ТЕХНОЛОГИЯ ВЫБОРА ИНФОРМАТИВНЫХ ПРИЗНАКОВ ДЛЯ АНАЛИЗА ДАННЫХ ИММУНОСИГНАТУР

DOI: 10.17691/stm2020.12.5.02
УДК 616–007:004.023:519.237
Поступила 10.02.2020 г.

© **А.А. Кошечкин,** ассистент кафедры теоретических основ информатики[1];
О.В. Романович, доцент кафедры теоретических основ информатики[1];
ведущий инженер Института прикладной математики и компьютерных наук[1];
D. Stamate, DSc, Senior Lecturer[2];
S.A. Johnston, DSc, Center Director and Professor[3];
А.В. Замятин, д.техн.н., зав. кафедрой теоретических основ информатики[1];
директор Института прикладной математики и компьютерных наук[1]

[1]Национальный исследовательский Томский государственный университет, проспект Ленина, 36, Томск, 634050;
[2]Data Science Department of Computing Goldsmiths, University of London, New Cross, London, SE14 6NW, UK;
[3]Biodesign Center for Innovations in Medicine, Arizona State University, Tempe, AZ 85281, USA

Основной сложностью практической работы с данными, полученными посредством иммуносигнатурного анализа, является высокая размерность и наличие значительного числа неинформативных или ложно-информативных признаков в связи с особенностью технологии. Для обеспечения практически значимого качества анализа и классификации по данным необходимо учитывать эти особенности.

Цель исследования — разработка и апробация технологии эффективного снижения размерности данных иммуносигнатурного анализа, которая, учитывая особенности получаемых данных, обеспечивает высокое, практически значимое качество классификации.

Материалы и методы. В исследовании использовались два нормализованных набора данных из публичного хранилища биомедицинских данных, содержащих результаты иммуносигнатурного анализа.

В рамках исследования предложена технология отбора информативных признаков, состоящая из трех последовательных шагов: 1) разбиение многоклассовой задачи на ряд бинарных задач с использованием стратегии «один против всех»; 2) для каждого бинарного сравнения отсев ложно-информативных признаков с помощью сопоставления значений медианы множеств «один» и «все»; 3) ранжирование оставшихся признаков по информативности и отбор лучших из них для каждого бинарного сравнения.

Для оценки качества предложенной технологии отбора информативных признаков используются результаты классификации по отфильтрованным данным после ее применения. В качестве модели классификации используется метод опорных векторов, положительно зарекомендовавший себя в задачах классификации данных высокой размерности.

Результаты. Оценена эффективность предложенной технологии отбора информативных признаков. Данная технология позволяет обеспечить высокое качество классификации при значительном сокращении признакового пространства. Количество признаков, отсеянных на втором шаге, составляет примерно 50% для каждого из рассмотренных наборов данных, что в значительной степени упрощает последующий анализ данных. После третьего шага для набора данных GSE52580 при сокращении признакового пространства до 15 признаков оценка качества классификации по метрике macro-average F1-score составляет 98,9%. Для набора данных GSE52581 при сокращении признакового пространства до 266 признаков качество классификации по метрике macro-average F1-score составляет 91,3%.

Заключение. Результаты работы демонстрируют перспективность предложенной технологии отбора информативных признаков применительно к данным иммуносигнатурного анализа.

Ключевые слова: ранняя диагностика заболеваний; иммуносигнатура; отбор информативных признаков в выборке; машинное обучение.

Как цитировать: Koshechkin A.A., Romanovich O.V., Stamate D., Johnston S.A., Zamyatin A.V. Technology of informative feature selection for immunosignature analysis. Sovremennye tehnologii v medicine 2020; 12(5): 19–27, https://doi.org/10.17691/stm2020.12.5.02

Для контактов: Кошечкин Александр Алексеевич, e-mail: kaa1994g@mail.ru

English

Technology of Informative Feature Selection for Immunosignature Analysis

A.A. Koshechkin, Assistant, Department of Theoretical Foundations of Informatics[1];
O.V. Romanovich, Associate Professor, Department of Theoretical Foundations of Informatics[1];
Leading Engineer, Institute of Applied Mathematics and Computer Science[1];
D. Stamate, DSc, Senior Lecturer[2];
S.A. Johnston, DSc, Center Director and Professor[3];
A.V. Zamyatin, DSc, Head of the Department of Theoretical Foundations of Informatics[1];
Director of the Institute of Applied Mathematics and Computer Science[1]

[1]National Research Tomsk State University, 36 Lenin Avenue, Tomsk, 634050, Russia;
[2]Data Science Department of Computing, Goldsmiths, University of London, New Cross, London, SE14 6NW, UK;
[3]Biodesign Center for Innovations in Medicine, Arizona State University, Tempe, AZ 85281, USA

The main difficulty in practical work with data obtained via immunosignature analysis is high dimensionality and the presence of a significant number of uninformative or false-informative features due to the specific character of the technology. To ensure practically relevant quality of data analysis and classification, it is necessary to take due account of this specific character.

