УДК 004.93

АДАПТИВНЫЙ МЕТОД РЕДУКЦИИ РАЗМЕЧЕННЫХ ВЫБОРОК ДАННЫХ ДЛЯ ПОСТРОЕНИЯ ДИАГНОСТИЧЕСКИХ МОДЕЛЕЙ

Каврин Д.А., Субботин С.А.

Запорожский национальный технический университет, г. Запорожье, Украина
ORCID: [1] https://orcid.org/0000-0002-8952-4067, [2] https://orcid.org/0000-0001-5814-8268
E-mail: [1] kavrin@gmail.com, [2] subbotin.csit@gmail.com

DOI: 10.15673/atbp.v10i3.1084

Аннотация: *Решена актуальная задача редукции размеченных выборок данных большого размера путем извлечения подвыборок меньшего размера для построения диагностических и распознающих моделей по прецедентам.*

Предложен детерминированный метод редукции размеченных выборок, который использует информацию о классах для извлечения репрезентативных выборок небольшого размера. Предложенный метод последовательно разбивает исходную выборку на гиперсферы, радиусы которых определяются расстояниями до ближайших экземпляров противоположного класса. Из центров полученных гиперсфер формируется подвыборка меньшего размера. Благодаря адаптивности радиуса каждой гиперсферы к расстоянию до ближайшего экземпляра противоположного класса в редуцированной выборке сохраняется большинство наиболее важных экземпляров, которые находятся вблизи границ классов. Это позволяет извлекать репрезентативные выборки с хорошо определенными межклассовыми границами. Метод базируется на гипотезе о компактности классов, поэтому объем сокращенной выборки сильно зависит от степени разделимости классов. Например, если классы компактны, объем редуцированной выборки может быть слишком малым с плохо определенными границами классов. Для решения данной проблемы, предлагается регулировать объем извлекаемой выборки, изменяя радиусы гиперсфер с помощью долевого коэффициента. Таким образом, можно более точно определять границы классов, повышая репрезентативность редуцированных выборок.

Для обработки очень больших исходных выборок, когда объем данных не позволяет загрузить их полностью в память ЭВМ, либо данные поступают динамически, предложенный метод позволяет обрабатывать исходную выборку пакетами заданного объема.

Разработано программное обеспечение, реализующее предложенный метод, которое позволяет проводить вычислительные эксперименты по исследованию его свойств, при решении задач редукции размеченных выборок данных большого размера.

Abstract: *The urgent problem of reducing the large labeled dataset by extracting smaller samples for constructing diagnostic and recognition models using precedents is solved.*

A deterministic method of reduction of labeled datasets is proposed, which uses information about classes to extract representative samples of small size. The proposed method successively splits the target dataset into hyperspheres, the radii of which are determined by the distances to the nearest instances of the opposite class. From the centers of the obtained hyperspheres, a smaller sub-sample is formed. Due to the adaptability of the radius of each hypersphere to the distance to the closest instance of the opposite class, the most related instances that are near the class boundaries remain in the reduced sample. This allows us to extract representative instances with well-defined interclass boundaries. The method is based on the compactness hypothesis of classes; therefore the volume of a reduced sample strongly depends on the degree of class separation. For example, if the classes are compact, the volume of the reduced sample may be too small with poorly defined class boundaries. To solve this problem, it is proposed to regulate the volume of the extracted sample by changing the radii of the hyperspheres using the proportional coefficient. Thus, it is possible to more accurately determine the boundaries of classes, increasing the representativeness of the reduced samples.

For processing original large datasets, when the amount of data does not allow loading them completely into the computer's memory or the data is dynamically received, the proposed method allows processing the original sample with packets of a given volume.

The software has been developed that implements the proposed method, which makes it possible to carry out computational experiments on the study of its properties in solving the reduction problems of the large labeled datasets.

Ключевые слова: выборка, диагностирование, классификация, класс, кластер, метрика, экземпляр.
Keywords: class, classification, cluster, diagnosis, instance, metric, sample.

Введение

Современные вычислительные системы позволяют собирать большие массивы данных, характеризующие работу сложных технических объектов и процессов в реальном времени. Это предоставляет большие возможности для эффективного решения задач технической диагностики и прогнозирования в оперативном режиме.

Объектом исследования данной работы является процесс построения диагностических моделей по прецедентам.

При решении задач построения диагностических моделей на практике часто приходится сталкиваться с выборками очень большого размера, обработка которых может потребовать значительных затрат времени и ресурсов памяти ЭВМ. Поэтому актуальной является задача сокращения объема выборок данных.

Предмет исследования составляют методы формирования обучающих выборок из исходных размеченных выборок большого объема.

Применение известных методов формирования выборок [1–4] для извлечения данных из исходных выборок, имеющих информацию о классах, не могут гарантировать, что полученная выборка малого размера будет отображать свойства исходной выборки, особенно вблизи границ классов.

Целью работы являлась разработка метода извлечения выборок, позволяющего минимизировать размер и сохранить топологическую карту основных свойств исходных данных в пространстве признаков.

