

ORIGINAL

## Advanced weighted approach for class imbalance learning

### Enfoque avanzado de ponderación para el aprendizaje con desequilibrio de clases

Lamyae Benhlima<sup>1</sup>  , Mohammed El Haj Tirari<sup>1</sup>  

<sup>1</sup>National Institute of Statistics and Applied Economics, Laboratory of Methods Applied in Statistics, Actuaries, Finance and Quantitative Economics. Rabat, Morocco.

Cite as: Benhlima L, El Haj Tirari M. Advanced Weighted Approach for Class Imbalance Learning. Data and Metadata. 2025; 4:719. <https://doi.org/10.56294/dm2025719>


Submitted: 03-07-2024

Revised: 11-11-2024

Accepted: 14-05-2025

Published: 15-05-2025

Editor: Dr. Adrián Alejandro Vitón Castillo 

Corresponding Author: Lamyae Benhlima 

#### ABSTRACT

Predictive models derived from statistical learning techniques often assume that data originate from simple random sampling, thus assigning equal weight to all individuals. However, this assumption faces two significant challenges: it overlooks the complexity of real samples, where individuals may have different sampling weights, and it introduces a bias toward the majority class in imbalanced datasets. In this study, we propose an innovative approach that introduces differentiated weights for individuals by adjusting sample weights through calibration. This method aims to address class imbalance issues while improving the representativeness of samples. We applied it to the Support Vector Machine. Additionally, we developed an improved adjusted weighting approach to further enhance model performance, particularly for the minority class. This improved version combines two widely used techniques for handling class imbalances (resampling and cost-sensitive learning) by first balancing the classes through resampling, then applying adjusted sample weights during training. We evaluated the performance of our approach on real datasets with varying levels of imbalance using multiple evaluation metrics. The results were compared with various conventional methods commonly employed to address class imbalance. Our findings demonstrate the relevance and generalizability of our proposed algorithms, which often achieve performance equal to or better than that of established competing methods. Overall, our methodology not only corrects sample imbalances but also ensures a more accurate representation of the target population in the model, making it a robust and flexible solution for real-world imbalanced classification challenges.

**Keywords:** Representativeness; Sample Weights; Calibration; Class Imbalance; Support Vector Machine.

#### RESUMEN

Los modelos predictivos derivados de técnicas de aprendizaje estadístico a menudo asumen que los datos provienen de un muestreo aleatorio simple, asignando así un peso igual a todos los individuos. Sin embargo, esta suposición enfrenta dos desafíos significativos: pasa por alto la complejidad de las muestras reales, donde los individuos pueden tener diferentes pesos de muestreo, e introduce un sesgo hacia la clase mayoritaria en conjuntos de datos desequilibrados. En este estudio, proponemos un enfoque innovador que asigna pesos diferenciados a los individuos ajustando los pesos de la muestra mediante calibración. Este método tiene como objetivo abordar los problemas de desequilibrio de clases al mismo tiempo que mejora la representatividad de las muestras. Lo aplicamos a la Máquina de Vectores de Soporte. Además, desarrollamos un enfoque de ponderación ajustada mejorado para potenciar aún más el rendimiento del modelo, especialmente para la clase minoritaria. Esta versión mejorada combina dos técnicas ampliamente utilizadas para tratar el desequilibrio de clases (reejemplado y aprendizaje sensible al costo) al equilibrar inicialmente las clases mediante reejemplado, para luego aplicar pesos ajustados durante el entrenamiento.

Evaluamos el rendimiento de nuestro enfoque en conjuntos de datos reales con diversos niveles de desequilibrio utilizando múltiples métricas de evaluación. Los resultados se compararon con varios métodos convencionales comúnmente empleados para abordar el desequilibrio de clases. Nuestros hallazgos demuestran la relevancia y la generalizabilidad de los algoritmos propuestos, los cuales a menudo alcanzan un rendimiento igual o superior al de los métodos competidores establecidos. En general, nuestra metodología no solo corrige los desequilibrios de la muestra, sino que también asegura una representación más precisa de la población objetivo en el modelo, constituyéndose en una solución robusta y flexible para los desafíos de clasificación desequilibrada en contextos reales.

**Palabras clave:** Representatividad; Pesos de Muestra; Calibración; Desequilibrio de Clases; Máquina de Vectores de Soporte.

## INTRODUCTION

As machine learning models are increasingly employed in critical decision-making areas such as medical diagnostics, fraud detection, and predictive maintenance, the need for highly accurate and reliable models becomes paramount. The effectiveness of these models fundamentally depends on the quality of their data. This aspect has been analyzed from various perspectives, including data complexity,<sup>(1)</sup> missing values,<sup>(2)</sup> noise,<sup>(3)</sup> and dataset shift.<sup>(4)</sup> A key concept in ensuring data quality is data representativity, which ensures that models reflect the diverse conditions and scenarios they will encounter in real-world applications.<sup>(4,5)</sup> Without representativity, models are prone to poor performance and limited generalizability, highlighting the critical need for training data that accurately mirrors the target environments.

