ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 20–27

УДК 004.8

КАЛИБРОВКА ВЕРОЯТНОСТЕЙ С ПРИМЕНЕНИЕМ ТЕОРИИ НЕЧЕТКИХ МНОЖЕСТВ НА ПРИМЕРЕ УЛУЧШЕНИЯ РАННЕЙ ДИАГНОСТИКИ РАКА

© 2023 г. О. А. Филимонова[1,*], А. Г. Овсянников[1], Н. В. Бирюкова[1]

Представлено академиком РАН А.И. Аветисяном
Поступило 29.08.2023 г.
После доработки 06.09.2023 г.
Принято к публикации 15.10.2023 г.

Рак занимает лидирующие позиции в списке причин смерти людей возрастом до 70 лет. Важным шагом для снижения смертности является выявление заболевания на ранних стадиях. Для улучшения ранней диагностики рака мы предлагаем алгоритм калибровки вероятностей бинарных классификаторов с применением нечетких множеств. Наша идея проверена на распознавании рака молочной железы у женщин и рака легкого. Первый случай осложняется небольшим набором данных, второй – сильно несбалансированными данными. В обоих случаях наш метод калибровки вероятностей, в отличие от стандартных, улучшил логарифмическую потерю (лучший результат – на 48.86%), оценку Брайера (лучший результат – на 13.24%) и площадь под кривой Precision-Recall (лучший результат – на 13.94%). Сфера применения нашего алгоритма может быть расширена на любые прогрессирующие заболевания и события без четкой границы принадлежности.

Ключевые слова: калибровка вероятностей, теория нечетких множеств, бинарная классификация, ранняя диагностика заболеваний

DOI: 10.31857/S2686954323601367, **EDN:** IDVYYR

1. ВВЕДЕНИЕ

По оценкам Всемирной организации здравоохранения в 2019 г. рак занимал лидирующие места среди причин смерти людей возрастом до 70 лет в 112 странах. Самыми диагностируемыми являются рак легкого (РЛ) и рак молочной железы у женщин (РМЖ), занимающие соответственно первое и пятое места среди причин смерти от злокачественных новообразований [1].

Важным шагом для снижения смертности от рака является улучшение распознавания заболевания на ранней стадии [2]. Одним из основных способов диагностики рака являются гистологическое исследование и низкодозовая компьютерная томография (КТ).

При традиционной диагностике врач просматривает снимки КТ и гистологические изображения, что требует большого количества времени. Для упрощения этого процесса могут использоваться алгоритмы классификации, которые выдают вероятности принадлежности изображений к какому-либо классу [3]. Иными словами, это прогностические модели, определяющие вероятность наличия заболевания.

Важное свойство прогностической модели – корреляция между предсказанными вероятностями и наблюдаемыми результатами [2]. Для улучшения этого параметра применяется калибровка, стандартные методы которой корректируют вероятности с позиции теории вероятностей [3–7]. Однако это лишь один из способов описания неопределенности величины [8]. Мы предлагаем калибровку вероятностей бинарных классификаторов с применением теории нечетких множеств.

Как и теория вероятностей, теория нечетких множеств описывает неопределенность. Отличие в том, что первая передает информацию о частоте возникновения события; а вторая – описывает степень принадлежности объекта ко множеству с неточно определенными свойствами [8].

В медицине теория нечетких множеств используется для решения ряда задач: анализа диабетической нейропатии и выявления ранней диабетической ретинопатии, улучшения процессов принятия решений при лучевой терапии, контроля гипертонии во время анестезии, визуализации нервных волокон в головном мозге человека [8, 9].

При этом теория нечетких множеств применяется непосредственно к алгоритмам анализа и обработки данных: в предварительной обработке

[1]*ФГАОУ ВО Первый МГМУ им. И.М. Сеченова, Ресурсный центр "Медицинский Сеченовский Предуниверсарий", Москва, Россия*
*****E-mail: olga.a.filimonova@gmail.com*

изображений [10] и сегментации снимков маммографии для выявления РМЖ [11]; в схеме выбора признаков для идентификации биомаркеров РЛ, РМЖ и рака толстой кишки [12].

Недостаток этих методов – отсутствие универсальности. Мы применяем теорию нечетких множеств напрямую к вероятностям, что позволяет использовать наш алгоритм при работе с любой прогностической моделью, которая выдает вероятности. Существующая калибровка вероятностей, основанная на теории нечетких множеств, базируется на нечеткой модели Такаги–Сугено (Takagi–Sugeno) [13]. Мы представляем другой подход, объединяющий теорию вероятностей и теорию нечетких множеств.

