Assessments Metrics for Multi-class Imbalance Learning: A Preliminary Study*

R. Alejo[1], J.A. Antonio[1], R.M. Valdovinos[2], and J.H. Pacheco-Sánchez[3]

[1] Tecnológico de Estudios Superiores de Jocotitlán, Carretera Toluca-Atlacomulco KM. 44.8, Col. Ejido de San Juan y San Agustín, 50700 Jocotitlán, México
[2] Centro Universitario UAEM Valle de Chalco, Universidad Autónoma del Estado de México Hermenegildo Galena No.3, Col. Ma. Isabel, 56615 Valle de Chalco, México
[3] Instituto Tecnológico de Toluca
Av. Tecnolgico s/n Ex-Rancho La Virgen, 52140, Metepec, México

Abstract. In this paper we study some of the most common global measures employed to measure the classifier performance on the multi-class imbalanced problems. The aim of this work consists of showing the relationship between global classifier performance (measure by global measures) and partial classifier performance, i.e., to determine if the results of global metrics match with the improved classifier performance over the minority classes. We have used five strategies to deal with the class imbalance problem over five real multi-class datasets on neural networks context.

Keywords: Multi-class imbalance, global measures, accuracy by class.

1 Introduction

Nowadays, the multi-class classification methods are increasingly required by modern applications (p.e. protein function classification, music categorization and semantic scene classification) [1]. However, while two-class classification problem has been widely studied, multi-class classification problem has not received much attention [2].

In addition, many multi-class classification applications suffer the class imbalance problem [3]. Class imbalance learning refers to a type of classification problems, where some classes are highly underrepresented compared to other classes. Several studies have shown that the class imbalance problem causes seriously negative effects on the classification performance [2,4,3], since the classifier algorithms are often biased towards the majority classes [5].

Much research has been done in addressing the class imbalance problem [4], and many new algorithms, methods and techniques have been presented [6]. Often the researchers use global measures to compare the classifier performance.

On two-class imbalance problems, global metrics to assessments of the classifier performance have been amply studied [7], and in the multi-class imbalance problems some of them have been modified to be adapted at the multi-class imbalance context

* This work has been partially supported under grants of: Projects 3072/2011 from the UAEM, PROMEP/103.5/12/4783 from the Mexican SEP and SDMAIA-010 of the TESJO.

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2013, LNCS 7914, pp. 335–343, 2013.

[8], for example the geometric mean, F-measure or measures of the area under curve family.

The use of global measures in order to evaluate the classifier performance on multi-class imbalance problem lead to some open interesting questions: a) the global measures can measure if the classifier performance over minority and majority classes is improved? b) Global measures can give us useful information about the true behavior of the classifier over minority and majority classes? c) How could we measure efficiently whether the archive classifiers allow a tradeoff between minority and majority classes performance? Or what measure can give us real information about of true behavior of the classifier?

In this paper, we presented a preliminary study about of some of the most often global metrics applied to evaluate the classifier performance over multi-class imbalance problems on the neural networks context. Our main contribution consists in to show the relationship between the results obtained by global metrics applied to evaluate the classifier performance and the classifier behavior over the minority classes, i.e., to determine if the results of global metrics match with the improve classifier performance over the minority classes.

2 Assessments Metrics for Multi-class Imbalance Learning

The most studied metrics for assessment the classifier performing in class imbalance domains have been focused a two class imbalance problems and some of them have been modified to accommodate them at the multi-class imbalanced learning problems [6]. In this section we present some of the most common two-class imbalance metrics adapted at multi-class imbalance scenarios.

Macro average geometric (MAvG): This is defined as the geometric average of the partial accuracy of each class.

$$MAvG = (\prod_{i=1}^{J} ACC_i)^{\frac{1}{J}}, \tag{1}$$

where $ACC_j = (\ correctly\ classified\ of\ class\ j)/(\ total\ of\ samples\ of\ class j)$, i.e., the accuracy on the class j. J is the number of classes.

