Measuring the Class-imbalance Extent of Multi-class Problems

Jonathan Ortigosa-Hernández, Iñaki Inza, and Jose A. Lozano

Abstract

Since many important real-world classification problems involve learning from unbalanced data, the challenging class-imbalance problem has lately received considerable attention in the community. Most of the methodological contributions proposed in the literature carry out a set of experiments over a battery of specific datasets. In these cases, in order to be able to draw meaningful conclusions from the experiments, authors often measure the class-imbalance extent of each tested dataset using imbalance-ratio, i.e. dividing the frequencies of the majority class by the minority class.

In this paper, we argue that, although imbalance-ratio is an informative measure for binary problems, it is not adequate for the multi-class scenario due to the fact that, in that scenario, it groups problems with disparate class-imbalance extents under the same numerical value. Thus, in order to overcome this drawback, in this paper, we propose *imbalance-degree* as a novel and normalised measure which is capable of properly measuring the class-imbalance extent of a multi-class problem. Experimental results show that imbalance-degree is more adequate than imbalance-ratio since it is more sensitive in reflecting the hindrance produced by skewed multi-class distributions to the learning processes.

1 Introduction

Most of the well-known traditional machine learning techniques are designed to solve classification problems showing reasonably balanced class distributions [24]. However, this assumption does not always hold in reality. Occasionally, real-world problems have skewed class distributions and, due to this, they present training datasets where several classes are represented by an extremely large number of examples, while some others are represented by only a few. This particular situation is known as the class-imbalance problem, a.k.a. learning from unbalanced data [17], and it is considered in the literature as a major obstacle to building precise classifiers: the solutions obtained for problems showing class-imbalance through the traditional learning techniques are usually biased towards the most probable classes showing a poor prediction power for the least probable classes [10]. Thus, in an attempt to overcome this obstacle, hundreds of methodological solutions have been proposed recently in order to balance the prediction powers for both the most and the least probable classes.

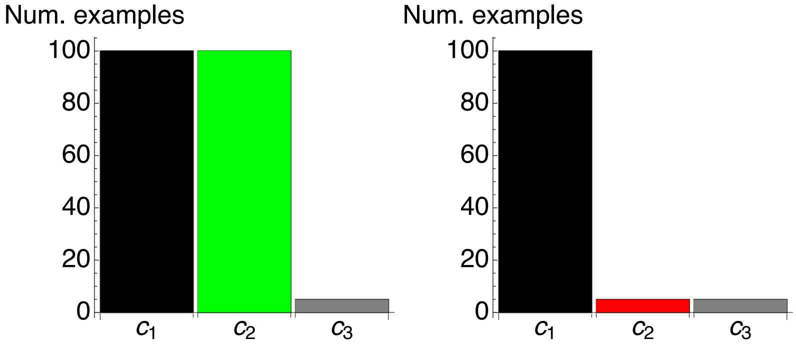
According to [28], the proposed solutions can be mainly categorised into the following three major groups: (i) the development of *inbuilt mechanisms* [11], which change the classification strategies to impose a bias toward the minority classes, (ii) the usage of *data sampling methods* [3], which modify the class distribution to change the balance between the classes, and (iii) the adoption of *cost-sensitive learning techniques* [22] which assume higher misclassification costs for examples of the minority classes.

Usually, every paper proposed within those categories shares the same experimental setup: the proposed method is compared against one or several competing methods over a dozen or so datasets. However, although this experimental setup is reasonable enough to support an argument that the new method is “as good as” or “better than” the state-of-the-art, it still leaves many unanswered questions [27]. Besides, it is costly in computing time [30]. Thus, in order to be able to perform more meaningful analyses, some authors complement this experimental schema with a study of the inherent properties of the checked datasets by extracting from them a set of informative measures [30, 31]. By means of this data characterisation, more solid empirical conclusions may be efficiently extracted: on the one hand, a better understanding of the problem faced may be achieved since it is a structured manner of investigating and explaining which intrinsic features of the data are affecting the classifiers [2]. On the other hand, the measured data can be related to the classifier performance so that the applicability and performance of a classifier based upon the data can be predicted, avoiding a great amount of computing time [30].

In the literature, authors often measure the class-imbalance extent. In those works, *imbalance-ratio* is the most frequently used summary of the class-imbalance extent due to its simplicity [11]. It reflects the (expected) number of instances of the most probable class for each instance of the least probable class. However, in this paper, we state that whilst it is a very informative summary of the class-imbalance extent for binary problems, it is not capable of completely and honestly describing the disparity among the frequencies of more than two classes. In the multi-class scenario, there exists other classes rather than the most and least probable classes and they are not taken into account for the calculation of this summary. This may lead to the undesired situation of characterising multi-class problems with disparate class-imbalance extents using the same imbalance-ratio.

In order to clarify this drawback, let’s consider the toy example presented in Figure 1; Imagine that a 3-class problem with an imbalance-ratio of 20 $(100 : 5)$ is provided. This means that there are 20 examples of the most probable class (c_1) for each example of the least probable class (c_3). However, by means of just imbalance-ratio, little knowledge can be extracted regarding the remaining class c_2, i.e. the number of examples of c_2 can vary from 5 to 100, and all these 95 different possible scenarios share an imbalance-ratio equal to 20.

As can be easily noticed, the scenario with 100 examples for the second class – Figure 1a –, is far less problematic than having only 5 examples of the second class – Figure 1b –. While there is only one minority class in the former scenario, we find two minority classes in the latter. So, it can be straightforwardly concluded that imbalance-ratio is not



(a) Best unbalanced scenario.

(b) Worst unbalanced scenario.

Figure 1: Extreme cases of an unbalanced ternary toy example showing an imbalance-ratio of 20.

a proper summary of the class-imbalance extent in the multi-class scenario as it groups diverse problems with different class-imbalance extents under the same numerical value.

