УДК 004.93

Д. А. Каврин, С. А. Субботин
Запорожский национальный технический университет
ул. Жуковского, 64, 69063 Запорожье, Украина

Метод редукции мажоритарного класса в несбалансированных выборках

Рассмотрены проблемы формирования обучающих выборок для построения диагностических и распознающих моделей по прецедентам в условиях несбалансированности классов. Предложен метод автоматизации формирования обучающих выборок из исходных несбалансированных выборок большого размера. Метод позволяет значительно сократить размер исходной выборки с сохранением важных топологических свойств путем редукции мажоритарного класса и восстановить количественный баланс классов. Разработано программное обеспечение, реализующее предложенный метод, которое было использовано при проведении вычислительных экспериментов на синтетических и реальных данных. Проведенные эксперименты подтвердили работоспособность и эффективность предложенного метода и реализующего его программного обеспечения.

Ключевые слова: *выборка, классификация, метрика, мажоритарный класс, миноритарный класс, сэмплинг, экземпляр.*

Введение

Формирование обучающей выборки является важнейшим этапом создания диагностических и распознающих моделей по прецедентам. От таких характеритик обучающей выборки как размерность, репрезентативность и сбалансированность будет зависеть точность и скорость работы построенной модели, а также ее эффективность, определяемая соотношением полезного эффекта (качества) модели и количества ресурсов, требуемых на построение и использование модели.

Объектом исследования данной работы являлся процесс формирования обучающих выборок для построения диагностических и распознающих моделей по прецедентам.

Практически все реальные наборы данных являются несбалансированными, когда в наборе данных экземпляров одного класса больше чем экземпляров другого класса. Соотношение классов больше чем 10:1 приводит к серьезному снижению производительности классификаторов, если интерес представляет меньший

класс [1]. Другой проблемой является большой размер выборки, т.к. в этом случае увеличиваются затраты машинного времени на обработку данных. Данные проблемы могут значительно снизить эффективность модели.

Предметом исследования являлись методы формирования обучающих выборок из исходных выборок большого размера для построения диагностических и распознающих моделей по прецедентам в условиях несбалансированности классов.

Участие человека-эксперта при формировании выборки не всегда оказывается возможным, поскольку либо экспертный опыт ограничен, либо объем выборки столь велик, что за ограниченное время невозможно силами имеющихся специалистов осуществить ее обработку. Разработка метода формирования выборок, способного автоматически решать указанные выше проблемы, позволит повысить качество работы существующих и создаваемых моделей в реальных приложениях.

Целью работы являлась разработка метода предварительной обработки обучающей выборки, позволяющего восстановить баланс классов и минимизировать размер выборки, сохранив при этом основные топологические свойства исходной выборки.

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. При этом один из классов представлен значительно меньшим числом прецедентов (экземпляров), чем другие классы.

Тогда задача сокращения объема и восстановления баланса выборки состоит в создании из исходной несбалансированной выборки $X = \langle x, y \rangle$ такой сбалансированной по классам подвыборки меньшего объема $X' = \langle x', y' \rangle$, которая сохранила бы наиболее важные топологические свойства исходной выборки. Формально 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 сбалансированной выборке; x' — набор признаков в сбалансированной выборке; y' — выходной признак (класс) в сбалансированной выборке.

Обзор литературы

Известно много методов классификации, которые успешно применяются в различных практических задачах [2]. Однако, если выборка является несбалансированной, то производительность большинства стандартных классификаторов падает. Выборка называется несбалансированной, если в ней число экземпляров одного класса, называемого мажоритарный, значительно превосходит число экземп-

ляров другого класса, называемого миноритарный [3]. Поскольку в такой выборке экземпляров миноритарного класса значительно меньше, классификаторы имеют тенденцию редко предсказывать либо игнорировать экземпляры миноритарного класса, поэтому они неверно классифицируются чаще, чем экземпляры мажоритарного класса. Проблема несбалансированности классов требует разрешения, когда интерес представляет миноритарный класс.

