КОМПЬЮТЕРНЫЕ НАУКИ

УДК 004.93

ИЗВЛЕЧЕНИЕ СБАЛАНСИРОВАННЫХ ОБУЧАЮЩИХ ВЫБОРОК МЕТОДОМ ПСЕВДОКЛАСТЕРИЗАЦИИ

КАВРИН Д.А., СУББОТИН С.А.

Решается задача формирования обучающих выборок из размеченных несбалансированных наборов данных большого объема для построения диагностических и распознающих моделей по прецедентам. Предлагается метод восстановления баланса классов, который позволяет извлечь из исходных данных сбалансированные обучающие подвыборки значительно меньшего объема. Проведенные эксперименты подтверждают работоспособность разработанного математического обеспечения и позволяют рекомендовать его для использования на практике при решении задач технической диагностики и распознавания образов по признакам.

Ключевые слова: выборка, несбалансированность классов, мажоритарный класс, миноритарный класс, отбор экземпляров, экземпляр.

Key words: sample, imbalance, majority class, minority class, instance selection, instance.

Введение

Для построения диагностических и прогнозирующих моделей по прецедентам необходимо выделять обучающую и тестовую выборки из заданной исходной выборки [1, 2].

Объектом исследования является процесс построения моделей на основе набора прецедентов. На практике исходные выборки данных часто имеют большой объем и, соответственно, требуют значительных затрат времени и машинных ресурсов на обработку. Другой проблемой исходных данных является дисбаланс классов, который встречается в подавляющем большинстве наборов данных [3, 4]. Большинство существующих методов классификации [2–5] не рассчитаны на работу в условиях дисбаланса классов, поэтому чем выше уровень дисбаланса классов в выборке, тем больше вероятность неверной работы построенной модели.

Предметом исследования біли методы восстановления баланса в выборках данных [3–4, 6] и методы извлечения обучающих выборок [7–12] для построения диагностических моделей по прецедентам.

Существует два базовых подхода к решению проблемы дисбаланса классов [3–4, 8]. Первый подход предполагает генерирование искусственных экземпляров миноритарного класса (класса меньшинства). При использовании данного подхода увеличивается объем исходных данных, поэтому он не может рассматриваться в контексте сокращения объема данных. Второй подход основан на сокращении числа экземпляров мажоритарного класса (класса большинства) и, соответственно, уменьшает объем исходной выборки данных в целом.

Цель данной работы – создание метода, позволяющего автоматически извлекать из исходных данных сбалансированные обучающие выборки меньшего объема, содержащие наиболее важные экземпляры для построения диагностических моделей на основе прецедентов.

1. 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 формирования сбалансированной выборки меньшего объема для построения модели зависимости $y(x)$ состоит в создании на основе исходной несбалансированной по классам выборки $X = \langle x, y \rangle$ такой подвыборки $X' = \langle x', y' \rangle$, чтобы выполнялись условия:

$$x' \in \{x^s\}, y' = \{y^s \mid x^s \in x'\}, S' \le S, S^{ma} \approx S^{mi}, \\ f(\langle x', y' \rangle, \langle x, y \rangle) \to opt,$$

где S^{ma}, S^{mi} – число экземпляров мажоритарного и миноритарного класса, соответственно; S' – число экземпляров в результирующей подвыборке; x' – набор признаков в сбалансированной выборке; y' – выходной признак (класс) в сбалансированной выборке.

Таким образом, необходимо из исходной несбалансированной выборки извлечь подвыборку меньшего объема, сбалансированную по классам, которая имеет точность не меньше исходного набора данных. Это означает, что при построении распознающей модели точность любого классификатора, использующего в качестве обу-

чающей выборки набор X', должна быть не меньше, чем точность, полученная при использовании исходной выборки X [5].

2. Литературный обзор

Наиболее точными методами извлечения обучающих выборок, которые могут гарантировать выполнение условия репрезентативности обучающей подвыборки, являются методы полного перебора всех возможных подвыборок исходного набора данных. Однако данные методы являются слишком затратными с точки зрения времени и ресурсов ЭВМ. Поэтому данные методы не применимы при обработке данных большого размера [13].

