Информационные процессы, Том 25, № 2, 2025, стр. 213–223.

Алгоритм классификации и его применение для прогноза рецидива при папиллярной микрокарциноме щитовидной железы[1]

А. П. Вайнштok[*], Е. А. Ващенко[*], Н. С. Кузнецов[**], М. В. Скибицкая[**]

[*] *ФГБУН «Институт проблем передачи информации им. А. А. Харкевича», Российская академия наук, Москва, Россия*

[**] *ГНЦ РФ ФГБУ «Национальный медицинский исследовательский центр эндокринологии» Минздрава России, Москва, Россия*

Поступила в редколлегию 28.05.2025 г. Принята 14.07.2025 г.

Аннотация—В статье представлены основные идеи и возможности алгоритма обучения классификации с учителем Фрагмент, который позволяет находить ансамбль классифицирующих решателей, и алгоритма классификации (принятия решения) Потенциал, в котором осуществляется голосование по множеству ансамбля решателей. В работе представлены результаты применения программ Фрагмент и Потенциал для прогноза рецидива после операции по поводу папиллярной микрокарциномы щитовидной железы (ПМЩЖ) по данным дооперационного обследования. Возможность возникновения рецидива является наиболее значимым результатом отдалённых последствий оперативного вмешательства. Полученные результаты свидетельствуют о перспективности использования разработанных методов для прогноза рецидива.

КЛЮЧЕВЫЕ СЛОВА: классификация, машинное обучение, метод фрагмент-потенциал, щитовидная железа, папиллярный рак, микрокарцинома, дооперационное обследование, хирургическое лечение, рецидив, прогноз.

DOI: 10.53921/18195822_2025_25_2_213

1. ВВЕДЕНИЕ

Одной из основных задач искусственного интеллекта и машинного обучения является классификация, которая соотносит объекты с заранее определёнными категориями или классами на основе их признаков. Классификация помогает принимать обоснованные решения на основе данных. Классификация необходима практически во всех видах деятельности; медицина, финансы, обработка и анализ текстов и изображений, применение естественного языка и многих других.

Применяются две основные модели классификации:

1. Модель обучения с учителем – модель обучается на наборе данных (обучающей выборке), в котором для каждого объекта известна принадлежность к правильному классу. Алгоритм находит зависимости между признаками и классами.
2. Модель обучения без учителя - выделяются группы объектов на основе их схожести в пространстве признаков без заранее известной принадлежности к классам (кластеризация).

После обучения производится оценка модели (экзамен) и она может быть использована для прогнозирования принадлежности к классам новых объектов.

[1] Посвящается памяти Переверзева-Орлова В. С. - автор идеи алгоритмов ФРАГМЕНТ и ПОТЕНЦИАЛ.

В статье представлены основные идеи и возможности алгоритма обучения классификации с учителем Фрагмент, который позволяет находить ансамбль классифицирующих решателей, и алгоритма классификации (принятия решения) Потенциал, в котором осуществляется голосование по множеству ансамбля решателей.

Первые версии алгоритмов и программ были разработаны в ИППИ РАН в начале 80-ых годов прошлого века. Наиболее полно идеи алгоритмов и их свойства описаны в монографии [1]. Первое описание программной реализации алгоритма Фрагмент представлено в [2]. В работе [3] рассматриваются свойства решателей, используемых в процедуре голосования. В последующие годы программа развивалась и применялась для ряда областей (различные разделы медицины, биржевая аналитика, свойства неорганических соединений) как средство машинного обучения [4–7].

В работе представлено применение программ **Фрагмент** и **Потенциал** для прогноза рецидива при папиллярной микрокарциноме щитовидной железы (ПМЩЖ) по данным дооперационного обследования. Возможность проявления рецидива является наиболее информативно значимым результатом отдалённых последствий оперативного вмешательства (гемитиреоидэктомии).

2. АЛГОРИТМ ФРАГМЕНТ

2.1. Построение классифицирующего решателя.

В основу разработки алгоритмов **ФРАГМЕНТ и ПОТЕНЦИАЛ** была положена “идея кликовости”, в соответствии с которой одновременно должны удовлетворяться требования по составу подмножеств объектов и признаков (первичных описаний), на которых максимизируется некий критерий смысловой однородности состава клики.