2. ОБОЗНАЧЕНИЯ И ОПРЕДЕЛЕНИЯ

Определим пространство элементарных исходов Ω: пусть $\Omega = \{M, 1 - M\}$, где событие M – постановка диагноза рака любой стадии. Введем событие N – постановка диагноза рака любой стадии, кроме первой. То есть, $N \subseteq M$. Обозначим вероятность наступления события M и события N как $P(M)$ и $P(N)$ соответственно. Тогда из условия $N \subseteq M$ следует справедливость неравенства $P(M) \geqslant P(N)$ [14]. Так как значения $P(M)$ и $P(N)$ неотрицательные, то при умножении обеих сторон неравенства $P(N) \leqslant 1$ на $P(M)$, получим $P(M) \geqslant P(N) \cdot P(M)$.

Описанное является концепцией из теории вероятностей. Мы рассмотрим ситуацию с позиции теории нечетких множеств.

Нечеткое множество – это набор объектов без четкой границы принадлежности [15]. Этот принцип может быть использован в медицине в связи с тем, что диагностика заболеваний сопряжена с несколькими уровнями неопределенности и неточности [9]. Например, нормальная клетка превращается в раковую в ходе неконтролируемых мутаций [16]. Следовательно, невозможно определить точный момент возникновения рака. Значит, множество людей, у которых может быть диагностирован рак, можно рассматривать как нечеткое множество.

Для применения теории нечетких множеств введем обозначения. Положим, $\hat{X}$ – множество всех данных пациентов: $\hat{X} = \{\hat{x}_i \mid i \in [1, K]\}$, где K – количество пациентов. Пусть нечеткое множество $\tilde{M}$ – множество пациентов, у которых “ нечетко” диагностировали любую стадию рака, а нечеткое множество $\tilde{N}$ – множество пациентов, у которых “ нечетко” диагностировали любую стадию рака кроме первой. То есть, существуют наборы упорядоченных пар таких, что $\tilde{M} = \{(\hat{x}, \mu_{\tilde{M}}(\hat{x})) \mid \hat{x} \in \hat{X}\}$ и $\tilde{N} = \{(\hat{x}, \mu_{\tilde{N}}(\hat{x})) \mid \hat{x} \in \hat{X}\}$, где $\mu_{\tilde{M}}(\hat{x})$ и $\mu_{\tilde{N}}(\hat{x})$ – соответствующие функции принадлежности. Чем выше значение функции принадлежности, тем выше степень принадлежности элемента к множеству [8].

Было предпринято несколько попыток установить связь между функцией принадлежности и классической вероятностью [8]. Мы воспользуемся следующим определением функции принадлежности.

Определение 1. Пусть X – множество всех действительных чисел, лежащих на отрезке от 0 до 1. Введем нечеткое множество A. То есть, существуют наборы упорядоченных пар таких, что $A = \{(x, \mu_A(x)) \mid x \in X\}$.

Функция принадлежности элемента x множеству A, которое “нечетко” больше чем параметр a, определяется как

$$\mu_{\geqslant a}(x) := \frac{P(t \leqslant x)}{P(t \leqslant a)} \quad \text{для} \quad x < a, \qquad := 1 \quad \text{для} \quad x \geqslant a, \tag{1}$$

где $t \in X$, $P(t \leqslant x)$ – вероятность, что какой-либо элемент t меньше или равен элементу x, $P(t \leqslant a)$ – вероятность, что какой-либо элемент t меньше или равен a, где параметр a – действительное число, с которым сравниваются элементы t и x [17].

Иными словами, мы включаем в множество A все элементы, которые больше или равны параметру a. Некоторый элемент x, меньший чем параметр a, также может быть включен в множество A со степенью принадлежности $\mu_{\geqslant a}(x)$.

3. МАТЕМАТИЧЕСКАЯ МОДЕЛЬ КАЛИБРОВКИ ВЕРОЯТНОСТЕЙ

Рассмотрим два бинарных классификатора, реализованных любыми способами. Пусть первый классификатор определяет пациентов, у которых или не обнаружен рак, или обнаружен рак любой стадии. Набор вероятностей, который выдает этот классификатор, обозначим как множество вероятностей $P_M = \{P_M(\hat{x}) \mid \hat{x} \in \hat{X}, \text{ и } 0 \leqslant P_M(\hat{x}) \leqslant 1\}$.

Пусть второй классификатор определяет пациентов, у которых или не обнаружен рак, или обнаружен рак любой стадии, кроме первой. Набор вероятностей, который выдает этот классификатор, обозначим как множество вероятностей $P_N = \{P_N(\hat{x}) \mid \hat{x} \in \hat{X}, \text{ и } 0 \leqslant P_N(\hat{x}) \leqslant 1\}$.