Существует два основных подхода к решению проблемы несбалансированности классов [3, 4]. Первый подход подразумевает создание методов предварительной обработки обучающей выборки (сэмплинга) для восстановления баланса классов. Достоинства данного подхода в том, что он применяется на этапе формирования обучающей выборки и не требует модификации существующей модели. Другой подход предполагает создание специализированных алгоритмов классификации нечувствительных к дисбалансу классов. Данный подход менее универсальный и практически всегда требует создания новых алгоритмов классификации в каждом частном случае, что требует дополнительных затрат.

Целью методов сэмплинга является формирование выборки с относительно сбалансированным распределением классов для облегчения работы традиционных классификаторов. Обычно методы сэмплинга применяют для облегчения классификации экземпляров миноритарного класса. Все методы сэмплинга построены на двух базовых стратегиях: удаление экземпляров мажоритарного класса (undersampling) и генерирование экземпляров миноритарного класса (oversampling) [5]. Обе стратегии имеют свои достоинства и недостатки. Однако очевидно, что при больших размерах исходной выборки и высоком уровне соотношения классов предпочтительным будет выбор стратегий undersampling, так как данные стратегии могут значительно сократить размер выборки и увеличить скорость классификации.

Разработано множество методов, реализующих стратегию undersampling [6]. В основном данные методы объединяет наличие в разной мере стохастической составляющей, например метод CNN (Condensed Nearest Neighbor rule) [7], т.е. при использовании таких методов существует вероятность потери важных экземпляров мажоритарного класса, что может ухудшить производительность классификатора, возможно нивелировав эффект от применения метода. Поэтому актуальным является поиск решений, которые бы позволили применять стратегию undersampling, сохраняя при этом важные топологические свойства мажоритарного класса обучающей выборки.

Для оценки эффективности разработанных методов необходимо иметь механизм оценки работы классификатора в условиях несбалансированности классов. Дело в том, что традиционная метрика истинности (accuracy) $A = S^{true} / S$, где S^{true} — число верно предсказанных экземпляров, S — число экземпляров в исходной выборке, которая показывает долю правильных ответов классификатора, не может быть подходящей метрикой в условиях несбалансированности классов. Например, в ситуации, когда миноритарный класс представлен только 0,1 % обучающей выборки, простейший классификатор, всегда предсказывающий мажоритарный класс, будет иметь истинность 99,9 %. Однако этот показатель не будет иметь смысла для приложений, в которых задача обучения состоит в определении миноритарного класса.

Общим подходом определения производительности классификаторов является использование матрицы ошибок (confusion matrix) [8], которая представляет собой способ группировки экземпляров в зависимости от комбинации истинного ответа и ответа классификатора и позволяет получить множество различных метрик [3].

При работе с несбалансированными данными миноритарный класс обычно представляют как позитивный. В данном случае интерес представляют характеристики точности (precision) и полноты (recall). Точность $P = TP/(TP + FP)$, где TP — верно классифицированные позитивные экземпляры, FP — неверно классифицированные позитивные экземпляры, показывает долю верно предсказанных позитивных экземпляров. Полнота $R = TP/(TP + FN)$, где FN — неверно классифицированные негативные экземпляры показывает долю верно предсказанных позитивных экземпляров из всех экземпляров, предсказанных как позитивные экземпляры. 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) $F = 2PR/(P + R)$ [3, 9].

Материалы и методы

Предварительная обработка несбалансированных выборок с помощью стратегии undersampling [3, 5, 6] предполагает удаление экземпляров мажоритарного класса. Сохранить топологические свойства исходной несбалансированной выборки возможно с помощью сохранения в обработанной выборке значимых экземпляров мажоритарного класса, которые находятся на границе классов. Для решения данной задачи в [10] предлагается использовать метод минимизации набора экземпляров STOLP. В общем случае данный метод решает задачу уменьшения объема всех классов выборки, оставляя пограничные экземпляры каждого класса. Но при работе с несбалансированными выборками, когда интерес представляет миноритарный класс, данный метод может быть применен только к экземплярам мажоритарного класса. В таком случае его можно рассматривать как разновидность стратегии undersampling.

