Araştırma Makalesi
Adıyaman Üniversitesi Mühendislik Bilimleri Dergisi
15 (2021) 343-352

A NEW RESAMPLING APPROACH BASED ON WEIGHTED GEOMETRIC MEAN FOR UNBALANCED DATA

Abdullah DAL[1], İbrahim Halil GÜMÜŞ[2], Serkan GÜLDAL[3], Mustafa YAVAŞ[4*]

[1]Adıyaman University, Graduate Education Institute, Department of Mathematics, Adıyaman, 02040, Turkey

[2]Adıyaman University, Faculty of Arts and Sciences, Department of Mathematics, Adıyaman, 02040, Turkey

[3]Adıyaman University, Faculty of Arts and Sciences, Department of Physics, Adıyaman, 02040, Turkey

[4]Adıyaman University, Vocational School of Technical Sciences, Department of Computer Technologies, Adıyaman, 02040, Turkey

Geliş Tarihi/Received Date: 26.05.2021 Kabul Tarihi/Accepted Date: 23.08.2021 DOI: 10.54365/adyumbd.940539

ABSTRACT

In recent years, there have been great improvements in data classification processes using machine learning methods. As technological advances increase, the size of data in the internet and other environments also increases rapidly. With these developments, unbalanced and unclassified data has emerged. The problem of imbalance is that one of the two classes has fewer samples than the other. Most of the datasets, especially used in the medical field, have an unbalanced distribution. A dataset with unbalanced distribution negatively affects the performance of classification algorithms. Many studies have been conducted to balance and classify this distribution. These studies are at the data and algorithm level and are undersampling and oversampling processes. In this study, the existing samples belonging to the minority class were resampled synthetically, and the datasets were balanced. For the resampling process, among the samples belonging to the minority class, the closest neighbors were determined for all data points using the Euclidean distance metric. Based on these neighbors, the desired number of new synthetic samples were created between each sample using the Weighted Geometric Mean. As a result of this process, the dataset has been balanced. In addition, Random Undersampling (RUS), Random Oversampling (ROS), and Synthetic Minority Sampling Technique (SMOTE) methods are also used to balance the datasets. The raw and balanced datasets are classified using the Random Forest algorithm, and the results are compared. As a result of the study, an increase is observed in all performance values of the datasets balanced with the new resampling approach. Using the approach proposed in the study, it is shown that the balanced datasets using the new resampling method improve the classification performance compared to the raw dataset and other methods.

Keywords: *Resampling, Weighted Geometric Mean, Unbalanced Data, SMOTE*

DENGESİZ VERİLER İÇİN AĞIRLIKLI GEOMETRİK ORTALAMA TABANLI YENİ BİR YENİDEN ÖRNEKLEME YAKLAŞIMI

ÖZET

Son yıllarda makine öğrenmesi yöntemleri kullanılarak veri sınıflandırma işlemlerinde büyük gelişmeler yaşanmıştır. Teknolojik gelişmeler arttıkça, internet ortamında ve diğer ortamlarda verilerin boyutu da hızla artmaktadır. Bununla beraber dengesiz ve sınıflandırılmamış veriler ortaya çıkmıştır. Dengesizlik problemi iki sınıftan birinin diğerine göre daha az örneğe sahip olması durumudur. Özellikle tıbbi alanda kullanılan veri kümelerin çoğu dengesiz dağılıma sahiptir. Dengesiz dağılıma sahip bir veri kümesi sınıflandırıcı algoritmaların başarım performansını olumsuz yönde etkilemektedir. Bu dağılımı dengelemek ve sınıflandırmak için birçok çalışma yapılmıştır. Bu çalışmalar veri ve algoritma düzeyinde olup, yeniden örnekleme yöntemi ile örneklem azaltma ve örneklem çoğaltma işlemleridir. Bu çalışmada azınlık sınıfa ait mevcut örnekler, yeniden sentetik olarak çoğaltılmıştır ve veri kümeleri dengelenmiştir. Yeniden örnekleme işlemi için, azınlık sınıfa ait örnekler arasında, Öklid uzaklık metriğiyle tüm data noktaları için en yakın komşular tespit edilmiştir. Bu komşular baz alınarak, her örnek arasında Ağırlıklı Geometrik Ortalama kullanılarak istenen sayıda yeni sentetik örnekler oluşturulmuştur. Bu işlem sonucunda veri kümeleri dengeli hale getirilmiştir. Ayrıca, veri setlerini dengelemek

