НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
июль–август 2026 Том 26 № 4 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
July–August 2026 Vol. 26 No 4 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

КОМПЬЮТЕРНЫЕ СИСТЕМЫ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

COMPUTER SCIENCE

doi: 10.17586/2226-1494-2026-26-4-763-770
УДК 519.7

Комбинированный отбор признаков в машинном обучении для диагностики заболеваний в медицине

Кирилл Сергеевич Дёмин[1]✉, Илья Васильевич Гермашев[2]

[1,2] Волгоградский государственный университет, Волгоград, 400062, Российская Федерация

[1] dominkirill@yandex.ru✉, https://orcid.org/0009-0002-4571-3437
[2] germashev@volsu.ru, https://orcid.org/0000-0001-5507-8508

Аннотация

Введение. Исходные наборы данных для машинного обучения содержат избыточное количество признаков, среди которых могут присутствовать как высокоинформативные, так и малозначимые, коррелирующие или шумовые переменные. Решение проблемы отбора признаков не только оптимизирует процессы машинного обучения, но и открывает новые возможности для внедрения машинного обучения в практико-ориентированные сферы, требующие высокой точности и доверия к алгоритмам. Это явление приводит к ряду критических проблем: искаженному обучению модели на нерелевантных закономерностях, снижению ее обобщающей способности, резкому росту вычислительных затрат на обучение и, как следствие, к затруднению интерпретации полученных результатов. **Метод.** Предложен метод оптимизации признакового пространства, который использует комбинированный подход, включающий в себя максимизацию меры эффективности моделей, отбор наиболее качественных признаков на основе информативности и корреляционного анализа. Помимо оптимизации признакового пространства, модель предлагает наилучший классификатор для используемого набора данных. В качестве решения целевой функции применен генетический алгоритм с элитизмом для нахождения оптимального значения. **Основные результаты.** Показана работа представленной модели на наборе данных в области медицинской диагностики в сравнении с известным методом рекурсивного отбора признаков (Recursive Feature Elimination, RFE), корреляционным анализом, а также с результатами, полученными при использовании полного набора данных. Набор данных представляет собой измерения молочных желез при помощи метода микроволновой радиотермометрии и метки, характеризующие выраженность температурных аномалий. Используемый набор данных содержит 62 признака и 6 меток, и 9310 записей измерений. Результаты демонстрируют высокую эффективность предложенной модели, близкую по значению с методом RFE или превосходя его. В результате оптимизации для набора данных удалось сократить признаковое пространство с 62 признаков до 15, при этом не только не уменьшив точность модели, но даже немного ее увеличив, при этом лучшей моделью классификатора оказалась модель логистической регрессии. Например, показатели точности составили 0,79 для полного набора данных, 0,7879 для 15 оптимизированных признаков на основе метода RFE, 0,7175 для 29 признаков, полученных в результате корреляционного анализа, и 0,7911 для 15 признаков — с использованием предлагаемой модели. **Обсуждение.** Предложенный метод не только эффективно сокращает признаковое пространство, но и увеличивает точность модели классификации, несмотря на существенное уменьшение количества признаков.

Ключевые слова

машинное обучение, интеллектуальный анализ данных, оптимизация признакового пространства, классификация, рак молочной железы

Благодарности

Работа выполнена при финансовой поддержке Минобрнауки Российской Федерации в рамках проекта № FZUU-2026-0005 «Разработка методов моделирования сложных систем на основе интеграции прямых вычислительных экспериментов и интеллектуального анализа данных».

Ссылка для цитирования: Дёмин К.С., Гермашев И.В. Комбинированный отбор признаков в машинном обучении для диагностики заболеваний в медицине // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 4. С. 763–770. doi: 10.17586/2226-1494-2026-26-4-763-770

Combined feature selection in machine learning for disease diagnosis in medicine

Kirill S. Dyomin[1]✉, Ilya V. Germashev[2]

[1,2] Volgograd State University, Volgograd, 400062, Russian Federation

[1] diominkirill@yandex.ru✉, https://orcid.org/0009-0002-4571-3437

[2] germashev@volsu.ru, https://orcid.org/0000-0001-5507-8508

Abstract

Often, the initial data sets for machine learning contain an excessive number of features, among which there may be highly informative and insignificant, correlating or even noise variables. Solving the feature selection problem not only optimizes machine learning processes, but also opens up new opportunities for implementing machine learning in practice-oriented areas that require high accuracy and trust in algorithms. This phenomenon leads to a number of critical problems: distorted training of the model based on irrelevant patterns, a decrease in its generalizing ability, a sharp increase in computational costs for training and, as a result, difficulty in interpreting the results obtained. The paper proposes a method for optimizing the feature space which uses a combined approach that includes maximizing the measure of model effectiveness, selecting the highest-quality features based on information content and correlation analysis. In addition to optimizing the feature space, the model offers the best classifier for the data set used. As a solution to the objective function, a genetic algorithm with elitism is used to find the optimal value. The proposed model is demonstrated on a dataset in the field of medical diagnostics in comparison with the well-known method of recursive feature selection (RFE), correlation analysis, as well as with the results obtained using the full dataset. The data set consists of measurements of mammary glands using the method of microwave radiothermometry and labels characterizing the severity of temperature anomalies. The dataset contains 62 features and 6 labels and 9,310 measurement records. The results demonstrate the high efficiency of the proposed model either close to or exceeding the value of RFE. As a result of optimizing the feature space for the data set, it was possible to reduce from 62 features to 15, while not only not reducing the accuracy of the model, but even slightly increasing it, the best model of the classifier turned out to be the logistic regression model. Thus, the accuracy indicators were 0.79 for the complete data set, 0.7879 for 15 optimized features based on the RFE method, 0.7175 for 29 features obtained as a result of correlation analysis, and 0.7911 for 15 features obtained using the proposed model. The proposed method not only effectively reduces the feature space, but also increases the accuracy of the classification model, despite a significant decrease in the number of features.

