p-ISSN 1607-3274. Радіоелектроніка, інформатика, управління. 2018. № 1
e-ISSN 2313-688X. Radio Electronics, Computer Science, Control. 2018. № 1

НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

УДК 004.93

Каврин Д. А.[1], Субботин С. А.[2]

[1]Аспирант кафедры программных средств Запорожского национального технического университета, Запорожье, Украина

[2]Д-р техн. наук, проф., заведующий кафедрой программных средств Запорожского национального технического университета, Запорожье, Украина

МЕТОДЫ КОЛИЧЕСТВЕННОГО РЕШЕНИЯ ПРОБЛЕМЫ НЕСБАЛАНСИРОВАННОСТИ КЛАССОВ

Актуальность. Решена задача восстановления баланса классов в несбалансированных выборках для повышения эффективности диагностических и распознающих моделей.

Цель работы – модификация существующего метода восстановления баланса классов и сравнительный анализ показателей его производительности с некоторыми современными методами.

Метод. Предложен метод предварительной обработки несбалансированной выборки, который базируется на объединении стратегии undersampling и технологии кластерного анализа. Метод позволил восстановить баланс классов и сократить объем выборки, при этом были сохранены важные топологические свойства выборки, высокий показатель точности и приемлемое время работы.

Результаты. Разработано программное обеспечение, реализующее предложенный метод, которое было использовано при проведении вычислительных экспериментов по исследованию свойств метода и сравнительному анализу с другими методами восстановления баланса классов.

Выводы. Проведенные эксперименты подтвердили работоспособность, предложенного метода и реализующего его программного обеспечения. Метод позволил уменьшить мажоритарный класс до размеров миноритарного класса, соответственно уменьшив обучающую выборку (выборка считается несбалансированной, если размер миноритарного класса составляет менее 10% от размера исходной выборки), при этом продемонстрировал самые лучшие среди исследуемых методов показатели точности модели и сравнимую скорость формирования выборки. Это позволяет рекомендовать их для применения на практике при решении задач формирования обучающих выборок в условиях несбалансированности классов для диагностических и распознающих моделей.

Ключевые слова: выборка, классификатор, метрика качества, мажоритарный класс, миноритарный класс, экземпляр.

НОМЕНКЛАТУРА

C_{ma}^{i} – i-й центр масс кластера мажоритарного класса выборки;

C_{j}^{q} – j-й признак центра масс q-го кластера;

K – число классов в выборке;

k – число ближайших соседей;

m – отношение числа кластеров мажоритарного класса к числу экземпляров миноритарного класса;

N – число входных признаков характеризующих экземпляры выборки;

Q_{ma} – число кластеров мажоритарного класса в исходной выборке;

q – номер текущего кластера;

S – число экземпляров в исходной выборке;

S' – число экземпляров в сбалансированной выборке;

S_{ma} – число экземпляров мажоритарного класса в исходной выборке;

S_{ma}^{q} – число экземпляров q-го кластера мажоритарного класса;

S_{mi} – число экземпляров миноритарного класса в исходной выборке;

s – номер текущего экземпляра;

X – исходная выборка;

X' – сбалансированная выборка;

X'_{ma} – множество прецедентов о зависимости мажоритарного класса в сбалансированной выборке;

X_{mi} – множество прецедентов о зависимости миноритарного класса;

DOI 10.15588/1607-3274-2018-1-10

x – набор признаков в исходной выборке;

x' – набор признаков в сбалансированной выборке;

x^s – s-й экземпляр выборки;

y – выходной признак (класс) в исходной выборке;

y' – выходной признак (класс) в сбалансированной выборке;

y^s – выходной признак s-го экземпляра выборки;

CBU – cluster based undersampling.

ВВЕДЕНИЕ

Для построения диагностических и распознающих моделей по экспериментально полученным наблюдениям (прецедентам) необходимо из набора имеющихся наблюдений большого объема выделить обучающую выборку, обладающую меньшим объемом, но отражающую основные свойства исходной совокупности наблюдений.

Объектом исследования являлись методы автоматического формирования выборок для построения диагностических и распознающих моделей по прецедентам.

Обучающая выборка является одним из важнейших компонентов диагностических и распознающих моделей. От объема выборки и представительности ее данных будет зависеть производительность построенной модели, ее точность и скорость. Большинство стандартных алгоритмов классификации предполагают равномерное распределение данных в обучающих выборках, однако, в реальной жизни это достаточно редкое явление [1]. Поэтому актуальным является применение различных подходов для восстановления равномерного распределения данных в обучающих выборках. Одним из таких подходов, являются методы восстановления баланса классов в несбалансированных выборках.

Предметом исследования являлись методы формирования сбалансированных выборок.

