

Article

Extreme Sample Imbalance Classification Model Based on Sample Skewness Self-Adaptation

Jie Xue *  and Jinwei Ma * 

Department of Statistics, School of Economics, Hangzhou Dianzi University, Hangzhou 310018, China

* Correspondence: xjsnow@hdu.edu.cn (J.X.); 202150064@alu.hdu.edu.cn (J.M.)

Abstract: This paper aims to solve the asymmetric problem of sample classification recognition in extreme class imbalance. Inspired by Krawczyk (2016)'s improvement direction of extreme sample imbalance classification, this paper adopts the AdaBoost model framework to optimize the sample weight update function in each iteration. This weight update not only takes into account the sampling weights of misclassified samples, but also pays more attention to the classification effect of misclassified minority sample classes. Thus, it makes the model more adaptable to imbalanced sample class distribution and the situation of extreme imbalance and make the weight adjustment in hard classification samples more adaptive as well as to generate a symmetry between the minority and majority samples in the imbalanced datasets by adjusting class distribution of the datasets. Based on this, the imbalance boosting model, the Imbalance AdaBoost (ImAdaBoost) model is constructed. In the experimental design stage, ImAdaBoost model is compared with the original model and the mainstream imbalance classification model based on imbalanced datasets with different ratio, including extreme imbalanced dataset. The results show that the ImAdaBoost model has good minority class recognition recall ability in the weakly extreme and general class imbalance sets. In addition, the average recall rate of minority class of the mainstream imbalance classification models is 7% lower than that of ImAdaBoost model in the weakly extreme imbalance set. The ImAdaBoost model ensures that the recall rate of the minority class is at the middle level of the comparison model, and the F1-score comprehensive index performs well, demonstrating the strong stability of the minority class classification in extreme imbalanced dataset.

Keywords: ImAdaBoost model; extreme imbalance classification; skewness adjustment function; asymmetric problem



Citation: Xue, J.; Ma, J. Extreme Sample Imbalance Classification Model Based on Sample Skewness Self-Adaptation. *Symmetry* **2023**, *15*, 1082. <https://doi.org/10.3390/sym15051082>

Academic Editor: José Carlos R. Alcantud

Received: 6 April 2023
Revised: 4 May 2023
Accepted: 11 May 2023
Published: 14 May 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Symmetry has played and will continue to play a very important role in the world. Over the years, there has been a growing focus on detecting and exploiting symmetry and asymmetry in all aspects of theoretical and applied computing. Several studies involving symmetry and asymmetry have been published in Internet technology, machine learning, data analysis, and many other applications [1]. In the field of machine learning classification, asymmetry is widely involved. Bejjanki (2020) proposed a novel class imbalance reduction (CIR) algorithm to create a symmetry between the defect and non-defect records in the imbalanced datasets by considering distribution properties of the datasets [2]. Aiming at the asymmetry problem in customer credit identification, a novel sample-based online learning ensemble (SOLE) model for client credit assessment is proposed [3]. Faced with the problem of imbalanced multi-classification in text dataset, Li (2022) used an asymmetric cost-sensitive support vector machine to generate balanced dataset [4].

In essence, class imbalance is a common asymmetric problem in machine learning and has been widely studied. It mainly refers to the asymmetric fact that the number of minority class samples is much smaller than that of majority class samples [5]. There is a common assumption in most classification learning algorithms, that is the number of training

samples is the same for different classes. The small difference in the number of classes has a weak impact on the generalization performance of the learned classification model. Yet the large difference in number of classes can influence the learning process. Specifically, due to a general classification algorithm with the same loss weights for each sample in its objective function, the classification algorithm aims to maximize the classification accuracy when the samples are imbalanced. As a result, the classifier will favor the majority class samples and will sacrifice the recognition performance of the minority class, resulting in a model with good recognition for the majority class and poor recognition for the minority class. In addition, the minority class samples are easily recognized as noise by the anomaly detection model, resulting in the loss of sample information.

The main difficulty of imbalance samples classification is that when the traditional classification algorithm is directly applied to imbalanced samples, the global classification accuracy is very high, but the classification recall of minority class samples is very low. Related studies point out that the loss function of traditional machine learning algorithm, the relative scarcity of samples, noise, boundary overlap and so on are the reasons of classification difficulties [6]. Among them, inappropriate evaluation criteria and data sparsity are significant causes of classification difficulties. On the one hand, traditional classification algorithms usually assume class balance [7]. When the classes are imbalanced, the learning process tends to sacrifice the recognition performance of the minority class samples due to the sparsity of the minority class samples. On the other hand, when the data set is relatively sparse, using traditional machine learning will often sacrifice the recognition performance of the minority class samples [8]. Meanwhile, absolute sparsity of the dataset can also lead to classification difficulties which will make the minority class samples unrepresentative of the entire minority class distribution and the classifier will fail to learn the relevant information. In addition, some of the noise filtering methods consider minority classes as noise, and the number of minority classes will be more sparse after removing them [9]. Moreover, it is difficult to distinguish minority classes from noise. Therefore, the noisy data are included in the training process, resulting in the fact that some true minority classes cannot be trained well. In addition, boundary overlap refers to the overlap of sample spaces of different classes. In this state, because of the class imbalance, the traditional classifiers extend the boundaries of the majority class to the minority class in order to seek to minimize the global error rate, resulting in the misclassification of the minority class on the boundary into the majority class [10].

The asymmetric problem of imbalanced classification of data widely exists in many fields such as industrial production, finance, information security, etc. There are also significant differences in sample imbalance ratio in different fields. Krawczyk (2016) [11] pointed out that the ratios of imbalance classes in current studies range from 1:4 to 1:100, failing to classify extreme class imbalance. However, fraud detection in real-life production corresponds to imbalance ratios of 1:100 or more. This poses a new challenge for data preprocessing and classification algorithms to adapt to extremely imbalanced data types.

The contributions of this paper are proposed an ImAdaBoost model with the help of the feature of its dynamic adjustment of sample weights to handle the imbalanced problem and extremely imbalanced problem. These contributions include: (1) This paper attempts to verify the validity of the ImAdaBoost classification model based on self-adaptive sample skewness by using different degrees of imbalanced datasets including extreme imbalanced datasets. (2) This model introduces the proportion parameters of majority class samples and minority class samples in the samples with classification errors in one round of iteration. By using the adaptive function of AdaBoost model, namely, its sample weight update function, its corresponding weight distribution are adjusted to adaptively strengthen the next round weights of minority class samples that are difficult to classify with classification errors. (3) This method is based on the AdaBoost model, which mainly refers to the sample weight adjustment function, and improves the problem of giving the same sample weight to the minority class samples and the majority class samples with classification errors. Finally, the convergence of the improved imbalance classification model is proved theoretically.

The remainder of this paper is organized as follows. Section 2 mainly discusses various classification methods in the imbalanced dataset, and further leads to the problems encountered in the extreme imbalance problem that the academic community is concerned about and the related algorithm research published. Based on this, an advanced method based on the Boosting framework is proposed. Section 3 focuses on describing the ImAdaBoost algorithm and obtaining upper bounds on the error of its training sample misclassification to justify the correction of the weight adjustment function. Section 4, based on the optimal base model, firstly, the law and optimal values of the weight adjustment factor c in the weight adjustment function on the classification effect of the final model are explored to investigate its generality. Then, the paper experimentally compares the classification effects of ImAdaBoost and AdaBoost, as well as the commonly used imbalance processing algorithms for extreme class imbalanced datasets. Section 5 presents conclusions and prospects.