Thus, in order to bridge this gap, in this paper, we propose a new summary which is capable of properly shortening the class distributions of both binary and multi-class classification problems into a single value. This measure, which we name *imbalance-degree*, represents the existing difference between a purely balanced distribution and the studied unbalanced problem, and it has the following three interesting properties:

1. By means of a single real value in the range $[0, K)$, where K is the number of classes, it not only summarises the class distribution of a given problem but also inherently expresses the number of majority and minority classes.

2. Depending on the requirements of the experimental setup and the degree of sensitivity sought, this measure can be instantiated with any common distance between vectors or divergence between probability distributions.

3. A unique mapping between the class distributions and the numerical value of imbalance-degree is ensured for problems showing different numbers of majority and minority classes. Therefore, diverse problems cannot share a common numerical value as happens with imbalance-ratio.

Experimental results show that imbalance-degree is a more appropriate summary than imbalance-ratio. In the multi-class framework, the former is not only able of differentiating class distributions than the latter groups with the same value but it also achieves a greater correlation with the hindrance that skewed class distributions cause in the learning processes.

The rest of the paper is organised as follows: Section 2 introduces the framework, notation, and a review of the most-commonly used measures and summaries of the class distribution. In Section 3, we introduce imbalance-degree as a more informative measure for the multi-class scenario. After that, Section 4 presents an empirical study of the adequateness of the proposed measure. Finally, Section 5 sums up the paper.

2 Problem Formulation and State-of-the-art Measures for the Class-imbalance Extent

Let γ_K be a K-class classification problem with a generative model given by the generalised joint probability density function

$$\rho(\mathbf{x}, c) = p(c)\rho(\mathbf{x}|c), \tag{1}$$

where $p(c)$ is a multinomial distribution representing the class probabilities and $\rho(\mathbf{x}|c)$ is the conditional distribution of the feature space. For convenience, henceforth, we rewrite the former as $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_K)$, where each $\eta_i = p(c_i)$ stands for the probability of each categorical class c_i. Also, we denote the special case of equiprobability as $\mathbf{e} = (e_1, e_2, \ldots, e_K)$, where $\forall i, \eta_i = 1/K = e_i$. Then, depending on the outline of its class distribution $\boldsymbol{\eta}$, every classification problem γ_K can be catalogued into one of the following groups: (i) γ_K may be a balanced problem, (ii) an unbalanced problem showing multi-majority, or (iii) a multi-minority unbalanced problem. The formal definitions for these groups, as expressed in [17] and [31], are the following:

Definition 1. *A K-class classification problem, γ_K, is balanced if it exhibits a uniform distribution between its classes. Otherwise, it is considered to be unbalanced. Formally,*

$$\gamma_K \textit{ is balanced } \iff \boldsymbol{\eta} = \mathbf{e}. \tag{2}$$

Definition 2. *A multi-class classification problem ($K > 2$), γ_K, shows a multi-majority class-imbalance if most of the classes have a higher or equal probability than equiprobability, i.e.*

$$\gamma_K \textit{ is multi-majority } \iff \sum_{i=1}^{K} \mathbb{1}\left(\eta_i \geq \frac{1}{K}\right) \geq \frac{K}{2}. \tag{3}$$

Definition 3. *An unbalanced classification problem, γ_K with $K > 2$, shows a multi-minority class-imbalance when most of the class probabilities are below the equiprobability. Formally,*

$$\gamma_K \textit{ is multi-minority } \iff \sum_{i=1}^{K} \mathbb{1}\left(\eta_i < \frac{1}{K}\right) > \frac{K}{2}. \tag{4}$$

Here, $\mathbb{1}(\mathcal{E})$ is the indicator function, 1 if the event $\mathcal{E}$ is true, 0 otherwise. Note that Figure 1a and Figure 1b correspond to multi-majority and multi-minority problems respectively, and that only when facing multi-class problems do Definition 2 and 3 make sense.

Unfortunately, in most of the real-world cases, the generative model, along with the real class distribution, is unknown. Thus, authors must estimate $\boldsymbol{\eta}$ from a training dataset D in order to not only classify γ_K into one of the groups proposed in the definitions, but also to be capable of using a close approximation of the real class distribution to properly validate the conclusions exposed in their experimental schemas.

Then, let $D = \{(\mathbf{x}^{(1)}, c^{(1)}), \ldots, (\mathbf{x}^{(l)}, c^{(l)})\} = \{(\mathbf{x}^{(n)}, c^{(n)})\}_{n=1}^{l}$ be defined as a supervised training dataset of size l drawn from the generative function[1]. There, let the class labels $\{c^{(n)}\}_{n=1}^{l}$ be i.i.d. random values drawn from $\boldsymbol{\eta}$ and let each observation $\{\mathbf{x}^{(n)}\}_{n=1}^{l} \in D$ be also an i.i.d. random value but drawn from $\rho(\mathbf{x}|c_i)$. In order to estimate the class distribution $\boldsymbol{\eta}$, we define the empirical distribution $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_K)$. $\boldsymbol{\zeta}$ is a multinomial distribution with K categories, which exhibits the information available in the dataset about the class distribution of the problem γ_K. There, each statistic ζ_i estimates each class probability η_i by just determining the frequency of the class c_i in the dataset. Formally, the statistic is defined as follows:

$$\zeta_i = \frac{1}{l} \sum_{n=1}^{l} \mathbb{1}(c^{(n)} = c_i). \tag{5}$$

Unless otherwise stated, henceforth, we only use the estimator $\boldsymbol{\zeta}$ of the class distribution since having an unknown generative model is the most common scenario. Anyhow, in the event of knowing the generative model, all the methodologies presented can be directly used with $\boldsymbol{\eta}$ by just substituting the empirical class distribution by the real class distribution in the formulae.