The aim of the study is to create and test the technology for effective reduction of immunosignature data dimensionality, which provides practically relevant and high quality of classification with due regard for the properties of the data obtained.

Materials and Methods. The study involved the use of two normalized data sets obtained from the public biomedical repository and containing the results of immunosignature analysis.

The technology for selecting informative features was proposed within the framework of the study. It consisted of three successive steps: 1) breaking a multiclass task into a series of binary tasks using the “one vs all” strategy; 2) screening of false-informative features is performed for each binary comparison by comparing the values of the median of the sets “one” and “all”; 3) ranking of the remaining features according to their informative value and selection of the most informative ones for each binary comparison.

To assess the quality of the proposed technology for informative feature selection, we used the results obtained after application of classification based on the filtered data. Support vector method that proved itself in the problems of high-dimensional data classification was used as a classification model.

Results. Effectiveness of the proposed technology for informative feature selection was determined. This technology allows us to provide high quality of classification while significantly reducing the feature space. The number of features eliminated in the second step is approximately 50% for each data set under consideration, which greatly simplifies subsequent data analysis. After the third step, when the feature space is reduced to 15 features, the quality of classification by the macro-average F1-score metric is assessed as 98.9% for the GSE52580 dataset. For the GSE52581 dataset, with the feature space reduced to 266 features, the quality of classification by the macro-average F1-score metric is 91.3%.

Conclusion. The results of the work demonstrate the promising outlook of the proposed technology for informative feature selection as applied to the data of immunosignature analysis.

Key words: early diagnosis of diseases; immunosignature; feature selection in the sample; machine learning.

Введение

В 2018 г. в России пациентов, у которых впервые в жизни диагностировали онкологическое заболевание, насчитывалось 624 тыс. человек: I стадии — 30,6%, II — 25,8%, III — 18,2%, IV — 20,3%. Смертность в России от онкологических заболеваний за 2018 г. составила более 293 тыс. человек. В то же время за последние 5 лет не наблюдается статистически значимого изменения абсолютного числа умерших от злокачественных новообразований [1].

Эффективность лечения онкологических заболеваний напрямую зависит от своевременной диагностики. Для раннего их выявления нужны эффективные и специфичные, удобные в применении, приемлемые для пациентов и недорогие методы диагностики [2]. Одним из перспективных является технология иммуносигнатурного (immunosignature) анализа, основанная на идее профилирования антител человека [3]. В основе данной технологии — микрочип, представляющий собой набор пептидов со случайными аминокислотными последовательностями, которые при взаимодействии с сывороткой крови человека дают карту иммунной активности. Пептидные микрочипы разнообразны и содержат от 10 тыс. до 330 тыс. пептидов.

На данный момент для анализа и интерпретации данных, полученных посредством технологии иммуносигнатурного анализа, активно исследует-

ся применимость различных методов интеллектуального анализа и классификации данных. Для построения эффективных моделей классификации исследователи нуждаются в релевантных и высококачественных данных. В связи с тем, что признаковое пространство основано на случайно созданных пептидах, не все признаки будут информативными, поэтому их отбор является одним из важнейших этапов анализа данных. При отбрасывании бесполезных и избыточных признаков не только улучшается производительность модели, но и облегчается ее интерпретация [4]. В связи с этим в каждой статье по исследованию данных иммуносигнатурного анализа так или иначе присутствует этап отбора информативных признаков.

В работе [5] рассматривается применимость иммуносигнатурного анализа для выявления четырех различных заболеваний поджелудочной железы (рак и предраковое состояние, диабет 2-го типа и панкреатит). На ранней стадии эти заболевания имеют одинаковые симптомы, что усложняет их диагностику. Авторы применяли t-критерий Стьюдента с целью отбора лучших признаков для дальнейшего анализа. Средняя точность классификации достигала 92%. В то же время было выявлено, что каждая болезнь обладает уникальными иммунологическими характеристиками.

Авторы исследования [6] демонстрируют, что технология иммуносигнатурного анализа потенциально может соответствовать требованию универсального теста для диагностики онкологических заболеваний. Проведен интеллектуальный анализ двух наборов данных 6-го и 15-го классов. В результате экспериментально показано, что с помощью иммуносигнатурного анализа возможно разделить различные типы заболеваний с высокой точностью. Для отбора информативных признаков использовали U-test.

В работе [7] рассматривались возможности применения технологии иммуносигнатурного анализа на примере микрочипа с 330 тыс. пептидов для диагностики рака молочной железы. Основная идея исследования заключалась в использовании метода проекции на латентные структуры для выявления эффективной размерности данных. Это должно уменьшить негативный эффект от переобучения модели и улучшить качество распознавания объектов. Данный подход противоречит основной идее иммуносигнатурного анализа, направленного на поиск возможных антигенов для различных заболеваний. После применения метода проекции на латентные структуры происходит трансформация исходного признакового пространства в новое пространство латентных структур. В связи с этим становится невозможной какая-либо интерпретация признакового пространства в контексте взаимодействия «антиген–антитело». В качестве противовеса проекции на латентные структуры рассматривается применение U-test для отбора лучших признаков, что повторяет исследователей из предыдущей статьи.