This paper explores a significant challenge in machine learning: class imbalance, a prominent example of data non-representativity that severely impairs model efficacy. Often resulting from selection bias, class imbalance manifests through a disproportionate representation of classes within the dataset.<sup>(6)</sup> This skew in the learning process affects both the model's performance and its capacity for generalization. Consequently, models trained on such data tend to exhibit diminished sensitivity towards minority classes, display misleadingly high accuracy metrics, struggle with generalization on unseen data, and face heightened risks of overfitting. Addressing this imbalance is crucial for achieving equitable and accurate predictive outcomes in machine learning applications. The issue of class imbalance, prevalent in various real-world applications, has attracted significant interest from researchers.<sup>(7)</sup> Various strategies like data resampling, cost-sensitive methods, and active learning have been proposed to address this issue. In this work, we introduce a novel approach inspired by cost-sensitive learning, involving the use of a differential system for weighting individuals based on calibrated sample weights. We specifically focus on the Support Vector Machine (SVM) model, to demonstrate the effectiveness of our approach.<sup>(8)</sup>

Traditional SVMs often favor the majority class, which reduces their generalization performance due to uniform penalties for misclassifications across all samples.<sup>(9)</sup> To address this, we propose the Adjusted Weighted SVM (AW-SVM) model, which enhances data representativity by assigning weights according to each sample's complexity and representativity relative to the target population. By leveraging prior knowledge of the overall population, calibrated sample weights help to bridge the gap between individual samples and the broader population, mitigating the effects of data imbalance. Building on this, we introduce an Improved Adjusted Weighted SVM (IAW-SVM) model that combines cost-sensitive learning with data resampling. This improved version first addresses class distribution imbalances through resampling and then applies the Adjusted Weighted SVM with calibrated sample weights. Our approach, which integrates calibrated weights with data resampling, achieves superior performance compared to traditional weighted methods, such as Weighted Support Vector Machine (WSVM), as well as data sampling-based preprocessing techniques, including Data Sampling applied to SVM (DS-SVM). This combination ensures a more representative dataset, leading to better classification results, improved generalization, and more reliable machine learn outcomes. Thus, it offers a valuable solution for various real-world applications requiring accurate data representation.

The remainder of the paper is structured as follows: first, we provide a brief overview of SVM, including its cost-sensitive training approaches WSVM, and a focused discussion on data resampling techniques. We then present the sample weight generation approach, the proposed AW-SVM models, and their enhanced version. Next, we detail the experimental results. Finally, we conclude with a summary of key findings.

## Background study

SVM and its variants, such as WSVM, are widely used in machine learning for classification tasks. To address class imbalance, techniques like data resampling and weighted learning have been developed, improving model

performance by ensuring better representation of minority classes. This section aims to deepen the understanding of the modeling process of these models and present the main approaches for handling imbalanced data.

### Support Vector Machine

In the field of machine learning, SVM models have gained widespread adoption among practitioners. Initially introduced by Cortes and Vapnik for binary classification tasks,<sup>(10)</sup> SVM is recognized as a robust supervised kernel-based method. Central to its effectiveness is the principle of structural risk minimization, which helps minimize the upper bound of generalization error and thereby grants SVM superior generalization capabilities compared to many alternative supervised learning approaches.<sup>(11)</sup>

SVM has been successfully applied in many application areas, addressing a wide range of problems, including both classification and regression tasks.<sup>(12,13)</sup>

Giving a set of training pairs:

$$(x_i, y_i), i = 1, 2, \dots, l$$

Where:

$x_i \in \mathbb{R}^n$

Represents the  $i$ th  $n$  dimensional input point.

$y_i$  is its corresponding binary label (-1 or 1), SVM aims to classify data points into two classes based on their features.

The main concept of SVM involves transforming the input data from their original low-dimensional space to a higher-dimensional feature space using a nonlinear function  $\Phi(X)$ . Subsequently, the objective is to identify an optimal classifier in this higher-dimensional feature space that linearly separates the input data, typically represented as:

$$(w \cdot \Phi(x)) + b = 0 \quad (1)$$

Where:

“.” is a scalar product.

$b$  represents the offset of the hyperplane from the origin.

$w = (w_1, \dots, w_n)$  denotes the weight vector of  $n$  elements that determine the direction of the optimal separating hyperplane.

The classifier should satisfy the following constraints in the high-dimensional feature space:

$$\begin{cases} (w \cdot \Phi(x_i)) + b \geq 1 & \text{if } y_i = 1 \\ (w \cdot \Phi(x_i)) + b \leq -1 & \text{if } y_i = -1 \end{cases} \quad (3)$$

The quest for the optimal classifier involves maximizing the margin  $2/||W||$  between the hyperplanes, based on the principle of structural risk minimization. This maximization is reformulated to streamline problem-solving as the minimization of  $||W||/2$ . To address misclassifications, an error term  $\psi$  also known as slack variable is introduced to relax the constraints. With these adjustments, the pursuit of the optimal classifier is reframed as the following optimization problem:

$$\min \frac{1}{2} ||w||^2 + C \sum_{i=1}^l \psi_i \quad (4)$$

$$\text{s.t., } y_i(w \cdot \Phi(x_i) + b) \geq 1 - \psi_i, \quad \psi_i \geq 0, \quad i = 1, \dots, l$$

Where:

$C$  denotes a regularization parameter of this model, influencing the balance between maximizing the margin and classification violation.

