Relevance-based Evaluation Metrics for Multi-class Imbalanced Domains

Paula Branco[1,2], Luís Torgo[1,2], and Rita P. Ribeiro[1,2]

[1] LIAAD - INESC TEC
[2] DCC - Faculdade de Ciências - Universidade do Porto
paula.branco@dcc.fc.up.pt ltorgo@dcc.fc.up.pt rpribeiro@dcc.fc.up.pt

Abstract. The class imbalance problem is a key issue that has received much attention. This attention has been mostly focused on two-classes problems. Fewer solutions exist for the multi-classes imbalance problem. From an evaluation point of view, the class imbalance problem is challenging because a non-uniform importance is assigned to the classes. We propose a relevance-based evaluation framework that incorporates user preferences by allowing the assignment of differentiated importance values to each class. The presented solution is able to overcome difficulties detected in existing measures and increases discrimination capability. The proposed framework requires the assignment of a relevance score to the problem classes. To deal with cases where the user is not able to specify each class relevance, we describe three mechanisms to incorporate the existing domain knowledge into the relevance framework. These mechanisms differ in the amount of information available and assumptions made regarding the domain. They also allow the use of our framework in common settings of multi-class imbalanced problems with different levels of information available.

1 Introduction

The class imbalance problem is a relevant problem with extensive research literature. It occurs in many application domains like medical, financial, meteorological, and others. Assessing performance in these contexts has been studied and several metrics were proposed. However, most proposals for this type of problems are only applicable to binary classification problems [7]. Recently, the multi-class imbalance problem has received increased attention.

In this paper, we address the key issue of performance assessment for multi-class imbalanced domains. These domains require special purpose evaluation metrics that are able to adequately reflect the preference biases of the users concerning prediction errors. In imbalanced domains, the user is typically more interested in the minority class(es) while the majority class(es) are usually less relevant. Therefore, traditionally used measures, such as Accuracy, are not suitable for this type of problems due to their inability of taking into account the user preferences. For multi-class imbalanced domains the few solutions that exist are essentially extensions of metrics used for the binary case.

There is a direct connection between imbalanced domains and cost-sensitive learning. However, when we face a cost-sensitive problem we have a cost matrix defined for the task at hand that is used to assess the models performance. The model with minimum cost (or maximum benefit) is the best. The tasks we are addressing in this paper are different because there is a class imbalance but no cost matrix is available. This is the usual setting when dealing with imbalanced classes. Typically, the only information available regarding the user preferences is informal and can be expressed as: "*the minority class(es) is(are) the most important one(s)*". This is an important class of applications as it is well known that cost/benefit information is frequently hard to obtain or simply not available.

When the user preference bias is not uniform across the domain of the target variable it is important to transfer this information to the evaluation metrics. We propose a new evaluation framework that incorporates this information. The proposed measures are based on the existence of different relevance/importance scores for the problem classes and try to mirror the user preference bias in the evaluation of the predictions of a model. This means that the same errors made in two different classes with different importance scores can have different weights in the final evaluation score. We also propose three mechanisms for estimating the expected domain preferences in a typical imbalanced multi-class setting. These mechanisms can be used when the user is not able to precisely specify each class relevance. The proposed mechanisms differ in the assumptions regarding the domain and amount of information that the user is able to provide.

The main contributions of this work are: i) highlight that existing metrics for handling multi-class imbalanced domains are not always adequate ; ii) propose a new evaluation framework that acounts for user preferences in multi-class imbalanced domains; iii) propose three mechanisms for estimating the preference bias in typical multi-class imbalance settings; and iv) compare the discrimination capability of existing and new proposed metrics for this problem. This paper is organized as follows. Section 2 describes existing metrics for handling multi-class imbalance domains. Section 3 explains why these metrics are unsuitable for this problem providing three examples where those metrics show unreliable results. Section 4 presents our framework for performance assessment on multi-class imbalance problems, and mechanisms to deal with different information levels. Section 5 evaluates our framework regarding performance and discrimination capability under different scenarios. Section 6 concludes the paper.

2 Evaluation Metrics for Multi-class Imbalanced Learning

Several metrics have been proposed to evaluate the performance within the problem of class imbalance for two classes. However, only a few have been successfully adapted to address the more difficult problem of multi-class imbalanced domains.