Применение статистических критериев для отбора информативных признаков является примером использования методов фильтрации. Для этих методов характерны такие проблемы, как неочевидность выбора порога для отсечения неинформативных признаков и сохранение избыточности признакового пространства. Анализ данных с избыточным числом признаков в общем случае требует большой памяти и вычислительной мощности, а также способен вызвать такой нежелательный эффект, как переобучение модели классификации [8]. В то же время никак не учитывается природа возникновения данных, что может приводить к неочевидным ошибкам.

Цель исследования — разработка и апробация технологии эффективного снижения размерности данных иммуносигнатурного анализа, которая обеспечивает высокое, практически значимое качество классификации.

Материалы и методы

В исследовании использовались два нормализованных набора данных из публичного хранилища биомедицинских данных GSE52580 и GSE52581, содержащих результаты иммуносигнатурного анализа пациентов с различными онкологическими и инфекционными заболеваниями, а также контрольной группы здоровых лиц [9, 10]. Предварительно наборы данных подверглись операции транспонирования, чтобы соответствовать формату tidy data [11]. Итоговые материалы представляют собой набор данных (таблицу) значений интенсивности флуоресценции пептидов, где названия пептидов являются столбцами (признаками), а метки классов — строками (экземплярами).

Набор данных GSE52580 имеет следующие характеристики:

количество экземпляров — 240;

количество признаков — 9787;

количество классов — 6.

Количество экземпляров каждого класса в наборе данных GSE52580 одинаково.

Набор данных GSE52581 имеет следующие характеристики:

количество экземпляров — 1516;

количество признаков — 10 372;

количество классов — 15.

Количество экземпляров каждого класса в этом наборе различно (табл. 1).

Описание технологии. В рамках исследования предложена технология отбора информативных признаков, состоящая из трех последовательных шагов:

1) разбиение многоклассовой задачи на ряд бинарных задач с использованием стратегии «один против всех»;

2) выполнение для каждого бинарного сравнения отсева ложно-информативных признаков с помощью сравнения значений медианы множеств «один» и «все»;

Таблица 1

Описание набора данных GSE52581

Класс	Количество экземпляров
Здоровые	249
Астроцитома	166
Кокцидиоидомикоз	142
Рак молочной железы	141
Рак поджелудочной железы	136
Множественная миелома	112
Рак легкого	107
Смешанная олигоастроцитома	97
Карцинома яичника	86
Панкреатит	82
Повторный рак молочной железы	61
Олигодендроглиома	48
Рак молочной железы IV стадии	42
Мультиформная глиобластома	27
Саркома Юинга	20

3) ранжирование оставшихся признаков по информативности и отбор лучших из них для каждого бинарного сравнения.

Первый этап технологии: применение стратегии «один против всех». В контексте отбора информативных признаков многоклассовые задачи раскладываются на несколько бинарных задач с помощью стратегии «один против всех» или «один против одного» с последующим отбором лучших признаков по каждому бинарному сравнению [12]. В данном исследовании будет использоваться стратегия «один против всех». Данный выбор объясняется особенностью технологии иммуносигнатурного анализа, заключающейся в имитации антигенов заболеваний с помощью пептидов случайной аминокислотной последовательности. Цель анализа на данном этапе — найти пептид, который выполняет роль антигена определенного заболевания, т.е. антитела, выработанные против искомого заболевания, связываются с данным пептидом, в то время как для пациентов из групп с другим диагнозом этого не происходит.

Второй этап технологии: медианный фильтр. Методы фильтрации — это методы ранжирования признаков, которые оценивают их релевантность, рассматривая внутренние свойства данных [13]. После ранжирования происходит удаление признаков, чьи оценки информативности находятся ниже порогового значения. Результирующее подмножество признаков используется для дальнейшего анализа данных (например, классификации посредством методов машинного обучения). Методы фильтрации легко масштабируются для высокоразмерных данных и не обладают высокой вычислительной сложностью, но большинство из них применимы только к бинарным задачам и рассматривают каждый признак отдельно, игнорируя при этом зависимости признаков, что может привести к ухудшению последующего анализа данных.

Первый фильтр представляет собой сравнение значений медианы по каждому признаку для множества «один» и множества «все». Смысл данного фильтра заключается в удалении из набора данных всех признаков, у которых значение медианы интенсивности флуоресценции множества «один» меньше медианы интенсивности флуоресценции множества «все», так как они признаны неинформативными. Причина этой неинформативности проиллюстрирована на примере двух признаков из набора данных GSE52580 (рис. 1).

