Vitor Modesto Leitão

Instance Hardness–Based Relevance for Imbalanced Regression

Recife

2026

Vitor Modesto Leitão

Instance Hardness–Based Relevance for Imbalanced Regression

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de bacharel em Ciência da Computação.

Orientador (a): Juscimara Gomes Avelino

Recife

2026

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Leitão, Vitor Modesto.
Instance Hardness?Based Relevance for Imbalanced Regression / Vitor Modesto Leitão. - Recife, 2026.
9 p : il., tab.

Orientador(a): Juscimara Gomes Avelino
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Informática, Ciências da Computação - Bacharelado, 2026.
Inclui referências.

1. Imbalance Problems. 2. Regression. 3. Instance Hardness. I. Avelino, Juscimara Gomes. (Orientação). II. Título.

000 CDD (22.ed.)

Vitor Modesto Leitão

Instance Hardness–Based Relevance for Imbalanced Regression

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de bacharel em Ciência da Computação..

Aprovado em: 15/01/2026

Banca Examinadora:

Prof. Dr. Juscimara Gomes Avelino (Orientador)
Universidade Federal de Pernambuco

Prof. Dr. George Darmiton da Cunha Cavalcanti (Examinador Interno)
Universidade Federal de Pernambuco

Recife

2026

Instance Hardness–Based Relevance for Imbalanced Regression

Vitor Modesto Leitao, Juscimara Gomes Avelino
Centro de Informatica
Universidade Federal de Pernambuco
Recife-PE, Brazil
vml2, jga2@cin.ufpe.br

***Abstract*—Imbalanced regression problems occur when the target variable exhibits an asymmetric distribution, resulting in certain value ranges being underrepresented in the dataset. Traditional techniques for identifying these rare instances have limitations because they rely on the relevance function (ϕ) to define what is considered rare in the data distribution. This approach presents shortcomings in more complex scenarios, such as bimodal distributions, as it fails to adequately capture the notion of rarity by assigning fixed relevance values to all examples in the dataset, thereby compromising the distinction between truly rare and normal instances. Given these limitations, this study proposes an Instance Hardness-based relevance function, called the IH-based function, to identify rare instances in regression problems, as traditional relevance functions may fail in scenarios such as bimodal distributions, where rarity cannot be accurately inferred from the target values alone. By incorporating learning difficulty, the IH-based function offers a more reliable identification of truly rare instances.. Through experiments, we demonstrate that the IH-based function is able to correctly identify rare regions even in scenarios with bimodal distributions. Furthermore, the results show a significant improvement in model performance when using the IH-based function in balancing strategies such as Random Oversampling (RO) and Gaussian Noise (GN), compared to the traditional relevance function.**

***Index Terms*—Imbalance Problems, Regression, Instance Hardness.**

I. Introduction

In machine learning, imbalanced problems occur when the data distribution is asymmetric, such that certain regions of the space of interest contain a significantly smaller number of instances compared to other, more represented regions. This scenario has been widely studied in classification tasks [1]–[3], but it also arises in regression problems [4], where the target variable exhibits non-uniform distributions characterized by the presence of rare and hard-to-model values [5]. Imbalance in regression can negatively impact the performance of predictive models, as they tend to favor denser regions of the distribution at the expense of less frequent regions [6].

The identification of these rare instances is a fundamental step and, in the literature, has traditionally been carried out through the relevance function proposed in [7], which classifies instances as rare or normal based on the distribution of the target variable. This identification is mandatory for the application of data balancing strategies such as Random Oversampling (RO) and Gaussian Noise (GN) [8]. However, despite being widely used, the relevance function has limitations, especially in scenarios with bimodal distributions, as discussed in the literature [9], [10]. Figure 1 illustrates this type of distribution, in which the relevance function fails to properly identify underrepresented instances, as it either misclassifies instances or is unable to distinguish between rare and normal examples across different regions of the distribution.