С практической точки зрения самым распространенным подходом являются стохастические методы извлечения обучающих подвыборок данных. В основе данных методов лежит идея случайного отбора экземпляров. К этим методам в первую очередь относятся такие методы сэмплинга [5, 9–11] , как простой случайный отбор с заменой (SRSWR) и без замены (SRSWOR), систематический отбор, стратифицированный отбор (Stratification sampling) и другие [5, 9]. Важным преимуществом семплинга для сокращения данных является отсутствие необходимости изучения всей исходной выборки для отбора конкретного экземпляра. Сложность методов сэмплинга прямо пропорциональна размеру извлекаемой подвыборки X' и не зависит от размера исходных данных. Недостатками являются неопределенность объема выборки и вероятность исключения важных информативных экземпляров и включения незначимых малоинформативных экземпляров.

Отдельной категорией методов извлечения обучающих подвыборок являются стратегии, основанные на эвристиках [7–8, 12–13]. Это методы интеллектуальной категоризации экземпляров, согласно степени значимости и в зависимости от выполняемой задачи. Они построены на таких предположениях, как компактность классов, решающие границы. Такие стратегии достаточно эффективны при правильном выборе метода для конкретной задачи. Однако в большинстве случаев они требуют загрузки в память всей исходной выборки и имеют высокую вычислительную сложность [8].

Другим подходом к извлечению обучающих выборок являются методы кластерного анализа [11], которые позволяют выделить все основные типы экземпляров в пространстве признаков. Недостатком данных методов является изначально неизвестное число типов экземпляров. Базовые методы кластерного анализа требуют загрузки в память исходной выборки и неоднократного пересчета расстояний между экземплярами в пространстве признаков, поэтому их применение для больших выборок также является достаточно ресурсоемким.

В условиях несбалансированности классов, когда число экземпляров одного класса (мажоритарный класс) значительно превышает число экземпляров другого класса (миноритарный класс), для восстановления баланса классов возможно применение всех перечисленных выше стратегий для сокращения экземпляров мажоритарного класса. Такой подход решил бы одновременно задачу восстановления баланса и сокращения выборки. Обычно выборка считается несбалансированной, когда объем миноритарного класса в выборке составляет менее 5% ее объема выбрки [14]. Такое соотношение классов в обучающей выборке негативно влияет на эффективность распознающих моделей. Это обусловлено тем, что классификаторы, которые не учитывают дисбаланс классов, могут просто игнорировать миноритарный класс, демонстрируя высокую точность. Например, если при объеме мажоритарного класса 95% классификатор неверно распознает только экземпляры миноритарного класса, его точность классификации составит 95%, и такая модель может представляться достаточно точной и эффективной. В то же время, в практических задачах технической диагностики интерес обычно представляет именно миноритарный класс, так как ему принадлежат редкие события, которые отображают любые отклонение от нормальной работы устройства.

Недостатки описанных выше стратегий характерны и для несбалансированных данных. Поэтому для формирования обучающих выборок из исходных выборок большого размера в условиях несбалансированности классов необходимо разработать методы, лишенные таких недостатков и способные в автоматическом режиме извлекать из исходных выборок сбалансированные выборки минимального объема, содержащие максимальное число экземпляров миноритарного класса и наиболее значимые экземпляры мажоритарного класса для построения диагностических моделей.

3. Материалы и методы

Для автоматического принятия решений в задачах технической диагностики необходимо в первую очередь обеспечить сохранение экземпляров, находящихся на границах классов в пространстве признаков. В общем случае такая задача решается методами кластерного анализа, который позволяет обнаружить экземпляры разных классов вблизи решающих границ. Однако методы кластерного анализа требуют определения расстояний между всеми экземплярами выборки и загрузки этих данных в память ЭВМ, многократных проходов по исходной выборке. Поэтому при работе с исходными данными большого размера обработка может потребовать значительных затрат памяти, вычислительных ресурсов и времени выполнения.

