Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2023;15(1):19-29

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2022-93961

Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research

Makine Öğrenmesinde Dengesiz Sınıf Problemiyle Başa Çıkmak İçin Sınıf Ağırlıklandırma Tekniği Kullanımı: Metodolojik Araştırma

Batuhan BAKIRARAR[a], Atilla Halil ELHAN[a]

[a]Department of Biostatistics, Ankara University Faculty of Medicine, Ankara, Türkiye

ABSTRACT Objective: Machine learning algorithms are based upon the assumption that data are balanced and so they do not provide good results in imbalanced datasets. This study aimed to explain the methods to be used for fitting a highly accurate model which better classifies the class of interest in imbalanced datasets with the class having a lower number of samples. **Material and Methods:** The study was planned as a methodological research. There are several weighting methods to calculate the class weight. This study included 4 most frequently used weighting methods. These are inverse of number of samples, inverse of square root of number of samples, effective number of examples and sample based class weight methods. In our study, 4 different class weighting methods were used on random forest and support vector machine, and it was explained how those methods affected class-based performances and the overall performance. **Results:** In simulated datasets, the best performance was achieved using the using the inverse of square root of number of samples class weighting method both on random forest and support vector machine. In real dataset, the best performance was achieved using the sample based class weight class weighting method on support vector machine. **Conclusion:** It was seen that all of the class weighting methods used in both machine learning methods were found to increase the performance of the class where recurrence was seen, therefore increasing the overall performance. It has been seen how effective the class weighting method is in dealing with the class imbalance problem.

Keywords: Class weighting; imbalanced class; machine learning

ÖZET Amaç: Makine öğrenmesi algoritmaları, verilerin dengeli olduğu varsayımı altında ve dengesiz veri setlerinde iyi sonuçlar vermez. Bu çalışma, dengesiz veri setlerinde, daha az örnekleme sahip, ilgilenilen sınıfı daha iyi sınıflandıran bir modelin oluşturulması için kullanılacak yöntemleri açıklamayı amaçlamıştır. **Gereç ve Yöntemler:** Çalışma metodolojik bir araştırma olarak planlanmıştır. Sınıf ağırlığını hesaplamak için çeşitli ağırlıklandırma yöntemleri vardır. Bu çalışma, en sık kullanılan 4 ağırlıklandırma yöntemini içermektedir. Bunlar örneklem sayısının tersi, örneklem sayısının karekökünün tersi, efektif örneklem sayısı ve örneklem bazlı sınıf ağırlığı yöntemleridir. Çalışmamızda, random forest ve destek vektör makinesi üzerinde 4 farklı sınıf ağırlıklandırma yöntemi kullanılmış ve bu yöntemlerin sınıf bazlı performansları ve genel performansı nasıl etkilediği açıklanmıştır. **Bulgular:** Simüle edilmiş veri kümelerinde, hem random forest hem de destek vektör makinesi üzerinde örneklem sayısının karekökünün tersi sınıf ağırlıklandırma yöntemi kullanılarak en iyi performans elde edildi. Gerçek veri setinde en iyi performans, destek vektör makinesi üzerinde örneklem bazlı sınıf ağırlığı yöntemi kullanılarak elde edilmiştir. **Sonuç:** Her iki makine öğrenmesi yönteminde kullanılan sınıf ağırlıklandırma yöntemlerinin tamamının, düşük örnekleme sahip sınıfın performansını artırdığı, dolayısıyla genel performansı artırdığı görülmüştür. Çalışma sonuçları, sınıf dengesizliği problemiyle başa çıkmada sınıf ağırlıklandırma yönteminin ne kadar etkili olduğunu göstermiştir.