2. Related Work and the Proposal of ImAdaBoost Algorithm

2.1. The Solution to Class Imbalance

The traditional methods to solve class imbalance can be divided into data-level methods and algorithm-level methods. (1) Data-level approaches mainly focus on balancing the number of majority and minority classes from the data distribution level, followed by training and learning the model. It mainly includes undersampling of majority samples (Random Under-Sampling, RUS), oversampling of minority samples (Random Over-Sampling, ROS) [12], hybrid sampling [13], and using various sample synthesis techniques to achieve sample rebalancing, such as SMOTE [14], ADASYN [15], etc. However, some data may be lost after rebalancing the sample data at the data level such as RUS, leading to information loss or changing the true distribution of the dataset. This may affect the classification authenticity. Therefore, after training and learning the rebalanced data, the original dataset is used for cross-validation to check the classification effect of the model. SMOTE did not resolve the issue of boundary overlap between classes, although SMOTE could correct the balance on majority and minority class distribution, and resolving the issue of classification on class boundaries was the next development in the design of this class of sample synthesis model. ADASYN did not take the initiative to find boundary classes, but find them according to the normalized proportion of majority samples in the nearest neighbor samples of each minority sample and determine the number of synthesized minority samples based on the proportion, thus taking the boundary sample distribution into account; (2) Meanwhile, there is no concern about information loss at the algorithm level. The methods mainly include cost-sensitive learning methods and ensemble learning methods. The former improves the recognition effect of the model for minority classes by adding a penalty factor for minority class recognition errors to the model loss function. The introduction of the surrogate value into the loss function can adjust the loss after misclassification of each class in imbalanced samples. The loss cost of minority class samples is generally set higher than majority class samples to achieve the purpose of balancing the number of samples [16,17]. The latter deals with the problem of imbalanced samples by combining data-level sampling methods with an integrated learning framework. Among the most common frameworks for integrated learning are boosting and bagging methods, such as RUSBoost [18], UnderBagging [19], BalanceCascade [20], etc. Table 1 provides a categorization of traditional imbalance methods used for comparison. In addition, according to the previous analysis, the strategies and basic ideas used by the respective methods is summarized and introduced in Table 1.

Table 1. Traditional methods to imbalanced dataset.

Category	Strategy	Method	Basic Idea of the Methods
Data Level	Under-Sampling	RUS	This method proposes to use random undersampling to rebalance the imbalanced training data.
		ROS	This method proposes to use random oversampling to rebalance the imbalanced training samples.
	Over-Sampling	SMOTE	This method proposes to synthesize artificial minority class samples to rebalance the imbalanced training data.
		ADASYN	The method determines the synthetic quantity of each sample quantitatively and solve the problem of sampling proportion.
Algorithm Level	Cost-Sensitive	CostDT	The method combines cost value with decision tree.
	Ensemble	RUSBoost	This method introduces the RUS sampling method into Boosting.
		UnderBagging	This method proposes random undersampling based on Bagging.
		BalanceCascade	This method uses sampling without replacement instead of RUS.

2.2. The Extreme Class Imbalance

For data preprocessing and classification algorithms to adapt to extremely imbalanced data types, three major difficulties are proposed as follows [11]: (1) Minority class tends to perform poorly and lack clear structure in such highly imbalanced situations. Therefore, direct application of preprocessing methods that rely on relationships between a few objects (e.g., SMOTE) can actually worsen classification performance. It seems to be a promising direction that methods can assign weights to a minority class and predict or reconstruct the underlying class structure. (2) The second point is to decompose the original problem into a set of sub-problems characterized by a reduced imbalance rate, and then use the canonical algorithm. Yet this approach considers it from an algorithmic point of view, namely, it requires the development of algorithms for multi-layer partitioning of sample classes and algorithms for reconstructing the original extreme imbalance task. (3) It is also a valid direction to effectively extract the problem features. In extreme class imbalance, internet transaction data or protein data features are often accompanied by high dimensionality and sparsity, extracting effective features from the feature extraction perspective. This improves the model's ability to identify minority class samples under extreme imbalance.

Aiming to extremely imbalanced dataset classification problem, Sharma(2018) holds that popular oversampling methods ignore majority class samples and lose a global perspective on the imbalanced dataset classification problem [21]. It may negatively affect the ability to learn by generating bounded or overlapping instances while mitigating class imbalance. This becomes even more critical when faced with extreme imbalances, where the representation of minority class is seriously insufficient and does not contain sufficient information on its own to conduct the oversampling process. Based on these issues, Sharma proposes a new synthetic oversampling method, which uses the rich information inherent in the majority class to

synthesize minority class data, that is, by generating synthetic data that has the same Mahalanobis distance between the known minority class and the majority class.

In recent years, scholars have paid extensive attention to solving the problem of asymmetric classification based on boosting model framework. Ge (2008) not only thoroughly discussed the various discrete asymmetric AdaBoost algorithms and their relationships based on the three different upper bounds of the asymmetric training error, but also deduced the real-valued asymmetric boosting algorithms with analytical solutions in the form of additive logistic regression [22]. Sun et al. (2020) [23], Wang et al. (2021) [24] and other scholars have used the AdaBoost modeling framework to construct models such as ADASVM-TW and Reinforced AdaBoost to solve imbalanced classification problems with temporal features or other characteristics in the last three years. As is known to all, AdaBoost [25] is the most commonly used model among boosting models, which is based on a forward learning algorithm to learn a model integrated with a large number of weak classifiers iteratively. Each weak classifier is a sample with the same data set but different distributions. At the same time, each iteration trains a different weak classifier and adjusts the data distribution characteristics after each training round to boost the weight of misclassified samples. It is worth noting that AdaBoost can accept initialized sample distributions with arbitrary sample and only focus on boosting samples that are prone to misclassification. The boosting sample weights is essentially re-sampling all the time, but here is more dynamic and directed “re-sampling”, which has more robust and universal. However, AdaBoost treats minority samples and majority samples in the misclassified samples equally, which is also the disadvantage of this algorithm in handling imbalanced samples.

2.3. The Proposal of ImAdaBoost

Specifically, in term of the distribution of misclassified samples, although AdaBoost will iteratively increase the next round of sampling weights for misclassified samples, it is still unable to determine the distribution in the error classification samples, that is, the misclassified samples may still be in imbalanced state. The so-called error samples not only include minority class samples, but also include many boundary overlapping majority class samples in imbalanced data. In addition, the sampling weights that are increased after minority class classification error are the same, essentially does not change the algorithm’s preference for the minority class samples in the incorrectly classified samples. Unless there are only a few types of samples in the data distribution that will cause more misclassification, but the number of minority type samples in the imbalanced data set is still much smaller than the number of majority type samples, and the weight update of minority type samples will not change their proportion in the next weighting round. Therefore, it is necessary to change the weight allocation method for various types of samples. In terms of weighting error, boosting training sample mainly comes from a data set composed of a majority class, so the weighting error of boosting training set may still be dominated by a majority of classes after training. The boosting algorithm, although it reduces the bias of the final set, may not be as effective for data sets with skewness distributions. In an imbalanced dataset, there is a very strong learning bias for majority classes of samples. With the training iteration, majority classes of samples can be fully learned.