Let C represent the total number of classes of a problem. Consider a $C \times C$ confusion matrix, mat, for which $mat_{k,l}$ represents the examples of the true class k that were predicted as class l. For a class i, tp_i represents the true positives for class i; tn_i are the true negatives for class i, i.e., all the examples that were

correctly predicted and are not from class i; fp_i is the number of false positive for class i, i.e. all the examples incorrectly predicted as class i, and fn_i are the false negatives for class i. We use t_i and p_i for the total number of true and predicted examples for class i respectively, i.e., $t_i = tp_i + fn_i$ and $p_i = tp_i + fp_i$. The indexes M and μ represent respectively a Macro and Micro averaging strategy for a metric, where the first strategy averages the metric results over all classes while the second uses the pooled results. With this notation, we define the following metrics for a class i:

$$recall_i = \frac{tp_i}{t_i} \quad (1) \quad precision_i = \frac{tp_i}{p_i} \quad (2) \quad F_{\beta i} = \frac{(1+\beta^2) precision_i \cdot recall_i}{\beta^2 \cdot precision_i + recall_i} \quad (3)$$

where β sets the relative importance of $recall_i$ in comparison with $precision_i$.

Table 1 presents a description of the existing metrics for multi-class imbalance tasks. For a more comprehensive overview, we also include some multi-class measures which were not specifically developed for imbalanced domains. The Area Under the ROC Curve (AUC) is not considered in this paper. Although some attempts have been made to also adapt AUC to a multi-class context [9] we opted not to include it here for two reasons. The first reason is related to the demonstrated incoherence of AUC metric [8]. The second reason concerns the nonexistence of a well-developed ROC analysis for multi-class problems [14].

The metrics described in Table 1 can be clustered into recall-based ($MAvG$, Rec_M, Rec_μ), precision-based ($Prec_M$, $Prec_\mu$) or general metrics ($AvAcc$, $F_{\beta M}$ $F_{\beta\mu}$, AvF_β, CBA, MCC, RCI and CEN) depending on the information used. Thus, each type of metric presents a different evaluation perspective. While recall-based metrics are focused on the true class labels, precision-based metrics consider the predicted class labels and the general metrics aggregate both perspectives into a single value providing a global performance overview. An alternative solution to Table 1 metrics consist of not aggregating the $precision_i$, $recall_i$ and $F_{\beta i}$ measures. However, this has the disadvantage of generating a large number of results increasing the complexity of the analysis of the results.

The metrics in Table 1 present differences in both the range of values they may take and the representation of the best performing classifier. For a straightforward comparison we present the metric **value** and a **normalized value**. This **normalized value** corresponds to the metric value in a percentage, where 0% matches the worst possible performance and 100% the best.

3 Unsuitability of the Existing Evaluation Metrics

The so-called "imbalanced problems" are based on the assumption that the user has a differentiated interest in the problem classes. In two-class problems the user preference bias is, usually, towards the minority class. This also happens in the multi-class context.

Several metrics have been proposed (cf. Table 1) to assess the performance in multi-class imbalanced domains. We claim that these solutions are not adequate for these domains because they fail to reflect the user preferences in several situations and therefore can be misleading. To demonstrate this, we use the

Table 1: Performance assessment metrics for imbalanced domains with C classes.