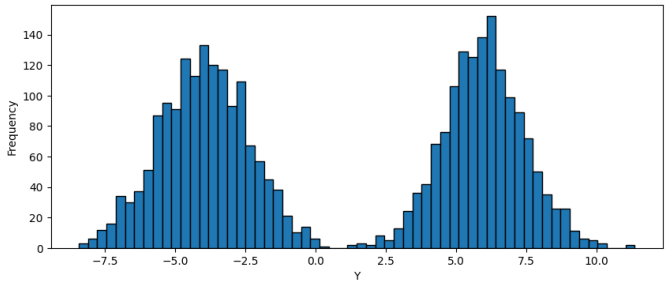

Fig. 1. Histogram of the target variable in the synthetic bimodal ensemble.

In light of this limitation, this work proposes an Instance Hardness-based relevance function (IH-based function) for regression problems, grounded in the Instance Hardness (IH) concept [11], to identify rare instances and overcome the limitations of traditional relevance functions, particularly their inability to distinguish different degrees of rarity in scenarios with bimodal distributions. In such situations, the relevance function tends to assign a static value to all instances, whereas the IH-based function allows continuous values to be assigned within the interval [0,1], where values closer to 1 indicate a higher degree of rarity. Unlike relevance functions based exclusively on statistical rarity, the concept of Instance Hardness captures the inherent difficulty of the learning process, such that an instance may be frequent in the data distribution and still be systematically mispredicted, revealing complex regions of the input space. Thus, it takes into account the difficulty of predicting an instance when determining whether it should be considered rare or common.

The evaluation of the proposed approach was conducted from three main perspectives: (i) the analysis of the correlation between Instance Hardness (IH) and the traditional relevance

function (ϕ); (ii) the assessment of the IH-based function in scenarios with bimodal distributions; and (iii) the use of the IH-based function in balancing strategies. In the first perspective, we investigated whether there is a significant correlation between the concepts of rarity and instance difficulty, that is, whether instances considered rare, associated with high values of ϕ, also correspond to difficult instances. In the second perspective, the effectiveness of the IH-based function in correctly identifying different degrees of rarity in regressions with bimodal distributions was validated, highlighting its ability to overcome the limitations of the traditional relevance function in these scenarios. Finally, in the third perspective, the use of the IH-based function as a criterion in balancing methods was analyzed, replacing the traditional relevance function in the definition of rare and normal instances. The results show that, in approximately 70% of the tests performed, the proposed approach achieved superior performance compared to the traditional approach.

II. Related Work

Imbalance in regression problems is widely discussed in the literature and occurs when certain regions of the target variable distribution are underrepresented, compromising the models' ability to predict rare values [5]. Approaches to deal with imbalanced regression problems can be organized into three main categories: regression models, modifications to the learning process, and evaluation metrics [6]. While individual and ensemble models have been widely used, their performance in imbalanced scenarios is often improved through modifications to the learning process, especially by means of preprocessing strategies based on resampling and cost-sensitive learning [7], [8]. Among these, methods such as Random Over-sampling, Gaussian Noise, SmoteR and its variants, SMOGN and SmoteR-Geometric, stand out [12], [13]. In addition, traditional global metrics, such as MSE, may not adequately reflect performance in rare regions of the distribution, which motivates the proposal of metrics specifically designed for imbalanced regression, such as the Squared Error Relevance Area (SERA) [14].

In this taxonomy, balancing strategies constitute the most explored approach in the literature for imbalanced regression [6]. Most of these techniques rely on the relevance function to identify rare instances and guide resampling procedures. Despite the advances achieved, recent studies highlight the limitations of these approaches, especially in scenarios with multimodal distributions, where the relevance function tends to assign the same degree of rarity to distinct instances [9]. To mitigate this problem, recent works have proposed distance-based alternatives, such as the Distance-Based Relevance Function [10], which aims to differentiate instances based on the local density of the distribution. Even so, difficulties persist in correctly identifying rare regions, since the definition of rarity depends exclusively on the marginal distribution of the target variable, disregarding information from the attribute space and the learning difficulty associated with the instances.

Given this scenario, a gap can be observed in the literature regarding the definition of rarity criteria that do not rely exclusively on the statistical distribution of the target variable. In this context, this work introduces a novel alternative grounded in the concept of Instance Hardness [15], proposing a new criterion for identifying rare cases based on the learning difficulty of instances. By exploring this perspective, the proposal seeks to contribute to the advancement of imbalanced regression techniques, expanding the scope of traditional approaches based on the relevance function.

