УДК 519.254
Doi: 10.31772/2712-8970-2021-22-1-18-31

Для цитирования: Денисов М. А., Сопов Е. А. Генетический алгоритм условной оптимизации для проектирования информативных признаков в задачах классификации // Сибирский аэрокосмический журнал. 2021. Т. 22, № 1. С. 18–31. Doi: 10.31772/2712-8970-2021-22-1-18-31.

For citation: Denisov M. A., Sopov E. A. Constraint handling genetic algorithm for feature engineering in solving classification problems // Siberian Aerospace Journal. 2021, Vol. 22, No. 1, P. 18–31. Doi: 10.31772/2712-8970-2021-22-1-18-31.

Генетический алгоритм условной оптимизации для проектирования информативных признаков в задачах классификации

М. А. Денисов*, Е. А. Сопов

Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева
Российская Федерация, 660037, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
*E-mail: denisov.maksim.work@gmail.com

Проектирование признаков в машинном обучении является перспективным, но недостаточно изученным направлением. Создание нового пространства признаков из исходного набора позволяет повысить эффективность алгоритма машинного обучения, применяемого для решения сложных задач интеллектуального анализа данных. Некоторые методы отбора часто способны одновременно при увеличении точности классификации уменьшить исходное пространство, что особенно актуально в эпоху больших данных.

В работе предлагается новый подход машинного обучения к решению задачи классификации на основе методов проектирования информативных признаков. Проектирование информативных признаков осуществляется с помощью методов извлечения и отбора. На основании исходных данных созданы новые множества признаков, которые включают исходные признаки и признаки, полученные методом главных компонент. Выбор эффективного подмножества информативных признаков реализуется с использованием генетического алгоритма. Для того чтобы избежать переобучения и создания тривиальных классификаторов, на функцию пригодности генетического алгоритма накладываются ограничения, требующие определенного количества признаков исходной выборки, а также определенного количества признаков, полученных методом главных компонент. Проведен сравнительный анализ эффективности следующих алгоритмов классификации: k-ближайших соседей, метод опорных векторов и случайный лес. Эксперименты по исследованию эффективности проводятся путем решения прикладных задач бинарной классификации из репозитория задач машинного обучения UCI Machine Learning. В качестве критерия эффективности выбрана мера macro F1-score.

Результаты численных экспериментов показали, что точность классификации предложенным подходом превосходит решения, полученные на исходном наборе признаков и при случайном отборе (оценка границы снизу). Причем, увеличение точности характерно для всех типов задач (выборки, у которых количество признаков больше числа объектов, а также объемом 500 значений и более). Подтверждена статистическая значимость результатов.

Ключевые слова: отбор признаков, извлечение признаков, генетический алгоритм, условная оптимизация.

Constraint handling genetic algorithm for feature engineering in solving classification problems

M. A. Denisov*, E. A. Sopov

Reshetnev Siberian State University of Science and Technology
31, Krasnoiarskii Rabochi Prospekt, Krasnoyarsk, 660037, Russian Federation
*E-mail: denisov.maksim.work@gmail.com

Feature engineering in machine learning is a promising but still insufficiently studied domain. Creating new feature space from an original set allows increasing the accuracy of the machine learning algorithm chosen to solve complex data mining problems. Some existing selection methods are capable of simultaneously increasing the accuracy and reducing feature space. The reduction is an urgent task for big data problems.

The paper considers a novel machine learning approach for solving classification problems based on feature engineering methods. The approach constructs informative features using feature selection and extraction methods. Original data and features obtained by principal component analysis form a new set of features. The genetic algorithm selects an effective subset of informative features. It is important to avoid overfitting and builng a trivial classifier. Therefore, the fitness function is constrained for producing the given number of original features and the given number of features obtained by principal component analysis. The paper describes a comparative analysis of three classifiers, namely k-nearest neighbors, support vector machine and random forest. In order to prove the accuracy improvement, the authors examine several real-world problems chosen from the UCI Machine Learning repository. The accuracy measure in the study is the macro F1-score.

The results of numerical experiments show that the proposed approach outperforms the performance obtained using the original data set and the performance of random feature selection (the low bound for the results). Moreover, the accuracy enhancement is obtained for all types of problems (data sets that have more features than values). All results are proved to be statistically significant.

Keywords: feature selection, feature construction, genetic algorithm, constraint optimization.

Введение. Машинное обучение является неотъемлемой частью современных информационных технологий и находит активное применение во многих областях. Например, для распознавания текста, написанного от руки, классификации изображений, спам фильтрации [1–3]. Наука и техника, медицина, экономика и другие отрасли также активно используют алгоритмы машинного обучения при решении сложных прикладных задач [4; 5]. Обучающие данные являются ключевой составляющей для алгоритмов машинного обучения. На практике при анализе данных может оказаться, что часть признаков неинформативна. Такие признаки либо нерепрезентативны, либо имеют сильную корреляцию друг с другом. При наличии нерепрезентативных признаков, вклад которых в итоговую точность незначителен либо отсутствует, обычно используются методы из класса отбора признаков (*Feature Selection*) [6; 7]. В ситуациях, когда признаки сильно взаимосвязаны, т. е. одинаковым образом влияют на предсказательную способность системы, используются методы конструирования признаков (*Feature Construction*) или их извлечения (*Feature Extraction*) [8; 9]. На современном этапе упомянутые подходы обобщены в единый термин – проектирование признаков (*Feature Engineering*) [10; 11].