Метод STOLP [10] эффективно производит редукцию мажоритарного класса, но не позволяет задавать количество удаляемых экземпляров, поэтому не решает проблему несбалансированности классов. Причем, если классы достаточно хорошо разделены (компактны), метод может сильно уменьшить мажоритарный класс таким образом, что число экземпляров мажоритарного класса окажется значительно меньше числа экземпляров миноритарного класса. В этом случае, несмотря на хорошо определенную границу между классами, классификатор, например, такой как метод ближайших соседей kNN (k Nearest Neighbor) [11], может неверно распознавать новые экземпляры из-за разреженности мажоритарного класса. Поэтому для восстановления баланса классов в полученной выборке на следующем этапе предлагается случайным образом продублировать экземпляры мажоритарного класса, внося незначительные коррективы в их координаты.

Результатом работы полученного метода будет сбалансированная обучающая выборка размера $2S_{mi}$, где S_{mi} — число экземпляров миноритарного класса в исходной выборке (т.е., если доля миноритарного класса в исходной выборке составляет 1 %, размер обработанной выборки будет составлять 2 % от размера исходной выборки), в которой будут сохранены важные топологические свойства исходной выборки.

В общем виде предлагаемый метод может быть представлен как последовательность следующих этапов.

Этап 1. *Инициализация*. Установить: $X' = X_{mi}$, $V = X_{ma}$, где X_{mi} — множество прецедентов о зависимости миноритарного класса в исходной выборке; X_{ma} — множество прецедентов о зависимости мажоритарного класса в исходной выборке; V — временное множество прецедентов о зависимости.

Этап 2. *Анализ экземпляров мажоритарного класса*. Найти экземпляр мажоритарного класса x^w с максимальным значением величины риска неверной классификации w

$$w = \arg\max_{s=1,S_{ma}} \left\{ \frac{\rho_{in}^s}{\rho_{out}^s} \right\},$$

где $\rho_{in}^s = \arg\min\limits_{\substack{p=1,S_{ma} \\ p \neq s}} \left\{ \sqrt{\sum_{j=1}^{N} \left(x_j^s - x_j^p \right)^2} \right\}$ — расстояние от s-го экземпляра до ближайшего соседа своего класса; $\rho_{out}^s = \arg\min\limits_{p=1,S_{mi}} \left\{ \sqrt{\sum_{j=1}^{N} \left(x_j^s - x_j^p \right)^2} \right\}$ — расстояние от s-го экземпляра до ближайшего соседа другого класса; x_j^s (x_j^p) — j-й признак s-го (p-го) экземпляра выборки.

Этап 3. *Поиск критических экземпляров мажоритарного класса, находящихся на границе классов*. Сформировать две выборки: $X' = X' \bigcup x^w$ и $V = V \setminus x^w$. Классифицировать подвыборку V с помощью метода одного ближайшего соседа [11] по обучающей выборке X'.

Этап 4. *Переопределение исходной выборки*. Выделить множество неверно классифицированных экземпляров мажоритарного класса $V = V^{fn}$, где V^{fn} — множество неверно классифицированных негативных (мажоритарных) экземпляров. Если $V \neq \varnothing$, тогда переопределить выборку $X = V \bigcup X_{mi}$ и перейти к этапу 2.

Этап 5. *Восстановление баланса выборки* X'. Если $X'_{ma} < X'_{mi}$, продублировать каждый экземпляр мажоритарного класса $m = round\left(X'_{mi} / X'_{ma} \right)$ раз:

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

где X'_{ma} — множество прецедентов о зависимости мажоритарного класса в обработанной выборке; X'_{mi} — множество прецедентов о зависимости миноритарного класса в обработанной выборке; x_{ma}^i — i-й дублируемый экземпляр мажоритарного класса.

При этом желательно, внести небольшие коррективы в координаты

$$x_j = x_j(1 + 0{,}1\text{rand} - 0{,}1\text{rand}),$$

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

Предложенная модификация базового метода STOLP [10] на заключительном этапе формирования обучающей выборки восстановит баланс классов за счет случайного дублирования экземпляров мажоритарного класса, что позволит повысить точность классификации построенной модели.

Эксперименты

Для исследования свойств предложенного метода использован программный комплекс «Автоматизированная система отбора оптимального метода восстановления баланса классов при формировании обучающей выборки» [14], в которую был интегрирован разработанный модуль STOLP_b, представляющий собой функцию редукции мажоритарного класса обучающей выборки и восстановления баланса классов.