Aiming at the shortcomings encountered by the original AdaBoost model in dealing with extreme imbalanced datasets, this paper proposes ImAdaBoost model. In this model, the proportion parameters of majority class samples and minority class samples in the sample are introduced in each iteration round’s weight updated function. According to the skewness proportion parameters in each round of misclassification samples, the next round weights of minority classes samples in the corresponding misclassification samples are adaptively strengthened. Based on adjusting the asymmetry of the class samples in each iteration, the symmetry between the minority class and the majority class samples can be achieved in the whole training process. The improved model is more suitable

for imbalanced distribution of sample classes, and even for extremely imbalanced cases. In essence, the improvement of ImAdaBoost model is to bring the adjusted weight of each sample into the loss function of each layer at the algorithm level, so as to realize the tendency of samples with high weight and add penalty coefficient to them. Compared with AdaBoost, using the adjusted weight update function makes the weight adjustment in hard samples more adaptive and produces more differences.

3. The Design of ImAdaBoost Algorithm

3.1. The ImAdaBoost Alogrithm Description

This section will describe the algorithm in this section. Firstly, the following parameters are defined.

Definition 1. Define $\beta'_t(i) = \{\beta_t(i) + 1\} \cdot c$ and $\beta_t(i) = \beta_t(\text{sign}\{y_i h_t(x_i)\}, \frac{n_{y=1}}{n})$

h is the predicted value of classifier and $h_t(x_i)$ represents the predicted value of sample x_i during the round t iteration. $\beta'_t(i)$ is named as sample skewness adjustment function and c is named as sample skewness adjustment factor. For each $h_t(x)$, $\beta'_t(i)$ is the adjustment variance of the all samples in D , which is affected by the sample skewness. $\beta_t(i)$ is the function before $\beta'_t(i)$ range correction and its range is 0 to 1, which specific function content is detailed in Appendix A.1 and Formulas (A1) and (A2). In addition, c is the adjustment factor of $\beta'_t(i)$, which adjust the amplitude of weight fluctuations to adapt to different sample sizes and distributions, with a range greater than or equal to 1. The following is the pseudo code of ImAdaBoost algorithm (Algorithm 1).

Algorithm 1 The ImAdaBoost.

Input: Training data set $D = \{(x_i, y_i)\}, x_i \in X \subseteq \mathbf{R}^l$, and $y_i \in \{-1, 1\}, i \in \{1, 2, 3, \dots, n\}$, choose the number of classifiers T .

1: **for** $t = 1$ to T **do**

2: Let $t = 1$ and initialize the sample weight $D_t(i) = \frac{1}{n}, i = (1, \dots, n)$;

3: Find the weak classifier $h_t(x) \in \{-1, 1\}$ minimizing the weighted error rate $e_t = \sum_{i: h_t(x_i) \neq y_i} D_t(i)$. If $e_t < 0.5$, repeat this step;

4: Calculate the weak classifier weight $\alpha_t = \frac{1}{2} \log(\frac{1-e_t}{e_t})$;

5: Update

$$D_{t+1}(i) = \frac{D_t(i) \exp\{-\alpha_t y_i h_t(x_i) \beta'_t(i)\}}{\sum_{i=1}^n D_t(i)} \quad (1)$$

β' is defined as sample skewness adjustment function. Let $\beta'_t(x) = \{\beta_t(x) + 1\} \cdot c$, where β and c are two parameters and $c \geq 1$. $\beta_t(x) = \beta_t(\text{sign}\{y_i h_t(x_i)\}, \frac{n_{y=1}}{n})$ and c is named as sample skewness adjustment factor;

6: **end for**

Output: The final strong classifier is $G(x) = \text{sign}\{f(x) - m\}$, where

$$f(x) = \sum_{t=1}^T \alpha_t h_t(x) \quad (2)$$

and m is the threshold.

3.2. Properties of the ImAdaBoost

In Appendix A.1, when $\text{sign}\{y_i h_t(x_i)\} = 1$, let $\beta_t(\text{sign}\{y_i h_t(x_i)\}, \frac{n_{y=1}}{n}) = \beta_+ = 0$, when $\text{sign}\{y_i h_t(x_i)\} = -1$, let $\beta_t(\text{sign}\{y_i h_t(x_i)\}, \frac{n_{y=1}}{n}) = \beta_- \geq 0$, namely, it ensures that the weight adjustment function β_+ for correctly classified samples is always smaller than the weight adjustment function β_- for misclassified samples; Secondly, it differentiates the Sample skewness adjustment function of minority samples from majority samples

in misclassified samples, as shown in Table 2. Please refer to Appendix A.1 for specific formula derivation.

Table 2. β in different conditions of y_i and $h_t(x_i)$.

		$h_t(x_i) = 1$	$h_t(x_i) = -1$
$y_i = 1$	$n_{y=1} < n_{y=-1}$	1	$\beta_{+-} = -0.5 \times \frac{n_{y=1}}{n} + 1$
	$n_{y=1} = n_{y=-1}$	1	$\beta_{+-} = 0.5$
	$n_{y=1} > n_{y=-1}$	1	$\beta_{+-} = -0.5 \times \frac{n_{y=1}}{n} + 0.5$
$y_i = -1$	$n_{y=1} < n_{y=-1}$	$\beta_{-+} = 0.5 \times \frac{n_{y=1}}{n}$	1
	$n_{y=1} = n_{y=-1}$	$\beta_{-+} = 0.5$	1
	$n_{y=1} > n_{y=-1}$	$\beta_{-+} = 0.5 \times \frac{n_{y=1}}{n} + 0.5$	1

The following three assumptions are set here: (1) The value of the weight adjustment function of the minority class is always greater than that of the majority class in the misclassified samples. (2) As the proportion of the actual minority class in the misclassified sample increases, the value of the weight adjustment function β for the minority class decreases and the corresponding value of the weight adjustment function for the majority class increases. (3) the paper set the scenario in extreme imbalanced, so the weight adjustment function β for the minority class decreases more slowly and the corresponding weight adjustment function for the majority class increases more slowly. It makes sure that the weight of the minority class always remains dominant in the sample when it is present.

Similar to AdaBoost model, an upper bound on the training error is proved in the ImAdaBoost model to theoretically demonstrate the convergence of its algorithm.

Theorem 1. Let $f'(x) = \sum_{t=1}^T \alpha^t h_t \beta'(\text{sign}\{y_i h_t(x_i)\}, \frac{n_{y=1}}{n})$ and $H'(x) = \text{sign}(f'(x))$. If $\forall i$, $\beta'_-(i) \geq \beta'_+(i)$, then:

$$\forall X \in S, (H'(x) = y \Rightarrow H(x) = y) \quad (3)$$

The proof process of this theorem is shown in Appendix A.2. Theorem 1 shows that for each class that is correctly classified by $H'(x)$ classifier, it is also correctly classified by $H(x)$. The difference between the two classifiers is that the former introduces a new parameter $\beta'(\text{sign}\{y_i h_t(x_i)\})$.

Theorem 2. According to Theorem 1, The upper bound formula of ImAdaBoost loss function can be derived:

$$\sum_i c_i \cdot I(H(x) \neq y) \leq d \prod_t Z_t, d = \sum_j c_j \quad (4)$$

The proof process of this theorem is shown in Appendix A.3.

When adjusting the β function, namely, to ensure that $\beta_- \geq \beta_+$, i.e., the value of β for the misclassified sample is greater than β for the correctly classified sample. Although the weight of the misclassified sample is always greater than the weight of the correctly classified sample in the weight adjustment formula, this setting helps us to remove the effect of the class skewness adjustment function β and the specific correctly classified sample on the proof in Theorem 1, and these factors are not present in our test set.

Meanwhile, $\beta_{+-} \geq \beta_{-+}(n_{y=1} < n_{y=-1})$, which means that in the misclassified samples (Hard-examples), if the minority class samples are assigned β values greater than the majority class assigned β values, it makes the weak classifier in each round focus more on the minority class samples that were misclassified by the previous classifier. It is similar in $\beta_{+-} \leq \beta_{-+}(n_{y=1} > n_{y=-1})$. Such an optimized weight adjustment function ensures that each weak classifier will focus more on the minority class of hard samples.

