поколения и напрямую переходят в следующее для сохранения наилучших решений.

3) Шаг скрещивания.

На основе предыдущего этапа формируются пары особей-родителей A и B с векторами генов

$$w^A = (w_1^A, \dots, w_r^A, w_{r+1}^A, \dots, w_d^A), \quad w^B = (w_1^B, \dots, w_r^B, w_{r+1}^B, \dots, w_d^B),$$

соответственно, для обмена генами по следующему принципу:

$$\begin{aligned} S_{\text{child1}}(w) &= (w_1^A, \dots, w_r^A, w_{r+1}^B, \dots, w_d^B), \\ S_{\text{child2}}(w) &= (w_1^B, \dots, w_r^B, w_{r+1}^A, \dots, w_d^A). \end{aligned}$$

4) Шаг мутации.

С некоторой заданной вероятностью происходит изменение генов особей путем добавления случайной величины с нулевым математическим ожиданием:

$$w'_j = w_j + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0; \sigma^2),$$

где ε — случайное приращение; σ^2 — дисперсия мутации.

Поскольку гены предполагают ограничение в заданном интервале, отрицательные значения заменяются на 0. Значения, превышающие единицу на 1.

5) Критерий остановки.

Шаги 2–4 повторяются до достижения заданного поколения T . В результате работы алгоритма выбирается лучшая особь — подмножество оптимизированных признаков — и соответствующая модель машинного обучения из первого слагаемого функции (2) для последующей проверки на тестовом наборе данных, не участвовавшем на этапе отбора.

4. Вычислительный эксперимент оценки работы модели оптимизации признаков пространства

Вычислительный эксперимент предлагаемой модели оптимизации признаков пространства проведен на открытом наборе данных, агрегирующем демографические, социально-экономические и эпидемиологические показатели по географическим регионам США [4]. Исходный набор данных содержит 26 признаков и целевую переменную — среднее ежегодное количество смертей, обусловленных онкологическими заболеваниями, в расчете на один регион. Данный показатель представляет собой целочисленную величину со значениями от 3 до 14 010. Всего исходный набор данных содержит 3 047 записей. Имеющиеся признаки можно разделить на 4 категории:

- 1) эпидемиологические, например, среднее количество заболеваний раком;
- 2) демографические, например, средний возраст пациентов в регионе;
- 3) социально-экономические, например, средний доход в регионе;
- 4) доступность здравоохранения, например, процент населения, имеющего медицинскую страховку.

Для оценки эффективности предложенной модели данные были разделены на обучающую (80 %) и тестовую (20 %) выборки. Все этапы оптимизации признаков выполнялись только на обучающей выборке для последующей проверки на тестовых данных.

В качестве набора моделей H машинного обучения будут использоваться случайный лес, линейная регрессия и гребневая регрессия (Ridge), структура которых описана в [16]. Данные модели машинного обучения были выбраны из-за имеющегося механизма интерпретации результатов, что важно для рассматриваемой области.

В качестве дополнительной метрики эффективности моделей машинного обучения была использована средняя абсолютная ошибка (MAE), которая вычисляется следующим образом:

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|.$$

Для сравнения, изначально были обучены модели машинного обучения на полном наборе признаков. В результате лучшей моделью оказался случайный лес с результатом $R^2 = 0,936$, $MAE = 18,92$.

Далее производился процесс оптимизации признакового пространства, где были заданы следующие параметры генетического алгоритма: количество особей – 50, количество эпох – 50.

Так как значения η_z зависят от используемого набора данных, были произведены дополнительные вычисления с помощью перекрестной проверки методом перебора по сетке параметров (GridSearchCV). Были получены оптимизированные значения $\eta_1 = 0,005$, $\eta_2 = 0,005$, $\eta_3 = 0,1$, которые давали наилучшее значение функции (2).

В результате работы алгоритма был получен оптимизированный набор, состоящий из 5 признаков, и лучшая модель – случайный лес.