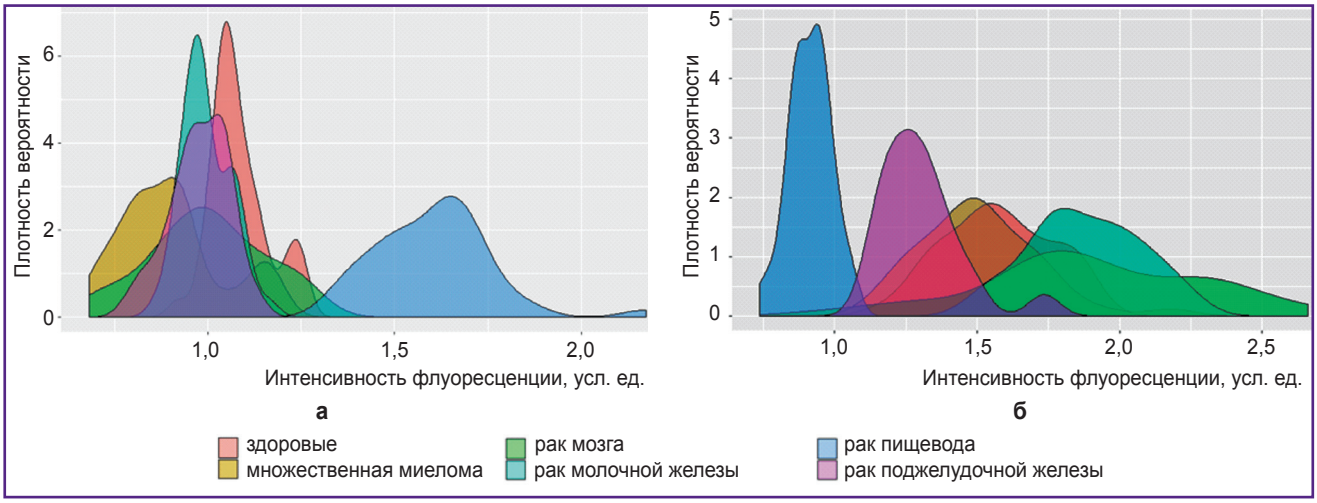

Рис. 1. Сравнение информативного (пептид CSGTMNSEFQNTTRHVYIMS) (*а*) и ложно-информативного (пептид CSGVFMLSHHQFHPSWYQPN) (*б*) признаков для класса «рак пищевода»

Если рассматривать два этих признака в отрыве от предметной области, то они оба будут информативными для отделения класса «рак пищевода» от всех остальных. Однако в случае, когда распределение интенсивности флуоресценции класса «рак пищевода» находится правее всех остальных классов, это означает, что антитела из крови пациентов с диагнозом «рак пищевода» соединялись с пептидом в большем количестве, чем в остальных представленных классах (рис. 1, *а*). Случай же, когда распределение интенсивности флуоресценции находится левее всех остальных классов, будет означать, что к данному пептиду присоединялись антитела всех классов, включая «здоровые», исключая класс «рак пищевода» (рис. 1, *б*). Это значит, что зафиксированы антитела, выработанные на какое-либо другое заболевание, которое как-то объединяет людей этих групп, и оно никак не связано с диагнозом «рак пищевода». Во всех рассмотренных нами более ранних исследованиях по иммуносигнатурному анализу данный аспект никак не освещался и оба варианта признаков использовались в качестве информативных. В свою очередь, для объектов класса «здоровые» вообще не существует информативных признаков, фактически это исключения, не попавшие ни в один другой класс.

Третий этап технологии: ранжирование и отбор признаков. В качестве второго фильтра для оценки признаков используется критерий symmetric uncertainty (SU). Этот критерий оценки корреляции между и признаками и целевой переменной является улучшенной версией критерия information gain [14]. Значения SU находятся в диапазоне [0; 1], где 0 означает полное отсутствие корреляции и, как следствие, — нерелевантность признака.

$$\mathrm{SU}(X;Y)=\frac{2\cdot IG(X;Y)}{H(X)+H(Y)},$$

где $IG(X;Y)=H(X)-X(X|Y)$ — information gain для признака и метки класса Y; $H(X)$ — энтропия признака X; $H(Y)$ — энтропия признака Y.

Следующим шагом является выбор подмножества информативных признаков. С этой целью для каждого бинарного сравнения выбираются лучшие признаки на основе оценок метода SU.

Оценка эффективности отбора информативных признаков. Основная задача иммуносигнатурного анализа — диагностика заболеваний, что в терминах анализа данных является задачей классификации. В связи с этим для оценки качества предложенной технологии отбора информативных признаков используются результаты классификации по полученным данным после ее применения.

Существует множество различных методов классификации, которые могут использоваться для решения данной задачи. Рассмотрим метод машинного обучения, который ранее уже показал высокую эффективность в предыдущих исследованиях [15], — метод опорных векторов (support vector machine, SVM), основанный на построении гиперплоскости, максимально разделяющей классы между собой [16]. В зависимости от настроек ядра модели возможно построение разделяющих поверхностей различного рода. Не существует общего подхода к автоматическому выбору ядра, в связи с чем в данном исследовании оценивается эффективность каждого из них.

