Imbalanced Data Problem In Machine Learning

Marina Lukashevich
Information Management Systems Department
Belarusian State University
Minsk, Belarus
LukashevichMM@bsu.by

Sergei Bairak
Electronics Computers Department
Belarusian State University of Informatics and Radioelectronics
Minsk, Belarus
bairak@bsuir.by

Ilya Malochka
Discrete Mathematics and Algorithmics Department
Belarusian State University
Minsk, Belarus
ilja.molochko97@gmail.com

***Abstract*—Imbalanced data is a common challenge in real-world classification tasks. This study analyzes methods and algorithms for addressing class imbalance in binary classification models. Experimental evaluations are conducted on data balancing techniques, including oversampling the minority class and undersampling the majority class. The experiments cover both tabular data and image datasets. Based on the results, the impact of these methods on model performance is assessed, and practical recommendations for effective data balancing are provided.**

***Keywords*—imbalanced data, binary classification, sampling**

I. Introduction

In machine learning tasks, both in research and practical applications, we often encounter imbalanced datasets. In medical diagnosis, rare diseases like cancer often have far fewer positive cases than healthy samples, making early detection difficult. Fraud detection systems face severe imbalance, as fraudulent transactions may represent less than 1% of all transactions. Spam filtering also deals with skewed data, where spam emails are vastly outnumbered by legitimate ones. Similarly, in manufacturing quality control, defective products are typically rare compared to non-defective ones, requiring specialized techniques to identify anomalies. Class imbalance presents a significant challenge, as traditional classifiers tend to bias their predictions toward the majority class, which is often the least important class. This makes them unsuitable for handling imbalanced learning tasks [1]. A dataset is considered imbalanced when there is a substantial disproportion between the number of instances in different classes. The degree of imbalance can range from a slight bias to extreme cases where one minority-class instance corresponds to hundreds, thousands, or even millions of majority-class instances. To identify class imbalance, one can visually inspect a histogram showing the distribution of instances across classes or visualize the data on a graph when the feature space has low dimensionality. To quantify class imbalance, the imbalance ratio (IR) is commonly defined as the ratio between the number of instances in the minority class and the majority class.

$$imbalance\ ratio\ (IR) = \frac{len(Y_{minor})}{len(Y_{major})} \qquad (1)$$

It is possible to distinguish datasets with no imbalance, minor imbalance, major imbalance and huge imbalance. The Figure 1 shows the synthetic datasets for two classes with different imbalance levels and the estimated imbalance ratio [2]–[4].

It is particularly important to emphasize the need for a semantic approach to class imbalance, which focuses on preserving and utilizing meaningful (semantic) relationships in the data rather than simply mechanically balancing class distributions. The semantic approach is especially effective in tasks such as: Natural language processing (NLP), Computer vision (where semantic differences matter), Recommender systems, Medical diagnosis (where interpretability is crucial). In OSTIS technology, data is represented through semantic networks and ontological models, ensuring flexibility and interoperability in intelligent systems. Combining classical approaches to handling imbalanced data with the advantages of OSTIS technology could be a promising direction for further research [5].

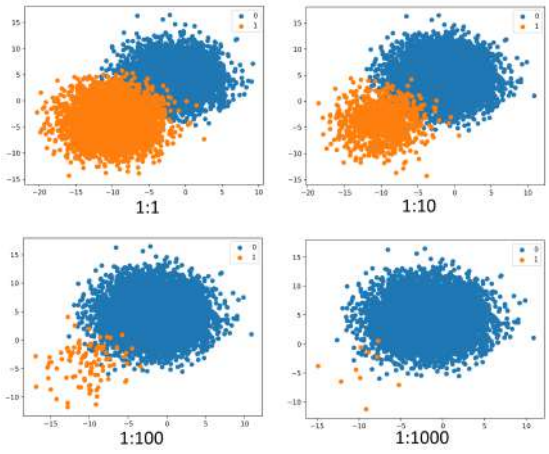

Figure 1. Scatter Plot of a Binary Classification Dataset with Different Class Distributions

The goal of this paper is to examine data balancing algorithms and evaluate their effectiveness for tabular data and images. Both classical machine learning algorithms and deep convolutional neural networks are explored in this work.

II. Related Work

According to [6], [7], methods for handling imbalanced data can be categorized as follows, Figure 2.

