УДК 004.85 **DOI** 10.15622/ia.24.4.5

А.Г. Лосев, И.Е. Попов, А.С. Резникова

ПОВЫШЕНИЕ РЕПРЕЗЕНТАТИВНОСТИ ОБУЧАЮЩЕГО НАБОРА ДАННЫХ ЗА СЧЕТ ПРОСТРАНСТВЕННОЙ БАЛАНСИРОВКИ

Лосев А.Г., Попов И.Е., Резникова А.С. **Повышение репрезентативности обучающего набора данных за счет пространственной балансировки.**

Аннотация. В работе исследуются некоторые проблемы машинного обучения, связанные с анализом данных малого объема. А именно, решается задача формирования обучающего набора, обладающего высокой репрезентативностью в задачах многоклассовой классификации. Известно, что межклассовая и внутриклассовая несбалансированность, свойственная малым наборам данных, негативно влияет на качество работы алгоритмов. С целью их устранения в машинном обучении разработаны некоторые методы синтеза данных, дополняющие имеющийся набор и уравнивающие количество объектов каждого из классов. Однако такие методы не всегда решают проблему недостаточной репрезентативности. В данной статье предлагается метод построения репрезентативного обучающего набора данных за счет задания распределения, в наибольшей степени соответствующего действительности. Распределение формируется по каждому признаку в пределах информативных областей. Информативные области содержат характерные значения признаков, наиболее значимые для различения классов объектов. Предложенный метод построения областей основан на идее поэтапного расширения, сопровождающегося ростом информативности областей. При этом под информативностью понимается мера, отражающая то, насколько хорошо с помощью рассматриваемой области можно разделить объекты различных классов. С целью формирования дополняющего набора данных разработан метод генерации. В результате его применения дополняющий набор данных объединяется с исходным и образует в информативной области указанное распределение. Это распределение может быть задано либо на основе экспертных знаний о предметной области, если известно истинное распределение, либо получено в результате вычислительных экспериментов, направленных на поиск наиболее эффективного варианта. Применимость метода продемонстрирована на примере решения задачи определения уровня температурных аномалий молочных желёз. Показано, что для рассматриваемых температурных признаков характерно нормальное распределение. Повышение репрезентативности обучающего набора позволило обучить классический алгоритм классификации – логистическую регрессию – с точностью, сопоставимой с многослойной нейронной сетью. Такой подход к формированию обучающего набора данных открывает возможность создания более прозрачных и интерпретируемых систем искусственного интеллекта.

Ключевые слова: машинное обучение, малые наборы данных, репрезентативность данных, синтез данных, нейронные сети, логистическая регрессия.

1. Введение. Одной из ключевых проблем машинного обучения является разработка методов правильного формирования обучающего набора данных [1–3]. Особенно остро данный вопрос стоит при работе с "малыми" данными, хотя и в интеллектуальном анализе "больших" данных указанная задача тоже является актуальной.

В подавляющем большинстве исследований формирование обучающей выборки производится на основе "натурных" данных, т.е. данных, полученных в результате каких-то измерений и т.п. При этом известно, что натурные данные могут обладать следующими недостатками.

1. Наличие ошибок результатов измерений (пропуски, некорректные измерения и т.п. [4])

2. Наличие ошибок в метках [5,6]. Например, по мнению некоторых специалистов, в медицинских картах пациентов может содержаться более 30 процентов неточных диагнозов, что порождает неверные метки [7].

3. Наличие персональных данных. Последнее может породить трудности с распространением набора данных и валидацией полученных результатов.

4. Малый объем набора данных. В зависимости от предметной области сложность составления набора данных варьируется. Тем не менее исследователи сталкиваются с проблемой недостатка данных во многих дисциплинах [8–14]. В то время как большие данные подразумевают терабайты или петабайты объема, малые наборы данных не столь велики в объеме и обычно измеряются сотнями или тысячами записей. Однако понятие малых данных определяется не только объемом записей. Так, например, достаточно много данных может быть в национальной переписи населения, тем не менее такой набор данных не может быть определен понятием большие данные. Ключевые отличия малых наборов данных от больших заключаются в том, что малые данные накапливаются достаточно медленно, имеют однообразную структуру, низкую детализацию и поступают из одного или небольшого числа источников. В химии, физике и материаловедении трудности с получением больших выборок данных обусловлены высокими затратами и сложностью проведения экспериментов. Кроме того исследователи сталкиваются с малыми наборами данных и при решении задача по обработке естественного языка в области программной инженерии (классификация отзывов о работе приложений, классификация комментариев к коду и т.д.). При этом объемы используемых наборов данных варьируются по объему в диапазоне от 300 до 65 000 экземпляров. В медицине и биологии редкость некоторых заболеваний и состояний также приводит к недостаточному количеству клинических случаев для обучения моделей машинного обучения и проведения статистически значимых исследований. Однако несмотря на свою редкость, эти явления могут существенно влиять на здоровье пациентов, поэтому их изучение становится необходимым. Так в одном из исследований, посвященном решению задачи прогнозирования

развития сахарного диабета, итак изначально малое количество равное 3000 записей, было снижено до 162 единиц после выбора полностью заполненных данных [38]. В исследовании по классификации МРТ снимков головного мозга количество записей тоже было мало и составило 380 единиц [8].

Следовательно, проблема малого набора данных охватывает многие сферы деятельности и имеет междисциплинарный характер. При этом работа с малым набором данных имеет свои особенности: высокая вероятность переобучения, повышенная чувствительность к шуму в данных, отсутствие разнообразия наблюдений, ограничения в проведении качественной валидации и тестирования.

5. Несбалансированность данных. Работа с наборами данных малого объема практически всегда обостряет проблему их несбалансированности. В ситуациях, когда количество объектов одного или нескольких классов значительно превосходит количество объектов других классов, производительность и точность большинства стандартных алгоритмов классификации снижается [15]. По умолчанию эти алгоритмы рассчитаны на равномерное распределение данных по классам. Однако, при решении различных задач, модели машинного обучения должны оставаться непредвзятыми и одинаково эффективно определять объекты как мажоритарного, так и миноритарного классов. Кроме проблемы несбалансированного соотношения данных различных классов существует также проблема внутриклассовой несбалансированности, проявляющаяся в неравномерном распределении данных внутри каждого отдельного класса. Таким образом, важность учета расположения объектов в признаковом пространстве и равнозначности их оценки становится особенно значимой.

Обозначенные проблемы в известной степени можно решать на основе синтезирования набора данных различными методами. Например, статистическими методами в какой-то мере можно пытаться решить проблемы № 3, № 4 и № 5, а проблемы № 1 и № 2 стандартными методами можно решить до степени распространенности соответствующих ошибок в обучающем наборе данных.

Весьма перспективным выглядит подход синтеза на основе создания компьютерных моделей [16]. Однако такой подход не нацелен напрямую на решение проблемы несбалансированности данных. Компьютерные модели работают с теми данными, которые им предоставляются, и если эти данные изначально несбалансированы, то набор данных, полученный в результате моделирования, также может сохранить этот недостаток.

Другое решение проблем, связанных с малыми данными и несбалансированностью выборок, заключается в выполнении стандартной количественной балансировки. При таком подходе на основе статистического анализа создаются дополнительные объекты, которые имитируют свойства исходного набора данных. Например, при использовании алгоритма SMOTE, для каждого объекта x_i из класса меньшинства новый объект синтезируется вдоль отрезков, соединяющих его с соседями x_{ni} из этого же класса [17]:

$$x_{new} = x_i + \lambda(x_i - x_{ni}),$$

здесь λ – некоторое случайное число, $\lambda \in [0; 1]$.

В случае использования алгоритма ADASYN синтезация производится преимущественно для тех объектов миноритарного класса, которые располагаются в области с повышенной концентрацией объектов мажоритарного класса. Это позволяет увеличить плотность объектов миноритарного класса именно в областях наибольшей неопределенности.

Таким образом, синтез данных является мощным подходом, способствующим улучшению качества выборок и решению некоторых проблем, связанных с их формированием.

Тем не менее, синтезирование набора данных порождает свои проблемы, требующие дополнительного внимания. Целью синтеза должно быть не просто увеличение объема данных за счет создания подобных данных (что наблюдается при количественной балансировке, например, при использовании SMOTE и ADASYN), а именно улучшение их репрезентативности. Сравнение различных методов синтеза приведено в таблице 1.

Таким образом, такие методы синтезации как SMOTE и ADASYN демонстрируют ограниченные возможности, создавая объекты похожие на уже имеющиеся в данных, основываясь на подходе поиска ближайшего соседа, что может также привести к увеличению количества объектов-выбросов. Особенностью Gaussian Copula Synthesizer является учет зависимостей между объектами на основе всего набора данных, однако отсутствие приоритета синтезирования в определенной области может привести к сохранению или даже усилению шума. Предлагаемый в данной работе метод пространственная балансировка, напротив, целенаправленно синтезирует объекты в значимых областях, что способствует устранению шума и выявлению новых закономерностей в данных и существенно повышает качество и репрезентативность данных.

Таблица 1. Сравнение методов синтезации

SMOTE	ADASYN	Gaussian Copula Synthesizer	Пространственная балансировка
Создание синтетических объектов путём линейной интерполяции между соседними объектами миноритарного класса (синтез на локальном уровне ближайших соседей).	Адаптивная синтезация объектов в области трудноклассифицируемых объектов миноритарного класса (адаптивный синтез на локальном уровне ближайших соседей).	Преобразование исходных данных в заданное распределение. Синтез с учетом зависимостей между характеристиками объектов на основе всего набора данных.	Синтез данных в определенных значимых с точки зрения классификации областях путем задания нужного распределения
Объекты миноритарного класса, на основе которых осуществляется синтез, выбираются случайным равномерным образом, поэтому исходное распределение данных практически не меняется.	Увеличение числа объектов на границе классов и в проблемных зонах (незначительное повышение репрезентативности).	Синтез по всей области значений, характеризующих состояние объектов.	Синтез в заданной области значений, способствующий выявлению новых закономерностей с учетом предметной области (предметно обоснованный выбор области синтеза, способствующий повышению репрезентативности данных).
Возможно усиление шума при создании синтетических объектов возле объектов-выбросов			Снижение влияния шума за счет учета логики размещения объектов согласно предметной области.

Репрезентативность – это соответствие распределения признаков объектов выборки аналогичному истинному распределению в генеральной совокупности. Репрезентативность характеризуется такими понятиями, как полнота и разнообразие данных [18, 19] и свидетельствует о способности выборки достоверно отражать структуру и особенности исходной совокупности. При обнаружении недостаточно представленных областей исходного распределения выборка дополняется синтетическими данными, что позволяет существенно повысить её репрезентативность по сравнению с чистым набором натурных эмпирических данных. При этом оценка репрезентативности является нетривиальной задачей ввиду

того, что зачастую в задачах машинного обучения нам неизвестно истинное распределение данных. В работе оценка репрезентативности производится опосредованно через модели машинного обучения. В условиях, когда меняются только характеристики обучающего набора данных, а не модели, логично предполагать, что на изменения точности классификации в первую очередь повлияли изменения в данных и их качественных характеристиках. Также осуществляется оценка репрезентативности через энтропию распределения объектов в признаковом пространстве.

2. Термометрические данные как класс медицинских данных. Как уже отмечалось ранее, одной из распространенных сфер деятельности, где ведется работа преимущественно с малыми наборами данных, является медицина. При этом особое значение в медицине имеют термометрические данные, получаемые после проведения соответствующих обследований, например микроволновой радиотермометрии.