4. Experimental Setup and Result Analysis

4.1. Data Collection

Krawczyk (2016) proposed that the imbalanced datasets with class ratios above 100 is called extremely imbalanced datasets [11]. Based on this criterion, the paper selected Credit Card dataset from the Université Libre de Bruxelles (ULB), Page Blocks dataset from the University of California Irvine (UCI) Data Platform, and Fraud Detection dataset from the Kaggle Data Platform to verify the effectiveness of the ImAdaBoost algorithm on the extreme class imbalance and the general imbalance data, respectively. The first two datasets basically meet the conditions of extreme imbalance and the differences are obvious to verify the adaptability of ImAdaBoost model under the conditions of extreme imbalance. In addition, the dataset with common imbalance ratios is selected to verify the advantages and disadvantages of the ImAdaBoost model and the traditional imbalance model in the conventional imbalance state. The specific distribution of the datasets is shown in Table 3.

Table 3. Datasets and sample description.

Dataset	n	n _{minority}	n _{majority}	Imbalanced Ratio	Features
ULB: Credit Card	285,299	492	284,807	1:570	28
UCI: Page Blocks	5028	115	4913	2.3:100	10
Kaggle: Fraud Detection	307,511	24,825	282,686	1:11	120

Firstly, the Credit Card dataset is obtained from ULB Machine Learning Group. It includes the 28 features and a binary imbalanced label. This dataset presents transactions that occurred in two days, where we have 492 minority classes out of 285,299 samples. The dataset is extremely imbalanced and the minority class account for 0.172% of all transactions. Secondly, the Page Blocks dataset is weakly extreme imbalance sample and is obtained from UCI machine learning repository data sets. It includes the 10 features and a binary imbalanced label. This dataset presents transactions that occurred in two days, where the dataset have 115 minority classes out of 5028 samples. The dataset is imbalanced and the minority class accounts for 2.287% of all transactions. Thirdly, the Fraud Detection dataset is little imbalance sample and is obtained from Kaggle Data Platform. It includes the 120 features and a binary imbalanced label. This dataset presents transactions that occurred in two days, where the dataset have 24,825 minority classes out of 307,511 samples. The dataset is a little imbalanced, the minority class accounts for 8.073% of all transactions.

4.2. Model Evaluation Metrics

The evaluation metrics of general classification problems can be roughly divided into two categories: threshold-based metrics and threshold-free metrics. The former mainly depends on the threshold, mainly including accuracy, recall, specificity, precision and F1-score metrics, and these corresponding metrics value is determined according to the determined threshold; while the latter is independent of the threshold, mainly including ROC curve and Precision-Recall curve, and its corresponding quantitative metrics AUC values and Average-Precision values, they determine the corresponding metrics under each threshold by traversing all possible decision thresholds, so as to construct the corresponding curve evaluation metrics. In the experiment of this paper, due to the extreme imbalance of the samples, the number of minority class samples is too small. Although the model will improve the recognition of minority classes, the overall sample size is too small, resulting in this improvement is diluted in the AUC and AP metrics, so the selection of our metrics needs to reflect the changes in the minority class. In summary, the paper adopt common evaluation metrics, and comprehensively selects the minority class recall rate (TPR), the majority class recall rate (TNR), as well as F1-score and G-mean (Equations (5)–(9)), see Table 4 for details.

$$TPR = recall = \frac{TP}{TP + FN} \quad (5)$$

$$TNR = \frac{TN}{TN + FP} \quad (6)$$

$$precision = \frac{TP}{TP + FP} \quad (7)$$

$$F1\text{-score} = \frac{(1 + \beta^2) \times recall \times precision}{\beta^2 \times recall + precision} \quad (8)$$

$$G\text{-mean} = \sqrt{TPR \times TNR} \quad (9)$$

Table 4. The confusion matrix.

True Class	Predicted Class	
	Positive/Minority	Negative/Majority
positive/minority	TP	FN
negative/majority	FP	TN

4.3. Experimental Design

Firstly, by training the extreme imbalanced data, the paper will calculate the optimal weight adjustment factor c of the ImAdaBoost model. Secondly, two comparative experiments will be conducted to analyze the effect of ImAdaBoost model from multiple perspectives. On the one hand, the paper study the differences between ImAdaBoost and AdaBoost models in terms of training metrics under the same parameters. On the other, the paper compare ImAdaBoost with other mainstream imbalance algorithms based on three different types of imbalanced datasets. Among them, since the imbalance algorithms at the data level are all data preprocessing methods, the paper use AdaBoost as the base model to train the preprocessed dataset. Meanwhile, in the analysis process, the paper further demonstrate the advantages and disadvantages of ImAdaBoost in dealing with imbalance problems.

4.4. Experimental Results and Analysis

4.4.1. The Choice of the Adjustment Factor c

ImAdaBoost model sets the adjustment factor c in the weight adjustment function to control the magnitude of the weight change. The super parameter range is set for training to obtain the best results of the ImAdaBoost model, which also improves the adaptability of the improved model to other data. Because the UBL: Credit Card dataset is in an extreme imbalanced state, this dataset is selected to calculate the weight adjustment factor, so as to ensure that the ImAdaBoost model can better handle the extremely imbalanced data set and show the difference from the original imbalance model in the extremely imbalanced data set. Based on this dataset, the paper study the weight adjustment factor within the range of [1, 5], select 70% of the sample size as the training set to train the ImAdaBoost model, and obtain the combined G-mean and F1-score evaluation index values on the test set. Finally, the paper obtain the most suitable weight adjustment factor and the best classification result for this dataset.

As is shown in Figure 1, with the change of the weight adjustment factor, the G-mean value does not fluctuate significantly. It is mainly due to the small number of minority class, and the change of weight adjustment factor in ImAdaBoost model is weak on the minority class recall with a fluctuation interval of [0.88, 0.92]. Yet the weight adjustment factor has a significant impact on the precision of ImAdaBoost model's minority class recognition, i.e., the precision of model recognition decreases rapidly with the increase of adjustment

factor, leading to the rapid decrease of F1-score value. Therefore, the value of the weight adjustment factor should be between the range $[1, 3]$. After comprehensive consideration, the weight adjustment factor is set to 2.5, then the F1-score is 0.88 and the G-mean is 0.91, which are both at a high level.

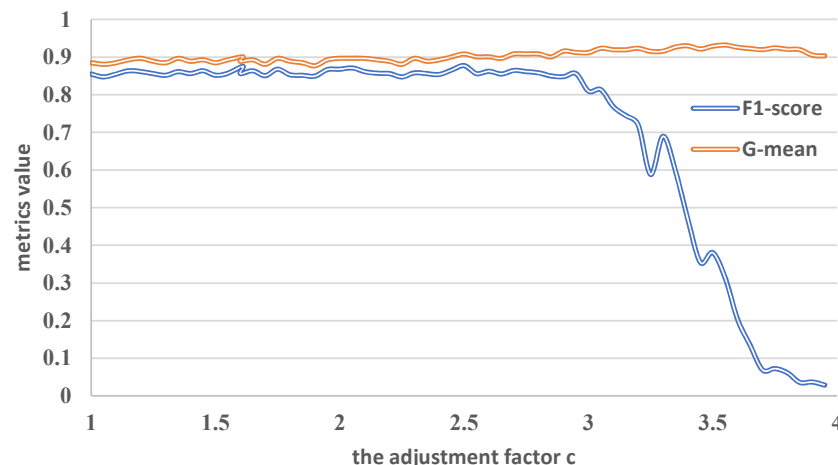


Figure 1. Metrics value change with the adjustment factor c .

