Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2023;15(2):79-90

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/biostatic.2023-96293

A New Class-Weighting Formulation for the Class Imbalance Problem: A Methodological Research

Sınıf Dengesizliği Problemi İçin Yeni Bir Sınıf Ağırlıklandırma Formülasyonu: Metodolojik Bir Araştırma

Batuhan BAKIRARAR[a], Selen YILMAZ IŞIKHAN[b]

[a]Department of Biostatistics, Ankara University Faculty of Medicine, Ankara, Türkiye
[b]Department of Economics and Administrative Programs, Hacettepe University Vocational School of Social Sciences, Ankara, Türkiye

ABSTRACT Objective: Many of the machine learning classification algorithms are not robust against unbalanced classes and result in poorly accurate and biased models. One way to address class imbalance is to assign weights to classes. This article proposes a new class-weighting approach to improve the classification problem when there is an imbalance between two class. **Material and Methods:** The performances of the new formulation were compared with the previously proposed Inverse of Square Root of Number of Samples, effective number of samples weighting formula and unweighted Random Forest solutions. A simulation study was performed using performances of 3 imbalance rates (0.10, 0.20, 0.30), 6 different sample sizes (250, 300, 350, 400, 450, 500) and 4 different methods with 1,000 repetitions. Additionally, the methods were analyzed on the lung cancer dataset with 39 samples in the minority group and with 270 samples in the majority group. **Results:** Experimental results demonstrated that our proposed weighting formula, least number of ratio and range multiplier, performed equal to or better solution than Inverse of Square Root of Number of Samples in both simulations and real data. Generally, minority class accuracy and balanced accuracy of our formulation were either very close to or higher than that of Inverse of Square Root of Number of Samples. **Conclusion:** The new formulation provided accuracy estimates of the 2 classes in a balanced way for each sample size and for each imbalance rate. Additionally, as the sample size increased from 250 to 500, stable decreasing weights could be obtained for the patient and control groups.

Keywords: Class Imbalance;
class-weighting methods;
classification;
Random Forest algorithm

ÖZET Amaç: Makine öğrenimi sınıflandırma algoritmalarının birçoğu, dengesiz sınıflara karşı güçlü değildir ve doğruluğu düşük ve yanlı modeller ile sonuç verir. Sınıf dengesizliğini çözmenin bir yolu, sınıflara ağırlık atamaktır. Bu makale, 2 sınıf arasında bir dengesizlik olduğunda sınıflandırma problemini iyileştirmek için yeni bir sınıf ağırlıklandırma yaklaşımı önermektedir. **Gereç ve Yöntemler:** Yeni formülasyonun performansları, daha önce önerilen Örnek Sayısının Karekökünün Tersi, etkin örnek sayısı ağırlıklandırma formülü ve ağırlıklandırılmamış Random Forest çözümleri ile karşılaştırılmıştır. Üç dengesizlik oranı (0,10, 0,20, 0,30), 6 farklı örneklem büyüklüğü (250, 300, 350, 400, 450, 500) ve 4 farklı yöntemin 1.000 tekrarlı performansları kullanılarak simülasyon çalışması yapılmıştır. Ayrıca yöntemler azınlık grubunda 39 örnek ve çoğunluk grubunda 270 örnek ile akciğer kanseri veri setinde analiz edilmiştir. **Bulgular:** Deneysel sonuçlar, önerilen ağırlıklandırma formülümüz olan en az sayı oranı ve açıklık çarpanının hem simülasyonlarda hem de gerçek veride Örnek Sayısının Karekökünün Tersi'ninkine eşit veya daha iyi bir performans gösterdiğini belirtmiştir. Genel olarak formülümüzün azınlık sınıfı doğruluğu ve dengeli doğruluğu, Örnek Sayısının Karekökünün Tersi formülünün doğruluğuna çok yakın ya da daha yüksektir. **Sonuç:** Yeni formülasyon, her örneklem büyüklüğü ve her bir dengesizlik oranı için 2 sınıfın doğruluk tahminlerini dengeli bir şekilde sağlamıştır. Ayrıca örneklem büyüklüğü 250'den 500'e çıkarıldığında hasta ve kontrol grupları için tutarlı azalan ağırlıklar elde edilebilmiştir.

Anahtar kelimeler: Sınıf dengesizliği;
sınıf ağırlıklandırma yöntemleri;
sınıflandırma;
Rastgele Orman algoritması

Correspondence: Selen YILMAZ IŞIKHAN
Hacettepe University Vocational School of Social Sciences, Ankara, Türkiye
E-mail: seleny@hacettepe.edu.tr

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 23 Feb 2023 ***Received in revised form:*** 28 Mar 2023 ***Accepted:*** 28 Mar 2023 ***Available online:*** 31 Mar 2023

In a problem that considers the classification of data samples in two groups, class imbalance occurs especially when the sample size of the patient group is less than the sample size of the healthy group.[1] Thus, the dataset contains two types of samples: minority and majority. In most problems, the minority class is the positive group. Based on the studies, it can be said that the class imbalance problem arises frequently in classification problems in the field of health.[2] In a sample where most patients are healthy and the detection of the disease is critical, the most of healthy patients can be identified as negative class. Learning from these unstable datasets can be overwhelming, specifically when analysing with big data. This causes the minority class to be underrepresented in the dataset and the trained classification models often produce biased predictions in favor of the majority class.[3]