Полученная модель была протестирована на ранее отделенном тестовом наборе. В результате было получено значение $R^2 = 0,936$, $MAE = 20,11$.

Также была произведена оптимизация признакового пространства при помощи известного метода – рекурсивного отбора признаков (RFE), описанного в работе [9]. Так как данный метод требует задания явного количества признаков и модели машинного обучения, для адекватной оценки была выбрана модель случайного леса с 5-ю оптимизированными признаками. Полученная оптимизированная модель машинного обучения также была протестирована на ранее отделенном тестовом наборе, показав результаты $R^2 = 0,934$, $MAE = 24,41$.

Таким образом, предложенная модель позволяет значительно сократить исходное признаковое пространство, при этом незначительно снизить качество модели, обученной на полном наборе данных. Также модель показывает лучшие значения на отобранных признаках в сравнении с известным методом отбора признаков.

Заключение

В работе представлен гибридный метод оптимизации признакового пространства для задач с использованием медицинских данных, включающий в себя основные методы оптимизации, такие как фильтрующие, оберточные и встроенные. Авторами предложена целевая функция, которая объединяет в себе использование набора моделей, состоящего из случайного леса, линейной регрессии и гребневой регрессии. В функцию также входят статистические методы оценки значимости признаков, а также корреляционный анализ для устранения избыточности между признаками. Используется открытый набор данных о заболеваемости раком в Соединенных Штатах Америки. Эффективность предложенного метода сравнивается с моделями машинного обучения, обученными на полном наборе признаков, а также известным методом отбора признаков RFE. Сокращение исходного пространства с 26 до 5 признаков демонстрирует эффективность предложенной гибридной модели в устранении избыточных и зашумленных переменных, при этом значение качества модели машинного обучения осталось на уровне, сопоставимом с моделью, обученной на полном наборе из 26 признаков.

ПРИМЕЧАНИЕ

¹ Работа выполнена при финансовой поддержке Минобрнауки РФ в рамках проекта FZUU-2026-0005 «Разработка методов моделирования сложных систем на основе интеграции прямых вычислительных экспериментов и интеллектуального анализа данных».

СПИСОК ЛИТЕРАТУРЫ

1. Попов, И. Е. Интерпретация математических моделей по данным микроволновой радиотермометрии / И. Е. Попов // Математическая физика и компьютерное моделирование. — 2025. — № 2. — С. 62–80. — DOI: <http://dx.doi.org/10.15688/mpcm.jvolsu.2025.2.6>
2. Berisha, V. Digital Medicine and the Curse of Dimensionality / V. Berisha // NPJ Digital Medicine. — 2021. — Vol. 4, № 153. — DOI: <https://doi.org/10.1038/s41746-021-00521-5>
3. Breast Cancer Classification Based on Hybrid CNN with LSTM Model / M. Kaddes, Y. M. Ayid, A. M. Elshewey, Y. Fouad // Scientific Reports. — 2025. — Vol. 15, № 4409. — DOI: <https://doi.org/10.1038/s41598-025-88459-6>
4. Cancer Data of United States of America. — URL: <https://www.kaggle.com/datasets/tanishal604/cancer-data-of-united-states-of-america>
5. Dyomin, K. Detection of Breast Cancer Risk Using Machine Learning Methods Based on Features Selected According to Diagnostic Models / K. Dyomin, I. Germashev, I. Niang // Human-Centric Cyber-Society. Studies in Systems, Decision and Control. — 2026. — Vol. 639. — DOI: https://doi.org/10.1007/978-3-032-06787-6_12
6. Ellsworth, S. Declarations of Independence: How Embedded Multicollinearity Errors Affect Dosimetric and Other Complex Analyses in Radiation Oncology / S. Ellsworth // International Journal of Radiation Oncology. — 2023. — DOI: <https://doi.org/10.1016/j.ijrobp.2023.06.015>
7. Feature Selection / J. Li, K. Cheng, S. Wang, F. Morstatter, R. Trevino, J. Tang, H. Liu // ACM Computing Surveys. — 2016. — P. 1–45. — DOI: <https://doi.org/10.1145/3136625>
8. Fira, M. Evaluating Sparse Feature Selection Methods: A Theoretical and Empirical Perspective / M. Fira, L. Goras, H. Costin // Applied Sciences. — 2025. — DOI: <https://doi.org/10.3390/app15073752>