Metric	Description	Definition
$AvAcc$	Classes average accuracy.	$\frac{1}{C}\sum_{i=1}^{C}\frac{tp_i+tn_i}{tp_i+tn_i+fp_i+fn_i}$
$MAvG$	Geometric average of recall in each class [15].	$\sqrt[C]{\prod_{i=1}^{C} recall_i}$
Rec_M	Arithmetic Macro-average of recall in each class.	$\frac{1}{C}\sum_{i=1}^{C} recall_i$
$Prec_M$	Arithmetic Macro-average of precision in each class.	$\frac{1}{C}\sum_{i=1}^{C} precision_i$
Rec_μ	Arithmetic Micro-average of recall in each class.	$\sum_{i=1}^{C} tp_i / \sum_{i=1}^{C} t_i$
$Prec_\mu$	Arithmetic Micro-average of precision in each class.	$\sum_{i=1}^{C} tp_i / \sum_{i=1}^{C} p_i$
$F_{\beta M}$	Mean F_β measure evaluated with Macro-averaged precision and recall [14].	$\frac{(1+\beta^2)\cdot Prec_M \cdot Rec_M}{\beta^2 \cdot Prec_M + Rec_M}$
$F_{\beta\mu}$	Mean F_β measure evaluated with Micro-averaged precision and recall [14].	$\frac{(1+\beta^2)\cdot Prec_\mu \cdot Rec_\mu}{\beta^2 \cdot Prec_\mu + Rec_\mu}$
AvF_β	Extension for any value of β of the definition for F_1 measure to multi-class [4].	$\frac{1}{C}\sum_{i=1}^{C}\frac{(1+\beta^2)\cdot precision_i \cdot recall_i}{\beta^2 \cdot precision_i + recall_i}$
CBA	Class Balance Accuracy [12].	$\frac{\sum_{i=1}^{C}\frac{mat_{i,i}}{max\left(\sum_{j=1}^{C} mat_{i,j}, \sum_{j=1}^{C} mat_{j,i}\right)}}{C}$
MCC	Matthews Correlation Coefficient introduced for two-class problems and extended to multi-class [11, 6].	$\frac{X}{Y\,Z}$, where $X=\sum_{k,l,m=1}^{C}(mat_{k,k}mat_{m,l}-mat_{l,k}mat_{k,m})$ $Y=\sqrt{\sum_{k=1}^{C}\left(\sum_{l=1}^{C} mat_{l,k}\right)\left(\sum_{\substack{f,g=1\\ f\neq k}}^{C} mat_{g,f}\right)}$ $Z=\sqrt{\sum_{k=1}^{C}\left(\sum_{l=1}^{C} mat_{k,l}\right)\left(\sum_{\substack{f,g=1\\ f\neq k}}^{C} mat_{f,g}\right)}$
RCI	Relative Classifier Information [13]	$\frac{H_d-H_o}{H_d}$, where $H_d=-\sum_{i=1}^{C}\left(\frac{\sum_{l=1}^{C} mat_{i,l}}{C} log \frac{\sum_{l=1}^{C} mat_{i,l}}{C}\right)$ $H_o=\sum_{j=1}^{C}\left(\frac{\sum_{k=1}^{C} mat_{k,j}}{C} H_{oj}\right)$ and $H_{oj}=-\sum_{i=1}^{C}\left(\frac{mat_{i,j}}{\sum_{k=1}^{C} mat_{k,j}} log \frac{mat_{i,j}}{\sum_{k=1}^{C} mat_{k,j}}\right)$
CEN	Confusion Entropy [16].	$\sum_{j=1}^{C}(P_j CEN_j)$, where $P_j=\frac{\sum_{k=1}^{C} mat_{j,k}+mat_{k,j}}{2*\sum_{k,l=1}^{C} mat_{k,l}}$, $CEN_j=-\sum_{\substack{k=1\\ k\neq j}}^{C}(P^j_{j,k}\log_{2(C-1)}(P^j_{j,k})+P^j_{k,j}\log_{2(C-1)}(P^j_{k,j}))$ $P^i_{i,i}=0,\quad P^i_{i,j}=mat_{i,j}/\left(\sum_{k=1}^{C}(mat_{i,k}+mat_{k,i})\right), i\neq j$

three cases described below. The user can also follow the strategy of observing each class precision, recall and F_β. To show this perspective, we also include the evaluation provided by these measures for each class in the next examples.

Multi-class imbalance problems can be grouped into: multi-minority, multi-majority and complete. In a multi-minority scenario one class has significantly more examples than the mean number of examples of all classes, i.e., $t_{maj} >> \bar{t}$,

Table 2: Cases 1 to 3 confusion matrix (top) and $prec_i$, rec_i and F_{1i} (bottom).

Case 1

trues \ preds	c_1	c_2	c_3
c_1	5	0	0
c_2	0	10	0
c_3	0	300	0

Class	rec_i	$prec_i$	F_{1i}
c_1	1	1	1
c_2	1	0.032	0.063
c_3	0	n. def.	n. def.

Case 2

trues \ preds	c_1	c_2	c_3
c_1	1	0	3
c_2	0	100	0
c_3	0	0	200

Class	rec_i	$prec_i$	F_{1i}
c_1	0.25	1	0.4
c_2	1	1	1
c_3	1	0.985	0.993

Case 3

trues \ preds	c_1	c_2	c_3	c_4
c_1	1	3	0	0
c_2	9	1	0	0
c_3	0	0	100	0
c_4	0	0	0	200

Class	rec_i	$prec_i$	F_{1i}
c_1	0.25	0.1	0.14
c_2	0.1	0.25	0.14
c_3	1	1	1
c_4	1	1	1

Table 3: Performance assessment metrics in Case 1,2 and 3. N.Val: normalized value; Ac.: Accordance with user preferences (misleading: ×, suitable: ✓).