The expression  $2/||W||^2$  is a smoothed version of  $2/||W||$ , making it suitable for convex quadratic programming (QP). Typically, to solve this QP problem, equation (4) is transformed into its Wolfe dual form by adding the Lagrangian multiplier  $0 \leq \alpha$  and solving it by applying the Karush-Kuhn-Tucker condition,<sup>(14)</sup> the obtained dual problem can be expressed as:

$$\min \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j (\Phi(x_i) \cdot \Phi(x_j)) \alpha_i \alpha_j - \sum_{j=1}^l \alpha_j \quad (5)$$

$$\text{s.t., } 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^l \alpha_i y_i = 0, \quad i = 1, \dots, l$$

The product of  $\Phi(x_i)$  and  $\Phi(x_j)$  can be defined as kernel function  $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$  knowing the exact form of the kernel function is essential in the problem-solving process, as long as it meets the Mercer condition.<sup>(15)</sup>

After solving the quadratic problem (5) and obtaining the optimal values  $\alpha^*(\alpha_1, \dots, \alpha_l)$ , we determine the unknown variables and as mentioned in the equations below:

$$w^* = \sum_i \alpha_i^* y_i \Phi(x_i)$$

$$b^* = - \frac{\sum_i \sum_j y_i \alpha_i^* \alpha_j^* K(x_i, x_j)}{\sum_i \alpha_i^*} \quad (6)$$

For any new instance the decision function is obtained as:

$$f(x) = \text{sign}(w^* \cdot \Phi(x) + b^*) \quad (7)$$

SVM models must solve the quadratic programming problem to find the optimal hyperplane. The complexity of this problem, as revealed by equation 4, directly correlates with the number of training instances. Consequently, resolution time is significantly prolonged for large-scale problems, leading to increased computational costs. In response to this challenge, various effective algorithms have been developed, including chunking and sequential minimal optimization algorithms to improve classification accuracy by selecting optimal parameters via heuristic search.

### Weighted SVM (WSVM)

WSVM enhance traditional SVM models by assigning different weights to instances based on their importance. Unlike conventional SVM, which treats all instances equally, WSVM addresses noise, outliers, and class imbalance issues by giving more influential data points greater impact on the decision boundary. This is achieved by adjusting the regularization parameter  $c$ , allowing the model to emphasize key data points and create a more reliable decision boundary. This results in improved classification performance, especially in noisy or imbalanced datasets. The constrained optimization problem for WSVM is described as follows:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l W_i \psi_i \quad (8)$$

$$\text{s.t., } y_i (w \cdot \Phi(x_i) + b) \geq 1 - \psi_i, \quad \psi_i \geq 0, \quad i = 1, \dots, l$$

In the given formulation, the data point  $x_i$  is assigned a weighting factor  $W_i$  based on its class membership, leading to the following dual formulation:

$$\min \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j K(x_i, x_j) \alpha_i \alpha_j - \sum_{j=1}^l \alpha_j \quad (9)$$

$$\text{s.t., } \sum_{i=1}^l \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C W_i, \quad i = 1, \dots, l$$

By calculating  $\alpha_i$  we determine the unknown variables  $w$   $b$  as outlined in the following equations:

$$w^* = \sum_i \alpha_i^* y_i \Phi(x_i),$$

$$b^* = \frac{1}{N_{NSV}} \sum_{x_i \in NSV} (y_i - \sum_{x_j \in NSV} \alpha_j^* y_j K(x_i, x_j)) \quad (10)$$

Where:

$N_{NSV}$  in the number of normal support vectors, for any new instance the decision function is obtained as:

$$f(x) = \text{sign}(w^* \cdot \Phi(x) + b^*) \quad (11)$$

By emphasizing important data points, WSVM improves the robustness of the model and offers a more adaptive and effective solution for complex datasets with varying data point importance.

#### Data sampling applied to SVM (DS-SVM)

This paper focuses primarily on data resampling techniques applied to SVM (DS-SVM), which include oversampling, undersampling, and hybrid approaches.<sup>(16,17)</sup> These methods adjust class ratios to ensure that each class contributes equally to the learning process, thereby enhancing model fairness and accuracy. Both oversampling and undersampling aim to adjust the ratios between majority and minority classes. Hybrid techniques combine both approaches to leverage the benefits of each, creating a more balanced dataset. By rebalancing the training data, resampling allows different classes to have a more equal impact on the outcomes of a classification model. These traditional rebalancing methods have shown some potential in mitigating the class imbalance problem. However, they often disrupt the original data structure and fail to preserve the distribution of data during resampling. As a result, these methods tend to lose important information from the original datasets, which may limit classification accuracy. This limitation further intensifies the problem of datasets not being representative of the overall population.

