УДК 681.518
DOI: 10.15593/2224-9397/2020.4.05

А.Г. Марахтанов, Е.О. Паренченков, Н.В. Смирнов

Петрозаводский государственный университет, Петрозаводск, Россия

ОПРЕДЕЛЕНИЕ ЭЛЕКТРОННОГО МОШЕННИЧЕСТВА МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ В СЛУЧАЕ НЕСБАЛАНСИРОВАННОГО НАБОРА ДАННЫХ

Существуют различные типы электронного мошенничества, связанные с хищением средств с банковских счетов и карт, легализацией полученных незаконным путем финансовых средств, получением кредитов, и пр. Один из типов мошенничества состоит в регистрации фиктивных аккаунтов пользователей и получении финансового вознаграждения от заинтересованной в увеличения популярности и прибыльности электронного ресурса стороны. В случае обнаружения заявки на регистрацию предположительно фиктивного аккаунта электронный ресурс может подвергнуть заявку дополнительной проверке. Набор данных в задаче определения электронного мошенничества, как правило, сильно несбалансирован: количество экземпляров записей о регистрациях фиктивных аккаунтов значительно меньше количества записей о регистрации действительных аккаунтов. **Цель исследования:** построение классификатора, использующего методы машинного обучения для выявления регистрации фиктивных аккаунтов в случае обучения на сильно несбалансированном наборе данных. **Результаты:** рассмотрены существующие решения задачи противодействия электронному мошенничеству с использованием методов машинного обучения. Для решения поставленной задачи на языке программирования Python написан комплекс программ. Были произведены предобработка данных и анализ полученных из них признаков. В работе представлены результаты корреляционного анализа признаков и результаты применения различных алгоритмов классификации (случайного леса, бэггинга, сбалансированного бэггинга и LinearSVC) в задаче бинарной классификации аккаунтов на действительные и фиктивные. Также представлены результаты применения различных алгоритмов для уменьшения влияния на процесс обучения несбалансированности классов набора данных: SMOTE, ADASYN, NCR, Tomek Links, CNN. Для указанных алгоритмов в работе приведены полученные матрицы ошибок и значения метрик. Произведено сравнение метрик и выбран алгоритм, обеспечивающий лучшие результаты классификации. **Практическая значимость:** использование примененных в работе методов машинного обучения позволит автоматизировать процесс выявления при регистрации фиктивных аккакунтов.

Ключевые слова: электронное мошенничество, машинное обучение, бустинг, случайный лес, несбалансированный набор данных.

A.G. Marakhtanov, E.O. Parenchenkov, N.V. Smirnov

Petrozavodsk State University, Petrozavodsk, Russian Federation

FRAUD DETECTION BY MACHINE LEARNING METHODS IN THE CASE OF AN IMBALANCED DATASET

There are various types of fraud: related to the theft of funds from bank accounts and cards, money laundering, mortgage fraud, etc. One of the fraud types is registering fictitious user accounts and receiving financial rewards from people or companies, those interested in increasing popularity and profitability of an electronic resource. If an application for registration of a presumably fictitious account is found, the electronic resource may subject the application to additional verification. The dataset in the identifying electronic fraud task is usually highly imbalanced: the number of applications for fictitious accounts registration is significantly less than the number of applications for valid accounts registration. **Purpose**: Creating a classifier using machine learning methods to detect registration of fictitious accounts in the case of using a highly imbalanced training dataset. **Results**: We considered the existing solutions to the task of combating fraud, witch using machine learning methods. The complex of programs has been written in the Python programming language to solve the task. Data preprocessing and analysis of the features obtained from them were carried out. The paper presents the results of the features correlation analysis and the results of applying various classification algorithms (random forest, bagging, balanced bagging, and LinearSVC) in the task of binary classification of accounts into real and fictitious. Also the results of using various algorithms to reduce the impact on the learning process of the dataset classes imbalance: SMOTE, ADASYN, NCR, Tomek Links, CNN are presented in the paper. The obtained confusion matrices and metrics for the specified algorithms are presented. A comparison of the metrics was made and the algorithm that provides the best classification results was selected. **Practical relevance:** The use of machine learning methods used in the work can automate the process of detecting fictitious accounts during registration.

Keywords: fraud, machine learning, boosting, random forest, imbalanced dataset.