* e-mail: [1] m.abdullah.dal@gmail.com ORCID ID: https://orcid.org/0000-0001-9306-6276
[2] igumus@adiyaman.edu.tr ORCID ID: https://orcid.org/0000-0002-3071-1159
[3] sguldal@adiyaman.edu.tr ORCID ID: https://orcid.org/0000-0002-4247-0786
[4] myavas@adiyaman.edu.tr ORCID ID: https://orcid.org/0000-0002-9111-9095 (Sorumlu Yazar)

için Rastgele Az Örnekleme (RUS), Rastgele Aşırı Örnekleme (ROS) ve Sentetik Azınlık Örnekleme Tekniği (SMOTE) yöntemleri de kullanılmıştır. Orijinal ve dengelenmiş veri kümeleri Random Forest algoritması ile sınıflandırılmış ve sonuçları kıyaslanmıştır. Çalışma sonucunda, yeniden örnekleme yaklaşımı ile dengelenen veri setlerinin tüm performans değerlerinde artış gözlemlenmiştir. Çalışmada önerilen yaklaşım ile yeniden örneklenerek dengelenen veri kümesi, ham veri kümesi ve diğer yöntemlere kıyasla sınıflandırma performansını iyileştirdiği gösterilmiştir.

Anahtar Kelimeler: *Yeniden Örnekleme, Ağırlıklı Geometrik Ortalama, Dengesiz Veri, SMOTE*

1. Introduction

It is not possible to manually process and analyze the increased data on the Internet and other digital platforms. In recent years, these data can be classified with artificial intelligence-based machine learning methods, and predictive analysis can be made from past data [1]. Machine learning methods are continuously developed to find the most suitable model for new data using existing data. The processes of analyzing the data, extracting and interpreting the useful information from it can be done with machine learning-based data mining [2]. Machine learning methods are used in many areas such as medical diagnostics, face recognition, text classification, fake transactions, and spam filtering [3]. The purpose of usage in these areas is to produce feasible solutions to complex events in our lives. However, machine learning and deep learning-based algorithms are insufficient in studies conducted in these areas. One of the problems in this regard is the problems caused by unbalanced datasets. An unbalanced dataset is an unequal distribution of data across classes. There is a noticeable decrease in the performance values of machine learning algorithms if the dataset is unbalanced. For this reason, resampling methods that stabilize unbalanced datasets are an important step in increasing the performance values of classifiers before starting the classification phase. There are two well-known methods emerging to eliminate this imbalance [4]. The first method is oversampling and making synthetic samples, the second method is undersampling. In the oversampling method, the data of the party with less class is generated by the existing dataset and reused to balance the two sets. In the synthetic data generation method, the classes are to be balanced by resampling the data of the minority class using certain algorithms. In the undersampling method, it is to ensure that the majority class in the dataset is reduced and then balanced.

According to the results of many studies, it has been observed that the applications that give more successful results in balancing the unbalanced dataset are the methods used by increasing the samples in the dataset. It has been observed that the results of datasets with fewer samples are more inconsistent. It was seen that the SMOTE algorithm of Chawla, Bowyer [5] is one of the methods that systematically balances the samples in the dataset and provides the most accurate operation of the application.

In this study, a synthetic sample generation method provides higher performance values in the classification phase of the datasets. In the developed method, the neighbor pairs of the samples belonging to the minority group in the predefined range are determined by using Euclidean distance metric, and new synthetic data is generated in the desired number of samples using Weighted Geometric Mean. The defined range changes respect to the balance of the overall dataset. With the applied new method, the datasets are balanced, and the new datasets are classified using the Random Forest algorithm. In addition, Random Undersampling (RUS), Random Oversampling (ROS), and Synthetic Minority Sampling Technique (SMOTE) methods are also used to balance the datasets. As a result of the experiments conducted in the study, it was found that the dataset balanced using the proposed method increased the performance values as a result of classification and improves the classification performance compared to the raw dataset and other sampling methods. As the performance value, overall Accuracy, Precision, Recall, F-Measure, and AUC-ROC values were taken into account.