Therefore, modeling from imbalanced data sets reveals great difficulties for traditional classification algorithms. Imbalanced class problem occurs naturally in many situations where the positive class has smaller samples such as disease diagnosis, fault detection, computer security, fraud detection and image recognition.[4-8]

Class imbalance is generally measured by the imbalance ratio (IR)=M/m, where m and M are observations size in the positive and negative group, respectively. Based on the value of IR, the imbalanced datasets are separated into 3 groups: datasets with high imbalance (IR is higher than 9), datasets with moderate imbalance (IR is between 3 and 9), datasets with low imbalance (IR is between 1.5 and 3).[9] At the same time, methods for solving class imbalance in machine learning are divided into 3 types: data-level techniques, algorithm-level techniques, and hybrid techniques. Data level methods try to reduce the imbalance level with the help of different data sampling techniques. Algorithm-level techniques to address class imbalance, often done with a weight or cost scheme, involve modifying the base learner or its output to reduce bias in favor of the majority group. Lastly, hybrid techniques assemble both sampling and algorithmic techniques strategically.[1,10]

In resampling, the number of samples is arranged straightly by oversampling (adding repetitive data) for the minority class or undersampling (removing the data) for the majority class, or both. In cost-sensitive reweighting, the loss function was affected by appointing comparatively higher costs to samples in the small class. Mostly, it has been proposed to appoint sample weights inversely proportional to the class frequency. This procedure has been widely used.[11]

In this study, a new weighting formula was developed to balance the classification performance of classes. The performances of the new formulation will be compared with the Inverse of Square Root of Number of Samples (ISNS), effective number of samples (ENS) weighting formula and unweighted Random Forest solutions.

MATERIAL AND METHODS

DATA-LEVEL METHODS

Data-level methods for considering class imbalance consist of oversampling and undersampling. These methods modify the training distributions of mislabeled samples or anomalies to reduce the imbalance level or noise. Basically, random samples are excluded from the majority group in random undersampling (RUS), whereas in random oversampling (ROS) random samples in the minority group are amplified.[1]

OVERSAMPLING

ROS is the process of multiplying existing minority samples to expand the size of a minority class. In this method, additional samples are added to the minority class to balance the data set (Figure 1). Oversampling successfully balances the input dataset and ensures maximum accuracy by multiplying the minority class samples until both classes have an equal number. Oversampling causes the training time to increase due to the increase in the size of the training set, and it has also been demonstrated to lead to overfitting.[1] To over-

come this problem, the Synthetic Minority Oversampling Technique (SMOTE) method, which creates "synthetic" samples instead of oversampling with replacement, has been proposed.[12] SMOTE randomly selects a minority class sample and starts by randomly finding its k nearest minority class neighbors.

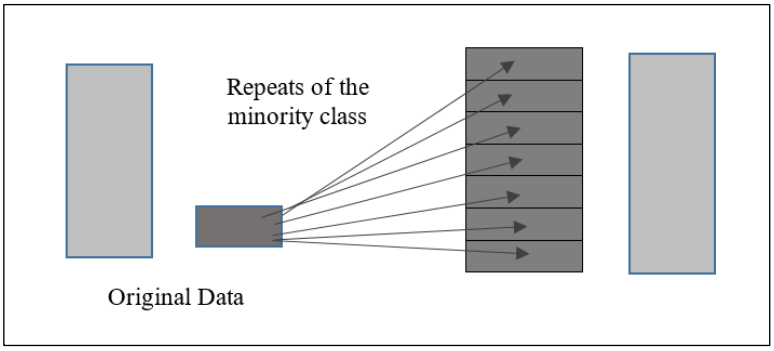

FIGURE 1: Representation of oversampling.[13]

UNDERSAMPLING

In this technique, majority class samples are decreased to balance the data set. RUS intentionally excludes data and reduces the overall amount of information the model can learn (Figure 2). Undersampling successfully balances the input dataset, ignoring the remaining samples available, only considering specific samples in the majority class that meet the number equal to that of the minority class.[13]

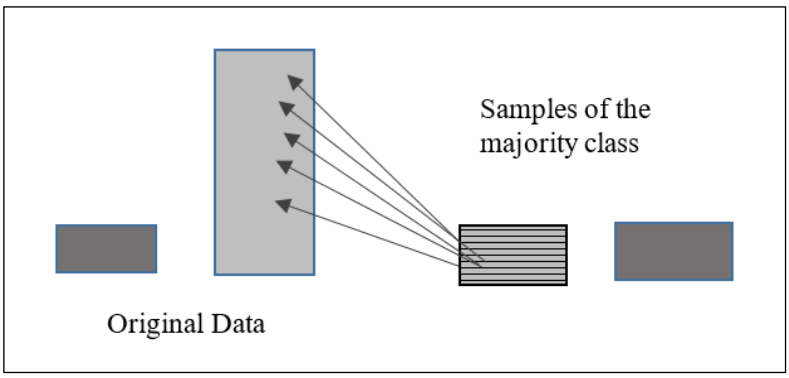

FIGURE 2: Representation of undersampling.[13]

CLASS WEIGHTING METHOD

Oversampling reveals duplicate samples, causing overfitting in the model, while undersampling disqualifies a certain number of samples. Another solution to overcome problems such as overfitting and underfitting is to play with the loss function. It is a good option to assign a higher weight to the loss faced by instances associated with small classes to make up for the class imbalance.[14]

Various weighting methods are available to calculate the class weight. In this study, Random Forest classification performances will be compared on the proposed weighting formula [Least Number of Ratio and Range Multiplier (LNR+RM)], ISNS weighting formula, ENS weighting scheme and unweighted data.