В последнее время методы проектирования признаков активно исследуются и развиваются. С появлением области больших данных задача снижения размерности пространства признаков

стала еще более актуальной [12]. Методы отбора признаков позволяют значительно снизить требуемую вычислительную мощность компьютера, сохраняя или увеличивая при этом точность прогноза. В то же самое время снизить исходную размерность признакового пространства пытаются путем его трансформации, преобразования в новое, меньшей размерности [13]. Однако исследований в этом направлении по-прежнему недостаточно. В данной работе предлагается объединить техники извлечения и отбора признаков вместе, чтобы получить новое представление исходных данных, которое увеличивает предсказательную способность. Рассматривается задача бинарной классификации. В качестве техники извлечения используется метод главных компонент (МГК) [14]. Далее полученные признаки объединяются с исходной выборкой. Последний шаг – это отбор информативных признаков с помощью генетического алгоритма (ГА), на который дополнительно накладываются ограничения, заданные пользователем с учетом практических целей решения задачи или программной, или аппаратной реализацией.

Статья организована следующим образом. В первом разделе рассматриваются существующие работы по тематике исследования. Второй раздел нацелен на подробное описание предлагаемого способа проектирования признаков с использованием МГК и ГА. В третьем разделе представлено описание вычислительных экспериментов. В заключении подводится итог и обсуждаются дальнейшие перспективы исследования.

1. Анализ литературы по теме. Несмотря на то, что задачами конструирования и извлечения признаков занимаются со второй половины XX в., терминология до сих пор не устоялась. Одни авторы используют единый термин «конструирование признаков» понимая под ним также и «извлечение признаков». Другие отдают предпочтение только «извлечению признаков». В данной работе решено разделять эти два понятия, поскольку они решают принципиально разные и, в общем случае, независимые задачи.

1.1. Конструирование признаков. Под конструированием будем понимать процесс создания новых признаков с помощью некоторых преобразований. В роли таких преобразований могут выступать как математические операции (сложение, вычитание, умножение и другие), так и логические (конъюнкция, дизъюнкция, импликация и т. д.). Обычно выбранный набор математических операторов является уникальным для каждой конкретной задачи и не поддается обобщению [15; 16]. В [17] используется специальный критерий для поиска признаков, которые при объединении могли бы образовать новый, способный дать лучшую точность отклика. В работе [18] для прикладной экономической задачи алгоритм классификации, использующий выборку сконструированных признаков, показывает лучшие результаты по сравнению с классификатором, использующим исходные данные. Тем не менее все упомянутые подходы не могут быть обобщены на произвольные задачи.

В связи с этим, в конце XX – начале XXI в. развиваются алгоритмы, применение которых становится возможным в различных прикладных задачах. Среди таких можно упомянуть, например, *FRINGE* [19] и *CITRE* [20], которые используют бинарные операции и деревья решений для создания новых признаков. Авторы *FICUS* [15] решили усовершенствовать существующие подходы и помимо бинарных операций добавили стандартные математические, а также другие функции, которые могут быть предложены экспертом предметной области. Недостаток таких методов заключается в их вычислительной сложности. На каждой итерации в исходную выборку добавляются все новые и новые признаки, которые необходимо подавать на вход дереву решений. В результате дерево становится слишком большим.

Примерно в этот же период начали развиваться алгоритмы, основанные на генетическом программировании. Например, в работах [21; 22] популяция состоит из индивидов, представляющих собой закодированный набор арифметических и логических операторов. В ходе

эволюции с их помощью образуется новое пространство признаков, которое впоследствии подается на классификатор.

Существует также метод конструирования признаков с использованием индуктивного логического программирования для формирования предикатов на основе некоторых априорных знаний. В прикладных задачах его используют для устранения смысловой неоднозначности слов в процессе обработки и анализа компьютером естественного языка [23].

1.2. Извлечение признаков. Второй тип из данного класса задач – извлечение признаков. Под извлечением понимается изменение исходного пространства признаков путем уменьшения его размерности. Классическим методом является МГК и его различные вариации [24]. В общем смысле данная техника с помощью сингулярного разложения матрицы данных позволяет построить новые признаки, которые являются линейной комбинацией исходных. Полученные признаки некоррелированы, а исходная выборка не содержит избыточной информации, что является значительным преимуществом метода. Данный подход относят к классу обучения без учителя. Он не требует дополнительных сведений предметной области. Недостаток заключается в том, что новые данные больше не отражают исходного представления, т. е. их интерпретация становится почти невозможной.

Авторы данной статьи в своей работе используют метод МГК для извлечения признаков, которые впоследствии добавляются к исходному набору. Логика такой манипуляции заключается в самом принципе алгоритма. В процессе трансформации пространства первая главная компонента отражает наибольшую часть дисперсии всей совокупности данных. Каждая последующая компонента отражает эту дисперсию в меньшем объеме. Если отобрать несколько первых (наиболее значимых) компонент и добавить их к исходной выборке, можно усилить ее предсказательную способность. По сравнению с методами конструирования такой подход не требует никакого дополнительного знания предметной области и тем самым может быть использован для любой прикладной задачи.

1.3. Отбор признаков. Задача отбора признаков подразумевает сокращение их числа с целью повысить вычислительную производительность, улучшить интерпретируемость модели, при этом сохранив высокую точность классификации. Выполнение этих требований осуществляется за счет удаления из данных избыточных, нерелевантных признаков, а также тех, которые можно отнести к шуму. Работы [25; 26] показывают, что с помощью отбора можно не только сохранить, но и повысить точность прогноза.