A few measures for the class-imbalance extent of the class distribution using the empirical class distribution $\boldsymbol{\zeta}$ have been already utilised in the experimental setups of the state-of-the-art literature: the most simple manner to measure the class-imbalance extent of a given problem is just to write down the **empirical class distribution** [3], $\boldsymbol{\zeta}$, or to directly transcribe **the occurrences of all the classes** [13, 31] in the dataset, i.e. $\mathbf{l} = (l_1, l_2, \ldots, l_K)$ s.t. $\forall i, l_i = l\zeta_i$.

These descriptions seem to be a good choice due to the fact that they contain all the information available in the dataset with regards to the class-imbalance extent of the generative class distribution $\boldsymbol{\eta}$. However, analysing them can be quite tedious in problems with a large number of class values, especially in highly multi-class problems ($K \geq 1,000$, [15]). In those cases, it is very common to find unbalanced distributions among the classes. Additionally, these solutions are also more difficult to read and/or compare than single value summaries. Therefore, functions $d(\cdot)$ which assign different single real numbers to disparate values of $\boldsymbol{\zeta}$, i.e. $d : \boldsymbol{\zeta} \mapsto \mathbb{R}$ and which are somehow correlated with the hindrance that skewed class distributions cause on learning algorithms mainly dominate the class-imbalance literature [11, 27]. Regarding the summaries, **imbalance-ratio** (IR) between the majority and minority classes is, to the best of our knowledge, the only summary for $\boldsymbol{\zeta}$ used for multi-class problems. It is calculated by dividing the maximum statistic ζ_i by the minimum. Formally,

$$IR(\boldsymbol{\zeta}) = \frac{\max_i \zeta_i}{\min_j \zeta_j}. \tag{6}$$

[1] Note that we assume that D is i.i.d. from eq. (1). Therefore, in this work, we only focus on the case that the nature of the class-imbalance is in the probability distribution, not on the case of having a biased training dataset.

Table 1: Summary of measures for the class-imbalance extent of the class distribution $\boldsymbol{\eta}$ used in the literature and our proposal.

Measure	Formula	Strength	Weakness	Ref
Empirical distribution	$\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_K)$	It is the most informative measure.	Difficult to read and/or compare in highly multi-class problems.	[3]
Frequency of the classes	$\mathbf{l} = (l_1, l_2, \ldots, l_K)$ s.t. $\forall i, l_i = l\zeta_i$	Very informative (equivalent to the empirical class distribution).	Difficult to read and/or compare in highly multi-class problems.	[31]
Imbalance-ratio	$IR(\boldsymbol{\zeta}) = \max_i \zeta_i / \min_j \zeta_j$	It is a single value and easily readable *summary.*	Inappropriate summary for multi-class problems since the injection is lost.	[27]
Imbalance-degree	$ID(\boldsymbol{\zeta}) = d_\Delta(\boldsymbol{\zeta}, \mathbf{e})/d_\Delta(\boldsymbol{\iota}_m, \mathbf{e}) + (m-1)$	It is a single easily readable *summary* appropriate for binary and multi-class problems.	A total injection can only be achieved by the proper choice of the metric/divergence Δ.	This paper

It is trivial to prove that $IR : \boldsymbol{\zeta} \mapsto \mathbb{R}$ is an injective function for binary problems. This property makes this summary appropriate for such scenarios due to the fact that all possible unbalanced scenarios yield to different IR values and that any $\boldsymbol{\zeta}$ can be easily recovered from the $IR(\boldsymbol{\zeta})$. However, when the number of classes outnumbers 2, the injection is lost (as previously shown in the toy example of Figure 1, where multi-majority and multi-minority problems share the same numerical value). This is an inappropriate characteristic for a summary of the class-imbalance extent since previous papers have shown that multi-minority problems are harder than multi-majority [31]. This may imply that IR is not correlated with the hindrance produced by skewed multi-class distributions.

Therefore, it can be concluded that neither of the presented measures (summarised in Table 1) for the class-imbalance extent are appropriate for multi-class unbalanced problems.

3 Imbalance-degree

In this section, our aim is to propose a new and more suitable summary for any empirical class distribution $\boldsymbol{\zeta}$ with $K \geq 2$ which, at least, fulfils the following properties: (i) it must be an easily readable finite single valued summary of a multinomial distribution and (ii) it needs to be correlated with the hindrance that highly unbalanced datasets cause in the learning processes.

Thus, since the class distribution does harm the learning processes as it extremely diverges from the balanced one [27], it is immediate to use a distance/similarity function, $d_\Delta(\boldsymbol{\zeta}, \mathbf{e})$, between both the empirical and balanced distributions, $\boldsymbol{\zeta}$ and $\mathbf{e}$, to summarise the degree of skewness of a classification problem γ_K. Here, Δ stands for any chosen distance between vectors or divergence between probability distributions which can be found in the literature.

However, just relying on the direct usage of a distance/similarity function has, for our purpose, two undesirable properties which may clash with our aim of having an informative easily readable or comparable summary function:

1. Similar to IR, *different values for different number of majority/minority classes cannot be assured.* For instance, imagine we use the Kullback-Leibler divergence [19] as a summary of two diverse class distributions $\boldsymbol{\zeta}^{(1)} = (0.027009, 0.486495, 0.486495)$ and $\boldsymbol{\zeta}^2 = (0.712853, 0.143573, 0.143573)$. There, both calculi reach the same value: $d_{KL}(\boldsymbol{\zeta}^{(1)}, \mathbf{e}) = d_{KL}(\boldsymbol{\zeta}^{(2)}, \mathbf{e}) = 0.273$.

2. Although a measure is always a finite positive value, *it is not necessarily upper bounded.* For example, Kullback-Leibler divergence may be unbounded, and Manhattan and Euclidean distances [9], in this context, are upper bounded by the values 2 and 1, respectively.