Mean F-measure (MFM): This measure has been widely employed in information retrieval

$$F - \text{measure}(j) = \frac{2 \cdot recall(j) \cdot precision(j)}{recall(j) + precision(j)}, \tag{2}$$

where $recall(j) = (correctly\ classified\ positives)/(total\ positives)$ and $precision\ (j) = (correctly\ classified\ positives)/(total\ predicted\ as\ positives)$; j is the index of the class considered as $positive$. Finally, mean F-measure is defined for multi-class in Reference [8] as follow:

$$MFM = \sum_{j=1}^{J} \frac{FMeasure(j)}{J}. \tag{3}$$

Macro average arithmetic (MAvA): This is defined as the arithmetic average of the partial accuracies of each class.

$$MAvA = \frac{\sum_{i=1}^{J} ACC_i}{J} . \tag{4}$$

One the most widely used techniques for the evaluation of binary classifiers in imbalanced domains is the Receiver Operating Characteristic curve (ROC), which is a tool for visualizing, organizing and selecting classifiers based on their trade-offs between true positive rates and false positive rates. Furthermore, a quantitative representation of a ROC curve is the area under it, which is known as AUC [9]. The AUC measure has been adapted at multi-class problems [8] and can be defined as follow.

AUC of each class against each other, using the uniform class distribution (AU1U):

$$AU1U = \frac{2}{\|J\|(\|J\|-1)} \sum_{j_i, j_k \epsilon J} AUC_R(j_i, j_k) \ , \tag{5}$$

where $\text{AUC}_R(j_i, j_k)$ is the AUC for each pair of classes j_i and j_k.

AUC of each class against each other, using the a priori class distribution (AU1P):

$$AU1P = \frac{2}{\|J\|(\|J\|-1)} \sum_{j_i, j_k \epsilon J} p(j) AUC_R(j_i, j_k) \ , \tag{6}$$

where $p(j)$ is a priori class distribution.

AUC of each class against the rest, using the uniform class distribution (AUNU):

$$AUNU = \frac{1}{J} \sum_{j \epsilon J} AUC_R(j, rest_j) \ , \tag{7}$$

where $rest_j$ gathers together all classes different from class j, i.e., the area under the ROC curve is computed in the approach one against all.

AUC of each class against the rest, using the a priori class distribution (AUNP):

$$AUNP = \frac{1}{J} \sum_{j \epsilon J} p(j) AUC_R(j, rest_j) \ , \tag{8}$$

this measure takes into account the prior probability of each class $(p(j))$.

3 Experimental Protocols

3.1 Database Description

We have used in our experiments five remote sensing datasets: Cayo, Feltwell Satimage, Segment and 92AV3C. Feltwell is related to an agricultural region near Felt Ville, Feltwell (UK) [10], Cayo represents a particular region in the Gulf of Mexico, and Satimage consists of the multi-spectral values of pixels in 3x3 neighborhoods in a

satellital image [11]. The Segment contains instances drawn randomly from a dataset of seven outdoor images [11]. 92AV3C dataset[1] corresponds to a hyperspectral image (145x145 pixels) taken over Northwestern Indianas Indian Pines by the AVIRIS sensor.

In order to cover Cayo in a highly imbalanced dataset some of their classes were merged as follows: join classes 1, 3, 6, 7 and 10 to integrate class 1; join classes 8, 9 and 11 to integrate class 3, finally, the rest of classes (2, 4 and 5) were obtained from the original dataset. M92AV3C is a subset of 92AV3C, it contains six classes (2, 3, 4, 6, 7 and 8) and 38 attributes. The attributes were selected using a common features selection algorithm (Best-First Search [12]) implemented in WEKA[2].

Feltwell, Satimage, Segment and 92AV3C were random under-sampled with the goal of generating severe class imbalanced datasets. A brief summary of these multi-class imbalance datasets is shown in the Table 1. Note that they are highly imbalanced datasets. For each database, a $10-$fold cross$-$validation was applied. The datasets were divided into ten equal parts, using nine folds as training set and the remaining block as test set.

Table 1. A brief summary of some basic characteristics of the datasets. The bold numbers represent the samples of minority classes. Observe that in these datasets is very easy to identify the minority and majority classes.