Anahtar kelimeler: Sınıf ağırlıklandırma; dengesiz sınıf; makine öğrenmesi

Classification in machine learning is based on the principle of estimating the target (dependent) variable over some inputs (independent variables). It is quite possible that the number of subject in each category of class variable is very different. Due to such difference in each class, algorithms tend to be biased towards current majority values and fail to perform well in minority values. That difference in class frequencies impacts general predictiveness of the model.[1,2]

Correspondence: Batuhan BAKIRARAR
Department of Biostatistics, Ankara University Faculty of Medicine, Ankara, Türkiye
E-mail: batuhan_bakirarar@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 21 Oct 2022 ***Received in revised form:*** 17 Nov 2022 ***Accepted:*** 19 Nov 2022 ***Available online:*** 12 Dec 2022

It is not so difficult to achieve a good accuracy in such problems; however, it is not always about obtaining a good prediction performance. We need to check whether performance of these models have any commercial meaning or have any value at all. Thus, it is very important to understand your inquiry and data very well so that you can use the right metrics and optimize it through appropriate methods.[1,3]

Such datasets occur in several real datasets where class distributions of data are highly imbalanced. For an independent variable that has two classes, it is assumed that minority, or rare class is the positive class (e.g. presence of a disease), and majority class is the negative class (e.g. absence of a disease). In general, minority class only has 1% of the dataset. If traditional (cost-insensitive) classifiers are to be implemented, those classifiers will likely predict everything as being negative (majority class). This problem usually arises when learning is performed from highly imbalanced datasets. To put it in simple terms, there is skewness towards the majority class in the data. Class imbalance can be observed in several different areas such as medical diagnosis, spam filtering and detection of fraud.[1,3]

Most of the machine learning algorithms are based upon the assumption that data are balanced, in other words, data are equally distributed among all classes. When training a model over an imbalanced dataset, learning becomes biased towards majority classes. As the number of samples is high, the model learns how to perform well in the class but fails to learn meaningful patterns in minority classes due to having insufficient samples. When there is class imbalance in training data, algorithms usually overestimate the majority class. Consequently, samples from the minority group are more frequently misclassified than those from the majority group.[2,4]

In the case of class imbalance, machine learning algorithms classify the majority class well. When looking at the overall performance using performance criteria that provide misleading results in imbalanced distribution such as accuracy, it can also be concluded that a good model has been obtained. Since overall performance is determined by calculating the weighted average of the classes, it is more affected by results of the majority class, and such result may be misleading for the researcher. However, when assessing the class-based results, it is seen that the class with lower number of samples do not meet performance criteria sufficiently.[1]

Hence, it is important to consider the representation of minority and majority classes when learning from the imbalanced database. It was argued by Krawczyk that good results can be obtained independently from the imbalance of the class if both groups are well-represented and come from non-overlapping distributions.[5] Japkowicz and Shah examined the effects of class imbalance by creating artificial datasets through several combinations of complexity, training set sizes, and degrees of imbalance.[6] The results show that sensitivity to imbalance increases with increased complexity of problem and non-complex, linearly separable problems are not affected by any of the class imbalance levels.[6]

Especially in medicine, there is a data imbalance due to low incidence of a given event (disease). It is of great importance to learn from majority class's data with extreme class imbalance, which represent about 10% of training data because these are usually the rare events of interest.

There are many methods used to prevent class imbalance in the literature. The most frequently used methods to address class imbalance is undersampling for majority classes and oversampling for minority samples. When undersampling is performed for majority class, a certain number of examples from the majority class are omitted from the dataset. Examples from the minority classes are duplicated when oversampling is implemented for the minority class.[7]

Even though either of the two strategies balances the dataset, they cannot directly solve the problems caused by class imbalance; instead, they even pose the risk of causing new problems. Since oversampling presents iterative examples, it may slow down the training, resulting in overfitting in the model. On the other hand, examples concluded from undersampling may lead to deficiency in learning some of the important concepts.[7]

The best method to implement for overcoming such problems is class weighting. In this method, different weights are applied to classes in the dataset depending on the number of samples. That is to say, the class with smaller sample is less weighted and the class with larger sample is more weighted to try to eliminate data imbalance.[7]

Most machine learning algorithms do not provide good results in imbalanced datasets. Yet, the training algorithm in use can be changed to take the skewed distribution of classes into consideration. This can be achieved by weighting both majority and minority classes differently. Difference in the weights affect their classification performance in the training stage. The purpose here is to mitigate the misclassification performed by the minority class by designating a higher class weight to the class, and for the majority class, to penalize this class by lowering its class weight.[3,4]

This study aimed to explain the methods to be used for fitting a highly accurate model which better classifies the class of interest in imbalanced datasets with the class having a lower number of samples. With models to be fitted through these methods, one can create decision support systems and use them in real life problems (e.g. disease diagnosis and prediction at hospitals).