Metric	Case 1			Case 2			Case 3		
	N.Val.(%)	Value	Ac.	N.Val.(%)	Value	Ac.	N.Val.(%)	Value	Ac.
$AvAcc$	36.5	0.365	×	99.3	0.993	×	98.1	0.981	×
$MAvG$	0.0	0.000	×	63.0	0.630	✓	39.8	0.398	✓
Rec_M	66.7	0.667	✓	75.0	0.750	×	58.8	0.588	×
$Prec_M$	not defined		×	99.5	0.995	✓	58.8	0.588	×
Rec_μ	4.8	0.048	×	99.0	0.990	×	96.2	0.962	×
$Prec_\mu$	4.8	0.048	×	99.0	0.990	✓	96.2	0.962	×
F_{1M}	not defined		×	85.5	0.855	×	58.8	0.588	×
$F_{1\mu}$	4.8	0.048	×	99.0	0.990	×	96.2	0.962	×
AvF_1	not defined		×	79.8	0.798	×	57.1	0.571	×
CBA	34.4	0.344	×	74.5	0.745	✓	55.0	0.550	×
MCC	65.1	0.301	✓	98.9	0.978	×	96.2	0.923	×
RCI	36.8	0.368	×	92.6	0.926	×	97.9	0.979	×
CEN	97.8	0.022	✓	98.1	0.019	×	98.5	0.015	×

where $\bar{t} = \sum_{i=1}^{C} t_i/C$ is the mean number of examples of all classes. On a multi-majority case a single class is significantly less frequent than the others, i.e., $t_{min} << \bar{t}$. In the complete case, several classes can have a significantly larger size than other classes which have a significantly smaller size relatively to $\bar{t}$.

The cases described below exemplify the depicted scenarios. They illustrate the unsuitability of the existing metrics and show the need of a more adequate framework for this context. We assume that the most relevant classes are the less populated. Tables 2 and 3 describe these cases.

Case 1: Multi-minority Example - In this case, the two minority classes are correctly predicted and the majority class is completely mispredicted.

Case 2: Multi-majority Example - In this case both majority classes are correctly predicted and the minority class is nearly always mispredicted.

Case 3: Complete Example - In this case two majority classes are correctly predicted while the two minority classes are almost always mispredicted.

Table 3 includes a summary of the misleading metrics for the cases presented. Generally, we observe that the metrics fail to correctly represent the user preferences. Either by providing an over- or under-estimated value, the metrics are not able to correctly incorporate the domain knowledge, and therefore, the results

obtained are not reliable. In more detail, for case 1, $MAvG$, provides a result of zero which is clearly not adequate given that both minority classes have a perfect score regarding the recall metric, a problem also observed by [12]. The remaining metrics marked in Table 3 for case 1 are misleading because they present a normalized value approximately below 45%. In case 2, the minority and most important class was almost always incorrectly predicted. However, all metrics, with exception of $MAvG$, CBA, $Prec_M$ and $Prec_\mu$, over-estimate the value of the confusion matrix which can be misleading. In case 3 the metrics are unable to show that both minority and important classes were almost always incorrectly predicted. Although big mistakes occur on all minority classes, most metrics normalized value is high or moderate which is misleading.

The cases described show that no metric provides reliable results in all situations. When the classes have a distinct relevance to the user it is unavoidable to consider this relevance in the evaluation. Thus, a new framework is required for embedding the relevance into the existing metrics. This framework should also be usable when the user has a more informal information. So, mechanisms for embedding different levels of information provided by the user are necessary.

4 A Framework for Relevance-based Evaluation

4.1 Relevance-based Metrics for Multi-class Imbalance Learning

Our proposal is based on the assumption that classes have different relevance for the user. A certain number of classes may be extremely important while the performance on other classes may be negligible. The key idea is to use the relevance values as weights for the classes when evaluating the models performance.

The use of weights is a well-known strategy. However, only two metrics were proposed using this notion. A weighted macro-averaging recall [2] was proposed for multi-class although it was only used in binary classification. Moreover, no guidelines for defining/choosing the weights were provided. A weighted AUC for multi-class was presented [10], with weights determined by the classes prevalence.

Our relevance-based metrics proposal assumes that the user assigns an importance score to each problem class. Let us suppose that this domain information is converted into a function $\phi()$ that maps each class into a relevance score in the interval $[0, 1]$. The value 0 is assigned to a class with zero relevance, and the value 1 is assigned to a class with maximum relevance to the user. For instance, a relevance function for a four-class problem can be define as: $\phi(c_1) = 0.2$, $\phi(c_2) = 0$, $\phi(c_3) = 0.9$ and $\phi(c_4) = 1$. Fromthis illustrative $\phi()$ function, class c_1 has a very low relevance, class c_2 is irrelevant, and classes c_3 and c_4 are very relevant.