Микроволновая радиотермометрия (РТМ метод) – это метод диагностики различных заболеваний, основанный на измерении температурного поля $T(x, y, z)$ обследуемого органа, где x, y, z – координаты точек органа. Обследование РТМ методом является неинвазивной и совершенно безопасной для пациента процедурой. Обследование проводится в инфракрасном и микроволновом диапазоне. Инфракрасный диапазон измеряет температуру подкожного слоя, а микроволновый диапазон позволяет выявлять тепловые аномалии внутри организма на глубине нескольких сантиметров. Далее на основе полученных измерений выполняется построение, визуализация и анализ температурного поля. При визуализации данных, характерными особенностями могут быть горячие области, различие температурных полей парных органов. К примеру, при раке молочной железы, в скрининговых обследованиях которого данный метод применяется наиболее активно, у пациента наблюдается повышенная температура в области нахождения опухоли [22]. Диагностику молочных желез методом микроволоноовой радиотермометрии впервые начали проводить в 70-ых годах [23–26]. В итоге было установлено, что благодаря данному методу рак молочной железы можно косвенно выявлять на ранней стадии.

Однако РТМ метод находит применение не только в диагностике молочных желез, но и в выявлении других заболеваний, таких как патологии головного мозга, заболевания сердечно-сосудистой системы, вен нижних конечностей, болезни костно-мышечной системы и др. [27,28]. Для каждого типа заболевания существует своя схема обследования, согласно которой измерение температур производится в определенных

точках. Таким образом, термометрические данные представляют собой значения температур, измеренных в соответствующих точках. Структура данных в таком случае может быть задана таблицей. Колонки таблицы соответствуют номерам точек, в которых производились измерения. Строки - значениям измерений обследуемого органа. При этом схема обследования ограничена заданным количеством точек измерения, в таком случае функционалы, описываемые далее, заменяются функциями. Помимо этого термометрические данные также можно представить в виде изображения температурных полей, которые строятся путем цветового кодирования схематично обозначенной области измерения температур от среднего значения температуры (зеленый цвет) к максимальной (красный цвет) и минимальной (синий цвет).

Термометрические данные, получаемые при использовании метода РТМ, можно отнести к отдельному классу медицинских данных. Проведение анализа таких данных и постановка на их основе диагноза во многом субъективны и значимо зависят от опыта и искусства врача. Поэтому на практике анализ таких данных доступен только высококвалифицированным специалистам достаточно узких специальностей. Данное обстоятельство нивелирует потенциальную возможность широкого использования термометрического оборудования. Решение данной проблемы видится в использовании методов математического и компьютерного моделирования, возможностей современных информационных технологий и, прежде всего, методов машинного обучения. Применение машинного обучения позволит сделать РТМ-диагностику доступной для максимально широкого круга специалистов и пациентов. Интеллектуальный анализ данных упростит процесс анализа результатов обследования, что позволит расширить доступ к качественной диагностике при массовых скрининговых обследованиях и снизить нагрузку на опытных высококвалифицированных врачей, направив их усилия на решение наиболее сложных задач.

3. Предшествующие результаты. Важный вклад в развитие методов интеллектуального анализа термометрических данных внесли работы [16, 29–36], в которых были созданы математические модели описания диагностического состояния пациента, представлена оригинальная концепция построения признаковых пространств на основе дескриптивных моделей тепловых полей биологических тканей и реализованы эффективные алгоритмы классификации.

Первым шагом к созданию математических моделей для описания диагностического состояния пациента послужило использование не самих температурных данных, а функционалов, т.е. отображений из множества

температурных полей на числовую прямую. Отметим, что на практике производится аппроксимация температурных полей путем измерения их значений в определенных точках. Выбор точек осуществляется врачами на основе знаний предметной области. В итоге при аппроксимации поля конечным множеством температур функционалы заменяются функциями. Так, поля глубинных температур i-го пациента представимы в виде векторов:

$$T_c^{i,mw} = (t_{c,0}^{i,mw}, t_{c,1}^{i,mw}, ..., t_{c,9}^{i,mw});$$
$$T_p^{i,mw} = (t_{p,0}^{i,mw}, t_{p,1}^{i,mw}, ..., t_{p,9}^{i,mw}),$$

где первый нижний индекс обозначает температуры текущей ("c") или парной молочной железы ("p"), второй нижний индекс – номер точки согласно заранее заданной схеме обследования. Тогда значения функции 1 считается как разность среднего значения множества глубинных температур текущей молочной железы и парной. Для температур подкожного слоя вид векторов аналогичен:

$$T_c^{i,ir} = (t_{c,0}^{i,ir}, t_{c,1}^{i,ir}, ..., t_{c,9}^{i,ir});$$
$$T_p^{i,ir} = (t_{p,0}^{i,ir}, t_{p,1}^{i,ir}, ..., t_{p,9}^{i,ir}),$$

В таблицах 2 – 5 показан пример записи температур для нескольких молочных желез.

Таблица 2. Значения глубинных температур рассматриваемой молочной железы

i	$t_{c,0}^{i,mw}$	$t_{c,1}^{i,mw}$	$t_{c,2}^{i,mw}$	$t_{c,3}^{i,mw}$	$t_{c,4}^{i,mw}$	$t_{c,5}^{i,mw}$	$t_{c,6}^{i,mw}$	$t_{c,7}^{i,mw}$	$t_{c,8}^{i,mw}$	$t_{c,9}^{i,mw}$
0	35.6	35.4	35.9	35.8	35.6	35.6	35.3	35.3	35.6	35.6
1	33.6	33.8	33.2	33.2	33.0	34.0	33.2	33.8	34.3	34.1

Таблица 3. Значения температур подкожного слоя рассматриваемой молочной железы

i	$t_{c,0}^{i,ir}$	$t_{c,1}^{i,ir}$	$t_{c,2}^{i,ir}$	$t_{c,3}^{i,ir}$	$t_{c,4}^{i,ir}$	$t_{c,5}^{i,ir}$	$t_{c,6}^{i,ir}$	$t_{c,7}^{i,ir}$	$t_{c,8}^{i,ir}$	$t_{c,9}^{i,ir}$
0	33.9	33.9	34.6	34.5	33.9	33.3	32.7	33.2	33.2	32.8
1	32.7	32.7	31.7	31.6	30.5	32.1	31.4	32.9	33.3	32.9

Таблица 4. Значения глубинных температур молочной железы парной к рассматриваемой

i	$t^{i,mw}_{p,0}$	$t^{i,mw}_{p,1}$	$t^{i,mw}_{p,2}$	$t^{i,mw}_{p,3}$	$t^{i,mw}_{p,4}$	$t^{i,mw}_{p,5}$	$t^{i,mw}_{p,6}$	$t^{i,mw}_{p,7}$	$t^{i,mw}_{p,8}$	$t^{i,mw}_{p,9}$
0	34.9	35.4	35.6	35.7	35.6	35.2	35.0	35.1	35.6	34.9
1	35.2	34.2	34.4	34.3	34.5	34.3	34.5	34.7	34.9	34.5

Таблица 5. Значения температур подкожного слоя молочной железы парной к рассматриваемой

i	$t^{i,ir}_{p,0}$	$t^{i,ir}_{p,1}$	$t^{i,ir}_{p,2}$	$t^{i,ir}_{p,3}$	$t^{i,ir}_{p,4}$	$t^{i,ir}_{p,5}$	$t^{i,ir}_{p,6}$	$t^{i,ir}_{p,7}$	$t^{i,ir}_{p,8}$	$t^{i,ir}_{p,9}$
0	33.2	33.5	33.8	34.0	33.6	32.9	32.5	32.7	33.5	32.7
1	33.8	32.8	32.7	33.2	32.8	32.4	31.8	32.6	33.3	32.8

При создании математических моделей рассматривались качественные характеристики патологий, основанные на базовых медицинских принципах. Описание данных характеристик возможно с помощью математических функционалов, приведенных в работах [20, 32, 33]. В частности, известно, что здоровые железы обладают «зеркальной» симметрией температур в одноименных точках. Данное свойство может быть описано, например, следующим функционалом

$$sym = \overline{T^{mw}_c} - \overline{T^{mw}_p}, \tag{1}$$

Здесь T^{mw}_c – поле глубинных температур молочной железы, T^{mw}_p – поле глубинных температур парной молочной железы, $\overline{T}$ – среднее значение температур.

Для здоровых молочных желез характерны невысокие меры разброса температур. Данное свойство может быть описано, например, следующим функционалом

$$osc = \max_{t \in T^{mw}_c} t - \min_{t \in T^{mw}_c} t, \tag{2}$$

Кроме того отметим, что температура внутри органа почти всегда выше, чем на поверхности. Разность внутренней и поверхностной температур ($t^{i,g}_{c,b} = t^{i,mw}_{c,b} - t^{i,ir}_{c,b}$, b – номер точки измерения) принято называть внутренним градиентом, величина которого является важной

характеристикой состояния пациентов. Пример записи о внутреннем градиенте молочных желез приведён в таблице 6.

Таблица 6. Значения внутреннего градиента рассматриваемой молочной железы

i	$t^{i,g}_{c,0}$	$t^{i,g}_{c,1}$	$t^{i,g}_{c,2}$	$t^{i,g}_{c,3}$	$t^{i,g}_{c,4}$	$t^{i,g}_{c,5}$	$t^{i,g}_{c,6}$	$t^{i,g}_{c,7}$	$t^{i,g}_{c,8}$	$t^{i,g}_{c,9}$
0	1.7	1.5	1.3	1.3	1.7	2.3	2.6	2.1	2.4	2.8
1	0.9	1.1	1.5	1.6	2.5	1.9	1.8	0.9	1.0	1.2

Обозначенную характеристику разности температур внутри и на поверхности органа можно описать, например, функционалом

$$max_g = \max_{t \in T_c^{mw} - T_c^{ir}} |t|, \tag{3}$$

где T_c^{ir} – температуры подкожного слоя молочной железы.

Таким образом, на основе приведённых формул (1) – (3) для каждой из рассматриваемых молочных желез были найдены соответствующие значения функционалов (таблица 7).

Таблица 7. Найденные значения функционалов

i	$\overline{T_c^{mw}}$	$\overline{T_p^{mw}}$	$\max_{t \in T_c^{mw}} t$	$\min_{t \in T_c^{mw}} t$	sym	osc	max_g
0	35.6	35.3	35.9	35.3	0.3	0.6	2.8
1	33.6	34.6	34.3	33.0	-1	1.3	2.5

Итоговое признаковое пространство для решения задачи классификации составляется из наиболее информативных признаков и для различных органов и различных задач классификации может меняться.

Отметим, что информативные области найденные в работе [20] не были напрямую задействованы в алгоритме классификации, а использовались как инструмент для обоснования решения. Благодаря этим данным эксперт может установить, по каким признакам и насколько сильно термометрические показатели пациентки отклоняются от нормы.

Созданный в работе [20] алгоритм выполнял классификацию двух классов: здоровых молочных желез и больных раком. Полученное значение эффективности классификации $eff = 0.85$, чувствительности $sens = 0.89$, специфичности $spec = 0.81$. Здесь

$$sens = \frac{TP}{TP + FN},$$

$$spec = \frac{TN}{TN + FP},$$

$$eff = \sqrt{sens \cdot spec},$$

а TP – количество истинно-положительных результатов, или объектов класса «Больные», классифицированных как «Больные»; TN – количество истинно-отрицательных результатов, или объектов класса «Здоровые», классифицированных как «Здоровые»; FP – количество ложно-положительных результатов, или объектов класса «Больные», не классифицированных как «Больные»; FN – количество ложно-отрицательных результатов, или объектов класса «Здоровые», классифицированных как «Больные».

Учитывая, что микроволновая радиотермометрия является, так называемым, скрининговым методом, решалась и другая задача классификации молочных желез по термометрическим данным. А именно, производилось не выявление рака, а принадлежность пациентки к группе риска. Группа риска выделяется на основе тепловых аномалий по специальной термометрической шкале. Данная шкала содержит шесть групп, которые характеризуют уровень выраженности тепловых процессов и присваиваются молочным железам, исходя из РТМ-заключения. Далее эти группы будем именовать th-классами, которые принимают следующие значения: th0, th1, th2, th3, th4, th5. Согласно термометрической шкале чем выше уровень аномалий (значение th-класса), тем выше риск наличия рака молочной железы. Основные характеристики каждого из значений th-классов приведены в таблице 8.