Для преодоления перечисленных недостатков предлагается методы кластерного анализа заменить на псевдокластеризацию, которая представляет собой объединение результатов частных кластеризаций выборки в одномерных проекциях на оси признаков [15].

В условиях несбалансированности классов для извлечения обучающей подвыборки предлагается отбирать все экземпляры миноритарного класса и, применяя метод псевдокластеризации, отбирать только те экземпляры мажоритарного класса, которые находятся на границе классов. Такой подход позволит получить относительно сбалансированную по классам обучающую подвыборку, в которой будут сохранены все важные экземпляры вблизи решающих границ. При этом число мажоритарных экземпляров в извлеченной подвыборке будет стремиться к числу экземпляров миноритарного класса. Таким образом, из несбалансированной исходной выборки, в которой число экземпляров миноритарного класса составляет 5%, будет извлечена выборка объемом порядка 10% от исходного объема.

Формально предлагаемый метод может быть представлен следующими этапами.

1. Инициализация. Задать исходную выборку X и результирующую выборку $X' = \varnothing$.

2. Псевдокластеризация. Для каждого i-го признака $(i = 1,2,...,N)$ выполнить.

2.1. Упорядочить экземпляры исходной выборки в порядке неубывания значений i-го признака.

2.2. Экземпляры с максимальным и минимальным значениями признака i добавить в выборку X', если они принадлежат мажоритарному классу: $y^s = majoritary$.

2.3. Рассмотреть попарно множество экземпляров, двигаясь по оси i-го признака, в порядке возрастания. Если оба экземпляра принадлежат разным классам, добавить экземпляр мажоритарного класса в выборку X'. В целях исключения дублирования для каждого экземпляра-претендента x^p на добавление в выборку X' найти такой экземпляр $x^* \in X'$, для которого будет выполнено условие: $d(x^*, x^p) = \arg\min\left\{d\left(x^s, x^p\right) \mid x^s \in X', s = 1,2,...,S'\right\}$, где расстояние $d(x^s, x^p) = \sqrt{\sum_{j=1}^{N}\left(x_j^s - x_j^p\right)^2}$, S' – число экземпляров в выборке X'. Если минимальное расстояние до ближайшего соседа больше нуля: $d(x^*, x^p) > 0$, экземпляр x^p добавляется в выборку X': $X' = X' \bigcup x^p$.

3. Добавить в выборку X' все экземпляры миноритарного класса:

$$X' = \left\{X' \bigcup x^s \mid x^s y^s \neq majoritary, s = 1,2,...,S\right\}.$$

4. Завершение. Возвратить сформированную обучающую подвыборку X' как результат.

В основе метода лежит предположение о компактности классов в пространстве признаков [12], поэтому если условия компактности не соблюдаются по нескольким признакам, это может привести к незначительному сокращению экземпляров мажоритарного и, соответственно, к неэффективной работе метода.

При выполнении условия компактности достоинством данного метода является существенное сокращение объема исходной выборки. При этом сохраняются все экземпляры миноритарного класса, который в большинстве случаев представляет основной интерес для задач диагностики (класс интереса), исключаются неинформативные и избыточные экземпляры мажоритарного класса и восстанавливается баланс классов. Недостатком метода может быть сохранение в извлеченной подвыборке незначимых или шумных экземпляров миноритарного класса.

Для оценки эффективности метода псевдокластеризации несбалансированных выборок использовались несколько наборов данных. Каждый набор делился на обучающую и тестовую выборки методом стратификации. Полученная обучающая выборка обрабатывалась несколькими известными методами сокращения несбалан-

сированных по классам данных, использующими различные стратегии. В качестве классификатора для синтеза распознающей модели применялся метод *k*-ближайших соседей (kNN), в основе которого также лежит гипотеза о компактности классов [12]. Решающие правила строились по принципу большинства голосов, поэтому для однозначности выбора в работе использовались методы с нечетным числом ближайших соседей ($k = 1, 3, 5$).

В качестве функционала качества применялась метрика гармоничного среднего (F-measure) [16]. Это связано с тем, что данная метрика позволяет объективно оценивать несбалансированные данные в исходной выборке, которая также использовалась для сравнения производительности.