Постановка задачи

Пусть задана несбалансированная выборка $X = \langle x, y \rangle$ – набор S прецедентов о зависимости $y(x), x = \{x^s\}, y = \{y^s\}, s = 1,2,...,S$, характеризующихся набором N входных признаков $\{x_j\}, j = 1,2,...,N$, и выходным признаком y. Каждый s-й прецедент представим как $\langle x^s, y^s \rangle, x^s = \{x_j^s\}, y^s \in \{1,2,...,K\}, K > 1$, где K – число классов в выборке. Тогда задача сокращения объема исходной выборки состоит в извлечении из исходной выборки $X = \langle x, y \rangle$ такой подвыборки меньшего объема $X' = \langle x', y' \rangle$, которая сохранила бы наиболее важные топологические свойства исходной выборки. Формально данное условие может быть представлено в следующем виде: $x' \in \{x^s\}, y' = \{y^s \mid x^s \in x'\}, S' \le S, f(\langle x', y' \rangle, \langle x, y \rangle) \to opt$, где S' – число экземпляров в редуцированной выборке; x' – набор признаков в редуцированной выборке; y' – выходной признак (класс) в редуцированной выборке, opt – условное обозначение оптимума.

Литературный обзор

Методы редукции исходных выборок путем извлечения подвыборок меньшего размера условно можно разделить на две основные категории: вероятностные и детерминированные методы [1–12] (рис. 1).

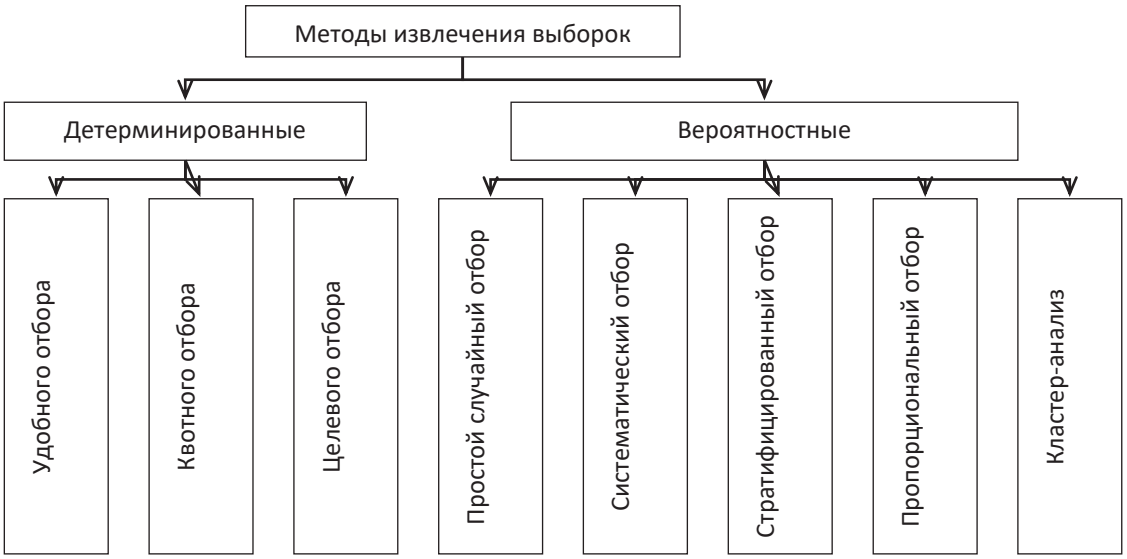

Рис. 1 – Классификация методов извлечения выборок

Вероятностные методы предполагают случайное извлечение каждого экземпляра (группы экземпляров) из исходной выборки с известной ненулевой вероятностью, которая может быть точно определена.

Вероятностные методы извлечения подвыборок [1–10] включают в себя:

– простой случайный отбор (simple random sampling) [1–5]: из исходной выборки случайным образом отбирается заданное число экземпляров. Причем все экземпляры исходной выборки имеют одинаковую вероятность быть выбранными;

– систематический отбор (systematic sampling) [1–4, 6]: исходная выборка упорядочивается определенным образом и разбивается на последовательные группы экземпляров, затем из каждой выбирается объект с заданным порядковым номером в группе для включения в формируемую подвыборку;

– стратифицированный отбор (stratification sampling) [1–5, 7]: исходная выборка делится на непересекающиеся однородные подмножества (страты), включающие все виды экземпляров, затем к каждому подмножеству применяется случайный или систематический отбор;

– вероятностно-пропорциональный отбор (probability proportional to size sampling) [1–4, 8]: применяется при наличии дополнительной информации о классах и их объеме, при этом вероятность выбора каждого элемента исходной выборки будет пропорциональна объему класса, к которому он принадлежит;

– отбор на основе кластер-анализа (cluster sampling) [9–10]: экземпляры исходной выборки делятся на кластеры, из каждого кластера случайно выбирается некоторое подмножество экземпляров для формируемой выборки.

Достоинствами вероятностных методов [2, 5] являются их относительная простота и возможность оценки ошибки выборки, а недостатками – то, что они не гарантируют, что сформированная выборка малого объема будет хорошо отображать свойства исходной выборки, либо не будет избыточной, и не будет искусственно упрощать задачу.

Детерминированные методы формирования выборок [2, 11–12] предполагают извлечение экземпляров на основе предположений об их информативности, которая формирует критерии отбора. При этом данные выборки содержат экземпляры, которые могут не быть выбраны или вероятность их выбора не может быть точно определена. Поэтому к таким выборкам неприменима теория, разработанная для вероятностных выборок. К детерминированным методам формирования выборок относят следующие методы:

– удобного отбора (convenience sampling) [2, 12]: формирует нерепрезентативную выборку из наиболее легко доступных для исследования объектов;

– квотного отбора (quota sampling) [2]: исходная выборка разделяется на непересекающиеся подгруппы с отличающимися свойствами, после чего из каждой подгруппы выбираются объекты на основе заданной пропорции и на основании предпочтений исследователя;

– целевого отбора (purposive sampling) [2, 12]: объекты извлекаются из исходной выборки исследователем в соответствии с его мнением относительно их пригодности для исследования.