В алгоритме ФРАГМЕНТ реализуется метод «сострогивания» — такое проведение разделяющей плоскости в пространстве описаний, когда по одну сторону преимущественно оказываются объекты какого-либо одного из классов, а по другую объекты других классов. Практически, однако, допускаются и не совсем «чистые сострогивания», когда в сострогиваемый класс попадает небольшое количество объектов других классов. Последовательное применение сострогиваний приводит к созданию кусочно-линейного решающего правила (решателя) при любом числе классов.

Выбор плоскости для сострогивания — итерационная процедура. Сначала поиск ведем на множестве плоскостей, каждая из которых определяется значением какого-либо одного из признаков. Отбор ведётся по качеству сострогивания, характеризующему упрощение задачи разделения после выбора данного сострогивания. Простейший показатель качества — доля сострганного. Пробное сострогивание с наибольшим качеством определяет некий признак. Когда такой признак найден, можно попытаться улучшить сострогивание по нему, несколько поворачивая координатную ось, соответствующую ему в пространстве, что достигается поочередным прибавлением к этому признаку оставшихся и оценкой получаемых сострогиваний. Парная комбинация признаков, соответствующая наилучшему сострогиванию, включается в число исходных признаков в качестве еще одного, после чего выясняется, нельзя ли улучшить и этот признак, и т. д. Когда достигается предел улучшения качества или допустимого числа итераций, по последнему синтезированному признаку проводится рабочее сострогивание, исключающее из последующего рассмотрения сострогиваемые точки (в наших экспериментах с алгоритмом ФРАГМЕНТ обычно допускалась только одна такая итерация). Затем, с учетом всех уже имеющихся признаков для оставшихся точек ищется следующее сострогивание и т.д., пока не будет завершено построение решающего правила. Обычно это осуществляется за один—пять шагов. В реализованном нами варианте этого алгоритма объекты с пропусками

значений признаков, входящих в синтезированный признак, при проверке по нему игнорируются, и поэтому разделение ведётся всегда для объектов с известными значениями.

Таким образом, алгоритм ФРАГМЕНТ позволяет находить маломерные (обычно их размерность два - четыре) просто интерпретируемые кусочно-линейные решающие правила на основании известных характеристик объектов. Время синтеза таких правил почти пропорционально размерности описания и числу объектов в выборке.

2.2. Построение ансамбля (множества) решателей.

Для генерации множества классифицирующих решателей у алгоритма ФРАГМЕНТ, помимо повторного поиска решения с учётом признаков, синтезированных в предыдущем поиске, имеются обширные возможности, основанные на вариациях основных параметров алгоритма, определяющих конкретный вид оценки качества, приоритеты классов, положение сострагивающей плоскости, допустимую долю пропуска «чужих» объектов при сострагивании и т. д. Простой способ - поиск с вычёркиванием, когда перед построением следующих решающих правил исключаются признаки, с которых начинался синтез предшествующих правил или запрещается их использовать в качестве первых. Таким образом, в получаемый в результате ансамбль решателей, хотя каждый из них маломерен, включается значительная часть признаков исходного пространства и тем самым реализовать избыточность исходных показателей.

Примеры из нашей практики подтверждают эффективность такого подхода. Например, в задаче со сверхпроводниками в 48-мерном пространстве описаний для отделения высокотемпературных соединений от низкотемпературных по такой методике удалось построить более 20 в основном одношаговых и двушаговых решающих правил, пока сложность их не начала резко увеличиваться, свидетельствуя об исчерпании множества существенных признаков. Отметим, что исключение части признаков из рассмотрения при таком подходе в силу градиентного характера поиска алгоритмом ФРАГМЕНТ решающего правила, видимо, не очень сильно влияет на возможности последующего синтеза. При этом обычно наблюдается неповторимость последующих решателей, что особенно заметно при поиске таким способом первых решателей, когда ещё много неиспользованных признаков.

3. АЛГОРИТМ ПОТЕЦИАЛ

Принятие решений методом голосования — это один из простых, но эффективных методов принятия решений в задачах классификации новых объектов (не участвовавших в обучении). Метод основан на агрегировании предсказаний нескольких моделей (классификаторов), в нашем случае решателей. Этот подход может повысить точность результатов и уменьшить вероятность ошибок, связанных со случайными флуктуациями в данных. Метод голосования делает систему более устойчивой к ошибкам, поскольку даже если отдельные решатели ошибаются, общее предсказание может оказаться верным благодаря множеству решателей, использующих избыточность данных.

