Pattern Recognition Letters 98 (2017) 32–38

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Measuring the class-imbalance extent of multi-class problems

Jonathan Ortigosa-Hernández[a,*], Iñaki Inza[a], Jose A. Lozano[a,b]

[a] Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, Computer Science Faculty, The University of the Basque Country UPV/EHU, P. Manuel Lardizabal 1, 20018, Donostia-San Sebastián, Spain
[b] Basque Center for Applied Mathematics BCAM, Alameda de Mazarredo 14, 48009, Bilbao, Spain

ARTICLE INFO

Article history:
Received 2 October 2016
Available online 8 August 2017

MSC:
41A05
41A10
65D05
65D17

Keywords:
Class-imbalance problem
Skewed class distribution

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

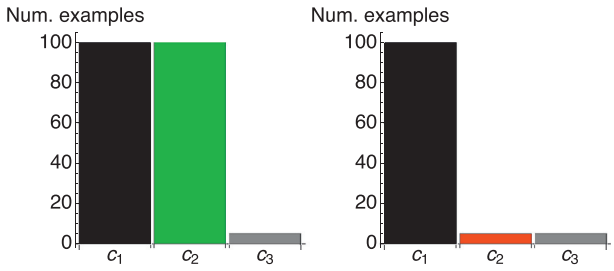
Most of the well-known traditional machine learning techniques are designed to solve classification problems showing reasonably balanced class distributions [24]. However, this assumption does not always hold in reality. Occasionally, real-world problems have skewed class distributions and, due to this, 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 appli-

* Corresponding author.
E-mail address: jonathan.ortigosa@ehu.es (J. Ortigosa-Hernández).

http://dx.doi.org/10.1016/j.patrec.2017.08.002
0167-8655/



(a) Best unbalanced scenario. (b) Worst unbalanced scenario.

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

cability 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 Fig. 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 – Fig. 1a –, is far less problematic than having only 5 examples of the second class – Fig. 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 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,

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

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.

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

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,

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

Here, $\mathbb{1}(\mathcal{E})$ is the indicator function, 1 if the event $\mathcal{E}$ is true, 0 otherwise. Note that Fig. 1a and Fig. 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