4.4.2. Comparative Analysis of ImAdaBoost and AdaBoost Model

Because Credit Card dataset has the smallest imbalance ratio, it can better reflect the difference between the improved model and the original model in the iterative process. Then, 70% of the samples are divided into the training dataset and 30% of the samples are divided into the test dataset. So, based on the Credit Card train dataset from ULB, the paper used ImAdaBoost and AdaBoost models for training, and obtained loss function data during model iteration.

Figure 2 shows the loss of sample weights in the training set after each iteration of the ImAdaBoost and AdaBoost models for 150 iterations. The sample weight loss of AdaBoost is smaller than that of ImAdaBoost in each iteration. This is because the paper assign greater weights to the misclassified samples in ImAdaBoost compared to AdaBoost when calculating its weight log loss function. It also proves that the weights assigned by the improved model are effective. However, this does not mean that obtaining a larger estimator error in each round leads to poor training. On the contrary, the paper let the model learn the features of minority classes in the sample by extending this loss to allow the model to train more on this part.

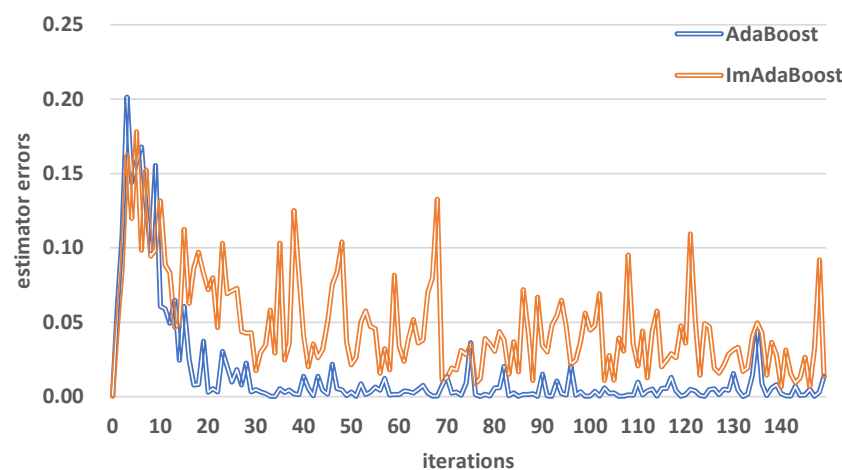


Figure 2. Estimator errors trend for AdaBoost and ImAdaBoost model with different iterations.

Figure 3 shows the recall of the minority class samples in the test dataset for AdaBoost and ImAdaBoost under the same training conditions. According to Figure 3, the percentage of minority class samples correctly identified in each iteration in AdaBoost is unstable with increasing iterations compared to ImAdaBoost. The overall recall is observed to be smaller than that of ImAdaBoost. It shows that it is effective in the training set by assigning more weight to the next iteration under incorrect identification in minority class samples.

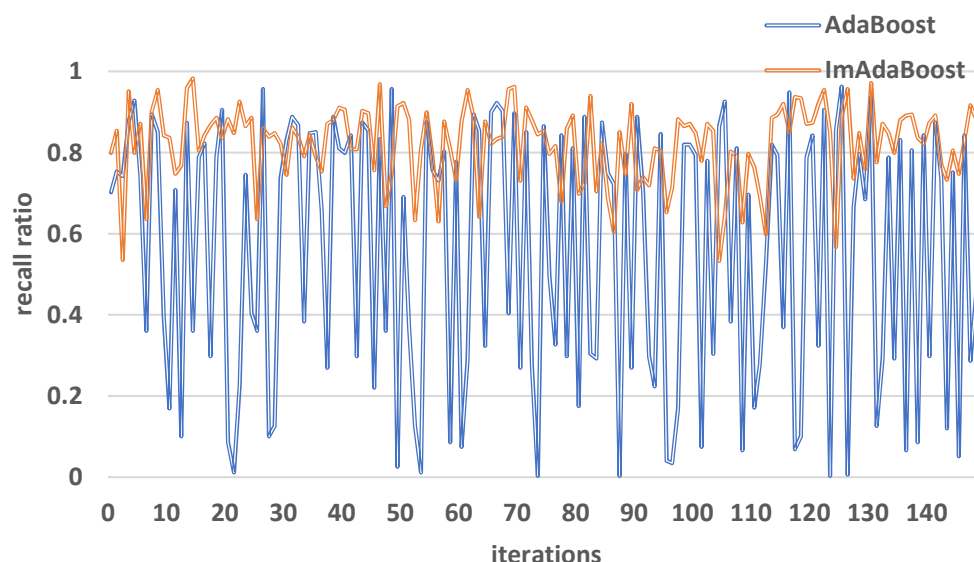


Figure 3. Training sample recall rate change for AdaBoost and ImAdaBoost model.

4.4.3. Comparative Analysis of ImAdaBoost and Traditional Imbalance Models

Next, the paper selected three types of imbalance models for comparative experiments to verify the advantages and disadvantages of ImAdaBoost in processing imbalanced datasets with different proportions. The three types of imbalance models are as follows: (1) Sampling type imbalance model: RUS [12], ROS [12]; (2) Synthetic type imbalance model: SMOTE [14], ADASYN [15], Borderline-SMOTE [26] and KMeans-SMOTE [27]; (3) Ensemble type imbalance models: EasyEnsemble [20], Balanced-Bagging [28], RusBoost [18] and Balanced-RandomForest [29].

The advantages and disadvantages of ImAdaBoost are fully validated against the other mainstream models mentioned above. In addition to TPR (minority class) and TNR (majority class), the paper also select another two comprehensive metrics to judge the models in this paper. One is a comprehensive metrics G-mean that measures the comprehensive ability of the model to identify the recall metrics of each class; The other is the comprehensive metrics F1-score, which is harmonic mean of precision and recall of minority class, and measures the comprehensive ability of the model to identify minority samples. Tables 5–7 show the 20-fold cross validation mean data for each evaluation metrics of ImAdaBoost and other mainstream models under three different datasets.

Table 5 show that in the extremely imbalanced dataset, the F1-score of ImAdaBoost model is slightly larger than that of AdaBoost model, and is significantly larger than the other three types imbalance models. The majority class recall rate (TNR) of ImAdaBoost model inherits the excellent classification effect of AdaBoost model and outperforms the other imbalance ensemble models. It shows that in the sampling type imbalance model, the synthetic type imbalance model and the ensemble type imbalance model which all sacrifice the recognition of majority class samples and recognize relatively more samples as minority class. This results in the phenomenon of “winning by quantity” to ensure the minority class recall rate is higher than ImAdaBoost model. Although these three types models improve the recognition of minority samples, the recognition recall of majority samples and the recognition accuracy of minority samples are lower than that of ImAdaBoost model.

Table 5. The classification results for ULB: Credit Card dataset.