Существуют разные способы реализации метода голосования, однако все они основаны на одном основном принципе: комбинирование результатов нескольких моделей (решателей) для получения финального предсказания. Наиболее простой способ голосования - результаты классификации всех решателей объединяются, и класс, который получил наибольшее количество голосов, становится финальным прогнозом или вводится порог, определяющий разность между голосами за класс, получивший максимальное количество голосов и ближайшего по количеству голосов класса.

4. ПРОГРАММА ФРАГМЕНТ – ПОТЕНЦИАЛ.

Алгоритмы Фрагмент и Потенциал реализованы в одной программе, в которой осуществлены 3 режима:

1. Поиск признаков и конструирование решающих правил.
2. Экзамен по полной выборке для разных порогов принятия решения при голосовании (разница голосов за различаемые классы), на основании чего выбирается порог Tr для классификации новых прецедентов, обычно $Tr \geq 2$.
3. Скользящий контроль, который позволяет оценить качество найденных признаков и решателей.

В конфигурационном файле задаются следующие параметры для поиска признаков и решающих правил для двух классов:

1. число классов,
2. минимальное и максимальное число плоскостей в одном правиле - определяют удовлетворение правила условиям модели и условие останова для построения правила для пары классов,
3. параметр, запрещающий или разрешающий использовать признак, вошедший первым в уже построенное правило для данной пары классов, в последующем построении других правил,
4. параметр, который устанавливает режим использования любого признака только один раз в правилах,
5. доля неотделенных объектов в любом из пары классов, при которой построение правила завершается,
6. критерий качества разделения классов для включения показателя в правило (три критерия):

- отделение максимального числа объектов класса при заданной допустимой доле объектов противоположного класса;

- отделение числа объектов класса не менее заданной величины при числе объектов противоположного класса не более заданной величины (абсолютные значения);

- максимум разности двух показателей, связанных с оценкой качества классификации в задачах машинного обучения и статистики

$$\frac{TP}{TP+FN} - \frac{FP}{FP+TN}, \text{ где}$$

TP (True Positives) истинно положительные — количество правильно предсказанных (классифицированных) положительных случаев при заданной допустимой доле объектов противоположного класса,

FN (False Negatives) ложно отрицательные — количество положительных случаев, неправильно классифицированных как отрицательные,

FP (False Positives) ложно положительные — количество отрицательных случаев, неправильно классифицированных как положительные,

TN (True Negatives) истинно отрицательные — количество правильно предсказанных отрицательных случаев.

Формула может быть интерпретирована как разность между чувствительностью (точностью) и специфичностью.

Для экзамена по полной выборке задается минимальное, максимальное значения и шаг для перебора по порогу для голосования Tr, по результатам которого выбирается порог для принятия решения при классификации новых прецедентов в реальных случаях.

Для скользящего контроля случайным образом выбираются прецеденты, исключаемые из обучения, доля которых задается. Также может быть задан отбор из классов одинакового количества объектов.

В результате работы программы формируется текстовый файл с решающими правилами и выводится информация, позволяющая пользователю оценивать качество правил и результаты классификации. В частности, для всех показателей, вошедших в ансамбль правил выводится частота их включения и редко используемые показатели могут быть запрещены для применения, после чего производится новый поиск решателей. Имеются и другие приемы управления пользователем настройкой решателей.

Результаты экзамена для каждого объекта обучающей выборки и скользящего контроля содержат: идентификатор объекта, исходный номер класса объекта, если он известен, и номер класса, к которому отнесен объект моделью. Если объект не набирает порогового количества голосов, чтобы быть отнесенным к определенному классу, то он считается нераспознанным сформированной моделью.

5. ПАПИЛЛЯРНАЯ МИКРОКАРЦИНОМА ЩИТОВИДНОЙ ЖЕЛЕЗЫ

Задачей и тенденцией современной медицины является персонализация лечебного процесса. Это возможно при анализе особенностей индивидуальных показателей больного с использованием современного инструментария, информационных технологий и компьютерных средств поддержки принятия решений, помогающих врачу в диагностике и выборе эффективной тактики лечения [8].

Папиллярный рак встречается в 85-90% случаев злокачественных опухолей щитовидной железы (ШЖ) [9]. Согласно исследованиям [10, 11] около 50 % новых случаев РЩЖ составляют злокачественные опухоли размером до 1 см, в Европе их доля за 30 лет увеличилась с 18 до 40% [10]. В Соединенных Штатах 25% новых случаев РЩЖ, диагностированных в 1988–1989 годах, были $\leq$1 см по сравнению с 39% новых случаев, диагностированных в 2008–2009 годах [11]. Это объясняется улучшением диагностики на более ранних стадиях заболевания. Согласно классификации Всемирной организации здравоохранения опухоль размером $\leq$1 см определена как папиллярная микрокарцинома щитовидной железы (ПМЩЖ) [11]. Американской национальной комплексной онкологической сетью (NCCN) при ПМЩЖ и отсутствии регионарных или отдалённых метастазов рекомендуется либо регулярное наблюдение, либо проведение операции - гемитиреоидэктомии [12].