Руководствуясь рекомендациями онкологов-маммологов, th-классы были разделены на 2 группы. Пациентки, температурные поля молочных желез которых определяются как th0-, th1- или th2-класс, относятся к группе пациенток, у которых отсутствуют онкологические заболевания. Пациентки, температурные поля молочных желез которых определяются как th3-, th4- или th5-класс, относятся к группе риска развития злокачественного новообразования.

Определение группы риска рака молочной железы по термометрическим признакам и интерпретация полученных результатов на языке понятном врачам выполнялась в работе [37]. Эффективность классификации достигала 0.92, специфичность – 0.90 и чувствительность – 0.93. Но в ходе проведения детального анализа результатов, несмотря на

высокую точность отделения этих двух групп данных, было замечено, что точность определения некоторых из th-классов оказалась критически низкой.

Таблица 8. Характеристики уровней выраженности температурных аномалий

th	Характеристика
0	Тепловых изменений практически нет.
1	Характерных признаков рака молочной железы на текущей момент нет. Но присутствует область пониженной температуры.
2	Характерных признаков рака молочной железы на текущей момент нет. Но присутствует область повышенной температуры.
3	Высокие глубинные температуры и температуры подкожного слоя, значение разности между ними не соответствует норме.
4	Температурные отклонения соответствуют одному из нижеперечисленных признаков: – заметно отличающаяся термоасимметрия (разность температурных полей молочных желез) при этом (глубинные температуры органа) t_{mw} и t_{ir} (температуры подкожного слоя) на одном уровне; – $t_{ir} \geqslant t_{mw}$; – сочетание повышенного уровня температур в одной из молочных желез с сохранением соотношения глубинных температур и температур подкожного слоя на одном уровне.
5	Температурные отклонения соответствуют одному или нескольким нижеперечисленным признакам: – заметно отличающаяся термоасимметрия при этом t_{mw} и t_{ir} на одном уровне; – на поле глубинных температур наблюдается «колокол»; – повышенный уровень температур в одной из желез, при этом $t_{ir} \geqslant t_{mw}$;

Также стоит принять во внимание тот факт, что th-классы, которые определяют молочные железы без онкологии вовсе не означают, что железы абсолютно здоровы. Так в молочных железах, определенных как th1, обнаруживаются следующие доброкачественные процессы: локализованный фиброз, липома, грубый послеоперационный рубец или фиброаденома без признаков роста. В случае th2 врачи диагностируют мастопатию, воспаленные кисты, участки ожога или воспаления на коже, а также фиброаденомы с признаками роста. Следовательно, обнаружение температурных отклонений, присущих всем этим 6 классам, будет способствовать своевременной диагностике различных заболеваний.

Исходя из написанного выше становится понятно, что возникает новая задача по определению, описанных th-классов, т.е. проведению многоклассовой классификации для выявления различных температурных отклонений в молочных железах. Решение данной задачи также рассматривается в настоящей работе. Признаки, по которым проводилась

классификация, брались из работы [20]. Обоснование выбора таких признаков заключается в схожести предметных областей данных исследований.

4. Постановка задачи. Предположим, что у нас имеется множество объектов обучающей выборки, разделенное на k классов. Будем считать, что объем выборки мал, а количество объектов некоторых классов значимо превосходит количество объектов в других классах. В связи с малым объемом данных, нам нежелательно решать проблему несбалансированности путем удаления объектов обучающей выборки. Следовательно, вместо удаления будем добавлять объекты. Далее считаем, что добавление объектов обучающей выборки возможно только путем их синтеза.

Пусть состояние каждого объекта выборки описывается некоторой функцией (или вектор-функцией). Далее будем считать, что в качестве элементов признакового пространства используется набор функционалов $(F_1, \ldots, F_n)$, определённых на множестве функций, описывающих состояние объектов [20]. Подобный подход к построению признакового пространства в последние годы стал активно применяться и уже доказал свою эффективность.

Рассмотрим функционал F_j, $j \in \{1, \ldots, n\}$. Эффективность разделения объектов на классы с помощью данного функционала определяется информативностью W. Традиционно под информативностью понимают величину, количественно характеризующую степень уменьшения неопределенности при классификации объектов с помощью данного функционала, т.е. меру того, насколько хорошо он помогает различать объекты разных классов. Для оценки информативности существует множество критериев. Так в работе [20] использовался критерий статистической информативности, но также рассматривались энтропийный, эвристический критерий и др. В данной работе информативность функционала оценивалась для каждого класса $y \in \{0, ..., k\}$ отдельно и основывалась на том, насколько хорошо по значениям F_j можно отделить объекты класса y от объектов остальных классов (метод "один против всех"). Область значений F_j, которая является типичной для класса объектов y, называется информативной областью и обозначается I_{jy}. Информативная область – это область, на которой обеспечивается достижение величины:

$$W_{jy} = \sup_{I_{j\alpha}} W(F_j, I_{jy}),$$

где $I_{j\alpha}$ – всевозможные подобласти значений F_j, на которых указанная информативность имеет смысл.

Информативность $W(F_j, I_{jy})$ является функцией от рассматриваемой информативной области I_{jy} по функционалу F_j. В настоящем исследовании информативность определяется следующим образом:

$$W(F_j, I_{jy}) = \sqrt{(1 - l_{j0}) \cdot l_{jy}}, \tag{4}$$

где:

$$l_{jy} = \frac{|\{x \in F_{jy} \mid i_{jy,\text{start}} < x < i_{jy,\text{end}}\}|}{|F_{jy}|}$$

– доля объектов класса y (множество F_{jy} содержит значения функционала F_j объектов, соответствующих этому классу), попавших в интервал значений $[i_{jy,\text{start}}; i_{jy,\text{end}}]$, характерный для данного класса; здесь $|\cdot|$ – мощность множества;

$$l_{j0} = \frac{|\{x \in F_{j0} \mid i_{jy,\text{start}} < x < i_{jy,\text{end}}\}|}{|F_{j0}|}$$

– доля объектов других классов (множество F_{j0} включает значения функционала F_j объектов, не относящихся к классу y), также попавших в указанный интервал $[i_{jy,\text{start}}; i_{jy,\text{end}}]$.

Выбор формулы определен устоявшейся формой определения качества разделения объектов разных классов в данной предметной области.

На основе введенных обозначений определим геометрическое строение подмножеств признакового пространства на основе определения признака из работы [20]:

$$\phi_j = (F_j, (I_{j0}, \ldots, I_{jk}), (W_{j0}, \ldots W_{jk})).$$

Здесь F_j – j-ый функционал; I_j – вектор информативных областей функционала F_j по классам объектов; W_j – вектор информативностей j-ого признака по классам объектов.

Учитывая геометрическое строение признакового пространства и несбалансированость классов в обучающей выборке, сформулируем задачу исследования.

Задача: повысить репрезентативность обучающего набора данных путем синтеза дополняющего набора, который при объединении с исходным набором данных, меняет расположение элементов обучающей выборки в признаковом пространстве в соответствии с некоторым распределением.

5. Метод решения задачи. Пусть в имеющейся обучающей выборке содержится m объектов, определенных с помощью матрицы функционалов F и вектор-столбца Y, содержащего классы, к которым принадлежат соответствующие объекты:

$$F = \begin{bmatrix} F_{11} & \dots & F_{1n} \\ \vdots & \ddots & \vdots \\ F_{m1} & \dots & F_{mn} \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}.$$

Здесь $(F_{i1}, \dots, F_{in})$ – описание объекта i из обучающей выборки; y_i – класс объекта i из обучающей выборки. Матрица F строится на основе измеренных температур и выбранных функционалов (например, 1, 3, 2). Пример расчета ее значений приведен в таблице 7. Элементы вектор-столбца Y как правило заранее определены врачом-диагностом и являются числом от 0 до 5, которое соответствует температурным аномалиям из таблицы 8.

Каждому классу объектов y по каждому функционалу F_j соответствует свое значение информативности W_{jy} и своя информативная область I_{jy} (в данной работе в качестве информативной области рассматривается числовой промежуток - информативный интервал). Это можно отразить в виде соответствующих матриц W^* и I размерностью $n \times k$. Строки матриц соответствуют функционалам, а столбцы – классам объектов:

$$W^* = \begin{bmatrix} W_{11} & \dots & W_{1k} \\ \vdots & \ddots & \vdots \\ W_{n1} & \dots & W_{nk} \end{bmatrix},$$

$$I = \begin{bmatrix} (i_{11,start}; i_{11,end}) & \dots & (i_{1k,start}; i_{1k,end}) \\ \vdots & \ddots & \vdots \\ (i_{n1,start}; i_{n1,end}) & \dots & (i_{nk,start}; i_{nk,end}) \end{bmatrix}.$$

Здесь $i_{j,start}$ – левая граница (начало) интервала соответствующего F_j; $i_{j,end}$ – правая граница (конец) интервала соответствующего F_j.

Построение информативных интервалов I_{jy} признакового пространства выполняется в соответствии с предметной областью исследования, согласно которой у каждого термометрического класса должна существовать единственная связная информативная область. Опишем метод построения информативных областей (интервалов) подробнее:

1. Определяется стартовая точка, от нее будет проводиться расширение интервала. В качестве стартовой точки для большинства классов бралась их медиана. Для $th0$-класса, которому характерно отсутствие температурных аномалий, в качестве стартовой точки по некоторым признакам выбиралось нулевое значение. Так, например, исходя из медицинского положения о симметричности температур в одноименных точках парных молочных желез значение стартовой точки интервала $th0$-класса было равно 0 для признака, характеризующего отличие температурных полей правой и левой молочных желез.

2. Определяется способ расширения интервала. Исходя из особенностей выбора стартовой точки, расширение интервала происходило либо в обе стороны от нее, либо только в одну. Расширение в одну сторону применялось, когда известно, что значения левее точки, от которой начинается расширение, признак принимать не может.

3. Осуществляется расширение интервала от стартовой точки на шаг δ до тех пор, пока не уменьшается информативность интервала W_{ij}, а границы не выходят за пределы области распределения класса. Информативность интервалов определяется формулой 4, ей же задается матрица W^*.

Выбор шага δ влияет на скорость решения и точность границ интервала. При низких значения шага необходимо большее количество итераций для сходимости метода, но будет найдено наибольшее значение величины 4. Так как на малых наборах данных метод не является ресурсоемким, рекомендуемым значением δ является минимальное значение расстояния между элементами множества F_{jy}, по которому строится интервал:

$$\min_{\substack{x,y \in F_{jy} \\ x \neq y}} |x - y|.$$

Выбор меньшего значения шага не имеет смысла, так как при расширении интервала может не изменяться соотношение количество элементов в нем, т.е. информативность будет оставаться прежней. В работе шаг δ определялся данным способом.

Следует отметить, что интервалы различных классов по ряду функционалов могут пересекаться, что указывает на их общие характеристики по ним.

Можно считать, что выбор начального варианта обучающей выборки для каждого класса y по каждому функционалу F_j задал нам некоторый закон распределения данных d_{jy} внутри информативного интервала I_{jy}. Под законом распределения понимается некоторое соотношение, которое устанавливает связь между возможными значениями $\xi = F_j$ и соответствующей им вероятностью. Задать такой закон можно, например, через функцию распределения или плотность распределения. Функция распределения выражает вероятность того, что ξ примет значение, меньше некоторого x: $f_\xi(x) = P(\xi < x)$. Плотность распределения в свою очередь находится как f'_ξ [21].

Таким образом исходная матрица распределений D имеет следующий вид:

$$D = \begin{bmatrix} d_{11} & \dots & d_{1k} \\ \vdots & \ddots & \vdots \\ d_{n1} & \dots & d_{nk} \end{bmatrix}.$$

Распределение объектов выборки в информативных интервалах может не соответствовать распределению, имеющемуся в реальности. Следовательно, возникает задача восстановления репрезентативности обучающей выборки данных с помощью дополнения имеющегося набора данных синтезированным набором и приведения закона распределения D к некоторому D^*:

$$D^* = \begin{bmatrix} d^*_{11} & \dots & d^*_{1k} \\ \vdots & \ddots & \vdots \\ d^*_{n1} & \dots & d^*_{nk} \end{bmatrix},$$

где d^*_{jy} определяется реальным распределением изучаемых объектов.