Для P_M и P_N справедливы соотношения, введенные для $P(M)$ и $P(N)$:

$$P_M \geqslant P_N, \tag{2}$$

$$P_M \geqslant P_N \cdot P_M. \tag{3}$$

Применим определение (1) к результатам первого классификатора. Положим параметр a равным единице. Так как все элементы множества P_M принадлежат отрезку от 0 до 1, то для любого элемента $t' \in P_M$, меньшего или равного любому элементу $P_M(\hat{x})$, знаменатель $P(t' \leqslant 1)$ равен единице.

Рассмотрим числитель $P(t' \leqslant P_M(\hat{x}))$. Отметим на отрезке от 0 до 1 точки, соответствующие значениям t' и $P_M(\hat{x})$ в случае $t' \leqslant P_M(\hat{x})$. Тогда, как изображено на рис. 1, вероятность $P(t' \leqslant P_M(\hat{x}))$ будет равна отношению длины отрезка $l_{t'}$ к длине отрезка $l_{\hat{x}}$. Для упрощения в качестве длины отрезка $l_{t'}$ возьмем квадрат значения вероятности $P_M(\hat{x})$. В свою очередь длина отрезка $l_{\hat{x}}$ равна значению $P_M(\hat{x})$.

В итоге функция принадлежности – которую обозначим $\mu_{\tilde{M}}(\hat{x})$ – элемента $\hat{x}$ множеству пациентов, у которых "нечетко" диагностировали любую стадию рака, определяется

$$\mu_{\tilde{M}}(\hat{x}) = \frac{P_M^2(\hat{x})}{P_M(\hat{x})} = P_M(\hat{x}).$$

Аналогично получим функцию принадлежности $\mu_{\tilde{N}}(\hat{x})$ множества пациентов, у которых " нечетко" диагностировали любую стадию рака, кроме первой

$$\mu_{\tilde{N}}(\hat{x}) = P_N(\hat{x}).$$

Теперь неравенство (2) перепишем в виде

$$P_M \geqslant \mu_{\tilde{N}}(\hat{x}), \tag{4}$$

а неравенство (3) можно представить как $P(M) \geqslant \mu_{\tilde{N}}(\hat{x}) \cdot \mu_{\tilde{M}}(\hat{x})$.

С другой стороны, произведение функций принадлежности $\mu_{\tilde{N}}(x)$ и $\mu_{\tilde{M}}(x)$ можно записать в виде $\max\{\mu_{\tilde{N}}(x), \mu_{\tilde{M}}(x), \alpha\} \cdot \mu_{\tilde{N}\cap\tilde{M}}(x)$, где параметр α такое действительное число, что $\mu_{\tilde{N}}(x), \mu_{\tilde{M}}(x) \in [0, \alpha]$ [8]. Для упрощения положим, что параметр α равен значению $\mu_{\tilde{M}}(x)$ или $\mu_{\tilde{N}}(x)$. Степень принадлежности $\mu_{\tilde{N}\cap\tilde{M}}(x)$ может быть записана в виде $\min\{\mu_{\tilde{N}}(x), \mu_{\tilde{M}}(x)\}$ [8]. Тогда неравенство $P(M) \geqslant \mu_{\tilde{N}}(x) \cdot \mu_{\tilde{M}}(x)$ принимает вид

$$P_M \geqslant \max\{\mu_{\tilde{N}}(x), \mu_{\tilde{M}}(x)\} \cdot \min\{\mu_{\tilde{N}}(x), \mu_{\tilde{M}}(x)\}. \tag{5}$$

Заменим в неравенстве (4) один из множителей на логистическую функцию [8]. Так как область значения логистической функции принадлежит отрезку от 0 до 1, то для удовлетворения условия (2) заменим множитель $\min\{\mu_{\tilde{N}}(x), \mu_{\tilde{M}}(x)\}$. В итоге неравенство (5) можно переписать как

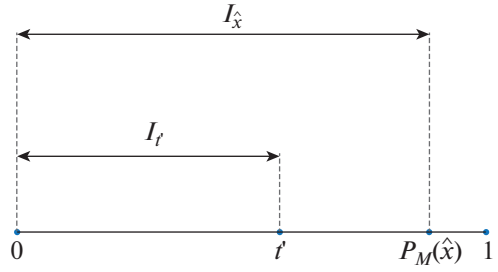

Рис. 1. Геометрическое представление $P(t' \leqslant P_M(\hat{x}))$.

$$P_M \geqslant \frac{\max\{\mu_{\tilde{N}}(x), \mu_{\tilde{M}}(x)\}}{1 + \exp\left(E \cdot \min\{\mu_{\tilde{N}}(x), \mu_{\tilde{M}}(x)\} + B\right)},$$

где E и B – действительные числа, различные в каждом конкретном случае.