Для врачей значительный интерес представляет прогноз отдалённых результатов гемитиреоидэктомии ПМЩЖ по дооперационным наблюдениям. Наиболее информативно значимым результатом является возможность проявления рецидива.

Дооперационный качественный прогноз возможности рецидива при планировании гемитиреоидэктомии может быть критерием для выбора адекватной тактики хирургического лечения, определяющей объём оперативного вмешательства. Эта проблема может быть решена созданием формализованного способа выделения комплекса диагностически значимых показателей из данных протокола дооперационного обследования и прогнозирования по ним возможности рецидива с последующим принятием решения по объёму операции, при котором при прогнозировании отсутствия рецидива пациенту проводят гемитиреоидэктомию, а при прогнозировании рецидива выполняют гемитиреоидэктомию с ипсилатеральной лимфаденэктомией.

6. МАТЕРИАЛ ИССЛЕДОВАНИЯ (ВЫБОРКА)

В статье [13] была представлена методика составления выборки, метод прогнозирования и пилотные эксперименты, которые показали наличие закономерностей, позволяющих решать задачу прогноза рецидива по показателям дооперационного обследования.

В настоящее время подготовлена и выверена ретроспективная выборка из базы данных (БД) «Национального медицинского исследовательского центра эндокринологии (ЭНЦ)» Минздрава России. В выборку вошли показатели дооперационного обследования 115 пациентов, проходивших оперативное лечение в хирургическом отделении ЭНЦ в 2017–2024 гг., которым был впервые поставлен диагноз ПМЩЖ и проведена гемиотериодиктомия. В выборке больных микрокарциномой были представлены пациенты в возрасте от 17 до 83 лет, средний возраст 43.27 лет, среди которых, 64 пациента без рецидива и 51 с рецидивом. Класс без рецидива составили больные, у которых в течение 5 лет после гемиотериодиктомии не возник рецидив, в эту группу отобраны пациенты, у которых операция была в 2017–2018 гг.

Комплекс показателей (признаков), используемых для прогноза рецидива, выявлен из 18 первичных показателей дооперационного обследования. Данные включают следующие показатели из историй болезни: пол, возраст, размер опухоли, неровность контура опухоли - да/нет, наличие кальцинатов в ЩЖ - да/нет, эхоструктура опухоли - изоэхогенная, гиперэхогенная, гипоэхогенная, количество метастазов в поражённой доли, наличие патологии лимфатических узлов (ЛУ) - да/нет, неровность/нечеткость контура ЛУ - да/нет, наличие кальцинатов в ЛУ - да/нет, эхоструктура ЛУ - изоэхогенная, гиперэхогенная, гипоэхогенная, количество метостатических ЛУ, классификация по международной системе TNM, классификация узлового образования щитовидной железы по системе TIRADS, результаты тонкоигольной аспирационной биопсии узлов щитовидной железы, риск по диагностическим категориям BETHESDA, кальцитонин - кровь пг/мл, морфология - количество поражённых ЛУ (показатель включен для контроля).

7. ПРИМЕНЕНИЕ МЕТОДОВ ФРАГМЕНТ И ПОТЕНЦИАЛ ДЛЯ ПРОГНОЗИРОВАНИЯ РЕЦИДИВОВ ПРИ ПМРЩЖ.

В ансамбле классифицирующих правил проверяется соответствие значений показателей пациента диагностически значимым показателям (признакам) и условиям правил, найденным в программе **Фрагмент,** и относящим пациента к одному из классов: **1 – нет рецидива, 2 - есть рецидив.** Решение принимается за класс 1, если $N_1 - N_2 \geq Tr$, за класс 2 если $N_2 - N_1 \geq Tr$, (N_1, N_2 – количество правил, относящих прецедент к 1-му или 2-му классу соответственно, рекомендуемый порог $Tr \geq 2$). Т.е. рецидив не прогнозируется, если число правил за 1-й класс N_1 на два или более чем за 2-й класс N_2 и рецидив прогнозируется, если число правил за 2-й класс N_2 на два или более чем за 1-й класс N_1. Чем больше разница, тем больше надёжность соответствующего прогнозирования, если указанное условие не выполняется, решение не принимается.