В основе гармоничного среднего лежит понятие матрицы ошибок (табл. 1), которая представляет собой способ группировки экземпляров в зависимости от комбинации истинного ответа и ответа классификатора и позволяет получить множество различных метрик [4].

Таблица 1. Матрица ошибок

	$y = 1$	$y = 0$
$f(x) = 1$	True Positive (TP)	False Positive (FP)
$f(x) = 0$	False negative (FN)	True Negative (TN)

При работе с несбалансированными данными миноритарный класс обычно представляют как позитивный. В данном случае интерес представляют характеристики точности (precision) и полноты (recall) [4, 6]. Точность $P = TP/(TP + FP)$, где *TP* – верно классифицированные позитивные экземпляры (True Positive), *FP* – неверно классифицированные позитивные экземпляры (False Positive), показывает долю верно предсказанных позитивных экземпляров. Полнота $R = TP/(TP + FN)$, где *FN* – неверно классифицированные негативные экземпляры (False negative), показывает долю верно предсказанных позитивных экземпляров из всех экземпляров предсказанных как позитивные. Очевидно, что чем выше значения данных метрик, тем лучше классификатор. Однако на практике невозможно одновременно достигнуть максимальных значений точности и полноты, поэтому приходится выбирать, какая характеристика важнее для конкретной задачи, либо искать баланс между этими величинами. Объединить показатели точности и полноты позволяет характеристика гармонического среднего (F-measure) $F = 2PR/(P + R)$ [4, 16].

4. Эксперименты

Для проверки работоспособности предложенного метода и сравнения его с другими методами было разработано программное обеспечение, с помощью которого проводились эксперименты по автоматизации процесса извлечения обучающих выборок из исходных данных большого размера в условиях несбалансированности классов.

Исследования проводились на трех бинарных выборках, которые имели различное распределение классов:

1. Синтетическая выборка со случайным распределением экземпляров в пространстве признаков. Признаки генерировались случайным образом (Random sample).
2. Синтетическая выборка, в которой классы располагались в компактных областях пространства признаков. Признаки генерировались случайным образом в компактных областях пространства признаков (Compact sample).
3. Реальная выборка для задачи определения пульсаров (Pulsar) [17] .

В табл. 2 представлены такие основные характеристики исследуемых выборок данных, как число экземпляров выборки (*S*), число признаков (*N*) и доля миноританого класса (*M*).

Таблица 2.Характеристики исходных выборок данных

Выборка	*S*	*N*	*M*
Random sample	30000	2	5%
Compact sample	30000	2	5%
Pulsar	17898	8	10%

Каждая выборка делилась методом стратификации на обучающую и тестовую выборку в соотношении 75/25 соответственно. Далее каждая обучающая выборка обрабатывалась методом псевдокластеризации [15] (PCP – pseudo clustering procedure) и тремя методами восстановления баланса классов, использующими разные стратегии отбора экземпляров: Under – метод случайного удаления экземпляров мажоритарного класса из исходной выборки [9–10], CNN – метод, позволяющий извлекать подвыборку меньшего размера, которая имеет точность не меньше оригинальной выборки, на основе правила одного ближайшего соседа и предположения о значимости экземпляров на границах классов [13] и CBU – метод сокращения экземпляров с помощью кластеризации экземпляров мажоритарного класса и отбора по одному экземпляру из каждого кластера [6]. Также для

сравнения использовалась необработанная базовая выборка (Orig).

Затем на основе полученных обучающих выборок и kNN классификатора синтезировались распознающие модели, с помощью которых классифицировались экземпляры тестовых выборок и рассчитывались параметры гармоничного среднего (F-measure) для различных значений числа ближайших соседей *k*.

Сравнение методов производилось по точности синтезированных моделей и затратам ресурсов на формирование сбалансированных обучающих подвыборок.

5. Результаты

В табл. 3–5 представлены результаты сравнения затрат ресурсов, предложенных методов восстановления баланса классов обучающих выборок данных при решении задач извлечения обучающих подвыборок данных в условиях несбалансированности классов. Для сравнения использовались такие характеристики, как степень сокращения выборки (S/S'), время обработки выборки определенным методом (*t*), значение гармоничного среднего при классификации с использованием полученной обучающей подвыборки.