В ходе диагностики может быть найдена любая стадия рака, но при раннем выявлении необходимо учитывать минимальный обоснованный риск наличия онкологического процесса. Поэтому в качестве функции калибровки возьмем равенство. Таким образом, для i-го пациента вероятность обнаружения θ_i рака любой стадии с учетом приоритета ранней диагностики калибруется по формуле

$$\theta_i = \frac{\max\{P_N(\hat{x}_i), P_M(\hat{x}_i)\}}{1 + \exp\left(E \cdot \min\{P_N(\hat{x}_i), P_M(\hat{x}_i)\} + B\right)}. \tag{6}$$

Верхняя граница $\mu_{\tilde{M}}(x)$ и $\mu_{\tilde{N}}(x)$ равна единице, и множества $\tilde{M}$ и $\tilde{N}$ являются нормальными [18]. Тогда функции принадлежности дополнений $\tilde{M}$ и $\tilde{N}$ определяется как $1 - \mu_{\tilde{M}}(x)$ и $1 - \mu_{\tilde{N}}(x)$ соответственно [8]. Применим рассуждения, аналогичные вышеописанным, к вероятностям $1 - P_M$ и $1 - P_N$ и получим

$$1 - \theta_i = \frac{\min\{1 - P_N(\hat{x}_i), 1 - P_M(\hat{x}_i)\}}{1 + \exp\left(C \cdot \max\{1 - P_N(\hat{x}_i), 1 - P_M(\hat{x}_i)\} + D\right)}, \tag{7}$$

где C и D – действительные числа, различные в каждом конкретном случае.

Вычтем выражение (7) из выражения (6). Это даст окончательную формулу для нашей модели калибровки вероятностей

$$\theta_i = \frac{1}{2}\left(\frac{\max\{P_N(\hat{x}_i), P_M(\hat{x}_i)\}}{1 + \exp\left(E \cdot \min\{P_N(\hat{x}_i), P_M(\hat{x}_i)\} + B\right)} - \right.$$
$$\left. - \frac{\min\{1 - P_N(\hat{x}_i), 1 - P_M(\hat{x}_i)\}}{1 + \exp\left(C \cdot \max\{1 - P_N(\hat{x}_i), 1 - P_M(\hat{x}_i)\} + D\right)} + 1\right).$$