MATERIAL AND METHODS

The study was conducted in accordance with the principles of the Declaration of Helsinki. The study was planned as a methodological research and the details of the class weighting methods used in the study are given below.

WEIGHTING METHODS

There are several weighting methods to calculate the class weight. This study included 4 most frequently used weighting methods. Using Inverse of number of samples (INS) and inverse of square root of number of samples (ISNS) as weights are two of the simplest and most popular weighting schemes. The third method in the study is a rather recent weighting method known as effective number of examples weighting scheme. The last method is the sample based class weight (SBCW) method that utilizes example-dependent weight classes.

USING THE INS AS WEIGHT

In this method, examples are weighted based on the inverse of class frequency for the class they belong to; hence the name.[8]

$$W = 100 * \frac{1}{Number\ of\ Samples\ in\ Class}$$

W: Sample weighting.

USING THE ISNS AS WEIGHT

In this method, examples are weighted based on the inverse of square root of class frequency for the class they belong to.[9]

$$W = 10 * \frac{1}{\sqrt[2]{Number\ of\ Samples\ in\ Class}}$$

W: Sample weighting.

EFFECTIVE NUMBER OF SAMPLES

The main idea here is to associate each example with a small neighboring region instead of a single point because benefit of a newly added data will diminish with higher number of samples. With this approach, each example is associated with a small neighboring region, and weight is calculated based on the size of data overlap. Effective number of samples (ENS) is defined as the sample volume and can be calculated with the formula below.[10]

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

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

$$W = 100 * \frac{1}{ENS}$$

where n: Number of samples in class; ENS: Effective number of samples; W: Sample weighting.

SBCW

In this method, class weighting is performed based on the examples in the study, and class weights are calculated using the formula below.[11]

$$\mathrm{W} = \mathrm{N}/(\mathrm{class} * \mathrm{sample})$$

where N: Total number of samples in the dataset; Class: Total number of unique classes in the dataset; Sample: Total number of samples of the respective class.

PERFORMANCE CRITERIA

In general, one will want to strike a balance between false positive and false negative rates in imbalanced datasets. Accuracy cannot provide a good assessment in such cases. The following are the most frequently used performance criteria in the literature that attempt to balance between false positives and false negatives.[12]

GEOMETRIC MEAN

Geometric mean (G-mean) is a criterion which measures the balance between the classification performances in majority and minority classes. A low G-mean is the indicator of poor performance in the classification of positive cases even if the negative cases have been correctly classified. This measure is important for determining the classification with low negative class and high positive class.[12]

$$G - Mean = \sqrt{Sensitivity * Specificity}$$

BALANCED ACCURACY

Balanced accuracy is the mean between sensitivity and selectivity that measures the mean accuracy obtained from both minority and majority classes. If a classifier performs equally well in both classes, this measure yields a similar result as accuracy. However, if high accuracy value is due to classifier's good classification of the majority class, balanced accuracy value is lower than the accuracy measure.[4,12]

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

MATTHEWS CORRELATION COEFFICIENT

Matthews correlation coefficient (MCC) is the measure least affected by imbalanced data. It is a correlation coefficient between the observed and predicted classes. Its value varies between -1 and +1. Whereas a value of +1 refers to a perfect prediction, a value of 0 refers to a random prediction, and a value of -1 represents the worst possible prediction.[4,12]

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

DATASETS

Two simulated datasets with 250 samples consisting of 6 independent and one dependent variables were created. The first data set is derived so that the dependent variable has a distribution of 70% no and 30% yes, and the second dataset has a distribution of 80% no and 20% yes for the dependent variable.

The real dataset used in the study was retrieved from Ljubljana University, Medical Center, Institute of Oncology and included 286 patient data. The dataset is comprised of a total of ten variables, with nine of them being independent variables and one of them being dependent variable.

STATISTICAL ANALYSIS

Random Forest and WeightSVM packages (https://CRAN.R-project.org/package=WeightSVM) in the R programming language were used for the analysis. The number of patients (percentage) was used as descriptors for qualitative variables. Random Forest and Support Vector Machine (SVM) were used as a classification methods. The dataset was evaluated using the 10-fold cross validation test option and all analyzes were repeated 1,000 times. Accuracy, MCC, G-mean and balanced accuracy were used as machine learning performance criteria.