Недостатком данных методов [11–12] является невозможность оценивания ошибки сформированных выборок. Достоинством детерминированных методов является то, что они могут выявить наиболее значимые для решения задачи построения диагностической модели прецеденты, которые также могут быть использованы для инициализации распознающих моделей и ускорения процесса обучения.

Исходные выборки данных, используемые при решении задачи построения диагностической модели, могут быть очень большого объема, либо иметь избыточные данные. Работа модели с использованием таких выборок может потребовать значительных вычислительных и временных ресурсов. Извлечение подвыборок меньшего объема, при построении модели, является естественным и достаточно эффективным решением данной проблемы. При этом важно сохранить значимые экземпляры исходной выборки для получения репрезентативной подвыборки меньшего объема. Для извлечения подвыборок меньшего размера, при построении диагностических моделей, наиболее подходящими являются детерминированные методы на основе кластер-анализа, так как они позволяют выявить наиболее значимые экземпляры. При решении задач технической диагностики, как правило, присутствует информация о классах. Это значительно упрощает задачу кластеризации, которую можно выполнить более точно, особенно на границе классов, опираясь на данные о классах. Однако, в общем случае, кластер-анализ решает задачу кластеризации для неразмеченных данных и не позволяет выделять важные экземпляры, расположенные вблизи границ классов.

Поэтому для создания репрезентативных выборок небольшого размера с хорошо определенными границами классов необходимо разработать метод формирования выборок из исходных размеченных данных, который позволит уменьшить вычислительную нагрузку и при этом сохранить топологическую репрезентативность исходной выборки в пространстве признаков за счет сохранения значимых экземпляров на границе классов.

Адаптивный метод редукции выборок большого размера

Для формирования сокращенной подвыборки в предлагаемом методе исходная размеченная выборка разбивается на гиперсферы различных радиусов в пространстве признаков, зависящих от расстояний между экземплярами различных классов. Для этого выбирается наиболее перспективный экземпляр класса, который находиться ближе остальных к центру масс данного класса. Затем в пространстве признаков строится гиперсфера с центром в этой точке и радиусом равным расстоянию до ближайшего экземпляра противоположного класса. Все экземпляры класса, попавшие в гиперсферу, удаляются из дальнейшего рассмотрения. Процедура повторяется до тех пор, пока все экземпляры класса не будут исключены из рассмотрения.

Сокращенная выборка формируется из центров полученных гиперсфер. Построение гиперсфер производится отдельно для каждого класса исходной выборки. Таким образом, метод как бы адаптируется к распределению данных в выборке, автоматически регулируя число экземпляров в редуцированной выборке. Благодаря адаптивности радиусов гиперсфер, ближе к границе классов формируется больше гиперсфер меньшего радиуса (рис. 2).

Другими словами, в обработанной выборке плотность экземпляров на границе классов будет выше, чем вдали от границ, что позволяет более точно определить границы классов.

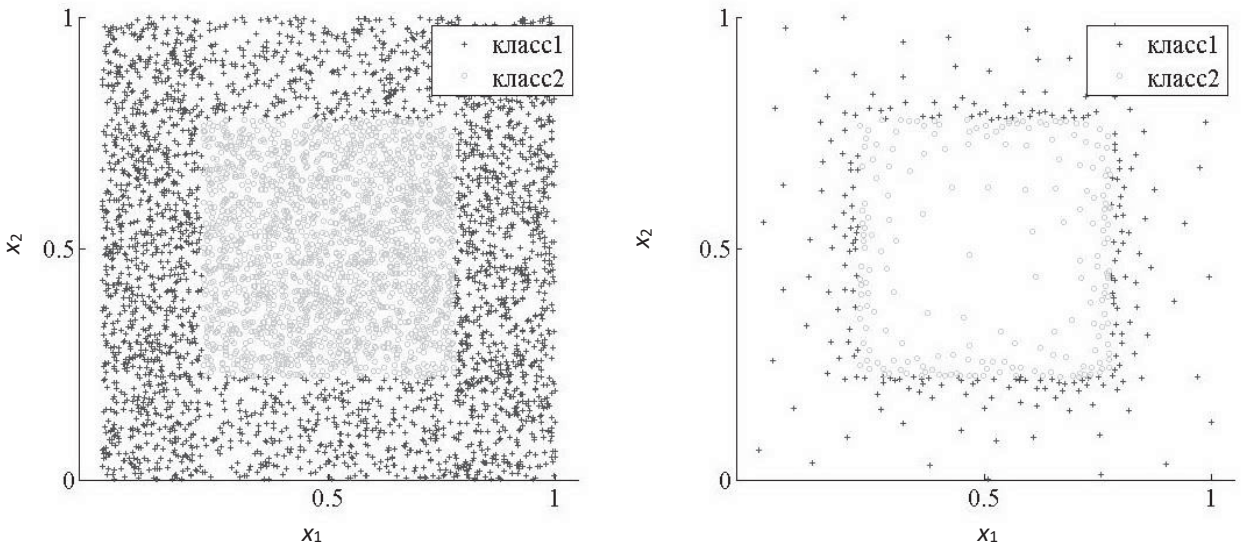

Рис. 2 – Визуализация синтетической выборки в пространстве признаков: а – до обработки, б – после обработки предложенным методом

В основе предложенного метода лежит гипотеза о компактности классов [13]. В идеальном случае, когда классы не пересекаются (компактны), предложенный метод очень сильно сокращает исходную выборку, в том числе, удаляя значимые экземпляры на границах классов, искажая тем самым границы классов. В этом случае, для формирования более представительной выборки предлагается регулировать число гиперсфер долевым изменением их радиусов. При уменьшении радиусов гиперсфер, число кластеров увеличивается, причем основная масса кластеров накапливается именно на границах классов. Таким образом, в формируемой выборке, можно регулировать число экземпляров, особенно на границе классов, повышая ее репрезентативность.