Стратегии поиска принято подразделять на подвиды: методы фильтрации (*Filter methods*), методы обертки (*Wrapper Methods*) и встраиваемые методы (*Embedded methods*). Методы фильтрации не зависят от классификатора (никаким образом в него не встроены). Несмотря на то, что данный подход обычно требует небольшие вычислительные мощности, а также может быть относительно просто обобщен на различные прикладные задачи, он никак не учитывает итоговую точность, полученную в результате отбора. Методы обертки и встраиваемые наоборот оценивают полученный набор признаков с помощью метрики точности прогноза и на основании этой оценки улучшают процесс отбора в ходе, например, процесса эволюции. Отличительной чертой встраиваемых методов является встроенный (embedded) в структуру классификатора поиск оптимального набора признаков. Другими словами, отбор – это часть процесса обучения [27]. Для методов обертки критерием отбора признаков является точность работы классификатора, т. е. классификатор является как бы «оберткой» (wrapper) для алгоритма поиска.

Важно отметить, что отбор признаков – достаточно непростая задача по своей сути. Пространство поиска содержит $(2^n–1)$ возможных решений, где n – количество признаков в задаче. Если учесть, что человечество вступило в эру больших данных [12], когда признаки могут

исчисляться уже даже не в тысячах, а миллионах, можно представить, насколько сильно увеличивается трудоемкость. В исследовании [28] показано, что генетические алгоритмы эффективно применяются для решения задачи отбора признаков.

1.4. Адаптивный штраф. Генетические алгоритмы для решения задач условной оптимизации исследованы достаточно подробно. В работе [29] показаны различные варианты, включающие использование штрафных функций, специальных генетических операторов или алгоритмов на основе коэволюции.

В статье рассматривается задача оптимизации вида

$$f(x) \to \max_{x \in B_2^n}, \tag{1}$$

где $x = (x_1, \ldots, x_n)$, $x_i \in \{0,1\}$, $i = \overline{1,n}$.

В своей работе мы используем адаптивный штраф [30], который в среднем превосходит другие методы на основе штрафов. Математическое представление функции пригодности выглядит следующим образом:

$$F(x) = \begin{cases} f(x), & \text{если } x \text{ допустимо}, \\ \tilde{f}(x) - \sum_{j=1}^{l} k_j v_j(x), & \text{иначе}, \end{cases} \tag{2}$$

где $F(x)$ – функция пригодности, полученная методом штрафных функций; $f(x)$ – целевая функция; l в сумматоре равно числу ограничений, накладываемых на целевую функцию; v_j – численный размер нарушения ограничения j; k_j – параметр штрафной функции для ограничения j, который рассчитывается как

$$k_j = \frac{\left| \sum_{i=1}^{pop} f(x_i) \right|}{\sum_{s=1}^{l} \left[\sum_{i=1}^{pop} v_s(x_i) \right]^2} \sum_{i=1}^{pop} v_j(x_i), \tag{3}$$

где pop – размер популяции. $\tilde{f}(x)$ определяется как

$$\tilde{f}(x) = \begin{cases} f(x), & \text{если } f(x) > \langle f(x) \rangle, \\ \langle f(x) \rangle, & \text{иначе}, \end{cases} \tag{4}$$

где $\langle f(x) \rangle = \sum_{i=1}^{pop} f(x_i) \Big/ pop$.

2. Предлагаемый подход. Ограничение пространства поиска. В работе исследуется задача классификации, математическая постановка которой может быть представлена следующим образом. Пусть U – множество атрибутов, а Y – множество меток классов, т. е. наименований классов. Предполагается, что существует неизвестное отображение

$$y^* : U \to Y, \tag{5}$$

значение которого известно только на объектах обучающей выборки:

$$U^m = \{(u_1, y_1), \ldots, (u_m, y_m)\}. \tag{6}$$

Требуется построить алгоритм:

$$a : U \to Y, \tag{7}$$

способный классифицировать произвольное значение $u \in U$, образованное множеством признаков $u \in \{\overline{Attr} \in attr_i,\ i = 1, \ldots, n_1\}$. Обозначим множество признаков, полученных с помощью МГК, как $\overline{Attr}^{МГК} \in attr_i,\ i = 1, \ldots, n_2$.

В работе рассматривается классификация с использованием нескольких подходов проектирования признаков:

1) извлечение признаков с помощью МГК;

2) создание нового пространства признаков путем объединения исходных с МГК;

3) отбор генетическим алгоритмом признаков из пространства, полученного в 2 (исходные с МГК);

4) отбор признаков исходного множества случайным образом;

5) отбор признаков случайным образом из пространства, полученного в 2 (исходные с МГК).

Опишем подходы более подробно. В первом эксперименте оценивается точность классификации объектов, описываемых сконструированными признаками МГК. Для второго на классификатор подается новое признаковое пространство, полученное путем объединения исходных признаков выборки с МГК: $\overline{Attr} \cup \overline{Attr}^{МГК}$. В третьем эксперименте осуществляется отбор признаков генетическим алгоритмом. В качестве входных данных используется аналогичная второму эксперименту выборка: $\overline{Attr} \cup \overline{Attr}^{МГК}$. Используемая в третьем эксперименте стратегия поиска – метод обертки, где классификатор является оберткой для генетического алгоритма поиска. Ограничения, накладываемые на целевую функцию, описываются следующим образом:

$$\begin{cases} r_1 - \sum\limits_{i \in \overline{Attr}} x_i \le 0, \\ \sum\limits_{i \in \overline{Attr}} x_i - w_1 \le 0, \\ r_2 - \sum\limits_{i \in \overline{Attr}^{МГК}} x_i \le 0, \\ \sum\limits_{i \in \overline{Attr}^{МГК}} x_i - w_2 \le 0, \end{cases} \tag{8}$$