RESULTS

SIMULATED DATASETS

Weights for the 4 methods used in class weighting for simulated datasets in the study are shown in Table 1. As seen in the table, weights of the class with larger sample are lower while weights of the class with smaller sample are higher.

TABLE 1: Class weights for methods.

Datasets	**Methods**	**Class**	
		No	**Yes**
80%-20% distribution dataset	INS	0.500	2.000
	ISNS	0.707	1.414
	ENS	0.790	3.145
	SBCW	0.625	2.500
70%-30% distribution dataset	INS	0.571	1.333
	ISNS	0.756	1.155
	ENS	0.902	2.101
	SBCW	0.714	1.667
Real dataset	Methods	Class	
		No recurrence events	Recurrence events
	INS	0.500	1.200
	ISNS	0.700	1.080
	ENS	0.700	1.300
	SBCW	0.711	1.682

INS: Inverse of number of samples; ISNS: Inverse of square root of number of samples; ENS: Effective number of samples; SBCW: Sample based class weight.

Random forest and SVM of machine learning methods with the feature of class weight were utilized in the study. In the random forest method, number of trees was chosen to be 1,000, the analyses for were established so as to consist of 1,000 iterations, and the calculated performance criteria were averaged, which is presented in Table 2.

TABLE 2: Random forest performance results based on class weights for simulated datasets.

Datasets	Methods	Class	Accuracy	MCC	G-mean	Balanced accuracy
80%-20% distribution dataset	No method	No	0.957	0.525	0.682	0.233
		Yes	0.487			
		Overall	0.863			
	INS	No	0.764	0.546	0.826	0.341
		Yes	0.893			
		Overall	0.790			
	ISNS	No	0.813	0.590	0.845	0.357
		Yes	0.878			
		Overall	0.826			
	ENS	No	0.765	0.547	0.827	0.342
		Yes	0.893			
		Overall	0.791			
	SBCW	No	0.764	0.545	0.826	0.341
		Yes	0.892			
		Overall	0.790			
70%-30% distribution dataset	No method	No	0.951	0.599	0.743	0.276
		Yes	0.580			
		Overall	0.839			
	INS	No	0.812	0.570	0.800	0.320
		Yes	0.789			
		Overall	0.805			
	ISNS	No	0.852	0.610	0.813	0.331
		Yes	0.777			
		Overall	0.829			
	ENS	No	0.812	0.571	0.801	0.320
		Yes	0.789			
		Overall	0.805			
	SBCW	No	0.812	0.571	0.800	0.320
		Yes	0.789			
		Overall	0.805			

MCC: Matthews correlation coefficient; G-mean: Geometric mean; INS: Inverse of number of samples; ISNS: Inverse of square root of number of samples; ENS: Effective number of samples; SBCW: Sample based class weight.

Considering the results obtained without using the class weighting methods, accuracy was low in the class where recurrence was seen whereas those values were found to be high in the class where recurrence was not seen and overall. The reason is that the overall performance was calculated via weighted averaging and the results were closer to the class with larger sample. As for the performance criteria with class weighting methods implemented, a decrease was observed in the performance of the class with larger sample, and an increase was found in the performance of the class with smaller sample. Regarding the criteria of MCC, G-mean and balanced accuracy which are suggested for balanced datasets, there was an overall increase. The class weighting method with the best performance for random forest was found to be ISNS in 80%-20% and 70%-30% distribution datasets. According to this method, in 80%-20% distribution dataset accuracy values were found to be 0.813 in the no response, 0.878 in the yes response and overall accuracy value was found to be 0.826. In 70%-30% distribution dataset accuracy values were found to be 0.852 in the no response, 0.777 in the yes response and overall accuracy value was found to be 0.829. When compared to the results obtained without using any of the class weighting methods, increases were observed in the values of accuracy and in the overall performance values of MCC, G-mean and balanced accuracy.

For the SVM method, as with the random forest method, analyses were performed by forming cycles so as to consist of 1,000 iterations, and the performance criteria were average; the results are shown in Table 3. The class weighting method with the best performance for SVM was found to be ENS in 80%-20% distribution dataset and ISNS in 70%-30% distribution dataset. According to this method, in 80%-20% distribution dataset accuracy values were found to be 0.810 in the no response, 0.880 in the yes response and overall accuracy value was found to be 0.824. In 70%-30% distribution dataset accuracy values were found to be 0.937 in the no response, 0.693 in the yes response and overall accuracy value was found to be 0.864. When compared to the results obtained without using any of the class weighting methods, increases were observed in the values of accuracy and in the overall performance values of MCC, G-mean and balanced accuracy.