Ниже представлено поэтапное описание предлагаемого метода.

Этап 1. Инициализация. Задать исходную выборку данных $X\langle x, y\rangle$, инициализировать редуцированную выборку $X'\langle x', y'\rangle = \varnothing$.

Этап 2. Разбиение выборки по классам. Разбить исходную выборку X на K отдельных подвыборок $X(k)$ для экземпляров каждого класса: $X(k) = \bigcup_{s=1}^{S}\{x^s \mid y^s = k\}$, где $k = 1,...,K$. Определить объем (число экземпляров) каждой k-й выборки S_k.

Этап 3. Редукция подвыборок. Установить $k = 0$. Затем, пока $k < K$, выполнять в цикле: принять $k = k + 1$, установить основную подвыборку $A = X(k)$, объединить все остальные подвыборки в одну виртуальную подвыборку $B = \bigcup_{s=1}^{S}\{x^s \mid y^s \neq k\}$, далее, пока $A \neq \varnothing$ выполнять в цикле:

– найти координаты центра масс (эталона) подвыборки A:

$$C_j^A = \frac{1}{S_A}\sum_{s=1}^{S_A}\left\{x(k)_j^s\right\}, j = 1,2,...,N, k = 1,2,...,K;$$

– определить наиболее перспективную точку (экземпляр) x^a подвыборки A, для которой будет наименьшим расстояние до центра масс подвыборки A:

$$d(x^a, C^A) = \arg\min_{a=1,2,...,S_k}\left\{\sqrt{\sum_{j=1}^{N}(x_j^a - C_j^A)^2}\right\};$$

– добавить наиболее перспективный экземпляр x^a в выборку X': $X' = X' \bigcup x^a$;

– рассчитать расстояние от x^a до ближайшего к нему экземпляра подвыборки B:

$$r = \arg\min_{b=1,2,...,S_B} \left\{ \sqrt{\sum_{j=1}^{N} (x_j^a - x_j^b)^2} \right\};$$

– определить радиус гиперсферы с центром в точке x^a: $R = \lambda r$, где λ – долевой коэффициент;

– удалить из подвыборки A все экземпляры, попавшие в гиперсферу радиуса R: $A = A \setminus A^R$, где A^R – множество экземпляров удаленных от центра гиперсферы x^a на расстояние $d(x^a, x^s) < R$;

Этап 4. Построить распознающую модель по редуцированной выборке X'.

Предлагаемый метод позволяет автоматизировать процесс сокращения объема исходной выборки, имеющей информацию о классах, для решения задач технической диагностики.

Недостатком метода является необходимость расчета и хранения в памяти попарных расстояний между экземплярами противоположных классов.

Поэтому, если объем исходной выборки достаточно большой и не позволяет одновременно загрузить в память ЭВМ все попарные расстояния, либо данные поступают динамически, можно производить обработку исходной выборки пакетами.

Для корректной работы метода, каждый пакет должен быть представлен всеми классами. Применив к первому пакету данных описанный метод, получаем первичное распределение извлеченной выборки. Экземпляры следующего пакета данных объединяются с экземплярами выборки, полученной на предыдущем этапе. Объединенная выборка снова обрабатывается методом адаптивной редукции. Процедура повторяется, пока не будут обработаны все данные исходной выборки, либо пока не будет предпринята остановка в заданной фазе потока данных, если данные динамические.

Метод пакетной обработки данных можно представить следующими этапами.

Этап 1. Инициализация. Задать исходную выборку данных $X\langle x, y\rangle$ и редуцированную выборку $X'\langle x', y'\rangle = \varnothing$.

Этап 2. Инициализация пакетов. Определить число пакетов в выборке $P = \text{round}(S/Q)$, где Q – число экземпляров в пакете, задаваемое пользователем, round – функция округления аргумента до ближайшего целого числа. Разбить исходную выборку X на P пакетов $X(\rho)$, где $\rho = 1,...,P$.

Этап 3. Обработка пакетов. Установить $\rho = 0$. Пока $\rho < P$, выполнять в цикле: принять $\rho = \rho + 1$, объединить пакет $X(\rho)$ с редуцированной выборкой X': $X(\rho) = X' \bigcup X(\rho)$, затем обработать пакет $X(\rho)$, описанным выше, адаптивным методом редукции, установить $X' = X(\rho)$.

Этап 4. Построить распознающую модель по редуцированной выборке X'.

Предложенный пакетный метод обработки выборки позволяет обрабатывать очень большие исходящие выборки, пакетами заданного размера.

Пакетная обработка данных в комплексе с методом адаптивной редукции позволяет получать представительные сокращенные выборки из достаточно больших выборок данных, либо динамических наборов данных, не требуя значительных вычислительных ресурсов.

Эксперименты и результаты

Для экспериментального исследования работы предложенного метода редукции выборок было разработано программное обеспечение, реализующее предложенный метод.