где r_1, r_2, w_1, w_2 – это параметры, обозначающие количество признаков, которые останутся в выборке; x_i, $i = 1, \ldots, n$ является хромосомой генетического алгоритма. Размер хромосомы складывается из мощности множества $\overline{Attr}$ и множества $\overline{Attr}^{МГК}$: $n = n_1 + n_2$. Ноль в хромосоме обозначает признак, который не будет учитываться в классификаторе, а единица – наоборот. В (8) первые два условия требуют наличия от r_1 до w_1 признаков из исходной выборки, третье и четвертое условия требуют дополнительного наличия от r_2 до w_2 признаков МГК. Итоговое значение точности классификации определяется как медианное значение результатов точности, полученных на основании серии из 40 запусков третьего эксперимента.

В последних двух экспериментах осуществляется отбор случайным образом. Здесь используется маска, которая заполняется 0 или 1 генератором случайных чисел с вероятностью $p = 0{,}5$. Аналогично ГА, данная операция повторяется 40 раз для всего набора признаков, сравнение происходит по медиане.

3. Результаты экспериментов. В работе используются задачи из репозитория UCI Machine Learning [31]. Их основные характеристики представлены в табл. 1.

Ниже приведены используемые в работе классификаторы, гиперпараметры которых настраиваются в процессе обучения:

– k ближайших соседей (kNN). Количество соседей настраивается в интервале [2, 100];

– метод опорных векторов (SVM). Тип ядра настраивается среди [‘linear’, ’poly’, ’rbf’, ’sigmoid’];

– случайный лес (RFC). Количество деревьев настраивается в интервале [1, 100].

Таблица 1

Основные характеристики данных, выбранных для исследования

	Количество классов	Количество признаков	Объем выборки
Breast Cancer	2	30	569
LSVT Voice Rehabilitation	2	310	126
Australian Credit	2	14	690
Heart Disease	2	13	270

Точность упомянутых классификаторов в среднем является высокой, а настройка их гиперпараметров не требует большого количества времени, что главным образом позволяет сконцентрироваться на задаче поиска информативных признаков. Эти свойства обуславливают их использование в работе.

Лучшим гиперпараметром является тот, для которого значение медианы максимально. При настройке классификаторов используется стратифицированный метод кросс-валидации *k-Fold* с количеством разбиений $k = 5$. В качестве метрики оценки точности выбрана *macro F1-score* [32], которая рассчитывает невзвешенное среднее по каждому классу. Исходные данные предварительно нормируются в интервале [0, 1]. Параметры r_1, r_2 и w_1, w_2, ограничивающие целевую функцию при отборе признаков генетическим алгоритмом, равны $r_{1,2} = 2$ и $w_{1,2} = 4$. При использовании МГК учитываются только первые 4 признака с наибольшими значениями объясняющей дисперсии, которые впоследствии формируют множество $\overline{Attr}^{МГК}$.

Описанные подходы реализованы с помощью языка программирования *Python* версии 3.8.2 и библиотеки *Scikit-learn* [33] версии 0.23.2. Для классификации используются функции *KNeighborsClassifier* с заданными параметрами по умолчанию, *SVC* с заданными параметрами по умолчанию, кроме *max_iter* = 1000000 и *RandomForestClassifier* с параметром *random_state* = 1, а остальными по умолчанию. Для расчета признаков МГК используются функции класса *PCA* в модуле *sklearn.preprocessing*. Количество компонент *n_components*, рассчитываемых функцией, равно количеству признаков исходной выборки. В случае использования данных LSVT Voice Rehabilitation, где количество признаков превышает количество точек, значение *n_components* = 126. Кросс-валидация осуществляется с помощью функции *StratifiedKFold*, а нормирование выборок с помощью функции *MinMaxScaler*.

Ниже описаны параметры и особенности генетического алгоритма, с помощью которых отбирается необходимое количество признаков, увеличивающих точность классификации:

1) инициализация происходит следующим образом. Признаки отбираются равновероятно в два этапа. Напомним, что используемая выборка сконструирована из исходных признаков и МГК. На первом этапе отбирается не больше 4 признаков исходной выборки. На втором, не больше 4 признаков МГК. Это необходимо, чтобы обеспечить сходимость алгоритма. В случае, когда пространство поиска большое, как, например, для выборки LSVT Voice Rehabilitation – 310 признаков, сходимость алгоритма медленная, если использовать стандартную случайную инициализацию;

2) используется турнирная селекция. Размер турнира равен 2;

3) скрещивание – одноточечное;

4) вероятность мутация гена обратно пропорциональна количеству признаков исходной выборки;

5) функция пригодности – значение точности классификации, полученное после стратифицированной кросс-валидации *k-Fold* с адаптивным штрафом, описанным ранее;

6) размер популяции равен 100, количество индивидов в популяции равно 100.

Результаты решения задач представлены в табл. 2, 3, где столбцы отражают выборку и использованный для нее классификатор, а строки – тип эксперимента. В каждой ячейке приведены значения метрики точности классификации *macro F*1-*score*, полученные на тестовой выборке (медианное значение серии из 40 экспериментов). Эксперимент № 4, где признаки для классификации были отобраны генетическим алгоритмом с ограничениями, показал преимущество перед остальными подходами.