TABLE 3: SVM performance results based on class weights for simulated datasets.

Datasets	Methods	Class	Accuracy	MCC	G-mean	Balanced accuracy
80%-20% distribution dataset	No method	No	0.965	0.605	0.735	0.270
		Yes	0.560			
		Overall	0.884			
	INS	No	0.840	0.596	0.840	0.353
		Yes	0.840			
		Overall	0.840			
	ISNS	No	0.840	0.565	0.820	0.336
		Yes	0.800			
		Overall	0.832			
	ENS	No	0.810	0.588	0.844	0.356
		Yes	0.880			
		Overall	0.824			
	SBCW	No	0.830	0.597	0.845	0.357
		Yes	0.860			
		Overall	0.836			
70%-30% distribution dataset	No method	No	0.977	0.604	0.713	0.254
		Yes	0.520			
		Overall	0.840			
	INS	No	0.777	0.529	0.782	0.306
		Yes	0.787			
		Overall	0.780			
	ISNS	No	0.937	0.665	0.806	0.325
		Yes	0.693			
		Overall	0.864			
	ENS	No	0.777	0.541	0.788	0.311
		Yes	0.800			
		Overall	0.784			
	SBCW	No	0.777	0.541	0.788	0.311
		Yes	0.800			
		Overall	0.784			

SVM: Support vector machine; MCC: Matthews correlation coefficient; G-mean: Geometric mean; INS: Inverse of number of samples; ISNS: Inverse of square root of number of samples; ENS: Effective number of samples; SBCW: Sample based class weight.

REAL DATASET

Of the patients enrolled in the study, 12.9% were below 40 years of age whereas 31.5% were in the age range of 40-49 years, 33.6% in the age range of 50-59 years, 19.9% in the age range of 60-69 years, and 2.1% were at the age of 70 years and above. 52.4% of the women were in the pre-menopause period whereas menopause age of 2.4% was <40 years and menopause age of 45.2% was ≥40 years. While 4.2% of the patients had tumors smaller than 1 cm, 20.3% had tumors with sizes between 1.00-1.99 cm, 36.4% had tumors with sizes between 2.00-2.99 cm, 27.6% had tumors with sizes between 3.00-3.99 cm, 8.7% had tumors with sizes between 4.00-4.99 cm, and 2.8% had tumors with a size of 5 cm and above. 23.8% of the patients had received radiotherapy and 29.7% of them had experienced recurrence. Identifiers for other data are given in Table 4.

TABLE 4: Demographic variables for real dataset.

Variables		n	%
Age (years)	<40	37	12.9
	40-49	90	31.5
	50-59	96	33.6
	60-69	57	19.9
	≥70	6	2.1
Menopause status	Pre-menopause	150	52.4
	<40	7	2.4
	≥40	129	45.2
Tumor size (cm)	<1.00	12	4.2
	1.00-1.99	58	20.3
	2.00-2.99	104	36.4
	3.00-3.99	79	27.6
	4.00-4.99	25	8.7
	≥5.00	8	2.8
Inv nodes	0-2	213	74.5
	3-5	36	12.6
	6-8	18	6.3
	9-11	10	3.5
	12-14	3	1.0
	15-17	6	2.1
Node caps	No	222	79.9
	Yes	56	20.1
Degree of malignancy	1	71	24.8
	2	130	45.5
	3	85	29.7
Breast side	Left	152	53.1
	Right	134	46.9
Breast quadrant	Left low	110	38.7
	Left up	97	33.9
	Right low	24	8.4
	Right up	33	11.6
	Central	21	7.4
Irradiation	No	218	76.2
	Yes	68	23.8
Recurrence	No	201	70.3
	Yes	85	29.7

Weights for the 4 methods used in class weighting for the presence of recurrence in the study are shown in Table 1. As seen in the table, weights of the class with larger sample are lower while weights of the class with smaller sample are higher.