Каждое правило - это последовательность узлов, каждый из которых содержит имя признака (идентификатор), условие (“<=” или “>”), значение признака и номер класса к которому относится прецедент.

Программа отобрала 5 показателей (признаков) и построила 5 правил с максимальным числом условий в одном правиле равным 4. Для каждого прецедента число сработавших правил может быть разным. Для принятия решения о включении показателя в правило использовался критерий максимума разности между чувствительностью и специфичностью.

Ниже приведены показатели, их идентификаторы и числовые коды значений показателей, используемые в классифицирующих правилах. Порядок показателей соответствует их зна-

чимости для прогнозирования, значимость определяется количеством их срабатываний при обучении модели. Список диагностически значимых признаков включает следующие 5 характеристик дооперационного обследования пациентов:

1. УЗИ л/у, количество метостатических узлов, (qty_{nodes}).
2. УЗИ л/у, патология выявлена [нет – 1, да - 2], (uzi_{nodes}).
3. УЗИ опухоли, количество выявленных метастазов, ($mets$).
4. УЗИ л/у, эхоструктура [не выявлена -1, изоэхогенная - 2, гиперэхогенная - 3, гипоэхогенная - 4], (eho_{nodes}).
5. УЗИ л/у, неровность\нечёткость контура [нет – 1, да - 2], ($cont_{nodes}$).

Ансамбль классифицирующих правил схематично представлен следующими комбинациями показателей с соответствующими отношениями (в скобках указан класс, к которому относит узел):

$qty_{nodes} = 0(1) \rightarrow qty_{nodes} > 0(2)$

$uzi_{nodes} <= 1(1) \rightarrow qty_{nodes} = 0(1) \rightarrow qty_{nodes} > 0(2)$

$mets =0(1) \rightarrow mets > 0(2)$

$eho_{nodes} = 1(1) \rightarrow eho = 1(1) \rightarrow eho > 1(2)$

$cont_{nodes} >1(2) \rightarrow mets >1(2) \rightarrow mets = 0(1) \rightarrow qty_{nodes} < = 2(2)$

Пример структуры и свойств правила представлены в виде таблицы 1. В выборке 115 пациентов, из них прецедентов 1-го класса (пациенты без рецидива) – 64, прецедентов 2-го класса (пациенты с рецидивом) – 51.

Таблица 1. Пример обучения классификации

Узел	Условие	Класс	$N_{\text{кл}}$	$N_{\text{альт}}$	$N_{\text{ост\_прог}}$	$N_{\text{ост\_альт}}$
1	$cont_{nodes} > 1$	2	30	6	21	58
2	$mets >1$	2	5	0	16	58
3	$mets = 0$	1	55	6	3	10
4	$qty_{nodes} <= 2$	2	10	2	0	1

Обозначения: Класс – номер класса $(1 \vee 2)$, к которому относятся пациенты, удовлетворяющие условию, 1 – нет рецидива, 2 – рецидив; $N_{\text{кл}}$ - количество правильно классифицированных прецедентов; $N_{\text{альт}}$ - количество неправильно классифицированных прецедентов альтернативного класса; $N_{\text{ост\_прог}}$ – текущее количество оставшихся прецедентов в прогнозируемом классе, $N_{\text{ост\_альт}}$ – текущее количество оставшихся прецедентов в альтернативном классе.

Выполнены экзамены на полной (обучающей) выборке и контрольных выборках методом скользящего контроля. Результаты экзамена на учебной выборке Gthdfze представлены в таблице 2.

Таблица 2. Результаты экзамена на учебной выборке

$Tresh$	$True$	$Error$	$True_1$	$True_2$	$False_1$	$False_2$	Un_1	Un_2	Sum_1	Sum_2
1	100	15	55	45	9	6	0	0	64	51
2	100	15	55	45	9	6	0	0	64	51
3	100	15	55	45	9	6	0	0	64	51
4	94	11	55	39	6	5	3	7	64	51
5	94	11	55	39	6	5	3	7	64	51

Обозначения: $Tresh$ – порог (разница между количеством правил, голосующих за каждый класс, по которой принимается решение), $True$ – общее количество правильно классифицируемых пациентов в обоих классах, $Error$ – общее количество ошибок в двух классах, $True_1$ и $True_2$ - количество правильно классифицируемых пациентов за 1-й и 2-й классы соответственно, $False_1$ и $False_2$ - количество ошибок за 1-й и 2-й классы соответственно, Un_1 и Un_2 – «не знаю» за 1-й и 2-й классы соответственно, Sum_1 и Sum_2 - количество пациентов в 1-ом и 2-ом классах соответственно.