ISNS

In this weighting formula, samples are weighted as the inverse of the Square Root of the class frequency for the class they belong to.

$$W_{n,c} = 10 * \frac{1}{\sqrt[2]{Number\ of\ Samples\ in\ Class\ c}}$$

ENS

The basis of this method is to represent each sample not by a single point, but by a small neighboring region, because the exstra profit of a newly added data point decreases as the number of samples increases. ENS is defined as the sample size, and the simple formulation of the ENS is as below:

$$\beta = \frac{(N-1)}{N}$$

$$ENS = \frac{(1-\beta^N)}{(1-\beta)}$$

$$W = \frac{100}{ENS}$$

where N: Sample size in the class, ENS: Number of effective Samples, W: Sample weight.[11]

PROPOSED CLASS WEIGHTING FORMULA

LNR+RM

The weighting formula suggested in the study is as follows:

$$W_{LNR+RM} = \frac{2\times class}{\sqrt{n_j \times r_{maj}}} Ln\left(\frac{n}{n_j}\right) + \left(\frac{1}{2}\sqrt{\frac{n_{maj}-n_{min}}{n}}\right), \qquad r_{maj}=0.90,\ 0.80,\ 0.70$$

Here; nj: Sample size of the relevant class, n: Total sample size, class: Number of classes, rmaj: Ratio of majority class (0.90, 0.80, 0.70, respectively), nmaj: Sample size of the majority class, nmin: Sample size of the minority class.

The first term of the weight formula we developed is the logarithm of the ratio of the total sample size to the jth class sample size, and the second term is adjusted to be a factor of the range of class sizes. The logarithm transformation in the first term of the formula is a data transformation method that changes each variable x with a Log(x). When the original numerical data do not fit the symmetrical curve, we take the logarithm of this data to bring it closer to "normal", so the analysis findings of this data will be more valid. Namely, the logarithm transform reduces or eliminates the skewness of our original data. Therefore, if nj is close to n, the logarithm of the ratio N/nj will be closer to 0. If nj is smaller than n, the first term of the weight will most likely be greater than 1 (Ln (n/nj)>1) and will increase proportionally to this ratio. The coefficient before Ln is a constant that provides balance by slightly decreasing the weight to be obtained in each IR proportionally to rmaj.

The second term of the formulation ($\frac{1}{2}\sqrt{\frac{n_{maj}-n_{min}}{n}}$) is added as a multiplier depending on the range of the class size for both of them.

PERFORMANCE EVALUATION CRITERIA

The classical general accuracy metric is not convenient to evaluate classification performance in case of imbalanced class because the positive class occupies lesser importance in this measure than the negative class.[8] Therefore, Matthews correlation coefficient (MCC), balanced accuracy, accuracy, and geometric mean (G-mean) were used as performance evaluation measures in this study.[15]

G-MEAN

The G-mean formula measures performance by merging both true positive rate (TPR) and true negative rate (TNR) metrics using the square root of products. The G-mean is a measure that detects the balance between classification performances in both classes. Although negative cases are correctly classified, a low G-mean is indicative of weak performance in classifying positive cases.[16] Confusion matrix for formulation is as follows:

		Predicted Condition	
Actual Condition		**Positive**	**Negative**
	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

$$\text{Sensitivity}=\frac{\text{TP}}{\text{TP+FN}}\times 100$$

$$\text{Specificity}=\frac{\text{TN}}{\text{TN+FP}}\times 100$$

$$\text{G-mean}=\sqrt{\text{Sensitivity*Specificity}}$$

BALANCED ACCURACY

This measure merges TPR and TNR values to calculate a measure that is more susceptible to the minority group. If a classifier performs equally well in both classes, balanced accuracy coincides with conventional accuracy. Conversely, if the high value of conventional accuracy is due to the classifier uses the majority class distribution, the balanced accuracy will be smaller than accuracy.

$$\text{Balanced Accuracy}=\frac{1}{2}(\text{Sensitivity}*\text{Specificity})$$

MCC

The MCC is a special form of the coefficient ϕ phi.[15] F1 score and accuracy can be inaccurate in imbalanced datasets because they do not take into account the ratio between positive and negative groups. A classifier must yield accurate predictions on both the majority of negative cases and the majority of positive cases, regardless of their ratio in the entire dataset. Therefore, the MCC provided more informative and accurate scores than the other 2 metrics.[16]

The MCC ranges from -1 to 1, that here -1 represents a completely incorrect binary classifier whereas 1 represents the opposite.

$$\text{MCC}=\frac{(\text{TP}*\text{TN})-(\text{FP}*\text{FN})}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}$$

OVERALL ACCURACY

This metric is the ratio of the total true positive and true negative results to the total number of observations found in the study. In the case of unbalanced datasets, using accuracy to evaluate the performance of the classification model can be misleading. The overall accuracy value ranges from 0 to 1. If this value is close to 1, it represents high performance.[17]

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

STATISTICAL ANALYSIS

The data were analyzed using the R (Version 4.2.2) programming language. The data were analyzed over 1,000 iterations, and the 10-fold cross validation was used as the data splitting method. Accuracy, MCC, G-mean and balanced accuracy were utilized as performance metrics. In the R programming language, RWeka, randomForest, caret and Metrics libraries were used.