Также в линейных моделях (например, SVM) нужно стандартизировать признаки. Это необходимо сделать в связи со следующим обстоятельством. Одно из наиболее важных допущений при работе с линейными моделями, параметры которых оцениваются методом наименьших квадратов, состоит в том, что остатки модели независимы (т.е. не коррелируют) и имеют нормальное распределение со средним значением 0 и некоторым фиксированным стандартным отклонением (например, 1). В связи с этим в данном исследовании проведена стандартизация признаков по следующей формуле:

$$Z=\frac{X_i-\bar{X}}{\sigma},$$

где X_i — индивидуальное значение по признаку; $\bar{X}$ — среднее значение по признаку, σ — стандартное отклонение по признаку.

Известно множество различных метрик оценки качества классификации, которые подходят для представленной задачи. В данном исследовании будут использоваться precision, recall, balanced accuracy и F1-score [17], это обусловлено целями анализа: необходимо максимально эффективно отделить группу людей с определенным заболеванием от всех остальных, а представленные метрики как раз позволяют это измерить. Они будут вычисляться для каждого бинарного сравнения «один против всех», где «один» — позитивный класс, а «все» — негативный класс.

Экспериментальные исследования. В связи с небольшим количеством образцов в наборе данных необходимо провести анализ с применением перекрестной проверки. Таким образом, исходный набор данных разбивается на 5 приблизительно равных частей с соблюдением пропорции классов. На каждой итерации 4 части образуют тренировочную выборку (train), а 5-я — тестовую выборку с последующей сменой. На каждой итерации производятся поиск информативного множества признаков на основе train и его оценка с помощью тестовой выборки. Далее приводится описание одной итерации перекрестной проверки.

1. Оценка информативности признаков в train с помощью метода SU и применения стратегии «один против всех».

2. Обращение в 0 оценок ложно-информативных признаков по результатам фильтрации на основе сравнения медиан.

3. Отбор *N* лучших признаков для каждого бинарного сравнения, пока не будут достигнуты приемлемые результаты классификации.

4. На основе полученного подмножества признаков проведение обучения SVM на train и оценка качества классификации на тестовой выборке.

5. Оценка полученных результатов.

Вся работа выполнена с использованием языка программирования R и библиотек из публичного репозитория CRAN и Bioconductor.

Результаты

В первую очередь рассмотрим результаты экспериментов по оценке эффективности различных ядер SVM и отбору информативных признаков предложенной технологией для набора данных GSE52580. На рис. 2 представлены результаты классификации для каждого ядра SVM по метрике macro-average F1-score.

Из информации, представленной на рис. 2, видно, что наилучшие результаты показывают линейное, сигмоидальное и радиально-базисное ядро, однако линейное ядро многократно превосходит по производительности аналоги. В связи с этим в дальнейшем будем использовать только линейное ядро.

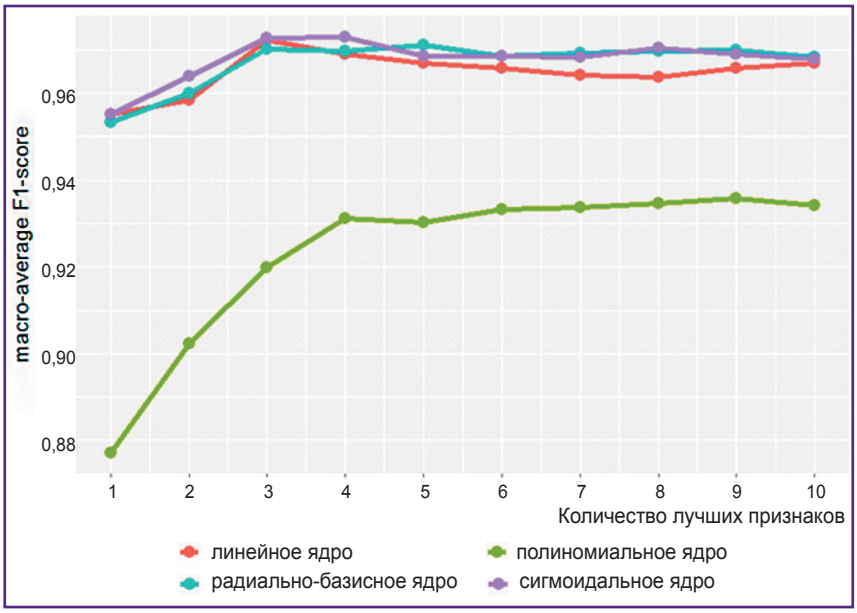

Рис. 2. Результаты классификации для различных ядер SVM в зависимости от количества лучших признаков (набор GSE52580)