In order to overcome these drawbacks, we purposely divide the space of class distributions so that we can operate on the distance/similarity function and obtain an adequate summary: let $\mathcal{Z}^K$ be defined as the set containing all the possible empirical distributions $\boldsymbol{\zeta}$ of a K-class problem and let $\mathcal{Z}_m^K \subset \mathcal{Z}^K$, $m \in \{0, 1, \ldots, K-1\}$ be a subset containing all the empirical class distributions containing exactly m minority classes. Formally,

$$\mathcal{Z}_m^K \triangleq \left\{ \boldsymbol{\zeta} \in \mathcal{Z}^K : m = \sum_{i=1}^{K} \mathbb{1} \left(\zeta_i < \frac{1}{K} \right) \right\}. \tag{7}$$

Straightaway, this severance of $\mathcal{Z}^K$ into K different subsets $\mathcal{Z}_m^K$ allows us to tackle both problems:

1. On the one hand, *different values for different numbers of minority/majority classes can be directly provided* in the summary function by just forcing different ranges of values to different subsets. Here, the range $(m-1, m]$ is assigned to each subset $\mathcal{Z}_m^K$ in the summary (0 for $\mathcal{Z}_0^K$).

2. On the other hand, a *common upper bound* for each subset, and consequently to the summary, can also be assured by applying a 0-1 normalisation to the distance of the empirical class distribution (a range of size 1 has been assigned to each subset). This is achieved through the division of $d_\Delta(\boldsymbol{\zeta}, \mathbf{e})$ by $d_\Delta(\boldsymbol{\iota}_m, \mathbf{e})$, being $\boldsymbol{\iota}_m$ the distribution in $\mathcal{Z}_m^K$ most distant to $\mathbf{e}$.

Then, through the application of these amendments on the distance/similarity function, we define our main proposal as:

Definition 4. *The **imbalance-degree** (ID) of a multi-class dataset showing an empirical class distribution $\boldsymbol{\zeta}$ is given by*

$$ID(\boldsymbol{\zeta}) = \frac{d_\Delta(\boldsymbol{\zeta}, \mathbf{e})}{d_\Delta(\boldsymbol{\iota}_m, \mathbf{e})} + (m-1), \tag{8}$$

where m is the number of minority classes, d_Δ is the chosen distance/similarity function to instantiate ID, and $\boldsymbol{\iota}_m$ is the distribution showing exactly m minority classes with the highest distance to $\mathbf{e}$ ($\arg\max_{\boldsymbol{\zeta}\in\mathcal{Z}_m^K} d_\Delta(\boldsymbol{\zeta}, \mathbf{e})$).

In eq. (8), the term $m-1$ is intentionally added to the normalisation term to ensure different values for different values of m, i.e. $ID(\boldsymbol{\zeta}) \in (m-1, m]$ when $\boldsymbol{\zeta} \in \mathcal{Z}_m^K$. Moreover, in the purely balanced scenario $\boldsymbol{\zeta} = \mathbf{e}$, our proposal $ID(\mathbf{e}) = 0$ due to the fact that, conventionally, $d_\Delta(\mathbf{e}, \mathbf{e})/d_\Delta(\mathbf{e}, \mathbf{e}) = 1$.

4 Empirical Study

In order to determine the appropriateness of ID (over IR) as a summary of the class-imbalance extent in the multi-class framework, we define two different sets of experiments to empirically corroborate the following hypotheses:

- H_1: *While IR has a deficient resolution to summarise the class-imbalance extent in the multi-class scenario, ID offers a wide variety of high resolution summaries.*
- H_2: *When used on real-world multi-class classification problems, ID is more sensitive to the class-imbalance extent than IR.* I.e. ID is more accurate than IR in informing about a poor performance of traditional learning systems.

Since ID can be instantiated with any chosen distance/similarity function, we first introduce the measures used in the experiments: from the metrics in the vector space [9], Manhattan[2], Euclidean and Chebyshev distances are chosen. Together, the f-divergences [7], the most utilised measures for probability distributions, are also included. Within the latter group, we introduce Kullback-Leibler divergence [20], Hellinger [18] (closely related to, although different from, the Bhattacharyya distance [4]) and total variation distances, and χ^2-divergence [26]. These measures are mathematically defined in Table 2.

Additionally, in order to use eq. (8), the furthest distribution $\boldsymbol{\iota}_m = (\iota_1, \iota_2, \dots, \iota_K)$ to $\mathbf{e}$ must be calculated for every instantiation and every subset $\mathcal{Z}_m^K$. Opportunely, this class distribution coincides for all the considered measures and for all values of m. It satisfies that

$$\sum_{i=1}^{K} \mathbb{1}\left(\iota_i = 0\right) = m \wedge \sum_{i=1}^{K} \mathbb{1}\left(\iota_i = \frac{1}{K}\right) = K - m - 1, \tag{9}$$

i.e. the furthest distribution is composed of (i) m minority classes with zero probability, (ii) $K-m-1$ (all but one) majority classes with probability $1/K$, and (iii) a majority class with the remaining probability $1-(K-m-1)/K$. This distribution always shows the lowest entropy [29] in the subset $\mathcal{Z}_m^K$, whilst the balanced setting $\mathbf{e}$ corresponds to the distribution with the highest entropy in $\mathcal{Z}$. Note that, by symmetry, there may be

[2] Manhattan distance has been left out of the experimentation due to the fact that, for our purposes, it is equivalent to total variation distance for any $K \geq 2$.