Algorithms	TNR	Evaluation Measures		F1-Score
		TPR	G-Mean	
ImAdaBoost	0.9999	0.8330	0.9114	0.8781
AdaBoost	0.9999	0.7864	0.8859	0.8474
RUS	0.9424	0.8062	0.9237	0.0530
ROS	0.9764	0.7894	0.9100	0.1136
ADASYN	0.9279	0.8882	0.9070	0.0414
SMOTE	0.9714	0.8839	0.9260	0.0994
Borderline-SMOTE	0.9858	0.8678	0.9243	0.1763
KMeansSMOTE	0.9817	0.8904	0.9343	0.1451
EasyEnsemble	0.9676	0.9169	0.9414	0.0890
BalancedBagging	0.9826	0.8945	0.9369	0.1502
RUSBoost	0.9645	0.8314	0.8947	0.0807
BalancedRF	0.9768	0.9087	0.9416	0.1194

As shown in Table 6, ImAdaBoost model's identification recall (TNR) in the majority class sample is slightly lower than that of the synthetic type imbalance model represented by ADASYN and the ensemble class imbalance model represented by BalancedBagging. ImAdaBoost model outperforms the other models in terms of minority class recall (TPR), thus showing the best performance of G-mean. The mainstream imbalance models do not lose as much precision in minority class identification as in the extreme class imbalance sample and the F1-score of a few models is still higher than that of ImAdaBoost model, but the minority class recall is weaker than ImAdaBoost model in all imbalance models.

Table 6. The classification results for UCI: Page Blocks dataset.

Algorithms	TNR	Evaluation Measures		F1-Score
		TPR	G-Mean	
ImAdaBoost	0.9553	0.8945	0.9143	0.8032
AdaBoost	0.9435	0.8064	0.7859	0.8214
RUS	0.9142	0.8532	0.8421	0.6241
ROS	0.9021	0.8213	0.8242	0.6915
ADASYN	0.9623	0.7923	0.8453	0.7324
SMOTE	0.8938	0.8421	0.8660	0.7994
Borderline-SMOTE	0.9264	0.7537	0.8542	0.8065
KMeansSMOTE	0.9153	0.8065	0.8754	0.8053
EasyEnsemble	0.8946	0.7865	0.8143	0.7968
BalancedBagging	0.9616	0.8857	0.8567	0.8389
RUSBoost	0.9287	0.8675	0.9053	0.6775
BalancedRF	0.9213	0.7856	0.8753	0.7194

Table 7 show that the effect of this experiment is similar to that in the weak extremely imbalanced sample of UCI: Page Blocks dataset. When the imbalance ratio is 1:11, the ImAdaBoost model also has better classification recall in the minority sample. This is due to the fact that the absolute number of minority samples in this sample is relatively larger and the model can fully learn the features of the minority samples. In the composite evaluation metrics F1-score of minority class samples, ImAdaBoost model and ensemble type imbalance model RUSBoost and BalancedBagging perform best.

Based on the previous experiments, it can be concluded that the improved model ImAdaBoost proposed in this paper outperforms most traditional imbalance models in terms of the classification metrics F1-score and G-mean obtained on extreme imbalanced and imbalanced data sets. On the two extreme imbalanced datasets, Credit Card and Page Blocks, the difference shown by the ImAdaBoost model is mainly reflected in that compared to the comparative model, the improved model obtains the best TPR on the Page Blocks dataset, indicating that the model has the best high recall performance for minority class at

this sample ratio. In addition, ImAdaBoost model also has the best high recall performance for minority class sample in the imbalanced dataset Fraud Detection.

Table 7. The classification results for Kaggle: Fraud Detection dataset.

Algorithms	Evaluation Measures			
	TNR	TPR	G-Mean	F1-Score
ImAdaBoost	0.9754	0.8923	0.9231	0.9054
AdaBoost	0.9593	0.8364	0.8632	0.8530
RUS	0.9233	0.8064	0.8642	0.8213
ROS	0.9394	0.8106	0.8543	0.8532
ADASYN	0.9723	0.8346	0.8654	0.8432
SMOTE	0.9553	0.8574	0.8908	0.8654
Borderline-SMOTE	0.9675	0.8643	0.8762	0.8532
KMeansSMOTE	0.9568	0.8543	0.8876	0.8573
EasyEnsemble	0.9246	0.8325	0.8543	0.8214
BalancedBagging	0.9865	0.8643	0.8905	0.9123
RUSBoost	0.9887	0.8776	0.9343	0.8932
BalancedRF	0.9475	0.8556	0.8906	0.8234

All in all, from the applicability view of the ImAdaBoost model in different ratio datasets, the classification effect of ImAdaBoost is somewhat differentiated on the different imbalanced datasets. In the dataset with an extreme imbalance ratio, the main advantage of ImAdaBoost model over other imbalance classification models is its strongest stability in minority class precision and recall, ensuring the highest F1-score value; In the dataset with a weakly extreme imbalance ratio, it shows excellent minority class recall ability. In addition, it also has the best minority class recall ability in the generally imbalanced dataset. Just in terms of extreme imbalances with a ratio which is below 1%, ImAdaBoost model is stronger than the sampling-based classification models for minority class recall. Nonetheless, it is significantly weaker than the synthetic and ensemble type imbalance models. The traditional imbalance classification models show unstable values for minority class precision. At the weakly extremely imbalanced dataset with a ratio which is above 1%, ImAdaBoost model outperforms all three classification models in terms of recall for the minority class. The majority class recall is weaker than the ADASYN model which belongs to the synthetic type imbalance models. In addition, the minority class precision identification is weaker than the BalancedBagging model in the ensemble class imbalance model. In addition, it is noted that the weaknesses of the ImAdaBoost model are that the performance in different types of extreme class imbalanced data varies for higher ratio of extreme imbalance, the recall effect is not superior to the mainstream imbalance model.

5. Conclusions and Prospects

It has been an important asymmetric topic how to improve the performance of identifying minority and majority class samples in imbalanced sample, especially in extreme sample imbalance. It varies from sampling to cost-sensitive and single-model research on a single method to combining data level with algorithm level, data rebalancing or cost-sensitive is embedded in the integrated model for balancing sample number or adjusting the weight update formula. Finally, the mainstream method of deeply embedding sample balancing strategy is applied with the help of ensemble framework. At the same time, deep learning model framework is introduced in sample imbalance processing. However, the evolution of these methods fails to consider the state of extreme sample imbalance. Based on the weight adjustment function in the AdaBoost ensemble model, this paper modifies this function and embed the weight adjustment factor to make this model more adaptable to the imbalanced sample class distribution. It also retains the sampling weights of misclassified samples while paying more attention to the classification of minority class samples in misclassified samples. Additionally, it optimized the weight adjustment factor to make the model applicable to the extreme imbalance situation. Through empirical analysis, there

are three key findings: (1) ImAdaBoost model performs strongest classification stability in extreme imbalanced dataset. It can be concluded that It has strong stability to minority class precision compared to other models in data with extreme imbalance ratio which is below 1%; (2) ImAdaBoost model has good recall for minority class samples in a weakly extreme imbalance and generally imbalance state with a ratio which is above 1%; (3) Compared with traditional imbalance models, the ImAdaBoost model can still maintain good minority recall ability in extreme imbalanced datasets without losing the accuracy of minority class recognition, while traditional models often lose the accuracy of minority class recognition in extreme imbalanced datasets.

From the previous model comparison discussion, ImAdaBoost model's recall and precision stability for minority samples are characteristics that other models do not have. It is because of the improved weight adjustment function in this paper, which constantly follows up the proportion of imbalance in misclassified samples in each round of iterations to dynamically assign different weights. It always ensures the stability of minority samples and majority samples. However, the selection of weight adjustment factor in this paper is determined based on a single data set. Whether the weight adjustment factor c selection can be further determined adaptively based on sample skewness to further improve the model's generalization ability and minority sample recall in extremely imbalanced datasets needs to be further explored. Meanwhile, the selected extreme imbalance ratios in the dataset are not refined to each stratum, which requires more collection of extreme imbalanced dataset with different ratios for training validation to further improve the applicability of ImAdaBoost model in extreme imbalanced datasets with different ratios.