III. The Proposed Relevance Function

The proposed relevance function, called the IH-based function, is based on the concept of Instance Hardness for regression [15], as defined in Equation 1, in which $h_j(x_i)$ represents the output of regressor j belonging to the set $\mathcal{L}$ for instance x_i. In this formulation, the difficulty of each instance is estimated from the error made by a set of models, adopting the normalization term defined as $\gamma = \frac{1}{n}\sum_i y_i^2$ in this work. By using instance difficulty as a criterion to identify rare cases in imbalanced regression problems, the method begins to consider information associated with the behavior of instances in the attribute space, offering an alternative to the relevance functions traditionally used in this context. In this sense, the IH-based function allows assigning continuous values in the interval [0, 1], where values closer to 1 indicate a higher degree of instance rarity, that is, instances that are more difficult to predict.

$$IH_{\mathcal{L}}(x_i, y_i) = 1 - \frac{1}{|\mathcal{L}|}\sum_{j=1}^{|\mathcal{L}|} \exp\left(-\frac{d\big(y_i, h_j(x_i)\big)}{\gamma}\right) \quad (1)$$

The IH-based function can be employed as a relevance criterion in data balancing techniques, such as Random Over-sampling (RO) and Gaussian Noise (GN), allowing the separation between rare and normal instances based on predictive difficulty. The Instance Hardness IH values are calculated individually for each instance, and those associated with the highest levels of difficulty can be considered rare through the adoption of a relative threshold, defined as a method parameter. In this work, the threshold is established based on the 70th percentile of the IH distribution, corresponding to the 30% most difficult instances in the dataset. Thresholds between 50% and 80% were experimentally evaluated, with the 70% threshold yielding the best results. The complete procedure is presented in Algorithm 1.

Algorithm 1 IH-based function

Require: Dataset $D = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, threshold τ
Ensure: Set of rare instances D_R, set of normal instances D_N
Calculate the values of IH:
    $IH \leftarrow \texttt{instance\_hardness}(D)$
Initialize $D_R \leftarrow \emptyset$, $D_N \leftarrow \emptyset$
for each instance $i \in D$ **do**
  if $IH_i \geq \tau$ **then**
    Label the instance i as rare
    Add i to the set D_R
  else
    Label the instance i as normal
    Add i to the seto D_N
  end if
end for
return D_R, D_N

IV. EXPERIMENTAL PROTOCOL

A. DATASET

To evaluate the behavior of the balancing method guided by the IH-based function in different imbalanced regression scenarios, we conducted a comprehensive experimental investigation involving 29 datasets. The main information about the datasets used is available in Table I, where instances whose ϕ [16] value is greater than 0.7 are considered rare.

TABLE I
SUMMARY OF THE 29 DATASETS USED, INCLUDING THE NUMBER OF INSTANCES, PREDICTOR ATTRIBUTES, AND THE PERCENTAGE OF RARE DATA.

Dataset	Rows	Features (X)	% rare data
a1	198	11	14.14
a2	198	11	11.11
a3	198	11	16.16
a7	198	11	13.64
abalone	4177	8	16.26
acceleration	1732	14	5.14
airfoild	1503	5	10.71
analcatdata_apnea3	450	11	22.89
available_power	1802	15	8.71
boston	506	13	17.98
cocomo_numeric	60	56	16.67
compactiv	8192	21	8.70
concreteStrength	1030	8	5.34
cpu_small	8192	12	8.70
debutanizer	2394	7	10.03
fuel_consumption_country	1764	37	9.30
forestFires	517	12	15.28
heat	7400	11	8.97
kdd_coil_1	316	18	10.76
lungcancer_shedden	442	24	5.66
maximal_torque	1802	32	7.16
meta	528	65	20.45
mortgage	1049	15	10.10
pdgfr	79	320	12.66
sensory	576	11	11.98
space_ga	3107	6	5.57
treasury	1049	15	10.39
triazines	186	60	10.75
wine	6497	11	23.44

B. MACHINE LEARNING MODELS

To evaluate the impact of balancing strategies based on RO and GN [17], different Machine Learning models were employed: RandomForestRegressor (RF), BaggingRegressor (BG), XGBRegressor (XGB), Support vector machine (SVR), and MLPRegressor, which were trained on the original dataset, on the dataset balanced with traditional balancing methods, and on the dataset balanced with the IH-based function.