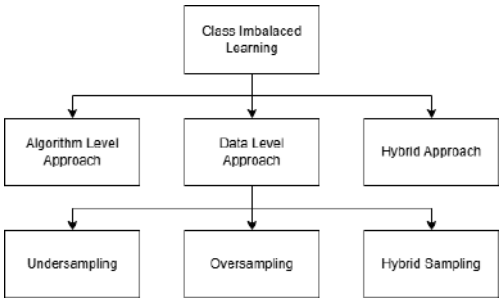

Figure 2. Approaches for class imbalansed learning

Algorithm-level approaches involve modifying existing machine learning algorithms to make them more suitable for imbalanced datasets by reducing their inherent bias toward the majority class. These modifications may include cost-sensitive learning techniques, where algorithms are adjusted to incorporate misclassification costs that penalize errors on minority class instances more heavily. Data-level approaches address class imbalance by directly modifying the dataset distribution prior to model training. These methods include oversampling techniques (e.g., SMOTE) that artificially increase minority class instances, and undersampling strategies that reduce majority class samples. Advanced variations combine both approaches or incorporate synthetic data generation through GANs. The hybrid approach combines the aforementioned methods in varying proportions to optimize classification performance. Different techniques can be integrated either across categories (e.g., combining data-level and algorithm-level methods) or within the same category (e.g., using multiple data-balancing techniques simultaneously). The literature presents numerous approaches for addressing class imbalance in machine learning. While various techniques exist, this study specifically examines oversampling and undersampling methods for data balancing.

The fundamental data-level approach for addressing class imbalance is known as data balancing. Among these techniques, increasing the minority class size (oversampling) is particularly valuable for small datasets, where reducing the majority class could compromise classification accuracy. The simplest method, Random Oversampling, duplicates randomly selected minority-class instances. These algorithms progressively refine synthetic data generation to improve model performance while mitigating overfitting risks. More advanced techniques include:

- SMOTE: Generates synthetic minority-class instances by interpolating between existing examples, avoiding mere duplication.
- Borderline-SMOTE: Focuses synthetic sample generation near class boundaries to enhance decision boundary learning.
- Borderline-SMOTE-SVM: Uses Support Vector Machines to identify critical boundary regions for synthetic data generation.
- ADASYN: Adaptively creates synthetic samples based on local density and class distribution, prioritizing difficult-to-learn areas.

Undersampling methods reduce the majority class size through three strategic approaches: (1) retaining informative instances, (2) removing redundant/noisy instances, or (3) hybrid combinations. Key algorithms include:

- Random Undersampling: Randomly eliminates majority-class instances (Note: Corrected erroneous "duplicating" description from original text)
- CNN (Condensed Nearest Neighbor): Preserves a subset that maintains original decision boundaries
- NearMiss Variants: NearMiss-1: Keeps majority instances with smallest average distance to 3 nearest minority instances. NearMiss-2: Retains majority instances farthest from minority clusters. NearMiss-3: Stores majority instances closest to each minority instance.
- Tomek Links: Removes overlapping majority-class instances in boundary pairs
- ENN (Edited Nearest Neighbors): Deletes misclassified majority instances based on 3-NN evaluation
- OSS (One-Sided Selection): Hybrid of Tomek Links (boundary cleaning) and CNN (redundancy removal)
- NCR (Neighborhood Cleaning Rule): Combines ENN (noise removal) and CNN (redundancy reduction)

These methods systematically address imbalance while preserving critical data patterns. Figure 3 shows the concept of resampling [6]–[10].

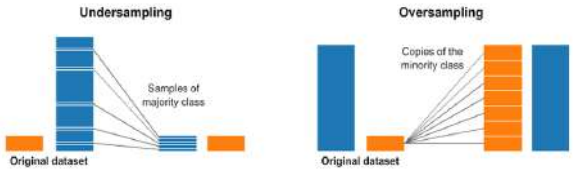

Figure 3. Resampling Methods

III. Experiments with Tabular Data

For our experiments, we evaluated three imbalanced binary classification datasets (Table 1). The class imbal-

ance ratios are visually demonstrated through histograms showing the distribution of instances between classes for: (1) abalone_19, (2) mammography, and (3) car_eval_34 datasets.

- Abalone_19: Predicts abalone age (marine mollusks), where the "19" class (oldest specimens) represents the rare minority, creating significant classification challenges.
- Mammography: A medical imaging dataset for breast cancer detection with extreme imbalance, where calcification clusters (minority class) require specialized detection approaches.
- Car_eval_34: Derived from vehicle evaluations, with merged "good"/"vgood" acceptability classes forming the minority, testing boundary-learning capabilities (Figure 4).