Введение. Разновидностью мошенничества в области информационных технологий является фрод (от анг. fraud). Часто с понятием фрода ассоциируют операции, связанные с хищением финансовых средств с различных банковских карт и счетов. Существуют и другие типы фрода, например, предназначенные для получения финансовой выгоды за привлечение и регистрацию новых пользователей, увеличение популярности веб-ресурса. Метрики популярности веб-ресурсов влияют на частоту и порядок их отображения в поисковых запросах, а формулы для их расчета часто включают такие показатели, как количество зарегистрированных пользователей, посещений/просмотров веб-ресурса, оставленных сообщений, время, проведенное на веб-ресурсе, и пр. Для искусственного увеличения указанных показателей используют компьютерные программы (боты) и труд людей. Увеличение степени использования различных электронных информационных ресурсов обусловливает актуальность задачи нахождения фрода и необходимость автоматизации этого процесса.

Новизна и оригинальность работы состоят в предложенном алгоритме предобработки данных, выделения признаков и применения методов машинного обучения в задаче классификации аккаунтов пользователей. Целью работы является построение на основе методов машинного обучения классификатора для решения задачи бинарной классификации: отнесения зарегистрированного аккаунта к классу действительных (не фрод) или фиктивных (фрод) в случае обучения на сильно несбалансированном наборе данных. Для достижения поставленной цели:

– произведен анализ работ по применению методов машинного обучения в задаче определения других типов фрода (не связанных с регистрацией фиктивных аккаунтов);

– разработан и применен алгоритм предобработки входного набора данных и выделения признаков, значения которых передавались в методы машинного обучения;

– получены и проанализированы результаты классификации различными методами машинного обучения;

– произведены численные эксперименты и представлены результаты применения различных алгоритмов балансировки данных;

– выбран метод машинного обучения, обеспечивающий максимизацию указанных ниже метрик классификации.

В широком спектре задач анализа данных используются методы машинного и глубокого обучения [1–7]. Решение задачи классификации фрода с помощью различных эвристических методов часто влечет чрезмерно большие затраты времени, в этой задаче все чаще используют методы машинного обучения [8–11]. Работа [8] посвящена анализу проблем в существующих системах для противодействия легализации полученных преступным путем финансовых средств. Авторы в [8] используют методы обнаружения аномалий в данных, выделяют шаблоны мошеннических схем, полученные результаты предлагают использовать в методах машинного обучения. Авторы [9] предлагают при классификации транзакции вместе с данными о транзакции использовать дополнительные данные (информацию об устройстве, с которого произведена транзакция, характеристики аппаратного и программного обеспечения), для идентификации пользователя предлагают использовать Finger Print (цифровой отпечаток на основе множества параметров устройства, обеспечивающий высокую точность иденти-

фикации клиента). Наиболее перспективными алгоритмами для поиска фрода авторы [9] считают ансамблевые (бэггинг), «слабые» классификаторы (бустинг) и сочетания различных алгоритмов (стекинг). В [10] авторы производят анализ различных типов мошенничества, связанного с легализацией денежных средств, кредитными картами и ипотечным кредитованием, а также предлагают использовать метод опорных векторов для выявления мошенничества, при этом определение мошенничества происходит в два этапа: на первом, используя обучение с учителем, выявляется мошенничество, если мошенничество не выявлено, то на втором этапе используется обучение без учителя для поиска аномальных транзакций, которых не было зафиксировано ранее.

Набор данных. Для анализа был получен набор данных о регистрациях пользователей в системе. Данные о пользователях обезличены, при этом известны такие параметры, как домен электронной почты, длина пароля, год рождения, идентификаторы страны пользователя и языка, различные параметры устройства: IP-адрес, тип устройства, операционной системы и браузера, а также сделанная экспертом отметка, является ли аккаунт действительным или фиктивным.

В наборе данных 1328794 записи, из которых 1328181 отмечены как не фрод, а 40612 как фрод. Фроду соответствует 3,06 % записей исходного набора данных, что обусловливает сильную несбалансированность последнего. Данные подверглись предобработке: удалены неинформативные признаки, такие как системные идентификаторы, не влияющие на работу классификатора. Все признаки были преобразованы к числовому виду. Признак isFraud принимает значение 0, если регистрация отмечена как немошенническая, и 1 – иначе.

Для удобства отображения признаки переименованы в p1, p2,..., p29. Для признаков с числовым значением была построена матрица корреляции (рис. 1). У различных признаков наибольший по модулю коэффициент корреляции между признаками p16 и p22, но он приблизительно равен – 0,4, следовательно, зависимости между признаками не наблюдается. Признаки подверглись предобработке. Все признаки были приведены к числовому типу данных либо с помощью one-hot encoding в случае категориальных признаков с малым числом возможных значений, либо с помощью хеширования. One-hot encoding подразумевает сопоставление каждой категории признака переменной, которая равна 1, если исходный признак имеет соответствующую катего-

рию, иначе 0. Хеширование позволяет преобразовать строковое значение признака в числовое путем вычисления хеш-функции от исходного значения. Значение хеш-функции уникально для каждого конкретного значения признака.