Keywords

machine learning, data mining, feature space optimization, classification, breast cancer

Acknowledgements

The work was supported by the Ministry of Science and Higher Education of the Russian Federation (project No. FZUU-2026-0005).

For citation: Dyomin K.S., Germashev I.V. Combined feature selection in machine learning for disease diagnosis in medicine. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 4, pp. 763–770 (in Russian). doi: 10.17586/2226-1494-2026-26-4-763-770

Введение

В последние годы машинное обучение стало одним из ключевых инструментов анализа данных, широко применяемым в науке, промышленности и социальной сфере [1, 2]. Рост объемов и размерности данных, получаемых из различных источников, приводит к необходимости обработки высокоразмерных признаковых пространств. Однако использование большого числа признаков не всегда повышает качество моделей машинного обучения [3]. В основном исходные наборы данных содержат избыточное количество признаков, среди которых могут присутствовать как высокоинформативные, так и малозначимые, коррелирующие или шумовые переменные [4]. Грамотный отбор признаков является важной задачей по нескольким причинам. Во-первых, оптимизация набора признаков напрямую повышает качество предсказаний: модели, обученные на релевантных данных, демонстрируют большую точность и обобщающую способность. Во-вторых, сокращение размерности данных снижает вычислительную сложность моделей машинного обучения, что особенно важно для систем и устройств с ограниченными ресурсами. В-третьих, интерпретируемость моделей, особенно в таких регулируемых областях, как, например, медицина, требует прозрачности принимаемых решений, что невозможно без понимания вклада ключевых признаков.

Методы отбора признаков

Методы отбора признаков можно поделить на три основные категории [5].

Методы фильтрации. В данных методах используются независимые от модели статистические показатели, такие как коэффициенты корреляции или оценки взаимной информации для исключения несущественных признаков. Отметим, что такие методы игнорируют взаимозависимость некоторых признаков и напрямую не оптимизируют производительность модели машинного обучения. Можно выделить несколько основных методов данной категории.

Корреляционный анализ используется для оценки линейной зависимости между числовым признаком и целевой переменной [6]. Признаки с абсолютным значением коэффициента, близким к нулю, считаются неинформативными и отсеиваются.

Критерий Хи-квадрат, применяемый в задачах классификации с категориальными признаками. Метод проверяет гипотезу о независимости признака и целевой

метки [7]. Чем выше значение статистики, тем сильнее зависимость и тем важнее признак.

Дисперсионный анализ (Analysis of Variance, ANOVA [8]) используется для выявления влияния признаков на целевую переменную. Данный метод оценивает, насколько значимо различаются средние значения признака в разных классах. Если средние значения одинаковы для всех классов, признак отбрасывается.

Встроенные методы. Методы применяются внутри конкретных моделей при обучении, используя штрафы для увеличения разреженности нерелевантных коэффициентов признаков. Можно выделить несколько основных методов данной категории.

В методе Лассо-регрессии (L1-регуляризация [9]) к линейной модели добавляется штраф, равный сумме абсолютных значений коэффициентов. Многие коэффициенты признаков в ходе обучения обнуляются, что приводит к автоматическому исключению неинформативных признаков.

В L2-регуляризации к функции потерь модели машинного обучения добавляется штраф, пропорциональный квадрату весов признаков.

Алгоритмы на основе деревьев рассчитывают важность признака по критерию Gini Importance [10]. Признаки, по которым реже всего происходит разбиение на ранних этапах, автоматически получают низкий ранг.

Оберточные методы. Методы отбирают признаки итеративно до заданного количества, максимизируя производительность классификационной модели по заданной метрике. Можно выделить несколько основных методов данной категории.

Последовательные методы получают оптимальное подмножество переменных путем итеративного добавления или исключения признаков на основе оценки качества модели. В отличие от методов фильтрации, оценивающих признаки независимо от классификатора, последовательные подходы напрямую учитывают взаимодействие между переменными и их совместный вклад в предсказательную способность модели. Примерами таких методов являются метод последовательного отбора признаков [11] и метод рекурсивного отбора признаков (Recursive Feature Elimination, RFE [12]).

Популяционные методы отбора признаков, основанные на эвристических алгоритмах, таких как генетические алгоритмы, представляют собой подход к поиску оптимального подмножества признаков, преодолевающий ограничения локальной оптимизации, характерные для последовательных методов. В данных подходах процесс отбора признаков моделируется как эволюция популяции особей, где каждая особь представляется в виде определенного набора признаков, а ее качество оценивается через фитнес-функцию, учитывающую как точность модели, так и размерность признакового пространства.

Заметим, что оберточные методы не гарантируют нахождения глобального оптимума из-за вычислительной сложности задачи перебора всех возможных вариаций множества признаков, но они обеспечивают эффективный баланс между вычислительной сложностью и точностью отбора, позволяя существенно снизить размерность пространства признаков при сохранении информативности данных.

Гибридные методы отбора признаков

Эффективность перечисленных методов в разделе «Методы отбора признаков» часто зависит от специфики данных или выбранной модели машинного обучения, поэтому на практике применяется не один из методов, а гибридные решения. Так, в работе [13] предложен гибридный метод выбора признаков Genetic Algorithm and Recursive Feature Elimination, объединяющий генетический алгоритм и метод RFE. Результаты демонстрируют эффективность предложенной модели в сокращении признакового пространства, а также лучшей точности классификатора на оптимизированном наборе по сравнению с 8-ю другими методами оптимизации.