Author Contributions: Conceptualization, J.X.; methodology, J.X. and J.M.; writing—original draft preparation, J.M.; writing—review and editing, J.X.; software, J.M.; formal analysis, J.X. and J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available on request from ULB: Credit Card Worldline and the Machine Learning Group accessed on 15 January 2022 (<http://mlg.ulb.ac.be>) of ULB (Université Libre de Bruxelles) Credit Card dataset, the UCI Data Platform Page Blocks dataset accessed on 19 January 2022 (<http://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification>) and the Kaggle Data Platform Fraud Detection dataset accessed on 6 February 2022 (<https://www.kaggle.com/datasets/mishra5001/credit-card>).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Formula Expansion of Definition 1

$$\beta_i = \begin{cases} \beta_{y_i=h_t(x)} = \beta_+ = 0; \\ \beta_{y_i \neq h_t(x)} = \beta_- = \begin{cases} \beta_{y_i=1, h_t(x)=-1} = \beta_{+-} = \begin{cases} 0.5 \times \frac{n_{y=-1}}{n} + 0.5, & n_{y=1} < n_{y=-1}; \\ -0.5 \times \frac{n_{y=1}}{n} + 0.5, & n_{y=1} > n_{y=-1}; \end{cases} \\ 0.5 & n_{y=1} = n_{y=-1}; \\ \beta_{y_i=-1, h_t(x)=1} = \beta_{-+} = \begin{cases} -0.5 \times \frac{n_{y=-1}}{n} + 0.5, & n_{y=1} < n_{y=-1}; \\ 0.5 \times \frac{n_{y=1}}{n} + 0.5, & n_{y=1} > n_{y=-1}; \end{cases} \end{cases} \end{cases} \quad (A1)$$

$\Rightarrow$

$$\beta_i = \begin{cases} \beta_{y_i=h_t(x)} = \beta_+ = 0; \\ \beta_{y_i \neq h_t(x)} = \beta_- = \begin{cases} \beta_{y_i=1, h_t(x)=-1} = \beta_{+-} = \begin{cases} -0.5 \times \frac{n_{y=1}}{n} + 1, & n_{y=1} < n_{y=-1}; \\ -0.5 \times \frac{n_{y=1}}{n} + 0.5, & n_{y=1} > n_{y=-1}; \end{cases} \\ 0.5 & n_{y=1} = n_{y=-1}; \\ \beta_{y_i=-1, h_t(x)=1} = \beta_{-+} = \begin{cases} 0.5 \times \frac{n_{y=1}}{n}, & n_{y=1} < n_{y=-1}; \\ 0.5 \times \frac{n_{y=1}}{n} + 0.5, & n_{y=1} > n_{y=-1}; \end{cases} \end{cases} \end{cases} \quad (\text{A2})$$

Appendix A.2. Proof of Theorem 1

$$H'(x) = y \quad (\text{A3})$$

$$\Leftrightarrow yf'(x) = y \sum_{t=1}^T \alpha_t h_t \beta'(\text{sign}\{y_i h_t(x_i)\}, \frac{n_{y=1}}{n}) > 0 \quad (\text{A4})$$

As above Formula (A4), the sum of the correct prediction part and the prediction error part on the right side, and the prediction error part is split into a false negative class with a true class of 1 and a predicted class of -1 , and the false positive class with a true class of -1 , predicted class 1 is as follows:

$$y \sum_{y h_t(x)=1 \Rightarrow t+}^T \alpha_{t+} h_{t+} \beta'_{t+} + y \sum_{y h_t(x)=-1 \Rightarrow t-, t- = \{t+ -, t- +\};}^T \alpha_{t-} h_{t-} \beta'_{t-} \quad (\text{A5})$$

$$= y \sum_{y h_t(x)=1 \Rightarrow t+}^T \alpha_{t+} h_{t+} \beta'_{t+} + y \sum_{t+-}^T \alpha_{t+-} h_{t+-} \beta'_{t+-} + y \sum_{t-+}^T \alpha_{t-+} h_{t-+} \beta'_{t-+} \quad (\text{A6})$$

because $\beta'_{t+} \geq 0, \beta'_{t+-} \geq 0, \beta'_{t-+} \geq 0, \alpha_t \geq 0$, then

$$y \sum_{t+}^T \alpha_{t+} h_{t+} \beta'_{t+} > 0, \quad (\text{A7})$$

$$(y \sum_{t+-}^T \alpha_{t+-} h_{t+-} \beta'_{t+-} + y \sum_{t-+}^T \alpha_{t-+} h_{t-+} \beta'_{t-+}) < 0 \quad (\text{A8})$$

$$y \sum_{t+-}^T \alpha_{t+-} h_{t+-} \beta'_{t+-} < 0, y \sum_{t-+}^T \alpha_{t-+} h_{t-+} \beta'_{t-+} < 0 \quad (\text{A9})$$

because $\beta'_{t+-} \geq \beta'_{t+} > 0, \beta'_{t-+} \geq \beta'_{t+} > 0$

$$\begin{aligned} & y \sum_{t+}^T \alpha_{t+} h_{t+} \beta'_{t+} + y \sum_{t+-}^T \alpha_{t+-} h_{t+-} \beta'_{t+-} + y \sum_{t-+}^T \alpha_{t-+} h_{t-+} \beta'_{t-+} \\ & \geq y \sum_{t+}^T \alpha_{t+} h_{t+} \beta'_{t+} + y \sum_{t+-}^T \alpha_{t+-} h_{t+-} \beta'_{t+-} + y \sum_{t-+}^T \alpha_{t-+} h_{t-+} \beta'_{t-+} \end{aligned} \quad (\text{A10})$$

which further translates to

$$y \beta'_{t+} \left\{ \sum_{t+}^T \alpha_{t+} h_{t+}(x) + \sum_{t+-}^T \alpha_{t+-} h_{t+-}(x) + \sum_{t-+}^T \alpha_{t-+} h_{t-+}(x) \right\} > 0 \quad (\text{A11})$$

$$\Rightarrow y\beta'_{t+} \sum_t^T \alpha_t h_t > 0 \quad (\text{A12})$$

$$\Rightarrow y \sum_t^T \alpha_t h_t > 0 \quad (\text{A13})$$

$$\Rightarrow yf(x) > 0 \quad (\text{A14})$$

$$\Leftrightarrow y = \text{sign}\{f(x)\} = H(x) \quad (\text{A15})$$

This proves that:

$$y = H'(x) \Rightarrow y = H(x) \quad (\text{A16})$$

Appendix A.3. Proof of Theorem 2

From Theorem 1, we can obtain:

$$\sum_i c_i \cdot I(H(x) \neq y) \leq \sum_i c_i \cdot I(H'(x) \neq y) \quad (\text{A17})$$

The transformation of the weight adjustment function $D(i)$ in the boosting process,

$$\begin{aligned} D_{T+1}(i) &= \frac{D_T(i) \exp\{-\alpha_T y_T h_T(x_i) \beta'(i)\}}{\sum_{i=1}^n D_T(i)} \\ &= \frac{D_T(i) \exp\{-\alpha_T y_T h_T(x_i) \beta'(i)\}}{Z_T} \\ &= \frac{D_1(i) \exp\{-y_i (\sum_{t=1}^T \alpha_t h_t(x_i) \beta'(i))\}}{\prod_{t=1}^T Z_t} \\ &= \frac{D_1(i) \exp\{-y_i f'(x_i)\}}{\prod_{t=1}^T Z_t} \end{aligned} \quad (\text{A18})$$

