Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering
Elektronik/Online ISSN: 2547-958X

RESEARCH ARTICLE / ARAŞTIRMA MAKALESI

An Improved Version of Edited Nearest Neighbor Undersampling Method Based on the kNN Approach

kNN Yaklaşımını Temel Alan Düzenlenmiş En Yakın Komşu Veri İndirgeme Yönteminin İyileştirilmiş Bir Versiyonu

Alican Doğan

Bandırma Onyedi Eylül Üniversitesi Uygulamalı Bilimler Fakültesi Yönetim Bilişim Sistemleri Bölümü, Balıkesir, TÜRKİYE
Corresponding Author / Sorumlu Yazar *: alicandogan@bandirma.edu.tr

Abstract

In the field of machine learning, handling imbalanced datasets remains a critical challenge, often addressed through various sampling techniques. Among these techniques, the Edited Nearest Neighbor (ENN) undersampling method is widely recognized for its ability to enhance classifier performance by reducing class imbalance. However, the traditional ENN method has limitations, such as the removal of potentially informative instances and suboptimal performance in complex datasets. This paper presents an improved version of the ENN undersampling method, leveraging the k-Nearest Neighbors (kNN) approach to refine the selection process for instance removal. The proposed method improves upon the traditional ENN by incorporating a more sophisticated neighbor evaluation criterion based on the k-NN algorithm, which better preserves informative instances while effectively reducing noise. Through extensive experiments on multiple benchmark datasets, we demonstrate that our improved ENN method achieves superior performance in terms of classification accuracy, F1-score, and AUC, compared to the traditional ENN and other state-of-the-art undersampling techniques. The results indicate that the improved ENN method not only mitigates the class imbalance problem more effectively but also maintains a higher level of data integrity, thereby enhancingthe robustness and reliability of machine learning models. This advancement provides a valuable tool for practitioners dealing with imbalanced datasets, contributing to the development of more accurate and efficient predictive models.

Keywords: *ENN, CxKNN, Undersampling, Classification*

Öz

Makine öğrenimi alanında, dengesiz veri kümeleriyle başa çıkmak önemli bir zorluk olmaya devam etmekte olup, genellikle çeşitli örnekleme teknikleriyle ele alınmaktadır. Bu teknikler arasında, Düzeltilmiş En Yakın Komşu (ENN) alt örnekleme yöntemi, sınıflandırıcı performansını artırma ve sınıf dengesizliğini azaltma yeteneğiyle geniş çapta tanınmaktadır. Ancak geleneksel ENN yönteminin, potansiyel olarak bilgilendirici örneklerin kaldırılması ve karmaşık veri kümelerinde yetersiz performans gibi sınırlamaları vardır. Bu makale, k-Nearest Neighbors (k-NN) yaklaşımını kullanarak örnek kaldırma sürecini iyileştiren ENN alt örnekleme yönteminin geliştirilmiş bir versiyonunu sunmaktadır. Önerilen yöntem, geleneksel ENN'yi, bilgilendirici örnekleri daha iyi korurken aynı zamanda gürültüyü etkili bir şekilde azaltan k-NN algoritmasına dayalı daha gelişmiş bir komşu değerlendirme kriteri ekleyerek iyileştirmektedir. Birçok benchmark veri kümesinde yapılan kapsamlı deneylerle, geliştirilmiş ENN yöntemimizin sınıflandırma doğruluğu, F1 skoru ve AUC açısından geleneksel ENN ve diğer en son alt örnekleme tekniklerine kıyasla üstün performans sergilediğini gösteriyoruz. Sonuçlar, geliştirilmiş ENN yönteminin yalnızca sınıf dengesizliği sorununu daha etkili bir şekilde hafifletmekle kalmayıp, aynı zamanda veri bütünlüğünü daha yüksek seviyede koruduğunu ve böylece makine öğrenimi modellerinin dayanıklılığını ve güvenilirliğini artırdığını göstermektedir. Bu ilerleme, dengesiz veri kümeleriyle çalışan uygulayıcılar için değerli bir araç sunarak, daha doğru ve verimli tahmin modellerinin geliştirilmesine katkıda bulunmaktadır.