Достаточно распространенным явлением является ситуация, когда в выборке экземпляров одного класса значительно больше (мажоритарный класс) чем экземпляров другого класса (миноритарный класс) [1]. В таких условиях большинство методов машинного обучения приводят к получению моделей, которые неправильно определяют редкие экземпляры миноритарного класса из-за подавления экземплярами мажоритарного класса экземпляров миноритарного класса при обучении модели. Для примера рассмотрим бинарную выборку, в которой 99% миноритарных экземпляров и 1% мажоритарных. Если после построения модели на основе такой выборки модель отнесет все экземпляры к мажоритарному классу, то ошибка классификации составит всего 1%, т. е. при очень высокой точности классификатор не сможет правильно определить экземпляры миноритарного класса. Однако именно миноритарный класс может иметь первостепенную важность в таких прикладных задачах, как медицинская диагностика, кредитный скоринг, выявление мошенничества с кредитными картами, защита компьютерных сетей [2, 3]. Поэтому актуальной является проблема формирования обучающих выборок для построения моделей при несбалансированных классах в исходной выборке.

Известно множество методов решения проблемы несбалансированности классов [1], которые можно разделить на два основных вида: уровня данных и уровня методов [3].

Уровень методов предполагает создание новых или модификацию существующих классификаторов при построении модели для каждого нового набора данных или новой задачи, что может потребовать дополнительных ресурсов и затрат.

В отличие от уровня методов, уровень данных не требует модификации методов классификации, достаточно простой и может использоваться с любыми типами классификаторов. Методы уровня данных основаны на предварительной обработке данных с помощью сэмплинга [4], стратегии которого делятся на два типа: undersampling (удаляют экземпляры мажоритарного класса) и oversampling (добавляют (синтезируют) экземпляры миноритарного класса). При этом методы на основе сэмплинга достаточно эффективно решают проблему несбалансированности классов и оптимизируют производительность используемых классификаторов [4].

На практике стратегия undersampling работает более эффективно, чем стратегия oversampling. Это связано с тем, что стратегия oversampling увеличивает размер выборки, что может повысить вероятность переобучения [5] и время работы классификатора. В свою очередь, при применении стратегии undersampling существует вероятность потери важной информации.

Для сохранения репрезентативности выборки в ряде работ предлагается использовать стратегию удаления экземпляров мажоритарного класса с применением кластерного анализа CBU [1, 5, 6]. При этом экземпляры мажоритарного класса сначала разбиваются на кластеры, а затем из каждого кластера по определенным правилам выбирается необходимое количество экземпляров. Такой подход уменьшает риск удаления значимых экземпляров, что позволяет увеличить производительность классификаторов.

Цель работы – усовершенствование метода восстановления баланса классов CBU и сравнительный анализ показателей его производительности с другими методами.

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 $\langle x^s, y^s \rangle, x^s = \{x_j^s\}, y^s \in \{1,2,...,K\}, K > 1.$

Тогда задача формирования сбалансированной выборки 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 выполнялось одно из следующих условий: для добавления экземпляров миноритарного класса (oversampling):

$$x' \in \{x^s\}, y' = \{y^s \mid x^s \in x'\}, S' \geq S, f(\langle x', y' \rangle, \langle x, y \rangle) \to opt;$$

для удаления экземпляров мажоритарного класса (undersampling):

$$x' \in \left\{x^s\right\}, y' = \left\{y^s \mid x^s \in x'\right\}, S' \le S, f\left(\langle x', y'\rangle, \langle x, y\rangle\right) \to opt.$$

Т.е. необходимо в несбалансированной выборке изменить распределение классов таким образом, чтобы получить сбалансированный набор данных.

2 ОБЗОР ЛИТЕРАТУРЫ

Для того, чтобы в несбалансированной выборке изменить распределение классов так, чтобы получить сбалансированный набор данных, применяют различные стратегии сэмплинга.

Стратегии сэмплинга применяются на этапе предварительной обработки данных. Они достаточно эффективны и просты в использовании, не требуют модификации методов классификации и могут использоваться с любыми классификаторами. Поэтому данные технологии широко используются для решения проблем несбалансированности классов.

Рассмотрим наиболее широко используемые стратегии сэмплинга разных типов с возможностью контроля количества удаляемых (добавляемых) экземпляров.

Случайное удаление экземпляров мажоритарного класса (random undersampling) – наиболее простая стратегия, в которой случайным образом удаляются экземпляры мажоритарного класса для достижения необходимого соотношения классов. Уровень соотношения классов подбирается эмпирическим путем. Достоинствами стратегии являются высокая скорость работы, уменьшение размера выборки и простота реализации, а недостатками – высокая вероятность потери значимых данных.

Удаление экземпляров мажоритарного класса с применением кластерного анализа (CBU – cluster based undersampling) – стратегия удаления экземпляров мажоритарного класса с применением методов кластерного анализа. На первом этапе множество экземпляров мажоритарного класса разбивается на число кластеров, равное числу экземпляров миноритарного класса. На следующем этапе выбирается по одному экземпляру из каждого кластера, и удаляются все остальные экземпляры мажоритарного класса [5]. Достоинствами стратегии являются уменьшение обучающей выборки до размера $2S_{mi}$ (т.е. если доля миноритарного класса составляет 1%, размер обработанной выборки будет 2% от размера исходящей выборки), сохранение важных топологических свойств выборки, а недостатком – низкая скорость работы.

Дублирование экземпляров миноритарного класса (oversampling) – это стратегия, в которой для достижения необходимого соотношения классов, дублируются экземпляры миноритарного класса. Достоинствами стратегии являются высокая скорость работы и простота реализации, а недостатками – возможность переобучения модели и увеличение размера выборки.