DATA AND APPLICATIONS

Real Dataset

As the real data set, lung cancer data collected from the online lung cancer prediction system website via the kaggle website[1] was used. The data set has a total sample size of 309, with 39 samples in the minority class and 270 samples in the majority class. The original dataset consisted of 16 variables (including gender, smoking, age yellow fingers, allergy, wheezing, alcohol, coughing, anxiety, peer_pressure, chronic disease, swallowing difficulty, chest pain fatigue, shortness of breath, lung cancer).

Simulation Study

Correlations between dependent and independent variables were defined as (0.30-0.65) for all distributions. It is assumed that there is a weak or no relationship between the independent variables with the range of (0.08-0.25). Six independent variables, together with the class variable, were first derived from the multivariate normal distribution using the "MASS" package "mvrnorm" functionin R Studio software, taking into account the relevant correlation structures.[18] The mean vector of the variables was taken into account as $\mu=(10, 5, 2, 16, 20, 30)$. Continuous independent variables derived from the normal distribution were then transformed into binary variables (0 and 1) by taking into account the different cut-off points (0.20, 0.25 and 0.30) of the data distribution with the help of the "mutate" function.[19] Class assignment was made to the class variable with the "mutate" function, considering the class imbalance rates of 9, 4 and 2.33. For each of these three class imbalances, datasets with sample sizes of 250, 300, 350, 400, 450 and 500 were created. Classification performances were obtained by applying Random Forest classification method to the datasets together with unweighted, ISNS, ENS and our newly developed weighting formula (LNR+RM). A total of 72 different scenario combinations were analyzed with 1,000 repetitions each with 4 different classification methods in each of 18 scenarios for 3 class IRs (9, 4, and 2.33) and 6 different sample sizes (250, 300, 350, 400, 450, 500).

Declaration

In this manuscript, there is no research involving "Human beings". The research was conducted in accordance with the principles set of the Helsinki Declaration 2008. Informed consent was not obtained as there were no subjects in the study. Accordingly, approval by the ethical board is not needed for this manuscript.

RESULTS

The class weights of each method for each sample size in each class distribution are given in Table 1.

[1]https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

TABLE 1: Class weights in different samples for methods.

Distribution	Sample size	Method					
		ISNS		ENS		LNR+RM	
		Class 1	Class 2	Class 1	Class 2	Class 1	Class 2
90%-10% (IR=9)	250	0.667	2.000	0.702	6.254	0.477	2.389
	300	0.609	1.826	0.585	5.222	0.474	2.220
	350	0.563	1.690	0.502	4.482	0.472	2.088
	400	0.527	1.581	0.439	3.926	0.471	1.982
	450	0.497	1.491	0.390	3.493	0.469	1.894
	500	0.471	1.414	0.351	3.145	0.468	1.820
80%-20% (IR=4)	250	0.707	1.414	0.790	3.145	0.458	1.405
	300	0.645	1.291	0.658	2.624	0.452	1.317
	350	0.598	1.195	0.564	2.251	0.447	1.248
	400	0.559	1.118	0.494	1.970	0.443	1.192
	450	0.527	1.054	0.439	1.752	0.440	1.146
	500	0.500	1.000	0.395	1.577	0.437	1.107
70%-30% (IR=2.33)	250	0.756	1.155	0.902	2.101	0.445	0.981
	300	0.690	1.054	0.752	1.752	0.434	0.923
	350	0.639	0.976	0.645	1.502	0.425	0.878
	400	0.598	0.913	0.564	1.315	0.418	0.842
	450	0.563	0.861	0.502	1.169	0.412	0.812
	500	0.535	0.816	0.452	1.053	0.407	0.786
Real dataset	309	0.609	1.601	0.585	4.026	0.467	1.850

Class 1: Class with large sample; Class 2: Class with low sample; ISNS: Inverse of Square Root of Number of Samples; ENS: Effective number of samples; LNR+RM: Least number of ratio and range multiplier; IR: Imbalance ratio.

In Table 2, the results of the methods in different samples are given for the simulated dataset with a distribution of IR=9. In the 250 sample dataset, ISNS and LNR+RM methods gave similar results, while the ENS method performed very poorly. On datasets with 300, 350, 400, 450 and 500 samples, all three methods had similar performance.

TABLE 2: Results of the methods for the simulated dataset with IR=9 distribution.