Dataset	Size	Attr.	Class	Class distribution
MCayo	6019	4	5	2941/**293**/2283/**322/133**
MFelt	10944	15	5	3531/2441/**91**/2295/**178**
MSat	6430	36	6	1508/1531/**104**/1356/**93/101**
MSeg	1470	19	7	330/**50**/330/330/**50/50**/330
M92AV3C	5062	38	6	**190/117**/1434/2468/747/**106**

3.2 Resampling Methods

SMOTE and random under sampling (RUS) are used in the empirical study, because they are popular approaches to deal with the class imbalance problem. However, these methods have an internal parameter that enables the user to set up the resulting class distribution obtained after the application of these methods. In this paper, we decided to add or remove examples until a balanced distribution was reached. This decision was motivated by two reasons: a) simplicity and b) effectiveness. Results obtained with the other classifiers [13], have shown that when AUC is used as a performance measure, the best class distribution for learning tends to be near the balanced class distribution.

Another common re-sampling method is the Gabriel Graph Editing (GGE). The GGE consists of applying the general idea of Wilson's algorithm [14], but using the graph neighbors of each sample instead of either the Euclidean or the norm-based distance neighborhood. The original GGE was proposed to improve the k-NN accuracy [15]. However, in Reference [16] the original GGE was adapted to do it effective in the back-propagation context.

[1] `engineering.purdue.edu/biehl/MultiSpec/hyperspectral.html`
[2] `www.cs.waikato.ac.nz/ml/weka/`

3.3 Modified Back-Propagation Algorithm (MBP)

The most popular training procedure for the MLP neural network is the back-propagation algorithm, which uses a set of training instances for the learning process. Several works have shown that the class imbalance problem generates unequal contributions to the mean square error (MSE) during the training phase in the MLP, where the major contribution to the MSE is produced by the majority class.

Given a training data set with two classes ($J = 2$), such that the size is denoted by $N = \sum_j^J n_j$ and n_j is the number of samples that belong to class j, the MSE by class can be expressed as

$$E_j(U) = \frac{1}{N} \sum_{n=1}^{n_j} \sum_{p=1}^{J} (t_p^n - z_p^n)^2 \,, \tag{9}$$

where t_p^n is the desired output and z_p^n is the actual output of the network for the sample n.

Then the overall MSE can be written in terms of $E_j(U)$ as follows:

$$E(U) = \sum_{j=1}^{J} E_j(U) = E_1(U) + E_2(U) \,. \tag{10}$$

When $n_1 << n_2$, the $E_1(U) << E_2(U)$ and $\|\nabla E_1(U)\| << \|\nabla E_2(U)\|$. Consequently, $\nabla E(U) \approx \nabla E_2(U)$. So, $-\nabla E(U)$ is not always the best direction to minimize the MSE in both classes.

The unequal contribution to the MSE can be compensated by introducing a cost function (γ) in order to avoid that the MLP ignores the minority class:

$$\begin{aligned} E(U) &= \textstyle\sum_{j=1}^{J} \gamma(j) E_j = \gamma(1) E_1(U) + \gamma(2) E_2(U) \\ &= \tfrac{1}{N} \textstyle\sum_{j=1}^{J} \gamma(j) \sum_{i=1}^{n_j} \sum_{p=1}^{J} (t_p^i - z_p^i)^2 \,, \end{aligned} \tag{11}$$

where $\gamma(1)\|\nabla E_1(U)\| \approx \gamma(2)\|\nabla E_2(U)\|$.

In this work, we define a cost function γ as $\gamma(j) = \|\nabla E_{max}(U)\| / \|\nabla E_j(U)\|$, where $\|\nabla E_{max}(U)\|$ corresponds to the majority class.

3.4 Neural Network Configuration

The MLP was trained with the standard back-propagation (SBP) and modified back-propagation (MBP) algorithm in batch mode. For each training data set, MLP was initialized ten times with different weights, i.e., the MLP was run ten times with the same training dataset. The results here included correspond to the average of those accomplished in the ten different initialization and of ten partitions. The learning rate (η) was set at 0.1 and only one hidden layer was used. The stop criterion was established at 25000 epoch or an MSE below to 0.001. The number of neurons for the hidden layer was obtained from the trial and error strategy. So, the number of neurons was 7, 6, 12, 10 and 10, for MCayo, MFelt, MSat, MSeg and M92AV3C datasets respectively.