Our proposal incorporates the user preference bias, expressed by the definition of a relevance function, in the metrics definition in the form of weights. This means that, if a class is very important to the user, then the performance on that class will also have a large weight in the evaluation. On the other hand, misclassification errors of less relevant classes have a reduced impact on the final evaluation. Equations 4 to 8 present an adaptation of recall, precision, $F_\beta - measure$ and CBA to incorporate relevance.

$$Rec^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \phi(i) \cdot recall_i \quad (4) \qquad Prec^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \phi(i) \cdot precision_i \quad (5)$$

$$F_{\beta}^{\phi} = \frac{(1+\beta^2) \cdot Prec^{\phi} \cdot Rec^{\phi}}{(\beta^2 \cdot Prec^{\phi}) + Rec^{\phi}} \quad (6)$$

$$AvF_{\beta}^{\phi} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \frac{\phi(i) \cdot (1+\beta^2) \cdot precision_i \cdot recall_i}{(\beta^2 \cdot precision_i) + recall_i} = \frac{1}{\sum_{i=1}^{C} \phi(i)} \sum_{i=1}^{C} \frac{\phi(i) \cdot (1+\beta^2) \cdot tp_i}{\beta^2 \cdot t_i + p_i} \quad (7)$$

$$CBA^{\phi} = \sum_{i=1}^{C} \phi(i) \cdot \frac{mat_{i,i}}{max\left(\sum_{j=1}^{C} mat_{i,j}, \sum_{j=1}^{C} mat_{j,i}\right)} \quad (8)$$

where $\phi(i)$ is the relevance of class i; t_i and p_i are the total number of true and predicted examples for class i; and tp_i is the number of true positives for class i.

With this framework we obtain the three evaluation perspectives: recall-based, precision-based and general measures. These metrics were selected because they cover all perspectives under a simple formulation.

4.2 Mechanisms for Relevance Estimation

The above evaluation framework depends on the availability of domain information regarding the classes relevance. However, this information may exist with different levels of detail. We will consider 4 types of information:

-**Informal**: characterized by completely informal domain knowledge. This is typical in imbalanced domains where no quantification regarding the importance of each class exists. Frequently, it is only stated that "the minority classes are the most important". This creates serious problems to the performance evaluation because the user does not specify the classes non-uniform importance.

-**Intermediate informal**: more information available although very limited. We assume the user provides a partial order of the classes by their importance.

-**Intermediate formal**: more complete information available. We consider that the user is able to provide a total order of the classes.

-**Formal**: the user provides a full specification of the relevance function. Although being the ideal setting, this is not so common in real world domains.

We will present mechanisms to estimate the relevance function from these different levels of available information. If the user fully specifies the relevance function (formal level) no mechanism is needed. To denote this situation we will add ϕ to the metrics name. The proposed mechanisms are pertinent because for most imbalance domains the full relevance function is unknown. Our goal is to incorporate the available domain knowledge in the evaluation framework.

Informal Level - Using Classes Prevalence (PREV)

When no preferences regarding the domain are provided, it is possible to use the observed frequency of the classes to obtain valid relevance scores. Our proposal sets the relevance of a class to be inversely proportional to its observed frequency in the available data:

$$\hat{\phi}(i) = \frac{1/t_i}{\sum_{i=1}^{C} 1/t_i} \quad (9)$$

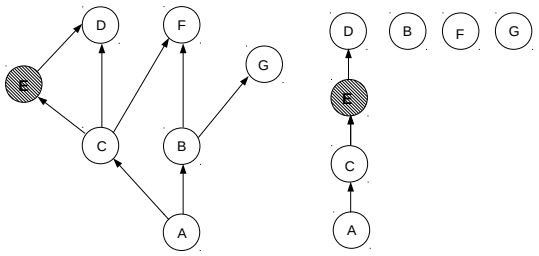

Fig. 1: An example of a partially ordered set (left hand side) and the construction of a LPOM for class E (right hand side).

where t_i is class i total number of examples. Using the estimated relevance we may obtain any of the proposed relevance-based metrics. We stress that the use of this method is not mandatory for applying our framework, provided that the user gives more domain information or specifies a relevance function.

Intermediate Informal Level - Using Classes Partial Order (PO)

A partial order specifies a binary relation which may hold between some pairs of classes. This relation is denoted as $c_1 < c_2$ and is read as "c_1 precedes c_2". In the context of relevance-based metrics, the relation $c_1 < c_2$ represents that c_1 has a lower relevance value than c_2, i.e. c_1 is less important than c_2. The relation is named partial because it does not provide a full relation between all the classes, i.e., there are pairs of classes named incomparable because the relation between both was not specified. More details regarding partially ordered sets can be obtained in [3]. Figure 1 shows on the left side an example of a partial order on a problem with 7 classes. Several studies have been conducted to estimate rankings from a partial order (e.g. [1]). However, as far as we known, no attempt has been made to use the partial order of classes to estimate their relevance. The main advantage of this method is that it is less demanding for the user when compared to a full specification of the relevance function. Moreover, to use a partial order of classes is preferable to not having any information at all .