6. Обсуждение

Как видно из табл. 3–5, традиционные методы имеют свои преимущества и недостатки. Самым точным оказался метод конденсированного ближайшего соседа (CNN) [13], однако данный метод практически не сокращает объем выборки.

По степени сокращения выборки, очевидно, лучшими в парадигме несбалансированности классов являются методы, которые точно определяют соотношение классов и сокращают точное число экземпляров мажоритарного класса.

Такими являются методы кластеризации (CBU) [6] и случайного сокращения (Undersampling) экземпляров мажоритарного класса [3, 8]. Однако метод кластеризации [6, 11] требует неоднократного пересчета расстояний между экземплярами в пространстве признаков, поэтому при работе с большими выборками требует больших вычислительных ресурсов и ресурсов памяти. Метод случайного сокращения экземпляров мажоритарного класса [3, 8] очень эффективен с точки зрения ресурсов ЭВМ, однако в силу своей стохастичной природы демонстрирует низкую точность, так как существует вероятность удаления значимых экземпляров мажоритарного класса.

Таблица 3. Результаты обработки выборки Random sample

Метод	*S/S'*	*t*, с	F-measure		
			k=1	*k*=3	*k*=5
CBU	10,00	313,0	0,09	0,10	0,09
CNN	1,00	0,24	0,05	0,04	0,02
Orig	1,00	0,00	0,05	0,02	0,01
PCP	7,04	0,05	0,08	0,09	0,08
Under	10,00	0,01	0,01	0,09	0,09

Таблица 4. Результаты обработки выборки Compact sample

Метод	*S/S'*	*t*, с	F-measure		
			k=1	*k*=3	*k*=5
CBU	10,00	301,5	0,91	0,85	0,83
CNN	1,05	0,17	0,98	0,96	0,96
Orig	1,00	0,00	0,97	0,96	0,96
PCP	10,73	0,03	0,86	0,83	0,83
Under	10,00	0,01	0,83	0,80	0,80

Таблица 5. Результаты обработки выборки Pulsar

Метод	*S/S'*	*t*, с	F-measure		
			k=1	*k*=3	*k*=5
CBU	5,46	298,2	0,69	0,78	0,79
CNN	1,01	0,41	0,85	0,87	0,88
Orig	1,00	0	0,85	0,88	0,88
PCP	2,97	0,30	0,74	0,84	0,86
Under	5,46	0,13	0,64	0,76	0,78

По сравнению с традиционными методами формирования сбалансированных обучающих выборок [3–4, 6, 8], метод псевдокластеризации [15] позволил комплексно сократить объем и время вычислений, при этом эффективно выполнив задачу восстановления баланса классов и сокращения данных и продемонстрировав достаточно высокую точность полученной модели. Однако предложенный метод чувствителен к распределению классов по признакам. Таким образом, невыполнение условия компактности классов по какому-либо признаку будет уменьшать число сокращенных экземпляров мажоритарного класса. На практике редко встречаются выборки, в которых классы компактны по всем признакам. При увеличении числа признаков в реальной выборке степень сокращения мажоритарного класса и выборки в целом будет уменьшаться, сохраняя дисбаланс классов исходной выборки. Поэтому при решении практических задач необходимо предварительно отбирать из исходной выборки неинформативные признаки, которые могут повлиять на результат работы предложенного метода. Таким образом, предлагается использовать метод в ансамбле с методами отбора признаков [1], что позволит комплексно сократить размерность выборки, обеспечив сохранение

наиболее важной информации исходной выборки в полученной подвыборке.

Выводы

В целях решения задачи сокращения размерности данных в условиях несбалансированности классов при построении диагностических и распознающих моделей разработано математическое обеспечение, позволяющее формирование сбалансированные обучающие выборки.

Научная новизна результатов работы заключается в том, что впервые предложен метод формирования обучающих выборок, который обеспечивает восстановление баланса классов при сохранении важнейших для последующего анализа топологических свойств исходной выборки. При этом метод значительно сокращает объем исходной выборки и не требует ее загрузки в память ЭВМ, что существенно уменьшает требования к ресурсам ЭВМ.