C. INSTANCE HARDNESS

The concept of Instance Hardness (IH) is based on calculating the prediction difficulty of each instance individually [18]. To perform these calculations, multiple formulas can be used, most of which employ Machine Learning models to predict the data and subsequently compare the predicted target with the actual one, considering as difficult those instances with higher error rates.

The main IH models tested use a pool of models [15] to train a Machine Learning [19] model, as described in Subsection IV-B, and subsequently compute the distance between the true and the predicted target, resulting in a value between 0 and 1 for each instance; the closer the value is to 1, the more difficult the instance is to predict.

D. METRICS

To evaluate the performance of the models on each dataset, two traditional metrics commonly used in regression problems were employed: the Mean Absolute Error (MAE) and the Mean Squared Error (MSE), as defined in Equations 2 and 3. These metrics allow for a more comprehensive analysis of model performance. We chose not to employ metrics specific to imbalanced regression, such as SERA [14] or adaptations of the F1-score [20], since these metrics depend directly on a previously defined relevance function. As this work proposes a new relevance function based on Instance Hardness, the use of metrics that explicitly incorporate a relevance function could favor the proposal or introduce additional biases into the evaluation. Thus, adopting metrics independent of the relevance function allows for a fairer and more neutral comparison among the evaluated methods.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

E. EVALUATION METHODS

For the experiments comparing the traditional balancing approach with the IH-based function, we employed the *Repeated K-Fold* cross-validation method. Specifically, a configuration with 5 folds and 2 repetitions (i.e., 2×5 cross-validation) was adopted, resulting in 10 distinct data partitions. At each repetition, the data are randomly redistributed among the folds, leading to more stable and reliable estimates of model performance.

F. BALANCING METHODS

To address data imbalance, two classical balancing methods were used: Random Oversampling (RO) and Introduction of Gaussian Noise (GN). RO consists of the random replication of rare instances, while GN generates new synthetic instances by adding Gaussian noise to the attributes of rare instances, while simultaneously removing part of the most frequent instances. These two strategies were chosen because they presented the best empirical performance in the comparative study conducted by Avelino et al. [6], making them suitable references for evaluating the proposed method.

V. RESULTS AND DISCUSSION

In this section, the experimental results obtained with the proposed approach are presented and discussed. The evaluation was conducted from three perspectives: (i) the analysis of the correlation between the Instance Hardness (IH) measure and the traditional relevance function (ϕ), with the aim of investigating the degree of agreement between them in identifying rare instances; (ii) the evaluation of the behavior of the IH-based function in scenarios with bimodal distributions of the target variable, aiming to analyze its ability to capture non-unimodal rarity patterns; and (iii) the application of the IH-based function in balancing strategies, examining its impact on the predictive performance of regression models.

A. Correlation between IH and ϕ

Initially, an analysis was conducted with the objective of verifying the existence of a correlation between the concepts, specifically whether rare instances [16], characterized by high values of ϕ, coincide with instances considered difficult, characterized by high values of the IH-based function. For this purpose, the Pearson correlation coefficient was used. This analysis is fundamental to understanding the degree of relationship between these concepts and, consequently, to justify conducting experiments using the IH-based function for the identification of rare cases.

The results indicate that, for most of the evaluated datasets, a moderate or strong correlation is observed between the analyzed measures. This behavior is illustrated in Figure 2, which presents the correlation between the two concepts for the 29 datasets analyzed. The correlation values were grouped into three distinct categories: green, representing strong and very strong correlations; orange, corresponding to moderate correlations; and purple, encompassing weak and very weak correlations. Overall, most datasets exhibit moderate to strong correlations, reinforcing the compatibility between the concepts and indicating that the *IH-based function* can effectively serve as an alternative to traditional relevance functions.

B. IH-based function for bimodal distributions

Bimodal distributions of the target variable represent a well-known challenge for traditional relevance functions in imbalanced regression. The relevance function proposed by Ribeiro (2011) [7], for example, presents limitations when the distribution has more than one mode, as it tends to assign static relevance values to instances, failing to capture local variations associated with rarity [9].

This behavior is illustrated in Figure 3, which presents the relevance function ϕ applied to the synthetic dataset shown in Figure 1, whose target variable follows a bimodal distribution. It can be observed that the function assigns a relevance value equal to zero to all instances, failing to identify rare regions between the modes of the distribution.