2. Related Studies

Many different approaches have been proposed to classify the dataset to improve model performance and get better results. The simplest method to reduce the imbalance rate is to remove some

random samples from the majority class from the dataset. This approach is included in the Random Undersampling (RUS) method proposed by Batista et al. [8]. Unlike RUS, instead of removing samples from the majority class, Random Oversampling (ROS) randomly selects samples in the minority class, copies them and balances the dataset [8]. In Chawla et al. [5], the SMOTE algorithm was developed to stabilize the unbalanced datasets, and it is among the best working algorithms on this subject. Han et al. [6] proposed the Borderline- SMOTE method in the study by using the borderlines of the datasets formed by the samples in the minority class using the SMOTE method. Nguyen et al. [7] developed the SVM-SMOTE model in order to create a new minority dataset element by using support vectors. Batista et al. [8] developed the SMOTE-ENN method with the combination of the Edited Nearest Neighbors (ENN) methods by using the random sampling reduction method with SMOTE. Mani and Zhang [9] proposed the NearMiss method in order to reduce the sample numbers of high-dimensional datasets in a more systematic way instead of randomly decreasing them. Sun et al. [10] obtained high performance values by using the AdaBoost method at the algorithm level, weighting strategies for different datasets and the activities in identifying rare cases through experiments on several real-world medical datasets with class imbalance problem. Yijing et al. [11] proposed an adaptive model based on local properties for multi-class unbalanced datasets. It is stated that they can make a successful classification for almost all kinds of unbalanced datasets in the model they named Adaptive Multiple Classifier System. Rahman and Davis [12] proposed an advanced sampling technique by examining the performances of oversampling and undersampling techniques to balance medical data. As a result of the study, it is shown that the proposed method is more successful than the existing methods.

3. Material ve Method

3.1. Datasets Used

In the study, 4 real-life datasets with different sizes, different sample numbers, and different imbalance rates were used. Datasets are taken from KEEL (Knowledge Extraction based on Evolutionary Learning) open source software tool site [13]. The characteristics of the datasets used in the study are given in Table 1.

Table 1. Characteristics of the datasets. The imbalance ratio is obtained by dividing the majority class by the minority class.

Datasets	No. of Samples	Attributes	No. of Majority	No. of Minority	Imbalance Ratio
Pima	768	8	500	268	1.87
Wisconsin	683	9	444	239	1.86
Vehicle2	846	18	628	218	2.88
Yeast1	1484	8	1055	429	2.46

Considering the sample distribution and imbalance rate of the datasets shown in Table 1, it is seen that the datasets have an unbalanced distribution. By using the new resampling method and other methods proposed in the study, the majority and minority classes are approximated, samples belonging to the minority class were artificially resampled, so the datasets were balanced.

3.2. Resampling Methods

Many methods have been proposed in the data level for solutions to deal with the class imbalance problem [14, 15]. To balance the datasets used in this study, sample reduction method RUS, sample duplicate method ROS and synthetic sampling method SMOTE were used.

RUS is a non-heuristic way of balancing classes by randomly removing majority class instances. It consists of minimizing random samples from the majority class by eliminating them. Due to the fact that this elimination is done in an unsupervised way, there is a risk that samples useful for the classifier will be removed from the dataset. This method is often used because of its simplicity and increases the speed of the learning phase [16].

ROS is the simplest and oldest method of dealing with this problem. In this method, while the classifier trains until it reaches the desired ratio, it is used to balance the distribution by randomly copying the minority class samples to bring them closer to the larger class [17].

SMOTE has an oversampling approach to balance the raw dataset [5]. Unlike ROS, instead of implementing a simple copy of minority class instances, SMOTE generates synthetic instances from minority class instances. In the SMOTE algorithm, each sample belonging to the minority class finds k close neighbors using the Euclidean distance and randomly selects from the samples it finds. The difference between the sample and k nearest neighbors is taken, a random number (α) between 0-1 is chosen and multiplied by the difference found. Using formula 1, new synthetic samples are obtained.

$$x_{new} = x_i + (x_j - x_i) \times \alpha \tag{1}$$

where, x_i represents each minority sample, x_j represents the randomly selected neighbor of x_i's nearest neighbors, and x_{new} represents the new synthetic sample.