В работе [14] представлен гибридный метод отбора признаков на основе роевого интеллекта и генетического алгоритма для поиска оптимального подмножества признаков [14]. Несмотря на то, что предложенный метод вычислительно затратный, он предлагает лучшие оптимизированные подмножества признаков на различных наборах данных.

В контексте настоящей работы рассматривается проблема оптимизации набора данных в медицинской сфере. Выбор признаков особенно важен в этой области, поскольку наборы данных часто имеют высокую размерность (гены, изображения, волновые формы), а размер выборки ограничен. Так как рассматриваемая область достаточно обширна, в качестве медицинских данных, рассматриваются данные, полученные в ходе диагностики пациентов на выявления рака молочной железы.

Целью работы является построение гибридной модели отбора признаков в машинном обучении, направленной на выявление наиболее информативного подмножества признаков, обеспечивающего повышение качества классификации, снижение вычислительной сложности и улучшение интерпретируемости моделей.

В отличие от описанных гибридных методов оптимизации, предлагаемая модель будет учитывать, не только оптимизированное множество признаков как единое целое, но также значимость и избыточность признаков в оптимизированном наборе за счет включения методов фильтрации в предлагаемую модель. Также в модель заложен инструмент отбора лучших классификаторов из заданного набора.

Модель отбора признаков

Пусть имеется некоторый набор данных $X = (x_{ij})_{i=1}^{N}{}_{j=1}^{d}$, где $X \in R^{N\times d}$; N — количество примеров; d — количество признаков. Каждому $x_i = (x_{ij})_{j=1}^{d}$ приписана метка $\mathbf{y} \in \{0, 1, \ldots, C-1\}$, C — количество классов, которая представляет собой вектор классов. Необходимо решить задачу отбора наиболее информативного подмножества признаков, обеспечивающего повышение качества классификации, снижение вы-

числительной сложности и улучшение интерпретируемости моделей. Для этого проведем формальную постановку задачи.

Пусть $D = \{1, 2, \ldots, d\}$ — множество номеров признаков, при этом множество отобранных номеров признаков обозначается через S, т. е. $S \subseteq D$. Определим оператор проекции: $\pi_S(x) = (x_j)_{j \in S}$, $x_j = (x_{ij})_{i=1}^{N}$. Рассмотрим ансамбль моделей $H = \{h_1, h_2, \ldots, h_k\}$ машинного обучения, где K — количество рассматриваемых моделей. Зададим обученную модель на признаках S в виде: $\hat{h}(S) = h^*(\pi_S(x_j)) = \arg\min_{h \in H} \frac{1}{N} \sum_{i=1}^{N} \lambda(h(\pi_S(x_j)_i), y_i)$, где h^* — лучшая модель из ансамбля; λ — функция потерь модели машинного обучения. Определим качество конкретной модели: $Q(\hat{h}(S)) = F1_{\text{macro}}(\hat{h}(S))$, где $F1_{\text{macro}}$ — средняя F1-мера по C классам. В итоге задача сводится к максимизации качества моделей

$$\max_{S \subseteq D} Q(\hat{h}(S)) \text{ при ограничении } 1 \le |S| \le d. \qquad (1)$$

Для нахождения подмножества признаков S введем вектор отбора признаков $\mathbf{w} = (w_1, w_2, \ldots, w_d)$, $w_j \in [0, 1]$. Вектор отбора признаков задает подмножество признаков S следующим образом: $S(\mathbf{w}) = \{j | w_j > 0{,}3\}$.

Для определения оптимального подмножества признаков $S(\mathbf{w})$ введем целевую функцию:

$$F(\mathbf{w}) = Q(\hat{h}(S(\mathbf{w}))) = \frac{1}{K} \sum_{k=1}^{K} F1_{\text{macro}}(h_k(S(\mathbf{w}))) - \\ - \eta_1 |S(\mathbf{w})| + \eta_2 \sum_{i \in S(\mathbf{w})} \sigma_i - \eta_3 Pcorr(S(\mathbf{w})), \qquad (2)$$
$$1 \le |S(\mathbf{w})| \le d,$$

где $\frac{1}{K} \sum_{k=1}^{K} F1_{\text{macro}}(h_k(S(\mathbf{w})))$ — средняя F1-мера для набора $S(\mathbf{w})$, берется из предположения, что если признаки значимы для классификации, то они будут примерно одинаково работать в разных моделях; η_1, η_2, η_3 — параметры регуляризации, отвечающие за влияние на конечный результат. Рассмотрим слагаемые уравнения (2) подробнее: $\eta_1 |S(\mathbf{w})|$ — штраф за количество признаков, так как оптимизируется набор, чем меньше, тем лучше; $\eta_2 \sum_{i \in S(\mathbf{w})} \sigma_i$ — поощрение за признаки с высоким статистическим качеством, $\sigma = \frac{1}{3}(q_1 + q_2 + q_3)$; $q_1 = 1 - p$, где p — значение из дисперсионного анализа (метод ANOVA [10]) признака j и целевой переменной, чем выше значение q_1, тем статистически более значима связь между признаком и целевой переменной; $q_2 = 1 - \mathbf{q}_{\text{row}}^{\text{scaled}}$, где $\mathbf{q}_{\text{row}} = \sum_{c=0}^{C-1} \sum_{j \in D} (x_j - \mu_c)^T (x_j^c - \mu_j^c)$, $\mu_{ij}^c = 1/N_c \sum_{i \in I_c} x_{ij}$ — инверсионная нормализированная ковариационная внутриклассовая матрица, чем выше значение q_2, тем лучше признак описывает классы C, $\mathbf{x}_j^c = (x_{ij})_{i \in I_c}$ — вектор-столбец значений признака x_j для примеров x_i, принадлежащих классу c ($y_i = c$), $I_c = \{i = 1, \ldots, N | y_i = c\}$; q_3 — коэффициент корреляции Пирсона признака с целевой переменной. $Pcorr(S(\mathbf{w}) = \frac{1}{|S(\mathbf{w})|/2} \sum_{j_1, j_2} \max\left(0, \frac{R_{j_1 j_2} - \tau}{1 - \tau}\right)$ — корреляционный штраф для множества выбранных признаков $S(\mathbf{w})$, R_{j_1, j_2} — коэффициент корреляции Пирсона между двумя признаками x_{j_1} и x_{j_2}, $\tau = 0{,}8$ — пороговое значение коэффициента корреляции, больше которого корреляция полагается высокой.