To estimate the classes relevance using a partial order we will apply the US-model [1]. This method builds a Local Partial Order Model (LPOM) for each class. A LPOM for a class node X represents all the successor (S), predecessor (P) and incomparable (U) nodes in relation to X. Then, the estimated average rank of node X is defined as $Rank(X) = \frac{(|S|+1)+(|S|+1+|U|)}{2} = |S| + 1 + \frac{|U|}{2}$. Figure 1 (on the right) shows the LPOM for node E. In this example node E has 2 successors (nodes A and C), 1 predecessor (node D) and 3 incomparable nodes (B, F and G). Node E ranking, according to the proposed US-model, is $Rank(E) = 4.5$. Our proposal, uses the classes ranks derived from the partial order provided by the user and estimates the relevance of each class i as follows:

$$\hat{\phi}(i) = \frac{Rank(i)}{\max_{\forall i \in C} Rank(i)} \tag{10}$$

Intermediate Formal Level - Using Classes Total Order (TO)

Table 4: Case 1, 2 and 3 information for each mechanism.

	Case 1				Case 2				Case 3				
	$\phi(c_1)$	$\phi(c_2)$	$\phi(c_3)$	order	$\phi(c_1)$	$\phi(c_2)$	$\phi(c_3)$	order	$\phi(c_1)$	$\phi(c_2)$	$\phi(c_3)$	$\phi(c_4)$	order
PREV	0.66	0.33	0.01		0.94	0.04	0.02		0.64	0.32	0.03	0.02	
PO	1	1	0.4	$c_3 < c_1$ $c_3 < c_2$	1	0.5	0.5	$c_3 < c_1$ $c_2 < c_1$	1	0.86	0.57	0.42	$c_3 < c_1$ $c_4 < c_1$ $c_4 < c_2$
TO	1	0.67	0.33	$c_3 < c_2 < c_1$	1	0.67	0.33	$c_3 < c_2 < c_1$	1	0.75	0.5	0.25	$c_4 < c_3 < c_2 < c_1$
ϕ	1	0.9	0.1		1	0.2	0.1		1	0.9	0.2	0.1	

This mechanism is similar to the previous one, but now the user is required to provide a total order of the problem classes. This is a more demanding task for the user because no pair of classes can remain incomparable. Still, it is less demanding than fully specifying the relevance function. Given a total order, only the magnitude of the classes relevance remains unspecified. We use the US-model [1] previously used in PO mechanism. For a node X, $Rank(X) = |S| + 1$ because X has no incomparable nodes. The relevance is estimated with Equation 10. The $\phi()$ function values are equidistant and range from $\frac{1}{C}$ to 1, where C is the number of classes. The metrics obtained by each described mechanism, have respectively $PREV$, PO and TO appended to their name.

4.3 Implementation Issues

To maximize the number of valid results supplied by the metrics, we exclude from the calculations of precision and recall-based metrics, all classes i for which $recall_i$ or $precision_i$ are not defined and use the AvF_1^ϕ extension presented in Equation 7. This way, we can always obtain Rec^ϕ, $Prec^\phi$and F_1^ϕ and maximize the number of obtained results for AvF_1^ϕ. With the extension proposed in [5] AvF_1^ϕ is only undefined when class i has neither true values nor predictions. To allow a fairer comparison we also applied these strategies to existing metrics.

5 Experimental Evaluation

5.1 Agreement with User Preferences

We will now present the performance of the proposed metrics in the cases described in Section 3. Table 4 provides the user-defined relevance and the relevance inferred from a simulation of incomplete user information using the mechanisms defined in Section 4.2. The performance results are shown in Table 5. Generally, we observe that the metrics considering the proposed evaluation framework are able to overcome the difficulties detected on the other existing metrics. The new metrics are capable of reflecting the user preferences independently of the level of information considered. It is noteworthy that for the most informal levels of information (PREV and PO) the results obtained for all the cases are preferable to those of the other existing metrics. Moreover, the results become more adjusted to the user preferences with the increase of the information level. In summary, all the proposed mechanisms show results that are more in accordance with the user preferences than the previous existing metrics.

Table 5: Performance assessment metrics normalized value for Cases 1, 2 and 3 (in bold: values in accordance with user preferences).