Table I
Dataset Summary

Dataset	Samples	Features	Imbalanced Ratio
abalone_19	4 177	10	0.0077
mammography	11 183	6	0.0238
car_eval_34	1 728	21	0.0841

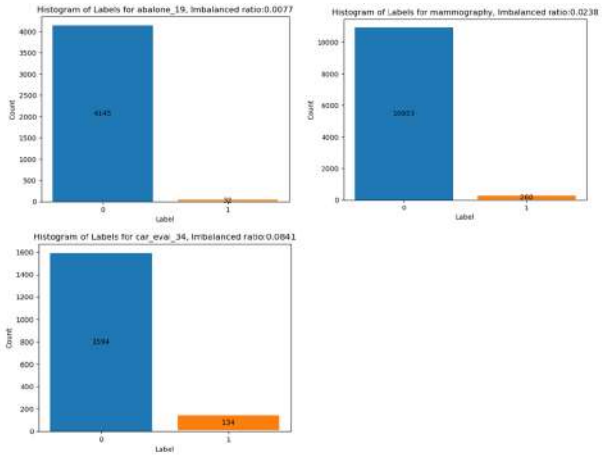

Figure 4. Abalone_19, Mammography and Car_eval_34 Datasets

Machine learning algorithms were selected for classifier construction, as detailed in Table II. These algorithms represent distinct approaches to classification - from linear separability to non-linear decision boundaries - while being computationally efficient for comparative analysis. Their performance metric (Accuracy) will be evaluated against the same imbalanced datasets to ensure consistent benchmarking conditions.

Our investigation systematically evaluated two balancing approaches: (1) undersampling methods, implemented through 13 distinct algorithms including random undersampling, Tomek Links, and neighborhood cleaning rule (Table III), and (2) oversampling techniques,

Table II
Classification Algorithms

Algorithm	Abbreviation
Logistic Regression	LR
SVC	SVC
KNN	KNN
DecisionTree	DT
Granient Boosting	GB
Random Forest	RF

comprising 7 algorithms such as SMOTE, Borderline-SMOTE, and ADASYN (Table IV). Each approach was rigorously tested under identical experimental conditions to ensure fair comparison of their effectiveness in handling class imbalance. The undersampling algorithms were selected to represent diverse strategies from random reduction to sophisticated instance selection, while the oversampling methods covered both basic interpolation and advanced adaptive synthesis techniques. This comprehensive framework allows for detailed analysis of how different balancing methodologies affect classifier performance across various imbalance scenarios.

All classifiers were implemented and evaluated using the specified algorithms under consistent experimental conditions. We employed 3-fold cross-validation to ensure robust performance estimation while maintaining computational efficiency.

Table III
Undersampling Balanced Algorithms

Balanced Approach	Designation
Random Undersampling	Under1
Condensed Nearest Neighbour	Under2
Tomek Links	Under3
One Sided Selection	Under4
Edited Nearest Neighbours – kind_sel=all	Under5
Edited Nearest Neighbours – kind_sel=mode	Under6
Repeated Edited Nearest Neighbours – kind_sel=all	Under7
Repeated Edited Nearest Neighbours – kind_sel=mode	Under8
All KNN – kind_sel=all	Under9
All KNN – kind_sel=mode	Under10
Neighbourhood Cleaning Rule	Under11
Instance Hardness Threshold	Under12

Table IV
Oversampling Balanced Algorithms

Balanced Approach	Designation
Random Oversampling	Over1
SMOTE	Over2
Borderline SMOTE-1	Over3
Borderline SMOTE-2	Over4
SVM-SMOTE	Over5
KMeans SMOTE	Over6
ADASYN	Over7

The results of experiments for abalone_19 dataset with undersampling and oversampling approaches are

presented in Table V and Table VI. The results of experiments for mammography dataset with undersampling and oversampling approaches are presented in Table VII and Table VIII. The results of experiments for car_eval_34 dataset with undersampling and oversampling approaches are presented in Table IX and Table X.

IV. Experiments With Image Data

The CIFAR-10 dataset was selected for experiments with imbalanced image classification tasks. As a standard benchmark in computer vision, it comprises 60,000 32×32 color images evenly distributed across 10 mutually exclusive categories. The dataset is split into 50,000 training images (5,000 per class) and 10,000 test images (1,000 per class), featuring objects from ten distinct classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck (Figure 5). This balanced original distribution was intentionally modified to create controlled imbalance conditions for our study.