В связи с использованием в (2) слагаемых на основе статистических оценок, на данную модель накладывается некоторые ограничения. В частности, предложенная модель может оптимизировать набор признаков, значения которых измерялись в интервальной шкале. Отметим, что при нарушении данного ограничения, необходимо эти слагаемые заменить на другие, применимые к используемым шкалам измерения.

Метод отбора признаков

Для нахождения оптимального подмножества признаков $S(\mathbf{w})$ необходимо максимизировать функцию (2), т. е. найти $\arg\max_{\mathbf{w} \in [0,1]^d} F(\mathbf{w})$. Так как полный перебор $S(\mathbf{w})$ невозможен из-за экспоненциального числа подмножеств, был выбран генетический алгоритм для эвристического решения данной задачи.

Проведем формализацию задачи (1) в терминах модели генетического алгоритма.

Геномом особи является вектор отбора признаков $\mathbf{w}^l = (w_1^l, w_2^l, \ldots, w_d^l)$, где $l = 1, \ldots, L$ — номер особи; L — количество особей. Особь — подмножество признаков $S(\mathbf{w})$. Целью отбора являются особи, для которых функция (2) достигает большего значения.

Инициализация популяции происходит следующим образом. Генерируется начальная популяция P_0 из $L = 100$ особей $S(w^{0l})$, $l = 1, \ldots, L$. Ген w^{0l} для каждой особи генерируется случайно с использованием равномерного распределения $U(0, 1)$ на интервале (0, 1):

$$w_i^{0l} \sim U(0, 1),\ l = 1, \ldots, L,\ i = 1, \ldots, N.$$

Для каждой особи вычисляется функция (2) и лучшие 5 % особей изымаются из P_0, составляют множество P_b и переходят в следующее поколение без каких-либо изменений. Здесь и далее будет считаться лучшей та особь, у которой функции (2) принимает большее значение.

С оставшимся множеством выполняется этап селекции. Проводится турнир, заключающийся в том, что случайно берутся три особи, из них выбирается лучшая и копируется в множество претендентов P, затем эти три особи возвращаются в свое поколение. Всего проводится L–5 % турниров для сохранения размера популяции.

Множество P случайным образом разбивается на пары $\{A, B\}$ (если $|P|$ нечетно, то оставшаяся одинокая особь не скрещивается). Далее каждая из пар $\{A, B\}$ с вероятностью 0,7 подвергаются скрещиванию по следующему правилу:

$$\mathbf{w}^{\text{child1}} = (w_1^A, \ldots, w_p^A, w_{p+1}^B, \ldots w_d^B),$$
$$\mathbf{w}^{\text{child2}} = (w_1^B, \ldots, w_p^B, w_{p+1}^A, \ldots w_d^A),$$

где p — случайная точка разреза гена (для каждой пары своя). Из получившихся особей $S(\mathbf{w}^{\text{child1}})$, $S(\mathbf{w}^{\text{child2}})$ и

тех особей, которые не подвергались скрещиванию, формируется следующее множество P' претендентов, а предыдущее множество P далее участие не принимает.

Затем с вероятностью 0,1 происходит мутация генов у каждой особи $A \in P'$ по следующему правилу:

$$w_i'^A = w_i^A + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0;\, 0{,}1),$$

где ε — шум; w_i^A и $w_i'^A$ — гены особи A до и после мутации; $\mathcal{N}(0;\, 0{,}1)$ — нормальное распределение с математическим ожиданием 0 и дисперсией 0,1. При этом отрицательные значения $w_i'^A$ заменяются на 0, а большие единицы — на 1. Получившиеся мутировавшие особи и особи, которые мутации не подвергались, составили множество P''.

В итоге, следующим поколением является $P_b \in P''$. Далее процедура отбора повторяется для получившегося и последующих поколений. Алгоритм завершает работу, когда количество поколений достигнет 30, так как на дальнейших поколениях целевая функция (2) не изменялась на рассмотренных далее наборах данных.

Упомянутые значения размера популяции, туров, а также вероятность мутации были определены эмпирическим путем в ходе нескольких вычислительных экспериментов, как значения, дающие лучший результат.

Например, для генетического алгоритма вычислялось оптимальное значение начальной популяции. Брались популяции размером от 10 до 200 с шагом 10, и на каждом шаге вычислялось значение функции (2). Вычислительный эксперимент останавливался, когда значение функции (2) переставало изменяться.

Для параметров функции η_1, η_2, η_3 для 30 экспериментов генерировались значения в диапазоне [0,01, 0,1] с использованием метода Latin Hypercube Sampling. Полученные значения параметров, которые давали лучшие результаты для классификации на тестовом наборе данных, использовались в данной работе.