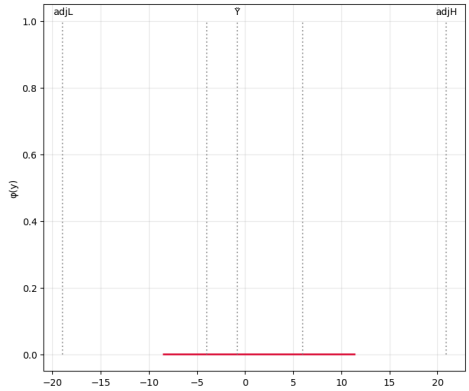

Fig. 3. Relevance function values for the bimodal target. It can be observed that the method produces constant values, failing to identify the rare region between the peaks.

In contrast, Figure 4 highlights the behavior of the IH-based function in this same scenario. The approach is able to correctly identify the less represented regions of the distribution, associated with higher levels of prediction difficulty, characterizing them as rare regions even in the presence of multiple modes.

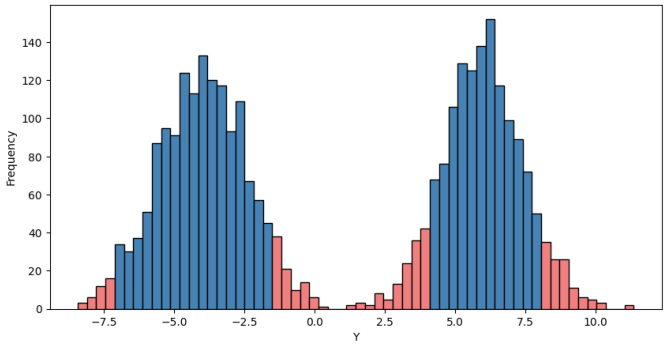

Fig. 4. Detection of rare regions in the bimodal target using the median IH per bin. Bins above the 70th percentile of IH were highlighted in red.

The identification of these regions is based on discretizing the target variable into equidistant bins, followed by calculating the median of the IH values for each bin. Next, a threshold is defined based on the 70% percentile of the global IH distribution, with bins whose median IH belongs to the top 30% of observed values being considered rare or difficult. In this way, regions of the distribution associated with higher prediction difficulty are highlighted as rare.

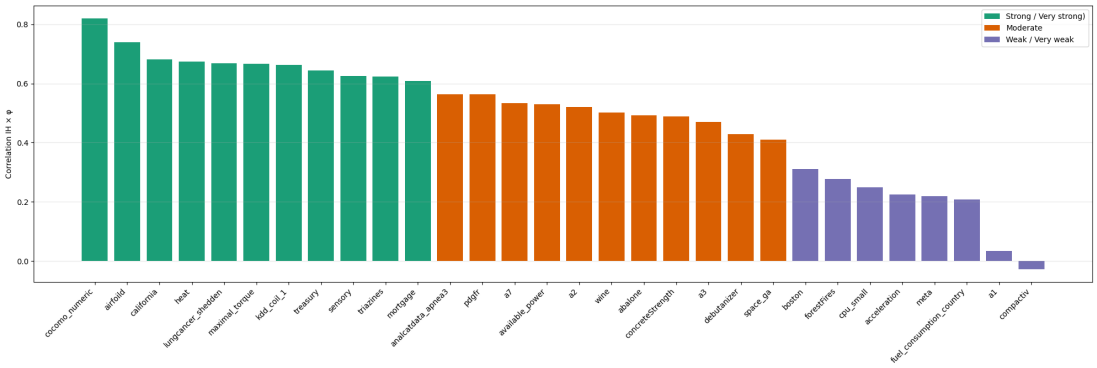

Fig. 2. Distribution of the correlation values between IH and ϕ.

It is important to note that the value of this threshold is a method parameter and is not fixed. Depending on the characteristics of the data distribution, percentiles such as 70%, 75%, or 80% may yield better performance, making it necessary to conduct experimental analyses to identify the most appropriate configuration for each dataset.