Исследования включали два этапа: на первом — была проанализирована работа базового метода STOLP [10] и модифицированного метода STOLP_b; на втором — проводился сравнительный анализ базовых методов восстановления баланса классов и метода STOLP_b для различных несбалансированных наборов данных.

В качестве наборов данных использовались бинарные выборки, которые были разделены методом стратификации [12] на обучающие и тестовые выборки в соотношении 4:1.

Для исследования методов при построении тестовой модели использовался метод *k*-ближайших соседей kNN, в основе которого лежит гипотеза о компактности классов [10], предполагающая, что тестируемый экземпляр будет относиться к тому же классу, что и экземпляры из его ближайшего окружения. Решающие правила строились по принципу большинства голосов, поэтому для однозначности выбора в работе использовались методы с нечетным числом ближайших соседей (k = 1, 9, 17, 25, 33, 41, 49, 57, 65, 77, 81, 97).

Оценка работы методов производилась с помощью метрики *F*-measure, значение которой рассчитывалось для различных параметров выборки и классификатора kNN, затем строились зависимости *F*-measure от параметра *k*-ближайших соседей метода kNN для набора исследуемых методов сэмплинга и разных наборов несбалансированных данных. Также для каждой тестовой модели строилась зависимость числа экземпляров каждого класса сформированной обучающей выборки от применяемого метода сэмплинга, которая позволяла оценить степень уменьшения исходной выборки.

Результаты

Результаты исследований представлены на рис. 1–4. Рис. 1,*а*, 2,*а*, 3,*а* и 4,*а* иллюстрируют зависимости метрики *F*-measure от числа ближайших соседей клас-

сификатора kNN для различных методов сэмплинга. Рис. 1,*б*, 2,*б*, 3,*б* и 4,*б* отображают объемы выборок по классам после обработки соответствующими методами.

Результаты первого этапа исследований представлены на рис. 1. Методы STOLP и STOLP_b были апробированы на синтетической несбалансированной бинарной выборке объемом сто тысяч экземпляров с соотношением классов 100:1.

Результаты второго этапа исследований представлены на рис. 3–4. На данном этапе сравнивались результаты работы тестовых моделей с использованием следующих методов сэмплинга: undersampling, oversampling, CNN (Condensed Nearest Neighbor Rule), STOLP_b, а также модель с исходной выборкой без применения семплинга (Original). Данные методы были исследованы на синтетических бинарных выборках объемом сто тысяч экземпляров с различными соотношениями классов и на реальной выборке (рис. 4) задачи определения пульсаров [13], которая имеет объем 3 580 экземпляров и соотношение классов 100:10.

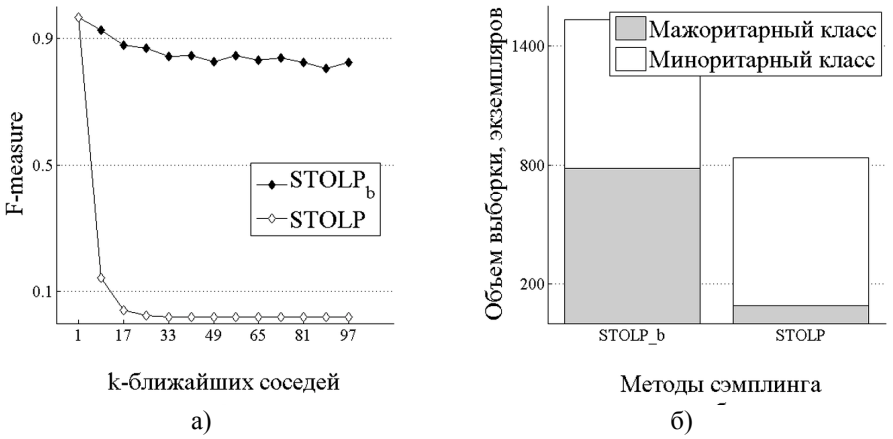

Рис. 1. Графики зависимостей параметров тестовой модели для синтетической выборки с соотношением классов 100:1 и объемом 100 000 экземпляров: а) *F*-measure от *k*-ближайших соседей метода kNN, б) числа экземпляров от применяемого метода сэмплинга