3.3. Proposed Method

In order to solve the problem of imbalance in datasets, a different method from previous approaches has been used. To balance the class distribution, minority class samples have been synthetically regenerated. In the approach suggested by the study, the samples belonging to the minority class are based on the closest neighbor samples while generating synthetic data. The distance between the nearest neighboring couples is measured with the Euclidean distance metric. Synthetic data is generated as much as the number of balances needed by using a weighted geometric mean among the samples identified in the predefined range. The method steps developed in the study are as follows;

- First, in the dataset, the imbalance ratio is determined by the ratio of the number of samples in the majority class to the number of samples in the minority class. If the dataset is unbalanced, other steps are applied.
- In order to balance the dataset, a sufficient number of synthetic data is generated from the minority class. At this phase, the Euclidean distance metric was used, which measures the distance between the two samples $x = [x_1, x_2, \dots . x_n]^T$ and $y = [y_1, y_2, \dots . y_n]^T$. The Euclidean distance metric is shown by formula 2.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

- A significant distance zone was defined for the samples and based on the remaining samples within the zone. The range of these areas is defined by brute force in the algorithm zone is expended until it reaches the number of missing data points. All the samples are selected within the zone.
- To generate new synthetic samples, we use weighted geometric mean. The weighted geometric mean for vectors with positive components is defined as follows: A random number between 0 and 1 is chosen for α. For $x = [x_1, x_2, \dots . x_n]^T$, $y = [y_1, y_2, \dots . y_n]^T \in R_+^n$ and $\alpha \in (0,1)$

$$x * y = \begin{bmatrix} x_1^{\alpha} y_1^{1-\alpha} \\ x_2^{\alpha} y_2^{1-\alpha} \\ \vdots \\ x_n^{\alpha} y_n^{1-\alpha} \end{bmatrix} \quad (3)$$

For generating new synthetic samples, we use the following formula,

$$S_{new} = x * y \quad (4)$$

- Formula 4 is repeated to generate the desired number of synthetic samples.

3.4. Random Forest Algorithm

The Random Forest algorithm is a powerful learning algorithm that generates results using multiple classifier estimates, instead of a single classifier, and classifies a new data sample with the votes from the classifier's predictions [18]. Random Forest is an ensemble classifier that uses multiple models of multiple decision trees to achieve better prediction performance. It searches for a random subset of variables to obtain a split at each node of the trees. For classification, the input vector is passed to each tree in the algorithm, and each tree votes for a class. The algorithm selects the class with the most votes [19]. The working system of the Random Forest algorithm is shown in Figure 1.