Закон распределения D^* может быть заранее известен, исходя из предметной области. В противном случае рассматривается конечное

множество различных законов распределений, из которых выбирается наилучшее по каждому функционалу. Наилучшим считается то распределение, по которому было получено максимальное значение меры эффективности на валидационном наборе данных. Множество законов распределений можно задавать как исходя из гипотез, сформированных на основе предметной области, так и выбирая из стандартных законов, например, нормальное, равномерное, биномиальное и т.п.

Для решения поставленной задачи предлагается метод управляемой синтезации данных. В результате получим новый набор данных, который, помимо натурных данных, также будет содержать и синтетические данные: $(F^*, Y^*) = (F \cup F_{syn}, Y \cup Y_{syn})$. Далее производится обучение на новом наборе данных алгоритма классификации и проверка его эффективности на валидационном наборе. Увеличение эффективности классификации при использовании нового обучающего набора позволит подтвердить правильность выбора распределения и, соответственно, правильность пространственной балансировки.

Приступим к синтезу на основе метода пространственной балансировки. Для каждого класса объектов y, интервала i_{jy} и соответствующего функционала F_j проведем следующие действия.

1. Разделим интервал i_{jy} на p подинтервалов. Значение p задается изначально. Его оптимальное значение определяется экспериментальным путем. Так, например, в случае нормального распределения, p может быть задано таким образом, чтобы ширина каждого подинтервала была не меньше среднеквадратического отклонения σ. В некоторых случаях, например в гауссовой смеси распределений с несколькими модами, имеет смысл увеличить количество подинтервалов, в работе интервал разбивался на 5 подинтервалов.

2. Определим долю объектов обучающего набора данных в каждом подинтервале $i^e_{jy}, e = \{1, ..., p\}$ на основе закона распределения d_{jy}.

В непрерывном случае доля объектов определяется формулой:

$$p^e_{jy} = P(i^e_{jy,start} < x < i^e_{jy,end}) = \int_{i^e_{jy,start}}^{i^e_{jy,end}} d(x)dx.$$

Здесь $i^e_{jy,start}$ и $i^e_{jy,end}$ – левая и правая границы подинтервала соответственно, $d(x)$ – функция плотности распределения.

В дискретном случае для известного набора данных доля объектов определяется по формуле:

$$p^e_{jy} = P(i^e_{jy,start} < x < i^e_{jy,end}) = \frac{|\{x \in F_{jy} \mid x \in i^e_{jy}\}|}{|\{x \in F_{jy} \mid x \in i_{jy}\}|}.$$

Таким образом имеется два вектора, соответствующих долям объектов в подинтервалах для распределений d_{jy} и d^*_{jy} соответственно:

$$P_{jy} = (p^1_{jy}, p^2_{jy}, ..., p^p_{jy}), \quad P^*_{jy} = (p^{1*}_{jy}, p^{2*}_{jy}, ..., p^{p*}_{jy}).$$

3. Определим количество имеющихся в каждом подинтервале объектов:

$$a^e_{jy} = |\{x \in F_{jy} \mid x \in i^e_{jy}\}|.$$

4. Определим количество объектов, которое должно быть в интервале i_{jy}. Поскольку обе плотности d_{ij} и d^*_{jy} на интервале i_{jy} непрерывны, неотрицательны и нормированы, найдется хотя бы один подинтервал, в котором $p^e_{jy} \geqslant p^{e*}_{jy}$.

Чтобы уменьшить долю объектов в таких подинтервалах, нужно либо удалить некоторое количество объектов в них, либо синтезировать новые объекты в других подинтервалах. Из-за малого объема данных нельзя удалять исходные объекты, поэтому имеет смысл сохранить в данном подинтервале имеющееся количество объектов и определить необходимое количество объектов на всем интервале i_{jy} следующим образом:

$$A^* = \left[\frac{a^e_{jy}}{p^{e*}_{jy}}\right].$$

В качестве значения e выбирается наименьшая разница среди элементов P_{jy} и P^*_{jy}:

$$e_{min} = arg \min_{i \in \{0,1,..p\}} (p^i_{jy} - p^{i*}_{jy}).$$

На основе этого рассчитаем общее количество объектов:

$$A^* = \left[\frac{a_{jy}^{e_{min}}}{p_{jy}^{e_{min}*}} \right],$$

где $[\cdot]$ – целая часть числа.

5. Определим количество объектов t_e, которых не хватает в каждом из подинтервалов для получения итогового распределения d_{jy}^* по формуле:

$$t_e = [A^* \cdot p_{jy}^{e*}] - a_{jy}^e.$$

6. Выполним синтез t_e объектов, таким образом, чтобы значения $F_{j,syn}$ попадали в подинтервал i_{jy}^e.

7. Добавим синтезированные объекты в F_{syn} и укажем для них соответствующее значение класса y. После завершения цикла синтеза, объединим изначальную и синтезированную выборки.

Пространственная балансировка позволяет получить более репрезентативный (с точки зрения заданного распределения D^*) набор данных. Тем самым данные начинают лучше отражать ключевые характеристики предметной области, минимизируется влияние пространственных аномалий, повышается точность прогнозов и устойчивость модели относительно распределения данных. Все это особенно актуально в условиях малых выборок.

Псевдокод методов приведен в Приложении.

Пример 1. На рассматриваемом интервале содержится 1000 объектов. При его разбиении на 5 равных подинтервалов получается следующий вектор долей p и количество объектов a:

$$p = (0.1, 0.2, 0.4, 0.2, 0.1), \quad a = (100, 200, 400, 200, 100).$$

Нужно синтезировать объекты, приводящие распределение на данном интервале к нормальному, т.е. к следующим долям p^*:

$$p^* = (0.027, 0.184, 0.578, 0.184, 0.027).$$

Подинтервалом e_{min} с наименьшей разницей долей будет являться первый (или последний, величины разностей долей в обоих случаях совпадают):

$$p^0 - p^{0*} = -0.073.$$

Тогда:

$$A^* = \left[\frac{100}{0.027}\right] = 3073,$$

и количество объектов, необходимых для синтеза в каждом подинтервале:

$$t = (0, 481, 1741, 481, 0).$$

Пример 2. С целью выявления взаимосвязи между эффективностью классификации и репрезентативность набора данных рассмотрим следующий пример. Зададим самостоятельно генеральную совокупность объектов $F_g \times Y_g$ объемом 10000 записей для каждого класса ($m = 30000$).

$$F_g = \begin{bmatrix} F_{11} & F_{12} & F_{13} \\ \vdots & \ddots & \vdots \\ F_{m1} & F_{m2} & F_{m3} \end{bmatrix}, Y_g = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}, y = \{0; 1; 2\}.$$

При этом пусть распределение объектов генеральной совокупности в информативных интервалах близким к нормальному (статистика Андерсона-Дарлинга принимает значения меньшие критического для уровня значимости в 5%). Пример распределения для функционала $F3$ и объектов из генеральной совокупности с классом 1 приведен на рисунке 1. В данном случае границы информативного интервала имели значения 9 и 10 соответственно.

Предположим, что на основе генеральной совокупности составлена некоторая выборка объектов. Причем составлена таким образом, что нормальное распределение значений функционалов в информативных интервалах было нарушено (рисунок 2).

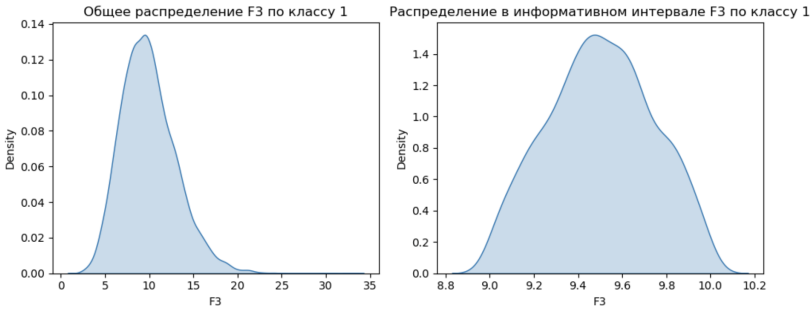

Рис. 1. Плотность распределения значений функционала $F3$ по объектам класса 1 из генеральной совокупности

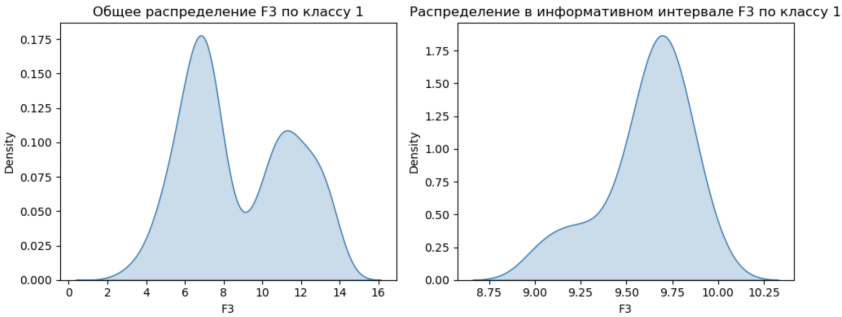

Рис. 2. Плотность распределения значений функционала $F3$ по объектам класса 1 из выборочной совокупности

Следовательно, данная выборка будет обладать пониженной репрезентативностью. Кроме того объем данных в выборке по сравнению с генеральной совокупностью был существенно уменьшен и составил 1374 записи для класса 0, 888 - для класса 1 и 536 - для класса 2; также в выборке наблюдается классовая несбалансированность.

Применим метод пространственной балансировки с целью синтезации данных и восстановления распределения в информативных интервалах.

Пример распределения значений для функционала $F3$ в информативном интервале после синтезации объектов класса 1 приведен на рисунке 3.

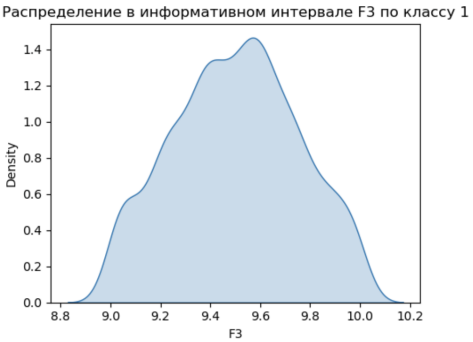

Рис. 3. Плотность распределения значений для функционала $F3$ в информативном интервале после синтезации объектов класса 1

Оценим значение дивергенции Кульбака-Лейблера (D_{KL_F}) по распределениям p генеральной совокупности и q – выборочной совокупности для каждого функционала F в пределах информативных интервалов до и после проведения синтезации:

$$D_{\text{kl}}(p||q) = \int p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx \, .$$

Также до и после синтезации вычислим эффективность классификации (применяемый алгоритм классификации k-ближайших соседей), исходя из точности определения каждого из классов выборочной совокупности (таблица 9):

$$m_{eff} = \sqrt[3]{\prod_{i=0}^{2} y(i)},$$

где $y(i)$ – точность определения класса i.

Таблица 9. Точность классификации и значения дивергенции до и после синтеза, осуществляемого с помощью пространственной балансировки

Этап	$D_{KL_{F1}}$	$D_{KL_{F2}}$	$D_{KL_{F3}}$	m_eff
до синтеза	1.1	0.83	0.45	0.84
после синтеза	0.18	0.11	0	0.94

Уменьшение значения дивергенции после проведения синтезации с помощью пространственной балансировки свидительствует повышении репрезентативности выборочной совокупности. Такой эффект наблюдается из-за того, что распределение выборочной совокупности в информативных интервалах после синтеза больше соответствует тому распределению, которое наблюдалось в генеральной совокупности. Кроме того с повышением репрезентативности увеличивается и эффективность классификации. При этом увеличение меры эффективности составляет 10% (с 84% до 94%).