Metric	Case			Metric	Case			Metric	Case		
	1	2	3		1	2	3		1	2	3
$AvAcc$	36.5	99.3	98.1	Rec^{PREV}	**98.9**	**29.2**	**24**	Rec^{TO}	**83.3**	**62.5**	**43**
$MAvG$	0	**63**	**39.8**	$Prec^{PREV}$	**67.7**	**100**	**17.8**	$Prec^{TO}$	**61.3**	**99.8**	**41.5**
Rec_M	**66.7**	75	58.8	F_1^{PREV}	**80.4**	**45.3**	**20.4**	F_1^{TO}	**70.6**	**76.9**	**42.2**
$Prec_M$	34.4 [a]	**99.5**	58.8	AvF_1^{PREV}	**68**	**43.4**	**17.8**	AvF_1^{TO}	**52.1**	**69.9**	**40**
Rec_μ	4.8	99	96.2	CBA^{PREV}	**67**	**29.2**	**13.7**	CBA^{TO}	**51.1**	**62.3**	**37**
$Prec_\mu$	4.8	**99**	96.2	Rec^{PO}	**83.3**	**62.5**	**46.7**	Rec^{ϕ}	**95**	**42.3**	**29.1**
F_{1M}	45.4 [a]	85.5	58.8	$Prec^{PO}$	**51.6**	**99.6**	**46**	$Prec^{\phi}$	**54.2**	**99.9**	**28.4**
$F_{1\mu}$	4.8	99	96.2	F_1^{PO}	**63.7**	**76.8**	**46.4**	F_1^{ϕ}	**69**	**59.4**	**28.7**
AvF_1	35.4 [a]	79.8	57.1	AvF_1^{PO}	44.3	**69.8**	**44.3**	AvF_1^{ϕ}	**52.8**	**53.8**	**26**
CBA	34.4	**74.5**	55	CBA^{PO}	43	**62.1**	**41.5**	CBA^{ϕ}	**51.5**	**42.2**	**22.3**
MCC	**65.1**	98.9	96.2								
RCI	36.8	92.6	97.9								
CEN	**97.8**	98.1	98.5								

[a] Evaluated using the strategies described in Section 4.3

We also tested the proposed metrics on 16 real world data sets*. Although we observe differences in the metrics results, it is not possible to assess the agreement with the user preferences because, in this case, we lack a ground truth.

5.2 Discrimination Capability

In this section we assess how well the metrics recognize different situations expressed in the confusion matrix. We consider problems with 3 or 4 classes and determine the percentage of different scores obtained by each metric in all possible confusion matrices for a problem.

We tested the multi-minority, multi-majority and complete scenarios, with problems with 3, 3 and 4 classes respectively. A problem with 3 classes with i, j, and k examples is denoted by $i-j-k$. For instance, problem *2-4-15* has 2 , 4 and 15 examples of classes c_1, c_2 and c_3. We tested multi-minority ($i-j-k$) and multi-majority ($i-k-l$) problems with $i \in \{2,3\}$, $j \in \{4,5\}$, $k \in \{15,16\}$ and $l \in \{17,18\}$.We only analysed problems *2-3-9-10* and *2-3-9-11* on the complete scenario due to the exponential number of confusion matrices generated.

Figure 2 shows the difference between the discrimination percentage of pairs of metrics (a relevance-based metric and an existing metric). Relevance-based metrics achieve a higher discrimination capability when compared to their corresponding initial proposals. Only metrics based on recall and CBA present some difficulty in improving the discrimination capability where we obtain differences of zero or negative in 5% and 2% of the results respectively. The results show that the proposed evaluation framework is able to better discriminate different setups in multi-class imbalanced problems. In summary, our experiments show that our proposal provides an enhanced discrimination capability and results more in accordance with the user preferences.

*The experimental framework, code and results of this evaluation is available in `https://github.com/paobranco/Relevance-basedMulticlassImbalanceMetrics`

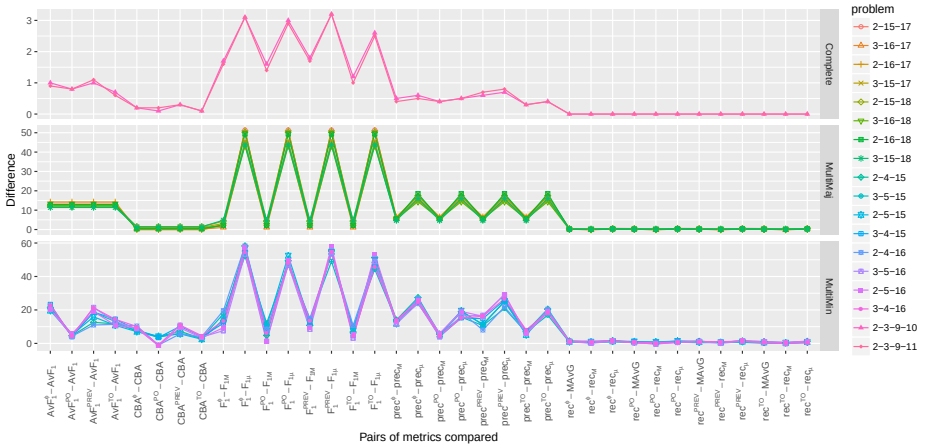

Fig. 2: Differences in the percentage of discrimination achieved between existing and corresponding proposed metrics in each scenario.