Random forest and SVM of machine learning methods with the feature of class weight were utilized in the study. In the random forest method, number of trees was chosen to be 1,000, the analyses for were established so as to consist of 1,000 iterations, and the calculated performance criteria were averaged, which is presented in Table 5.

TABLE 5: Random forest and support vector machine performance results based on class weights for real dataset.

Algorithms	Methods	Recurrence	Accuracy	MCC	G-mean	Balanced accuracy
Random forest	No method	No	0.893	0.293	0.561	0.157
		Yes	0.352			
		Overall	0.733			
	INS	No	0.817	0.281	0.609	0.186
		Yes	0.455			
		Overall	0.709			
	ISNS	No	0.836	0.299	0.611	0.187
		Yes	0.447			
		Overall	0.720			
	ENS	No	0.827	0.293	0.612	0.187
		Yes	0.453			
		Overall	0.716			
	SBCW	No	0.819	0.281	0.608	0.185
		Yes	0.452			
		Overall	0.710			
Support vector machine	No method	No	0.985	0.370	0.481	0.116
		Yes	0.235			
		Overall	0.762			
	INS	No	0.786	0.396	0.700	0.245
		Yes	0.624			
		Overall	0.738			
	ISNS	No	0.905	0.370	0.611	0.186
		Yes	0.412			
		Overall	0.759			
	ENS	No	0.841	0.402	0.682	0.232
		Yes	0.552			
		Overall	0.755			
	SBCW	No	0.781	0.422	0.717	0.257
		Yes	0.659			
		Overall	0.745			

MCC: Matthews correlation coefficient; G-mean: Geometric mean; INS: Inverse of number of samples; ISNS: Inverse of square root of number of samples; ENS: Effective number of samples; SBCW: Sample based class weight.

The class weighting method with the best performance for random forest was found to be ENS. According to this method, accuracy values were found to be 0.827 in the class where recurrence was not seen, 0.453 in the class where recurrence was seen, and overall accuracy value was found to be 0.716. When compared to the results obtained without using any of the class weighting methods, increases were observed in the values of accuracy and in the overall performance values of MCC, G-mean and balanced accuracy.

For the SVM method, as with the random forest method, analyses were performed by forming cycles so as to consist of 1,000 iterations, and the performance criteria were average; the results are shown in Table 5. Regarding the results obtained for the SVM method without using any class weighting methods, accuracy was very low in the class where recurrence was seen while those values were found to be high in the class where recurrence was not seen. As for the results obtained implementing the class weighting methods, an increase was observed in the performance of the class with smaller sample as desired. Looking at the overall performance criteria using the measures MCC, G-mean and balanced accuracy, the class weighting method with the best performance for SVM was found to be SBCW. According to this method, accuracy values were found to be 0.781 in the class where recurrence was not seen, 0.659 in the class where recurrence was seen, and overall accuracy value was found to be 0.745.

Given these criterias, SVM seems to have classified the class with smaller sample size better and corrected it better. It is recommended to use this method where the distribution between 2 classes is as different as 80%-20%.

DISCUSSION

Class imbalance is a common problem in health field. In this study, the methods used in the literature to solve this problem were mentioned and it was shown which of these methods had the best results and the most effective on classification performance. Although the performance of the methods used varies from data to data, class weighting is detected as the most appropriate method to deal with class imbalance.

Hinners et al. applied machine learning techniques in an attempt to predict and classify stellar properties in their study. They applied synthetic minority oversampling technique (SMOTE) to their minority class and this method improved their classification results. For the random forest method they used in the study, they achieved balanced accuracy values of 69.94% and 74.70% before and after implementing SMOTE, respectively.[13]

Hashemi and Karimi used Bayesian, least squares, SVM, decision tree, perceptron, and multilayer perceptron (MLP) methods in their study with breast cancer data. They compared the results obtained before and after implementing class weight for all methods and showed that class weighting increased the accuracy values for all methods. They found pre- and post-class weighting accuracy values of 46.71% and 76.72% for Bayesian method, 45.39% and 92.39% for least-squares, 51.24% and 64.86% for SVM, 48.17% and 89.90% for decision tree, 51.10% and 93.12% for perceptron, and 55.64% and 89.02% for MLP, respectively.[14]

Zong et al. applied both unweighted and weighted versions of extreme learning machine (ELM) method on multiple datasets. They showed that weighted ELM method yielded better results compared to unweighted ELM method for all datasets.[15]