Anahtar Kelimeler: *ENN, CxKNN, Veri İndirgeme, Sınıflandırma*

1. Introduction

In the realm of machine learning, the challenge of imbalanced datasets—where one class significantly outnumbers the others—is a pervasive issue that often diminishes the performance of predictive models. Imbalanced datasets can lead to favoring the majority class, resulting in poor classification performance for the minority class, which is often the class of greater interest [1]. Addressing this issue requires effective sampling techniques to ensure balanced training data, thereby enhancing the overall performance and reliability of machine learning models.

One widely recognized approach to deal with class imbalance is undersampling, which involves reducing the number of instances in the majority class. Among various undersampling techniques, the Edited Nearest Neighbor (ENN) method stands out for its ability to remove noisy and potentially misclassified instances. Edite Thus it improves classifier performance. However, the traditional ENN method is not without its limitations [2]. It can sometimes eliminate informative instances, leading to a loss of valuable data and suboptimal model performance, especially in complex datasets with overlapping class distributions.

DOI: https://doi.org/10.21205/deufmd.2025278105

Geliş Tarihi / Received: 02.10.2024
Kabul Tarihi / Accepted: 16.11.2024

There is extensive literature on undersampling techniques. Peiqi et al. utilized both nearest neighbors and density-based clustering to reduce the size of the majority class in addressing the class imbalance problem [3]. Similarly, Yun et al. applied the Fuzzy C-Means method and Multilayer Perceptron to improve the effectiveness of undersampling in their research [4]. Additionally, Daye et al. investigated the impact of various undersampling methods on the observation of coronal oscillations [5].

To address the limitations of traditional undersampling methods, this paper introduces an enhanced version of the ENN (Edited Nearest Neighbor) undersampling method. This approach is based on an adaptation of the k-Nearest Neighbors (kNN) algorithm, specifically the C x k – Nearest Neighborhood method proposed by Nasibov [6], which takes into account the classes of neighboring instances. The C x k -NN algorithm, known for its simplicity and effectiveness in various classification tasks, provides a robust framework for evaluating which instances to retain or remove. By incorporating a more sophisticated neighbor evaluation criterion, our improved ENN method aims to retain informative instances while effectively reducing noise and mitigating class imbalance.

This study presents a comprehensive methodology for the improved ENN method and demonstrates its effectiveness through extensive experiments on multiple benchmark datasets. The results show significant improvements in classification accuracy compared to the traditional ENN method and other state-of-the-art undersampling techniques. These findings suggest that our improved ENN method not only addresses the class imbalance problem more effectively but also enhances the robustness and reliability of machine learning models.

The remainder of this paper is structured as follows: Section 2 reviews related work in the field of undersampling techniques, particularly focusing on the ENN method and its variants. Section 3 describes the proposed methodology called CKNN-ENN, detailing the integration of the C x k-NN approach with the traditional ENN method. Section 4 outlines the experimental setup, including the datasets used and the evaluation metrics employed. Section 5 presents the results and discusses their implications. Finally, Section 6 concludes the paper and suggests directions for future research.

2. Related Work

The challenge of imbalanced datasets has spurred extensive research into various sampling techniques to improve classifier performance [3]. Among these techniques, undersampling methods, which reduce the number of instances in the majority class, have received significant attention. This section reviews key developments in undersampling with a focus on the Edited Nearest Neighbor (ENN) method and its variations.

Traditional undersampling methods, such as random undersampling, simply remove a random subset of the majority class instances to balance the dataset [4]. While effective in some cases, this approach can lead to the loss of important information and may not adequately address the issue of overlapping class distributions. More sophisticated techniques, like Tomek Links and Condensed Nearest Neighbor (CNN), aim to improve upon random undersampling by selectively removing instances that contribute to class overlap or noise. However, these methods still face challenges in maintaining the integrity of the dataset and preserving informative instances [7].

The Edited Nearest Neighbor (ENN) method, introduced by Wilson in 1972, is a more refined undersampling technique that aims to enhance data quality by removing misclassified instances [8]. ENN uses the k-Nearest Neighbors (kNN) classifier to identify and eliminate instances that differ from most neighbors, thus reducing noise and potentially improving classifier performance. Despite its advantages, the traditional ENN method has limitations, particularly in its tendency to remove instances that could be informative, leading to a loss of valuable data and suboptimal model performance [9].

Several variants of the ENN method have been proposed to address its limitations. The Repeated Edited Nearest Neighbors (RENN) method applies the ENN process iteratively until no more instances can be removed, aiming to further cleanse the dataset. However, this iterative approach can be computationally intensive and may still result in losing important instances [10]. Another variant, the All k-Nearest Neighbors (All-kNN), removes instances based on their classification by multiple kNN classifiers with varying values of k, which helps to stabilize the removal process but can increase computational complexity.