Стратегия искусственного увеличения экземпляров миноритарного класса (SMOTE – Synthetic Minority Over-sampling Technique) [7] – одна из популярных стратегий сэмплинга, которая базируется на технологии oversampling. Данная стратегия предполагает синтез искусственных экземпляров путем создания одного или нескольких ближайших соседей для экземпляров миноритарного класса, в зависимости от необходимого соотношения классов. Достоинствами стратегии являются высокая скорость работы и простота, а недостатками – возможное переобучения построенной модели, увеличение размера формируемой выборки.

Адаптивная технология искусственного увеличения экземпляров миноритарного класса (ASMO – Adaptive Synthetic Minority Oversampling) [8] – стратегия, являющаяся модификацией SMOTE, в которой поиск ближайших соседей производится для экземпляров мажоритарного класса, что позволяет лучше разделить классы. Достоинствами стратегии являются высокая скорость работы и простота реализации, а недостатками – возможность переобучения модели и увеличение размера формируемой выборки.

Существенное влияние на качество построенных моделей кроме стратегий формирования выборок также имеют метрики (способы оценивания) качества моделей.

Для оценки качества моделей с дискретным выходом традиционно используется ошибка [9]:

$$E = \sum_{s=1}^{S} \left\{1 \mid y^s \ne f(x^s)\right\} \to \min.$$

Это достаточно простая и эффективная метрика, которая широко используется на практике. Однако в условиях несбалансированности классов функция ошибки не является подходящей метрикой, поскольку миноритарный класс очень слабо влияет на ошибку по сравнению с мажоритарным классом. Например, в ситуации, когда миноритарный класс представлен только 1% выборки, простая стратегия может предсказывать мажоритарный класс для всех экземпляров. При этом ошибка будет составлять всего 1%. Однако это измерения не имеет смысла для приложений, в который задача обучения состоит именно в определении миноритарного класса. Поэтому для несбалансированных выборок целесообразно использование метрик, в основе которых лежит понятие матрицы ошибок [10]. Это способ группировки экземпляров в зависимости от комбинации истинного ответа и ответа алгоритма обучения.

В случае с бинарной выборкой, экземпляры можно разделить на четыре категории (табл. 1). Экземпляры класса, представляющего больший интерес, называют позитивными, экземпляры другого класса негативными.

С помощью матрицы ошибок (табл. 1) можно получить различные метрики [3]. Если актуальной задачей является изучение экземпляров миноритарного класса, его представляют как позитивный. В этом случае, интерес будут представлять такие характеристики, как точность и полнота. Точность (precision) показывает, сколь-

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)

ко из предсказанных позитивных объектов, оказались действительно позитивными:

$$P = \frac{TP}{TP + FP}.$$

Полнота (recall) показывает, сколько объектов из общего числа реальных позитивных объектов, было верно предсказано как позитивный класс

$$R = \frac{TP}{TP + FN}.$$

Чем выше значения этих характеристик, тем качественней 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. Дать оценку одновременно по точности и полноте позволяет характеристика гармоническое среднее (*F*-measure) [11]:

$$F = \frac{2PR}{P + R}.$$

В настоящей работе для оценки изучаемых методов восстановления баланса классов предлагается использовать *F*-measure, поскольку данная метрика позволяет явно выделить для анализа интересующий позитивный класс, в нашем случае миноритарный.

Для сравнения методов формирования выборок необходимо задать конкретный тип классификаторов. Метод *k*-ближайших соседей (kNN – *k* Nearest Neighbor) [12] является широко используемым, но при этом достаточно простым и эффективным методом. В его основе лежит гипотеза о компактности классов [13], которая предполагает, что тестируемый экземпляр будет относиться к тому же классу, что и экземпляры из его ближайшего окружения.

3 МАТЕРИАЛЫ И МЕТОДЫ

Метод удаления экземпляров мажоритарного класса с применением кластерного анализа CBU продемонстрировал высокую производительность в условиях несбалансированности классов, уменьшая размер обучающей выборки, что позволяет в дальнейшем снизить время работы построенной модели. Однако время формирования обучающей выборки данным методом оказывается значительно больше, чем для других рассмотренных методов. Поэтому, для уменьшения времени формирования обучающей выборки предлагается усовершенствовать данный метод следующим образом.

1. Задать коэффициент желаемого соотношения числа кластеров мажоритарного класса к числу экземпляров миноритарного класса m. Рекомендуется задавать $1 < m \le 10$, так как при значениях $m > 10$ производительность классификаторов достаточно быстро падает, и рассчитывать на достаточную точность классификации не приходится.

2. Разбить множество экземпляров мажоритарного класса на $Q_{ma} = \frac{S_{mi}}{m}$ кластеров. Таким образом, разделив множество экземпляров мажоритарного класса на компактные области в пространстве признаков. Для этого можно использовать простейший метод кластеризации *k*-средних (*k*-means) [14].