4 Experimental Results

In order to assessment the multi-class imbalance measures (sec. 2) we have carried out an experimental comparison over five popular strategies to deal with class imbalance problem: (*i*) Modified Back-Propagation Algorithm (MBP), (*ii*) Standard Back-Propagation with Gabriel Graph Editing (SBP+GGE), (*iii*) Modified Back-Propagation with Gabriel Graph Editing (MBP + GGE), (*iv*) SMOTE and (*vi*) Random Under Sampling (RUS).

The Standard Back-Propagation Algorithm (SBP) is presented as baseline method. The datasets that were preprocessed by the SMOTE and RUS strategies were applied to the SBP algorithm. In addition, we have used seven global measures: $MAvG$, MFM, $MAvA$, $AU1U$, $AU1P$, $AUNU$ and $AUNP$ (see sec. 2) to evaluate the global classifier performance.

In the Table 2 the experimental results are presented. The columns represent a strategies used to deal with class imbalance problem, the rows show values obtained for different measures. We use the average accuracy for majority and minority classes (acc^- and acc^+ respectively), because they can give us useful information about of the classifier performance.

The values between parentheses are the average rank (AR), they provide a useful tool to compare algorithms [17]. In the AR the best performing algorithm should have the rank of 1, the second best rank 2, the third best rank 3, etc. The values of AR in Table 2 might seem wrong but not so, consider that they are average of ten folds and ten run of the MLP for each dataset (see sec. 3).

AR-GM represents the average rank for the seven global measures ($MAvG$, MFM, $MAvA$, $AU1U$, $AU1P$, $AUNU$ and $AUNP$), i.e., it is the average of the ranks of the seven global measures. We make this because the global measures tend to give similar ranks. For example see M92AV3C with RUS strategy in Table 2. With other datasets and strategies it is not as clear as with the last one. However, in general terms the global measures show similar ranks. This is consistent with the presented in Reference [8].

Table 2 shows some interesting results. For example in M92AV3C the best AR-GM is obtained by RUS and the best AR for the partial measure acc^+ is presented by RUS too, i.e., the best ranks at classes level and at global level are showed for the same strategy (RUS). In contrast to M92AV3C, MSegment shows that the best AR-GM and the worst AR are obtained for MBP, i.e., the best ranks at classes level and at global level are showed for different strategies. This imply that global measures to indicate a good classifier performance but a classes level do not performance well. This might address at wrong conclusions.

In the rest of datasets MCayo, MFeltwell, and MSatimage, the behavior of the global and partial measures follow the same tendency (see Table 2), i.e., the best AR-GM and AR for the $acc+$ are not presented by the same strategy. The results presented in Table 2 pose some interesting facts: MSegment presents a best AR-GM but the worst performance at minority class's level. Some strategies improve the minority classes perform but damage the majority classes' performance and the global measures not always reflect this situation (for example, see the MBP+GGE results).

The global measures have been introduced for very different applications and, supposedly, measure quite different things [8] but in many researches they are applied to

Table 2. Performance on five datasets measured using $MAvG$, $MAvA$, $AU1U$, $AU1P$, $AUNU$, $AUNP$ and average rank (AR)