The kNN algorithm itself has been widely used in various sampling techniques due to its simplicity and effectiveness in classification tasks. It is particularly useful in evaluating the local structure of the data, making it a natural choice for enhancing undersampling methods [11]. For instance, the Neighborhood Cleaning Rule (NCR) combines kNN with undersampling by removing majority class instances that are misclassified by their neighbors, thus aiming to reduce class overlap and noise.

Undersampling techniques are widely used in many application fields. Maneerat et al. generated a network-based intrusion detection system to block malicious attacks on the internet using an undersampling approach [1]. Apart from this study, Yang et al. analyzed the impact of the random undersampling method on the classification accuracy of observational health data [11]. Furthermore, Kubicka et al. utilized Cartesian Undersampling in order to diagnose specific states of the head and neck region with their MRI results [12].

Hybrid approaches that combine undersampling with other techniques, such as oversampling or ensemble methods, have also been explored. The SMOTEENN method, for example, combines Synthetic Minority Over-sampling Technique (SMOTE) with ENN to balance the dataset while simultaneously cleaning it. These hybrid methods often achieve better performance by leveraging the strengths of both undersampling and oversampling techniques.

Building on the existing body of work, our study proposes an improved version of the ENN undersampling method that integrates a more sophisticated neighbor evaluation criterion based on the kNN approach. By refining the instance selection process, our method aims to preserve informative instances while effectively reducing noise and mitigating class imbalance. This paper provides a detailed methodology for the improved ENN method and demonstrates its efficacy through extensive experiments on multiple benchmark datasets, showing significant improvements in classification performance compared to the traditional ENN method.

Table 1. Some recent research studies that use undersampling methods to reduce negative effects of data imbalance.

Authors	Method	Aim
Zuo et al.	FCM-GRNN	Identification of lysine Crotonylation Sites
Wainer	Prototype Generator	Various imbalanced datasets
Bach	KNN_NEAR	Glass Dataset
Kim et al.	CNF Model	Personal Health
Hancock et al.	Random Undersampling	Medicare Data
Yang et al.	Random Undersampling	Health Data
Maneerat et al.	Clustering-based Undersampling	Artificial Bee Colony
Kubicka et al.	Cartesian Undersampling	MRI Data

This section provides a comprehensive overview of the existing literature on undersampling techniques, setting the stage for the introduction of our improved ENN method.

3. Materials and Method

The Edited Nearest Neighbor (ENN) method is a data preprocessing technique mainly used in machine learning to address class imbalance, where certain classes are underrepresented. The primary goal of ENN is to improve the quality of a dataset by removing instances (data points) that may lead to classification errors.

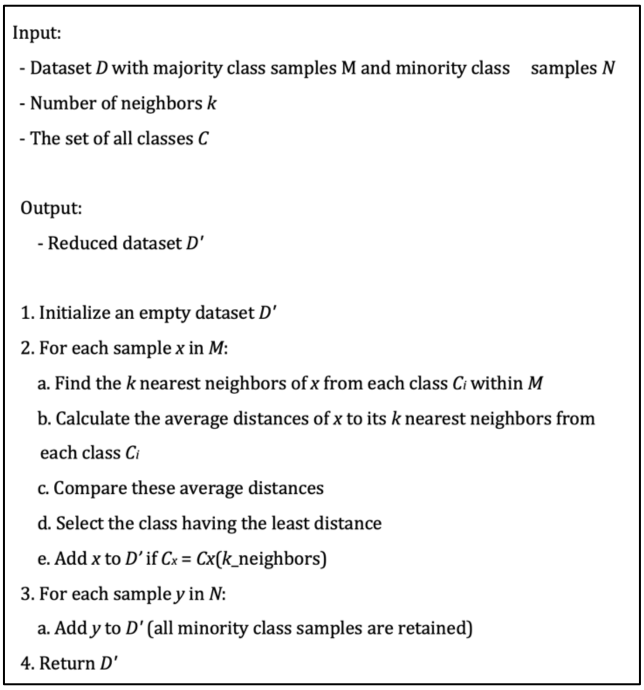
Despite extensive research on learning from imbalanced data, numerous challenges remain, keeping this problem both relevant and intriguing.