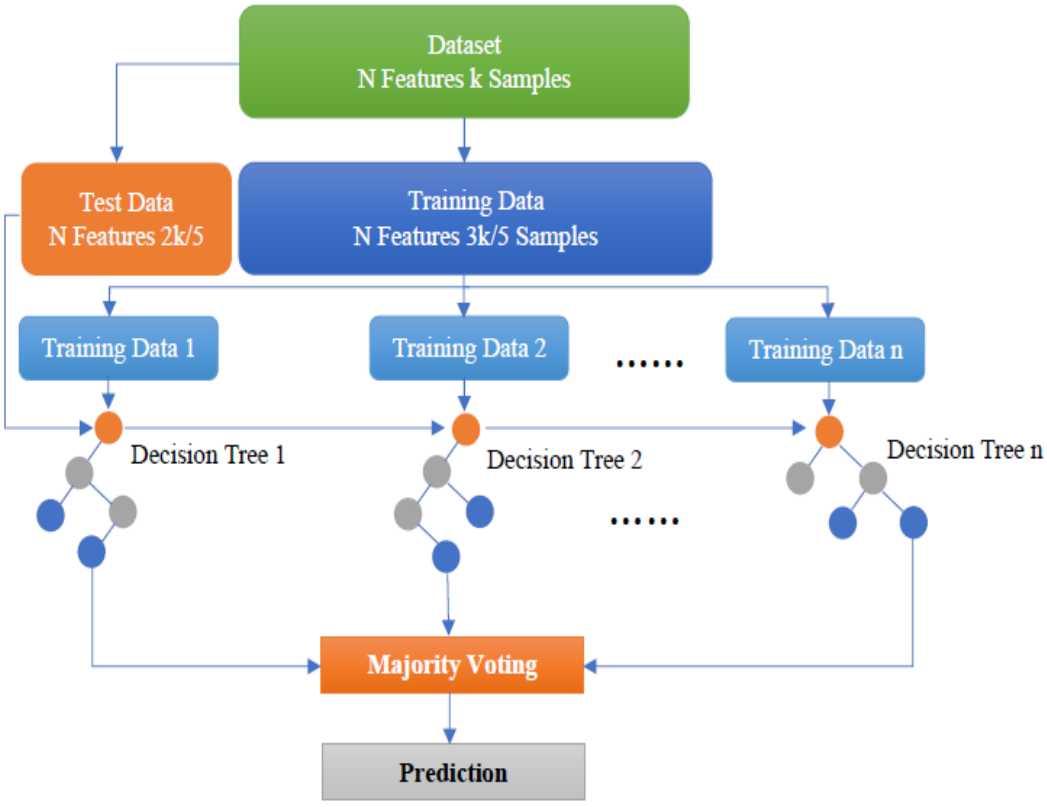


Figure 1. Random Forest algorithm working system

Random Forest method requires two different data groups as shown in Figure 1. These are training dataset (in-bag) and test dataset (out-of-bag, oob). The training dataset is 3/5 of the data and the test dataset is 2/5 throughout our study. The training dataset is used for training the tree. Test dataset is used to determine the generalized error rate (oob error) of the tree. A different training set is used for each tree in

the forest to be taken from the specified dataset. The training and test datasets of each tree are different to prevent any bias because of the selected dataset. In this study, the analysis is repeated 1500 times. If there is a data group reserved for testing purposes in the original dataset, this data group is used to detect the general fault of the forest. The average error rates of individual trees and the overall error rate of the forest are almost the same [20].

3.5. Model Performance Measurements

The performance of machine learning algorithms is typically evaluated using a confusion matrix, as shown in Table 2 for a binary class problem. The columns are the predicted classes and the rows are the actual classes. In the confusion matrix, TP represents true positive samples, FP represents false positive samples, TN represents true negative samples, and FN represents false negative samples.

Table 2. Confusion Matrix

	Predicted Positive Samples	Predicted Negative Samples
Actual Positive Samples	TP	FN
Actual Negative Samples	FP	TN

Different evaluation measures can be calculated using the confusion matrix. Accuracy, Recall, Precision, F-Measure, and AUC-ROC were used in this study.

Accuracy is the ratio of samples correctly classified by a classifier to the number of all samples.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (5)$$

Recall is obtained by dividing the number of samples correctly classified as positive by all positive samples.

$$Recall = TP/(TP + FN) \quad (6)$$

Precision is the calculation obtained by dividing the number of samples correctly classified as positive by all samples classified as positive.

$$Precision = TP/(TP + FP) \quad (7)$$

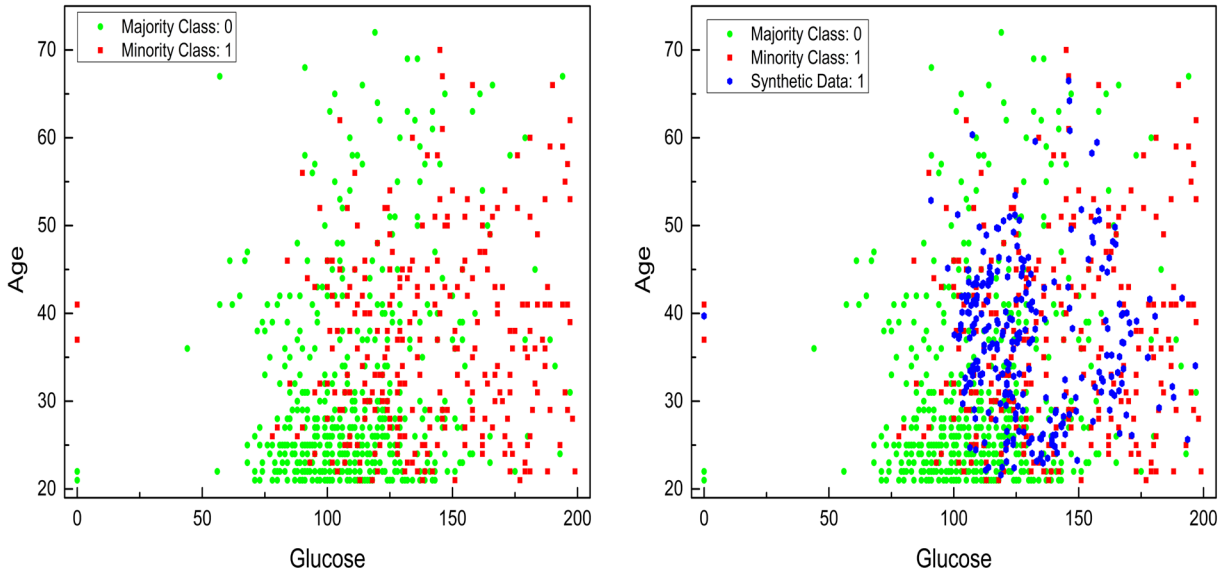
F-Measure is the harmonic mean of Precision and Recall values. Calculated with formula 8.