	Measure	GGE[1]	MBP+GGE	Imbalanced[1]	MBP	RUS[1]	SMOTE[1]
	acc^-	0.918 (2.8)	0.850 (5.0)	0.924 (1.8)	0.911 (3.5)	0.899 (3.0)	0.699 (3.0)
	acc^+	0.448 (5.0)	0.814 (2.0)	0.179 (6.0)	0.585 (4.0)	0.605 (3.0)	0.968 (1.0)
	$MAvG$	0.484 (5.0)	0.820 (2.0)	0.000 (6.0)	0.692 (4.0)	0.702 (3.0)	0.822 (1.0)
	MFM	0.466 (4.0)	0.503 (1.0)	0.386 (60)	0.501 (2.0)	0.491 (3.0)	0.393 (5.0)
	$MAvA$	0.636 (5.0)	0.828 (2.0)	0.477 (6.0)	0.715 (4.0)	0.722 (3.0)	0.860 (1.0)
MCayo	$AU1U$	0.636 (5.0)	0.828 (2.0)	0.477 (6.0)	0.715 (4.0)	0.722 (3.0)	0.860 (1.0)
	$AU1P$	0.840 (4.0)	0.841 (3.0)	0.804 (5.0)	0.853 (1.0)	0.844 (2.0)	0.738 (6.0)
	$AUNU$	0.743 (5.0)	0.833 (1.0)	0.649 (6.0)	0.788 (3.0)	0.786 (4.0)	0.790 (2.0)
	$AUNP$	0.840 (4.0)	0.841 (3.0)	0.804 (5.0)	0.853 (1.0)	0.844 (2.0)	0.738 (6.0)
	AR-GM	**4.6**	**2**	**5.7**	**2.7**	**2.8**	**3.1**
	Measure	GGE[1]	MBP+GGE	Imbalanced[1]	MBP	RUS[1]	SMOTE[1]
	acc^-	0.964 (2.7)	0.926 (5.3)	0.970 (2.0)	0.954 (2.8)	0.965 (2.2)	0.908 (5.0)
	acc^+	0.305 (5.2)	0.811 (2.0)	0.188 (5.8)	0.668 (3.0)	0.424 (4.0)	0.876 (1.0)
	$MAvG$	0.000 (5.5)	0.875 (2.0)	0.000 (5.5)	0.823 (3.0)	0.530 (4.0)	0.890 (1.0)
	MFM	0.614 (5.0)	0.631 (3.0)	0.598 (6.0)	0.661 (1.0)	0.641 (2.0)	0.617 (4.0)
	$MAvA$	0.700 (5.0)	0.880 (2.0)	0.658 (6.0)	0.839 (3.0)	0.749 (4.0)	0.895 (1.0)
MFeltwel	$AU1U$	0.700 (5.0)	0.880 (2.0)	0.658 (6.0)	0.839 (3.0)	0.749 (4.0)	0.895 (1.0)
	$AU1P$	0.942 (3.5)	0.923 (4.0)	0.942 (3.5)	0.944 (2.0)	0.947 (1.0)	0.909 (5.0)
	$AUNU$	0.822 (5.0)	0.902 (2.0)	0.801 (6.0)	0.892 (3.0)	0.849 (4.0)	0.903 (1.0)
	$AUNP$	0.942 (3.5)	0.923 (4.0)	0.942 (3.5)	0.944 (2.0)	0.947 (1.0)	0.909 (5.0)
	AR-GM	**4.6**	**2.7**	**5.2**	**2.4**	**2.9**	**2.6**
	Measure	GGE[1]	MBP+GGE	Imbalanced[1]	MBP	RUS[1]	SMOTE[1]
	acc^-	0.699 (5.0)	0.649 (5.3)	0.956 (1.5)	0.954 (1.8)	0.954 (2.3)	0.885 (3.7)
	acc^+	0.849 (2.0)	0.866 (1.2)	0.369 (4.8)	0.550 (4.3)	0.498 (4.8)	0.768 (2.5)
	$MAvG$	0.739 (2.0)	0.723 (3.0)	0.000 (5.5)	0.494 (4.0)	0.000 (5.5)	0.801 (1.0)
	MFM	0.346 (4.0)	0.317 (5.0)	0.512 (3.5)	0.563 (1.0)	0.546 (2.0)	0.512 (3.5)
	$MAvA$	0.774 (2.0)	0.757 (3.0)	0.663 (6.0)	0.752 (4.0)	0.726 (5.0)	0.826 (1.0)