3. Определить координаты центров масс полученных кластеров:

$$C_j^q = \frac{1}{S_{ma}^q} \sum_{s=1}^{S_{ma}} \left\{ x_j^s \mid y^s = q \right\}, j = 1,2,...,N, q = 1,2,...,Q_{ma}.$$

4. Для восстановления баланса классов продублировать центр каждого кластера m раз

$$X'_{ma} = \bigcup_{i=1}^{m} C_{ma}^i,$$

внося незначительные коррективы в координаты центров кластеров по формуле:

$$C_j^q = C_j^q (1 + 0{,}1\text{rand} - 0{,}1\text{rand}),$$

где rand – функция, возвращающая случайное число в диапазоне $\left[0; \left(\max_{s=1,...,S} \{x_j^s\} - \min_{s=1,...,S} \{x_j^s\} \right) / S^2 \right]$.

5. Удалить из выборки все экземпляры мажоритарного класса, и объединить множество центров кластеров мажоритарного класса с множеством экземпляров миноритарного класса $X' = X'_{ma} \bigcup X_{mi}$. Полученный набор данных с внесенными коррективами рассматривать как результирующую выборку для построения моделей.

Достоинством предложенной модификации метода CBU является то, что она значительно снижает время обработки выборки, при этом сохраняя основные особенности данных. Недостатком предложенной модификации метода CBU является то, что при увеличении коэффициента m, увеличивается вероятность потери важных экземпляров мажоритарного класса.

4 ЭКСПЕРИМЕНТЫ

Для исследования свойств рассмотренных методов они были программно реализованы как часть «Автоматизированной системы отбора оптимального метода восстановления баланса классов при формировании обучающей выборки» [15], в которую был интегрирован дополнительный модуль CBU. Модуль CBU представляет собой функцию, которая разбивает множество экземпляров мажоритарного класса исходящей выборки на заданное число кластеров, определяет центры масс полученных кластеров, формирует сбалансированную выборку из множества экземпляров миноритарного класса и множества центров кластеров мажоритарного класса, при необходимости, продублированных случайным образом для балансировки их числа с числом экземпляров миноритарным класса.

Модифицированное программное обеспечение использовалось при проведении вычислительных экспериментов, которые включали два этапа: на первом этапе проводился сравнительный анализ существующих методов восстановления баланса классов, на втором

этапе проводились исследования модифицированного метода CBU и сравнение его производительности с базовым методом CBU.

Исследования проводились на синтетических бинарных выборках, что позволило регулировать соотношения классов. На всех этапах синтезировалась бинарная выборка из 10 000 экземпляров, имеющих по два признака, принимавших значения из множества {0, 1}. Далее, для тестирования стратегий, выборка была разделена методом стратификации [16] на обучающую выборку (90% от исходной выборки) и тестовую (10% от исходной выборки).

Решающие правила строились по принципу большинства голосов. Поэтому для однозначности выбора в работе использовались методы с нечетным числом ближайших соседей ($k = 9, 25, 49$).

Для сравнительного анализа методов восстановления баланса классов для каждого метода рассчитывалось значение метрики *F*-measure для различных параметров выборки и классификатора: доля миноритарного класса в выборке (25%, 10%, 4%, 1%), число ближайших соседей классификатора kNN (3, 49). Затем строилась зависимость метрики *F*-measure от доли миноритарного класса, и сравнивались ее значения для различных подходов.

Далее были предприняты шаги для уменьшения времени работы метода CBU, модификация которого состояла в сокращении числа кластеров, предполагая, что это приведет к уменьшению времени формирования выборки при незначительном уменьшении точности полученной модели.

Значение *F*-measure рассчитывалось для различных параметров выборки и классификатора kNN: доля миноритарного класса в выборке (10%, 1%), число ближайших соседей классификатора kNN (3, 49). Строилась зависимость *F*-measure от числа кластеров в мажоритарном классе.

5 РЕЗУЛЬТАТЫ

Результаты исследований предложенной модификации метода CBU в сравнении с известными методами представлены на рис. 1 и в табл. 2.

Из рис. 1а и рис. 1б видно, что при различных настройках kNN классификатора, метод CBU сформировал самую представительную обучающую выборку.

Из табл. 2 следует, что время работы CBU значительно превысило время работы остальных методов. Таким образом, при работе с большими выборками, время работы метода может нивелировать преимущества метода, либо стать причиной отказа от подобного подхода.

На рис. 2 изображены графики зависимостей *F*-measure от числа кластеров мажоритарного класса. В табл. 3 представлена зависимость времени формирования выборки от *m*.

Как видно из рис. 2а и рис. 2б, при уменьшении числа кластеров мажоритарного класса качество построенной модели довольно быстро ухудшается. Одновременно уменьшается и время формирования обучающей выборки (табл. 3). Такая ситуация дает исследователю возможность находить компромисс между производительностью и скоростью построения модели, исходя из предъявленных требований.