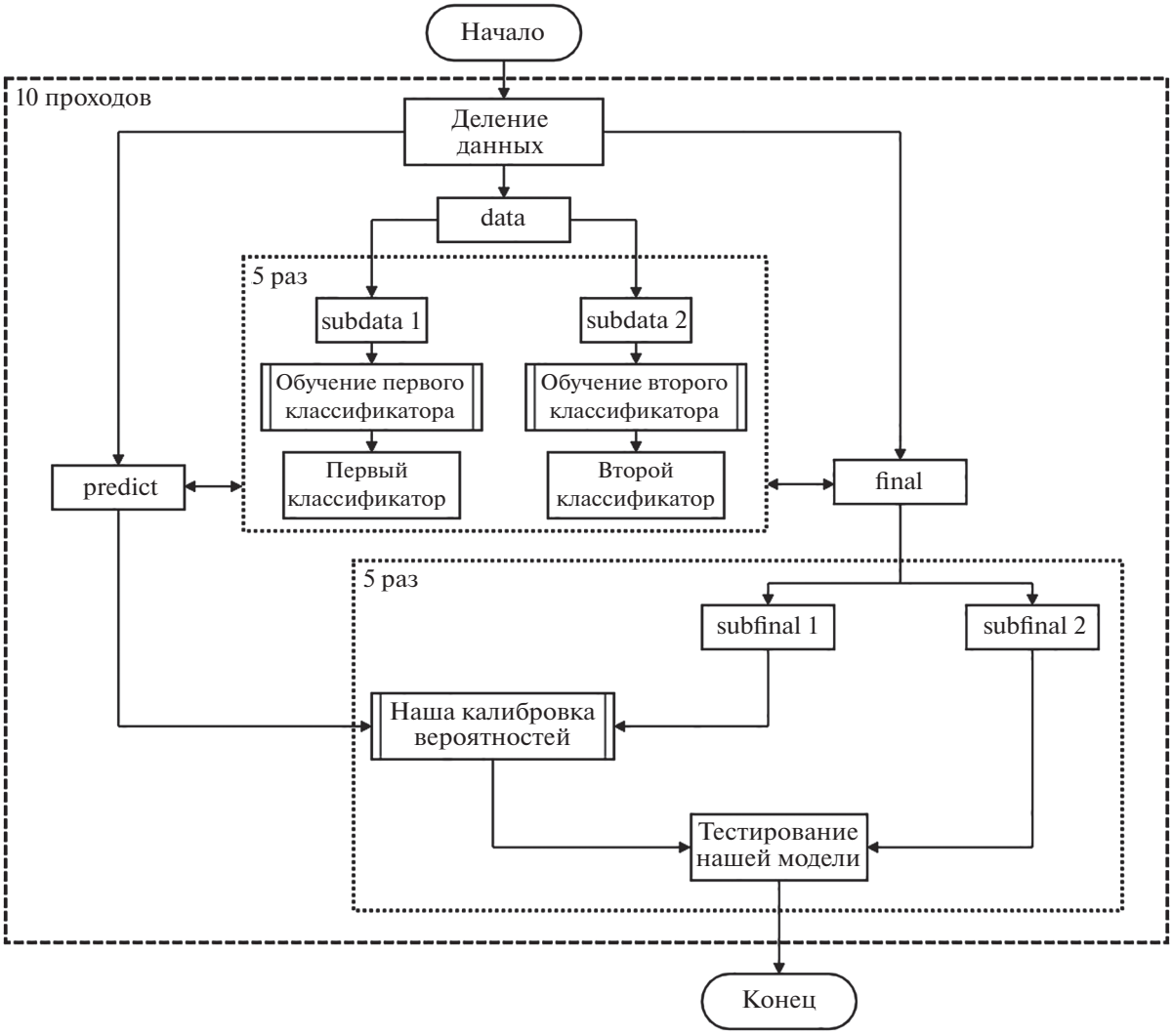

Рис. 2. Схема проверки нашего алгоритма.

4. ДАТАСЕТЫ

Для демонстрации работы нашего алгоритма независимо рассмотрены два случая: диагностика РМЖ по гистологическим изображениям и диагностика РЛ по снимкам КТ во фронтальной проекции.

В первом случае затронута проблема обучения на небольшом наборе данных. Мы взяли 400 гистологических изображений молочной железы [19] и поделили их на три класса: множество здоровых пациентов – 200 изображений или без патологий, или с выявленными доброкачественными новообразованиями; множество пациентов с ранней стадией рака – 100 изображений с выявленным преинвазивным раком (Carcinoma in situ); множество пациентов с поздней стадией рака – 100 изображений с выявленной инвазивной карциномой (Invasive carcinoma).

Во втором случае использованы два датасета из архива изображений рака (The Cancer Imaging Archive) [20, 21]. По статистике процент выявления РЛ в ходе проведения КТ составляет 0.42–2.1% [22], и задача осложняется несбалансированными данными. Мы оставили это соотношение на всех этапах проверки. Было взято 2159 снимков, среди которых у 2113 – не обнаружено рака, у 20 диагностирована первая стадия рака, и у 26 – вторая, третья или четвертая стадии.

5. СХЕМА ПРОВЕРКИ НАШЕГО АЛГОРИТМА

Во избежание переобучения проверка нашего алгоритма проводилась по схеме, указанной на рис. 2.

В начале каждого прохода все данные случайным образом разбивались на три части, две из ко-

Таблица 1. Описание классификаторов и объем выборок

Форма рака	Архитектура классификаторов	*subdata 1*	*subdata 2*	*predict*	*subfinal 1*	*subfinal 2*
РМЖ	Нейронные сети на основе предобученной модели *Inception v3* [23]. Использовались методы расширения исходного набора данных (Data Augmentation): поворот на 180°, сдвиг изображения, смещение по ширине и по высоте, отражение по горизонтали и по вертикали	120	160	160	40	40
РЛ	Нейронные сети с искусственным расширением минорного класса	1203 1042	1214 1052	405 567	270 270	270 270

торых (*predict* и *final*) не участвовали в обучении бинарных классификаторов. Эти классификаторы используются для получения вероятностей, предназначенных для калибровки. Первый классификатор определяет пациентов, у которых или не обнаружен рак, или обнаружен рак любой стадии. Второй – у которых или не обнаружен рак, или обнаружен рак любой стадии, кроме первой. Во время каждого прохода классификаторы обучались пять раз на заново случайно разделенных *subdata 1* и *subdata 2*.

Далее определялись параметры E, B, C, D. В каждом проходе наша модель пять раз обучалась на вероятностях, предсказанных классификаторами для *predict*. При этом каждый раз валидационная выборка *subfinal 1* и тестовая выборка *subfinal 2* разбивались заново случайным образом. Во время работы нашего алгоритма калибровались вероятности первого классификатора с учетом приоритета ранней диагностики.

В итоге наша модель была протестирована пятьдесят раз на нигде ранее не задействованных данных. Мы провели одно тестирование для РМЖ и два тестирования с разным объемом выборок для РЛ (табл. 1).