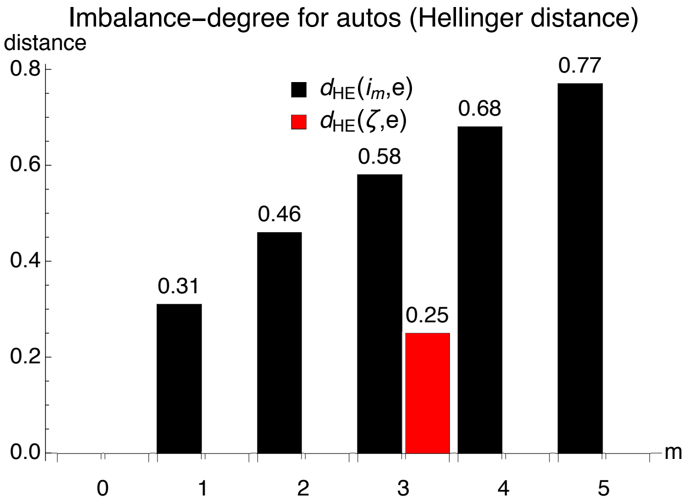


Figure 2: Calculating ID using the Hellinger distance [18] for the dataset `autos` ($K = 6$, $IR = 16$, $ID_{HE} = 2.44$).

up to $K!$ different furthest distributions $\boldsymbol{\iota}_m$. Fortunately, $ID(\boldsymbol{\zeta})$ is not affected by an arbitrary choice of $\boldsymbol{\iota}$ since the entropy, $H(\boldsymbol{\iota}_m)$, and distance values, $d_\Delta(\boldsymbol{\iota}_m, \mathbf{e})$, remain equal for all furthest distributions.

In order illustrate calculation of ID using the distance/similarity functions considered, Figure 2 shows, in a bar chart, an example to instantiate ID using the Hellinger distance, $d_{HE}(\boldsymbol{\zeta}, \mathbf{e})$, on the UCI dataset called `autos` [21]. The numbers above the black bars represent the value of each normaliser $d_\Delta(\boldsymbol{\iota}_m, \mathbf{e})$ for all possible scenarios of m of minority classes in a 6-class problem. Since `autos` has 3 (out of 6) minority classes and $d_{HE}(\boldsymbol{\zeta}, \mathbf{e}) = 0.25$, the problem has a (normalised) ID of 2.44 ($0.25/0.58 = 0.44$ plus $3 - 1$).

4.1 Study 1: Resolution and Diverseness of Imbalance-degree

The resolution of a measure is the smallest change which can be quantified. As previously put forward, IR cannot be considered as a measure which has a satisfactory resolution for multi-class problems; it only changes based on either the most or the least probable classes. In the toy example, for instance, it groups 95 different class distributions using the value $IR = 20$.

Thus, in order to corroborate the first hypothesis, those 95 scenarios are used to not only show that ID is capable of assigning diverse and reasonable values to them, but also to study the behaviours of the different instantiations of ID. Consequently, Figure 3 plots the values of ID for the indicated 95 different frequency scenarios, i.e. $\mathbf{l} = (100, l_2, 5)$, where $l_2 = \{5, \ldots, 100\}$, from the toy problem. The abscissa shows the number l_2 of instances of the second class and the ordinate shows the value of ID. From Table 2, Euclidean distance (ID$_{EU}$), Kullback-Leibler divergence (ID$_{KL}$), Hellinger distance (ID$_{HE}$), total variation distance (ID$_{TV}$) and chi-square divergence (ID$_{CS}$) are plotted.

Table 2: Mathematical formulae for the distance/similarity functions used to instantiate ID in the empirical studies.

Distance/Similarity Function	Δ	$d_\Delta(\boldsymbol{\zeta}, \mathbf{e})$
Metrics in the vector space		
Euclidean distance	EU	$\sqrt{\sum_{i=1}^{K}(\zeta_i - e_i)^2}$
Chebyshev distance	CH	$\max_i \lvert\zeta_i - e_i\rvert$
f-divergences		
Kullback-Leibler divergence	KL	$\sum_{i=1}^{K} \zeta_i \log \frac{\zeta_i}{e_i}$
Hellinger distance	HE	$\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{K}(\sqrt{\zeta_i} - \sqrt{e_i})^2}$
Total variation distance	TV	$\frac{1}{2}\sum_{i=1}^{K} \lvert\zeta_i - e_i\rvert$
Chi-square divergence	CS	$\sum_{i=1}^{K} \frac{(\zeta_i - e_i)^2}{e_i}$

Note that Chevbysev distance is not included[3].

Results show that ID is capable of differentiating each and every different scenario that IR groups with the value 20. Moreover, it can be seen that ID instantiations behave differently as result of the diversity of their distance/similarity functions: whilst all the instantiations share a similar monotonically decreasing shape up to the limiting point where the number the m changes from 2 to 1 ($l_2 = 53$), above that limit, two different groups of instantiations can be perceived. On the one hand, ID$_{EU}$, ID$_{KL}$ and ID$_{CS}$ show a convex shape since they descent down to a minimum and then slightly increase. On the other hand, ID$_{HE}$ and ID$_{TV}$ show a quasi-linear behaviour which starts increasing soon after reaching the limiting point. Thus, it can be straightforwardly concluded that there might be instantiations of ID which are more adequate to summarise the class-imbalance extent than others. Seemingly, the latter group of instantiations (ID$_{HE}$ and ID$_{TV}$) are more appropriate as they reflect the increased intricacy of the classification problem above the limiting point. When $l_2 > 53$, the probability of the minority class c_3 distance itself from the equiprobability causing an increase in the intricacy of the classification problem. In Section 4.2, we also deal with this issue by empirically determining which ID instantiations are more adequate summaries in real-world multi-class datasets. Finally, we believe that, in practise, the above mentioned diversity may also be potentially exploited to adapt ID to different requirements and constraints resulting from real-world unbalanced problems.

[3] It holds that instantiations of ID using Chevbysev and total variation distances are equivalent for the case $K = 3$.