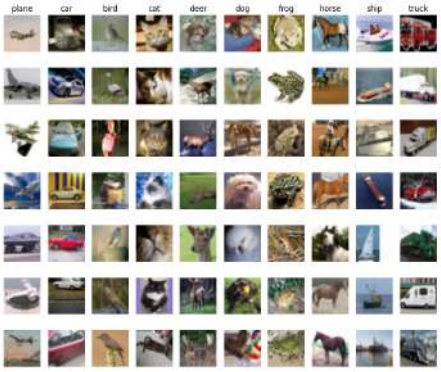

Figure 5. CIFAR-10 Dataset

Using the CIFAR-10 dataset, we constructed two specialized training sets for binary classification: (1) a balanced set containing 5,000 images per class, and (2) an imbalanced set with 5,000 majority-class images versus only 50 minority-class images (Figure 6). For evaluation, we created a separate test set comprising 2,000 images (1,000 per class) to ensure consistent benchmarking conditions across both balanced and imbalanced scenarios, (Figure 6).

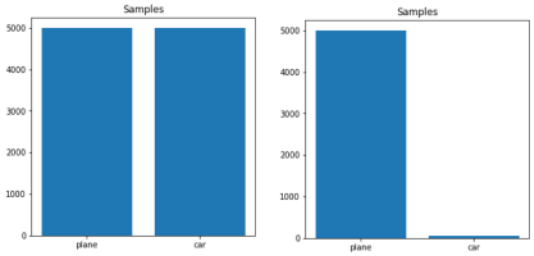

Figure 6. Balanced and Imbalanced CIFAR for 2 Classes

A convolutional neural network was chosen as a classifier. The neural network is a CNN with three convolutional blocks, each followed by ReLU activation and max-pooling. It processes 32x32x3 input images, gradually reducing spatial dimensions to 4x4x32 while increasing depth. A global average pooling layer flattens features into a 32-unit vector, followed by a dense layer and dropout for regularization. The final dense layer outputs 2 classes. The model has 15,458 trainable parameters and avoids overfitting through pooling and dropout. First, the classifier was trained on balanced and unbalanced datasets. For each of the two resulting models, an accuracy ('Accuracy') evaluation was performed on the testing dataset. The results clearly demonstrate that the imbalance in the training dataset caused the misclassification of the testing dataset of images, (Figure 7).

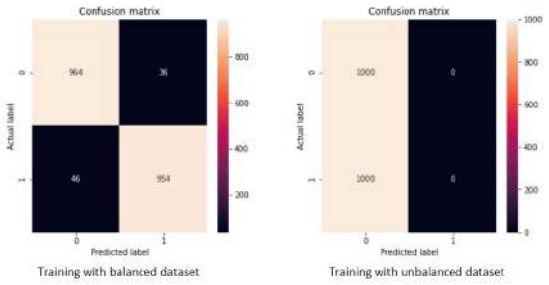

Figure 7. Testung Results for Training with Balanced and Imbalanced Data

We investigated three imbalance mitigation strategies: (1) cost-sensitive learning via class weights, (2) minority-class oversampling (duplication/SMOTE), and (3) majority-class undersampling. Each method was implemented through reproducible preprocessing pipelines while keeping other experimental parameters constant. This design enabled isolated measurement of how different balancing approaches affect model performance on our imbalanced image datasets. The evaluation was conducted using testing dataset comprising 2,000 images (1,000 per class), with detailed performance metric (Accuracy) reported in Table XI. The confusion matrices for the three approaches are shown in Figure 8.

V. Discussion

Oversampling increases instances in the minority class (e.g., SMOTE, ADASYN) to improve rare class representation, while undersampling reduces majority class samples (e.g., random removal, Tomek Links) to balance distributions. Oversampling preserves data but risks overfitting and longer training times, whereas undersampling speeds up training but may discard valuable patterns. Oversampling suits small datasets (e.g., medical data), while undersampling works better for large datasets (e.g., credit scoring). Key trade-offs include synthetic data