$$F - Measure = 2 * (Precision * Recall)/(Precision + Recall) \quad (8)$$

The Area Under the Curve-Receiver Operating Characteristic (AUC-ROC) curve is a performance measurement of data with binary or multi classes as a result of classification operations. The ROC is a probability curve, and the AUC represents the degree or measure of separability. In the AUC-ROC curve, the true positive rate-tpr (Recall) is drawn in the function of the false positive rate-fpr for different breakpoints. Each point in the ROC curve represents a tpr/fpr pair corresponding to a specific decision threshold. A test with perfect discrimination (no overlap in two distributions) has a ROC curve passing through the upper left corner. Therefore, as the ROC curve approaches the top left corner, the overall accuracy of the test increases [21].

4. Results and Discussions

In this section, we present the experimental results of the resampling approach developed based on weighted geometric mean. For the experimental results of the study, the datasets given in Table 1 is used. For example, in the Pima dataset, out of 768 patient samples, 500 belonged to the majority (0) class and 268 to the minority (1) class. Looking at these numbers, it seems that the dataset has an unbalanced distribution. In order to minimize the imbalance rate and obtain a balanced dataset, minority class samples were produced synthetically with the above-mentioned resampling approach. A total of 535 samples were obtained by producing 267 synthetic samples from minority class samples. In order to visualize the balanced dataset class distribution, Age and Glucose attributes are taken as a basis, and in Figure 2, before (raw data) and after resampling is shown in two-dimensional plane.



(a) Raw data (b) Resampled data using our approach

Figure 2. (Color online) Pima dataset is 500 majority and 268 minority. 267 synthetic data is generated by weighted geometric mean

As shown in Figure 2, it is seen that the minority classes are brought closer to the majority group and the closest neighbor values are sampled in the region where the minority samples are found in density.

In order to improve the classification performance, samples belonging to the minority group were resampled and all datasets were balanced. The datasets and algorithms we used in our study were coded in the Wolfram Mathematica program using Machine Learning functions. Windows Server 2019 Standard Edition is used for all classification and other operations. Server hardware information; System Type: x64-based PC, Processor(s): Intel(R) Xenon(R) CPU E5-2640 v3 @ 2.60 GHz 2.60 GHz (2 Processors), Total Physical Memory: 63.362 MB.

The datasets, which are balanced as a result of approximation of the majority and minority class samples, are classified by the Random Forest algorithm. In addition, raw and other resampled data sets were also classified, and the average performance values of both classes were compared. The average of Precision, Recall, F-Measure, and AUC-ROC of the Random Forest algorithm with the 4 different sampling methods are summarized in Table 3-6. The best performance of the sampling methods with a fixed classifier is in bold.

Table 3. Performance Values of Pima Dataset Classification Results

Dataset Status	Accuracy	Precision	Recall	F-Measure	AUC-ROC
Raw	0.749	0.731	0.699	0.706	0.813
RUS	0.737	0.739	0.737	0.736	0.818
ROS	0.757	0.698	0.741	0.707	0.817
SMOTE	0.786	0.787	0.782	0.785	0.863
Our Approach	**0.792**	**0.793**	**0.792**	**0.792**	**0.873**