Вычислительный эксперимент

Для эмпирической валидации предложенной модели отбора признаков и комплексной оценки ее эффективности был проведен вычислительный эксперимент. Эксперимент проводился на двух наборах данных, находящихся в предметной области — диагностика рака молочной железы. Набор данных 1 содержал данные микроволновой радиотермометрии молочных желез, полученных в результате обследований пациентов, проведенных медицинскими учреждениями.

Эффективность предлагаемого метода проанализирована в сравнении с несколькими методами. Из имеющихся категорий методов встроенные методы жестко привязаны к выбранной модели машинного обучения и не позволяют варьировать эти модели, поэтому эта категория не рассматривалась. В качестве представителя категории оберточных методов был выбран метод RFE, который часто используется для анализа медицинских данных. Из категории фильтрующих методов был взят корреляционный анализ, в результате которого удаляются сильно коррелируемые признаки в связи с тем, что для медицинских данных характерна довольно высокая корреляция измеряемых признаков. В качестве критерия оценки методов отбора признаков было выбрано количество отобранных признаков при условии сохранения точности модели классификации на тестовом наборе данных. Отметим, что, как и в любой области, в медицине важным аспектом является не только высокая предсказательная точность модели, но и ее интерпретируемость. При этом, чем меньше будет использоваться признаков в финальной модели, тем проще будет интерпретировать результаты для врачей.

Набор данных 1 содержал измерения в 18 точках молочных желез, две точки в аксилярной области и две опорные точки между молочными железами в инфракрасном и радиодиапазонах произведенных при помощи микроволнового радиотермометра РТМ-01М. Заметим, что метки, выставленные врачом на основе измеренных температур, включали в себя 6 категорий температурных аномалий, характерных для различных заболеваний молочной железы, в том числе и рака.

Категории характеризуют температурные аномалии в молочных железах и определялись следующим образом [15].

Категория 0 — температурных аномалий не выявлено (здоровый пациент).

Категория 1 — имеется участок понижения температуры на фоне умеренно выраженных тепловых изменений. Как правило, к этой категории относятся доброкачественные процессы (локализованный фиброз, липома).

Категория 2 — имеется участок повышения температуры на фоне умеренно выраженных тепловых изменений. Как правило, к этой категории также относятся доброкачественные процессы (участок аденоза, участок воспаления на коже).

Категория 3 — наблюдается высокий уровень температур в инфракрасном и радиодиапазонах. Как правило, к этой категории относятся здоровые пациенты и больные раком молочной железы с медленным ростом.

Категория 4 — имеется много признаков тепловой активности (наблюдается системное различие температур, т. е. температуры в одной из желез выше, чем в другой). Как правило, к этой категории относятся пациенты с подозрением на рак молочной железы.

Категория 5 — относятся уверенные заключения, когда наблюдается термоасимметрия в одинаковых точках молочной железы. Как правило, в категорию входят пациенты с острым воспалением или раком молочной железы.

На основании исходных температур в работе [16] было получено признаковое пространство, состоящее из 62 признаков, которое численно описывает те признаки (термоассиметрия, отклонение от температуры опорных точек и т. д.), исходя из которых врач принимает решение о выставлении категории температурных аномалий.

В результате набор данных содержит 9310 записей о молочных железах, разделенных на 6 категорий. Обозначим набор данных: $X = (x_{ij})_{i=1}^{N}{}_{j=1}^{d}$, где $X \in R^{N\times d}$; $N = 9310$ — количество наблюдений; $d = 62$ — количество признаков. Каждому $x_i = (x_{ij})_{j=1}^{d}$ приписана метка $y_i \in \{0, \ldots, C\}$, которая показывает номер класса, отра-

жающий наличие температурных аномалий в молочной железе, $i = 1, \ldots, N$, $C = 5$.

Пусть $I_k = \{i \in I | y_i = k\}$, где $I = \{1, \ldots, N\}$, $k = 0, \ldots, C$.

Разделим набор данных X на две выборки (обучающую X^{train} и тестовую X^{test}) так, что в каждой сохраняется исходное распределение классов. Таким образом, проведено разбиение $I_k = I_k^{\text{train}} \cup I_k^{\text{test}}$ случайным образом так, что $n_k^{\text{train}} \approx 0{,}8 n_k$, где $n_k^{\text{train}} = |I_k^{\text{train}}|$, $n_k = |I_k|$ и $X^{\text{train}} = (x_i)_{i \in I^{\text{train}}}$, $X^{\text{test}} = (x_i)_{i \in I^{\text{test}}}$, где $I^{\text{train}} = \bigcup_{k=0}^{C} I_k^{\text{train}}$, $I^{\text{test}} = \bigcup_{k=0}^{C} I_k^{\text{test}}$.

В результате получим два набора данных $N^{\text{train}} = \sum_{k=0}^{C} n_k^{\text{train}} = 7448$, $N^{\text{test}} = N - N^{\text{train}} = 1862$.

Набор данных N^{train} будет использоваться для оптимизации признакового пространства согласно модели (1), а N^{test} только для проверки работы модели машинного обучения по сравнению с полным набором признаков.

Определим ансамбль моделей машинного обучения

$$H = \{h_1, h_2, h_3\},$$

где h_1 — модель случайного леса [17]; h_2 — модель логистической регрессии [18]; h_3 — модель классификатора, в основе которого применяется метод опорных векторов [19].

Так как в настоящей работе рассматривается предметная область — медицинская диагностика, одним из основных критериев качества модели служит интерпретируемость результатов, поэтому в целевую функцию заложены именно интерпретируемые модели.