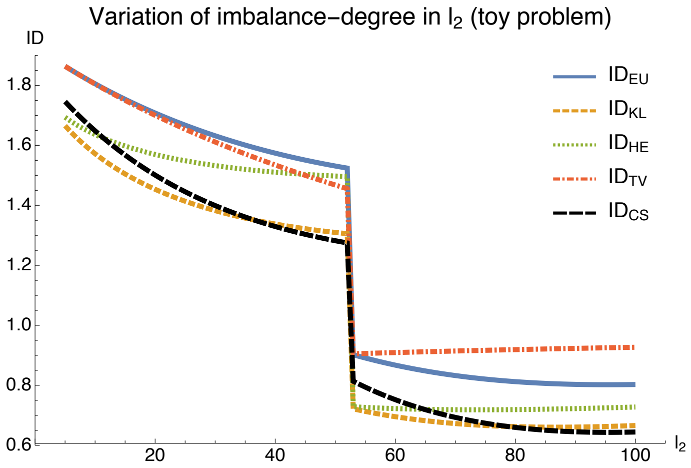


Figure 3: The variation of $ID_{\triangle}$ in all the 95 possible scenarios ($l_2 = \{5, \ldots, 100\}$) of the toy problem of Figure 1. For these scenarios, $IR = 20$.

4.2 Study 2: Sensitivity and Validity of Imbalance-degree

A measure is sensitive to recognise a given set of events if it is capable of valuing them differently. Specifically, we can consider a summary of the class-imbalance extent to be sensitive to recognise the hindrance that highly unbalanced data produce in the traditional learning systems if it is correlated with the performance of those learning systems. Thus, to determine which instantiation of ID is more sensitive to the exposed hindrance than IR, in this section, the following experiment is carried out:

A database containing the 15 unbalanced multi-class datasets recommended in the key work of [1] is assembled and the value of each summary presented in this paper (see Table 2) is calculated for each dataset. Their values, along with some main characteristics of the datasets, are presented in Table 3. There, each row corresponds to a dataset and each column stands for a characteristic (name, features and number of classes) or a summary (empirical class distribution, number of occurrences, IR and IDs). Afterwards, each dataset is used to feed a representative learning algorithm from the traditional major learning paradigms [27]. Specifically, for each problem, a different classifier is learnt using 5 different popular supervised algorithms[4]: C4.5 (Decision trees), RIPPER (Decision rules), Neural Networks (Connectionism), Naïve Bayes (Probabilistic), and SVM (Statistical learning). In order to assess the performance of each learnt classifier, three different performance scores which are highly recommended for multi-class unbalanced problems are used [24]: the arithmetic mean among the recall of the classes ($\mathcal{A}$), the geometric mean among the recalls ($\mathcal{G}$), and the minimum recall obtained (min). In order to obtain the values of these performance scores for each dataset, we estimate them using 10×10 fold cross-validation[5].

[4] In this experimentation, all learning and error estimation tasks have been performed using the software `Weka 3` [16].

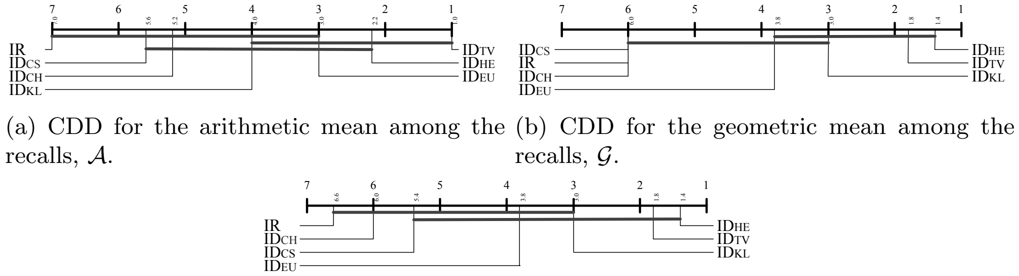
[5] These results can be downloaded, along with the source code.

Table 3: Characteristics of the studied unbalanced datasets [1] and the value of the summaries introduced in this paper for each dataset.

Dataset	$\|\mathbf{x}\|$	K	Empirical class distribution	Occurrences	IR	ID$_{EU}$	ID$_{CH}$	ID$_{KL}$	ID$_{HE}$	ID$_{TV}$	ID$_{CS}$
autos	25	6	0.02/0.13/0.30/0.29/0.18/0.08	3/20/48/46/29/13	16.0	2.44	2.30	2.24	2.44	2.55	2.19
balance	4	3	0.46/0.08/0.46	288/49/288	5.9	0.66	0.76	0.40	0.53	0.76	0.44
contraceptive	9	3	0.43/0.23/0.35	629/333/511	1.9	0.30	0.32	0.07	0.20	0.32	0.09
dermatology	34	6	0.31/0.17/0.20/0.13/0.14/0.05	112/61/72/49/52/20	5.6	2.32	2.28	2.11	2.29	2.34	2.10
ecoli	7	8	0.43/0.23/0.15/0.10/ 0.06/0.01/0.01/0.01	143/77/52/35/ 20/5/2/2	71.5	4.56	4.48	4.42	4.61	4.70	4.31
glass	9	7	0.33/0.36/0.08/0.00/0.06/0.04/0.14	70/76/17/0/13/9/29	8.4 (∞)	4.44	4.30	4.28	4.53	4.56	4.20
hayes-roth	4	3	0.39/0.39/0.23	51/51/30	1.7	0.28	0.32	0.06	0.19	0.32	0.08
lymphography	18	4	0.01/0.55/0.41/0.03	2/81/61/4	40.5	1.77	1.59	1.65	1.73	1.92	1.59
new-thyroid	5	3	0.70/0.16/0.14	150/35/30	5.0	1.55	1.55	1.25	1.40	1.55	1.30
pageblocks	10	5	0.90/0.06/0.01/0.02/0.01	492/33/8/12/3	164.0	3.87	3.87	3.73	3.75	3.87	3.76
penbased	16	10	0.10/0.10/0.10/0.10/0.10/ 0.10/0.10/0.10/0.10/0.10	115/114/114/106/114/ 106/105/115/105/106	1.1	4.02	4.01	4.00	4.02	4.04	4.00
shuttle	9	7	0.78/0.00/0.00/0.16/0.06/0.00/0.00	1706/2/6/338/123/0/0	853.0 (∞)	4.90	4.90	4.83	4.88	4.92	4.82
thyroid	21	3	0.02/0.05/0.93	17/37/666	39.2	1.89	1.89	1.72	1.73	1.89	1.79
wine	13	3	0.33/0.40/0.27	59/71/48	1.5	1.11	1.10	1.01	1.09	1.10	1.01
yeast	8	10	0.16/0.29/0.31/0.03/0.03/ 0.11/0.02/0.02/0.01/0.00	244/429/463/44/51/ 163/35/30/20/5	92.6	5.54	5.35	5.42	5.60	5.79	5.29