Таблица 2

Результаты вычислительных экспериментов (часть 1)

№ эксперимента*	Breast Cancer			LSVT Voice Rehabilitation		
	kNN	SVM	RFC	kNN	SVM	RFC
1	0,965	0,973	0,960	0,795	0,838	0,836
2	0,960	0,963	0,958	0,784	0,785	0,808
3	0,966	0,975	0,957	0,788	0,827	0,834
4	0,975	0,977	0,974	0,888	0,870	0,883
5	0,961	0,967	0,955	0,778	0,827	0,802
6	0,960	0,969	0,958	0,779	0,832	0,811

*Примечание: 1 – все признаки; 2 – МГК; 3 – конструкция признаков МГК с исходными; 4 – признаки, отобранные ГА из исходных + МГК; 5 – отбор случайным образом из исходных; 6 – отбор случайным образом из исходных + МГК.

Таблица 3

Результаты вычислительных экспериментов (часть 2)

№ эксперимента*	Australian Credit			Heart Disease		
	kNN	SVM	RFC	kNN	SVM	RFC
1	0,871	0,854	0,872	0,837	0,845	0,820
2	0,877	0,855	0,840	0,832	0,819	0,815
3	0,870	0,856	0,873	0,833	0,845	0,860
4	0,881	0,873	0,884	0,876	0,868	0,879
5	0,855	0,854	0,803	0,806	0,786	0,775
6	0,859	0,854	0,856	0,814	0,823	0,812

*Примечание: 1 – все признаки; 2 – МГК; 3 – конструкция признаков МГК с исходными; 4 – признаки, отобранные ГА из исходных + МГК; 5 – отбор случайным образом из исходных; 6 – отбор случайным образом из исходных + МГК.

На рис. 1 в виде диаграммы показан прирост точности в процентах для 4 эксперимента по сравнению с 1 для каждой выборки из табл. 1 и рассматриваемых в работе классификаторов.

Исходя из результатов, представленных на рис. 1, увеличение точности классификации зафиксировано для всех выборок. Наибольшее его значение можно отметить для выборки LSVT Voice Rehabilitation, которая в исходном варианте содержит 310 признаков. Важно отметить,

что прирост точности был достигнут со значительно меньшим количеством признаков – 8 (4 из них – это признаки МГК).

На рис. 2 представлена диаграмма размаха результатов точности 40 запусков для экспериментов 4–6.

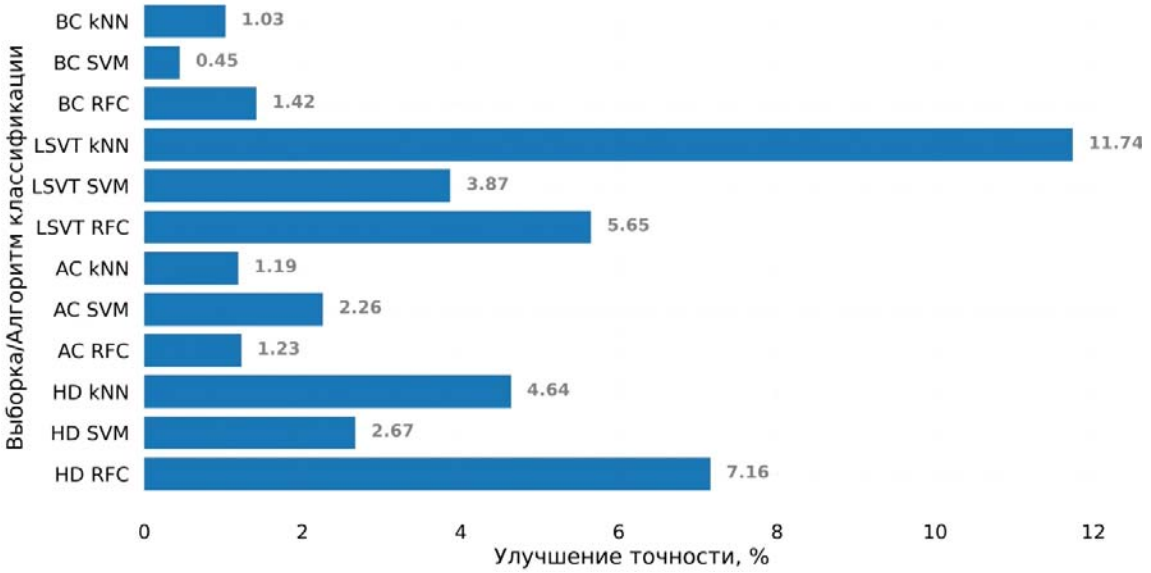

Рис. 1. Прирост точности в процентах для 4 эксперимента по сравнению с 1

Fig. 1. Percentage accuracy increase for the experiment 4 in comparison with the experiment 1