9. Guyon, I. Gene Selection for Cancer Classification Using Support Vector Machines / I. Guyon, J. Weston, V. Vapnik // *Machine Learning*. — 2002. — P. 389–422. — DOI: <https://doi.org/10.1023/A:1012487302797>
10. Kascenas, A. The Role of Noise in Denoising Models for Anomaly Detection in Medical Articles/Images / A. Kascenas // *Medical Image Analysis*. — 2023. — Vol. 90 — DOI: <https://doi.org/10.48550/arxiv.2301.08330>
11. Kraskov, A. Hierarchical Clustering Using Mutual Information / A. Kraskov // *EPL*. — 2003. — Vol. 70. — P. 278–284. — DOI: <https://doi.org/10.1209/epl/i2004-10483-y>
12. Larson, M. Analysis of Variance / M. Larson // *Circulation*. — 2007. — DOI: <https://doi.org/10.1161/circulationaha.107.654335>
13. Lin, Y. Mini-Review of Clinical Data Service Platforms in the Era of Artificial Intelligence: A Case Study of the iHi Data Platform / Y. Lin // *BioMedicine*. — 2025. — Vol. 15. — P. 6–22. — DOI: <https://doi.org/10.37796/2211-8039.1643>
14. Masoudi-Sobhanzadeh, Y. FeatureSelect: A Software for Feature Selection Based on Machine Learning Approaches / Y. Masoudi-Sobhanzadeh, H. Motieghader, A. Masoudi-Nejad // *BMC Bioinformatics*. — 2019. — Vol. 20, № 170. — DOI: <https://doi.org/10.1186/s12859-019-2754-0>
15. Sarwar, T. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges / T. Sarwar // *ACM Computing Surveys*. — 2022. — Vol. 55. — P. 1–40. — DOI: <https://doi.org/10.1145/3490234>
16. Supervised Learning. — URL: https://scikit-learn.org/stable/supervised_learning.html

REFERENCES

1. Popov I.E. Interpretation of Mathematical Models Based on Microwave Radiothermometry Data. *Matematicheskaya fizika i kompyuternoe modelirovanie*, 2025, no. 2, pp. 62–80. DOI: <http://dx.doi.org/10.15688/mpcm.jvolsu.2025.2.6>
2. Berisha V. Digital Medicine and the Curse of Dimensionality. *NPI Digital Medicine*, 2021, vol. 4, no. 153. DOI: <https://doi.org/10.1038/s41746-021-00521-5>
3. Kaddes M., Ayid Y.M., Elshewey A.M., Fouad Y. Breast Cancer Classification Based on Hybrid CNN with LSTM Model. *Scientific Reports*, 2025, vol. 15, no. 4409. DOI: <https://doi.org/10.1038/s41598-025-88459-6>
4. *Cancer Data of United States of America*. URL: <https://www.kaggle.com/datasets/tanishal604/cancer-data-of-united-states-of-america>
5. Dyomin K., Germashev I., Niang I. Detection of Breast Cancer Risk Using Machine Learning Methods Based on Features Selected According to Diagnostic Models. *Human-Centric Cyber-Society. Studies in Systems, Decision and Control*, 2026, vol. 639. DOI: https://doi.org/10.1007/978-3-032-06787-6_12
6. Ellsworth S. Declarations of Independence: How Embedded Multicollinearity Errors Affect Dosimetric and Other Complex Analyses in Radiation Oncology. *International Journal of Radiation Oncology*, 2023 DOI: <https://doi.org/10.1016/j.ijrobp.2023.06.015>
7. Li J., Cheng K., Wang S., Morstatter F., Trevino R., Tang J., Liu H. Feature Selection. *ACM Computing Surveys*, 2016, pp. 1–45. DOI: <https://doi.org/10.1145/3136625>
8. Fira M., Goras L., Costin H. Evaluating Sparse Feature Selection Methods: A Theoretical and Empirical Perspective. *Applied Sciences*, 2025 DOI: <https://doi.org/10.3390/app15073752>
9. Guyon I., Weston J., Vapnik V. Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, 2002, pp. 389–422. DOI: <https://doi.org/10.1023/A:1012487302797>
10. Kascenas A. The Role of Noise in Denoising Models for Anomaly Detection in Medical Articles/Images. *Medical Image Analysis*, 2023, vol. 90 DOI: <https://doi.org/10.48550/arxiv.2301.08330>
11. Kraskov A. Hierarchical Clustering Using Mutual Information. *EPL*, 2003, vol. 70, pp. 278–284. DOI: <https://doi.org/10.1209/epl/i2004-10483-y>