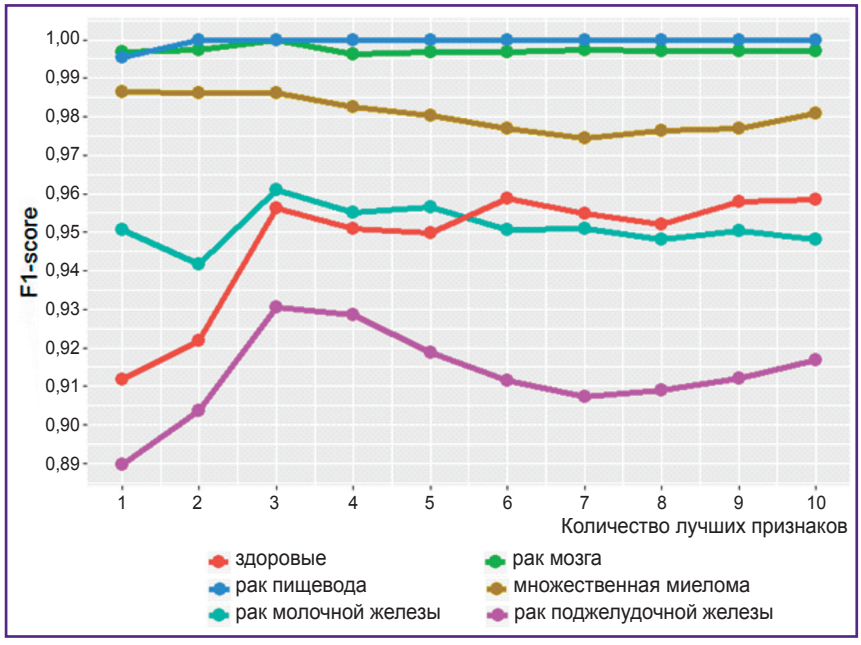

Рис. 3. Результаты классификации для каждого класса в зависимости от количества лучших признаков (набор GSE52580)

На рисунке отчетливо видно, что наилучшие результаты достигаются при выборе трех лучших признаков для каждого класса, кроме «здоровых» (для объектов класса «здоровые» не существует информативных признаков, фактически это исключения, не попавшие ни в один другой класс), что в сумме означает всего 15 признаков для 6 классов.

В табл. 2 представлены результаты классификации на тестовой выборке с использованием только трех лучших признаков для каждого заболевания. Значения в таблице — это различные метрики оценки качества, усредненные по перекрестной проверке. Представленные в таблице оценки качества классификации свидетельствуют о разделимости классов, несмотря на значительное сокращение признакового пространства.

Теперь рассмотрим результаты экспериментов по оценке эффективности различных ядер SVM и отбору информативных признаков предложенной технологией для набора данных GSE52581. На рис. 4 представлены результаты классификации для каждого ядра SVM по метрике macro-average F1-score.

На рисунке отчетливо видно преимущество линейного ядра над другими, в связи с этим в дальней-

шем анализе данного набора используется только оно.

На рис. 5 представлены результаты классификации по метрике F1-score в зависимости от количества отобранных лучших признаков для каждого класса.

В табл. 3 представлены результаты классификации на тестовой выборке с использованием только 19 лучших признаков для каждого класса, кроме «здоровых» (для объектов класса «здоровые» не существует информативных признаков, фактически это исключения, не попавшие ни в один другой класс). Значения в таблице — это различные метрики оценки качества, усредненные по перекрестной проверке. Отчетливо видно, что для подавляющего большинства классов достигнуто высокое качество классификации, несмотря на значительное сокращение признакового пространства.

Полученные результаты классификации соответствуют предшествующим исследованиям в данной области, однако ключевым аспектом работы является отсутствие ложно-информативных признаков в итоговом признаковом пространстве, что ранее не учитывалось. Это положительно повлияет на последующий анализ и поиск антигенов для различных заболеваний.

Таблица 2

Результаты классификации на тестовой выборке по трем лучшим признакам для каждого заболевания

Класс	Метрики			
	precision	recall	F1-score	balanced accuracy
Рак мозга	1	1	1	1
Рак молочной железы	0,976	1	0,988	0,997
Рак пищевода	1	1	1	1
Здоровые	0,969	0,993	0,981	0,993
Множественная миелома	1	0,975	0,987	0,987
Рак поджелудочной железы	0,968	0,944	0,956	0,969

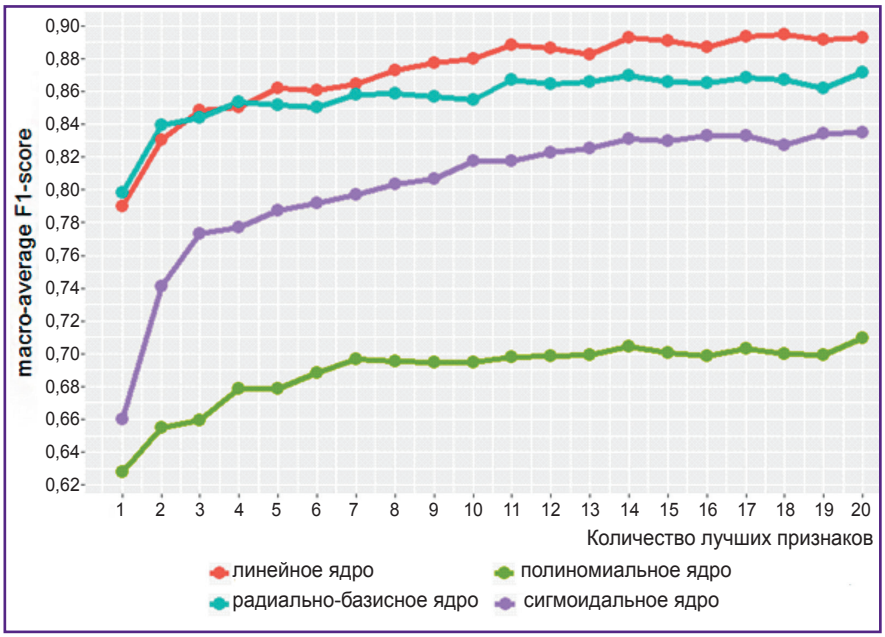

Рис. 4. Результаты классификации для различных ядер SVM в зависимости от количества лучших признаков (набор GSE52581)