Among them, $Z_T = \sum_{i=1}^n D_T(i)$. And when $H'(x_i) \neq y_i$, then $y_i \cdot f'(x_i) \leq 0$, you can obtain $\exp\{-y_i f'(x_i)\} \geq 0$. So,

$$I(H'(x_i) \neq y_i) \leq \exp\{-y_i f'(x_i)\} \quad (\text{A19})$$

Combining Formulas (A16)–(A18) and $D_1(i) = \frac{c_i}{\sum_k c_k}$ gives the upper bound formula of training error:

$$\begin{aligned} \sum_i c_i \cdot I(H'(x) \neq y) &\leq \sum_i c_i \cdot \exp\{-y_i f'(x_i)\} \\ &= \sum_i (c_i \cdot \frac{D_{T+1}(i)}{D_1(i)} \prod_t Z_t) \\ &= \prod_t Z_t \cdot \sum_i c_i \frac{D_{T+1}(i)}{D_1(i)} \end{aligned} \quad (\text{A20})$$

And $D_1(i) = c_i / \sum_k c_k$, then

$$\sum_i c_i \cdot I(H'(x) \neq y) \leq d \prod_t Z_t, d = \sum_k c_k \quad (\text{A21})$$

References

1. Garrido, A. Symmetry and Asymmetry Level Measures. *Symmetry* **2010**, *2*, 707–721. [\[CrossRef\]](#)
2. Bejjanki, K.K.; Gyani, J.; Gugulothu, N. Class Imbalance Reduction (CIR): A Novel Approach to Software Defect Prediction in the Presence of Class Imbalance. *Symmetry* **2020**, *12*, 407. [\[CrossRef\]](#)
3. Zhang, H.; Liu, Q. Online Learning Method for Drift and Imbalance Problem in Client Credit Assessment. *Symmetry* **2019**, *11*, 890. [\[CrossRef\]](#)
4. Li, D.C.; Chen, S.C.; Lin, Y.S.; Hsu, W.Y. A Novel Classification Method Based on a Two-Phase Technique for Learning Imbalanced Text Data. *Symmetry* **2022**, *14*, 567. [\[CrossRef\]](#)
5. Zhang, Y.Q.; Lu, R.Z.; Qiao, S.J.; Han, N.; Gutierrez, L.A.; Zhou, J.L. A Sampling Method of Imbalanced Data Based on Sample Space. *Zidonghua Xuebao/Acta Autom. Sin.* **2022**, *48*, 2549–2563. [\[CrossRef\]](#)
6. Galar, M.; Fernandez, A.; Barrenechea, E.; Bustince, H.; Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* **2012**, *42*, 463–484. [\[CrossRef\]](#)
7. Elkan, C. The Foundations of Cost-Sensitive Learning. In Proceedings of the 17th International Joint Conference on Artificial Intelligence, IJCAI'01, Seattle, WA, USA, 4–10 August 2001; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001; Volume 2, pp. 973–978. [\[CrossRef\]](#)
8. Guo, J. Research on Ensemble Approach for Classification of Imbalanced Data Sets. Master's Thesis, Harbin Institute of Technology, Harbin, China, 2017.
9. Napierała, K.; Stefanowski, J.; Wilk, S. Learning from Imbalanced Data in Presence of Noisy and Borderline Examples. In Proceedings of the Rough Sets and Current Trends in Computing: 7th International Conference, RSCTC 2010, Warsaw, Poland, 28–30 June 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 158–167. [\[CrossRef\]](#)
10. García, V.; Mollineda, R.; Sánchez, J. On the k-NN performance in challenging scenario of imbalance and overlapping. *Pattern Anal. Appl.* **2008**, *11*, 269–280. [\[CrossRef\]](#)
11. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [\[CrossRef\]](#)
12. Dey, T.; Giesen, J.; Goswami, S.; Hudson, J.; Wenger, R.; Zhao, W. Undersampling and oversampling in sample based shape modeling. In Proceedings of the Proceedings Visualization, VIS '01, San Diego, CA, USA, 21–26 October 2001; pp. 83–545. [\[CrossRef\]](#)
13. Sampath, S. Hybrid single sampling plan. *World Appl. Sci. J.* **2009**, *6*, 1685–1690.
14. Chawla, N.; Bowyer, K.; Hall, L.; Kegelmeyer, W. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)
15. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328. [\[CrossRef\]](#)
16. Chawla, N.; Cieslak, D.; Hall, L.; Joshi, A. Automatically countering imbalance and its empirical relationship to cost. *Data Min. Knowl. Discov.* **2008**, *17*, 225–252. [\[CrossRef\]](#)
17. Freitas, A.; Pereira, A.; Brazdil, P. Cost-Sensitive Decision Trees Applied to Medical Data. In Proceedings of the Data Warehousing and Knowledge Discovery: 9th International Conference, DaWaK 2007, Regensburg, Germany, 3–7 September 2007; Volume 4654, pp. 303–312. [\[CrossRef\]](#)
18. Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *IEEE Trans. Syst. Man Cybern. Part A Syst. Hum.* **2010**, *40*, 185–197. [\[CrossRef\]](#)
19. Raghuwanshi, B.S.; Shukla, S. Class imbalance learning using UnderBagging based kernelized extreme learning machine. *Neurocomputing* **2019**, *329*, 172–187. [\[CrossRef\]](#)
20. Liu, X.Y.; Wu, J.; Zhou, Z.H. Exploratory Undersampling for Class-Imbalance Learning. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2009**, *39*, 539–550. [\[CrossRef\]](#)
21. Sharma, S.; Bellinger, C.; Krawczyk, B.; Zaiane, O.; Japkowicz, N. Synthetic Oversampling with the Majority Class: A New Perspective on Handling Extreme Imbalance. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Hong Kong, China, 1–8 June 2018; pp. 447–456. [\[CrossRef\]](#)
22. Ge, J.F.; Luo, Y.P. A Comprehensive Study for Asymmetric AdaBoost and Its Application in Object Detection. *Acta Autom. Sin.* **2009**, *35*, 1403–1409. [\[CrossRef\]](#)
23. Sun, J.; Li, H.; Fujita, H.; Fu, B.; Ai, W. Class-imbalanced dynamic financial distress prediction based on Adaboost-SVM ensemble combined with SMOTE and time weighting. *Inf. Fusion* **2020**, *54*, 128–144. [\[CrossRef\]](#)
24. Wang, W.; Sun, D. The improved AdaBoost algorithms for imbalanced data classification. *Inf. Sci.* **2021**, *563*, 358–374. [\[CrossRef\]](#)
25. Freund, Y.; Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *Lect. Notes Comput. Sci.* **1995**, *904*, 23–37. [\[CrossRef\]](#)
26. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New over-Sampling Method in Imbalanced Data Sets Learning. In *Advances in Intelligent Computing, Proceedings of the 2005 International Conference on Advances in Intelligent Computing, ICIC'05, Hefei, China, 23–26 August 2005*; Springer: Berlin/Heidelberg, Germany, 2005; Volume Part I, pp. 878–887. [\[CrossRef\]](#)
27. Douzas, G.; Bacao, F.; Last, F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Inf. Sci.* **2018**, *465*, 1–20. [\[CrossRef\]](#)

-
28. Błaszczyński, J.; Stefanowski, J. Neighbourhood sampling in bagging for imbalanced data. *Neurocomputing* **2015**, *150*, 529–542. [[CrossRef](#)]
 29. Anaissi, A.; Kennedy, P.; Goyal, M.; Catchpoole, D. A balanced iterative random forest for gene selection from microarray data. *BMC Bioinform.* **2013**, *14*, 261. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.