На этапе предварительной подготовки признаки выборок нормировались по формуле:

$$x_j^s = \frac{x_j^s - x_j^{\min}}{x_j^{\max} - x_j^{\min}},$$

где x_j^s – j-й признак s-го экземпляра выборки, $x_j^{\min}$ – минимальное значение j-го признака, $x_j^{\max}$ – максимальное значение j-го признака.

Распознающая модель строилась с использованием метода kNN (k ближайших соседей) [14]. Решающие правила строились по принципу большинства голосов. Поэтому для однозначности выбора в работе использовались метод с нечетным числом ближайших соседей (k = 1, 3, 9, 25, 49).

Были использованы два набора данных для решения практических задач [15, 16]. Также, для чистоты эксперимента, использовалась синтетическая выборка с непересекающимися (компактными) классами [13].

Затем, для оценки работы модели каждый исходный набор данных был разделен на тестовую и обучающую выборки методом стратификации [1-3, 7] в соотношении 25/75. Таким образом, проводилась редукция обучающей выборки и далее, с помощью тестовой выборки, определялось качество работы полученной модели.

Анализ эффективности работы метода проводился путем измерения и сравнения точности классификации исходной и редуцированной выборок.

Оценка производительности производилась с помощью метрики гармоничного среднего F-measure. Данная метрика позволяет исключить влияние дисбаланса классов на оценку производительности [17].

Результаты исследований метода адаптивной редукции выборок представлены на рисунках 3-5. Рисунки 3а, 4а, 5а иллюстрируют зависимости гармоничного среднего (F-measure) от числа ближайших соседей классификатора kNN для различных долей радиусов гиперсфер. Рисунки 3б, 4б, 5б отображают зависимости объемов обработанных обучающих выборок от долей радиусов гиперсфер.

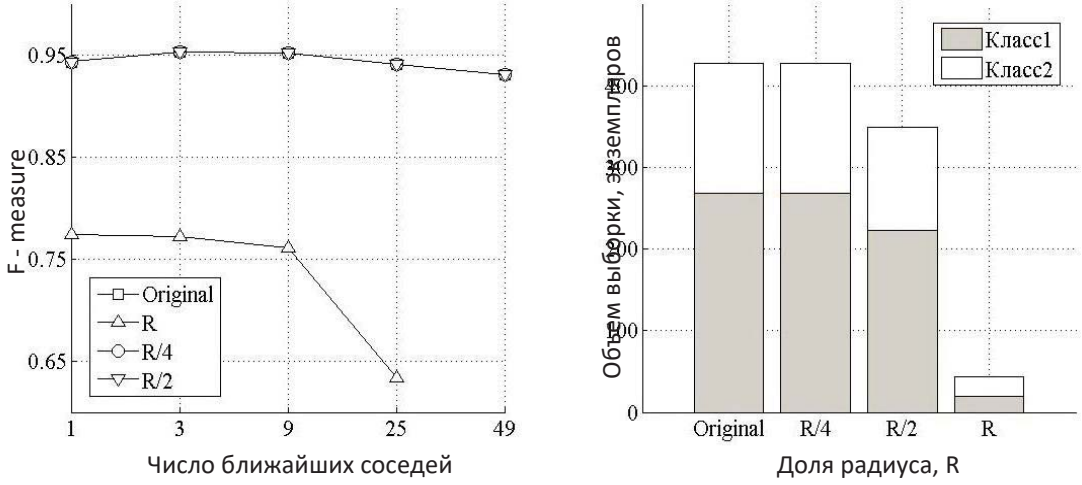

Рис. 3 – Графики зависимостей параметров тестовой модели для задачи предсказания рака молочной железы [16]: а – метрики F-measure от числа ближайших соседей метода kNN для различных долей радиуса гиперсферы, б – числа экземпляров от доли радиуса гиперсферы

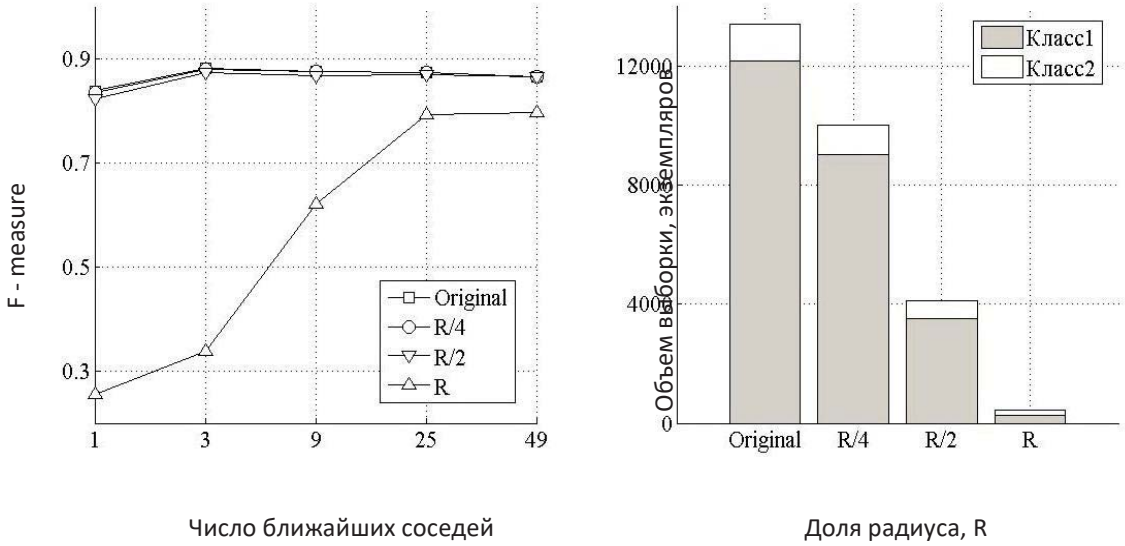

Рис. 4 – Графики зависимостей параметров тестовой модели для задачи для определения пульсаров [15]: а – метрики F-measure от числа ближайших соседей метода kNN для различных долей радиуса гиперсферы, б – числа экземпляров от доли радиуса гиперсферы