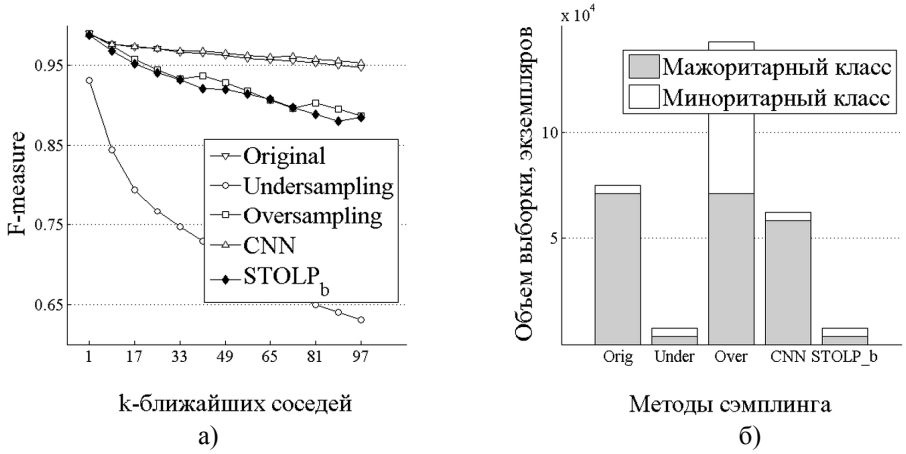

Рис. 2. Графики зависимостей параметров тестовой модели для синтетической выборки с соотношением классов 100:5 и объемом 100 000 экземпляров: а) *F*-measure от *k*-ближайших соседей метода kNN; б) числа экземпляров от применяемого метода сэмплинга

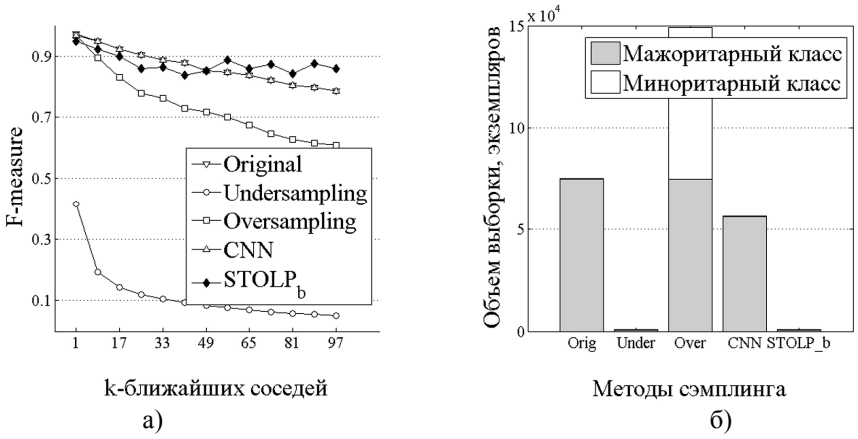

Рис. 3. Графики зависимостей параметров тестовой модели для синтетической выборки с соотношением классов 100:0,5 и объемом 100 000 экземпляров: а) *F*-measure от *k*-ближайших соседей метода kNN; б) числа экземпляров от применяемого метода сэмплинга

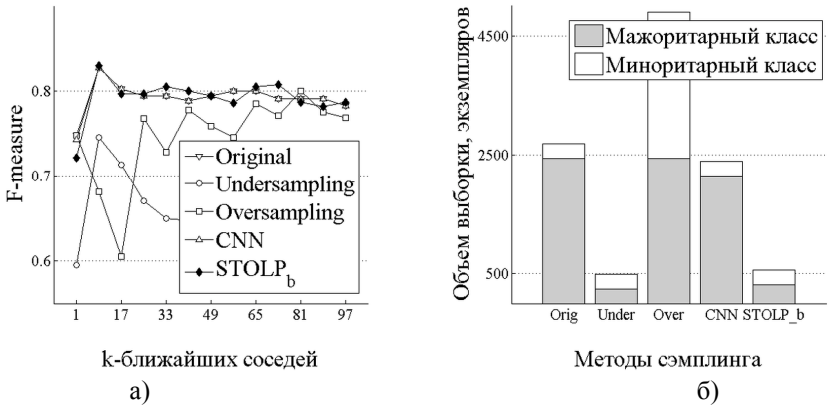

Рис. 4. Графики зависимостей параметров реальной выборки для задачи определения пульсаров [14] с соотношением классов 100:10 и объемом 3 580 экземпляров: а) *F*-measure от *k*-ближайших соседей метода kNN; б) числа экземпляров от применяемого метода сэмплинга