Table 4. Performance Values of Wisconsin Dataset Classification Results

Dataset Status	Accuracy	Precision	Recall	F-Measure	AUC-ROC
Raw	0.705	0.526	0.497	0.424	0.710
RUS	0.956	0.955	0.943	0.948	0.975
ROS	0.943	0.942	0.947	0.942	0.974
SMOTE	0.961	0.962	0.945	0.953	0.974
Our Approach	**0.969**	**0.970**	**0.953**	**0.961**	**0.981**

Table 5. Performance Values of Vehicle2 Dataset Classification Results

Dataset Status	Accuracy	Precision	Recall	F-Measure	AUC-ROC
Raw	0.972	0.972	0.957	0.963	0.996
RUS	0.961	0.962	0.961	0.961	0.991
ROS	0.977	0.918	0.984	0.946	**0.998**
SMOTE	**0.984**	0.983	**0.984**	**0.983**	**0.998**
Our Approach	**0.984**	**0.984**	0.983	**0.983**	**0.998**

Table 6. Performance Values of Yeast1 Dataset Classification Results

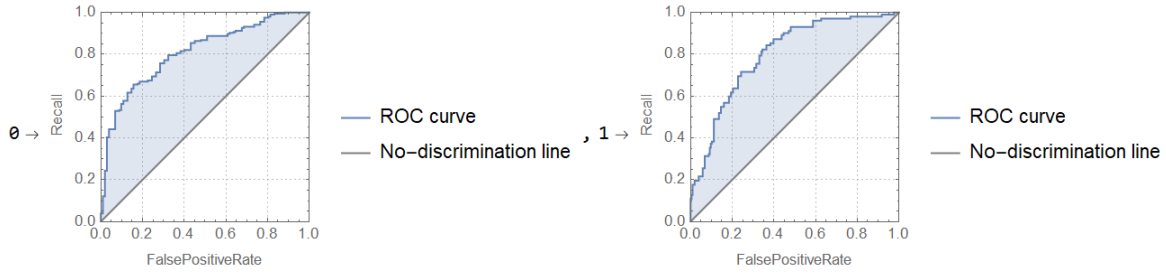
Dataset Status	Accuracy	Precision	Recall	F-Measure	AUC-ROC
Raw	0.761	0.725	0.650	0.662	0.780
RUS	0.709	0.713	0.709	0.707	0.780
ROS	0.766	0.642	0.722	0.655	0.796
SMOTE	0.772	0.774	0.773	0.772	0.855
Our Approach	**0.775**	**0.777**	**0.775**	**0.774**	**0.861**

When the performance values of the classification results shown between Table 3 and Table 6 are examined, it is clearly seen that our proposed method is more successful than raw and other methods. When we compared the raw and new approach resampled dataset performance values in Table 3, the overall accuracy value increased from 0.749 to 0.792. Other values, Precision increased from 0.731 to 0.793, Recall from 0.699 to 0.792, F-Measure from 0.706 to 0.792 and AUC-ROC from 0.813 to 0.873. When we examine the Accuracy, Precision, Recall, F-Measure, and AUC-ROC performance values obtained from the new approach between Table 4 and Table 6; In Table 4, the results are 0.969, 0.970, 0.953, 0.961, and 0.981 respectively. In Table 5, the results are 0.984, 0.984, 0.983, 0.983, and 0.998 respectively. In Table 6, the results are 0.775, 0.777, 0.775, 0.774, and 0.861 respectively.

There is improvement in all result values of the new approach compared to raw and other methods. It is especially striking that there is a higher increase in the Recall values for all datasets. In addition, in the resampled datasets, the Precision, Recall and F-Measure values produced close results and brought the

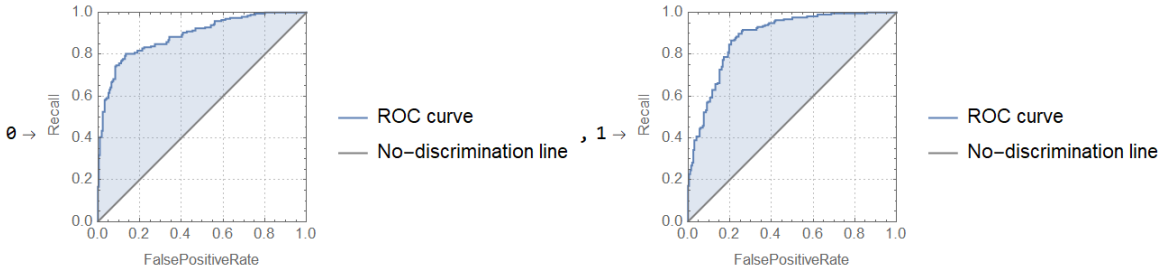
performance values closer together. Although the performance values of the RUS, ROS, and SMOTE methods are close to each other, the SMOTE method has been more successful than the others. However, it has been less successful than the method we proposed.