Many researchers highlight that the imbalance ratio is not the only source of learning difficulties. Even with a significant imbalance, good classification rates can be achieved if all classes are well-represented and come from non-overlapping distributions. Performance deterioration may be caused by the presence of difficult examples, particularly within the minority class. As a result, recent trends focus not only on the disproportion in class size but also on other inherent data challenges, such as the presence of noisy examples and class overlapping regions.

Our method has been created with the philosophy of conventional random undersampling techniques. However, instead of selecting objects to remove from the entire set of majority class objects, our method focuses only on a narrow group of k-nearest neighbors of each majority class sample. This approach prevents the removal of too many observations from a specific area, thereby reducing the risk of losing important information.

We presented an algorithm aimed at identifying and thinning clusters of majority class examples. By removing observations from high-density areas, this approach minimizes information loss compared to removing individual examples or those from lower-density areas.

The solution proposed in this paper combines the benefits of the previous two methods by removing the nearest neighbors for each majority object. The main idea is to ensure an even elimination of majority class samples while concentrating on the nearest objects.



```
Input:
 - Dataset $D$ with majority class samples M and minority class    samples $N$
 - Number of neighbors $k$
 - The set of all classes $C$

Output:
   - Reduced dataset $D'$

1. Initialize an empty dataset $D'$
2. For each sample $x$ in $M$:
   a. Find the $k$ nearest neighbors of $x$ from each class $C_i$ within $M$
   b. Calculate the average distances of $x$ to its $k$ nearest neighbors from
      each class $C_i$
   c. Compare these average distances
   d. Select the class having the least distance
   e. Add $x$ to $D'$ if $C_x = Cx(k\_neighbors)$
3. For each sample $y$ in $N$:
   a. Add $y$ to $D'$ (all minority class samples are retained)
4. Return $D'$
```

Figure 1. Pseudocode of the algorithm

For each object in the majority class M, the k nearest neighbors from each class including both majority and minority classes are identified. The distances between these k nearest neighbors and the object from the majority class are calculated. In total, k x C that is the number of neighbors times the number of classes distances are calculated. Next, the average distances between the instance and each class are obtained. That is to say, for each class, the mean of distances between the object and the k nearest neighbors are calculated. Totally, C (the number of classes) average distances are generated. Finally, these average distances are compared and the class having the smallest average distance is selected. If this class, the nearest class, is different from the real class of the instance, then this instance is extracted from the dataset.

The number of instances in the majority class decreases as a result of the undersampling process if necessary, according to the algorithm, but all instances from the minority class or classes are retained. Instead of comparing the class of the element with its k nearest neighbors like ENN method, the algorithm focuses on the average distance between the element and k nearest neighbors from each class. Overall, reduced dataset is returned as the output. Thus, the unbalanced dataset becomes balanced.

The nearest neighbors count *(k)* is an adjustable parameter that influences the algorithm's performance. The selection of the number of nearest neighbor k value does not depend on a rule. It is arbitrary. It can affect the performance of the method in terms of the experimented dataset as in the case of ENN method. Therefore, selecting appropriate values for this input parameter is crucial.

4. Experimental Results

The research was conducted using C# programming language and Microsoft Visual Studio software environments. We present the averaged results from five independent 10-fold stratified cross-validation experiments. Unbalanced datasets having high number of data in the majority class accordingly were chosen for this task. Also, all of the datasets are suitable for binary classification. They include categorical target attributes.

In the initial phase of experiments, the proposed algorithm was evaluated using artificial datasets. This approach allows for controlled alterations to the data's characteristics, facilitating an analysis of the impact of individual features on the algorithm's effectiveness. Two-dimensional datasets with two classes of varying shapes and distributions were generated. The total number of samples, the imbalance ratio, and the distance between classes were adjusted. Given that artificial data might not capture all issues present in real-world data, the experiments were also conducted on real datasets.