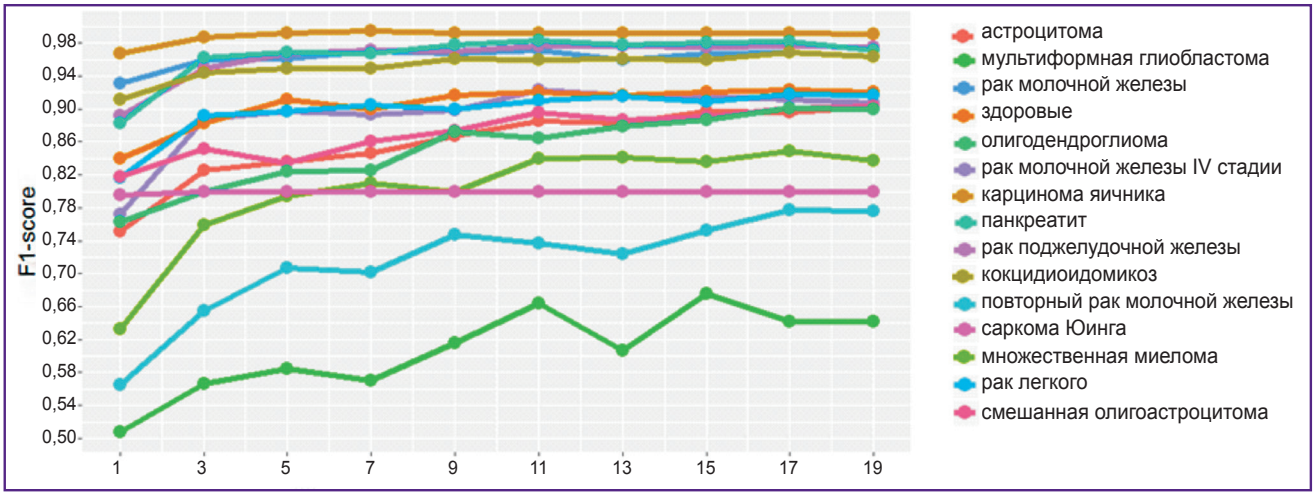

Рис. 5. Результаты классификации для каждого класса в зависимости от количества лучших признаков (набор GSE52581)

Таблица 3

Результаты классификации на тестовой выборке по 19 лучшим признакам для каждого заболевания

Класс	Метрики			
	precision	recall	F1-score	balanced accuracy
Астроцитома	0,915	0,903	0,908	0,946
Рак молочной железы	0,971	0,979	0,975	0,988
Рак молочной железы IV стадии	0,915	0,978	0,942	0,988
Мультиформная глиобластома	0,756	0,767	0,738	0,881
Здоровые	0,893	0,968	0,928	0,972
Рак легкого	0,867	0,963	0,912	0,976
Смешанная олигоастроцитома	0,99	0,865	0,922	0,932
Множественная миелома	0,962	0,757	0,84	0,877
Олигодендроглиома	0,938	0,85	0,887	0,924
Карцинома яичника	0,989	0,989	0,989	0,994
Рак поджелудочной железы	0,958	0,991	0,974	0,993
Панкреатит	0,973	0,962	0,967	0,980
Повторный рак молочной железы	0,786	0,771	0,768	0,881
Саркома Юинга	1	1	1	1
Кокцидиоидомикоз	0,964	0,956	0,960	0,976

Заключение

Основной сложностью практической работы с данными, полученными посредством иммуносигнатурного анализа, является высокая размерность и наличие значительного числа неинформативных или ложно-информативных признаков в связи с особенностью технологии. Для обеспечения практически значимого качества классификации необходимо учитывать эти особенности. Представленная технология отбора информативных признаков позволяет обеспечить высокие оценки качества классификации при значительном сокращении признакового пространства.

Количество признаков, отсеянных на втором шаге технологии (являющихся ложно-информативными), составляет примерно 50% для каждого из рассмотренных наборов данных, что в значительной степени упрощает последующий их анализ. После третьего шага для набора данных GSE52580 при сокращении признакового пространства до 15 признаков оценка качества классификации по метрике macro-average F1-score составляет 98,9%. Для набора данных GSE52581 при сокращении признакового пространства до 266 признаков оценка качества классификации по этой метрике составляет 91,3%.

Результаты работы демонстрируют перспективность предложенной технологии для отбора информативных признаков применительно к данным иммуносигнатурного анализа.

Финансирование исследования. Работа не получала финансовой поддержки.