#### METHOD

WSVM and WLSSVM models have been widely studied in machine learning to address various challenges, particularly in assessing model sparsity. In these models, weights are derived from error variables, and robust estimates are obtained using standard deviation.<sup>(18)</sup> Yang et al.<sup>(19)</sup> developed a new weighted SVM to reduce outlier insensitivity by employing a robust fuzzy clustering algorithm for weight generation. Tomar et al.<sup>(20)</sup> presented a weighted least squares twin SVM specifically for addressing class imbalance issues. Xia et al.<sup>(21)</sup> proposed a relative density-based SVM for noisy data classification, ranking points based on their relative density to assign higher values to more important data. Hazarika et al.<sup>(22)</sup> introduced a density-weighted SVM model for binary class imbalance issues.

Building on previous research, our approach aims to mitigate the effects of class imbalance by enhancing data representativity. To achieve this, we introduce the Adjusted Weighted Support Vector Machine (AW-SVM) and the Adjusted Weighted Least Squares Support Vector Machine (AW-LSSVM). Additionally, we propose their improved versions, IAW-SVM and IAW-LSSVM, which further refine the weighting strategy to optimize model performance.

#### Weight generation approach

To minimize bias and improve model performance in addressing class imbalance, we generate adjusted sample weights by combining sample design and calibration techniques based on available auxiliary information. These weights are crucial for enhancing the overall representativity and effectiveness of our models.

The concept of individual weighting has been extensively studied, with foundational work by Horvitz et al.<sup>(23)</sup> and Hansen<sup>(24)</sup> which involves weighting units by the inverse of their inclusion probabilities. The weight assigned to each individual, expressed as  $w_i = 1/\pi_i$ , where  $\pi_i$  is the probability of selection for individual  $i$ , reflects the number of target population members that the sampled individual represents.

In scenarios where sampling weights are unknown, such as in many machine learning datasets, we generate artificial weights for training samples. Initially, we assume the test sample data is selected using simple random sampling, equivalent to not applying weights for parameter estimation. We then refine these weights through calibration techniques, ensuring they align with known totals of specific auxiliary variables.<sup>(25)</sup> This refinement process incorporates appropriate weighting information into the parameter estimation, thereby enhancing the sample's representativity and improving the model's accuracy.

In classification problems, ensuring well-represented classes is crucial. Therefore, we employ a stratified sampling design instead of simple random sampling. This involves selecting independent samples from each stratum based on their respective sizes and the overall population size.

After determining these sampling weights, we apply calibration techniques to adjust for potential estimation biases and align sample characteristics with the population. Calibration adjusts weights using auxiliary information that is not included in the original datasets. This process corrects discrepancies between the sample and the population, thereby enhancing both representativity and precision in statistical estimates.<sup>(26)</sup> By aligning the sample characteristics more closely with those of the population, calibration helps to reduce potential biases in the estimation process. The specific method used for calibration weight adjustment in this study follows the Samplics approach.<sup>(27)</sup>



### Adjusted and Improved Weighted Support Vector Machine (AW-SVM and IAW-SVM)

The AW-SVM model searches for a classifying hyperplane in the input space  $w^* \cdot \Phi(x) + b^* = 0$  and applies a weight vector to the hinge loss term, adjusting the contribution of each sample based on its importance and representativity, thus the classifying hyperplane may be obtained by solving the optimization problem as:

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l S_i \psi_i \quad (12)$$

$$\text{s.t., } y_i(w \cdot \Phi(x_i) + b) \geq 1 - \psi_i, \quad \psi_i \geq 0, \quad i = 1, \dots, l$$

Unlike traditional weighting methods that allocate weights to individuals based on their class membership, our approach assigns varying weights according to each individual's representativity within the overall population. In this context,  $S = (S_1, \dots, S_l)$  represents the adjusted sample weight vector. To solve the constrained problem (12), the dual problem is determined by adding the Lagrangian multiplier and solving it by applying the KKT condition. The equation may be expressed as:

$$\min \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l y_i y_j K(x_i, x_j) \alpha_i \alpha_j - \sum_{j=1}^l \alpha_j$$

$$\text{s.t., } \sum_{i=1}^l \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C S_i, \quad i = 1, \dots, l \quad (13)$$

The unknown variables  $w$  and  $b$  can be computed after the calculation of  $\alpha_i$ . For any new datapoint, the decision classifier may be expressed as similar to equation (11).

The IAW-SVM applies the AW-SVM to preprocessed data samples using a selected data sampling technique. The objective is to enhance representativity while mitigating class imbalance, thereby improving predictive performance for the minority class. The implementation of IAW-SVM follows a structured sequence of steps, detailed as follows:

- Step 1: define the training dataset  $(x_i, y_i)$  for  $i = 1, 2, \dots, l$

Where

$X_i \in \mathbb{R}^n$ . This dataset consists of input vector  $x_i$

Their corresponding label  $y_i$ .