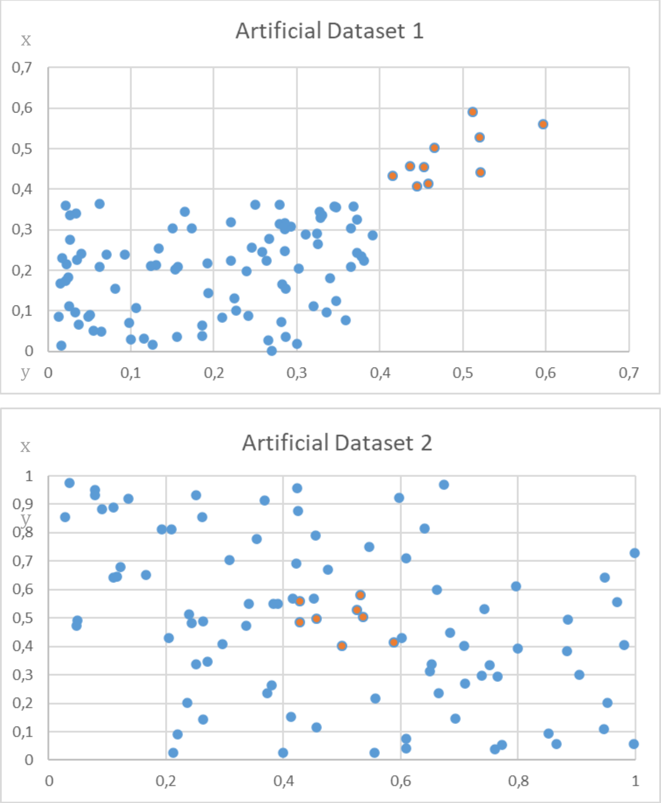


Figure 2. Data Distribution of Artificial Generated Datasets

The goal of the experiments was to evaluate the CxKNN-ENN method against various balancing techniques to determine its effectiveness in addressing class imbalance issues in practical applications. For comparison, the original random undersampling method and one heuristic approach based on the concept of nearest neighbors (ENN) was utilized.

All the aforementioned undersampling methods were applied to five classifiers: Ridge Classifier, Logistic Regression, Stochastic Gradient Descent (SGD), an optimization algorithm used to minimize a loss function in machine learning and deep learning models Perceptron, and Passive Aggressive Classifier (PA), a type of online learning algorithm used primarily for classification tasks where data arrives in a sequential manner. Additionally, for each dataset and tested classifier, the values of the undersampling parameters k (the number of nearest neighbors) and/or P (percentage of undersampling) were determined experimentally to achieve the best classification quality.

In the initial stage of the experiments, artificial datasets were utilized to demonstrate the performance of the new methods

First artificial dataset contains 92 samples in the majority class and 10 in the minority class, whereas the second artificial dataset includes 92 samples in the majority class and 8 in the minority class but they differ in the degree of separation between the classes.

The accuracy of the classification largely depends on the complexity of the data distribution. The first dataset, named Artificial Dataset1, is characterized by a greater distance between classes. In contrast, the classes in the Artificial Dataset 2 set are less separable, making the shape of the decision boundary more difficult to recognize.

The classification accuracy values in percentage (%) of the experimental results obtained for the original data (i.e., without balancing) and after applying our proposed balancing method are presented in Tables 2 and 3.

Table 2. Classification performance statistics for Artificial Dataset1.

	ENN-Accuracy (%)	CkNN-ENN Accuracy(%)
Ridge	83.76	90.52
Logistic	87.42	87.42
SGD	83.01	84.52
Perceptron	86.89	90.63
PA	89.41	89.41

Table 3. Classification performance statistics for Artificial Dataset2.

	ENN-Accuracy(%)	CkNN-ENN Accuracy(%)
Ridge	81.5	84.04
Logistic	83.79	89.65
SGD	89.08	93.72
Perceptron	85.73	96.28
PA	81.2	92.91

Despite the relatively high imbalance ratio, the results for the original Artificial1 set were very good for most classifiers. The only exception was the SGD, which performed poorly, assigning many objects to the majority class. For the other classifiers, the results were generally satisfactory in many areas, even without any balancing procedures.

For the Artificial2 dataset, the results are slightly worse, particularly for the Ridge and PA classifiers. However, using undersampling with the proposed method enhanced the outcomes.

In the initial part of the experiments, artificial data with varying imbalance ratios or class separability were analyzed. However, real-world datasets often face multiple issues simultaneously, such as rare sub-concepts, overlapping, noisy examples, and within-class imbalance, among others.

In the second part of the experiments, 5 real-world datasets from various domains, sizes, and class distributions were used. The tests were conducted on benchmark datasets obtained from the UCI (University of California, Irvine) [26] machine learning repository. The characteristics of these datasets, including the number of examples (#Ex.), number of attributes (#Atts.), and the imbalance ratio (IR) – which is the ratio of negative to positive instances – are detailed in Table 4.