Для большей наглядности проиллюстрируем метод на примере решения задачи многоклассовой классификации пациенток с заболеваниями молочных желез на основе данных микроволновой радиотермометрии.

6. Повышение сбалансированности. При создании модели машинного обучения, обучаемой для решения задачи многоклассовой классификации, а именно, определения уровня выраженности температурных аномалий, было выявлено, что простые модели зачастую склонны игнорировать объекты некоторых th-классов. Точность определения каждого из классов моделью логистической регрессии представлена на рисунке 4.

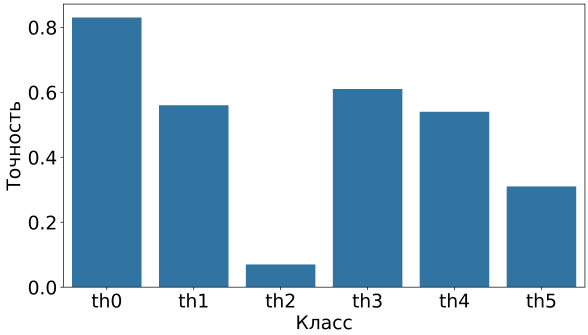

Рис. 4. Точность определения th классов моделью логистической регрессии

Отметим, что наименьшая точность наблюдается по объектам $th2$-класса – этот класс определяется лишь в 7% случаев, но для специалистов предметной области может представлять не меньший интерес, чем другие классы.

Основная причина неудовлетворительных результатов заключается в несбалансированности начального набора данных – распределение объектов по классам значительно отличается. Кроме того работа ведется

с выборкой данных малого объема (рисунок 5), что также влечет проблемы снижения репрезентативности данных.

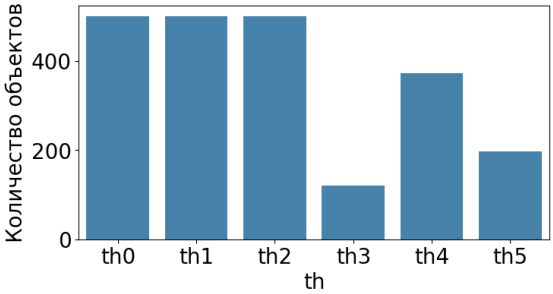

Рис. 5. Начальное распределение данных по th-классам

Заметим, что в оригинальном наборе данных объектов $th0$-, $th1$- и $th2$-классов значительно больше, чем остальных. В общей сумме их количество составляет 8428 объектов. Для упрощения задачи балансировки и в связи с низкой размерностью данных было произведено снижение их количества до 500 в каждом классе с сохранением оригинального распределения.

Для устранения всех вышеперечисленных недостатков предлагается применение метода балансировки, основанном на синтезации объектов, который позволит повысить точность многоклассовй классификации. Выполнять балансировку будем в соответствии с методом пространственной балансировки, описанном ранее в разделе 3. Перед ее осуществлением должны быть заданы:

1. матрица F значений функционалов по объектам (функционалы получены из работ [20, 32, 33]);

2. классы объектов Y, определенные специалистами;

3. матрица I с информативными интервалами, построенные описанным ранее алгоритмом;

4. распределение данных D, которое задано внутри интервалов по классам объектов.

В данном исследовании в качестве объектов рассматриваются температурные поля молочных желез пациенток. В качестве функционалов для описания термометрических процессов используются sym, osc, max_g, описывающие основные характеристические особенности классов. Данные функционалы были определены равенствами (1), (2), (3) выше. Матрица информативности, используемых признаков, представлена в таблице 10.

Таблица 10. Матрица информативности W^*

F	th0	th1	th2	th3	th4	th5
sym	0.765	0.699	0.652	0.598	0.735	0.689
max_g	0.529	0.544	0.531	0.685	0.512	0.593
osc	0.621	0.551	0.551	0.543	0.516	0.512

Метками объектов выступают значения th-классов. Задание информативных интервалов и определение распределения данных в них становится возможным после построения интервалов. Таким образом, в соответствии с постановкой задачи выполнение пространственной балансировки также связано с построением информативных интервалов признакового пространства. Полученные интервалы позволят выделить, во-первых, области характерных значений для объектов различных классов, а также выявить области, в которых наблюдается нехватка объектов. Последние области будут дополняться синтетическими объектами.

Для синтезации объектов использовалась модель Gaussian Copula Synthesizer. Она позволяет проводить генерацию новых данных двумя способами: создание всего набора данных с заданным количеством объектов и функцией распределения; создание группы объектов с заданными характеристиками. Благодаря второму способу создания объектов, становится возможной их генерация в найденных информативных интервалах I.

Обучение модели Gaussian Copula Synthesizer предполагает преобразование значений $(F_{i1}, \ldots, F_{in})$, описывающих объекты из обучающей выборки, к нормальному распределению с помощью Гауссовой копулы, получая таким образом значения в виде

$$[\Phi^{-1}(R_0(F_0)), \ldots, \Phi^{-1}(R_n(F_n))],$$

где R – это функция распределения (в случае использования нормального распределения R определяется через μ и σ); $\Phi^{-1}(R_i(F_i))$ – обратная функция распределения. Затем на основе вычисленных значений рассчитывается матрица ковариаций Q.

При создании синтетического объекта случайным образом генерируется вектор v размерностью n, компоненты которого распределены по нормальному закону. Последующие вычисления происходят с помощью разложения Холецкого, которое позволяет генерировать новые данные с заданной структурной зависимостью,

основанной на рассчитанной ковариационной матрицей: $LL^T = Q$. Таким образом, полученный вектор $u = Lv$ используется для синтезации нового объекта со значениями $[R_i^{-1}(\Phi(u_i)), \ldots, R_n^{-1}(\Phi(u_n))]$. Синтезация новых объектов проводится для каждого функционала, при этом объект добавляется в синтетический набор только в том случае, если значения рассматриваемого функционала попадают в информативный интервал.

7. Результаты. По представленным признакам для каждого класса были построены информативные интервалы, максимизирующие элементы матрицы W^*, результаты показаны на рисунке 6.

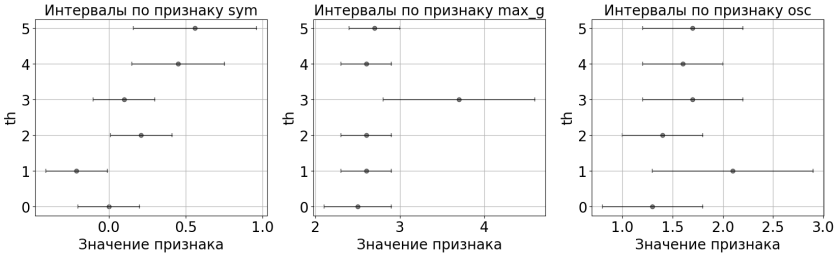

Рис. 6. Найденные информативные интервалы

По первому признаку (sym) наблюдается, что $th4$-, $th5$-классы имеют повышенную разность температурных полей молочных желез. Интервал по $th1$-классу заметно смещен влево. Это объясняется тем, что при вычитании из температур обследуемой молочной железы температур по парной молочной железе, найденная разность в большинстве случаев имеет отрицательный знак. Такое поведение температурных полей подтверждается положением о том, что температуры обследуемой молочной железы из $th1$-класса понижены по сравнению с парной железой.

По второму признаку (max_g) заметно, что интервал по $th3$-классу значительно смещён в сторону более высоких значений по сравнению с другими th-классами, что отражает аномалии в соотношении температур подкожного слоя и глубинных температурах.

По третьему признаку (osc) особенно выделяется интервал $th1$-класса. Вычисленные значения осцилляции свидетельствуют о существенном перепаде температур в пределах одной молочной железы, который не так заметен в $th2$- и особенно $th0$-классах. Таким образом, в объектах $th1$-класса наблюдается температурная неоднородность. Немного менее выраженные аномалии по признаку osc соответствуют классам $th3$, $th4$ и $th5$. Однако их отличие от классов с наименьшим

риском развития злокачественной опухоли (th-классы до $th3$) все же имеется и заметно также по признакам, описанным ранее.

На рисунке 7 показано сравнение распределений по первым двум признакам на изначальной выборке и сбалансированной для объектов $th2$-класса. Пунктирной областью обозначены границы информативных интервалов по данным признакам. Как видно, концентрация объектов в области увеличилась, а распределение объектов стало ближе к нормальному. Аналогичным образом была изменена характеристика распределения $th4$- и $th5$-классов.

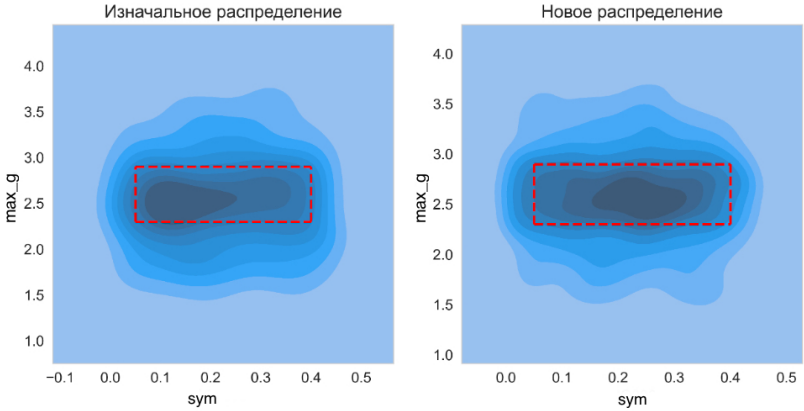

Рис. 7. Распределение объектов класса $th2$

Класс медицинских данных характеризуется высокой сложностью, обусловленной множеством факторов, влияющих на измеряемые характеристики. Одной из ключевых особенностей подобных данных является отсутствие информации о реальном распределении признаков в генеральной совокупности, что делает невозможным применение классических метрических критериев, таких как дивергенция Кульбака–Лейблера, для оценки степени репрезентативности выборки. Кроме того, применение таких дивергенционных мер требует знания априорных распределений, что в условиях ограниченных и неоднородных медицинских данных практически недостижимо.

В связи с этим в данной работе для количественной оценки репрезентативности используется мера информационной энтропии:

$$H = -\sum_{i=1}^{N} p_i \ln p_i,$$

где p_i – доля объектов в i-м подинтервале признакового пространства, N – число подинтервалов (пять). Энтропия, как мера неопределённости служит косвенной оценкой приближённости распределения к предполагаемому истинному распределению. Высокое значение энтропии может свидетельствовать о разнообразии выборки и охвате различных подгрупп в пределах исследуемой совокупности, тогда как низкая энтропия может указывать на смещённость или ограниченность представленных данных. Таким образом, энтропийный подход предоставляет инструмент для сравнения различных выборок по степени их потенциальной репрезентативности без необходимости априорного знания полной структуры генеральной совокупности.

Оценка энтропии по каждому интервалу и классу приведена в таблице 11. Как видно, количество информации в каждом интервале увеличилось.

Таблица 11. Сравнение энтропии в интервалах при количественной и пространственной балансировках. A^* – количество синтезируемых объектов в интервалах I

th-класс	Признак	Энтропия интервала по количественно сбалансированным данным	Энтропия интервала по пространственно сбалансированным данным	A^*	I
3	sym	1.28	1.29	173	$(-0.06; 0.14)$
3	max_g	1.14	1.22	357	$(2.8; 4.6)$
3	osc	1.08	1.15	422	$(1.2; 2)$
4	sym	1.25	1.29	366	$(0.13; 0.73)$
4	max_g	1.01	1.12	462	$(2.3; 2.9)$
4	osc	1.05	1.14	518	$(1.2; 2)$
5	sym	1.24	1.28	227	$(0.16; 0.96)$
5	max_g	1.12	1.20	281	$(2.4; 3)$
5	osc	1.03	1.15	205	$(1.25; 2.25)$

Для проверки качества полученного набора данных проводилось сравнение моделей классификации, обученных на пяти различных выборках:

– изначальной, не подвергавшейся изменениям;

– количественно сбалансированной с помощью применения метода SMOTE, в результате чего количество объектов каждого класса в выборке стало равным;

– количественно сбалансированной с помощью применения метода ADASYN;

– количественно сбалансированной с помощью применения метода Gaussian Copula;

– пространственно сбалансированной.