Исходя из результатов в таблице 2, определяется порог, используемый для прогноза в рабочем режиме. Для описываемого эксперимента целесообразно значение порога 3. При $Tresh$ =3 доля правильно классифицированных рецидивов (вероятность обнаружения рецидива, чувствительность) TPR=True_2/Sum_2=45/51=0.88, доля «пропущенных» случаев рецидива (вероятность пропуска цели) $FNR = 1 - TPR = 0.12$, доля правильно классифицированных «не рецидивов» (специфичность) TNR=True_1/Sum_1 = 55/64=0.86, доля (вероятность) ложных тревог (рецидива нет, а прогноз рецидива есть) $FPR = 1 - TNR = 0.14$. Доля ошибочных классификаций

$$Perr = \frac{Error}{Sum_1 + Sum_2} = \frac{15}{64 + 51} = 0.13.$$

Другими словами, чувствительность прогностического алгоритма достигает 88% при специфичности 86% и вероятности ошибки 13%.

Агрегированные результаты 10 испытаний методом скользящего контроля на 12 случайно выбранных прецедентах (∼10% от размера обучающей выборки) представлены в таблице 3.

Таблица 3. Агрегированные результаты 10 испытаний методом скользящего контроля на 12 случайно выбранных прецедентах.

№	$True$	$Error$	$True_1$	$True_2$	$False_1$	$False_2$	Un_1	Un_2	Sum_1	Sum_2
1	11	1	6	5	0	1	0	0	6	6
2	12	0	6	6	0	0	0	0	6	6
3	10	2	4	6	2	0	0	0	6	6
4	11	1	5	6	1	0	0	0	6	6
5	9	3	5	4	1	2	0	0	6	6
6	11	1	6	5	0	1	0	0	6	6
7	11	1	6	5	0	1	0	0	6	6
8	8	2	4	4	1	1	1	1	6	6
9	10	1	4	6	1	0	1	0	6	6
10	11	1	6	5	0	1	0	0	6	6
Total	*104*	**13**	**52**	**52**	**6**	**7**	**2**	**1**	**60**	**60**

Суммарные показатели по 2-ум классам: $Ntrue$=104, $Nerr$=13, $Nunkn$=3. Суммарная доля правильно классифицированных рецидивов по 10 испытаниям скользящего контроля (вероятность обнаружения рецидива, чувствительность)

$$TPR = \frac{True_2}{Total\,(Sum_2)} = \frac{52}{60} = 0.87,$$

доля «пропущенных» случаев рецидива (вероятность пропуска цели) $FNR = 1 - TPR$=0.13, суммарная доля правильно классифицированных «нет рецидивов» (специфичность)

$$TNR = \frac{True_1}{Total\,(Sum_1)} = \frac{52}{60} = 0.87,$$

суммарная доля (вероятность) ложных тревог (рецидива нет, а прогноз рецидива есть) $FPR = 1 - TNR = 0.13$, суммарная доля ошибочных классификаций и неклассифицированных прецедентов

$$Perr = \frac{Error + Unk_1 + Unk_2}{Total\,(Sum_1) + Total\,(Sum_2)} = \frac{13 + 2 + 1}{60 + 60} = 0.13.$$

Таким образом, можно считать, что для вновь поступающих пациентов вероятность правильного прогноза 0.87 а вероятности ошибки 0.13. Полученные результаты практически совпадают с результатами по полной выборке, что подтверждает эффективность используемого метода для прогноза рецидивов.

Приведём примеры клинических случаев. Решение принимается в пользу одного из классов при условии, если разница между голосами, отданными за классы, не менее 2 (порог может меняться при экзамене), в противном случае, решение о принадлежности к одному из классов не принимается. Каждый решатель представлен своими узлами, приведены узлы до момента принятия решения по данному пациенту. Ниже приведены 2 примера применения способа для пациентов с известной историей и 2 примера прогноза.