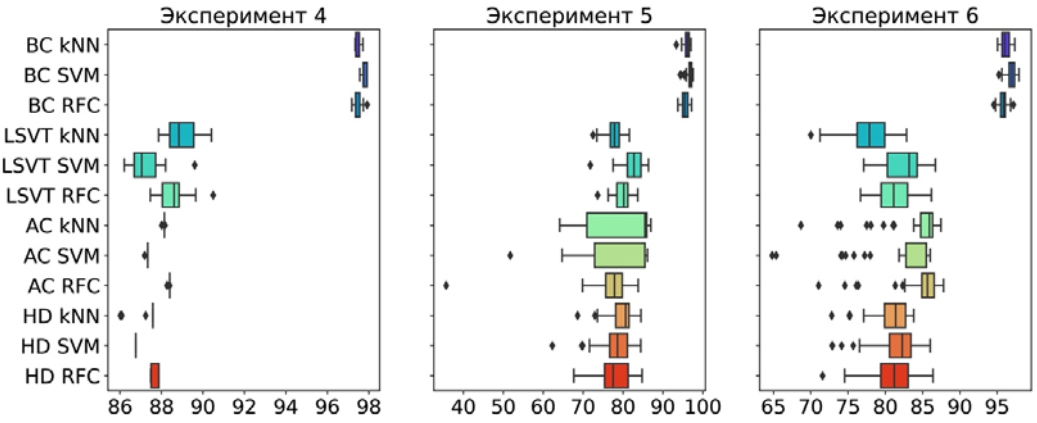

Рис. 2. Диаграмма размаха для экспериментов 4–6

Fig. 2. Box plot for experiments 4–6

На рис. 2 для экспериментов 5, 6 диапазон распределений значений точности выше, чем для эксперимента 4. Отсюда можно заключить, что предлагаемый в данной работе подход имеет устойчивое решение. Помимо этого, на основании одностороннего непараметрического U-критерия Манна – Уитни для эксперимента 4 по отношению к 5 и 6 была выявлена статистическая значимость различий в полученных результатах.

Заключение. В данной работе предложено объединить техники извлечения и отбора признаков, чтобы получить новое представление исходной выборки для увеличения точности классификации. Описанный подход извлечения признаков с помощью МГК с последующим

их добавлением к исходным данным и отбором генетическим алгоритмом с ограничениями показал большую эффективность по сравнению с другими методами проектирования признаков, использованными в работе. Зафиксировано увеличение точности при классификации выборок различного объема.

Помимо этого, была подтверждена статистическая значимость результатов предложенного подхода по сравнению с отбором признаков случайным образом (оценка границы точности снизу). Предложенный подход обладает меньшим разбросом значений метрики *macro F*1-*score* по серии независимых запусков.

Ограничения, накладываемые на функцию пригодности для отбора признаков, могут иметь практическую применимость в тех случаях, когда этого требует программная или аппаратная составляющая реализуемого проекта. Например, при определенных ограничениях канала связи в процессе передачи информации или недостаточном объеме памяти.

В дальнейшем планируется провести исследование других подходов к проектированию признаков. Например, нейронная сеть типа автоэнкодер (*autoencoder*) [34] для извлечения признаков. В отличие от МГК, такая сеть способна оперировать с нелинейными зависимостями, что может способствовать увеличению точности. Другой подход – метод генетического программирования для конструирования признаков, который позволяет не просто создать эффективный (с точки зрения точности) набор признаков, но и «обосновать» полученное решение в виде математической функции, что впоследствии позволяет увеличить не только интерпретируемость решения, но и количество знаний об исходных данных.

Библиографические ссылки

1. Guzella T. S., Caminhas W. M. A review of machine learning approaches to spam filtering // Expert Systems with Applications. 2009. Vol. 36, No. 7. P. 10206–10222.

2. Ballestar M. T., Grau-Carles P., Sainz J. Predicting customer quality in e-commerce social networks: a machine learning approach // Review of Managerial Science. 2019. Vol. 13, No. 3. P. 589–603.

3. Bahlmann C., Haasdonk B., Burkhardt H. Online handwriting recognition with support vector machines-a kernel approach // Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition. 2002. P. 49–54.

4. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective // Artificial Intelligence in medicine. 2001. Vol. 23, No. 1. P. 89–109.

5. Kouziokas G. N. Machine learning technique in time series prediction of gross domestic product // Proceedings of the 21st Pan-Hellenic Conference on Informatics. 2017. P. 1–2.

6. John G. H., Kohavi R., Pfleger K. Irrelevant features and the subset selection problem // Machine Learning Proceedings. 1994. P. 121–129.

7. Kira K., Rendell L. A. A practical approach to feature selection // Machine Learning Proceedings. 1992. P. 249–256.

8. Rendell L., Seshu R. Learning hard concepts through constructive induction: Framework and rationale // Computational Intelligence. 1990. Vol. 6, No. 4. P. 247–270.

9. Liu H., Motoda H. Feature extraction, construction and selection: A data mining perspective. Massachusetts : Kluwer Academic Publishers, 1998. 453 p.

10. Duboue P. The Art of Feature Engineering: Essentials for Machine Learning. Cambridge : Cambridge University Press. 2020. 270 p.

11. Zheng A., Casari A. Feature engineering for machine learning: principles and techniques for data scientists. Sebastopol : O'Reilly Media Inc., 2018. 193 p.

12. Feature selection: A data perspective / Li J., Cheng K., Morstatter F. et al. // ACM Computing Surveys (CSUR). 2017. Vol. 50, No. 6. P. 1–45.

13. Park M. S., Na J. H., Choi J. Y. PCA-based feature extraction using class information // 2005 IEEE International Conference on Systems, Man and Cybernetics. 2005. Vol. 1. P. 341–345.

14. Abdi H., Williams L. J. Principal component analysis // Wiley interdisciplinary reviews: computational statistics. 2010. Vol. 2, No. 4. P. 433–459.

15. Markovitch S., Rosenstein D. Feature generation using general constructor functions // Machine Learning. 2002. Vol. 49, No. 1. P. 59–98.

16. Hirsh H., Japkowicz N. Bootstrapping training-data representations for inductive learning: A case study in molecular biology // AAAI-94 Proceedings. 1994. P. 639–644.