Table V
Undersampling For Abalone_19 Dataset

	Baseline	Under1	Under2	Under3	Under4	Under5	Under6	Under7	Under8	Under9	Under10	Under11	Under12
LR	0,735	0,656	0,406	0,735	0,721	0,736	0,734	0,744	0,734	0,756	0,737	0,753	0,803
SVC	0,768	0,706	**0,62**	0,763	0,736	0,754	0,712	0,736	0,712	0,748	0,743	0,767	0,853
KNN	0,547	0,662	0,322	0,547	0,546	0,615	0,546	0,664	0,546	0,647	0,563	0,613	0,693
DT	0,527	0,689	0,457	0,528	0,527	0,512	0,528	0,525	0,528	0,51	0,527	0,511	0,59
GB	**0,798**	**0,759**	0,369	0,779	**0,776**	**0,806**	**0,785**	**0,811**	**0,785**	**0,784**	**0,773**	**0,808**	0,877
RF	0,753	0,758	0,269	**0,785**	0,737	0,737	0,69	0,798	0,69	0,77	0,752	0,766	**0,927**

Table VI
Oversampling For Abalone_19 Dataset

	Baseline	Over1	Over2	Over3	Over4	Over5	Over6	Over7
LR	0,735	0,872	0,885	0,933	0,94	0,933	0,945	0,873
SVC	0,768	0,907	0,919	0,984	0,994	0,983	0,955	0,902
KNN	0,547	0,995	0,975	0,992	0,995	0,991	0,991	0,96
DT	0,527	0,993	0,959	0,988	0,991	0,985	0,991	0,894
GB	**0,798**	0,999	0,989	0,997	**0,997**	**0,996**	0,997	0,975
RF	0,753	**1**	**0,999**	**0,998**	**0,997**	**0,996**	**0,998**	**0,987**

Table VII
Undersampling For Mammography Dataset

	Baseline	Under1	Under2	Under3	Under4	Under5	Under6	Under7	Under8	Under9	Under10	Under11	Under12
LR	0,921	0,938	0,814	0,922	0,922	0,924	0,923	0,926	0,923	0,925	0,922	0,923	0,941
SVC	0,881	0,949	0,84	0,893	0,893	0,918	0,9	0,928	0,901	0,925	0,907	0,908	**0,978**
KNN	0,9	0,942	0,793	0,901	0,901	0,915	0,901	0,917	0,904	0,917	0,902	0,916	0,945
DT	0,748	0,857	0,715	0,762	0,762	0,795	0,758	0,818	0,771	0,813	0,786	0,793	0,904
GB	0,94	0,944	**0,846**	0,944	0,944	0,951	**0,948**	**0,952**	**0,952**	**0,953**	**0,951**	0,952	0,972
RF	**0,944**	**0,957**	0,836	**0,946**	**0,942**	**0,954**	**0,948**	0,852	0,947	0,951	0,95	**0,955**	**0,978**

Table VIII
Oversampling For Mammography Dataset

	Baseline	Over1	Over2	Over3	Over4	Over5	Over6	Over7
LR	0,921	0,924	0,931	0,955	0,954	0,965	0,994	0,862
SVC	0,881	0,965	0,969	0,922	0,991	0,99	0,995	0,93
KNN	0,9	0,981	0,977	0,993	0,993	0,993	0,994	0,933
DT	0,748	0,99	0,968	0,987	0,983	0,951	0,989	0,904
GB	0,94	0,985	**0,985**	0,994	0,995	0,994	0,997	0,951
RF	**0,944**	**0,994**	0,992	**0,999**	**0,999**	**0,998**	**0,997**	**0,968**

Table IX
Undersampling For Car_eval_34 Dataset

	Baseline	Under1	Under2	Under3	Under4	Under5	Under6	Under7	Under8	Under9	Under10	Under11	Under12
LR	**0,958**	0,947	**0,893**	**0,958**	**0,954**	**0,958**	**0,958**	**0,958**	**0,958**	**0,958**	**0,958**	**0,998**	**0,999**
SVC	0,947	0,942	0,838	0,947	0,943	0,947	0,947	0,947	0,947	0,947	0,947	0,995	**0,999**
KNN	0,657	0,865	0,572	0,657	0,608	0,61	0,61	0,61	0,61	0,61	0,61	0,907	0,994
DT	0,559	0,802	0,47	0,559	0,506	0,559	0,559	0,559	0,559	0,559	0,559	0,532	0,669
GB	0,835	0,973	0,842	0,835	0,863	0,834	0,834	0,834	0,834	0,834	0,834	0,955	0,833
RF	0,899	**0,948**	0,615	0,899	0,891	0,892	0,892	0,892	0,892	0,892	0,892	0,985	0,992