12. Larson M. Analysis of Variance. *Circulation*, 2007 DOI: <https://doi.org/10.1161/circulationaha.107.654335>
13. Lin Y. Mini-Review of Clinical Data Service Platforms in the Era of Artificial Intelligence: A Case Study of the IHi Data Platform. *BioMedicine*, 2025, vol. 15, pp. 6-22. DOI: <https://doi.org/10.37796/2211-8039.1643>
14. Masoudi-Sobhanzadeh Y., Motieghader H., Masoudi-Nejad A. FeatureSelect: A Software for Feature Selection Based on Machine Learning Approaches. *BMC Bioinformatics*, 2019, vol. 20, no. 170. DOI: <https://doi.org/10.1186/s12859-019-2754-0>
15. Sarwar T. The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. *ACM Computing Surveys*, 2022, vol. 55, pp. 1-40. DOI: <https://doi.org/10.1145/3490234>
16. *Supervised Learning*. URL: https://scikit-learn.org/stable/supervised_learning.html

A MODEL FOR OPTIMIZING FEATURE SPACE IN REGRESSION PROBLEMS BASED ON MEDICAL DATA

Kirill S. Diomin

Postgraduate Student, Department of Mathematical Analysis and Function Theory,
Volgograd State University
diominkirill@volsu.ru
<https://orcid.org/0009-0002-4571-3437>
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Ilya V. Germashev

Doctor of Sciences (Engineering), Professor, Department of Mathematical Analysis
and Function Theory,
Volgograd State University
germashev@volsu.ru
<https://orcid.org/0000-0001-5507-8508>
Prosp. Universitetsky, 100, 400062 Volgograd, Russian Federation

Abstract. The paper presents a mathematical model for optimizing feature space in tasks using machine learning models based on medical data. A hybrid model was proposed that takes into account three main types of feature selection, combining both the assessment of the significance of features and their redundancy, and the quality of machine learning models. It was proposed to use a genetic algorithm to find the optimal subset of features, since a complete search in the tasks under consideration is impossible in an adequate amount of time. An open data set on cancer incidence in the United States of America was used in a computational experiment demonstrating the effectiveness of the proposed model. The initial data set contained 26 features and a target variable, the number of records was 3.047. Different models were trained and tested on the full set of features, the best of which turned out to be a random forest model with a coefficient of determination value of 0.936. As a result of the proposed model, it was possible to optimize the initial set of features to 7, with this coefficient of determination value being 0.917. The resulting model can be used as a tool both for selecting the optimal number of features and for choosing the best machine learning model.

Key words: mathematical models, machine learning, data mining, feature space optimization, regression.