Пусть параметры регуляризации в функции (2) равны $\eta_1 = 0{,}01$, $\eta_2 = 0{,}005$, $\eta_3 = 0{,}1$, значения которых были получены эмпирическим путем.

Выполним вычисления при помощи программного комплекса, написанного на языке Python с использованием библиотеки для моделей машинного обучения scikit-learn. В результате работы генетического алгоритма, максимизирующего функцию (2), удалось сократить исходное признаковое пространство с 62 до 15.

Лучшее значение целевой функции (2) было найдено на 17-м поколении. На рисунке показан график изменения значения целевой функции в ходе работы генетического алгоритма.

Наиболее перспективной моделью оказалась логистическая регрессия h_2(1). Благодаря высокому значению η_3 в оптимизированном множестве признаков большинство коррелирующих пар не были включены.

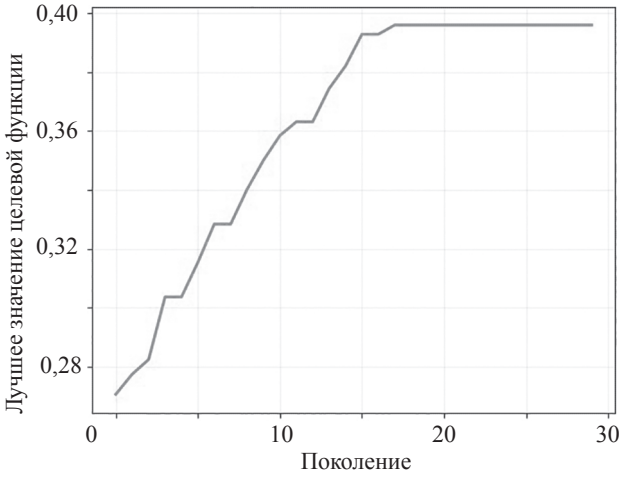

Рисунок. График изменения значения целевой функции в ходе работы генетического алгоритма

Figure. Graph of changes in the value of the objective function during the operation of the genetic algorithm

Это важно, поскольку в работе [20] было показано, что довольно много признаков с высокой корреляцией.

Проверим, что полученный набор признаков позволяет точно классифицировать данные. Выполним анализ полученного признакового пространства на наборе X^{test} в сравнении с тремя моделями, построенными на основе модели h_2:

1) с моделью h_2(full), обученной на полном исходном наборе из 62 признаков;
2) с моделью h_2(RFE), содержащей 15 признаков, отобранных методом RFE;
3) с моделью h_2(corr), обученной на наборе из 29 признаков, полученных в результате корреляционного анализа (признаки с коэффициентом корреляции выше 0,8 были удалены).

В табл. 1 показано сравнение верно классифицированных наблюдений, разделенных по классам (жирным выделены лучшие результаты).

В исследованных наборах признаков только 7 встречаются во всех трех наборах отобранных признаков, что говорит о разных подходах в механизме отбора рассмотренных методов классификации.

Предложенную модель протестировали на наборе данных 2 (открытый набор данных по раку молочной железы в Висконсине) в той же предметной области. Набор содержал 30 признаков, которые вычислялись на основе оцифрованного изображения тонкоигольной аспирационной биопсии (ТИАБ) образования в молоч-

Таблица 1. Сравнение верно классифицированных наблюдений на основе логистической регрессии

Table 1. Comparison of correctly classified observations based on logistic regression

Модель	Категория 0	Категория 1	Категория 2	Категория 3	Категория 4	Категория 5
h_2(full)	1064	238	103	15	29	**22**
h_2(RFE)	**1066**	239	89	19	**35**	19
h_2(corr)	925	230	**106**	**21**	32	**22**
h_2(1)	1062	**245**	97	17	32	20

Таблица 2. Сводная таблица точности разных моделей
Table 2. Summary table of accuracy of different models

Используемый набор данных	Модель			
	h_2(full)	h_2(RFE)	h_2(corr)	h_2(1)
Набор данных 1	0,7900	0,7879	0,7175 (29 признаков)	0,7911
Набор данных 2	0,9474	0,9386	0,9298 (14 признаков)	0,9561

ной железе. Также в наборе был определен диагноз пациента: доброкачественное или злокачественное образование и содержал 569 записей о ТИАБ, разделенных на два класса. В результате работы модели удалось оптимизировать набор до 10 признаков.

Обсуждение

По результатам проведенных вычислительных экспериментов можно сделать вывод об эффективности модели $h_2(1)$. В табл. 2 представлена сводная таблица результатов.

Из табл. 2 видно, что для исследованных наборов данных средняя точность гибридной модели $h_2(1)$ увеличилась по сравнению с исходной моделью h_2(full).

На основании полученных результатов эксперимента можно сделать вывод, что предложенный метод не только эффективно сокращает признаковое пространство, но и увеличивает точность модели классификации, несмотря на существенное уменьшение количества признаков.

Однако следует отметить, что вычислительная сложность предложенной модели является наиболее времязатратной из-за использования в целевой функции не одной, а трех моделей. Так, например, основная вычислительная сложность корреляционного анализа составляет $O(d^2 \times N)$.

Пренебрегая особенностями обучения моделей машинного обучения, вычислительная сложность метода RFE составила $O(d^2 \times N \times z)z \leq d$, где z — количество шагов рекурсии.