6. ИСПОЛЬЗУЕМЫЕ МЕТРИКИ И МЕТОДЫ КАЛИБРОВКИ ВЕРОЯТНОСТЕЙ

В качестве метрик взяты те, которые не зависят от порога принятия решения и рассматривают только предсказанные вероятности. Это логарифмическая потеря (Log Loss)

$$\text{Log Loss} = -\frac{1}{l}\cdot\sum_{i=1}^{l}\left(y_i\cdot\log(\hat{y}_i)+(1-y_i)\cdot\log(1-\hat{y}_i)\right),$$

где $\hat{y}$ – предсказанная вероятность для i-го пациента, y – истинная метка класса i-го пациента, l – размер выборки.

Еще одним показателем является оценка Брайера (BS) [2]

$$\text{BS} = \frac{1}{l}\sum_{i=1}^{l}(\hat{y}_i - y_i)^2.$$

Информативную картину о качестве предсказанных вероятностей может показать кривая Precision-Recall [24], и в качестве еще одной метрики взята площадь под этой кривой (AUC-PR).

Метрики считались после каждого прохода, после чего брались их относительные изменения (ОИ):

$$\text{ОИ} = \frac{M_c - M_o}{M_o}\cdot 100\%,$$

где M_c – метрика, посчитанная для откалиброванных вероятностей; M_o – метрика, посчитанная для исходных вероятностей. Таким образом, для каждого случая получено по пятьдесят показателей ОИ.

Калибровка вероятностей осуществлялась следующими способами: изотонической регрессией [4], моделью Платта [5], гистограммной калибровкой (Histogram Binning) [6], с помощью метода BBQ (Binning into Quantiles) [7] и нашей моделью.

Все способы, кроме нашего, реализованы посредством библиотек *scikit-learn* и *net:cal* [25]. В нашей модели параметры E, B, C, D оценивались с помощью метода максимального правдоподобия. Во избежание переобучения завершение обучения происходило тогда, когда логарифмическая потеря, посчитанная для *subdata 1*, становилась меньше, чем – посчитанная для *predict*.

7. РЕЗУЛЬТАТЫ

Во всех случаях изотоническая регрессия, гистограммная калибровка и метод BBQ привели к переобучению. Они занижали вероятности, и те не превышали 0,5.

В случае РМЖ модель Платта оставила метрики без изменений. Наш алгоритм улучшил все по-

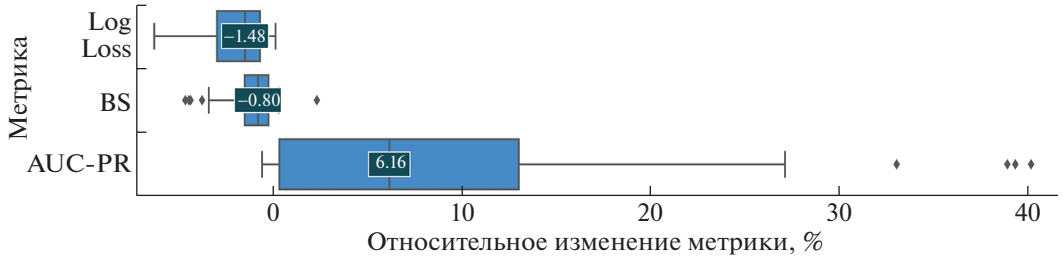

Рис. 3. Диграмма размаха ОИ после работы нашего метода в случае РМЖ. Среднеквадратичные отклонения ОИ: *Log Loss* – 1.59%; *BS* – 1.37%; *AUC-PR* – 11.31%.