Конфликт интересов отсутствует.

Литература/References

1. *Злокачественные новообразования в России в 2018 году (заболеваемость и смертность)*. Под ред. Каприна А.Д., Старинского В.В., Петровой Г.В. М: МНИОИ им. П.А. Герцена — филиал ФГБУ «НМИЦ радиологии» Минздрава России; 2019; 250 с.

Zlokachestvennye novoobrazovaniya v Rossii v 2018 godu (zabolevaemost' i smertnost') [Malignant neoplasms in Russia in 2018 (morbidity and mortality)]. Pod red. Kaprina A.D., Starinskogo V.V., Petrovoy G.V. [Kaprin A.D., Starinskiy V.V., Petrova G.V. (editors)]. Moscow: MNIOI im. P.A. Gertsena — filial FGBU "NMITs radiologii" Minzdrava Rossii; 2019; 250 p.

2. World Health Organization. *Guide to cancer early diagnosis*. World Health Organization; 2017. URL: https://apps.who.int/iris/bitstream/handle/10665/254500/9789241511940%20eng.pdf;jsessionid=F414948FB143C37513D7C21E675BA9C8?sequence=1.

3. Stafford P., Halperin R., Legutki J.B., Magee D.M., Galgiani J., Johnston S.A. Physical characterization of the "immunosignaturing effect". *Mol Cell Proteomics* 2012; 11(4): M111.011593, https://doi.org/10.1074/mcp.m111.011593.

4. Blum A.L., Langley P. Selection of relevant features and examples in machine learning. *Artificial Intelligence* 1997; 97(1–2): 245–271, https://doi.org/10.1016/s0004-3702(97)00063-5.

5. Kukreja M., Johnston S.A., Stafford P. Immunosignaturing microarrays distinguish antibody profiles of related pancreatic diseases. *J Proteomics Bioinform* 2013; S6(1): 1–5, https://doi.org/10.4172/jpb.s6-001.

6. Stafford P., Cichacz Z., Woodbury N.W., Johnston S.A. Immunosignature system for diagnosis of cancer. *Proc Natl Acad Sci U S A* 2014; 111(30): E3072–E3080, https://doi.org/10.1073/pnas.1409432111.

7. Анисимов Д.С., Подлесных С.В., Колосова Е.А., Щербаков Д.Н., Петрова В.Д., Джонстон С.А., Лазарев А.Ф., Оскорбин Н.М., Шаповал А.И., Рязанов М.А. Анализ многомерных данных пептидных микрочипов с использованием метода проекции на латентные структуры. *Математическая биология и биоинформатика* 2017; 12(2): 435–445, https://doi.org/10.17537/2017.12.435.

Anisimov D.S., Podlesnykh S.V., Kolosova E.A., Shcherbakov D.N., Petrova V.D., Dzhonston S.A., Lazarev A.F., Oskorbin N.M., Shapoval A.I., Ryazanov M.A. Projection to latent structures as a strategy for peptides microarray data analysis. *Matematicheskaa biologia i bioinformatika* 2017; 12(2): 435–445, https://doi.org/10.17537/2017.12.435.

8. Subramanian J., Simon R. Overfitting in prediction models — is it a problem only in high dimensions. *Contemp Clin Trials* 2013; 36(2): 636–641, https://doi.org/10.1016/j.cct.2013.06.011.

9. Stafford P., Zbigniew C., Johnston S. *An immunosignature system for diagnosis of cancer [Cancer immunosignaturing — test 1].* National Center for Biotechnology

Information Search database; 2013. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52580.

10. Stafford P., Zbigniew C., Johnston S. *An immunosignature system for diagnosis of cancer [Cancer immunosignaturing — test 2].* National Center for Biotechnology Information Search database; 2013. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52581.

11. Wickham H. Tidy data. *J Stat Softw* 2014; 59(10), https://doi.org/10.18637/jss.v059.i10.

12. Izetta J., Verdes P.F., Granitto P.M. Improved multiclass feature selection via list combination. *Expert Syst Appl* 2017; 88: 205–215, https://doi.org/10.1016/j.eswa.2017.06.043.

13. Bommert A., Sun X., Bischl B., Rahnenführer J., Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics & Data Analysis* 2020; 143: 106839, https://doi.org/10.1016/j.csda.2019.106839.

14. Shannon C.E. A mathematical theory of communication. *Bell System Technical Journal* 1948; 27(3): 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

15. Andryushchenko V.S., Uglov A.S., Zamyatin A.V. Statistical classification of immunosignatures under significant reduction of the feature space dimensions for early diagnosis of diseases. *Sovremennye tehnologii v medicine* 2018; 10(3): 14–20, https://doi.org/10.17691/stm2018.10.3.2.

16. Cortes C., Vapnik V. Support-vector networks. *Mach Learn* 1995; 20(3): 273–297, https://doi.org/10.1007/BF00994018.

17. Powers D. Evaluation: from precision, recall and F-factor to ROC, informedness, markedness & correlation. *J Mach Learn Tech* 2007; 2: 37–63.