Table 4. Classification performance statistics for Artificial Dataset2.

Name	# of Instances	# of Attributes	Imbalance Ratio
Adult	48842	15	5.22
Balloons	16	5	1.28
Breast Cancer	286	10	11.59
Chess (King-Rook vs. King-Pawn)	3196	36	1.09
Japanese Credit Screening	125	16	1.65

The classification tasks were conducted using five classifiers on five datasets, both in their original forms and after undersampling with ENN and CKNN-ENN methods.

The experimental results for the accuracy metric are summarized in Fig. 2. It is evident that the proposed CKNN-ENN solution outperforms the other tested methods in many instances. However, the results are influenced by various factors.

Let's examine this issue in more detail for these datasets: Balloons, which has the fewest minority class instances, and Breast Cancer, which has the highest imbalance ratio (IR). It should be noted that the results obtained for the original Adult dataset were poor. For the PA and SGD classifiers, nearly all instances were classified into the majority class (Fig. 2). The Ridge classifier performed slightly better on this dataset , where it reached around 57%. The proposed CKNN-ENN algorithm yielded the best results across all the metrics analyzed. While the improvement in classification quality compared to the original data and data undersampled with ENN method was significant, the results remain less than fully satisfactory. This is largely due to the complex structure of the dataset, particularly the issue of class overlap.

For the Balloons dataset, which very few minority class instances, one might expect an 'absolute rarity' problem—meaning there aren't enough minority-class examples to accurately learn the decision boundaries. However, the results show that this issue did not arise here. Minority class instances were identified fairly well, even in the original dataset. None of the heuristic undersampling methods led to a significant improvement in classification quality. The best results were achieved with CKNN-ENN undersampling. For instance, the Average Accuracy increased from 72% (with the original data) to 80%.

For the Breast Cancer dataset, which has the highest imbalance ratio among the analyzed sets (resulting in an IR of 11.59), the best results were achieved using the proposed CKNN-ENN method.

Several factors influence the selection of a sampling method, making it unrealistic to expect a single method to perform best for every dataset and classifier. Instead, it's more practical to identify the situations and conditions where a method is most likely to deliver good results.

The method introduced in this paper is particularly effective for datasets where the structure favors the even removal of majority class objects. This approach reduces overlap with minority class objects while minimizing the loss of important information from the majority class.

Table 5. Classification accuracies in percentage (%) of the selected classifiers for the selected datasets before implementing undersampling and after implementing undersampling with ENN and CKNN-ENN

Dataset	Classifier	Without Undersampling(%)	ENN(%)	CKNN-ENN(%)
Adult	Ridge	57	0.66	0.68
	LogisticRegression	52	0.56	0.61
	SGD	27	0.46	0.49
	Perceptron	51	0.54	0.62
	PassiveAggresive	15	0.38	0.44
Balloons	Ridge	100	100	100
	LogisticRegression	100	100	100
	SGD	100	100	100
	Perceptron	100	100	100
	PassiveAggresive	100	100	100
Breast Cancer	Ridge	62	72	76
	LogisticRegression	62	68	74
	SGD	62	62	62
	Perceptron	68	75	77
	PassiveAggresive	55	62	69
Chess (King-Rook vs. King-Pawn)	Ridge	93	93	95
	LogisticRegression	95	95	95
	SGD	89	91	93
	Perceptron	90	94	96
	PassiveAggresive	93	92	95
Japanese Credit Screening	Ridge	81	82	88
	LogisticRegression	84	81	85
	SGD	55	63	66
	Perceptron	60	65	72
	PassiveAggresive	72	71	74
Average		72	77	80

5. Conclusion and Future Work

The ongoing interest in class imbalance stems from its widespread occurrence and the increasing ability to collect vast amounts of data. Many areas of daily life, such as medical diagnosis, credit scoring, and fraud detection, are represented by datasets with skewed distributions. As a result, it's crucial to ensure that decision support systems perform effectively, especially for minority class instances.

This paper introduces a new undersampling method. Tests conducted on a large set of both artificial and real imbalanced datasets often showed that applying this method significantly improved classification performance.

Choosing the right sampling method depends on various factors. On one hand, certain classification algorithms tend to favor specific sampling techniques. On the other hand, challenges such as rare cases, noise, and overlapping distributions require a detailed analysis of class structure to select the most appropriate balancing method.