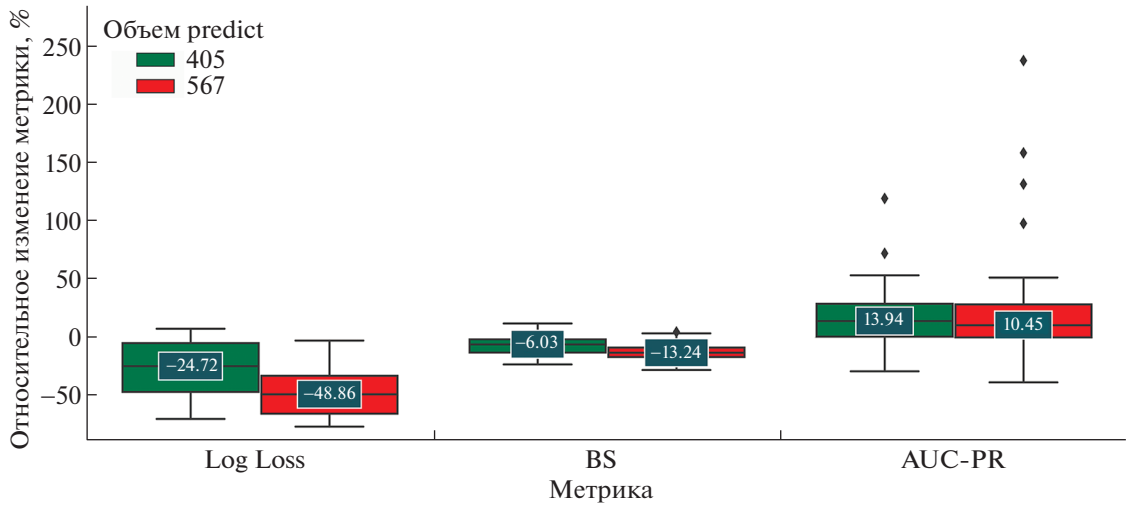

Рис. 4. Диаграмма размаха ОИ, полученная после работы нашего метода в случае РЛ. Среднеквадратичные отклонения ОИ для 405 снимков: *Log Loss* – 23.79%; *BS* – 8.67%; *AUC-PR* – 23.68%; для 567 снимков: *Log Loss* – 50.68%; *BS* – 7.13%; *AUC-PR* – 47.32%.

казатели, что продемонстрировано на рис. 3. В случае РЛ вне зависимости от объема выборок модель Платта не повлияла на оценку Брайера и не внесла значимых изменений: медианы всех ОИ равны нулю. Остальные показатели представлены в табл. 2.

Наш метод калибровки, как показано на рис. 4, снова улучшил все показатели.

8. ЗАКЛЮЧЕНИЕ

В этой статье мы представили новый метод калибровки вероятностей бинарных классификаторов, основанный на теории нечетких множеств. В отличие от других алгоритмов калибровки наша модель эффективна на небольших датасетах и сильно несбалансированных данных. По мере увеличения выборки, на которой обучается наш алгоритм, прослеживается тенденция повышения эффективности в улучшении логарифмической потери (от 1.48 до 48.86%) и оценки Брайера (от 0.8 до 13.24%). Рост площади под кривой Precision-Recall менее зависим от объема обучающей выборки и демонстрирует эффективность (увеличение на 6.16%) даже на небольшом наборе данных.

Таблица 2. Среднеквадратичные отклонения ОИ в случае РЛ, посчитанных после работы модели Платта

predict	*Log Loss*	*AUC-PR*
405	0.16%	6.36%
564	0.2%	5.16%

Несмотря на то что демонстрация работы проведена на примере улучшения ранней диагностики РМЖ и РЛ, наш метод калибровки можно использовать для улучшения ранней диагностики любых прогрессирующих заболеваний.

Сферу применения нашего алгоритма можно расширить на любую область, где присутствуют последовательные события без четкой границы

принадлежности. Например, кредитоспособность является субъективным параметром и может быть описана в терминах теории нечетких множеств. Вследствие чего нашу модель можно применять для улучшения оценки риска краткосрочных и долгосрочных кредитов.