Практическая значимость результатов работы состоит в том, что разработано программное обеспечение, реализующее предложенный метод формирования сокращенных сбалансированных выборок, а также проведены эксперименты по их исследованию при решении практических задач, результаты которых позволяют рекомендовать разработанный метод для использования на практике при решении задач технической диагностики и распознавания образов по признакам.

Перспективы дальнейших исследований состоят в том, чтобы изучить возможность использования предложенного метода в ансамбле с методами отбора информативных признаков для получения максимально сбалансированных обучающих выборок при сохранении минимальных требований к ресурсам ЭВМ.

Литература:

1. *Олійник А.О., Субботін С.О., Олійник С.О.* Інтелектуальний аналіз даних: навчальний посібник. Запоріжжя: ЗНТУ, 2012. 271 с. **2.** *Субботин С.А., Олейник Ан.А., Гофман Е.А., Зайцев С.А., Олейник Ал.А.* Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов: монография. Харьков: Компания СМИТ, 2012. 318 с. **3.** *Imbalanced Learning: Foundations, Algorithms, and Applications* / Ed. H. He., Y. Ma. Hoboken: Wiley-IEEE Press, 2013. 216 p. **4.** *Sun Y., Wong A.K.C., Kamel M.S.* Classification of imbalanced data: a review // International Journal of Pattern Recognition and Artificial Intelligence. 2009. Vol. 23, Issue 4. P. 687–719. **5.** *Encyclopedia of survey research methods* / Ed. P.J. Lavrakas. Thousand Oaks: Sage Publications, 2008. 968 p. **6.** *Lin W.C., Tsai C.F., Hu Y.H., Jhang J.S.* Clustering-based undersampling in class-imbalanced data // Information Sciences. 2017. Vol. 409-410. P. 17-26. **7.** *Leyva E., González A., Pérez R.* Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective // Pattern Recognition. 2015. Vol. 48, Issue 4. P. 1523–1537. **8.** *García S., Luengo J., Herrera F.* Data Preprocessing in Data Mining. Switzerland: Springer International Publishing AG, 2016. 320 p. **9.** *Thompson S.K.* Sampling. Hoboken: John Wiley & Sons, 2012. 472 p. **10.** *Кокрен У.* Методы выборочного исследования. Москва: Статистика, 1976. 440 с. **11.** *Chaudhuri A., Stenger H.* Survey sampling theory and methods. New York: Chapman & Hall, 2005. 416 p. **12.** *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. Новосибирск: ИИМ, 1999. 270 с. **13.** *Hart P.* The condensed nearest neighbor rule // IEEE Transactions on Information Theory. 1968. Vol. 14, Issue 3. P. 515–516. **14.** *He H., Garcia A.* Learning from Imbalanced Data // IEEE Transactions on Knowledge and Data Engineering. 2009. Vol. 21. P. 1263-1284. **15.** *Субботин С.А.* Методы формирования выборок для построения диагностических моделей по прецедентам // Вестник НТУ "ХПИ". Информатика и моделирование. 2011. № 17. С. 149-156. **16.** *Fawcett T.* An Introduction to ROC Analysis // Pattern Recognition Letters. 2006. Vol. 27, Issue 8. P. 861-874. **17.** *Lyon R.J.* HTRU2 [Electronic resource] // Access mode: https://figshare.com/articles/HTRU2/3080389/1.

Транслитерированный список литературы:

1. *Olijnyk A.O., Subbotin S.O., Olijnyk S.O.* Intelektual'nyj analiz danyh: navchal'nyj posibnyk. Zaporizhzhja: ZNTU, 2012. 271 s.

2. *Subbotin S.A., Olejnik An.A., Gofman E.A., Zajcev S.A., Olejnik Al.A.* Intellektual'nye informacionnye tehnologii proektirovanija avtomatizirovannyh sistem diagnostirovanija i raspoznavanija obrazov: monografija. Har'kov: Kompanija SMIT, 2012. 318 s.