Для вычисления сложности предложенной модели необходимо оценить сложность вычисления целевой функции. В частности, в нее заложено обучение трех моделей машинного обучения и проведение четырех статистических анализов (два корреляционных анализа, дисперсионный анализ и построение внутриклассовой матрицы), время, затраченное на учет слагаемого $\eta_1|S(\mathbf{w})|$ является константой. Вычислительная сложность одной особи составила $O(d^2 \times N)$. Так как количество эпох и особей задано, общая вычислительная сложность — $O(d^2 \times N \times L \times G)$, где G — количество эпох. Можно видеть, что вычислительная сложность предложенного метода будет сильно возрастать при увеличении L и G. По этой причине имеет смысл ограничиться их небольшими значениями, как было сделано в представленном вычислительном эксперименте ($G = 30$, $L = 100$).

Заключение

В работе предложена модель оптимизации признакового пространства с использованием комбинированного подхода, учитывающего такие факторы как заданную меру эффективности моделей, например, F1-мера, качество признаков на основе статистического анализа и учет корреляции между признаками. Помимо оптимизации признакового пространства, модель выявляет из числа заданных наилучший классификатор для используемого набора.

Показана эффективность предложенной модели на наборе данных, состоящего из 62 признаков, в результате которого удалось сократить с 62 до 15 наиболее эффективных, при этом, средняя точность с 0,7900 незначительно увеличилось до значения 0,7911 для полинарной классификации. Результаты доказали, что предложенный метод не только эффективен для сжатия набора данных, но и показывает значения эффективности лучше, чем методы рекурсивного отбора признаков и корреляционный анализ. В сравнении с рассмотренными моделями оптимизации признакового пространства, предложенная модель имеет некоторые преимущества, такие как выбор лучшей из моделей классификации, учет корреляционного анализа и статистическую значимость признака.

Отметим, что предложенный метод был апробирован только на двух наборах данных и не гарантирует нахождения оптимального набора в связи с использованием в его основе генетического алгоритма, поэтому требуются дополнительные исследования на разных типах данных и других моделях машинного обучения, например, неинтерпритируемых.

Литература

1. Pugliese R., Regondi S., Marini R. Machine learning-based approach: global trends, research directions, and regulatory standpoints // Data Science and Management. 2021. V. 4. P. 19–29. doi: 10.1016/j.dsm.2021.12.002
2. Sarker I.H. Machine learning: algorithms, real-world applications and research directions // SN Computer Science. 2021. V. 2. N 3. P. 160. doi: 10.1007/s42979-021-00592-x
3. Chen R-C.., Dewi C., Huang S.-W., Caraka R.E. Selecting critical features for data classification based on machine learning methods //

References

1. Pugliese R., Regondi S., Marini R. Machine learning-based approach: global trends, research directions, and regulatory standpoints. *Data Science and Management*, 2021, vol. 4, pp. 19–29. doi: 10.1016/j.dsm.2021.12.002
2. Sarker I.H. Machine learning: algorithms, real-world applications and research directions. *SN Computer Science*, 2021, vol. 2, no. 3, pp. 160. doi: 10.1007/s42979-021-00592-x
3. Chen R-C.., Dewi C., Huang S.-W., Caraka R.E. Selecting critical features for data classification based on machine learning methods.

Journal of Big Data. 2020. V. 7. N 1. P. 52. doi: 10.1186/s40537-020-00327-4
4. Cai J., Luo J., Wang S., Yang S. Feature selection in machine learning: A new perspective // Neurocomputing. 2018. V. 300. P. 70–79. doi: 10.1016/j.neucom.2017.11.077
5. Fira M., Goras L., Costin H.-N. Evaluating sparse feature selection methods: a theoretical and empirical perspective // Applied Sciences. 2025. V. 15. N 7. P. 3752. doi: 10.3390/app15073752
6. Pearson K. Notes on regression and inheritance in the case of two parents // Proceedings of the Royal Society of London. 1895. V. 58. P. 240–242.
7. Chernoff H., Lehmann E.L. The use of maximum likelihood estimates in χ^2 tests for goodness of fit // The Annals of Mathematical Statistics. 1954. V. 25. N 3. P. 579–586.
8. Larson M.G. Analysis of variance // Circulation. 2008. V. 117. N 1. P. 115–121. doi: 10.1161/circulationaha.107.654335
9. Santosa F., Symes W.W. Linear inversion of band-limited reflection seismograms // SIAM Journal on Scientific and Statistical Computing. 1986. V. 7. N 4. P. 1307–1330. doi: 10.1137/0907087
10. von Winterfeldt D., Edwards W. Decision Trees // Decision Analysis and Behavioral Research. 1986. P. 63–89.
11. Ferri F.J., Pudil P., Hatef M., Kittler J. Comparative study of techniques for large-scale feature selection // Machine Intelligence and Pattern Recognition. 1994. V. 16. P. 403–413. doi: 10.1016/B978-0-444-81892-8.50040-7
12. Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines // Machine Learning. 2002. V. 46. N 1-3. P. 389–422. doi: 10.1023/A:1012487302797
13. Rani P., Kumar R., Jain A., Chawla S.K. A Hybrid approach for feature selection based on genetic algorithm and recursive feature elimination // International Journal of Information System Modeling and Design. 2021. V. 12. N 2. P. 17–38. doi: 10.4018/IJISMD.2021040102
14. Benghazouani S., Nouh S., Zakrani A., Haloum I., Jebbar M. Enhancing feature selection with a novel hybrid approach incorporating genetic algorithms and swarm intelligence techniques // International Journal of Electrical and Computer Engineering. 2024. V. 14. N 1. P. 944–959. doi: 10.11591/ijece.v14i1.pp944-959
15. Тихомирова Н.Н. Техника проведения РТМ-обследования молочных желез. 2008. 64 с.
16. Лосев А.Г., Левшинский В.В. Интеллектуальный анализ данных микроволновой радиотермометрии в диагностике рака молочной железы // Математическая физика и компьютерное моделирование. 2017. Т. 20. № 5. С. 49–62. doi: 10.15688/mpcm.jvolsu.2017.5.6
17. Breiman L. Random forests // Machine Learning. 2001. V. 45. N 1. P. 5–32. doi: 10.1023/A:1010933404324
18. Yu H.-F., Huang F.-L., Lin C.-J. Dual coordinate descent methods for logistic regression and maximum entropy models // Machine Learning. 2011. V. 85. N 1-2. P. 41–75. doi: 10.1007/s10994-010-5221-8
19. Chang C.-C., Lin C.-J. LIBSVM: A library for support vector machines // ACM Transactions on Intelligent Systems and Technology. 2011. V. 2. N 3. P. 1–27. doi: 10.1145/1961189.1961199
20. Гермашев И.В., Дубовская В.И., Лосев А.Г., Попов И.Е. Факторный анализ влияния признаков на точность диагностики рака молочной железы по данным микроволновой радиотермометрии // Прикаспийский журнал: управление и высокие технологии. 2022. № 1 (57). С. 139–148.