17. Sutton R. S., Matheus C. J. Learning polynomial functions by feature construction // Machine Learning Proceedings. 1991. P. 208–212.

18. Zhao H., Sinha A. P., Ge W. Effects of feature construction on classification performance: An empirical study in bank failure prediction // Expert Systems with Applications. 2009. Vol. 36, No. 2. P. 2633–2644.

19. Pagallo G. Haussler D. Boolean feature discovery in empirical learning // Machine learning. 1990. Vol. 5, No. 1. P. 71–99.

20. Matheus C. J., Rendell L. A. Constructive Induction on Decision Trees // IJCAI'89: Proceedings of the 11th international joint conference on Artificial intelligence. 1989. Vol. 89. P. 645–650.

21. Krawiec K. Genetic programming-based construction of features for machine learning and knowledge discovery tasks // Genetic Programming and Evolvable Machines. 2002. Vol. 3, No. 4. P. 329–343.

22. Smith M. G., Bull L. Genetic programming with a genetic algorithm for feature construction and selection // Genetic Programming and Evolvable Machines. 2005. Vol. 6, No. 3. P. 265–281.

23. An investigation into feature construction to assist word sense disambiguation / Specia L., Srinivasan A., Sachindra J. et al. // Machine Learning. 2009. Vol. 76, No. 1. P. 109–136.

24. Khalid S., Khalil T., Nasreen S. A survey of feature selection and feature extraction techniques in machine learning // 2014 Science and Information Conference. 2014. P. 372–378.

25. Кривенко М. П. Критерии значимости отбора признаков классификации // Информатика и её применения. 2016. Т. 10, №. 3. С. 32–40.

26. Miao J., Niu L. A survey on feature selection // Procedia Computer Science. 2016. Vol. 91. P. 919–926.

27. Chandrashekar G., Sahin F. A survey on feature selection methods // Computers & Electrical Engineering. 2014. Vol. 40, No. 1. P. 16–28.

28. A survey on evolutionary computation approaches to feature selection / Xue B., Zhang M., Browne W. et al. // IEEE Transactions on Evolutionary Computation. 2015. Vol. 20, No. 4. P. 606–626.

29. Coello C. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art // Computer methods in applied mechanics and engineering. 2002. Vol. 191, No. 11–12. P. 1245–1287.

30. Barbosa H. J. C., Lemonge A. C. C. An adaptive penalty method for genetic algorithms in constrained optimization problems // Frontiers in Evolutionary Robotics, 2008.

31. UCI Machine Learning Repository [Электронный ресурс]. URL: https://archive.ics.uci.edu/ml/index.php (дата обращения: 09.01.2021).

32. Opitz J., Burst S. Macro f1 and macro f1. Препринт: arXiv:1911.03347. [Электронный ресурс]. URL: https://arxiv.org/abs/1911.03347 (дата обращения: 25.02.2021).

33. Scikit-learn: Machine learning in Python / Pedregosa F., Varoquaux G., Gramfort A. et al. // Journal of machine Learning research. 2011. Vol. 12. P. 2825–2830.

34. Dong G, Liao G., Liu H, Kuang G. A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images // IEEE Geoscience and Remote Sensing Magazine. 2018. Vol. 6, No. 3. P. 44–68.

References

1. Guzella T. S., Caminhas W. M. A review of machine learning approaches to spam filtering. *Expert Systems with Applications*. 2009, Vol. 36, No. 7, P. 10206–10222. Doi: 10.1016/j.eswa.2009.02.037.

2. Ballestar M. T., Grau-Carles P., Sainz J. Predicting customer quality in e-commerce social networks: a machine learning approach. *Review of Managerial Science*. 2019, Vol. 13, No. 3, P. 589–603. Doi: 10.1007/s11846-018-0316-x.

3. Bahlmann C., Haasdonk B., Burkhardt H. Online handwriting recognition with support vector machines-a kernel approach. *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*. 2002, P. 49–54. Doi: 10.1109/IWFHR.2002.1030883.

4. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*. 2001. Vol. 23, No 1, P. 89–109. Doi: 10.1016/S0933-3657(01)00077-X.

5. Kouziokas G. N. Machine learning technique in time series prediction of gross domestic product. *Proceedings of the 21st Pan-Hellenic Conference on Informatics*. 2017, P. 1–2. Doi: 10.1145/3139367.3139443.

6. John G. H., Kohavi R., Pfleger K. Irrelevant features and the subset selection problem. *Machine Learning Proceedings*. 1994, P. 121–129. Doi: 10.1016/B978-1-55860-335-6.50023-4.

7. Kira K., Rendell L. A. A practical approach to feature selection. *Machine Learning Proceedings*. 1992, P. 249–256. Doi: 10.1016/B978-1-55860-247-2.50037-1.

8. Rendell L., Seshu R. Learning hard concepts through constructive induction: Framework and rationale. *Computational Intelligence*. 1990, Vol. 6, No. 4, P. 247–270. Doi: 10.1111/j.1467-8640.1990.tb00298.x.

9. Liu H., Motoda H. *Feature extraction, construction and selection: A data mining perspective.* Massachusetts : Kluwer Academic Publishers, 1998, 453 p.

10. Duboue P. *The Art of Feature Engineering: Essentials for Machine Learning.* Cambridge : Cambridge University Press, 2020, 270 p. Doi: 10.1017/9781108671682.

11. Zheng A., Casari A. *Feature engineering for machine learning: principles and techniques for data scientists.* Sebastopol : O'Reilly Media Inc., 2018, 193 p.