3. *Imbalanced Learning: Foundations, Algorithms, and Applications* / Ed. H. He., Y. Ma. Hoboken: Wiley-IEEE Press, 2013. 216 p.

4. Sun Y., Wong A.K.C., Kamel M.S. Classification of imbalanced data: a review // International Journal of Pattern Recognition and Artificial Intelligence. 2009. Vol. 23, Issue 4. P. 687–719.

5. *Encyclopedia of survey research methods* / Ed. P.J. Lavrakas. Thousand Oaks: Sage Publications, 2008. 968 p.

6. *Lin W.C., Tsai C.F., Hu Y.H., Jhang J.S.* Clustering-based undersampling in class-imbalanced data // Information Sciences. 2017. Vol. 409-410. P. 17-26.

7. *Leyva E., González A., Pérez R.* Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective // Pattern Recognition. 2015. Vol. 48, Issue 4. P. 1523–1537.

8. *García S., Luengo J., Herrera F.* Data Preprocessing in Data Mining. Switzerland: Springer International Publishing AG, 2016. 320 p.
9. *Thompson S.K.* Sampling. Hoboken: John Wiley & Sons, 2012. 472 p.
10. Kokren U. Metody vyborochnogo issledovanija. Moskva: Statistika, 1976. 440 s.
11. *Chaudhuri A., Stenger H.* Survey sampling theory and methods. New York: Chapman & Hall, 2005. 416 p.
12. Zagorujko N. G. Prikladnye metody analiza dannyh i znanij. Novosibirsk: IIM, 1999. 270 s.
13. *Hart P.* The condensed nearest neighbor rule // IEEE Transactions on Information Theory. 1968. Vol. 14, Issue 3. P. 515–516.
14. *He H., Garcia A.* Learning from Imbalanced Data // IEEE Transactions on Knowledge and Data Engineering. 2009. Vol. 21. P. 1263-1284.
15. Subbotin S.A. Metody formirovanija vyborok dlja postroenija diagnosticheskih modelej po precedentam // Vestnik NTU "HPI". Informatika i modelirovanie. 2011. № 17. S. 149-156.
16. *Fawcett T.* An Introduction to ROC Analysis // Pattern Recognition Letters. 2006. Vol. 27, Issue 8. P. 861-874.
17. *Lyon R.J.* HTRU2 [Electronic resource] // Access mode: https://figshare.com/articles/HTRU2/3080389/1.

Поступила в редколлегию 11.06.2019
Рецензент: д-р техн. наук, проф. Кривуля Г.Ф.

Каврин Дмитрий Анатольевич, аспирант кафедры программных средств НУ «Запорожская политехника». Научные интересы: интеллектуальные системы технического диагностирования, оптимизация. Адрес: Украина, 69063, Запорожье, ул. Жуковского, 64, тел.: (091) 609-28-54, E-mail: kavrin@gmail.com.
Субботин Сергей Александрович, д-р техн. наук, проф., зав. кафедрой программных средств НУ «Запорожская политехника». Научные интересы: интеллектуальные системы технического диагностирования, нейронечеткие сети, оптимизация. Адрес: Украина, 69063, Запорожье, ул. Жуковского, 64, тел.: (061) 769-82-67, E-mail: subbotin.csit@gmail.com.
Kavrin Dmytro Anatoliiovych, PhD student, Department of Software Systems of National University “Zaporizhzhia Polytechnic”. Scientific interests: intelligent systems of technical diagnostics, optimization. Address: Zhukovsky str., 64, Zaporizhzhya, Ukraine, 69063, phone: (091) 609-28-54, E-mail: kavrin@gmail.com.
Subbotin Sergey Alexandrovich, Doctor of Technical Sciences, Professor, Head of Department of National University “Zaporizhzhia Polytechnic”. Scientific interests: intelligent systems of technical diagnostics, neuro-fuzzy networks, optimization. Address: Zhukovsky str., 64, Zaporizhzhya, Ukraine, 69063, phone: (061) 769-82-67, E-mail: subbotin.csit@gmail.com.