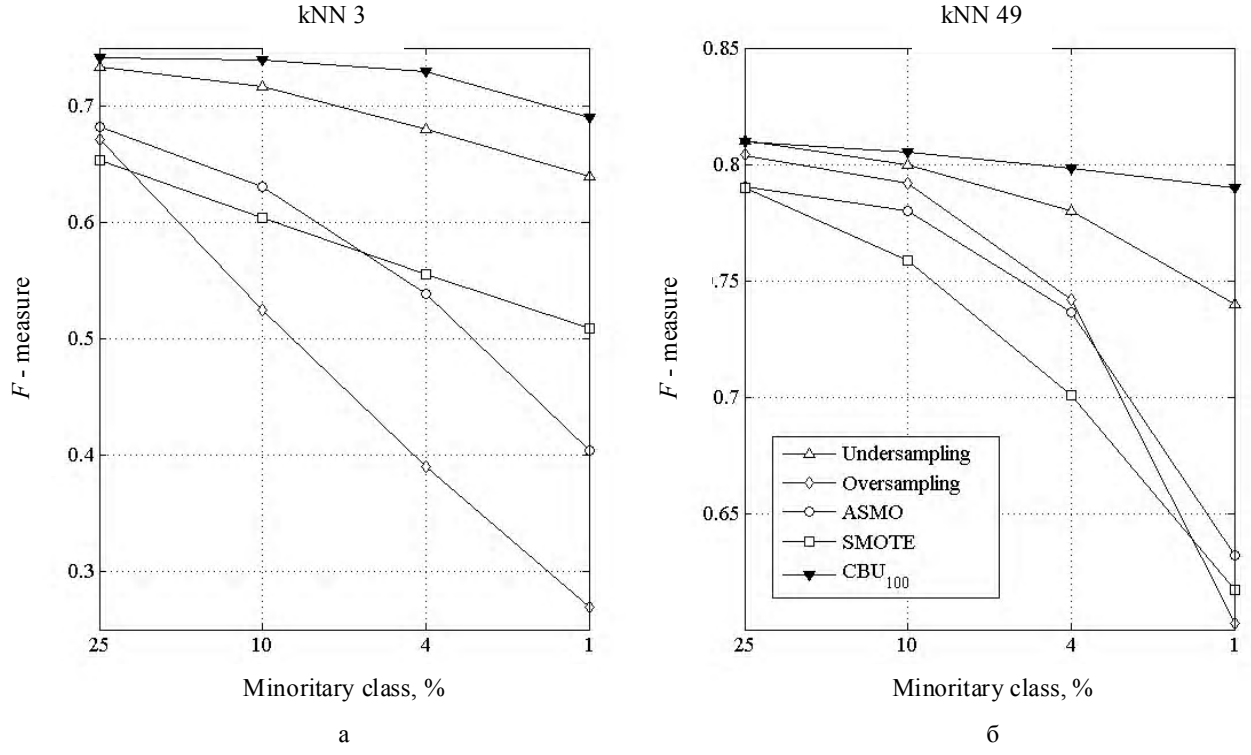

Рисунок 1 – Графики зависимостей F-measure от процента миноритарного класса:
а – kNN=3, б – kNN=49

Таблица 2 – Зависимость времени формирования выборки (с) от доли миноритарного класса

Методы сэмплинга	Доля миноритарного класса в выборке, %			
	25	10	4	1
Undersampling	0,00801	0,00212	0,00206	0,00204
Oversampling	0,00103	0,00087	0,00082	0,00081
ASMO	0,13606	0,10825	0,08979	0,09459
SMOTE	0,06319	0,02020	0,01242	0,01150
CBU_100	50,89648	13,17416	2,34512	2,33082

Таблица 3 – Зависимость времени формирования выборки (с) от m

Доля миноритарного класса, %	Соотношение m			
	1:1	1:10	1:100	1:1000
1	1,5964	0,2231	0,0063	0,0063
10	50,2719	3,1516	0,1852	0,0063