- Step 2: apply a data sampling technique. This preprocessing step can involve oversampling, undersampling, or hybrid techniques. The resulting training dataset is noted as  $(x_i, y_i)$  for  $i = 1, 2, \dots, s$ .
- Step 3: identify auxiliary variables to aid in weight calculation. These should include the variable most correlated with the target variable and a variable representing the original class sizes.
- Step 4: calculate the adjusted sample weights based on the chosen sample design and auxiliary variables.
- Step 5: select an appropriate kernel function for the AW-SVM.
- Step 6: train the AW-SVM using the selected kernel function and perform hyperparameter tuning.

Key hyperparameters to adjust include:

1. The penalty parameter  $C$ .
2. The polynomial degree  $d$  (for polynomial kernels).
3. The kernel coefficient  $\gamma$  (for RBF kernels).

- Step 7: compute the Lagrangian multipliers  $\alpha_i$  from the dual problem formulation. Use these multipliers to determine the optimal hyperplane parameters  $w^*$  and  $b^*$ .

- Step 8: to classify a new instance  $X_i \in \mathbb{R}^n$  use the trained IAW-SVM model with the decision function:

$$f(x) = \text{sign}(w^* \cdot \Phi(x) + b^*).$$

### RESULTS AND DISCUSSION

The results of our simulation experiments, designed to evaluate the effectiveness of the proposed models, are presented in this section. We begin by describing the benchmark datasets sourced from the UCI repository and outlining the evaluation metrics used to address class imbalance. Following this, we provide details on the simulation setup and present the results for binary classification tasks.

First, we compare the performance of AW-SVM with the original models, SVM. Then, we compare the Improved versions, IAW-SVM with W-SVM and DS-SVM as well as with the original models, SVM and LSSVM. The experiments are conducted in Jupyter 6.4.8 on a system equipped with 8 GB of RAM, 500 GB of storage, and an Intel Core i7 processor operating at 1,50 GHz.

### Datasets analysis

We evaluated our approach on ten real-world datasets, covering a wide range of sample sizes (from a few hundred to over 200 000 instances), feature counts (9 to 30), and class imbalance ratios (1,9 to 11,9). This diversity ensures a comprehensive assessment of our methods across various data scales and complexities, highlighting their robustness in different application scenarios.

Each dataset underwent rigorous preprocessing. First, we identified and mitigated outliers, which could otherwise distort model performance, particularly in SVM-based learning. Next, categorical features were encoded, and numerical attributes were standardized using min-max scaling. Finally, the datasets were split into training and testing sets to evaluate model performance on unseen data, ensuring a reliable assessment of generalization capability.

### Evaluation metrics

Choosing the right evaluation metrics is crucial for accurately assessing the performance of predictive models, especially in the context of imbalanced datasets. Standard metrics like accuracy often fail to provide meaningful insights in such scenarios, as they can be heavily biased towards the majority class.<sup>(28)</sup> Therefore, selecting evaluation metrics that appropriately address class imbalance is essential for developing robust and effective models.

In this study, we focus on the G-mean and FB Measure as our primary evaluation metrics. The G-mean is particularly valuable because it balances the true positive rate and the true negative rate, ensuring that the model performs well across both minority and majority classes. The F2-Measure is a specific case of the more general FB Measure, with  $B=2$ . It places more emphasis on recall than precision, which is crucial when the accurate identification of the minority class is of greater importance. By weighting recall more heavily, the F2-Measure ensures that the model prioritizes capturing the minority class, which is often the focus in imbalanced datasets.<sup>(29)</sup>

### Simulations analysis

In this section, we present the experimental results and compare the proposed algorithms for binary-class problems with various classic algorithms designed to address class imbalance issues. Our approach incorporates adjusted sample weights into predictive models. Since survey weights for each individual were not available in all datasets, we adopted an optimal stratified sampling approach to create a representative sample of the entire population, as explained in the weight generation subsection. The main goal was to generate a weight vector to study the impact of considering the differential weighting of individuals on the learning model.

After drawing the sample, we employed a calibration technique to adjust the weight vector and improve the precision of the estimators. This iterative process modifies the weights until the estimates from the calibrated sample align optimally with the true population values. To ensure consistency and reliability, we repeated this adjustment for different samples (100 samples in total) for each dataset. Through this rigorous approach, we could precisely evaluate the influence of survey weights on the learning model, leading to significant findings regarding the importance of this differentiated weighting approach. The average results for 100 samples from each dataset were obtained in terms of F2-Measure and G-Mean.

We conducted two comparison experiments. In the first experiment, we compared AW-SVM with the basic SVM algorithm. In the second experiment, we compared IAW-SVM with WSVM, DS-SVM, and the original model SVM.

### Comparison of AW-SVM versus SVM

Table 1 presents the average results for the algorithms AW-SVM and SVM, for the F2-Measure and G-Mean across 10 datasets.