Table 4: Pearson correlation coefficient (×100) among the performance of the major learning paradigms on the datasets of Table 3 and the studied summaries.

Summary	Decision trees (C4.5)			Decision rules (RIPPER)			Connectionism (Neural Net.)			Probabilistic (Naïve Bayes)			Statistical learn. (SVM)		
	$\mathcal{A}$	$\mathcal{G}$	min	$\mathcal{A}$	$\mathcal{G}$	min	$\mathcal{A}$	$\mathcal{G}$	min	$\mathcal{A}$	$\mathcal{G}$	min	$\mathcal{A}$	$\mathcal{G}$	min
IR	−9	−42	−41	14	−41	−38	−11	−50	−43	−4	−42	−39	−14	−34	−33
ID$_{EU} - (m+1)$	−27	−46	−52	−15	−46	−45	−51	−65	−61	−40	−56	−57	−60	−63	−65
ID$_{CH} - (m+1)$	−19	−39	−42	−4	−37	−34	−35	−50	−44	−29	−50	−51	−55	−56	−59
ID$_{KL} - (m+1)$	−25	−50	−54	−13	−50	−49	−54	−75	−72	−39	−59	−63	−54	−60	−61
ID$_{HE} - (m+1)$	−34	−56	−63	−25	−57	−59	−59	−77	−75	−48	−67	−68	−59	−69	−71
ID$_{TV} - (m+1)$	−42	−60	−66	−32	−60	−59	−62	−73	−69	−52	−64	−66	−61	−68	−70
ID$_{CS} - (m+1)$	−14	−38	−42	−1	−38	−36	−44	−63	−60	−31	−50	−54	−53	−55	−56



(a) CDD for the arithmetic mean among the recalls, $\mathcal{A}$.

(b) CDD for the geometric mean among the recalls, $\mathcal{G}$.

(c) CDD for minimum recall obtained, min.

Figure 4: Pearson correlation ranking between the performance of the supervised learning paradigms on the studied datasets and the summaries, $\alpha = 0.05$.

Then, the correlation between the estimated values for the performance scores and the summaries, IR and ID, are determined using the Pearson product-moment correlation coefficient [25] so that H_2 may be checked. Since a licit calculation of the correlation requires an ideal scenario with a fixed number of minority/majority classes, we emulate this requirement by subtracting $(m - 1)$ from the ID value before the calculation so that all considered classification problems are normalised in the same range $[0, 1]$. The results are presented in Table 4; rows represent the summaries and columns represent the estimated values for each score in each learning paradigm. Since the utilised scores assign higher values to better performance, an adequate summary is expected to have a negative correlation; the lowest the correlation, the better the sensitivity. We conclude from the results that summaries are, in general, negatively correlated with the performance of the classifiers, and that instantiations of ID are more sensitive than IR as the former obtain a lower negative correlation. The best results (highlighted in Table 4) are obtained by ID_{TV} and ID_{HE}.

Finally, to determine if there are summaries significantly more sensitive to the hindrance produced by skewed class distributions, a statistical hypothesis testing procedure is performed: Friedman test [8] with Shaffer's static post-hoc with $\alpha = 0.05$ [12]. The test results are represented by means of critical difference diagrams (CDD) [8], which show, in a numbered line, the arithmetic mean of the ranks of the correlation between each summary and the estimation of each score in the database. If there is no statistically significant difference between two summaries, they are connected in the diagram by a straight grey line. Figures 4a, 4b, and 4 show the CDD for the Pearson correlation between the summaries and $\mathcal{A}$, $\mathcal{G}$ and min, respectively. Results confirm the second hypothesis, in all rankings, IR shows the worst behaviour and significant differences are found between IR and other instantiations of ID for the performance scores. Moreover, they also show that instantiating ID using either Hellinger or total variation (Manhattan) distances produces significant robust summaries of the class-imbalance extent.

5 Summary

Authors often measure the class-imbalance extent in their experimental schemas when there is a reasonable suspicion of having unbalanced problems in the checked database. Up to now, the most utilised summary of the class-imbalance extent of a dataset was the imbalance-ratio, i.e. the (expected) number of instances of the most probable class for each instance of the least probable class. Although it is a powerful measure for binary problems, in this paper, we prove that it is a suboptimal summary for the multi-class scenario. For that reason, we propose a new more adequate and robust summary of the class-imbalance extent to deal with multiple classes, named *imbalance-degree.* It has three interesting properties: (i) it is a single easy-readable real value in the range $[0, K)$, where K is the number of classes. (ii) Depending on the requirements of the sensitivity sought in the tackled problem, it can be instantiated by any chosen metric or divergence. (iii) It is an injective function for different class distributions showing different numbers of majority/minority classes. Empirical results show that imbalance-degree has a higher resolution and is more sensitive to express the hindrance that skewed class distributions cause in the traditional supervised algorithms than imbalance-ratio. Additionally, it can also be concluded that either Hellinger, total variation or Manhattan distances are recommended distance/similarity functions to instantiate our proposal, imbalance-degree.