7.1. Клинический пример 1

Пациентка 39 лет, женщина, гемитиреоидэктомия ПМЩЖ выполнена в сентябре 2022 г. в хирургическом отделении ГНЦ РФ ФГБУ «НМИЦ эндокринологии» Минздрава России. В марте 2023 г. у пациентки выявлен рецидив. Результаты диагностически значимых признаков дооперационного обследования, используемых в решателях в кодированном виде приведены в таблице 4. Дополнительно для контроля приведено количество поражённых лимфоузлов по морфологии (qty_{morf}).

Таблица 4. Кодированные значения вошедших в правила показателей пациента

id	pol	age	$mets$	eho	uzi_{nodes}	$cont_{nodes}$	eho_{nodes}	qty_{nodes}	$reced$	$data_{rec}$	qty_{morf}
4	1	39	4	3	2	2	4	4	2	2023	7

Применим правила для значений показателей пациента, в $(\cdot)$ указан класс, к которому относит узел, в $[\cdot]$ указано значение показателя пациента:

$qty_{nodes} = 0(1)\ [4] \rightarrow qty_{nodes} > 0(2)\ [4]$

$uzi_{nodes} = 1(1)\ [2] \rightarrow qty_{nodes} = 0(1)\ [4] \rightarrow qty_{nodes} > 0(2)\ [4]$

$mets = 0(1)\ [4] \rightarrow mets > 0(2)\ [4]$

$eho_{nodes} = 1(1)\ [4] \rightarrow eho = 1(1)\ [3] \rightarrow \text{eho} > 1(2)\ [3]$

$cont_{nodes} > 1(2)\ [2]$

Все 5 правил относят пациентку ко 2-му классу – прогнозируется рецидив, что соответствует действительности. Правила приводятся до узла, в котором выполняется условие, т.е. значение показателя пациента соответствует условию узла,

Показатель патология л/у (uzi_{nodes}) принимает значение «не выявлена (1)»/«выявлена (2)» и мажорирует показатели количество метостатических узлов (qty_nodes), эхоструктура л/у (eho_{nodes}), неровность\нечёткость контуров ($cont_{nodes}$). В нашем случае паталогия л/у выявлена.

Сравнение значений показателей пациента со значениями в узлах правил показывает, что у пациента надёжно классифицируется рецидив и что ему изначально следовало делать гемитиреоидэктомию с ипсилатеральной лимфаденэктомией.

7.2. Клинический пример 2

Пациент 38 лет, женщина, гемитиреоидэктомия ПРЩЖ выполнена в июле 2017 г. в хирургическом отделении ГНЦ РФ ФГБУ «НМИЦ эндокринологии» Минздрава России, рецидив не выявлен. Результаты диагностически значимых признаков дооперационного обследования, используемых в решателях в кодированном виде приведены в таблице 5.

Таблица 5. Кодированные значения вошедших в правила показателей пациента

id	pol	age	$mets$	eho	uzi_{nodes}	$cont_{nodes}$	eho_{nodes}	qty_{nodes}	$receed$	$data_{rec}$	qty_{morf}
72	1	38	0	3	1	1	1	0	1	-	0

Применим правила для значений показателей пациента, в $(\cdot)$ указан класс, к которому относит узел, в $[\cdot]$ указано значение показателя пациента:

$qty_{nodes} = 0(1)$ [0]
$uzi_{nodes} = 1(1)$ [1]
$mets = 0(1)$ [0]
$eho_{nodes} = 1(1)$ [1]
$cont_{nodes} > 1(2)$ [1] $\rightarrow$ $mets > 1(2)$ [0] $\rightarrow$ $mets = 0(1)$ [0]

Значения соответствующих показателей у пациентки в первых 4-ёх правилах совпадают с условиями 1-го узла. Все 5 правил относят пациентку к 1-му классу – нет рецидива, что соответствует действительности. Поскольку у пациентки надёжно не прогнозируется рецидив, то ей и следовало делать гемитиреоидэктомию изначально.

8. ЗАКЛЮЧЕНИЕ.

В работе рассмотрены свойства и возможности оригинального метода классификации с учителем, объединяющем два алгоритма:

1. Нахождение признаков и построение множества (ансамбля) кусочно-линейных решающих правил попарно разделяющих прецеденты классов (ФРАГМЕНТ);
2. Голосование - комбинирование результатов нескольких решателей для получения финального предсказания (ПОТЕНЦИАЛ).

Важной характеристикой метода является способность работать с малыми объемами обучающей выборки, объекты которой могут содержать большое число показателей с возможными ошибками в описаниях.