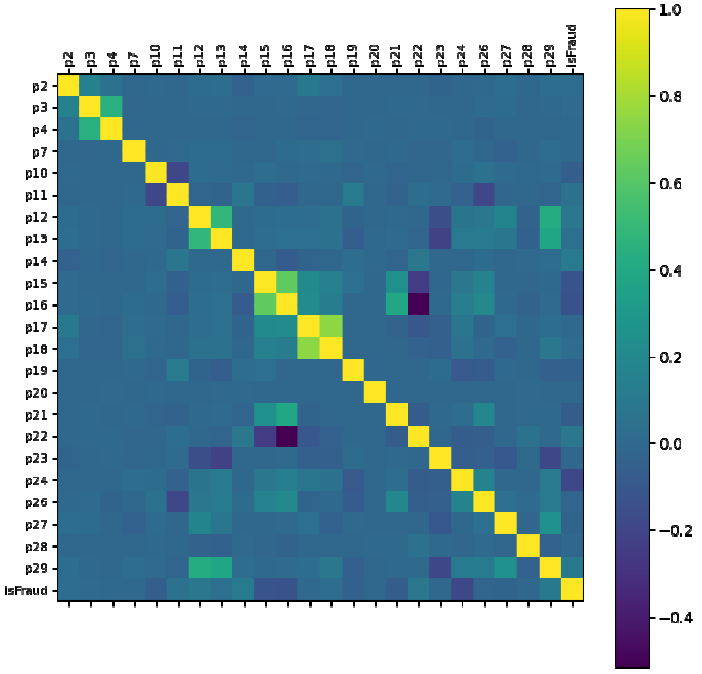

Рис. 1. Матрица корреляции числовых признаков

Классификация. К предобработанному набору данных были применены алгоритмы случайного леса, бэггинга, сбалансированного бэггинга, LinearSVC.

Случайный лес представляет собой множество решающих деревьев, каждое из которых обучается на определенной части обучающего множества. При построении деревьев леса для каждого дерева выбирается подвыборка обучающей выборки с возвращением или без, что зависит от параметров модели. При классификации алгоритмом случайного леса классом записи считается тот класс, который выбрали большинство деревьев.

В работе рассмотрен бэггинг на основе дерева решений. Идея бэггинга заключается в построении ансамбля классификаторов, которые обучаются на подвыборках из обучающего множества и принимают итоговое решение путем голосования, при этом базовый классифи-

катор может быть произвольным. Отличием от алгоритма случайного леса, в котором обучение происходит на случайно выбранном ограниченном подмножестве признаков, является то, что бэггинг использует для обучения все признаки.

Сбалансированный бэггинг – модификация бэггинга, при которой производится предварительная балансировка количества элементов разных классов путем удаления некоторого количества элементов мажоритарного класса. В остальном сбалансированный бэггинг ведет себя аналогично базовому алгоритму бэггинга.

LinearSVC – алгоритм семейства машин опорных векторов (SVM). Его отличия от других алгоритмов семейства SVM:

– разделяющая гиперплоскость является линейной;

– высокое быстродействие алгоритма.

Ввиду несбалансированности набора данных также были произведены эксперименты с различными алгоритмами балансировки данных (ресемплинга). Ресемплинг (resampling) – это добавление или удаление элементов выборки с целью балансировки количества элементов, принадлежащих к разным классам. Ресемплинг бывает двух типов: оверсемплинг (oversampling) и андерсемплинг (undersampling). Оверсемплинг – это добавление в выборку новых элементов, а андерсемплинг – удаление элементов из нее. В работе рассмотрены следующие алгоритмы ресемплинга: CNN, NCR, Tomek Links, SMOTE, ADASYN. Основные алгоритмы ресемплинга изложены в [12].

CNN (Condensed Nearest Neighbour) – это алгоритм, позволяющий произвести андерсемплинг мажоритарного класса на основе алгоритма одного ближайшего соседа (модификации алгоритма *k* ближайших соседей [13]). В результате выполнения этого алгоритма в выборке остается необходимое для решения задачи классификации количество элементов, остальные элементы можно рассматривать как выбросы [14].

NCR (Neighbourhood Cleaning Rule) позволяет производить андерсемплинг путем удаления из выборки элементов мажоритарного класса, которые в пространстве признаков близки к элементам миноритарного класса и затрудняют их верную классификацию. Его применение предлагается в [15].