Journal of Big Data, 2020, vol. 7, no. 1, pp. 52. doi: 10.1186/s40537-020-00327-4
4. Cai J., Luo J., Wang S., Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*, 2018, vol. 300, pp. 70–79. doi: 10.1016/j.neucom.2017.11.077
5. Fira M., Goras L., Costin H.-N. Evaluating sparse feature selection methods: a theoretical and empirical perspective. *Applied Sciences*, 2025, vol. 15, no. 7, pp. 3752. doi: 10.3390/app15073752
6. Pearson K. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 1895, vol. 58, pp. 240–242.
7. Chernoff H., Lehmann E.L. The use of maximum likelihood estimates in χ^2 tests for goodness of fit. *The Annals of Mathematical Statistics*, 1954, vol. 25, no. 3, pp. 579–586.
8. Larson M.G. Analysis of variance. *Circulation*, 2008, vol. 117, no. 1, pp. 115–121. doi: 10.1161/circulationaha.107.654335
9. Santosa F., Symes W.W. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 1986, vol. 7, no. 4, pp. 1307–1330. doi: 10.1137/0907087
10. von Winterfeldt D., Edwards W. Decision Trees. *Decision Analysis and Behavioral Research*, 1986, pp. 63–89.
11. Ferri F.J., Pudil P., Hatef M., Kittler J. Comparative study of techniques for large-scale feature selection. *Machine Intelligence and Pattern Recognition*, 1994, vol. 16, pp. 403–413. doi: 10.1016/B978-0-444-81892-8.50040-7
12. Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning*, 2002, vol. 46, no. 1-3, pp. 389–422. doi: 10.1023/A:1012487302797
13. Rani P., Kumar R., Jain A., Chawla S.K. A Hybrid approach for feature selection based on genetic algorithm and recursive feature elimination. *International Journal of Information System Modeling and Design*, 2021, vol. 12, no. 2, pp. 17–38. doi: 10.4018/IJISMD.2021040102
14. Benghazouani S., Nouh S., Zakrani A., Haloum I., Jebbar M. Enhancing feature selection with a novel hybrid approach incorporating genetic algorithms and swarm intelligence techniques. *International Journal of Electrical and Computer Engineering*, 2024, vol. 14, no. 1, pp. 944–959. doi: 10.11591/ijece.v14i1.pp944-959
15. Tikhomirova N.N. *The Technique of the RTM-Breast Examination*. 2008, 64 p. (in Russian)
16. Losev A., Levshinskiy V. Data mining of microwave radiometry data in the diagnosis of breast cancer. *Mathematical Physics and Computer Simulation*, 2017, vol. 20, no. 5, pp. 49–62. (in Russian). doi: 10.15688/mpcm.jvolsu.2017.5.6
17. Breiman L. Random forests. *Machine Learning*, 2001, vol. 45, no. 1, pp. 5–32. doi: 10.1023/A:1010933404324
18. Yu H.-F., Huang F.-L., Lin C.-J. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 2011, vol. 85, no. 1-2, pp. 41–75. doi: 10.1007/s10994-010-5221-8
19. Chang C.-C., Lin C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011, vol. 2, no. 3, pp. 1–27. doi: 10.1145/1961189.1961199
20. Germashev I.V., Dubovskaya V.I., Losev A.G., Popov I.E. Factor analysis of the influence of signs on the accuracy of breast cancer diagnosis according to microwave radiothermometry. *The Caspian Journal: Management and High Technologies*, 2022, no. 1 (57), pp. 139–148. (in Russian)

Авторы

Дёмин Кирилл Сергеевич — аспирант, Волгоградский государственный университет, Волгоград, 400062, Российская Федерация, https://orcid.org/0009-0002-4571-3437, diominkirill@yandex.ru
Гермашев Илья Васильевич — доктор технических наук, профессор, профессор, Волгоградский государственный университет, Волгоград, 400062, Российская Федерация, SC 6602890505, https://orcid.org/0000-0001-5507-8508, germashev@volsu.ru

Authors

Kirill S. Dyomin — PhD Student, Volgograd State University, Volgograd, 400062, Russian Federation, https://orcid.org/0009-0002-4571-3437, diominkirill@yandex.ru
Ilya V. Germashev — D.Sc., Full Professor, Volgograd State University, Volgograd, 400062, Russian Federation, SC 6602890505, https://orcid.org/0000-0001-5507-8508, germashev@volsu.ru

Статья поступила в редакцию 12.02.2026
Одобрена после рецензирования 02.06.2026
Принята к печати 21.07.2026

Received 12.02.2026
Approved after reviewing 02.06.2026
Accepted 21.07.2026