Следующим этапом исследований было изучение работы метода адаптивной редукции в комплексе с методом пакетной обработки данных. Для обеспечения необходимых условий эксперимента к предварительной обработке данных, используемой на первом этапе исследований, добавлялась процедура разбиения обучающей выборки на пакеты заданного размера. Для изучения работы данного подхода использовалась бинарная синтетическая выборка объемом 100 тыс. экземпляров и размерностью два признака, которая разбивалась на пакеты по 2000 экземпляров. Далее измерялись описанные выше параметры.

На рисунке 6 представлен результат работы адаптивного метода редукции в комплексе с пакетной обработкой.

Проведенные эксперименты показали, что благодаря своей адаптивности, предложенный метод хорошо работает в автоматических режимах, что является важным фактором при решении задач технической диагностики. Метод базируется на гипотезе о компактности классов [13], поэтому объем сокращенной выборки сильно зависит от степени разделимости классов (компактности). Так, для выборки с сильно перемешанными классами, объем обработанной выборки может практически не измениться, либо измениться незначительно. В свою очередь, если классы хорошо разделимы (компактны), объем редуцированной выборки будет достаточно небольшим и границы классов будут плохо определены, поэтому, изменяя радиусы гиперсфер, можно регулировать число экземпляров вблизи границ классов, повышая репрезентативность выборки. Таким образом, при решении практических задач, предложенный метод позволяет в зависимости от распределения данных в выборке регулировать объем и представительность редуцированной выборки, изменяя всего один параметр – долю радиусов гиперсфер.

Из рис. 3-6 видно, что репрезентативность редуцированных выборок всех исследуемых наборов данных, определяемая в тестовой модели параметром гармоничного среднего, при доле радиуса гиперсферы 50% практически была аналогичной репрезентативности оригинальной исходной выборки. При этом относительные объемы редуцированных выборок отличались, что было обусловлено различным распределением данных в исходных выборках.

Так в задаче предсказания рака молочной железы [16] (рис. 3), объем сокращенной выборки при радиусах гиперсфер $R/2$ уменьшился незначительно, в свою очередь в задаче определения пульсаров [15] (рис. 4) объем сокращенной выборки при радиусах гиперсфер $R/2$ уменьшился больше чем в три раза. При этом представительность сокращенных выборок оставалась практически эквивалентной исходным выборкам.

Синтетические выборки формировались в условиях абсолютной компактности классов, и поэтому объем сокращенных выборок значительно уменьшался. В этом случае уменьшение радиусов гиперсфер в два раза $R/2$ позволило сохранить значимые экземпляры на границах классов и получить репрезентативные выборки данных (рис. 5-6).

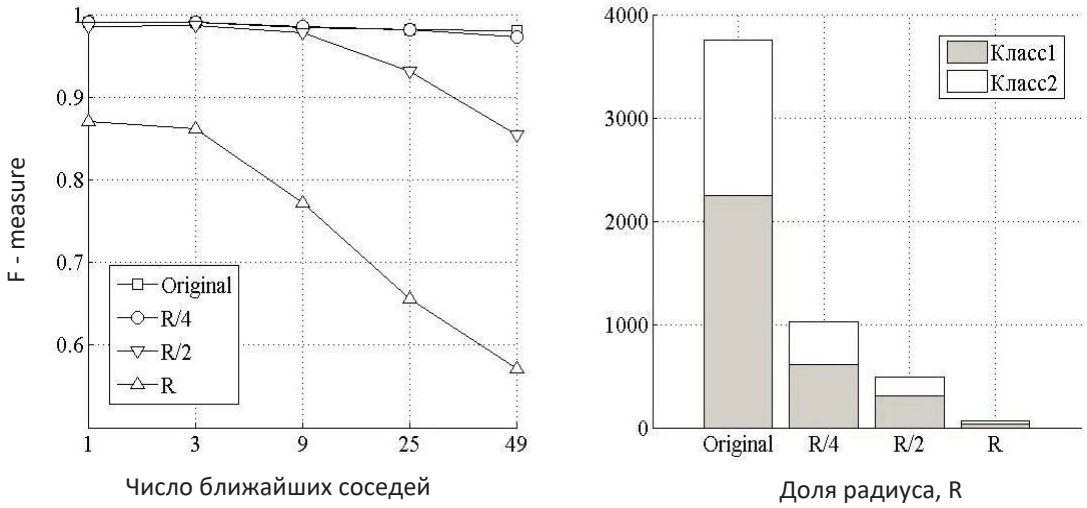

Рис. 5 – Графики зависимостей параметров тестовой модели для синтетической выборки объемом 5 тыс. экземпляров: а – метрики F-measure от числа ближайших соседей метода kNN для различных долей радиуса гиперсферы, б – числа экземпляров от доли радиуса гиперсферы

Регулируя долю радиусов гиперсфер, предложенный метод позволяет находить компромисс между объемом и репрезентативностью редуцированных выборок, в зависимости от поставленных задач. Исходя из проведенных экспериментов, для исходных выборок с хорошо разделенными классами, предлагается использование значений долевого коэффициента $\lambda = [0.25, 0.5]$, так как изменения радиусов гиперсфер в данных пределах несущественно влияет на точность классификации модели, при этом существенно уменьшая объем исходной выборки.