Еще одним алгоритмом, позволяющим произвести андерсемплинг выборки, является удаление связей Томека. Связь Томека

(Tomek link) – это пара элементов выборки x и y, принадлежащих к разным классам, такая, что для любого элемента z из этой выборки будет справедлива совокупность неравенств:

$$\begin{cases} d(x, y) < d(x, z), \\ d(x, y) < d(y, z), \end{cases}$$

где d – евклидово расстояние в пространстве признаков. Соответственно, андерсемплинг заключается в удалении элементов, между которыми есть связь Томека. Поиск связей Томека часто используется в задачах классификации [16].

SMOTE и ADASYN представляют собой алгоритмы оверсемплинга миноритарного класса. Synthetic Minority Oversampling Technique (SMOTE) заключается в создании новых элементов миноритарного класса на основе уже существующих, при этом новые элементы находятся между уже существующими элементами миноритарного класса в пространстве признаков. SMOTE часто применяется в сочетании с ансамблевыми классификаторами [17, 18]. Adaptive Synthetic Sampling Method (ADASYN) отличается от SMOTE тем, что добавляет к сгенерированным элементам миноритарного класса некоторые случайные отклонения (шум). Вариант его использования приведен в [19]. Авторами [20] произведен сравнительный анализ моделей с использованием SMOTE и ADASYN. Также алгоритмы оверсемплинга можно сочетать с алгоритмами андерсемплинга: в [21] предложено сочетание поиска связей Томека и SMOTE.

При проведении экспериментов использовались реализации вышеперечисленных алгоритмов из библиотек Scikit-Learn и Imbalanced-Learn для языка программирования Python. Первоначально были проведены эксперименты с алгоритмами андерсемплинга. Андерсемплинг алгоритмом NCR привел к удалению 389 записей из всего набора данных, а использование сочетания Tomek Links + CNN удалило только 53 записи, что пренебрежимо мало по сравнению со всем объемом данных.

Оверсемплинг производился с помощью SMOTE и ADASYN. Сочетание их с классификаторами на базе бэггинга и сбалансированного бэггинга привело к чрезмерно большим временным затратам на процесс обучения, поэтому дальнейшее исследование с использованием этих модификаций не проводилось. Для моделей, сочетающих SMOTE и ADASYN со случайным лесом и LinearSVC, метрики классификации приведены в таблице.

Результаты тестирования классификаторов

Алгоритм	Точность (%) на тестовой выборке	Полнота (%) на тестовой выборке	F_1-мера (%) на тестовой выборке	Коэффициент детерминации R^2 на тестовой выборке
Случайный лес	99	99	99	0,99
Сбалансированный бэггинг	88	99	94	0,86
Бэггинг	98	99	99	0,97
LinearSVC	18	90	29	–3,46
Случайный лес + SMOTE	99	99	99	0,98
Случайный лес + ADASYN	99	99	99	0,99
LinearSVC + SMOTE	18	91	31	–3,25
LinearSVC + ADASYN	16	92	28	–4,01

При построении классификаторов набор данных был разделен на обучающую и тестовую выборки в соотношении 70/30 %. Классам были сопоставлены метки: записям, отмеченным как действительные (не фрод), была присвоена метка 0 (отрицательный класс), записям, отмеченным как фиктивные (фрод), – метка 1 (положительный класс). Оценка работы алгоритмов производилась по нескольким метрикам: точность (precision) (1), полнота (recall) (2), F_1-мера (3), коэффициент детерминации R^2 (4), а также по матрице ошибок:

$$\text{precision} = \frac{TP}{TP + FP}, \tag{1}$$

$$\text{recall} = \frac{TP}{TP + FN}, \tag{2}$$

$$F_1 = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}}, \tag{3}$$

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y}_i)^2}, \tag{4}$$

где y_i – истинное значение метки класса для i-го элемента тестовой выборки, $\hat{y}_i$ – предсказанное значение метки класса для i-го элемента тестовой выборки, $\bar{y}$ – среднее значение метки класса для всех

элементов тестовой выборки, N – объем тестовой выборки, TP (true positive rate) – количество элементов выборки, верно классифицированных как принадлежащих к классу фиктивных регистраций, TN (true negative rate) – количество элементов выборки, верно классифицированных как принадлежащих к классу действительных регистраций, FP (false positive rate) – количество элементов выборки, принадлежащих к классу действительных регистраций, но классифицированных как фиктивные, FN (false negative rate) – количество элементов выборки, принадлежащих к классу фиктивных регистраций, но классифицированных как действительные.