Sample Size	Method	Performance Measures					
		Accuracy1	Accuracy2	Accuracy	MCC	G-mean	Balanced Accuracy
250	Unweighted	0.999	0.274	0.927	0.502	0.523	0.137
	ISNS	0.834	0.675	0.818	0.371	0.751	0.282
	ENS	0.361	0.675	0.392	0.023	0.493	0.122
	LNR+RM	0.827	0.674	0.812	0.362	0.747	0.279
300	Unweighted	0.999	0.362	0.935	0.573	0.601	0.181
	ISNS	0.806	0.767	0.802	0.396	0.786	0.309
	ENS	0.791	0.769	0.789	0.381	0.780	0.304
	LNR+RM	0.799	0.768	0.796	0.389	0.783	0.307
350	Unweighted	0.993	0.442	0.938	0.599	0.662	0.220
	ISNS	0.817	0.829	0.818	0.449	0.823	0.338
	ENS	0.806	0.829	0.808	0.435	0.817	0.334
	LNR+RM	0.811	0.829	0.813	0.441	0.820	0.336
400	Unweighted	0.991	0.463	0.939	0.603	0.677	0.229
	ISNS	0.821	0.850	0.824	0.468	0.836	0.349
	ENS	0.806	0.850	0.810	0.449	0.828	0.342
	LNR+RM	0.814	0.850	0.818	0.459	0.832	0.346
450	Unweighted	0.988	0.514	0.941	0.624	0.712	0.254
	ISNS	0.838	0.867	0.840	0.500	0.852	0.363
	ENS	0.823	0.867	0.827	0.481	0.845	0.357
	LNR+RM	0.831	0.867	0.835	0.491	0.849	0.360
500	Unweighted	0.984	0.559	0.942	0.639	0.742	0.275
	ISNS	0.845	0.880	0.849	0.519	0.862	0.372
	ENS	0.833	0.880	0.837	0.502	0.856	0.366
	LNR+RM	0.841	0.880	0.845	0.514	0.860	0.370

Accuracy1: Accuracy for majority class; Accuracy2: Accuracy for minority class; MCC: Matthews correlation coefficient; G-mean: Geometric mean; ISNS: Inverse of Square Root of Number of Samples; ENS: Effective number of samples; LNR+RM: Least number of ratio and range multiplier.

In Table 3, the results of the methods in different samples are given for the simulated dataset with IR=4 distribution. In the 250 sample dataset, ISNS had the best performance, while LNR+RM was the second best performing method. In the 300 sample dataset, LNR+RM had the best performance, while ISNS was the second best performing method. In the datasets with 350, 400, 450, and 500 samples, the LNR+RM method provided the best improvement in the performance measure of the low-sampling class, while it had similar performance with ISNS when general criteria were considered.

TABLE 3: Results of the methods for the simulated dataset with IR=4 distribution.

Sample Size	Method	Performance measures					
		Accuracy1	Accuracy2	Accuracy	MCC	G-mean	Balanced Accuracy
250	Unweighted	0.988	0.315	0.853	0.466	0.557	0.155
	ISNS	0.713	0.733	0.717	0.369	0.723	0.262
	ENS	0.607	0.758	0.637	0.293	0.678	0.230
	LNR+RM	0.661	0.753	0.680	0.336	0.705	0.249
300	Unweighted	0.987	0.331	0.857	0.483	0.572	0.164
	ISNS	0.729	0.790	0.741	0.428	0.759	0.288
	ENS	0.327	0.818	0.425	0.126	0.514	0.134
	LNR+RM	0.722	0.811	0.740	0.438	0.765	0.293
350	Unweighted	0.985	0.342	0.857	0.483	0.581	0.169
	ISNS	0.737	0.792	0.748	0.439	0.764	0.292
	ENS	0.252	0.892	0.380	0.139	0.474	0.112
	LNR+RM	0.719	0.818	0.739	0.441	0.767	0.294
400	Unweighted	0.977	0.425	0.866	0.528	0.644	0.207
	ISNS	0.763	0.788	0.768	0.462	0.775	0.300
	ENS	0.671	0.835	0.704	0.410	0.749	0.280
	LNR+RM	0.735	0.826	0.753	0.463	0.779	0.303
450	Unweighted	0.978	0.437	0.870	0.542	0.654	0.214
	ISNS	0.758	0.828	0.772	0.489	0.793	0.314
	ENS	0.720	0.852	0.746	0.468	0.783	0.307
	LNR+RM	0.739	0.838	0.759	0.477	0.787	0.310
500	Unweighted	0.974	0.508	0.880	0.587	0.703	0.247
	ISNS	0.779	0.870	0.797	0.544	0.823	0.339
	ENS	0.742	0.880	0.770	0.512	0.808	0.327
	LNR+RM	0.767	0.875	0.789	0.535	0.819	0.336

Accuracy1: Accuracy for majority class; Accuracy2: Accuracy for minority class; MCC: Matthews correlation coefficient; G-mean: Geometric mean; ISNS: Inverse of Square Root of Number of Samples; ENS: Effective number of samples; LNR+RM: Least number of ratio and range multiplier.

In Table 4, the results of the methods in different samples are given for the simulated dataset with a distribution of IR=2.33. In the 250-sampled dataset, LNR+RM method provided the best improvement in the performance criterion of the low-sampling class, while ISNS had the best performance based on the general criteria. In the dataset with 300 and 350 samples, the LNR+RM method provided the best improvement in the performance criterion of the low-sampling class, while it had similar performance with ISNS based on the general criteria. In the 400-sampled dataset, LNR+RM had the best performance, followed by ENS and ISNS, respectively. In the dataset with 450 samples, ISNS had the best performance, while LNR+RM and ENS method had similar results. In the dataset with 500 samples, LNR+RM and ENS methods provided the best improvement in the performance criterion of the low-sampling class, while ISNS had the best performance based on the general criteria.