Недостатком представленного метода является то, что он является затратным с вычислительной точки зрения, особенно для большеразмерных выборок.

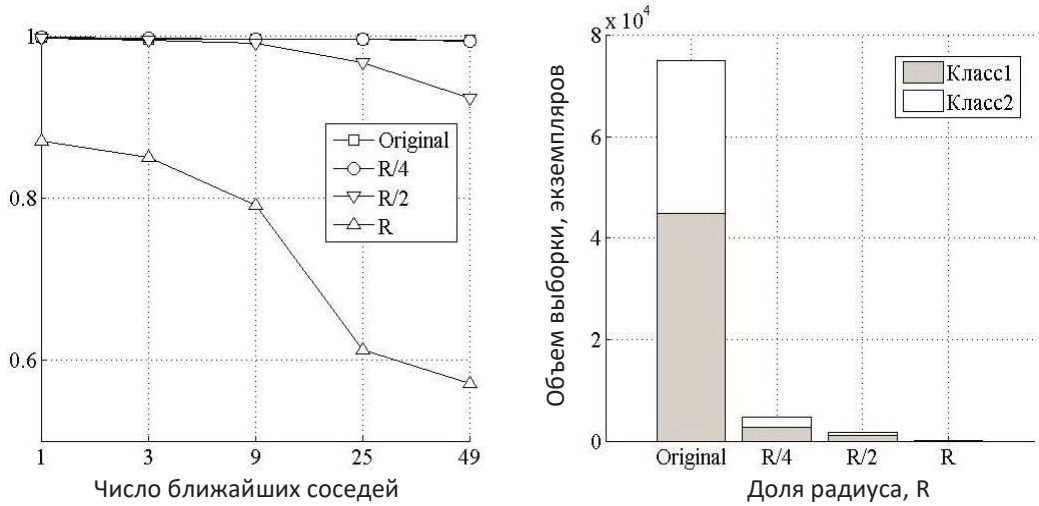

Рис. 6 – Графики зависимостей параметров тестовой модели для синтетической выборки объемом 100 тыс. экземпляров: а – метрики F-measure от числа ближайших соседей метода kNN для различных долей радиуса гиперсферы, б – числа экземпляров от доли радиуса гиперсферы

Поэтому существует необходимость применения подходов позволяющих уменьшить вычислительную нагрузку. Например, в случае большеразмерных выборок, возможно использование метода в ансамбле с методами сокращения размерности [18]. Также можно уменьшить вычислительную нагрузку, исключив из расчетов этап определения наиболее перспективной точки, заменив его случайным выбором любой точки класса. Однако такой подход требует дополнительного изучения и определения необходимых ограничений. Следующим подходом может быть распараллеливание вычислительной нагрузки с использованием многопроцессорных систем.

Выводы

Рассмотрена задача редукции размеченных выборок данных большого размера для построения диагностических и распознающих моделей по прецедентам.

Научная новизна полученных результатов состоит в том, что создан новый метод, редукции данных, позволяющий значительно сократить размер исходной размеченной выборки, сохраняя наиболее значимые экземпляры, находящиеся вблизи границ классов и удаляя менее информативные экземпляры, находящиеся внутри классов, либо экземпляры, удаленные от границ классов. Таким образом, предложенный метод позволяет в автоматическом режиме решать задачу редукции, адаптируясь к распределению данных в размеченной выборке.

Практическая значимость полученных результатов состоит в том, что разработано программное обеспечение, реализующее предложенный метод, с возможностью пакетной обработки выборок большого размера, либо выборок, которые формируются из динамически поступающих данных. Данное программное обеспечение было экспериментально исследовано при решении задач редукции синтетических и практических размеченных выборок. Проведенные эксперименты подтвердили работоспособность разработанного математического обеспечения. Результаты проведенных экспериментов позволяют рекомендовать использование разработанного метода и его программной реализации для решения задач технического и биомедицинского диагностирования.

Перспективы дальнейших исследований могут заключаться в изучении предложенного метода на более широком классе практических задач. Изучение возможностей метода в ансамблях с методами сокращения размерностей, при работе с выборками большой размерности. Разработка реализаций предложенного метода для многопроцессорных систем, работающих в параллельных режимах.

Список использованных источников

[1] Thompson S. K. Sampling. Hoboken: John Wiley & Sons, 2012. 472 p.
[2] Encyclopedia of survey research methods / ed. P. J. Lavrakas. Thousand Oaks: Sage Publications, 2008. Vol. 1–2. 968 p.
[3] Кокрен У. Методы выборочного исследования / пер. с англ. И. М. Сонина; под ред. А. Г. Волкова, Н. К. Дружинина. Москва: Статистика, 1976. 440 с.
[4] Chaudhuri A., Stenger H. Survey sampling theory and methods. New York: Chapman & Hall, 2005. 416 p.
[5] Tille Y., Wilhelm M. Probability Sampling Designs: Principles for Choice of Design and Balancing //Statistical Science. 2017. Vol. 32, Issue 2. P. 176–189.
[6] Kalton G. Systematic Sampling //Wiley StatsRef: Statistics Reference Online. 2017. [Электронный ресурс]. Режим доступа: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat03380.pub2.