This work can be extended in several ways. For example, only 8 different distance/similarity functions over 15 datasets are used in this paper. A more exhaustive analysis can be carried out using a larger number of distance/similarity functions [14, 5] over a larger set of unbalanced problems in order to statistically determine which functions behave differently and are recommended for highly different class-imbalanced scenarios.

Another straightforward future path to this research can be a study on the variation of the correlation between ID and the performance of the classifiers when class-imbalance techniques, such as SMOTE [6], are used. This could be a step forward in determining which intrinsic features of the data are affecting the classifiers [2], and whether the performance of a classifier can be predicted based upon the available data [30]. However, note that, although the negative correlation between ID and the performance is expected to decrease as long as the class-imbalance techniques alleviate the hindering effect of the class distribution, there might exist other hindering aspects [23] which may harm the performance of the classifiers.

Acknowledgments

This work is partially supported by the Basque Government (IT609-13 and Elkartek BID3A) and the Spanish Ministry of Economy and Competitiveness (TIN2013-41272P). Moreover, Jonathan Ortigosa-Hernández is partly financed by the Ministry of Science and Innovation (MEC-FPU AP2008-00766) and Jose A. Lozano by both the Basque Government (BERC program 2014-2017) and the Spanish Ministry of Economy and

Competitiveness (Severo Ochoa Program SEV-2013-0323).

References

[1] Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrara, F. Bkeel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing 17*, 2–3 (2011), 255–287.

[2] Anwar, N., Jones, G., and Ganesh, S. Measurement of data complexity for classification problems with unbalanced data. *Statistical Analysis and Data Mining* (2014), 194–211.

[3] Batista, G., Prati, R. C., and Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explorations Newsletter 6*, 1 (Jun. 2004), 20–29.

[4] Bhattacharyya, A. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society 35* (1943), 99–109.

[5] Cha, S. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences 1* (Jan. 2007), 300–307.

[6] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16* (Jan. 2002), 321–357.

[7] Csiszár, I. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutató Int. Közl 8* (1963), 85–108.

[8] Demsar, J. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research 7* (2006), 1–30.

[9] Deza, M. M., and Deza, E. *Encyclopedia of Distances.* Springer-Verlag Berlin Heidelberg, 2009.

[10] Drummond, C., and Holte, R. Severe class imbalance: Why better algorithms aren't the answer. In *Proc. of the 16th European Conf. on Machine Learning* (Porto, Portugal, Oct. 2005), pp. 539–546.

[11] Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. on Systems, Man, and Cybernetics - Part C: Applications and Reviews 42*, 4 (Jul. 2012), 463–484.

[12] García, S., and Herrera, F. An extension on "statistical comparisons of classifiers over multiple data sets". *Journal of Machine Learning Research 9* (2008), 2677–2694.

[13] Ghamen, A. S., Venkatesh, S., and West, G. Multi-class pattern classification in imbalanced data. In *Proc. of the Int. Conf. on Pattern Recognition 2010* (2010), pp. 2881–2884.

[14] Gibbs, A. L., and Su, F. E. On choosing and bounding probability metrics. *International Statistical Review 70*, 3 (2002), 419–435.

[15] Gupta, M., Bengio, S., and Weston, J. Training highly multiclass classifiers. *Journal of Machine Learning Research 15* (Apr. 2014), 1461–1492.

[16] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, R., and Witter, I. The WEKA data mining software: An update. *SIGKDD Explorations 11*, 1 (2009), 10–18.

[17] He, H., and Garcia, E. A. Learning from imbalanced data. *IEEE Trans. on Knowledge and Data Engineering 21*, 9 (Sep. 2009), 1263–1284.

[18] Hellinger, E. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik 136* (Jan. 1909), 210–271.

[19] Kullback, S. *Information Theory and Statistics*. Wiley, 1959.

[20] Kullback, S., and Leibler, R. A. On information and sufficiency. *The Annals of Mathematical Statistics 22*, 1 (Mar. 1951), 9–86.

[21] Lichman, M. UCI machine learning repository, 2013.

[22] Liu, X.-Y., and Zhou, Z.-H. The influence of class imbalance on cost-sensitive learning: An empirical study. In *Proc. of the 6th Int. Conf. on Data Mining (ICDM'06)* (Hong Kong, China, Dec. 2006), pp. 970–974.

[23] López, V., Fernández, A., García, S., Palade, V., and Herrera, F. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences 250* (Nov. 2013), 113–141.

[24] Ortigosa-Hernández, J., Inza, I., and Lozano, J. A. Towards competitive classifiers for the class-imbalance problem. *arXiv:1608.08984* (2016), 1–21.

[25] Pearson, K. Notes on regression and inheritance in the case of two parents. In *Proc. of the Royal Society of London 58*. Jun. 1895, pp. 240–242.

[26] Pearson, K. *On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling.* 1900.

[27] Prati, R. C., Batista, G. E., and Silva, D. F. Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. *Knowledge and Information Systems 45*, 1 (Nov. 2015), 247–270.

[28] Sáez, J., Krawczyk, B., and Woźniak, M. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition 57* (Sep. 2016), 164–178.

[29] Shannon, C. E. A mathematical theory of communication. *Bell system technical journal 27*, 3 (1948), 379–423.

[30] Sotoca, J., Sánchez, J., and Mollineda, R. A review of data complexity measures and their applicability to pattern classification problems. In *Proceedings of the III Simposio de Teoria y Aplicaciones de Mineria de Datos (TAMIDA 2005)* (2005), pp. 77–83.

[31] Wang, S., and Yao, X. Multi-class imbalance problems: Analysis and potential solutions. *IEEE Trans. on Systems, Man and Cybernetics - Part B 42*, 4 (Aug. 2012), 1119–1130.

Supplementary Material

The source code, in `Python 2.7`, used for the empirical studies of this manuscript can be downloaded from `http://github.com/jonathanSS/ImbalanceDegree`.