Table 1. Classification results based on G-mean and F2-Measure						
Datasets	G-mean			F2-Measure		
	SVM	AW-SVM	Err1	SVM	AW-SVM	Err2
Abalone	16,42	73,54	+57,12	6,01	37,0	+30,99
Pima	66,49	70,61	+4,12	56,56	60,89	+4,33
Segment	90,76	88,47	-2,29	84,69	82,86	-1,83
Wine quality-red	59,88	62,53	+2,65	40,39	44,2	+3,81

Bank marketing	35,24	53,16	+17,92	19,59	37,73	+18,14
Tyroid	90,09	91,96	+1,87	83,52	86,27	+2,75
Whine quality-white	63,22	64,49	+1,27	47,58	51,28	+3,7
Autism	94,93	97,04	+2,11	92,37	96,63	+4,26
Vehicle1	65,51	68,45	+2,94	52,03	56,71	+4,68
Diabetes	35,92	41,07	+5,15	17,61	20,98	+3,37

Based on these results, AW-SVM significantly outperform the basic models SVM in nearly all cases, achieving superior performance in 9 out of 10 datasets across both metrics, F2-Measure and G-Mean. Furthermore, our proposed models have effectively reduced the effect of data imbalance, as demonstrated by the enhanced performance in these metrics.

#### *Comparison of IAW-SVM with Basic Models and Popular Data Imbalance Techniques*

Although the approach proposed previously reduced the effect of class imbalance by improving performance indicators, the results were not sufficiently satisfactory in certain cases. This prompted us to develop an improved version to address the same issues of imbalance and representativity while also being competitive with popular techniques in this field. Therefore, in this section, we compare our improved imbalanced classification algorithms, IAW-SVM, with two widely used techniques for handling class imbalance: cost-sensitive learning and data resampling.

The cost-sensitive approach assigns weights to instances inversely proportional to their class distribution in the training data (W-SVM). In contrast, the data resampling technique involves adjusting the sample distribution before training the models (DS-SVM). Table 2 and table 3 presents the simulation results for four models, including our proposed model IAW-SVM, evaluated using F2-Measure and G-Mean, respectively.

**Table 2.** Classification results and rankings based on G-mean

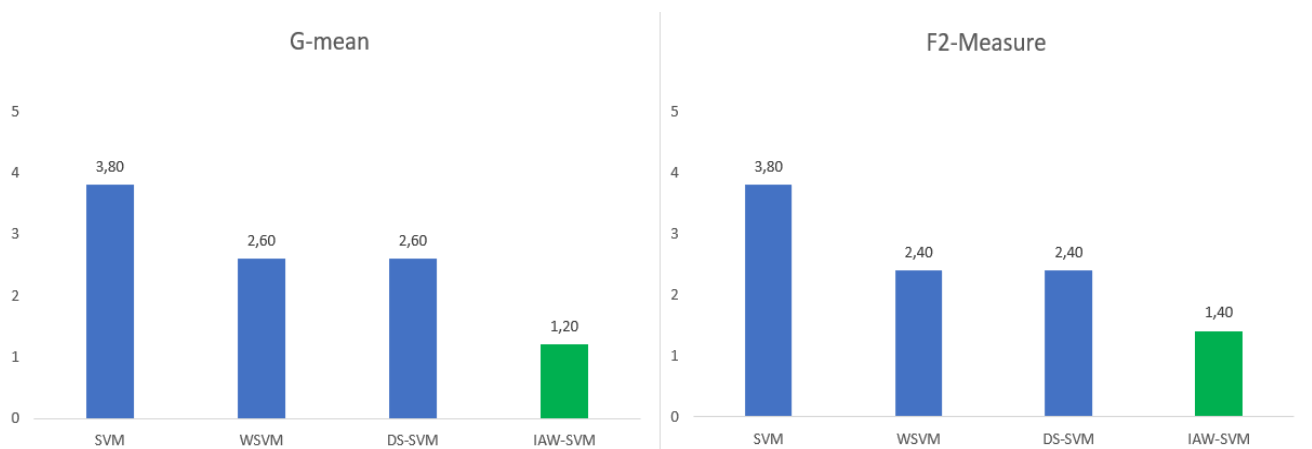
Datasets	SVM	Rank	WSVM	Rank	DS-SVM	Rank	IAW-SVM	Rank
Abalone	16,42	(4)	54,59	(3)	58,14	(2)	74,77	(1)
Pima	66,49	(4)	71,86	(2)	70,08	(3)	72,79	(1)
Segment	90,76	(3)	91,09	(2)	90,74	(4)	91,14	(1)
Wine quality-red	59,88	(4)	64,2	(3)	67,95	(2)	76,89	(1)
Bank marketing	35,24	(4)	69,41	(2)	66,62	(3)	70,0	(1)
Tyroid	90,09	(4)	92,81	(3)	95,15	(2)	97,68	(1)
Whine quality-white	63,22	(4)	67,13	(3)	71,78	(1)	71,55	(2)
Autism	94,93	(3)	96,78	(2)	69,0	(4)	97,15	(1)
Vehicle1	65,51	(4)	68,35	(3)	71,84	(2)	72,42	(1)
Diabetes	35,92	(4)	59,19	(3)	69,86	(1)	66,23	(2)
Average Rank		3,8		2,6		2,6		1,2