В матрицах ошибок (рис. 2 и 3) в строках указаны действительные классы экземпляров набора данных, в столбцах указаны предсказанные им классы, класс действительных заявок на регистрацию обозначен как не фрод, а фиктивных заявок – как фрод, в верхнем левом углу матрицы – значение TN, в верхнем правом – FP, в нижнем левом – FN, в нижнем правом – TP.

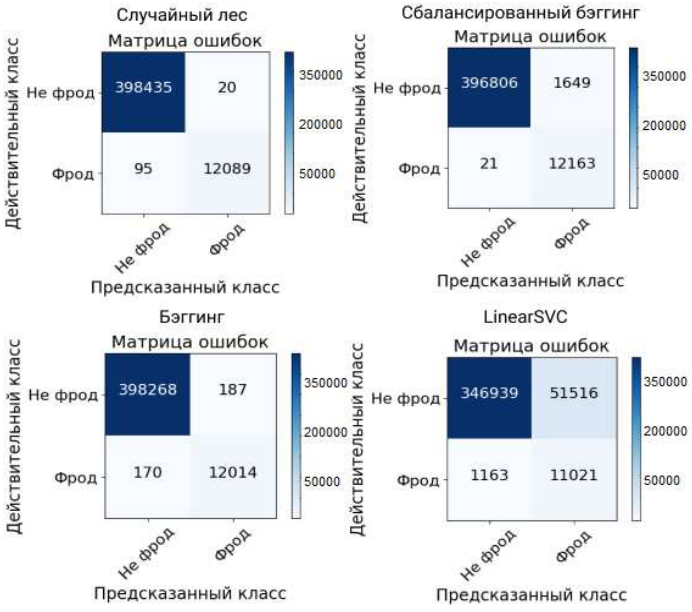

Рис. 2. Матрицы ошибок для моделей без использования алгоритмов оверсемплинга

Наибольшие значения метрик F_1 и R^2, а также наименьшее количество ошибок классификации были получены в результате использования алгоритма случайного леса.

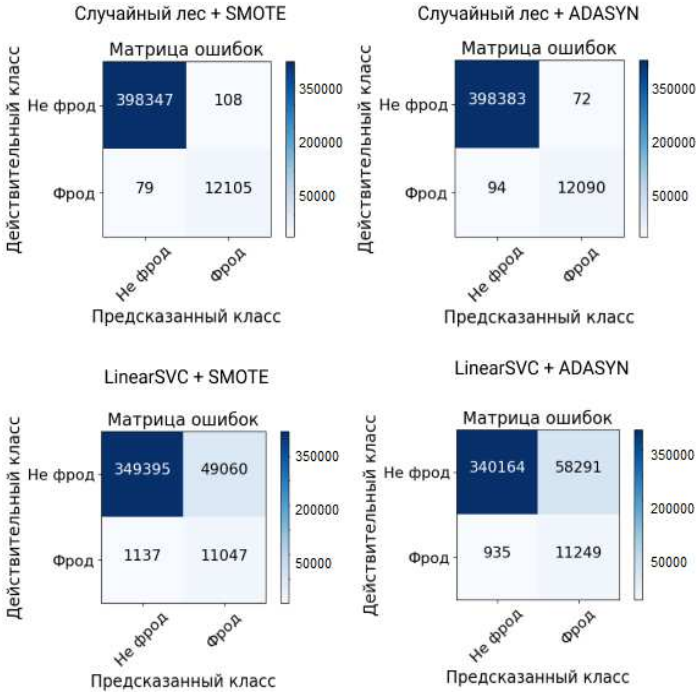

Рис. 3. Матрицы ошибок для моделей с использованием алгоритмов оверсемплинга

Применение оверсемплинга и андерсемплинга вместе со случайным лесом не обеспечило увеличение F_1 и R^2.

Выводы. Рассмотрена задача бинарной классификации аккаунтов пользователей на действительные и фиктивные с помощью методов машинного обучения. Проанализированы существующие методы решения этой задачи. Апробировано несколько классификаторов на основе различных алгоритмов машинного обучения в условиях несбалансированного набора данных. В результате анализа метрик, полученных при использовании различных классификаторов, сделан вывод, что наименьшее количество ошибок классификации обеспечивает классификатор на основе случайного леса. Для увеличения значения метрик планируется проведение оптимизации параметров алгоритма случайного леса, а также проведение экспериментов с применением в рассматриваемой задаче других методов машинного обучения. Применение указанного в статье алгоритма классификации аккаунтов позволит выявлять, отправлять на дополнительную проверку и отклонять заявки на регистрацию фиктивных аккакунтов. Практическая значи-

мость состоит в экономии значительных финансовых средств за счет отказа от выдачи финансового вознаграждения за регистрацию фиктивных аккаунтов.