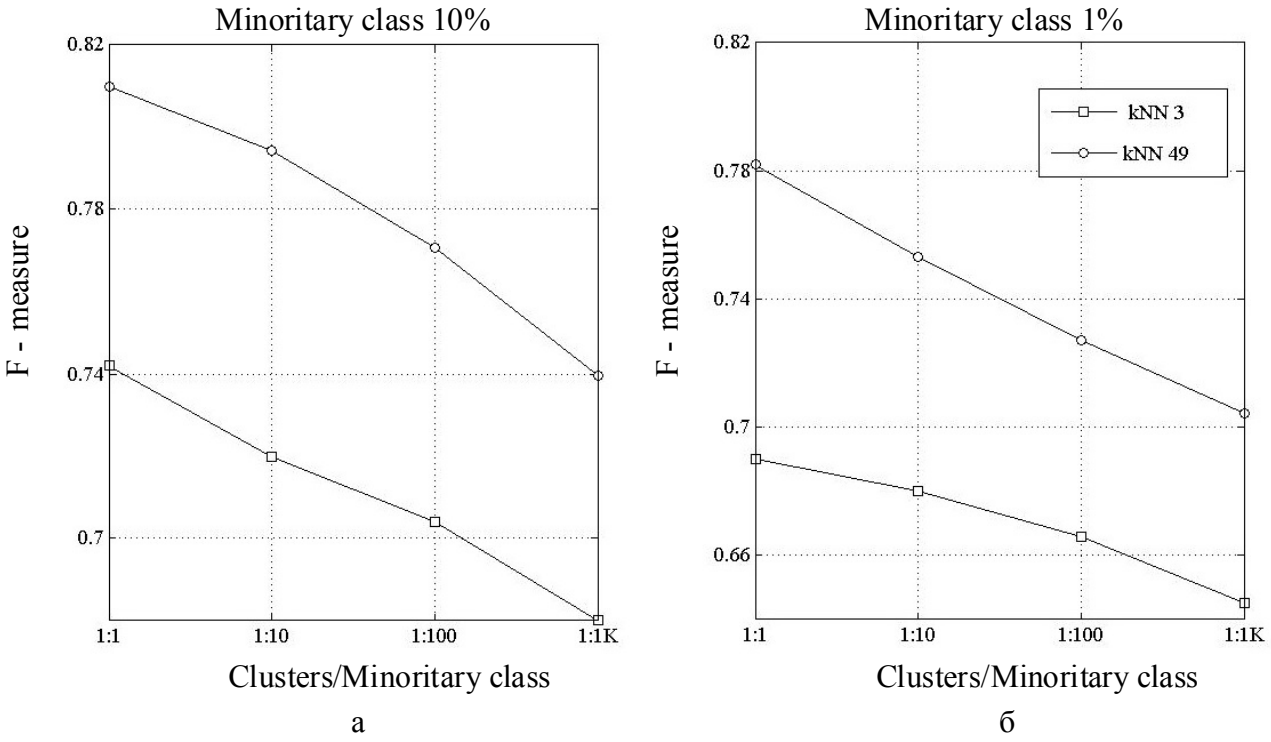

Рисунок 2 – Графики зависимостей F-measure от количества кластеров мажоритарного класса: а – миноритарный класс 10%, б – миноритарный класс 1%

6 ОБСУЖДЕНИЕ

Использование предложенной модификации метода CBU значительно уменьшило время формирования обучающей выборки при стабильной поддержки приемлемого значения показателя качества классификации. Однако, при уменьшении числа кластеров, увеличивается количество удаленных значимых экземпляров, что приводит к снижению качества обучающей выборки. Фактически миноритарный и мажоритарный классы меняются местами, когда для восстановления баланса необходимо уже синтезировать экземпляры мажоритарного класса, либо удалять экземпляры миноритарного класса. Естественно, такое положение вещей не может длиться бесконечно, в конечном итоге выборка может оказаться пустой. Поэтому в каждом конкретном случае важно найти предельные значения параметров метода, при которых модель будет демонстрировать требуемые показатели точности, скорости, объема и т. д.

Исходя из проделанных исследований, данный метод можно рекомендовать при количестве кластеров в соотношении не более 1:100 (1 кластер мажоритарного класса на 100 экземпляров миноритарного класса), т.к. при дальнейшем уменьшении числа кластеров производительность классификации резко падает.

В ходе исследования было замечено, что метод работает лучше других, если классы перемешаны (т.е. не выполняется условие компактности), по всей видимости, это связано с удалением экземпляров мажоритарного класса, которые могут рассматриваться как шум (экземпляры сильно удаленные от своего класса).

Таким образом, предложенный метод позволяет найти компромисс между временем создания выборки и ее репрезентативностью, а соответственно и точностью построенной распознающей модели.

ВЫВОДЫ

С целью повышения скорости построения и точности работы, диагностических и распознающих моделей по прецедентам решена задача восстановления баланса классов в несбалансированных обучающих выборках.

Научная новизна полученных результатов состоит в том, что модифицирован метод CBU, который уменьшает число экземпляров мажоритарного класса путем

разбиения их на кластеры и затем удаляя определенное число экземпляров из каждого кластера. Это позволяет восстановить баланс классов в несбалансированных выборках, используя такие преимущества базового метода undersampling, как высокая скорость работы и существенное уменьшение размера обучающей выборки. При этом модифицированный метод обеспечивает ускорение процесса построения моделей и повышает их адекватность, обеспечивая топологическую репрезентативность выборки данных.

Практическая значимость полученных результатов состоит в том, что разработано программное обеспечение, реализующее предложенный метод, а также другие методы формирования выборок (ASMO, CBU, Condensed Nearest Neighbor Rule, Neighborhood cleaning rule, One-side sampling, Oversampling, Random undersampling, SMOTE, Tomek Links), которое экспериментально исследовано при решении задач сэмплинга в условиях несбалансированности классов. 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, а также прогнозирования в различных областях.

Перспективы дальнейших исследований заключаются в том, чтобы исследовать и улучшить методы, учитывающих не только статистическую репрезентативность обучающей выборки, но и топологическую, что приведет к формированию малочисленных и при этом более качественных обучающих выборок. Также предполагается изучить свойства предложенного метода на более широком классе практических задач и разработать параллельную реализацию предложенного метода.

БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы «Методы и средства вычислительного интеллекта и параллельного компьютинга для обработки больших объемов данных в системах диагностирования» (номер гос. регистрации 0116U007419) кафедры программных средств Запорожского национального технического университета при частичной поддержке международного образовательного проекта “Internet of Things: Emerging Curriculum for Industry and Human Applications” (ALIOT, ref. number 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP) программы «Эразмус+» Европейского Союза.