TABLE 4: Results of methods for a simulated dataset with IR=2.33 distribution.

Sample Size	Method	Performance Measures					
		Accuracy1	Accuracy2	Accuracy	MCC	G-mean	Balanced Accuracy
250	Unweighted	0.939	0.538	0.817	0.542	0.711	0.253
	ISNS	0.745	0.840	0.774	0.542	0.791	0.313
	ENS	0.686	0.857	0.738	0.499	0.767	0.294
	New formula	0.687	0.856	0.738	0.499	0.767	0.294
300	Unweighted	0.953	0.624	0.854	0.638	0.771	0.297
	ISNS	0.769	0.850	0.793	0.576	0.809	0.327
	ENS	0.739	0.886	0.783	0.576	0.809	0.327
	New formula	0.760	0.869	0.793	0.583	0.813	0.330
350	Unweighted	0.936	0.653	0.851	0.631	0.782	0.306
	ISNS	0.767	0.905	0.809	0.621	0.833	0.347
	ENS	0.759	0.909	0.804	0.616	0.831	0.345
	New formula	0.761	0.905	0.803	0.614	0.829	0.344
400	Unweighted	0.964	0.737	0.896	0.746	0.843	0.355
	ISNS	0.794	0.896	0.824	0.642	0.843	0.356
	ENS	0.773	0.934	0.821	0.653	0.850	0.361
	New formula	0.778	0.933	0.825	0.658	0.852	0.363
450	Unweighted	0.959	0.799	0.910	0.783	0.875	0.383
	ISNS	0.828	0.920	0.855	0.700	0.872	0.380
	ENS	0.780	0.941	0.828	0.666	0.857	0.367
	New formula	0.785	0.941	0.832	0.672	0.859	0.369
500	Unweighted	0.960	0.813	0.916	0.795	0.883	0.390
	ISNS	0.844	0.926	0.869	0.725	0.884	0.391
	ENS	0.787	0.947	0.835	0.679	0.863	0.373
	New formula	0.812	0.947	0.853	0.706	0.877	0.384

Accuracy1: Accuracy for majority class; Accuracy2: Accuracy for minority class; MCC: Matthews correlation coefficient; G-mean: Geometric mean; ISNS: Inverse of Square Root of Number of Samples; ENS: Effective number of samples.

Using the Information Gain and Gain Ratio variable significance tests, the importance of 15 independent variables in the dataset for the outcome variable was checked. As a result of the 2 tests, the most important variables in common were allergy, alcohol consuming, swallowing difficulty, wheezing, coughing, chest pain, peer pressure, yellow fingers, anxiety and fatigue. Variable significance results are given in Figure 3. Gender, age, smoking, chronic disease and shortness of breath, which are less important than other variables in the dataset, were excluded from the analysis.

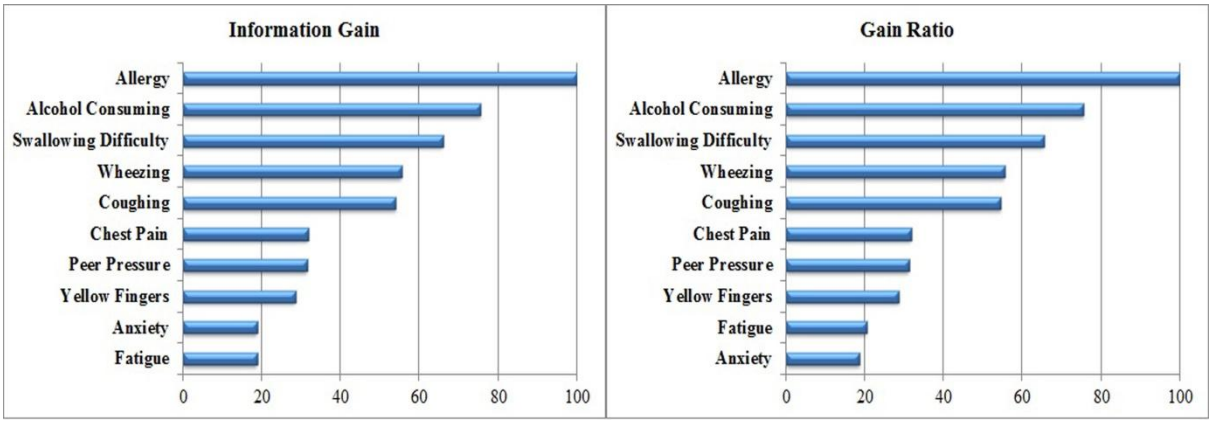

FIGURE 3: Information Gain and Gain Ratio variable importance test results for lung cancer.

Considering the results of the methods for the Lung Cancer dataset in Table 5, the best performance was achieved with the ISNS method. This method was followed by LNR+RM, ENS and unweighted RF methods, respectively.

TABLE 5: Results of methods for the lung cancer dataset.