Библиографический список

1. Eremenko I.I., Oliunina I.S. Use of Machine Learning Methods for Solving Problem of User Identifying by Keyboard Handwriting // 2019 International Russian Automation Conference (RusAutoCon). – 2019. DOI: https://doi.org/10.1109/RUSAUTOCON.2019.8867767

2. Bikmullina I., Andreyanov N., Medvedev M. Stand for Development of Tasks of Detection and Recognition of Objects on Image // 2019 International Russian Automation Conference (RusAutoCon). – 2019. DOI: https://doi.org/10.1109/RUSAUTOCON.2019.8867608

3. Face recognition based on an improved center symmetric local binary pattern / N.-N. Zhou, A.G. Constantinides, G. Huang, Z. Shaobai // Neural Computing and Applications. – 2017. – Vol. 30. – P. 3791–3797. DOI: https://doi.org/10.1007/s00521-017-2963-2

4. Papakostas M., Iannakopoulos T. Speech-music discrimination using deep visual feature extractors // Expert Systems with Applications. – 2018. – Vol. 114. – P. 334–344. DOI: https://doi.org/10.1016/j.eswa.2018.05.016

5. Robalinho M., Fernandes P. Software 2.0 for Scrap Metal Classification // 16th Int. Conf. on Informatics in Control, Automation and Robotics. – 2019. – P. 666–673. DOI: https://doi.org/10.5220/0007977506660673

6. Ojala T., Pietikainen M., Maenpaa T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – Vol. 24, No. 7. – P. 971–987. DOI: https://doi.org/10.1109/TPAMI.2002.1017623

7. Kudugunta S., Ferrara E. Deep Neural Networks for Bot Detection // Information Sciences. – 2018. – Vol. 467. – P. 312–322. DOI: https://doi.org/10.1016/j.ins.2018.08.019

8. Application of Machine Analysis Algorithms to Automate Implementation of Tasks of Combating Criminal Money Laundering / D. Dorofeev, M. Khrestina, T. Usubaliev [et al.] // DTGS 2018: Digital Transformation and Global Society. – 2018. – P. 3757–385. DOI: https://doi.org/10.1007/978-3-030-02843-5_30

9. Anti-fraud system on the basis of data mining technologies / M.U. Sapozhnikova, A.V. Nikonov, A.M. Vulfin [et al.] // 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT). – 2017. – P. 2437–248. DOI: https://doi.org/10.1109/ISSPIT.2017.8388649

10. Abdelhamid D., Khaoula S., Atika O. Automatic Bank Fraud Detection Using Support Vector Machines // Proceedings of the International conference on Computing Technology and Information Management. – 2014. – P. 10–17.

11. Gyamfi N. K., Abdulai J.-D. Bank Fraud Detection Using Support Vector Machine // 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). – 2018. – P. 37–41. DOI: https://doi.org/10.1109/IEMCON.2018.8614994

12. T. Ryan Hoens, Nitesh V. Chawla. Imbalanced Datasets: From Sampling to Classifiers // Imbalanced Learning: Foundations, Algorithms, and Applications. – 2013. – P. 43–59. DOI: https://dx.doi.org/10.1002/2F9781118646106.ch3

13. Sun M., Yang R. An efficient secure k nearest neighbor classification protocol with high-dimensional features // International Journal of Intelligent Systems. – 2020. – Vol. 35. – P. 17911–813. DOI: https://doi.org/10.1002/int.22272

14. Batchanaboyina M.R., Devarakonda N. Design and Evaluation of Outlier Detection Based on Semantic Condensed Nearest Neighbor // Journal of Intelligent Systems. – 2019. – Vol. 29. – P. 1416–1424. DOI: https://doi.org/10.1515/jisys-2018-0476

15. Agustianto K., Destarianto P. Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling // Proceedings – 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering, ICOMITEE 2019. – 2019. – P. 86–89. DOI: https://doi.org/10.1109/ICOMITEE.2019.8921159

16. Rodolfo M. Pereira, Yandre M.G. Costa, Carlos N. Silla Jr. MLTL: A multi-label approach for the Tomek Link undersampling algorithm // Neurocomputing. – 2020. – Vol. 383. – P. 95–105. DOI: https://doi.org/10.1016/j.neucom.2019.11.076

17. A new method of diesel fuel brands identification: SMOTE oversampling combined with XGBoost ensemble learning / S. Wang, S. Liu, J. Zhang [et al.] // Fuel. – 2020. – Vol. 282. – P. 1–9. DOI: https://doi.org/10.1016/j.fuel.2020.118848

18. Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods / L. Wang, Z.Q.