СПИСОК ЛИТЕРАТУРЫ

1. He H. Learning from Imbalanced Data / H. He, E. A. Garcia // IEEE Transactions on Knowledge and Data Engineering. – 2009. – Vol. 21. – P. 1263–1284. DOI: 10.1109/TKDE.2008.239
2. Паклин Н. Б. Построение классификаторов на несбалансированных выборках на примере кредитного скоринга / Н. Б. Паклин, С. В. Уланов, С. В. Царьков // Искусственный интеллект. – 2010. – № 3. – С. 528–534.
3. Sun Y. Classification of imbalanced data: a review / Y. Sun, A. K. C. Wong, M. S. Kamel // International Journal of Pattern Recognition and Artificial Intelligence. – 2009. – Vol. 23, Issue 4. – P. 687–719. DOI: 10.1142/S0218001409007326
4. Batista G. E. A. P. A. A study of the behavior of several methods for balancing machine learning training data / G. E. A .P. A. Batista, R. C. Prati, M. C. Monard // SIGKDD Explorations. – 2004. – Vol. 6, Issue 1. – P. 20–29. DOI: 10.1145/1007730.1007735
5. Clustering-based undersampling in class-imbalanced data / [W. C. Lin, C. F. Tsai, Y. H. Hu, J. S. Jhang] // Information Sciences. – 2017. – Vol. 409–410. – P. 17–26. DOI: 10.1016/j.ins.2017.05.008
6. Yen S. J. Cluster-based under-sampling approaches for imbalanced data distributions / S. J. Yen, Y. S. Lee // Expert Systems with Applications. – 2009. – Vol. 36, Issue 3. – P. 5718–5727. DOI: 10.1016/j.eswa.2008.06.108
7. Chawla N.V. SMOTE: Synthetic minority over-sampling technique / N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer // Journal of Artificial Intelligence Research. – 2002. –Vol. 16. – P. 321–357. DOI: 10.1613/jair.953
8. Wang B.X. Imbalanced Data Set Learning with Synthetic Samples [Electronic resource] / B. X.Wang, N. Japkowicz. – Access mode: http://www.iro.umontreal.ca/~lisa/workshop2004/ program.html
9. Субботін С. О. Інтелектуальні системи : навч. посіб. / С. О. Субботін, А. О. Олійник; під заг. ред. проф. С. О. Субботіна. – Запоріжжя : ЗНТУ, 2014. – 218 с.
10. Elkan C. The foundations of cost-sensitive learning / C. Elkan // 17th international joint conference on Artificial intelligence, Seattle, 4-10 August 2001 : Proceedings. – San Francisco : Morgan Kaufmann Publishers Inc., 2001. – Vol. 2. – P. 973–978.
11. Fawcett T. An Introduction to ROC Analysis / T. Fawcett // Pattern Recognition Letters. – 2006. – Vol. 27, Issue 8. – P. 861–874. DOI: 10.1016/j.patrec.2005.10.010
12. Cover T. Nearest neighbor pattern classification / T. Cover, P. Hart // IEEE Transactions on Information Theory. – 1967. – Vol. 13, Issue 1. – P. 21–27. DOI: 10.1109/TIT.1967.1053964
13. Загоруйко Н. Г. Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск : ИИМ, 1999. – 270 с.
14. Lloyd S. P. Least Squares Quantization in PCM / S. P. Lloyd // IEEE Transactions on Information Theory. – 1982. – Vol. 28. – P. 129–137. DOI: 10.1109/TIT.1982.1056489
15. Суботін С. О. Автоматизована система відбору оптимального методу відновлення балансу класів при формуванні навчальної вибірки / С. О. Суботін, Д. А. Каврін // Інформатика, управління та штучний інтелект. Матеріали четвертої міжнародної науковотехнічної конференції студентів, магістрів та аспірантів. – Харків: НТУ «ХПІ», 2017. – С. 94.
16. Кокрен У. Методы выборочного исследования / У. Кокрен. – М. : Статистика, 1976. – 440 с.

Статья поступила в редакцию 22.12.2017.
После доработки 25.01.2018.

Каврін Д. А.[1], Субботін С. О.[2]

[1]Аспірант кафедри програмних засобів Запорізького національного технічного університету, Запоріжжя, Україна

[2]Д-р техн. наук, проф., завідувач кафедри програмних засобів Запорізького національного технічного університету, Запоріжжя, Україна

МЕТОДИ КІЛЬКІСНОГО ВИРІШЕННЯ ПРОБЛЕМИ НЕЗБАЛАНСОВАНОСТІ КЛАСІВ

Актуальність. Вирішено завдання відновлення балансу класів в незбалансованих вибірках для підвищення ефективності діагностичних та розпізнавальних моделей.

Мета роботи – модифікація існуючого методу відновлення балансу класів та порівняльний аналіз показників його продуктивності з деякими сучасними методами.

Метод. Запропоновано метод попередньої обробки незбалансованої вибірки, який базується на об'єднанні стратегії undersampling та технології кластер-аналізу. Метод дозволив відновити баланс класів та зменшити об'єм вибірки, при цьому було збережено важливі топологічні властивості, високі показники точності та прийнятний час роботи.

Результати. Розроблено програмне забезпечення, що реалізує запропонований метод, яке було використано при виконанні обчислювальних експериментів властивостей методу та порівняльному аналізу з іншими методами відновлення балансу класів.