Обсуждение

Результат сравнения базового метода STOLP [10] и его модификации STOLP_b, представленной в данной работе, показан на рис. 1. Здесь хорошо видно, что для данной выборки оказалось недостаточным применение только базового метода STOLP, поэтому возникла ситуация, при которой классификатор не смог правильно классифицировать большинство экземпляров тестовой выборки. Использование же модифицированного метода STOLP_b позволило качественно повысить точность классификации, благодаря возможности балансировки классов.

Из рис. 2–4 видно, что метод односторонней редукции STOLP_b, в сравнении с другими методами сэмплинга [3, 5, 6], показал отличное соотношение показателя производительности тестовой модели, выраженного характеристикой гармоничного среднего (*F*-measure) и размера сформированной выборки. Причем, при увеличении соотношения дисбаланса классов в исходной выборке значение *F*-measure тестовой модели, построенной с помощью предложенного метода, повышалось относительно остальных методов. Это позволяет предположить, что

использование предложенного метода при формировании обучающих выборок из сильно несбалансированных наборов данных позволит повысить точность построенных моделей в сравнении с традиционными методами сэмплинга.

Выводы

В работе решена задача редукции и восстановления баланса классов с помощью метода автоматического формировании обучающих выборок из исходных несбалансированных выборок большого размера для построения моделей по прецедентам STOLP_b.

Перспективы дальнейших исследований могут заключаться в изучении предложенного метода на более широком классе практических задач, особенно при работе с несбалансированными выборками большого размера. Также возможны разработки реализаций предложенного метода для многопроцессорных систем, работающих в параллельных режимах с целью повышения скорости формирования обучающих выборок из исходных выборок большого размера.

1. Weiss G.M., He H., Ma Y. Foundations of Imbalanced Learning. Imbalanced Learning: Foundations, Algorithms, and Applications. Hoboken. NJ, USA: John Wiley & Sons, 2013. P. 13–42. DOI: 10.1002/9781118646106.ch2.

2. Fernandez-Delgado M, E. Cernadas S., Barro D. Amorim Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*. 2014. Vol. 15. P. 3133–3181.

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. P. 687–719. DOI: 10.1142/S0218001409007326.

4. Beyan C., Fisher R. Classifying imbalanced data sets using similarity based hierarchical decomposition. *Pattern Recognition*. 2015. Vol. 48, Issue 5. P. 1653–1672. DOI: 10.1016/j.patcog.2014.10.032.

5. 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. P. 20–29. DOI: 10.1145/1007730.1007735.

6. He H., Garcia E.A. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*. 2009. Vol. 21. P. 1263–1284. DOI: 10.1109/TKDE.2008.239.

7. Hart P. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*. 1968. Vol. 14. Issue 3. P. 515–516. DOI: 10.1109/TIT.1968.1054155.

8. Elkan C. The foundations of cost-sensitive learning. 17th Internation. joint Conf. on Artificial intelligence, Seattle, 4-10 August 2001: proceedings. San Francisco: *Morgan Kaufmann Publishers Inc*. 2001. Vol. 2. P. 973–978.

9. Fawcett T. An Introduction to ROC Analysis. *Pattern Recognition Letters*. 2006. Vol. 27. Issue 8. P. 861–874. DOI: 10.1016/j.patrec.2005.10.010.

10. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: ИИМ, 1999. 270 с.

11. 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.

12. Кокрен У. Методы выборочного исследования. Москва: Статистика, 1976. 440 с.

13. Lyon R.J. HTRU2. URL: https://figshare.com/articles/HTRU2/3080389/1. DOI: 10.6084/m9.figshare.3080389.v1.

14. Суботін С.О., Каврін Д.А. Автоматизована система відбору оптимального методу відновлення балансу класів при формуванні навчальної вибірки. *Інформатика, управління та штучний інтелект*. Матеріали четвертої міжнародної науковотехнічної конференції студентів, магістрів та аспірантів. Харків: НТУ «ХПІ», 2017. С. 94.

Поступила в редакцию 28.02.2018