Zhao, Y.H. Luo [et al.] // Computer Methods and Programs in Biomedicine. – 2020. – Vol. 196. – P. 1–14. DOI: https://doi.org/10.1016/j.cmpb.2020.105567

19. Liu C., Zhu L. A two-stage approach for predicting the remaining useful life of tools using bidirectional long short-term memory // Measurement. – 2020. – Vol. 164. – P. 1–12. DOI: https://doi.org/10.1016/j.measurement.2020.108029

20. Taneja S., Suri B., Kothari C. Application of Balancing Techniques with Ensemble Approach for Credit Card Fraud Detection // 2019 International Conference on Computing, Power and Communication Technologies, GUCON 2019. – 2019. – P. 753–758. – URL: https://ieeexplore.ieee.org/document/8940539

21. Detection of Non-Technical Losses Using SOSTLinkand Bidirectional Gated Recurrent Unit to SecureSmart Meters / Gul Hira, Javaid Nadeem, Ullah Ibrar [et al.] // Applied Sciences (Switzerland). – 2020. – Vol. 10. – P. 1–21. DOI: https://doi.org/10.3390/app10093151

References

1. Eremenko I.I., Oliunina I.S. Use of Machine Learning Methods for Solving Problem of User Identifying by Keyboard Handwriting. *2019 International Russian Automation Conference (RusAutoCon)*, 2019. DOI: https://doi.org/10.1109/RUSAUTOCON.2019.8867767

2. Bikmullina I., Andreyanov N., Medvedev M. Stand for Development of Tasks of Detection and Recognition of Objects on Image. *2019 International Russian Automation Conference (RusAutoCon)*, 2019. DOI: https://doi.org/10.1109/RUSAUTOCON.2019.8867608

3. Zhou N.-N., Constantinides A. G., Huang G., Shaobai Z. Face recognition based on an improved center symmetric local binary pattern. *Neural Computing and Applications*, 2017, vol. 30, pp. 3791-3797. DOI: https://doi.org/10.1007/s00521-017-2963-2

4. Papakostas M., Iannakopoulos T. Speech-music discrimination using deep visual feature extractors. *Expert Systems with Applications*, 2018, vol. 114, pp. 334-344. DOI: https://doi.org/10.1016/j.eswa.2018.05.016

5. Robalinho M., Fernandes P. Software 2.0 for Scrap Metal Classification. *16th Int. Conf. on Informatics in Control, Automation and Robotics*, 2019, pp. 666-673. DOI: https://doi.org/10.5220/0007977506660673

6. Ojala T., Pietikainen M., Maenpaa T. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 2002, vol. 24, no. 7, pp. 971-987. DOI: https://doi.org/10.1109/TPAMI.2002.1017623

7. Kudugunta S., Ferrara E. Deep Neural Networks for Bot Detection. *Information Sciences*, 2018, vol. 467, pp. 312-322. DOI: https://doi.org/10.1016/j.ins.2018.08.019.

8. Dorofeev D., Khrestina M., Usubaliev T. et al. Application of Machine Analysis Algorithms to Automate Implementation of Tasks of Combating Criminal Money Laundering. *DTGS 2018: Digital Transformation and Global Society*, 2018, pp. 3757-385. DOI: https://doi.org/10.1007/978-3-030-02843-5_30

9. Sapozhnikova M. U., Nikonov A. V., Vulfin A. M. [et al.]. Anti-fraud system on the basis of data mining technologies. *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2017, pp. 2437-248. DOI: https://doi.org/10.1109/ISSPIT.2017.8388649

10. Abdelhamid D., Khaoula S., Atika O. Automatic Bank Fraud Detection Using Support Vector Machines. *Proceedings of the International conference on Computing Technology and Information Management*, 2014, pp. 10-17.