MSatimage	$AU1U$	0.774 (2.0)	0.757 (3.0)	0.663 (6.0)	0.752 (4.0)	0.726 (5.0)	0.826 (1.0)
	$AU1P$	0.709 (5.0)	0.664 (6.0)	0.912 (3.0)	0.924 (1.0)	0.920 (2.0)	0.875 (4.0)
	$AUNU$	0.741 (5.0)	0.709 (6.0)	0.790 (4.0)	0.839 (2.0)	0.825 (3.0)	0.851 (1.0)
	$AUNP$	0.709 (5.0)	0.664 (6.0)	0.912 (3.0)	0.924 (1.0)	0.920 (2.0)	0.875 (4.0)
	AR-GM	**3.6**	**4.6**	**4.4**	**2.4**	**3.5**	**2.2**
	Measure	GGE[1]	MBP+GGE	Imbalanced[1]	MBP	RUS[1]	SMOTE[1]
	acc^-	0.905 (3.2)	0.890 (5.0)	0.973 (1.4)	0.961 (2.9)	0.921 (3.4)	0.914 (3.9)
	acc^+	0.905 (2.0)	0.958 (2.7)	0.736 (4.3)	0.857 (5.0)	0.905 (2.7)	0.835 (3.3)
	$MAvG$	0.892 (4.0)	0.913 (1.0)	0.666 (6.0)	0.901 (3.0)	0.904 (2.0)	0.788 (5.0)
	MFM	0.625 (6.0)	0.630 (4.0)	0.706 (2.0)	0.764 (1.0)	0.656 (3.0)	0.627 (5.0)
	$MAvA$	0.905 (4.0)	0.919 (1.0)	0.871 (6.0)	0.917 (2.0)	0.914 (3.0)	0.880 (5.0)
MSegment	$AU1U$	0.905 (4.0)	0.919 (1.0)	0.871 (6.0)	0.917 (2.0)	0.914 (3.0)	0.880 (5.0)
	$AU1P$	0.906 (5.0)	0.900 (6.0)	0.947 (2.0)	0.951 (1.0)	0.920 (3.0)	0.907 (4.0)
	$AUNU$	0.906 (4.0)	0.910 (3.5)	0.910 (3.5)	0.935 (1.0)	0.917 (2.0)	0.894 (5.0)
	$AUNP$	0.906 (5.0)	0.900 (6.0)	0.947 (2.0)	0.951 (1.0)	0.920 (3.0)	0.907 (4.0)
	AR-GM	**4.6**	**3.2**	**3.9**	**1.6**	**2.7**	**4.7**
	Measure	GGE[1]	MBP+GGE	Imbalanced[1]	MBP	RUS[1]	SMOTE[1]
	acc^-	0.764 (2.0)	0.723 (3.3)	0.773 (2.3)	0.688 (4.7)	0.732 (4.0)	0.634 (4.7)
	acc^+	0.467 (4.3)	0.837 (2.0)	0.252 (6.0)	0.493 (4.7)	0.860 (1.7)	0.773 (2.3)
	$MAvG$	0.322 (5.0)	0.733 (2.0)	0.000 (6.0)	0.494 (4.0)	0.778 (1.0)	0.555 (3.0)
	MFM	0.340 (3.0)	0.356 (2.0)	0.316 (4.0)	0.285 (5.0)	0.363 (1.0)	0.274 (6.0)
	$MAvA$	0.615 (4.0)	0.780 (2.0)	0.512 (6.0)	0.590 (5.0)	0.796 (1.0)	0.703 (3.0)
M92AV3C	$AU1U$	0.615 (4.0)	0.780 (2.0)	0.512 (6.0)	0.590 (5.0)	0.796 (1.0)	0.703 (3.0)
	$AU1P$	0.720 (2.0)	0.718 (3.0)	0.702 (4.0)	0.653 (5.0)	0.726 (1.0)	0.599 (6.0)
	$AUNU$	0.677 (3.0)	0.754 (2.0)	0.615 (6.0)	0.630 (5.0)	0.759 (1.0)	0.649 (4.0)
	$AUNP$	0.720 (2.0)	0.718 (3.0)	0.702 (4.0)	0.653 (5.0)	0.726 (1.0)	0.599 (6.0)
	AR-GM	**3.3**	**2.3**	**5.1**	**4.9**	**1**	**4.4**