[7] Parsons V. L. Stratified Sampling //Wiley StatsRef: Statistics Reference Online. 2017. [Электронный ресурс]. Режим доступа: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat05999.pub2.
[8] Skinner C. J. Probability Proportional to Size (PPS) Sampling //Wiley StatsRef: Statistics Reference Online. 2016. [Электронный ресурс]. Режим доступа: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat03346.pub2.
[9] Nelson G. A. Cluster Sampling: A Pervasive, Yet Little Recognized Survey Design in Fisheries Research //Transactions of the American Fisheries Society. 2014. Vol. 143, Issue 4. P. 926–938.
[10] Ly T., Cockburn M, Langholz B. Cost-efficient case-control cluster sampling designs for population-based epidemiological studies //Spatial and Spatio-temporal Epidemiology. 2018. Vol. 26. P. 95–105.
[11] Elliott M. R., Valliant R. Inference for Nonprobability Samples //Statistical Science. 2017. Vol. 32, Issue 2. P. 249–264.
[12] Etikan I., Musa S. A., Alkassim R. S. Comparison of Convenience Sampling and Purposive Sampling //American Journal of Theoretical and Applied Statistics. 2016. Vol. 5, Issue 1. P. 1–4.
[13] Загоруйко Н. Г. Прикладные методы анализа данных и знаний. Новосибирск: ИИМ, 1999. 270 с.
[14] Flach P. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. New York: Cambridge University Press, 2012. 409 p.
[15] Lyon R. J. HTRU2 [Электронный ресурс]. Режим доступа: https://figshare.com/articles/HTRU2/3080389/1.
[16] Breast Cancer Wisconsin (Diagnostic) Data Set [Электронный ресурс]. Режим доступа: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).
[17] Каврин Д. А., Субботин С. А. Метод редукции мажоритарного класса в несбалансированных выборках //Реєстрація, зберігання і обробка даних. 2018. Т. 20, № 1. С. 51–59.
[18] Субботін С. О., Олійник А. О. Інтелектуальні системи: навч. посіб. / під заг. ред. проф. С. О. Субботіна. Запоріжжя: ЗНТУ, 2014. 218 с.

References

[1] S. K. Thompson, *Sampling*. Hoboken: John Wiley & Sons, 2012.
[2] *Encyclopedia of survey research methods* / ed. P. J. Lavrakas. Thousand Oaks: Sage Publications, 2008.
[3] U. Kokren, *Metody vyborochnogo issledovanija*. Moskva: Statistika, 1976.
[4] A. Chaudhuri, H. Stenger, *Survey sampling theory and method*. New York: Chapman & Hall, 2005.
[5] Y. Tille, M. Wilhelm, “Probability Sampling Designs: Principles for Choice of Design and Balancing”, *Statistical Science*, vol. 32, issue 2, pp. 176–189, 2017.
[6] G. Kalton, “Systematic Sampling”, in *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, [online document], 2014. Available: Wiley Online Library, https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat03380.pub2 [Accessed: Feb. 15, 2017].
[7] V. L. Parsons, “Stratified Sampling”, in *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, [online document], 2014. Available: Wiley Online Library, https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat05999.pub2 [Accessed: Feb. 15, 2017].
[8] C. J. Skinner, “Probability Proportional to Size (PPS) Sampling”, in *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, [online document], 2014. Available: Wiley Online Library, https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat03346.pub2 [Accessed: Aug. 05, 2016].
[9] G. A. Nelson, “Cluster Sampling: A Pervasive, Yet Little Recognized Survey Design in Fisheries Research”, *Transactions of the American Fisheries Society*, vol. 143, issue 4, pp. 926–938, 2014.
[10] T. Ly, M. Cockburn and B. Langholz, “Cost-efficient case-control cluster sampling designs for population-based epidemiological studies”, *Spatial and Spatio-temporal Epidemiology*, vol. 26, pp. 95–105, 2018.
[11] M. R. Elliott, R. Valliant, “Inference for Nonprobability Samples”, *Statistical Science*, vol. 32, issue 2, pp. 249–264, 2017.
[12] I. Etikan, S. A. Musa and R. S. Alkassim, “Comparison of Convenience Sampling and Purposive Sampling”, *American Journal of Theoretical and Applied Statistics*, vol. 5, issue 1, pp. 1–4, 2016.
[13] N. G. Zagorujjko, *Prikladnye metody analiza dannykh i znanijj*. Novosibirsk: IIM, 1999.
[14] P. Flach, *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. New York: Cambridge University Press, 2012.
[15] R. J. Lyon, “HTRU2” [Online]. Available: https://figshare.com/articles/HTRU2/3080389/1. [Accessed: Mar. 01, 2016].
[16] Breast Cancer Wisconsin (Diagnostic) Data Set, “UCI Machine Learning Repository” [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic). [Accessed: Nov. 01, 1995]
[17] D. A. Kavrin, S. A. Subbotin, “Metod redukcii mazhoritarnogo klassa v nesbalansirovannyh vyborkah”, *Reiestratsiia, zberihannia i obrobka danykh*, t. 20, № 1, s. 51–59, 2018.
[18] S. O. Subbotin, A. O. Oliinyk, *Intelektualni systemy: navch. posib*. / pid zah. red. prof. S. O. Subbotina. Zaporizhzhia: ZNTU, 2014.