Bedi et al. used machine learning models to predict groundwater quality assessment. They utilized the methods of SVM, Extreme Gradient Boosting (XGB), artificial neural networks, and logistic regression. They also used the methods of oversampling, class weighting, and their combination for the problem of data imbalance. They reported that the best result was achieved by implementing the combination of oversampling and class weighting along with XGB algorithm.[16]

In our study, 4 different class weighting methods were used on random forest and SVM, and it was explained how those methods affected class-based performances and the overall performance. The best performance was achieved using the SBCW class weighting method on SVM in real dataset and using the ISNS method on random forest in 70%-30% distribution dataset.

CONCLUSION

It has been seen that all of the class weighting methods used in both machine learning methods were found to increase the performance of the class where recurrence was seen, therefore increasing the overall performance and how effective the class weighting method is in dealing with the class imbalance problem. Machine learning methods, which have become increasingly popular in recent years, can be more reliable and have higher classification performance by using them together with class weighting methods in case of class imbalance. Thus, machine learning can be used more frequently in real application areas.

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:** Atilla Halil Elhan, Batuhan Bakırarar; **Design:** Atilla Halil Elhan, Batuhan Bakırarar; **Control/Supervision:** Batuhan Bakırarar; **Data Collection and/or Processing:** Batuhan Bakırarar; **Analysis and/or Interpretation:** Batuhan Bakırarar; **Literature Review:** Batuhan Bakırarar; **Writing the Article:** Atilla Halil Elhan, Batuhan Bakırarar; **Critical Review:** Atilla Halil Elhan.*

REFERENCES

1. Sammut C, Webb GI. Encyclopedia of Machine Learning and Data Mining. 2nd ed. New York: Springer Publishing Company, Incorporated; 2017. [Crossref]
2. Solon G, Haider SJ, Wooldridge JM. What are we weighting for?. J Hum Resour. 2015;50(2):301-16. [Crossref]
3. He J, Cheng MX. Weighting methods for rare event identification from imbalanced datasets. Front Big Data. 2021;4:715320. [Crossref] [PubMed] [PMC]
4. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6(1):1-54. [Crossref]
5. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence. 2016;5(4):221-32. [Crossref]
6. Japkowicz N, Shah M. Performance evaluation in machine learning. In: El Naqa I, Li R, Murphy MJ, eds. Machine Learning in Radiation Oncology. 1st ed. Cham: Springer; 2015. p.41-56. [Crossref]
7. Anand A, Pugalenthi G, Fogel GB, Suganthan PN. An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids. 2010;39(5):1385-91. [Crossref] [PubMed]
8. Wang J, Zhang L, Cao JJ, Han D. NBWELM: naive Bayesian based weighted extreme learning machine. Int J Mach Learn Cybern. 2018;9(1):21-35. [Crossref]
9. Campos Almazán A. Bal images analysis for their automatic quantification [Degree thesis]. Spain: Universitat Politècnica de Catalunya; 2021. [08.11.2022]. [Link]
10. Cui Y, Jia M, Lin TY, Song Y, Belongie S. Class-balanced loss based on effective number of samples. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019;9268-77. [Crossref]
11. Analytics Vidhya [Internet]. © Copyright 2013-2022 Analytics Vidhya [Cited: October 20, 2022]. How to Improve Class Imbalance using Class Weights in Machine Learning. 2020. Available from: [Link]
12. Akosa J. Predictive accuracy: A misleading performance measure for highly imbalanced data. Proceedings of the SAS Global Forum. 2017;12. [Link]
13. Hinners TA, Tat K, Thorp R. Machine learning techniques for stellar light curve classification. Astron J. 2018;156(1):7. [Crossref]
14. Hashemi M, Karimi H. Weighted machine learning. Stat Optim Inf Comput. 2018:6(4):497-525. [Crossref]
15. Zong W, Huang GB, Chen Y. Weighted extreme learning machine for imbalance learning. Neurocomputing. 2013;101(1-3):229-42. [Crossref]
16. Bedi S, Samal A, Ray C, Snow D. Comparative evaluation of machine learning models for groundwater quality assessment. Environ Monit Assess. 2020;192(12):776. [Crossref] [PubMed]