12. Li J., Cheng K., Morstatter F. et al. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*. 2017, Vol. 50, No. 6, P. 1–45. Doi: 10.1145/3136625.

13. Park M. S., Na J. H., Choi J. Y. PCA-based feature extraction using class information. *2005 IEEE International Conference on Systems, Man and Cybernetics*. 2005, Vol. 1, P. 341–345. Doi: 10.1109/ICSMC.2005.1571169.

14. Abdi H., Williams L. J. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*. 2010, Vol. 2, No. 4, P. 433–459. Doi: 10.1002/wics.101.

15. Markovitch S., Rosenstein D. Feature generation using general constructor functions. *Machine Learning*. 2002, Vol. 49, No. 1, P. 59–98. Doi: 10.1023/A:1014046307775.

16. Hirsh H., Japkowicz N. Bootstrapping training-data representations for inductive learning: A case study in molecular biology. *AAAI-94 Proceedings*, 1994, P. 639–644.

17. Sutton R. S., Matheus C. J. Learning polynomial functions by feature construction. *Machine Learning Proceedings*. 1991, P 208–212. Doi: 10.1016/B978-1-55860-200-7.50045-3.

18. Zhao H., Sinha A. P., Ge W. Effects of feature construction on classification performance: An empirical study in bank failure prediction. *Expert Systems with Applications*. 2009, Vol. 36, No. 2, P. 2633–2644. Doi: 10.1016/j.eswa.2008.01.053.

19. Pagallo G. Haussler D. Boolean feature discovery in empirical learning. *Machine learning*. 1990, Vol. 5, No 1, P. 71–99. Doi: 10.1023/A:1022611825350.

20. Matheus C. J., Rendell L. A. Constructive Induction on Decision Trees. *IJCAI'89: Proceedings of the 11th international joint conference on Artificial intelligence*. 1989, Vol. 89, P. 645–650.

21. Krawiec K. Genetic programming-based construction of features for machine learning and knowledge discovery tasks. *Genetic Programming and Evolvable Machines*. 2002, Vol. 3, No. 4, P. 329–343. Doi: 10.1023/A:1020984725014.

22. Smith M. G., Bull L. Genetic programming with a genetic algorithm for feature construction and selection. *Genetic Programming and Evolvable Machines*. 2005, Vol. 6, No. 3, P. 265–281. Doi: 10.1007/s10710-005-2988-7.

23. Specia L., Srinivasan A., Sachindra J., et al. An investigation into feature construction to assist word sense disambiguation. *Machine Learning*. 2009, Vol. 76, No 1, P. 109–136. Doi: 10.1007/s10994-009-5114-x.

24. Khalid S., Khalil T., Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. *2014 Science and Information Conference*. 2014, P. 372–378. Doi: 10.1109/SAI.2014.6918213.

25. Krivenko M. P. [Significance tests of feature selection for classification]. *Informatics and Applications*. 2016, Vol. 10, No. 3, P. 32–40. Doi: 10.14357/19922264160305. (In Russ.)

26. Miao J., Niu L. A survey on feature selection. *Procedia Computer Science*. 2016, Vol. 91, P. 919–926. Doi: 10.1016/j.procs.2016.07.111.

27. Chandrashekar G., Sahin F. A survey on feature selection methods. *Computers & Electrical Engineering*. 2014, Vol. 40, No. 1, P. 16–28. Doi: 10.1016/j.compeleceng.2013.11.024.

28. Xue B., Zhang M., Browne W. et al. A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation*. 2015, Vol. 20, No. 4, P. 606–626. Doi: 10.1109/TEVC.2015.2504420.

29. Coello C. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art. *Computer methods in applied mechanics and engineering*. 2002, Vol. 191, No. 11–12, P. 1245–1287. Doi: 10.1016/S0045-7825(01)00323-1.

30. Barbosa H. J. C., Lemonge A. C. C. An adaptive penalty method for genetic algorithms in constrained optimization problems. *Frontiers in Evolutionary Robotics*. 2008. Doi: 10.5772/5446.

31. UCI Machine Learning Repository Available at: https://archive.ics.uci.edu/ml/index.php (accessed 09.01.2021).

32. Opitz J., Burst S. Macro f1 and macro f1. Preprint arXiv:1911.03347. Available at: https://arxiv.org/abs/1911.03347 (accessed 25.02.2021).

33. Pedregosa F., Varoquaux G., Gramfort A. et al. Scikit-learn: Machine learning in Python. *Journal of machine Learning research*. 2011, Vol. 12, P. 2825–2830.

34. Dong G., Liao G., Liu H, Kuang G. A review of the autoencoder and its variants: A comparative perspective from target recognition in synthetic-aperture radar images. *IEEE Geoscience and Remote Sensing Magazine*. 2018. Vol. 6, No. 3, P. 44–68. Doi: 10.1109/MGRS.2018.2853555.

Денисов Максим Андреевич – аспирант; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: max_denisov00@mail.ru.

Сопов Евгений Александрович – кандидат технических наук, доцент; Сибирский государственный университет науки и технологий имени академика М. Ф. Решетнева. E-mail: evgenysopov@gmail.com.

Denisov Maksim Andreevich – postgraduate; Reshetnev Siberian State University of Science and Technology. E-mail: max_denisov00@mail.ru.

Sopov Evgenii Aleksandrovich – PhD (CS), associate professor; Reshetnev Siberian State University of Science and Technology. E-mail: evgenysopov@gmail.com.