После обучения проводилась классификация одного и того же тестового набора данных. В результате определялись точность определения каждого из классов, а также мера эффективности – среднегеометрическое всех точностей:

$$m_{eff} = \sqrt[6]{\prod_{i=0}^{5} th_i},$$

где th_i – точность определения thi-класса. Определим точность, как долю верноопределенных объектов рассматриваемого класса. Для выявления наилучшего набора данных результаты классификации сравнивались по двум параметрам: минимальное значение точности (чем оно выше, тем модель классификации считается более сбалансированной), а также по эффективности, определяющей усреднённое качество классификации.

В ходе экспериментов были рассмотрены различные типы распределений, среди которых наиболее эффективным по всем признакам оказалось нормальное распределение, выбранное для дальнейшей работы в качестве D^* (таблица 12).

Таблица 12. Оценка распределений

Распределение	Равномерное	Нормальное	Гамма	Gaussian KDE
Эффективность	0.27	0.48	0.47	0.475

Также рассматривалась задача объективного задания параметра p, направленного на максимизацию меры эффективности классификации. В качестве эвристики был протестирован подход, основанный на правиле Стёрджеса:

$$p = 1 + log_2[N],$$

где N – общее количество объектов выборки.

В рамках вычислительного эксперимента использовался набор данных, ограниченный 500 объектами каждого класса, что соответствовало значению $p = 9$ для каждого подкласса. Результаты показали, что оптимальное значение параметра p зависит не только от

состава обучающей выборки, но и от используемой классификационной модели (таблица 13).

Таблица 13. Зависимость меры эффективности классификаторов от параметра p

p	m_{eff}^{df}	m_{eff}^{lr}	A^*
3	0.42	0.382	1994
4	0.441	0.404	2491
5	0.417	0.414	3011
6	0.421	0.426	3022
7	0.35	0.407	3414
8	0.411	0.407	3489
9	0.434	0.409	3857
10	0.426	0.418	2993
11	0.409	0.41	3420
12	0.436	0.406	3341
13	0.411	0.401	3497
14	0.41	0.42	3616

Более того, наивысшие значения меры эффективности m_{eff} были достигнуты при значениях $p < 1 + \log_2 N$. Это наблюдалось как для классификатора на основе решающего леса ($m_{\text{eff}}^{\text{df}}, p = 4$), так и для логистической регрессии ($m_{\text{eff}}^{\text{lr}}, p = 6$).

Дополнительно установлено, что увеличение значения параметра p приводит к росту числа синтезируемых объектов A^*, что, в свою очередь, увеличивает время выполнения процедуры синтеза данных.

Следует отметить, что параметр p по сути выполняет роль гиперпараметра, значение которого должно подбираться с учётом характеристик обучающей выборки, архитектуры модели, целей синтеза и допустимого времени выполнения. Это означает, что несмотря на попытки его объективного задания, оптимальный выбор p остаётся задачей, зависящей от множества факторов и требует валидации на конкретных данных.

В таблице 14 показаны результаты вычислительных экспериментов моделью логистической регрессии.

Как видно, количественная балансировка методом ADASYN понизила эффективность классификации, а при балансировке методами SMOTE и Gaussian Copula снизила точность определения $th4$-класса и повысила точность определения $th2$-класса, который практически никогда не определялся корректно. В результате эффективность классификации повысилась на 4-5%. Однако в целом качество классификации осталось на низком уровне, так как на двух классах модель оказывается верна

лишь в 17-20% случаев. Пространственная балансировка, в свою очередь, повысила точность определения каждого из этих классов более чем на 14% относительно количественной балансировки. Из-за чего наименьшая точность определения повысилась до 31%, а эффективность – на 7%, что согласуется с увеличением энтропии также на 0,07. Это связано с тем, что увеличение энтропии приводит к повышению информационной насыщенности информативных зон, определяющих характерные значения признаков для каждого класса.

Таблица 14. Результаты вычислительных экспериментов, проведенных логистической регрессией

Набор данных	m_{eff}	th_0	th_1	th_2	th_3	th_4	th_5
Начальный набор данных	0.38	0.83	0.56	0.07	0.61	0.54	0.31
Количественно сбалансированный методом SMOTE	0.42	0.61	0.71	0.19	0.68	0.17	0.69
Количественно сбалансированный методом ADASYN	0.36	0.74	0.54	0.06	0.74	0.41	0.37
Количественно сбалансированный методом Gaussian Copula	0.43	0.61	0.7	0.2	0.68	0.18	0.67
Пространственно сбалансированный	0.49	0.66	0.6	0.41	0.59	0.31	0.5

В случае более сложной модели машинного обучения, а именно полносвязной нейронной сети, разница между количественно и пространственно сбалансированными наборами данных, представленная в таблице 15, не столь значительна. Однако пространственная балансировка все же улучшает показатели количественной. Так можно отметить, что пространственная балансировка повысила минимальную точность определения $th5$-класса (относительно методов SMOTE и Gaussian Copula. ADASYN имеет высокую точность определения $th5$-класса, но практически никогда не определяет корректно $th4$-класс). Рост составил 42% (с 20% до 62%) относительно начального набора данных и 2% (с 60% до 62%) относительно количественно сбалансированного набора методом SMOTE. Устойчивость различных методов балансировки данных оценивалась на модели нейронной сети. Для этого использовалась процедура 5-кратной кросс-валидации: исходный набор данных

разделялся на 5 равных по объёму подвыборок с сохранением исходного соотношения классов в каждой из них. На четырёх подвыборках (объединённых в обучающую часть) выполнялось обучение модели нейронной сети и применялись соответствующие методы балансировки. Полученная модель затем тестировалась на оставшейся подвыборке. Процедура повторялась для всех пятёрок, в результате чего выполнялось 5 вычислительных экспериментов.

Таблица 15. Результаты вычислительных экспериментов, проведенных нейронной сетью

Набор данных	m_{eff}	th_0	th_1	th_2	th_3	th_4	th_5
Начальный набор данных	0.46	0.65	0.72	0.35	0.51	0.58	0.2
Количественно сбалансированный методом SMOTE	0.47	0.66	0.68	0.24	0.58	0.31	0.6
Количественно сбалансированный методом ADASYN	0.12	0.66	0.71	0.31	0.51	0.005	0.81
Количественно сбалансированный методом Gaussian Copula	0.48	0.7	0.65	0.27	0.6	0.34	0.53
Пространственно сбалансированный	0.48	0.66	0.65	0.28	0.58	0.31	0.62

Для количественной оценки устойчивости методов балансировки использовалась дисперсия величины m_{eff}, рассчитанная по результатам кросс-валидации. Наименьшую дисперсию (наиболее высокую устойчивость) продемонстрировал метод *Gaussian Copula*, для которого дисперсия составила $0{,}023$. Методы количественной балансировки *ADASYN* и *SMOTE* показали сопоставимую устойчивость – дисперсия m_{eff} составила $0{,}25$.

Наиболее чувствительным к составу выборки оказался метод пространственной балансировки. При параметре $p = 7$ дисперсия достигала значения $0{,}3$. При этом устойчивость данного метода существенно зависела от значения параметра p: при уменьшении p наблюдалось снижение дисперсии. Так, при $p = 4$ она составила $0{,}25$.

Полученные результаты подтверждают, что метод пространственной балансировки требует дополнительной настройки параметров и валидации для обеспечения устойчивости, в то время как

генеративные методы, такие как Gaussian Copula, демонстрируют более стабильное поведение.

8. Заключение. Таким образом, разработанный метод формирования обучающего набора данных на основе малой несбалансированной выборки обеспечил повышение качества данных и результатов многоклассовой классификации как по отношению к исходной выборке, так и к количественно сбалансированной. Такой эффект объясняется тем, что синтез объектов проводился именно в информативно значимых областях, что отличает данную методику от других существующих подходов синтеза.

Отметим, что обе рассматриваемые в работе модели классификации, а именно логистическая регрессия и полносвязная нейронная сеть, в результате пространственной балансировки достигли схожей эффективности в определении th-классов. Таким образом, при пространственной балансировке происходит снижение требований к моделям машинного обучения, благодаря чему становится возможным использование более интерпретируемых моделей в системах искусственного интеллекта. При этом повышение репрезентативности позволяет также улучшить качество обоснования по интерпретируемым моделям за счет более адекватного знания о предметной области, имеющегося у моделей.

Существенным фактором, влияющим на эффективность метода, является выбор распределений, к которым приводится обучающая выборка. В данной работе наилучшим распределением оказалось нормальное. Мы предполагаем, что в силу своего широкого распространения, зачастую именно оно будет являться адекватным выбором, если из предметной области не следует что-то другое.

Отметим, что в результате применения метода повышения репрезентативности также происходит снижение требований к изначальному обучающему набору данных, а именно к его объему и исходной репрезентативности. Описанный в работе метод позволяет повышать качество наборов данных по этим характеристикам на основе знаний предметной области. Таким образом, потенциально это может способствовать расширению применимости методов машинного обучения в областях, для которых характерны малые объемы данных, а именно в медицине, физике, химии и т.п.

Приложение

Алгоритм 1. Метод построения информативных интервалов

for all F in $Functional_set$ **do**
  for all y in $class_set$ **do**
    if $y = 0$ **then**
      $start_point \leftarrow 0$
    else
      $start_point \leftarrow$ медианное значение F в классе y
    end if
    $\delta \leftarrow$ минимальное расстояние между значениями F в классе y
    $direction \leftarrow$ направление расширения интервала для F по классу y
    $previos_interval_W \leftarrow 0$
    $interval_W \leftarrow 0$
    $i_start \leftarrow start_point$
    $i_end \leftarrow start_point$
    $min_F \leftarrow$ минимальное значение F в классе y
    $max_F \leftarrow$ максимальное значение F в классе y
    while $(interval_W >= previos_interval_W)$ and $(i_start >= min_F)$ and $(i_end <= max_F)$ **do**
      if $direction =$ 'one side' **then**
        $i_end \leftarrow i_end + \delta$
      else
        $i_start \leftarrow i_start - \delta$
        $i_end \leftarrow i_end + \delta$
      end if
      $interval \leftarrow [i_start; i_end)$
      $interval_W \leftarrow$ вычисленная информативность интервала
    end while
  end for
end for

Алгоритм 2. Метод пространственной балансировки

for all F in $Functional_set$ **do**
  for all y in $class_set$ **do**
    $i \leftarrow$ информативный интервал значений F по классу y
    $min_diff \leftarrow 1$
    for $subinterval = 1, \ldots, m$ **do**
      $p \leftarrow$ имеющаяся доля значений F в подинтервале
      $required_p \leftarrow$ требуемая доля значений F в подинтервале
      if $required_p - p < min_diff$ **then**
        $min_diff \leftarrow required_p - p$
        $a_e \leftarrow$ имеющееся число значений F в подинтервале
        $required_p_e \leftarrow required_p$
      end if
    end for
    $A \leftarrow a_e \div required_p_e$
    for $subinterval = 1, \ldots, m$ **do**
      $a \leftarrow$ имеющееся число значений F в подинтервале
      $t \leftarrow int(A \cdot required_p) - a$
      синтезируем t объектов с классом y и значениями F внутри подинтервала
    end for
  end for
end for

Литература

1. Jakubik J., Vossing M., Kuhl N., Walk J. Satzger G. Data-Centric Artificial Intelligence // Business & Information Systems Engineering. 2024. vol. 66. no. 4. pp. 507–515.
2. Kumar S., Datta S., Singh V., Singh S.K., Sharma R. Opportunities and Challenges in Data-Centric AI // IEEE Access. 2024. vol. 12. pp. 33173–33189.
3. Hamid O.H. Data-Centric and Model-Centric AI: Twin Drivers of Compact and Robust Industry 4.0 Solutions // Applied Sciences. 2023. vol. 13. no. 5. DOI: 10.3390/app13052753.
4. Emmanuel T., Maupong T., Mpoeleng D., Semong T., Mphago B., Tabona O. A survey on missing data in machine learning // Journal of Big Data. 2021. vol. 8. no. 1. DOI: 10.1186/s40537-021-00516-9.
5. Gerber S, Pospisil L, Sys S, Hewel C, Torkamani A, Horenko I Co-Inference of Data Mislabelings Reveals Improved Models in Genomics and Breast Cancer Diagnostics // Frontiers in Artificial Intelligence 2022. vol. 4. DOI: 10.3389/frai.2021.739432.
6. Feng W., Quan Y., Dauphin G. Label Noise Cleaning with an Adaptive Ensemble Method Based on Noise Detection Metric // Sensors. 2020. no. 20(23). DOI: org/10.3390/s20236718.
7. Чехомова О.В. О проблемах формирования административной ответственности за нарушения прав пациента // Юристъ - Правоведъ. 2010. № 3. С. 120–123.