Sample size	Method	Performance measures					
		Accuracy1	Accuracy2	Accuracy	MCC	G-mean	Balanced Accuracy
309	Unweighted	0.948	0.564	0.899	0.529	0.731	0.267
	ISNS	0.934	0.732	0.908	0.619	0.827	0.342
	ENS	0.941	0.599	0.989	0.539	0.751	0.282
	New formula	0.940	0.675	0.906	0.592	0.796	0.317

Accuracy1: Accuracy for majority class. Accuracy2: Accuracy for minority class; MCC: Matthews correlation coefficient; G-mean: Geometric mean; ISNS: Inverse of Square Root of Number of Samples; ENS: Effective number of samples.

DISCUSSION

In this study, a new formulation was proposed for the class-weighting methods used in solving the class imbalance problem. The ISNS method followed by our new proposal, LNR+RM weighting solution, provided the best performance according to both our 1,000 iterations of simulation and real data results. As the number of samples increased, the weighting method performances approached the unweighted performance, so n=1,000 was not included in the simulation scenario, especially since the class imbalance situation was insignificant in the case of 1,000 samples. Considering the accuracy of the low-sampling class (accuracy 2) and balanced accuracy, unweighted Random Forest gave the poorest performance.

A study proposed a general version of the SMOTE oversampling technique that involves a feature weighting process that considers relevancy/redundancy.[20] Their proposal utilizes the weighted Minkowski distance for exploring the K nearest objects of the minority class. According to the results of 42 real datasets, their suggested feature-weighted-SMOTE method outperformed the other SMOTE types on both high and low-dimensional datasets.

The classification of data with class imbalance attracted attention in medical application as well. Zhu et al. introduced a novel approach by allocating individual weights for each class rather than a single weight.[21] Their proposed method increased the accuracy of the classifier.

The class-weighting method has a different approach compared to pre-existing sampling methods. After proposed to consider the ENS instead of proportional frequency, designed a loss sensitive to label distribution to encourage larger margins for minority classes.[11,13,22] Developed a model in which, instead of generating new samples or ignoring existing samples, they fed composite inputs of the inverse of their true incidence. In a study, 2 class weight ratio plans, namely, INS and ISNS, were searched on the loss function to detect the best technique to handle imbalanced classes.[23] Based on the results, weighting the loss function with INS method produced a slightly better performance compared to ISNS. All three methods were able to improve the F1 score of the minority class from 0.09 to about 0.78-0.81.

The main opinion of re-weighting methods is to appoint weights for various training samples. In an imbalanced dataset, a basis strategy is to weight samples according to the inverse of class size or use inverse square root of class frequency.[24-26] Another study proposed aa easy but efficient loss, called equalization loss, to handle the problem of long-tailed rare categories by simply disregarding the gradients for rare categories.[27] In a study, it has been found that using the ISNS provided nearly the highest F1 score and accuracy

and is therefore accepted as the best weighting method.[28] His study demonstrated that the ISNS weighting formula provided the highest accuracy on the test dataset compared to INS and ENS.

Looking at our performance results; ISNS and LNR+RM gave the highest performances, followed by the ENS, and the unweighted RF, respectively. However, it was observed that the ENS gave lower accuracy and unbalanced performance between the accuracies of the 2 classes, especially in the sample size from 250 to 400. In line with our findings, in Table 2 of, the classification error rates calculated for the CIFAR-10 data (with a sample size of 10) and the CIFAR-100 data (with a sample size of 100) were shown for each of the IRs of ENS ranging from 1 to 200 (1, 10, 20, 50, 100, 200).[11] For example; while the classification error in the CIFAR-10 data was 16.40 at an imbalance rate of 20, this value was found to be 48.57 for the CIFAR-100 data. For each imbalance rate, the classification errors in the data with 10 samples were obtained from $\frac{1}{2}$ to $\frac{1}{5}$ times smaller than the classification errors in the data with 100 samples. On the other hand, for higher training samples (between 5,089 and 8,142); it was shown that the classification error steadily decreases as the number of samples increases. In our simulation results, ENS became stable when the sample size equal to or bigger than 400.

Our formulation, LNR+RM, showed a steady decrease for the patient and control groups as the sample size increased from 250 to 500 in experiments with 1,000 replicates. For 500 samples and IR=9 imbalance condition, the weight of the patient group was 1.820 and that of the control group was 0.468. While these weights were 1.414 and 0.471 respectively in ISNS, they were 3.145 and 0.351 in ENS. Especially, the fact that the weight of the control group in ENS was very low and far from the that of patient group made the formula weak. Additionally, while the patient/control group weights of ENS for 500 samples in an IR of 2.33 were 1.053/0.452, this difference between the weights of the 2 groups was higher than both our formula (0.786/0.407) and the ISNS (0.816/0.535), although the IR decreased. The balance of decrease and increase in LNR+RM weight in both groups was achieved by adding the ratio of majority class ratio ($\sqrt{n_j \times r_{maj}}$) to the denominator as a multiplier.