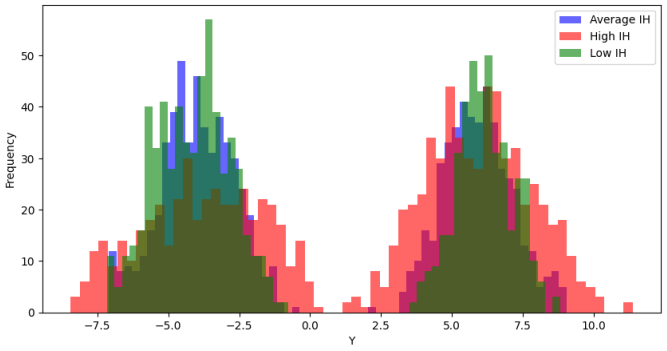

Fig. 5. Distribution of the target variable y segmented by levels of Instance Hardness. The instances were grouped into three categories — Low IH, Medium IH, and High IH — based on quantiles, allowing the visualization of the relationship between prediction difficulty and the concentration of instances across the target space.

Figure 5 illustrates the relationship between the distribution of the target variable and the different levels of Instance Hardness IH. For this analysis, the IH values were partitioned into three quantiles, corresponding to the categories Low IH, Medium IH, and High IH. It can be observed that instances associated with high levels of IH tend to concentrate in regions with lower data density, highlighting the association between prediction difficulty and rarity across the target space.

In a second analysis, the robustness of the proposed method was investigated through controlled variations in the generation of the synthetic dataset, resulting in the creation of six distinct datasets following this procedure. For this purpose, new datasets were generated while keeping the following characteristics fixed: (i) 3,000 instances, (ii) two input variables, and (iii) a bimodal distribution of the target variable. The only factor varied was the level of Gaussian noise added to the data, considered within the range from 0 to 20.

The results indicate that the IH-based function maintains its ability to correctly identify regions associated with rare data at lower noise levels. As the noise increases, it is observed that the distribution of the target variable ceases to exhibit a clearly bimodal shape, gradually assuming a smoother form. Even so, the proposed approach continues to correctly highlight the less represented regions of the distribution, associated with the extremes of the target space. This behavior is illustrated in Figure 6.

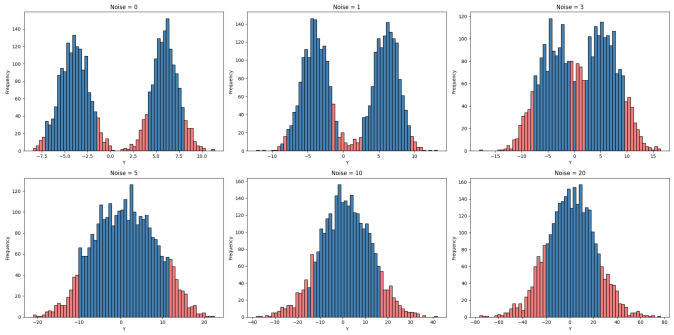

Fig. 6. Variation in the performance of the IH-based function with increasing noise levels. The method remains stable up to moderate noise levels, even in the presence of distribution bimodality loss.

Finally, tests were conducted on the dataset with a bimodal distribution (Figure 4), comparing the results obtained on the original dataset with those obtained after applying the IH-based function as a means of identifying rare data within the RO and GN balancing functions. For this test, traditional relevance functions were not used, as they are unable to define which data are rare in a bimodal distribution. As a result, as shown in Table II, for MAE, the balancing functions based on the IH-based function achieved 3 wins out of the 5 tested models, with 2 of them for RO and 1 for GN. For MSE, these functions reached the same result, with 3 wins in the 5 tests performed.In the table, the best statistical results for each model–metric combination are highlighted in bold.

TABLE II
COMPARATIVE RESULTS OF WINS IN THE BIMODAL DISTRIBUTION BY METHOD FOR MAE AND MSE.