СПИСОК ЛИТЕРАТУРЫ

1. *Sung H., Ferlay J., Siegel R. L., Laversanne M. et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. In CA: A Cancer Journal for Clinicians. 2021. V. 71. Issue 3. P. 209–249. Wiley. https://doi.org/10.3322/caac.21660
2. *Steyerberg E.W.* Clinical Prediction Models. A practical approach to development, validation and updating. New York: Springer, 2009. 508 c.
3. *Böken B.* On the appropriateness of Platt scaling in classifier calibration. In Information Systems. 2021. V. 95. P. 101641. Elsevier BV. https://doi.org/10.1016/j.is.2020.101641
4. *Chakravarti N.* Isotonic Median Regression: A Linear Programming Approach. Mathematics of Operations Research. 1989. V. 14. № 2. P. 303–308. http://www.jstor.org/stable/3689709
5. *Platt John*. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Adv. Large Margin Classif. 2000. 10.
6. *Bianca Zadrozny, Charles Elkan*. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In Proceedings of the Eighteenth International Conference on Machine Learning. 2001. P. 609–616.
7. *Naeini M.P., Cooper G.F., Hauskrecht M.* Obtaining Well Calibrated Probabilities Using Bayesian Binning. Proc AAAI Conf Artif Intell. 2015 Jan; 2015. P. 2901–2907. PMID: 25927013; PMCID: PMC4410090.
8. *Zimmermann H.-J.* Fuzzy Set Theory – and Its Applications. Springer Dordrecht, 2001. 514 c.
9. *Torres A., Nieto J.J.* Fuzzy Logic in Medicine and Bioinformatics. In Journal of Biomedicine and Biotechnology. 2006. V. 2006. P. 1–7. Hindawi Limited. https://doi.org/10.1155/jbb/2006/91908
10. *Hassanien A.* Fuzzy rough sets hybrid scheme for breast cancer detection. In Image and Vision Computing. 2007. V. 25. Issue 2. P. 172–183. Elsevier BV. https://doi.org/10.1016/j.imavis.2006.01.026
11. *Ghosh S.K., Mitra A., Ghosh A.* A novel intuitionistic fuzzy soft set entrenched mammogram segmentation under Multigranulation approximation for breast cancer detection in early stages. In Expert Systems with Applications. 2021. V. 169. P. 114329. Elsevier BV. https://doi.org/10.1016/j.eswa.2020.114329
12. *Ghosh S.K., Ghosh A., Bhattacharyya S.* Recognition of cancer mediating biomarkers using rough approximations enabled intuitionistic fuzzy soft sets based similarity measure. In Applied Soft Computing. 2022. V. 124. P. 109052. Elsevier BV. https://doi.org/10.1016/j.asoc.2022.109052
13. *Wang N., Yao W., Zhao Y., Chen X.* Bayesian calibration of computer models based on Takagi–Sugeno fuzzy models. In Computer Methods in Applied Mechanics and Engineering. 2021. V. 378. P. 113724. Elsevier BV. https://doi.org/10.1016/j.cma.2021.113724
14. Теория вероятностей : учебник для втузов / Печинкин А.В., Тескин О.И., Цветкова Г.М. и др.; ред. Зарубин В.С., Крищенко А.П. 3-е изд., испр. М.: Изд-во МГТУ им. Н.Э. Баумана, 2004. 455 с.
15. *Sadegh-Zadeh K.* The Logic of Diagnosis. In Philosophy of Medicine. 2011. P. 357–424. Elsevier. https://doi.org/10.1016/b978-0-444-51787-6.50012-x
16. *Castaneda M., den Hollander, Kuburich N. et al.* Mechanisms of cancer metastasis. In Seminars in Cancer Biology. 2022. V. 87. P. 17–31. Elsevier BV. https://doi.org/10.1016/j.semcancer.2022.10.006
17. *Beliakov G.* Fuzzy sets and membership functions based on probabilities. In Information Sciences. 1996. V. 91. Issues 1–2. P. 95–111. Elsevier BV. https://doi.org/10.1016/0020-0255(95)00291-x
18. *Chen G., Pham T.T.* Introduction to fuzzy sets, fuzzy logic, and fuzzy control systems. CRC Press. 2000. 329 c. ISBN 0-8493-1658-8
19. *Aresta G., Araújo T., Kwok S. et al.* BACH: Grand challenge on breast cancer histology images. In Medical Image Analysis. 2019. V. 56. P. 122–139. Elsevier BV. https://doi.org/10.1016/j.media.2019.05.010
20. *Armato III, Samuel G., McLennan et al.* Data From LIDC-IDRI (Version 4) [dataset]. The Cancer Imaging Archive. 2015. https://doi.org/10.7937/K9/TCIA.2015.LO9QL9SX
21. National Lung Screening Trial Research Team. Data from the National Lung Screening Trial (NLST) (Version 3) [dataset]. The Cancer Imaging Archive. 2013. https://doi.org/10.7937/TCIA.HMQ8-J677
22. *Pinsky P.F.* Lung cancer screening with low-dose CT: a world-wide view. In Translational Lung Cancer Research. 2018. V. 7. Issue 3. P. 234–242. AME Publishing Company. https://doi.org/10.21037/tlcr.2018.05.12
23. *Szegedy C., Vanhoucke V. et al.* Rethinking the Inception Architecture for Computer Vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. https://doi.org/10.1109/cvpr.2016.308
24. *Davis J., Goadrich M.* The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning - ICML '06. the 23rd international conference. ACM Press. 2006. https://doi.org/10.1145/1143844.1143874
25. *Küppers A. et al.* Parametric and Multivariate Uncertainty Calibration for Regression and Object Detection. European Conference On Computer Vision (ECCV) Workshops. 2022. 10.

PROBABILITY CALIBRATION ON THE EXAMPLE OF IMPROVING EARLY CANCER DETECTION: A FUZZY SET THEORY APPROACH

O. A. Filimonova[a], A. G. Ovsyannikov[a], and N. V. Biryukova[a]

[a]I.M. Sechenov First Moscow State Medical University (Sechenov University), Moscow, Russian Federation

Presented by Academician of the RAS A.I. Avetisyan

Cancer is the leading cause of death before the age of 70 years. Cancer mortality is reduced by early detection. To improve the early diagnosis of cancer, we propose a novel probability calibration method based on a fuzzy set theory. Our model for binary classification was tested on the detection of female breast cancer and lung cancer. The first case is complicated by a small data set problem, while the second – by highly imbalanced data. In both cases, our probability calibration method improved Log Loss (the best result is improved on 48.86%), Brier score (the best result is improved on 13.24%), and the area under the PR curve (the best result is improved on 13.94%). The application area of our algorithm can be extended to any progressive diseases and events without a clearly defined boundary.

Keywords: probability calibration, fuzzy set theory, binary classification, early diagnosis of diseases