CONCLUSION

As a result, with our new formula LNR+RM, the accuracy estimates of the 2 classes were obtained in a balanced way for each sample size and for each imbalance rate. Generally, minority class accuracy and balanced accuracy of LNR+RM were either very close to or higher than that of ISNS. These results showed that the LNR+RM method gives consistent results under all conditions and can be used as a new method in the literature, in addition to other methods. As a next study plan, the performances of LNR+RM in different correlation and variable structures can be evaluated with other methods.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Batuhan Bakırarar;* ***Design:*** *Batuhan Bakırarar, Selen Yılmaz Işıkhan;* ***Control/Supervision:*** *Batuhan Bakırarar, Selen Yılmaz Işıkhan;* ***Data Collection and/or Processing:*** *Batuhan Bakırarar;* ***Analysis and/or Interpretation:*** *Selen Yılmaz Işıkhan, Batuhan Bakırarar;* ***Literature Review:*** *Selen Yılmaz Işıkhan;* ***Writing the Article:*** *Selen Yılmaz Işıkhan, Batuhan Bakırarar;* ***Critical Review:*** *Selen Yılmaz Işıkhan;* ***References and Fundings:*** *Batuhan Bakırarar;* ***Materials:*** *Batuhan Bakırarar.*

REFERENCES

1. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6(1):27. [Crossref]
2. Wu Z, Lin W, Ji Y. An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. IEEE Access. 2018;6:8394-8402. [Crossref]
3. Barua S, Islam MM, Yao X, Murase K. MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. IEEE Trans know data Eng. 2012;26(2):405-25. [Crossref]
4. Hassan MM, Huda S, Yearwood J, Jelinek HF, Almogren A. Multistage fusion approaches based on a generative model and multivariate exponentially weighted moving average for diagnosis of cardiovascular autonomic nerve dysfunction. Inf. Fusion. 2018;41:105-18. [Crossref]
5. Han J, Yang Z, Zhang Q, Chen C, Li H, Lai S, et al. A method of insulator faults detection in aerial images for high-voltage transmission lines inspection. Appl Sci. 2019;9(10):2009. [Crossref]
6. Irtaza A, Adnan SM, Ahmed KT, Jaffar A, Khan A, Javed A, et al. An ensemble based evolutionary approach to the class imbalance problem with applications in CBIR. Appl Sci. 2018;8(4):495. [Crossref]
7. Fiore U, De Santis A, Perla F, Zanetti P, Palmieri F. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Inf. Sci. 2019;479:448-55. [Crossref]
8. Tao X, Chen W, Zhang X, Guo W, Qi L, Fan Z. SVDD boundary and DPC clustering technique-based oversampling approach for handling imbalanced and overlapped data. Knowl Based Syst. 2021;234:107588. [Crossref]
9. Mahani A, Ali ARB. Classification problem in imbalanced datasets. Recent Trends Comput Intell. 2019:1-23. [Crossref]
10. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221-32. [Crossref]
11. Cui Y, Jia M, Lin T-Y, Song Y, Belongie S. Class-balanced loss based on effective number of samples. Paper presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. [Crossref]
12. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intel Res. 2002;16:321-57. [Crossref]
13. Asundi RV, Prakash R, Kumar K. Class Weight technique for Handling Class Imbalance. ResearchGate. 18 Temmuz 2022. [Link]
14. Campos Almazán A. Bal images analysis for their automatic quantification [Degree thesis]. Barcelona: Universitat Politècnica de Catalunya; 2021. [Cited: 10.01.2023]. Available from: [Link]
15. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. [Crossref] [PubMed] [PMC]
16. Akosa J. Predictive accuracy: A misleading performance measure for highly imbalanced data. Paper Presented at: Proceedings of the SAS Global Forum. 2017. [Link]
17. Hen H, Garcia E. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263-84. [Crossref]
18. Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, et al. Package 'mass'. Cran R. 2013;538:113-20. [Link]
19. Yarberry W. DPLYR. CRAN Recipes: DPLYR, Stringr, Lubridate, and RegEx in R. 1st ed. Berkeley, CA: Springer; 2021. p.1-58. [Crossref]
20. Maldonado S, Vairetti C, Fernandez A, Herrera F. FW-SMOTE: A feature-weighted oversampling approach for imbalanced classification. Pattern Recognition. 2022;124:108511. [Crossref]
21. Zhu M, Xia J, Jin X, Yan M, Cai G, Yan J, et al. Class weights random forest algorithm for processing class imbalanced medical data. IEEE Access. 2018;6:4641-52. [Crossref]
22. Cao K, Wei C, Gaidon A, Arechiga N, Ma T. Learning imbalanced datasets with label-distribution-aware margin loss. Advances in Neural Information Processing Systems. 2019;32. [Link]
23. Divyanth L, Marzougui A, González-Bernal MJ, McGee RJ, Rubiales D, Sankaran S. Evaluation of effective class-balancing techniques for CNN-based assessment of aphanomyces root rot resistance in pea (Pisum sativum L.). Sensors. 2022;22(19):7237. [Crossref] [PubMed] [PMC]
24. Wang YX, Ramanan D, Hebert M. Learning to model the tail. Adv Neural Inf Process Syst. 2017;30. [Link]
25. Huang C, Li Y, Loy CC, Tang X. Learning deep representation for imbalanced classification. Paper Presented at: Proceedings of the IEEE Conference on Computer Vision And Pattern Recognition. 2016. [Crossref]
26. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural inf Process Syst. 2013;26. [Link]
27. Tan J, Wang C, Li B, Li Q, Ouyang W, Yin C, et al. Equalization loss for long-tailed object recognition. Paper Presented at: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. [Crossref]
28. Jonker J. Weighted convolutional neural networks rare electrocardiogram detection for real-time heart monitoring. 2021. [Link]