[1] Classification using SBP.

compare the effects of new algorithms over the classifier performance. We think it is not enough to evaluate the global performance whether it is not necessary to consider the classifier performance at classes' level too. Table 2 shows that not always the classifier performance over minority or majority classes are favorable when the global measures present a good performance of the classifier.

5 Conclusion

In this paper we study some commons global metrics ($MAvG$, $MAvA$, $AU1U$, $AU1P$, $AUNU$ and $AUNP$) used to assessment classifier performance over multiclass imbalance datasets. So we employ average accuracy of minority and majority classes (partial metrics) to contrast the results, and average rank to facility the algorithms comparison. The classifier used was a multilayer perceptron trained with the back-propagation algorithm. The study was made over five multi-class imbalanced datasets and five popular strategies to deal with the class imbalance problem.

The results obtained from five datasets used in this paper, show that not always is enough to use the global metrics to compare algorithms (over these datasets), because we observed that by one hand in some datasets these global metrics show well results and at the same time the partial metrics exhibit a bad classifier performance over minority classes. This implies that global metrics indicates a good classifier performance (on some of the datasets used in this work), but in some classes the classifier does not perform well and this might to address us at wrong conclusions. On the other hand, in some other datasets the global and partial metrics present good results, i.e., global and partial metrics agree in results. These differences in their results suggest that global metrics no always reflect the improving or damage of the strategies applied to deal with the class imbalance problem on the classifier performance over minority classes, so it is necessary to study other alternatives to assessment classifier performance over multi-class imbalance datasets.

Future work will be primarily addressed to get in depth in this topic. Priority is expand the research using more datasets and applies a significance statistical test to give better confidence to the conclusions. Also it is necessary the study of new metrics which help to assessment classifier performance over multi-class imbalance datasets and that they reflect the changes caused for the strategies used to deal with class imbalance problem on the classifier performance over the majority and minority classes.

References

1. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. Int. J. Data Warehousing and Mining, 1–13 (2007)
2. Ou, G., Murphey, Y.L.: Multi-class pattern classification using neural networks. Pattern Recognition 40(1), 4–18 (2007)
3. Wang, S., Yao, X.: Multi-class imbalance problems: Analysis and potential solutions. IEEE Transactions on IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics (99), 1–12 (2012)

4. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering 18, 63–77 (2006)
5. Pérez-Godoy, M.D., Fernández, A., Rivera, A.J., del Jesus, M.J.: Analysis of an evolutionary rbfn design algorithm, co2rbfn, for imbalanced data sets. Pattern Recogn. Lett. 31(15), 2375–2388 (2010)
6. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009)
7. García, V., Mollineda, R.A., Sánchez, J.S.: Theoretical analysis of a performance measure for imbalanced data. In: ICPR, pp. 617–620 (2010)
8. Ferri, C., Hernández-Orallo, J., Modroiu, R.: An experimental comparison of performance measures for classification. Pattern Recognition Letter 30(1), 27–38 (2009)
9. Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. 27, 861–874 (2006)
10. Bruzzone, L., Serpico, S.: Classification of imbalanced remote-sensing data by neural networks. Pattern Recognition Letters 18, 1323–1328 (1997)
11. A. Asuncion, D.N.: UCI machine learning repository (2007)
12. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. 97(1-2), 273–324 (1997)
13. Weiss, G.M., Provost, F.J.: Learning when training data are costly: The effect of class distribution on tree induction. J. Artif. Intell. Res. (JAIR) 19, 315–354 (2003)
14. Wilson, D.: Asymptotic properties of nearest neighbor rules using edited data. IEEE Transactions on Systems, Man and Cybernetics 2(4), 408–420 (1972)
15. Sánchez, J.S., Pla, F., Ferri, F.J.: Prototype selection for the nearest neighbour rule through proximity graphs. Pattern Recognition Letters 18(6), 507–513 (1997)
16. Alejo, R., Valdovinos, R., García, V., Pacheco-Sanchez, J.: A hybrid method to face class overlap and class imbalance on neural networks and multi-class scenarios. Pattern Recognition Letters 34(4), 380–388 (2012)
17. García, S., Herrera, F.: Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. Evolutionary Computation 17, 275–306 (2009)