Метод ФРАГМЕНТ-ПОТЕНЦИАЛ применен для построения правила прогнозирования рецидива по данным дооперационного обследования пациентов с ПМЩЖ, которым проведена гемитериоидэктомия. Скользящий контроль показал вероятность правильного прогноза 0.87, а вероятность ошибки 0.13. Полученные результаты свидетельствуют о перспективности использования метода для помощи врачам в принятии решения о выборе адекватной тактики хирургического лечения, определяющей объём оперативного вмешательства.

СПИСОК ЛИТЕРАТУРЫ

1. Левит В.Е., Переверзев-Орлов В.С. Структура и поле данных при распознавании образов. М.: Наука; 1984.

2. Вайнштok А.П., Переверзев-Орлов В.С. Программный комплекс «Распознавание образов с одновременным поиском подмножеств признаков и правил, попарно разделяющих классы». ГФАП СССР. Алгоритмы и программы. Информационный бюллетень. 1985; 4(67)

3. Вайнштok А.П., Левит В.Е. О непрерывности относительно метрики Хэмминга функций принадлежности к классу в методах распознавания образов использующих процедуры голосования. // Сборник “Обработка данных в информационных системах (часть II)”. ИППИ АН СССР, Москва, 1986.стр. 40-43.

4. Vashchenko E.A., Vitushko M.A., Gurov N.D., Pereverzev-Orlov V.S., Stenina I.I. Knowledge and Data Cooperation. Pattern Recognition and Image Analysis, 1998, vol.8, no. 2, pp. 25–41.

5. Vashenko E, Vitushko M, Pereverzev-Orlov V. Potentials of Learning on the Basis of Partner System. Pattern Recognition and Image Analysis. 2004; 14(1):84-91.

6. Ващенко Е.А., Витушко М.А., Переверзев-Орлов В.С., Стенина И.И. Советчик врача: технологии и возможности. // Электронный научный журнал "Информационные Процессы". 2009; 9(1):18-24.

7. Е.А.Ващенко, М.А.Витушко, В.А.Дударев, Н.Н.Киселева, В.С.Переверзев-Орлов. К возможности прогнозирования значений параметров многокомпонентных неорганических соединений. // Информационные процессы, Том 19, № 4, 2019, стр. 415–432.

8. Персонализированная эндокринология в клинических примерах. Под редакцией академика РАН И.И. Дедова. Издательская группа «ГЭОСТАР-Медиа», 2019 г., 328 с.

9. А.Д. Каприн. Состояние онкологической помощи населению России в 2021 году. М.: МНИОИ им. П.А. Герцена — филиал ФГБУ «НМИЦ радиологии» Минздрава России, 2022. 239 с.

10. Lin JD. Increased incidence of papillary thyroid microcarcinoma with decreased tumor size of thyroid cancer. Med Oncol. 2010 Jun;27(2):510-8. doi: 10.1007/s12032-009-9242-8. Epub 2009 Jun 9. PMID: 19507072.

11. Lloyd RV, Osamura RY, Klöppel G, Rosai J. WHO Classification of Tumours of Endocrine Organs. Lyon: «IARC»; 2017. 355p.

12. Haugen BR. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: What is new and what has changed? Cancer. 2017 Feb 1;123(3):372-381. doi: 10.1002/cncr.30360. Epub 2016 Oct 14. PMID: 27741354.

13. Кузнецов Н.С., Скибицкая М.В., Вайнштok А.П., Ващенко Е.А. Прогноз рецидива при папиллярном раке щитовидной железы по дооперационным данным. // Хирургия. Журнал им. Н.И. Пирогова. 2024;(9):76-85.

Classification Algorithm and Its Application to Predict Recurrence in Papillary Thyroid Microcarcinoma Based on Preoperative Examination Data

A. P. Vaynshtok, E. A. Vashenko, N. S. Kuznetsov, M. V. Skibitskaya

The article presents the main ideas and capabilities of the supervised classification learning algorithm Fragment, which allows finding an ensemble of classifying solvers, and the classification (decision-making) algorithm Potential, in which voting is carried out on a set of solver ensembles. The paper presents the results of applying the Fragment and Potential programs to predict recurrence after surgery for papillary thyroid microcarcinoma based on preoperative examination data. The possibility of recurrence is the most significant result of the long-term consequences of surgical intervention. The obtained results indicate the prospects of using the developed methods for recurrence prediction.

KEYWORDS: classification, machine learning, fragment-potential method, thyroid gland, papillary cancer, microcarcinoma, preoperative examination, surgical treatment, recurrence, prognosis.