Висновки. Проведені експерименти підтвердили працездатність запропонованого методу та програмного забезпечення, що його реалізує. Метод дозволив зменшити мажоритарний клас до розмірів міноритарного класу, зменшивши таким чином навчальну вибірку (вибірка вважається незбалансованою, коли розмір міноритарного класу становить менш ніж 10% від розміру вихідної вибірки), при цьому продемонстрував найкращі серед досліджених методів показники точності моделі та прийнятний час формування вибірки. Результати експериментів дозволяють рекомендувати їх для використання на практиці при вирішенні задач формування навчальних вибірок в умовах незбалансованості класів для діагностичних та розпізнавальних моделей.

Ключові слова: вибірка, екземпляр, метрика якості, класифікатор, кластер, мажоритарний клас, міноритарний клас

Kavrin D. A.[1], Subbotin S. A.[2]

[1]Post-graduate student of the Department of Software Tools of Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine

[2]Dr. Sc., Professor, Head of the Department of Software Tools of Zaporizhzhya National Technical University, Zaporizhzhya, Ukraine

THE METHODS FOR QUANTITATIVE SOLVING THE CLASS IMBALANCE PROBLEM

Context. The problem of recovery the classes' balance in imbalanced samples is solved to increase the efficiency of diagnostic and recognition models.

Objective. The purpose of the work is to modify the existing method of recovery classes' balance and to conduct comparative analysis of performance indicators with some modern methods.

Method. The proposed data preprocessing method is based on combining the undersampling and cluster-analysis technologies. The method has allowed restoring the balance and reducing the sample while maintaining important topological properties of the sample, high accuracy and acceptable operating time.

Results. The software that implements in proposed method has been developed and used in the computational experiments on the study of method's properties and comparative analysis with other methods of restoring classes' balance.

Conclusions. The experiments confirmed the efficiency of the proposed method and its implemented software. The method has allowed reducing the majority class to the size of the minority class, thus reducing the training sample (the sample is considered imbalanced if the size of the minority class is less than 10% of the original sample size), while demonstrating the best indicators of model accuracy and comparable sampling speed. It can be recommended for the practical application in solving problems of imbalance data for diagnostic and recognition models.

Keywords: sample, example, quality metric, cluster, classificatory, majority class, minority class.

REFERENCES

1. He H., Garcia E. A. Learning from Imbalanced Data, *IEEE Transactions on Knowledge and Data Engineering,* 2009, Vol. 21, pp. 1263–1284. DOI: 10.1109/TKDE.2008.239
2. Paklin N. B., Ulanov S. V., Car'kov S. V. Postroenie klassifikatorov na nesbalansirovannykh vyborkakh na primere kreditnogo skoringa, *Iskusstvennyjj intellekt*, 2010, No. 3, pp. 528–534.
3. 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, pp. 687–719. DOI: 10.1142/S0218001409007326
4. Batista G. E. A. P. A., Prati R. C., Monard M. C. A study of the behavior of several methods for balancing machine learning training data, *SIGKDD Explorations*, 2004, Vol. 6, Issue 1, pp. 20–29. DOI: 10.1145/1007730.1007735
5. 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, pp. 17–26. DOI: 10.1016/j.ins.2017.05.008
6. Yen S. J., Lee Y. S. Cluster-based under-sampling approaches for imbalanced data distributions, *Expert Systems with Applications*, 2009, Vol. 36, Issue 3, pp. 5718–5727. DOI: 10.1016/j.eswa.2008.06.108
7. 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
8. Wang B. X., Japkowicz N. Imbalanced Data Set Learning with Synthetic Samples [Electronic resource]. Access mode: http://www.iro.umontreal.ca/~lisa/workshop2004/ program.html
9. Subbotin S. O., Olijjnik A. O. Intelektual'ni sistemi : navch. posib. pid zag. red. prof. S. O. Subbotina. Zaporizhzhja, ZNTU, 2014, 218 p.
10. Elkan C. The foundations of cost-sensitive learning, *17th international joint conference on Artificial intelligence, Seattle, 4–10 August 2001 : Proceedings*. San Francisco, Morgan Kaufmann Publishers Inc., 2001, Vol. 2, pp. 973–978.
11. Fawcett T. An Introduction to ROC Analysis, *Pattern Recognition Letters*, 2006, Vol. 27, Issue 8, pp. 861–874. DOI: 10.1016/j.patrec.2005.10.010
12. Cover T., Hart P. Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, 1967, Vol. 13, Issue 1, P. 21–27. DOI: 10.1109/TIT.1967.1053964
13. Zagorujjko N. G. Prikladnye metody analiza dannykh i znanijj. Novosibirsk, IIM, 1999, 270 p.
14. Lloyd S. P. Least Squares Quantization in PCM, *IEEE Transactions on Information Theory,* 1982, Vol. 28, pp. 129–137. DOI: 10.1109/TIT.1982.1056489
15. Subotin S. O., Kavrin D. A. Avtomatizovana sistema vidboru optimal'nogo metodu vidnovlennja balansu klasiv pri formuvanni navchal'noï vibirki, *Informatika, upravlinnja ta shtuchnijj intelekt. Materiali chetvertoï mizhnarodnoï naukovotekhnichnoï konferenciï studentiv, magistriv ta aspirantiv*. Kharkiv, NTU "KhPI", 2017, P. 94.
16. Kokren U. Metody vyborochnogo issledovanija. Moscow, Statistika, 1976, 440 p.