**Table 3.** Classification results and rankings based on F2-Measure

Datasets	SVM	Rank	WSVM	Rank	DS-SVM	Rank	IAW-SVM	Rank
Abalone	6,01	(4)	29,24	(2)	29,12	(3)	36,82	(1)
Pima	56,56	(4)	65,22	(2)	64,54	(3)	68,51	(1)
Segment	84,69	(2)	84,91	(1)	83,32	(4)	83,97	(3)
Wine quality-red	40,39	(4)	43,46	(3)	46,59	(2)	56,24	(1)
Bank marketing	19,59	(4)	54,86	(2)	50,87	(3)	55,29	(1)
Tyroid	83,52	(4)	87,81	(3)	90,16	(2)	94,38	(1)
Whine quality-white	47,58	(4)	54,98	(3)	66,21	(1)	63,97	(2)
Autism	92,37	(4)	96,47	(2)	95,65	(3)	96,76	(1)
Vehicle1	52,03	(4)	56,16	(3)	62,61	(2)	67,7	(1)
Diabetes	17,61	(4)	39,35	(3)	53,67	(1)	50,4	(2)
Average Rank		3,8		2,4		2,4		1,4



The results show that IAW-SVM and IAW-LSSVM consistently achieve superior performance across most datasets, as indicated by their average ranks (noted in parentheses) based on F2-Measure and G-Mean values. Their effectiveness is further illustrated which presents a bar plot of the average performance across various real-world datasets. This visualization confirms that IAW-SVM outperform existing cost-sensitive and data resampling techniques, demonstrating a significant improvement in classification metrics for imbalanced datasets.



**Figure 1.** Average Ranking of Algorithms Based on F2-Measure and G-Mean

### Performances validation

Evaluating and comparing the performance of classification algorithms is crucial to determine their suitability for different datasets and tasks. To this end, we employed the Friedman test, a non-parametric statistical method designed to identify differences in the distributions of paired samples. This test is particularly effective for comparing algorithms across multiple datasets, as it allows us to assess their relative performance systematically. In this study, we analyzed the average ranks of eight classifiers across all datasets.

The Friedman test was conducted using G-mean and F2-measure as evaluation metrics, comparing the following algorithms: SVM, W-SVM, DS-SVM, and IAW-SVM. The results are summarized in table 4.

Metrics	Stat	CV	p-value	Significance Level	H0
G-mean	20,04	7,81	$1,40 \times 10^{-4}$	0,05	Rejected
F2-measure	17,51	7,81	$5,52 \times 10^{-4}$	0,05	Rejected

Given these extremely low p-values, combined with the fact that the test statistics exceed the critical thresholds, lead to the rejection of the null hypothesis. These findings confirm that the performance differences among the tested algorithms are statistically significant. Such insights highlight the importance of carefully selecting classification algorithms based on the evaluation metrics relevant to the task at hand.

To further analyze the observed performance differences, we conducted pairwise comparisons using the Wilcoxon test, based on both G-mean and F2-measure metrics. This non-parametric test identifies specific pairs of algorithms that exhibit statistically significant performance differences. The results of the Wilcoxon test revealed several significant differences between the evaluated algorithms. For the G-mean metric, IAW-SVM exhibited significant differences compared to SVM, WSVM, and DS-SVM ( $p < 0,05$ ). For the F2-measure, IAW-SVM also demonstrated statistically significant differences when compared to SVM, DS-SVM, and WSVM.

Overall, these results confirm that the tested algorithms do not perform equally, with certain models (particularly IAW-SVM) showing substantial differences when compared to traditional approaches like SVM. These findings highlight the importance of carefully selecting classification algorithms based on performance metrics relevant to the specific task.

### CONCLUSIONS

This paper presents a novel adjusted weighting strategy based on sample weighting to improve data representativity in handling imbalanced data classification problems. By focusing on the amelioration of data representativity through the incorporation of adjusted sample weights into SVM, we developed the AW-SVM approach. These methods stand out by using known population information not included in the dataset, through a calibration technique, to make samples more representative. Our study addresses the challenge of enhancing representativity in cases where datasets present class imbalance problems. Attributing differential system

weighting to individuals based on their degree of representativity significantly enhances model performance concerning the minority class. Experimental comparisons of our proposed models with traditional models on binary class imbalanced datasets demonstrate significant improvements in two key evaluation metrics: F2 measure and G-mean.

Furthermore, we developed an enhanced version of our models by integrating cost-sensitive techniques with adjusted survey weights and data resampling methods. Simulation results using this improved approach, IAW-cSVM, demonstrate competitive performance against well-established techniques, including W-SVM, DS-SVM, and DS-SVM. For future research, the adjusted weighting methodology could be extended to other SVM variants, such as Twin SVM, Least square SVM and Proximal SVM, to further enhance data representativity and generalization capabilities. Beyond SVM based classifiers, incorporating adjusted sample weights into other learning models such as decision trees and neural networks offers a promising avenue for addressing class imbalance across a broader range of applications.