In addition, the ROC curve of the Pima dataset is shown as an example in Figures 3 and 4 for raw and resampled respectively. Here, the performance values of each class are visualized separately with the ROC curve.



(a) (b)

Figure 3. ROC curves of the raw data. (a) is majority class (0) and (b) is minority class (1)



(a) (b)

Figure 4. ROC curves of the resampled data. (a) is majority class (0) and (b) is minority class (1)

In Figures 3 and 4, the AUC values of the raw and resampled datasets are shown with the ROC curve. As a result of the Random Forest classification, the AUC value of class 0 increased from 0.813 to 0.873 and the AUC value of class 1 increased from 0.814 to 0.874. When the AUC values of the classes and the ROC curves are examined, it is seen that the resampled dataset is more successful in both classes. In addition, the approach of the ROC curve to the upper left corner indicates that the Recall (tpr) ratio is high, so the area under the curve is high. Based on this, it is seen that the resampled dataset has more area under the curve than the raw dataset since the classification of resampled data makes a more successful distinction.

5. Conclusion

In this study, a synthetic sample replication method is proposed, which provides higher performance values as a result of balancing and classifying the datasets by using datasets with four different unbalance ratios. In the proposed method, the nearest neighbors of the samples belonging to the minority group were determined using Euclidean distance metric and new synthetic data was generated in the desired number of samples using Weighted Geometric Mean. With the applied method, the datasets were balanced and classified using Random Forest algorithm. The performance values of the raw and resampled datasets were compared and the resampled dataset was more successful in almost all metrics. It was also found that the

resampled dataset had more area under the curve than the raw dataset, making a more successful distinction. RUS, ROS, and SMOTE methods were also compared with the proposed method in the study and it was seen that the new method was more successful. As a result of the experiments conducted in the study, it was determined that the datasets that were balanced using the proposed method increased the performance values as a result of the classification.

References

[1] E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.

[2] D. T. Larose and C. D. Larose, *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

[3] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information,* vol. 10, no. 4, p. 150, 2019.

[4] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," *International Journal of Recent Trends in Engineering and Research,* vol. 3, no. 4, pp. 444-449, 2017.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[6] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*, 2005: Springer, pp. 878-887.

[7] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *International Journal of Knowledge Engineering and Soft Data Paradigms,* vol. 3, no. 1, pp. 4-21, 2011.

[8] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD explorations newsletter,* vol. 6, no. 1, pp. 20-29, 2004.

[9] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, 2003, vol. 126: ICML United States.

[10] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition,* vol. 40, no. 12, pp. 3358-3378, 2007.

[11] L. Yijing, G. Haixiang, L. Xiao, L. Yanan, and L. Jinling, "Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data," *Knowledge-Based Systems,* vol. 94, pp. 88-104, 2016.

[12] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing,* vol. 3, no. 2, p. 224, 2013.

[13] K. D. Repository. (12.04.2021). *Imbalanced Dataset* [Online]. Available: https://sci2s.ugr.es/keel/imbalanced.php.

[14] R. Longadge and S. Dongre, "Class imbalance problem in data mining review," *arXiv preprint arXiv:1305.1707,* 2013.

[15] S. M. Abd Elrahman and A. Abraham, "A review of class imbalance problem," *Journal of Network and Innovative Computing,* vol. 1, no. 2013, pp. 332-340, 2013.

[16] S. Vluymans, *Dealing with imbalanced and weakly labelled data in machine learning using fuzzy and rough set methods*. Springer, 2019.

[17] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from imbalanced data sets*. Springer, 2018.

[18] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[19] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news,* vol. 2, no. 3, pp. 18-22, 2002.

[20] M. Ercire, "Classification of short-term power quality disturbances with wavelet analysis and random forest method," Ph.D Doctoral 2019.

[21] S. Narkhede, "Understanding auc-roc curve," *Towards Data Science,* vol. 26, pp. 220-227, 2018.