11. Gyamfi N. K., Abdulai J.-D. Bank Fraud Detection Using Support Vector Machine. *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2018, pp. 37-41. DOI: https://doi.org/10.1109/IEMCON.2018.8614994

12. T. Ryan Hoens, Nitesh V. Chawla. Imbalanced Datasets: From Sampling to Classifiers. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 2013, pp. 43-59. DOI: https://dx.doi.org/10.1002/2F9781118646106.ch3

13. Sun M., Yang R. An efficient secure k nearest neighbor classification protocol with high-dimensional features. *International Journal of Intelligent Systems*, 2020, vol. 35, pp. 17911-813. DOI: https://doi.org/10.1002/int.22272

14. Batchanaboyina M.R., Devarakonda N. Design and Evaluation of Outlier Detection Based on Semantic Condensed Nearest Neighbor. *Journal of Intelligent Systems,* 2019, vol. 29, pp. 1416-1424. DOI: https://doi.org/10.1515/jisys-2018-0476

15. Agustianto K., Destarianto P. Imbalance Data Handling using Neighborhood Cleaning Rule (NCL) Sampling Method for Precision Student Modeling. *Proceedings - 2019 International Conference on Computer Science,*

Information Technology, and Electrical Engineering, ICOMITEE 2019, 2019, pp. 86-89. DOI: https://doi.org/10.1109/ICOMITEE.2019.8921159

16. Rodolfo M. Pereira, Yandre M.G. Costa, Carlos N. Silla Jr. MLTL: A multi-label approach for the Tomek Link undersampling algorithm. *Neurocomputing*, 2020, vol. 383, pp. 95-105. DOI: https://doi.org/10.1016/j.neucom.2019.11.076

17. Wang S., Liu S., Zhang J. et al. A new method of diesel fuel brands identification: SMOTE oversampling combined with XGBoost ensemble learning. *Fuel*, 2020, vol. 282, pp. 1-9. DOI: https://doi.org/10.1016/j.fuel.2020.118848

18. Wang L., Zhao Z.Q., Luo Y.H. et al. Classifying 2-year recurrence in patients with dlbcl using clinical variables with imbalanced data and machine learning methods. *Computer Methods and Programs in Biomedicine*, 2020, vol. 196, pp. 1-14. DOI: https://doi.org/10.1016/j.cmpb.2020.105567

19. Liu C., Zhu L. A two-stage approach for predicting the remaining useful life of tools using bidirectional long short-term memory. *Measurement*, 2020, vol. 164, pp. 1-12. DOI: https://doi.org/10.1016/j.measurement.2020.108029

20. Taneja S., Suri B., Kothari C. Application of Balancing Techniques with Ensemble Approach for Credit Card Fraud Detection. *2019 International Conference on Computing, Power and Communication Technologies, GUCON 2019*, 2019, pp. 753-758. URL: https://ieeexplore.ieee.org/document/8940539

21. Hira Gul, Nadeem Javaid, Ibrar Ullah et al. Detection of Non-Technical Losses Using SOSTLinkand Bidirectional Gated Recurrent Unit to SecureSmart Meters. *Applied Sciences (Switzerland),* 2020, vol. 10, pp. 1-21. DOI: https://doi.org/10.3390/app10093151

Сведения об авторах

Марахтанов Алексей Георгиевич (Петрозаводск, Россия) – директор Центра искусственного интеллекта Петрозаводского государственного университета (185910, Республика Карелия, Петрозаводск, пр. Ленина, 33, e-mail: marahtanov@petrsu.ru).

Паренченков Евгений Олегович (Петрозаводск, Россия) – программист Центра искусственного интеллекта Петрозаводского государственного университета (185910, Республика Карелия, Петрозаводск, пр. Ленина, 33, e-mail: parenche@cs.karelia.ru).

Смирнов Николай Васильевич (Петрозаводск, Россия) – кандидат технических наук, доцент кафедры «Теория вероятностей и анализ данных» Петрозаводского государственного университета (185910, Республика Карелия, Петрозаводск, пр. Ленина, 33, e-mail: nvsmirnov87@gmail.com).

About the authors

Marakhtanov Aleksey Georgiyevich (Petrozavodsk, Russian Federation) is a Director Artificial intelligence center Petrozavodsk State University (185910, Petrozavodsk, 33, Lenina ave., e-mail: marahtanov@petrsu.ru).

Parenchenkov Evgeny Olegovich (Petrozavodsk, Russian Federation) is a Programmer Artificial intelligence center Petrozavodsk State University (185910, Petrozavodsk, 33, Lenina ave., e-mail: parenche@cs.karelia.ru).

Smirnov Nikolai Vasilyevich (Petrozavodsk, Russian Federation) is a Ph. D. in Technical Sciences, Associate Professor Department of Probability Theory and Data Analysis Petrozavodsk State University (185910, Petrozavodsk, 33, Lenina ave., e-mail: nvsmirnov87@gmail.com).

Получено 07.10.2020