## BIBLIOGRAPHIC REFERENCES

1. Ramalingam PA, Fathima N, Supriya P, Shetty P, Sanyal M, Yeshaswini P, et al. Data Complexity for Identifying Suitable Algorithms. In: 2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC). IEEE; 2023. page 955-61.
2. Thomas T, Rajabi E. A systematic review of machine learning-based missing value imputation techniques. *Data Technologies and Applications* 2021;55(4):558-85.
3. Gupta S, Gupta A. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science* 2019;161:466-74.
4. Suresh H, Gutttag JV. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:190110002* 2019;2(8):73.
5. Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. *Pattern Recognition* 2012;45(1):521-30.
6. Abd Elrahman SM, Abraham A. A review of class imbalance problem. *Journal of Network and Innovative Computing* 2013;1:9-9.
7. Ali A, Shamsuddin SM, Ralescu AL. Classification with class imbalance problem. *Int J Advance Soft Compu Appl* 2013;5(3):176-204.
8. Rezvani S, Wang X. A broad review on class imbalance learning techniques. *Applied Soft Computing* 2023;143:110415.
9. Batuwita R, Palade V. Class imbalance learning methods for support vector machines. *Imbalanced learning: Foundations, algorithms, and applications* 2013;83-99.
10. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;20(3):273-97.
11. Wang L, Gao C, Zhao N, Chen X. A projection wavelet weighted twin support vector regression and its primal solution. *Appl Intell* 2019;49(8):3061-81.
12. Hazarika BB, Gupta D, Ashu, Berlin M. A Comparative Analysis of Artificial Neural Network and Support Vector Regression for River Suspended Sediment Load Prediction. In: Luhach A, Kosa J, Poonia R, Gao XZ, Singh D, éditeurs. *First International Conference on Sustainable Technologies for Computational Intelligence*. Singapore: Springer; 2020.
13. Borah P, Gupta D. Functional iterative approaches for solving support vector classification problems based on generalized Huber loss. *Neural Comput & Applic* 2020;32(13):9245-65.
14. Fletcher R. *Practical methods of optimization*. John Wiley & Sons; 2000.
15. Smola AJ, Schölkopf B, Müller KR. The connection between regularization operators and support vector kernels. *Neural Networks* 1998;11(4):637-49.

16. Haibo He, Garcia EA. Learning from Imbalanced Data. *IEEE Trans Knowl Data Eng* 2009;21(9):1263-84.
17. Vluymans S. Learning from Imbalanced Data. In: *Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods*. Cham: Springer International Publishing; 2019. page 81-110. [http://link.springer.com/10.1007/978-3-030-04663-7\\_4](http://link.springer.com/10.1007/978-3-030-04663-7_4)
18. Suykens JA, De Brabanter J, Lukas L, Vandewalle J. Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing* 2002;48(1-4):85-105.
19. Yang X, Song Q, Wang Y. A WEIGHTED SUPPORT VECTOR MACHINE FOR DATA CLASSIFICATION. *Int J Patt Recogn Artif Intell* 2007;21(05):961-76.
20. Tomar D, Singhal S, Agarwal S. Weighted Least Square Twin Support Vector Machine for Imbalanced Dataset. *IJDTA* 2014;7(2):25-36.
21. Xia S, Xiong Z, Luo Y, Dong L, Xing C. Relative density based support vector machine. *Neurocomputing* 2015;149:1424-32.
22. Hazarika BB, Gupta D. Density-weighted support vector machines for binary class imbalance learning. *Neural Comput & Applic* 2021;33(9):4243-61.
23. Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association* 1952;47(260):663-85.
24. Hansen MH, Hurwitz WN. The Problem of Non-Response in Sample Surveys. *Journal of the American Statistical Association* 1946;41(236):517-29.
25. Deville JC, Särndal CE. Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association* 1992;87(418):376-82.
26. Valliant R, Dever JA, Kreuter F. Calibration and Other Uses of Auxiliary Data in Weighting. *Practical Tools for Designing and Weighting Survey Samples* 2013;349-95.
27. Diallo M. *samplings*: a Python Package for selecting, weighting and analyzing data from complex sampling designs. *JOSS* 2021;6(68):3376.
28. 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 Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 2011;42(4):463-84.
29. Brownlee J. *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. v1.2. Online: Machine Learning Mastery; 2020.

## FINANCING

The authors did not receive funding for the development of this article.

## CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

## AUTHORSHIP CONTRIBUTION

*Conceptualization*: Mohammed El Haj Tirari, Lamyae Benhlima.

*Methodology*: Mohammed El Haj Tirari, Lamyae Benhlima.

*Data Collection and Analysis*: Lamyae Benhlima.

*Supervision*: Mohammed El Haj Tirari.

*Writing - Original Draft Preparation*: Lamyae Benhlima.

*Writing - Review Editing*: Mohammed El Haj Tirari, Lamyae Benhlima.