Model	MAE			MSE		
	None	IH-RO	IH-GN	None	IH-RO	IH-GN
BG	0.0593	**0.0592**	0.1132	**0.0124**	0.0126	0.0326
MLP	0.0467	**0.0423**	0.0657	0.0046	**0.0038**	0.0082
RF	**0.0474**	0.0480	0.1015	0.0103	**0.0102**	0.0286
SVR	0.1182	0.1154	**0.0849**	0.1155	0.1095	**0.0508**
XGB	**0.0684**	0.0691	0.1256	**0.0106**	0.0107	0.0326

C. IH-based function for balancing problems

From the perspective of regression balancing analysis, the IH-based function is investigated as an alternative criterion for identifying relevant instances in the resampling process. To evaluate the generalization capability of the proposed methodology, an experimental analysis was conducted on 29 datasets (Table I), comparing MAE and MSE across different training strategies for regression models. Specifically, the following were considered: (i) the use of the original dataset, without applying balancing; (ii) balancing guided by the traditional relevance function (ϕ) [4]; (iii) balancing based on the *Distance-Based Relevance Function* (DRF) [10], which takes into account the rarity of observations based on the distance between target values; and (iv) balancing guided by the IH-based function, proposed in this work, for which, in this test, an instance was considered rare if it belonged to the top 30% of IH [15] values in the complete dataset.

Table III presents an overview of the comparative results in terms of the number of wins achieved by each method for both MAE and MSE metrics, aggregated across learning models and datasets. This table provides an initial descriptive analysis of the performance tendencies before considering statistical significance.

In this comparison, five distinct models were evaluated on 29 datasets, totaling 145 model–dataset pairs. As shown in Table III, Considering the MAE metric, the balancing functions that used the IH-based function achieved superior performance in 68 pairs, with 48 associated with RO using the IH-based function (*IH-RO*) and 20 with GN (*IH-GN*). These results outperformed the other approaches in approximately 50% of the evaluated pairs. For the MSE metric, the number of pairs with better performance was 58, with 38 related to *IH-RO* and 20 to *IH-GN*, corresponding to about 40% of the tested pairs.

Table IV summarizes the pairwise Wilcoxon signed-rank results, showing a consistent advantage of the IH-based strategy across both RO and GN paradigms. For both MAE and MSE metrics, IH-based variants achieve a substantially larger number of wins over their respective baselines, with most of these improvements being statistically significant. When compared to the DRF variants, IH-based methods also present a higher number of wins, and a considerable portion of these differences are statistically significant. Importantly, when IH-based methods incur losses, these are predominantly non-significant, indicating that the proposed strategy rarely leads to statistically meaningful performance degradation. Overall, the results confirm that the IH-based relevance function provides robust and statistically reliable improvements for both RO and GN.

TABLE IV
PAIRWISE WILCOXON SIGNED-RANK ANALYSIS OF IH-BASED METHODS AGAINST BASELINE AND DENSITY-BASED VARIANTS. WINS AND LOSSES ARE REPORTED, WITH STATISTICALLY SIGNIFICANT RESULTS ($p < 0.05$) SHOWN IN PARENTHESES FOR MAE AND MSE.

Comparison	MAE		MSE	
	Win (sig.)	Loss (sig.)	Win (sig.)	Loss (sig.)
IH-RO vs RO	**27 (22)**	2 (0)	**20 (13)**	9 (1)
IH-RO vs DRF-RO	**17 (11)**	12 (0)	**13 (9)**	16 (1)
IH-GN vs GN	**24 (19)**	5 (4)	**20 (19)**	9 (5)
IH-GN vs DRF-GN	**23 (15)**	6 (3)	**16 (8)**	13 (8)

Finally, it is noteworthy that the balancing methods based on the IH-based function achieved superior performance for both RO and GN, across both metrics (*MAE* and *MSE*), when compared to the original dataset, traditional RO and GN methods using the relevance function, and those based on the Distance-Based Relevance Function (DRF). These results highlight the robustness and effectiveness of the IH-based function in identifying rare data for balancing purposes.

VI. CONCLUSION

In conclusion, the results demonstrate that the use of IH-based function as a means of identifying rare instances is an effective alternative to address some limitations of the traditional approach based on the relevance function, especially in bimodal distributions, where this function returns a static value for all instances in the dataset. The IH-based function proved capable of identifying these rare instances in datasets with bimodal distributions. In addition, it was demonstrated that the use of IH in balancing strategies yields superior overall performance compared to the original dataset and traditional balancing approaches.

Finally, it is also necessary to identify the limitations of the proposed approach. Since it is an alternative based on a pool of machine learning models, the computational cost may be high for datasets with a large number of records. Moreover, conducting tests to determine the optimal threshold for each dataset is essential for achieving better performance of the technique in question, as adjusting its value in the balancing functions may lead to adding or removing more or fewer data points, which can significantly impact model performance.