Table X
Oversampling For Car_eval_34 Dataset

	Baseline	Over1	Over2	Over3	Over4	Over5	Over6	Over7
LR	**0,958**	0,954	0,952	0,948	0,905	0,98	0,938	0,975
SVC	0,947	**1**	**1**	**1**	0,994	0,983	**0,992**	**1**
KNN	0,657	0,996	0,991	0,986	0,938	0,982	0,917	0,975
DT	0,559	0,798	0,79	0,801	0,772	0,796	0,698	0,729
GB	0,835	0,913	0,925	0,94	0,893	0,986	0,903	0,97
RF	0,899	**1**	0,999	0,999	0,995	**1**	0,981	0,984

Table XI
Evaluation Results for Image Classification on Imbalanced Dataset

	Accuracy
Unbalanced	0,5
Class Weights	0,8339
Undersampling	0,8305
Oversampling	0,8230

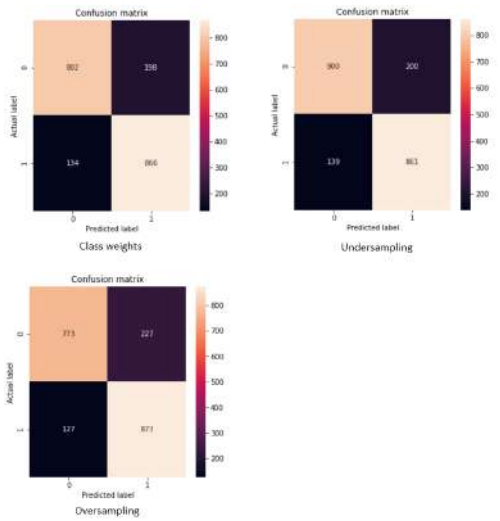

Figure 8. Training Strategies for Imbalanced Datasets: Class Weighting vs. Resampling Methods

realism (oversampling) and information loss (undersampling). The choice depends on data size, computational resources, and domain requirements. In most cases, classifiers trained on the balanced dataset performed as well as on the original dataset. In all cases it is possible to improve classification accuracy by combining a particular classifier model and data balancing technique.

VI. Conclusion

This paper investigates the effects of class imbalance on machine learning models across diverse data types. We systematically analyze principal approaches for handling imbalanced datasets, with particular focus on data-level techniques including oversampling and undersampling methods. The study presents two experimental frameworks. Through comprehensive comparative analysis, we demonstrate the quantitative impact of data balancing techniques on model performance metrics. Our results provide actionable insights for selecting appropriate imbalance mitigation strategies based on data characteristics and model architecture.

References

[1] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS Int. Trans. Comput. Sci. Eng.*, vol. 30, no. 1, pp. 25-36, 2006.

[2] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263-1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[3] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, p. 27, 2019, doi: 10.1186/s40537-019-0192-5.

[4] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 1-50, 2016, doi: 10.1145/2907070.

[5] V. V. Golenkov et al., "Intelligent computer systems of new generation and complex technology of their development, application and modernization," Doklady BGUIR, vol. 22, no. 2, pp. 70–79, 2024.

[6] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221-232, 2016, doi: 10.1007/s13748-016-0094-0.

[7] J. Brownlee, *Imbalanced Classification With Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery, 2020.

[8] R. Agarwal, "Sampling," 2020. [Online]. Available: https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets

[9] Y. Zhang et al., "Imbalanced learning: Foundations, algorithms, and applications," *arXiv*, preprint arXiv:2301.10319, 2023.

[10] J. M. Johnson et al., "Imbalanced learning in medical imaging: A comprehensive review," *Artif. Intell. Med.*, vol. 135, p. 102476, 2023, doi: 10.1016/j.artmed.2022.102476.

ПРОБЛЕМА НЕСБАЛАНСИРОВАННЫХ ДАННЫХ В МАШИННОМ ОБУЧЕНИИ

Лукашевич М. М., Байрак С. А., Молочко И. П.

Несбалансированные данные являются проблемой для реальных задач классификации. В работе анализируются методы и алгоритмы работы с несбалансированными данными при построении моделей машинного обучения для задач бинарной классификации. Проведены экспериментальные исследования алгоритмов балансировки данных, основанных на увеличении меньшего класса и уменьшении большего класса. Представлены эксперименты для табличных данных и изображений. По результатам экспериментов оценено влияние исследуемых методов и алгоритмов на качество моделей, получаемых в результате обучения. Даны рекомендации по применению методов балансировки данных.

Received 27.03.2025