6 Conclusions

Class imbalance is a problem appearing in many relevant application domains. Performance assessment under this situation is a key issue that has been addressed, mainly, for the two-classes case. For the multi-class imbalance problem, only a few solutions exist. We have shown that existing metrics for multi-class imbalance domains are not adequate in certain cases. We propose a new relevance-based evaluation framework that integrates the notion of a non-uniform importance across the target variable domain through a relevance function.

The evaluation of imbalanced domains is still an open issue in two-classes and multi-class problems. Relevance-based metrics are suitable for evaluating predictive tasks on imbalanced domains because they are able to reflect the user preferences. Such metrics easily adapt to different types of domain knowledge. We provide three mechanisms to facilitate the users task of embedding domain knowledge into the proposed relevance framework for performance assessment. This integration boosts the capability of correctly reflecting the performance of cases that other measures are not able to capture. We also show that these metrics present an enhanced discrimination capability. For reproducibility purposes, all the code used in this paper is available in `https://github.com/paobranco/Relevance-basedMulticlassImbalanceMetrics`.

Acknowledgements

This work is financed by the ERDF – European Regional Development Fund through the Operational Programme for Competitiveness and Internationali-

sation - COMPETE 2020 Programme within project POCI-01-0145-FEDER-006961, and by National Funds through the FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) as part of project UID/EEA/50014/2013. The work of P. Branco is supported by a PhD scholarship of FCT (PD/BD/105788/2014).

References

[1] R. Brüggemann, P. B Sørensen, D. Lerche, and L. Carlsen. Estimation of averaged ranks by a local partial order model#. *Journal of chemical information and computer sciences*, 44(2):618–625, 2004.

[2] G. Cohen, M. Hilario, H. Sax, S. Hugonnet, and A. Geissbuhler. Learning from imbalanced data in surveillance of nosocomial infection. *Artificial Intelligence in Medicine*, 37(1):7–18, 2006.

[3] B. Dushnik and E. W Miller. Partially ordered sets. *American journal of mathematics*, 63(3):600–610, 1941.

[4] C. Ferri, J. Hernández-Orallo, and R Modroiu. An experimental comparison of performance measures for classification. *Pat. Rec. L.*, 30(1):27–38, 2009.

[5] G. Forman and M. Scholz. Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *SIGKDD Explor. Newsl.*, 12(1):49–57, 2010.

[6] J. Gorodkin. Comparing two k-category assignments by a k-category correlation coefficient. *Comp. Biol. Chem.*, 28(5):367–374, 2004.

[7] Q. Gu, L. Zhu, and Z. Cai. Evaluation measures of the classification performance of imbalanced data sets. In *ISICA*, pages 461–471. Springer, 2009.

[8] D. J Hand. Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine learning*, 77(1):103–123, 2009.

[9] D.J. Hand and R.J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *MACH*, 45(2):171–186, 2001.

[10] K. Hempstalk and E. Frank. Discriminating against new classes: One-class versus multi-class classification. In *AI 2008*, pages 325–336. Springer, 2008.

[11] B. W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *BBA-Protein Struct*, 405(2):442–451, 1975.

[12] L. Mosley. A balanced approach to the multi-class imbalance problem. *Graduate Theses and Dissertations. Paper 13537.*, 2013.

[13] V. Sindhwani, P. Bhattacharya, and S. Rakshit. Information theoretic feature crediting in multiclass support vector machines. In *SDM*, pages 1–18. SIAM, 2001.

[14] M. Sokolova and G. Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Proc. & Manag.*, 45(4):427–437, 2009.

[15] Y. Sun, M. S Kamel, and Y. Wang. Boosting for learning multiple classes with imbalanced class distribution. In *ICDM*, pages 592–602. IEEE, 2006.

[16] J.M. Wei, X.J. Yuan, Q.H. Hu, and S.Q Wang. A novel measure for evaluating classifiers. *Expert Systems with Applications*, 37(5):3799–3809, 2010.