8. Собянин К.В., Куликова С.П. Unet-boosted classifier – мультизадачная архитектура для малых выборок на примере классификации МРТ снимков головного мозга // Информатика и автоматизация. 2024. Т. 23. № 4. С. 1022–1046.
9. Зуев Ю.Ф. Сравнительная оценка достоверности биологических данных при малых выборках методами дисперсионного, кластерного и графового анализа (нестохастический подход) // Актуальные проблемы гуманитарных и естественных наук. 2015. № 1-1. С. 36–45.
10. Моргунова А.В., Сажина О.С. Использование непараметрических методов математической статистики для поиска космических струн // Вестник Московского университета. Серия 3. Физика. Астрономия. 2019. № 5. С. 86–92.
11. Мокеев В.В., Томилов С.В. О решении проблемы выборки малого размера при использовании линейного дискриминантного анализа в задачах распознавания лиц // Бизнес-информатика. 2013. № 1(23). С. 37–43.
12. Алексахина К.А., Жуков В.В. Особенности построения моделей машинного обучения на малых наборах данных в биомедицине // Информационно-телекоммуникационные технологии и математическое моделирование высокотехнологичных систем. Российский университет дружбы народов, 2021. С. 151–155.
13. Hu Y., Sandt R., Spatschek R. Practical feature filter strategy to machine learning for small datasets in chemistry // Scientific Reports. 2024. vol. 14. no. 1. DOI: 10.1038/s41598-024-71342-1.
14. Safonova A., Ghazaryan G., Stiller S., Main-Knorn M., Nendel C., Ryo M. Ten deep learning techniques to address small data problems with remote sensing // International Journal of Applied Earth Observation and Geoinformation. 2023. vol. 125. DOI: 10.1016/j.jag.2023.103569.
15. He H., Garcia E.A. Learning from Imbalanced Data // IEEE Transactions on Knowledge and Data Engineering. 2009. vol. 21. no. 9. pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.
16. Polyakov M., Sirotin D. A Realistic Breast Phantom for Investigating the Features of the Microwave Radiometry Method Using Mathematical and Physical Modelling // Technologies. 2025. vol. 13(3). DOI: 10.3390/technologies13030106.
17. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: synthetic minority over-sampling technique // Journal of Artificial Intelligence Research. 2002. vol. 16. pp. 321–357. DOI: 10.1613/jair.953.
18. Hauptmann T., Fellenz S., Nathan L., Tuscher O., Kramer S. Discriminative Machine Learning for Maximal Representative Subsampling // Scientific Reports. 2023. vol. 13. no. 1. DOI: 10.1038/s41598-023-48177-3.
19. Priestley M., O'Donnell F., Simperl E. A Survey of Data Quality Requirements That Matter in ML Development Pipelines // ACM Journal of Data and Information Quality. 2023. vol. 15. no. 2. pp. 1–39. DOI: 10.1145/3592616.
20. Levshinskii V. Mathematical models for analyzing and interpreting microwave radiometry data in medical diagnosis // Journal of computational and engineering mathematics. 2021. vol. 8. no. 1. pp. 3–14.
21. Лебедев А.В., Фадеева Л.Н. Теория вероятностей и математическая статистика (4-е издание) // М.: Наука. 2018. 408 с.
22. Kwon S., Lee S. Recent Advances in Microwave Imaging for Breast Cancer Detection // International journal of biomedical imaging. 2016. vol. 2016. pp. 1–25. DOI: 10.1155/2016/5054912.
23. Barrett A.H., Myers P.C. Subcutaneous Temperature: A method of Noninvasive Sensing // Science. 1975. vol. 190. pp. 669–671. DOI: 10.1126/science.1188361.

24. Barrett A.H., Myers Ph.C., Sadovsky N.L. Microwave Thermography in the Detection of Breast Cancer // AJR. American journal of roentgenolog. 1980. vol. 134(2). pp. 365–368. DOI: 10.2214/ajr.134.2.365.
25. Gautherie M. Temperature and Blood Flow Patterns in Breast Cancer During Natural Evolution and Following Radiotherapy // Progress in Clinical and Biological Research. 1982. pp. 21–64.
26. Троицкий В.С. К теории контактных радиометрических измерений внутренней температуры тел // Радиофизика. 1981. Т. 2. № 9. С.v1054–1061.
27. Goryanin I., Karbainov S., Shevelev O., Tarakanov A., Redpath K., Vesnin S., Ivanov Yu. Passive microwave radiometry in biomedical studies // Drug Discovery Today. 2020. vol. 25. no. 4. pp. 757–763.
28. Попов И.Е., Крылова А.Е. Анализ термометрических данных головного мозга, полученных методом микроволновой радиотермометрии // Математическая физика и компьютерное моделирование. 2023. Т. 26. № 2. С. 32–42. DOI: 10.15688/mpcm.jvolsu.2023.2.3.
29. Sedankin M.K., Gudkov A.G., Nelin I.V., Skuratov V.A., Leushin V.Yu, Chizhikov S.V., Sidorov I.A. Mathematical modeling of heat exchange processes in the brain in the presence of pathology for the design of a microwave radiothermograph // Bio-Medical Engineering. 2023. vol. 57. no. 4. pp. 276–279. DOI: 10.1007/s10527-023-10314-7.
30. Li J., Galazis C., Popov L., Ovchinnikov L., Kharybina T., Vesnin S., Losev A., Goryanin I Dynamic Weight Agnostic Neural Networks and Medical Microwave Radiometry (MWR) for Breast Cancer Diagnostics // Diagnostics. 2022. vol. 12. no. 9. DOI: 10.3390/diagnostics12092037.
31. Веснин С.Г., Каплан М.А., Авакян Р.С. Современная микроволновая радиотермометрия молочных желез // Опухоли женской репродуктивной системы. 2008. № 3. С. 28–33. DOI: 10.17650/1994-4098-2008-0-3-28-33.
32. Лосев А.Г., Левшинский В.В. Интеллектуальный анализ данных микроволновой радиотермометрии в диагностике рака молочной железы // Математическая физика и компьютерное моделирование. 2017. Т. 20. № 5. С. 49–62. DOI: 10.15688/mpcm.jvolsu.2017.5.6.
33. Лосев А.Г., Левшинский В.В. Интеллектуальный анализ термометрических данных в диагностике молочных желез // Управление большими системами: сборник трудов. 2017. № 70. С. 113–135.
34. Levshinskii V., Galazis C., Ovchinnikov L., Vesnin S., Goryanin I. Application of Data Mining and Machine Learning in Microwave Radiometry (MWR) // Communications in Computer and Information Science (CCIS). 2020. vol. 1211. pp. 265–288. DOI: 10.1007/978-3-030-46970-2 _13.
35. Гермашев И.В., Дубовская В.И. Применение моделей нечеткой математики для решения задач медицинской диагностики // Математическая физика и компьютерное моделирование. 2021. Т. 24. № 4. С. 53–66. DOI: 10.15688/mpcm.jvolsu.2021.4.4.
36. Osmonov B., Ovchinnikov L., Galazis C., Emilov B., Karaibragimov M., Seitov M., Vesnin S., Losev A., Levshinskii V., Popov I., Mustafin C., Kasymbekov T., Goryanin I. Passive Microwave Radiometry for the Diagnosis of Coronavirus Disease 2019 Lung Complications in Kyrgyzstan // Diagnostics. 2021. vol. 11. no. 2. DOI: 10.3390/diagnostics11020259.
37. Лосев А.Г., Попов И.Е., Гудков А.Г., Чижиков С. В. Интеллектуальный анализ данных микроволновой радиотермометрии в медицинской диагностике // Нанотехнологии: разработка, применение – XXI век. 2023. Т. 15. № 1. С. 5–22. DOI: 10.18127/j22250980-202301-01.

38. Ahmad H.F., Mukhtar H., Alaqail H., Seliaman M., Alhumam A. Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning // Applied Sciences. 2021. vol. 11. no. 3. DOI: 10.3390/app11031173.

Лосев Александр Георгиевич — д-р физ.-мат. наук, профессор, директор института, институт математики и информационных технологий, Волгоградский государственный университет. Область научных интересов: искусственный интеллект, интеллектуальный анализ данных, математическое моделирование, эллиптические уравнения в частных производных на римановых многообразиях. Число научных публикаций — 191. alexander.losev@volsu.ru; проспект Университетский, 100, 400062, Волгоград, Россия; р.т.: +7(904)774-6821.

Попов Илларион Евгеньевич — аспирант, кафедра математического анализа и теории функций, Волгоградский государственный университет. Область научных интересов: искусственный интеллект, интеллектуальный анализ данных, математическое моделирование. Число научных публикаций — 28. popov.larion@volsu.ru; проспект Университетский, 100, 400062, Волгоград, Россия; р.т.: +7(937)088-8938.

Резникова Анастасия Сергеевна — магистрант, кафедра математического анализа и теории функций, Волгоградский государственный университет. Область научных интересов: искусственный интеллект, интеллектуальный анализ данных, математическое моделирование. Число научных публикаций — 11. a.reznikova@volsu.ru; проспект Университетский, 100, 400062, Волгоград, Россия; р.т.: +7(917)845-2535.

Поддержка исследований. Исследование выполнено за счет гранта Российского научного фонда № 25-21-00330, https://rscf.ru/project/25-21-00330/.

DOI 10.15622/ia.24.4.5

A. LOSEV, I. POPOV, A. REZNIKOVA

IMPROVING THE REPRESENTATIVENESS OF THE TRAINING DATASET BY MEANS OF SPATIAL BALANCING

Losev A., Popov I., Reznikova A. **Improving the Representativeness of the Training Dataset by Means of Spatial Balancing.**

Abstract. The study examines the specifics of training machine learning algorithms on small datasets and addresses the task of forming a training set with high representativeness. It is known that class imbalance in objects, typical for small datasets, negatively affects the performance of algorithms. To mitigate this issue, various data synthesis methods have been developed in machine learning to supplement existing datasets and equalize the number of objects per class. However, these methods do not solve the problem of insufficient representativeness. This article proposes a method for constructing a representative training dataset by specifying the distribution that best corresponds to reality. The distribution is formed for each feature within the informative areas. Informative areas contain characteristic values of features that are most significant for distinguishing classes of objects. The proposed method of constructing areas is based on the idea of gradual expansion, accompanied by an increase in the informativeness of the areas. At the same time, informativeness is understood as a measure reflecting how well objects of different classes can be separated using the considered area. To form a complementary dataset, a generation method has been developed. As a result of its application, the complementary dataset is combined with the original one and forms the specified distribution in the informative area. This distribution can be determined either based on expert knowledge about the subject area, if the true distribution is known, or obtained as a result of computational experiments aimed at finding the most effective option. The applicability of the method is demonstrated by solving the problem of determining the level of temperature anomalies of the mammary glands. It is shown that the considered temperature features are characterized by a normal distribution. Increasing the representativeness of the training set allowed training a classic classification algorithm – logistic regression – with an accuracy comparable to a multilayer neural network. This approach to the formation of a training dataset opens up the possibility of creating more transparent and interpretable artificial intelligence systems.