REFERENCES

[1] H. Guo, Y. Li, J. Shang, M. Gu, Y. Huang, and B. Gong, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017.

[2] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016.

[3] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[4] P. Branco, R. P. Ribeiro, and L. Torgo, "Ubl: an r package for utility-based learning," *arXiv preprint arXiv:1604.08079*, 2016.

[5] D. Kowatsch, N. M. Müller, K. Tscharke, P. Sperl, and K. Böttinger, "Imbalance in regression datasets," Fraunhofer AISEC, Garching near Munich, Germany, Tech. Rep.

TABLE III
COMPARATIVE RESULTS OF WINS BY METHOD FOR MAE AND MSE

Model	MAE							MSE						
	None	**RO**	**IH-RO**	**DRF-RO**	**GN**	**IH-GN**	**DRF-GN**	**None**	**RO**	**IH-RO**	**DRF-RO**	**GN**	**IH-GN**	**DRF-GN**
BG	5	5	**9**	3	0	6	1	**8**	4	6	5	0	5	1
MLP	4	6	**10**	4	1	4	0	7	5	**8**	4	1	4	0
RF	6	6	**8**	2	0	6	1	6	5	**9**	4	0	5	0
SVR	**11**	1	**11**	4	0	1	1	7	1	5	**9**	2	3	2
XGB	6	6	**10**	4	0	3	0	6	6	**10**	4	0	3	0
Total	32	24	**48**	17	1	20	3	34	21	**38**	26	3	20	3

[6] J. G. Avelino, G. D. C. Cavalcanti, and R. M. O. Cruz, "Resampling strategies for imbalanced regression: a survey and empirical analysis," *Artificial Intelligence Review*, vol. 57, no. 82, 2024.

[7] R. Ribeiro, "Utility-based regression," *Ph. D. dissertation*, 2011.

[8] P. Branco, L. Torgo, and R. P. Ribeiro, "Pre-processing approaches for imbalanced distributions in regression," *Neurocomputing*, vol. 343, pp. 76–99, 2019.

[9] S. Stocksieker and D. Pommeret, "A comprehensive survey on imbalanced regression: Definitions, solutions, and future directions," 2025, hAL open archive, hal-05213741.

[10] D. D. In and H. Kim, "Distance-based relevance function for imbalanced regression," *Stats*, vol. 8, no. 3, p. 53, 2025.

[11] P. Y. A. Paiva, C. C. Moreno, K. Smith-Miles, M. G. Valeriano, and A. C. Lorena, "Relating instance hardness to classification performance in a dataset: a visual approach," *Machine Learning*, vol. 111, no. 8, pp. 3085–3123, 2022.

[12] P. O. Branco, L. Torgo, and R. P. Ribeiro, "Smogn: a pre-processing approach for imbalanced regression," in *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, vol. 74, 2017, pp. 36–50.

[13] L. Camacho, G. Douzas, and F. Bacao, "Geometric smote for regression," *Expert Systems with Applications*, p. 116387, 2022.

[14] R. P. Ribeiro and N. Moniz, "Imbalanced regression and extreme value prediction," *Machine Learning*, vol. 109, pp. 1803–1835, 2020.

[15] G. P. Torquette, V. S. Nunes, P. Y. A. Paiva, and A. C. Lorena, "Instance hardness measures for classification and regression problems," Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, SP, Brazil, Tech. Rep., 2024, published: 27 February 2024.

[16] L. Torgo, R. P. Ribeiro, and B. Pfahringer, "Utility-based regression," in *Lecture Notes in Computer Science*. Springer, 2009, vol. 5812.

[17] J. A. Sáez, B. Krawczyk, and M. Woźniak, "Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets," *Pattern Recognition*, vol. 57, pp. 164–178, 2016.

[18] G. P. Torquette, V. S. Nunes, P. Y. A. Paiva, L. B. Neto, and A. C. Lorena, "Characterizing instance hardness in classification and regression problems," in *Proceedings of KDMile 2022*, 2022, arXiv:2212.01897.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] L. Torgo and R. Ribeiro, "Precision and recall for regression," in *International Conference on Discovery Science*. Springer, 2009, pp. 332–346.