Overall, there is no clear consensus on which approach yields the best results. While some guidelines exist, experimental evaluation of different methods is generally needed in practice. Our method is particularly effective for datasets where even thinning of the majority class minimizes the loss of valuable information.

Ethics committee approval and conflict of interest statement

Ethics committee approval is not required for the prepared article. There is no conflict of interest with any individual or institution in the prepared article.

References

[1] Maneerat, T., Iam-On, N., Boongoen, T., Kirimasthong, K., Naik, N., Yang, L., Shen, Q. 2025. Optimisation of Multiple Clustering-Based Undersampling Using Artificial Bee Colony: Application to Improved Detection of Obfuscated Patterns without Adversarial Training, Information Sciences, 687, Article 121407. https://doi.org/10.1016/j.ins.2024.121407

[2] Ghasemkhani, B., Yilmaz, R., Kut, A., Birant, D., 2023, Logistic Model Tree Forest for Steel Plates Faults Prediction, Machines, 11 (7), 679, https://doi.org/10.3390/machines11070679.

[3] Sun, P., Du, Y., Xiong, S. 2024. Nearest neighbors and density-based undersampling for imbalanced data classification with class overlap, Neurocomputing, 609, Article 128492. https://doi.org/10.1016/j.neucom.2024.128492.

[4] Zuo, Y., Wan, M., Shen, Y., Wang, X., He, W., Bi, Y., Liu, X., Deng, Z. 2024. ILYCROsite: Identification of lysine crotonylation sites based on FCM-GRNN undersampling technique, Computational Biology and Chemistry, 113, Article 108212. https://doi.org/10.1016/j.compbiolchem.2024.108212.

[5] Lim, D., Van Doorsselaere, T., Nakariakov, V. M., Kolotkov, D. Y., Gao, Y., Berghmans, D. 2024. "Undersampling Effects on Observed Periods of Coronal Oscillations," Astronomy & Astrophysics, 690(L8). https://doi.org/10.1051/0004-6361/202451684.

[6] Nasibov, E., Dogan, A. 2016. An Efficient Algorithm for Classification of EEG Eye State Data, Global Journal of Information Technology: Emerging Technologies, 6 (3), 158-165, https://doi.org/10.18844/gjit.v6i3.

[7] Wainer, J. 2024. An Empirical Evaluation of Imbalanced Data Strategies from a Practitioner's Point of View, Expert Systems with Applications, 256, Article 124863. https://doi.org/10.1016/j.eswa.2024.124863.

[8] Bach, M. 2022, New Undersampling Method Based on the kNN Approach, 26th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2022), Procedia Computer Science 207, 3397-3406, https://doi.org/10.1016/j.procs.2022.09.399

[9] Kim, Y., Choi, W., Choi, W., Ko, G., Han, S., Kim, H.-C., Kim, D., Lee, D.-G., Shin, D. W., Lee, Y. 2024. A Machine Learning Approach Using Conditional Normalizing Flow to Address Extreme Class Imbalance Problems in Personal Health Record,. BioData Mining, 17(1), Article 14. https://doi.org/10.1186/s13040-024-00366-0.

[10] Hancock, J. T., Wang, H., Khoshgoftaar, T. M., Liang, Q. 2024. Data Reduction Techniques for Highly Imbalanced Medicare Big Data, Journal of Big Data, 11(1), Article 8. https://doi.org/10.1186/s40537-023-00869-3.

[11] Yang, C., Fridgeirsson, E. A., Kors, J. A., Reps, J. M., Rijnbeek, P. R. 2024. Impact of Random Oversampling and Random Undersampling on the Performance of Prediction Models Developed Using Observational Health Data, Journal of Big Data, 11(1), Article 7. https://doi.org/10.1186/s40537-023-00857-7.

[12] Kubicka, F., Nitschke, L., Penzkofer, T., Tan, Q., Nickel, M.D., Wakonig, K.M., Fahlenkamp, U.L., Lerchbaumer, M., Michallek, F., Dommerich, S., Hamm, B., Wagner, M., Walter-Rittel, T. 2024. "Dynamic contrast enhanced MRI of the head and neck region using a VIBE sequence with Cartesian undersampling and compressed sensing," Magnetic Resonance Imaging, 113, Article 110220. https://doi.org/10.1016/j.mri.2024.110220.