Keywords: machine learning, small datasets, data representativeness, data synthesis, neural networks, logistic regression.

References

1. Jakubik J., Vossing M., Kuhl N., Walk J. Satzger G. Data-Centric Artificial Intelligence. Business & Information Systems Engineering. 2024. vol. 66. no. 4. pp. 507–515.
2. Kumar S., Datta S., Singh V., Singh S.K., Sharma R. Opportunities and Challenges in Data-Centric AI. IEEE Access. 2024. vol. 12. pp. 33173–33189.
3. Hamid O.H. Data-Centric and Model-Centric AI: Twin Drivers of Compact and Robust Industry 4.0 Solutions. Applied Sciences. 2023. vol. 13. no. 5. DOI: 10.3390/app13052753.
4. Emmanuel T., Maupong T., Mpoeleng D., Semong T., Mphago B., Tabona O. A survey on missing data in machine learning. Journal of Big Data. 2021. vol. 8. no. 1. DOI: 10.1186/s40537-021-00516-9.
5. Gerber S, Pospisil L, Sys S, Hewel C, Torkamani A, Horenko I Co-Inference of Data Mislabelings Reveals Improved Models in Genomics and Breast Cancer Diagnostics. Frontiers in Artificial Intelligence 2022. vol. 4. DOI: 10.3389/frai.2021.739432.

6. Feng W., Quan Y., Dauphin G. Label Noise Cleaning with an Adaptive Ensemble Method Based on Noise Detection Metric. Sensors. 2020. no. 20(23). DOI: org/10.3390/s20236718.
7. Chehomova O.V. [On the problems of forming administrative responsibility for violations of patient rights]. Jurist - Pravoved – Jurist - Lawyer. 2010. no. 3. pp. 120–123. (In Russ.).
8. Sobjanin K.V., Kulikova S.P. [Unet-boosted classifier – a multi-task architecture forn small samples using the example of classifying MRI images of the brain]. Informatics and automation. 2024. vol. 23. no. 4. pp. 1022–1046. (In Russ.).
9. Zuev Ju.F. [Comparative assessment of the reliability of biological data in small samples using methods of dispersion, cluster and graph analysis (non-stochastic approach)]. Aktual'nye problemy gumanitarnyh i estestvennyh nauk – Current issues in the humanities and natural sciences. 2015. no. 1-1. pp. 36–45. (In Russ.).
10. Morgunova A.V., Sazhina O.S. [Using Nonparametric Methods of Mathematical Statistics to Search for Cosmic Strings]. Vestnik Moskovskogo universiteta. Serija 3. Fizika. Astronomija – Bulletin of Moscow University. Series 3. Physics. Astronomy. 2019. no. 5. pp. 86–92. (In Russ.).
11. Mokeev V.V., Tomilov S.V. [On the solution of the problem of small sample size when using linear discriminant analysis in face recognition problems]. Biznes-informatika – Business informatics. 2013. no. 1(23). pp. 37–43. (In Russ.).
12. Aleksahina K.A., Zhukov V.V. Osobennosti postroenija modelej mashinnogo obuchenija na malyh naborah dannyh v biomedicine [Features of building machine learning models on small data sets in biomedicine]. Informacionno-telekommunikacionnye tehnologii i matematicheskoe modelirovanie vysokotehnologichnyh system [Information and telecommunication technologies and mathematical modeling of high-tech system]. Peoples' Friendship University of Russia, 2021. pp. 151–155. (In Russ.).
13. Hu Y., Sandt R., Spatschek R. Practical feature filter strategy to machine learning for small datasets in chemistry. Scientific Reports. 2024. vol. 14. no. 1. DOI: 10.1038/s41598-024-71342-1.
14. Safonova A., Ghazaryan G., Stiller S., Main-Knorn M., Nendel C., Ryo M. Ten deep learning techniques to address small data problems with remote sensing. International Journal of Applied Earth Observation and Geoinformation. 2023. vol. 125. DOI: 10.1016/j.jag.2023.103569.
15. He H., Garcia E.A. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering. 2009. vol. 21. no. 9. pp. 1263–1284. DOI: 10.1109/TKDE.2008.239.
16. Polyakov M., Sirotin D. A Realistic Breast Phantom for Investigating the Features of the Microwave Radiometry Method Using Mathematical and Physical Modelling. Technologies. 2025. vol. 13(3). DOI: 10.3390/technologies13030106.
17. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002. vol. 16. pp. 321–357. DOI: 10.1613/jair.953.
18. Hauptmann T., Fellenz S., Nathan L., Tuscher O., Kramer S. Discriminative Machine Learning for Maximal Representative Subsampling. Scientific Reports. 2023. vol. 13. no. 1. DOI: 10.1038/s41598-023-48177-3.
19. Priestley M., O'Donnell F., Simperl E. A Survey of Data Quality Requirements That Matter in ML Development Pipelines. ACM Journal of Data and Information Quality. 2023. vol. 15. no. 2. pp. 1–39. DOI: 10.1145/3592616.
20. Levshinskii V. Mathematical models for analyzing and interpreting microwave radiometry data in medical diagnosis. Journal of computational and engineering mathematics. 2021. vol. 8. no. 1. pp. 3–14.

21. Lebedev A.V., Fadeeva L.N. Teorija verojatnostej i matematicheskaja statistika (4-e izdanie) [Probability Theory and Mathematical Statistics (4th edition)]. M: Nauka, 2018. 408 p. (In Russ.).
22. Kwon S., Lee S. Recent Advances in Microwave Imaging for Breast Cancer Detection. International journal of biomedical imaging. 2016. vol. 2016. pp. 1–25. DOI: 10.1155/2016/5054912.
23. Barrett A.H., Myers P.C. Subcutaneous Temperature: A method of Noninvasive Sensing. Science. 1975. vol. 190. pp. 669–671. DOI: 10.1126/science.1188361.
24. Barrett A.H., Myers Ph.C., Sadovsky N.L. Microwave Thermography in the Detection of Breast Cancer. AJR. American journal of roentgenolog. 1980. vol. 134(2). pp. 365–368. DOI: 10.2214/ajr.134.2.365.
25. Gautherie M. Temperature and Blood Flow Patterns in Breast Cancer During Natural Evolution and Following Radiotherapy. Progress in Clinical and Biological Research. 1982. pp. 21–64.
26. Troickij V.S. [On the theory of contact radiometric measurements of the internal temperature of bodies]. Radiofizika – Radiophysics. 1981. vol. 2. no. 9. pp. 1054–1061. (In Russ.).
27. Goryanin I., Karbainov S., Shevelev O., Tarakanov A., Redpath K., Vesnin S., Ivanov Yu. Passive microwave radiometry in biomedical studies. Drug Discovery Today. 2020. vol. 25. no. 4. pp. 757–763.
28. Popov I.E., Krylova A.E. [Analysis of brain thermometric data obtained by microwave radiothermometry]. Matematicheskaya fizika i kompyuternoe modelirovanie – Mathematical Physics and Computer Modeling. 2023. vol. 26. no. 2. pp. 32–42. DOI: 10.15688/mpcm.jvolsu.2023.2.3. (In Russ.).
29. Sedankin M.K., Gudkov A.G., Nelin I.V., Skuratov V.A., Leushin V.Yu, Chizhikov S.V., Sidorov I.A. Mathematical modeling of heat exchange processes in the brain in the presence of pathology for the design of a microwave radiothermograph. Bio-Medical Engineering. 2023. vol. 57. no. 4. pp. 276–279. DOI: 10.1007/s10527-023-10314-7.
30. Li J., Galazis C., Popov L., Ovchinnikov L., Kharybina T., Vesnin S., Losev A., Goryanin I Dynamic Weight Agnostic Neural Networks and Medical Microwave Radiometry (MWR) for Breast Cancer Diagnostics. Diagnostics. 2022. vol. 12. no. 9. DOI: 10.3390/diagnostics12092037.
31. Vesnin S.G., Kaplan M.A., Avakjan R.S. [Modern microwave radiothermometry of mammary glands]. Opuholi zhenskoj reproduktivnoj sistemy – Female reproductive system tumors. 2008. no. 3. pp. 28–33. DOI: 10.17650/1994-4098-2008-0-3-28-33. (In Russ.).
32. Losev A.G., Levshinskii V.V. [Microwave radiothermometry data mining in breast cancer diagnostics]. Matematicheskaja fizika i komp'juternoe modelirovanie – Mathematical Physics and Computer Modeling. 2017. vol. 20. no. 5. pp. 49–62. DOI: 10.15688/mpcm.jvolsu.2017.5.6. (In Russ.).
33. Losev A.G., Levshinskii V.V. [Intelligent analysis of thermometric data in breast diagnostics]. Upravlenie bol'shimi sistemami: sbornik trudov – Management of large systems: collection of works. 2017. no. 70. pp. 113–135. (In Russ.).
34. Levshinskii V., Galazis C., Ovchinnikov L., Vesnin S., Goryanin I. Application of Data Mining and Machine Learning in Microwave Radiometry (MWR). Communications in Computer and Information Science (CCIS). 2020. vol. 1211. pp. 265–288. DOI: 10.1007/978-3-030-46970-2 _13.
35. Germashev I.V., Dubovskaja V.I. [Application of fuzzy mathematics models for solving medical diagnostics problems]. Matematicheskaya fizika i kompyuternoe modelirovanie

– Mathematical Physics and Computer Modeling. 2021. vol. 24. no. 4. pp. 53–66. DOI: 10.15688/mpcm.jvolsu.2021.4.4.

36. Osmonov B., Ovchinnikov L., Galazis C., Emilov B., Karaibragimov M., Seitov M., Vesnin S., Losev A., Levshinskii V., Popov I., Mustafin C., Kasymbekov T., Goryanin I. Passive Microwave Radiometry for the Diagnosis of Coronavirus Disease 2019 Lung Complications in Kyrgyzstan. Diagnostics. 2021. vol. 11. no. 2. DOI: 10.3390/diagnostics11020259.
37. Losev A.G., Popov I.E., Gudkov A.G., Chizhikov S. V. [Intelligent analysis of microwave radiothermometry data in medical diagnostics]. Nanotehnologii: razrabotka, primenenie - XXI vek – Nanotechnology: development, application - 21st century. 2023. vol. 15. no. 1. pp. 5–22. DOI: 10.18127/j22250980-202301-01. (In Russ.).
38. Ahmad H.F., Mukhtar H., Alaqail H., Seliaman M., Alhumam A. Investigating Health-Related Features and Their Impact on the Prediction of Diabetes Using Machine Learning. Applied Sciences. 2021. vol. 11. no. 3. DOI: 10.3390/app11031173.

Losev Alexandr — Ph.D., Dr.Sci., Professor, Head of the institute, Institute of mathematics and information technology, Volgograd State University. Research interests: artificial intelligence, data mining, mathematical modeling, elliptic partial differential equations on Riemannian manifolds. The number of publications — 191. alexander.losev@volsu.ru; 100, Universitetsky Av., 400062, Volgograd, Russia; office phone: +7(904)774-6821.

Popov Illarion — Postgraduate student, Department of mathematical analysis and theory of functions, Volgograd State University. Research interests: artificial intelligence, data mining, mathematical modeling. The number of publications — 28. popov.larion@volsu.ru; 100, Universitetsky Av., 400062, Volgograd, Russia; office phone: +7(937)088-8938.

Reznikova Anastasia — Master's student, Department of mathematical analysis and theory of functions, Volgograd State University. Research interests: artificial intelligence, data mining, mathematical modeling. The number of publications — 11. a.reznikova@volsu.ru; 100, Universitetsky Av., 400062, Volgograd, Russia; office phone: +7(917)845-2535.

textbfAcknowledgements. This work is supported by the Russian Science Foundation No. 25-21-00330, https://rscf.ru/project/25-21-00330/.