ON REPORTING PERFORMANCE OF BINARY CLASSIFIERS

Pavel Škrabánek, Petr Doležel

Abstract: *In this contribution, the question of reporting performance of binary classifiers is opened in context of the so called class imbalance problem. The class imbalance problem arises when a dataset with a highly imbalanced class distribution is used within the training or evaluation process. In such cases, only measures, which are not biased by distribution of classes in datasets, should be used; however, they cannot be chosen arbitrarily. They should be selected so that their outcomes provide desired information; and simultaneously, they should allow a full comparison of just evaluated classifier performance along, with performances of other solutions. As is shown in this article, the dilemma with reporting performance of binary classifiers can be solved using so called class balanced measures. The class balanced measures are generally applicable means, appropriate for reporting performance of binary classifiers on balanced as well as on imbalanced datasets. On the basis of the presented pieces of information, a suggestion for a generally applicable, fully-valued, reporting of binary classifiers performance is given.*

Keywords: *machine learning, binary classification, class imbalance problem, performance measures, reporting of results.*

JEL Classification: *C45, C83.*

Introduction

In general, classification is a process in which objects, either real or abstract, are recognized, differentiated, and understood. This issue is important in various fields such as by text classification (Nigam, et al., 2000), in medicine (Goeuriot et al., 2016), or within visual information retrieval (Lew, 2001). Naturally, economics is no exception. The relevance of classification in this field is obvious e.g. in applications related to machine learning (Qiao, et al., 2016) or data mining (Feelders, 2002).

From the perspective of machine learning or data mining, classification is the task of classifying elements of a given set into a predetermined number of groups, usually called **classes**. Although the number of classes might be any positive integer, just two classes are considered in many real-world applications. Such classification tasks are known as **binary classification**. Considering the status of the binary classification, let us focus on this issue in this contribution.

The process of the classification is carried out by a **classifier**. The classifier assigns each element into one of the considered classes. The assignment is accomplished on the basis of a **classification rule**. The classification rule is formed both by selection of a **classification method** and on the basis of **training data**. To this day, a number of supervised classification methods have been introduced, such as *k*-nearest neighbour, neural networks, support vector machine, random forest, and many others (Murphy, 2012). Selection of an appropriate classification method is carried out by an expert. Such a formed classifier is then trained on the training data.

The previous description of classifier development might give the wrong notion that design of classifiers is a simple task. The opposite is true. Depending on the used classification method, several parameters are usually needed to be set, in order to obtain a classifier of a desirable performance. However, the performance of the classifier is also affected by other factors, such as generality of the training data, or appropriateness of the selected methods for a particular task. It is obvious that information about performance of the classifier is highly desirable in order to fulfil the practical need to perform comparisons across various classifiers, settings and datasets.

Information about performance of a classifier can be acquired using a **performance measure**. In the case of binary classification, a variety of performance measures have been introduced, e.g. (Brodersen et al., 2010), (Garcia et al., 2010), (Hand, 2012), but not all of them are used in practice. To be honest, it is not so difficult to design a new performance measure; however, a successful measure needs to satisfy three basic criteria:

- it must coherently capture the aspect of performance of interest;
- it must be intuitive enough to become widely used, so that the same measure is consistently reported by a majority of researchers;
- it must be simple to report, preferably as a single number, for each method-setting-dataset combination.

As already implied, even if a measure meets all the stated out requirements, its universal acceptance is not guaranteed. It might be pointed here that different application areas have different preferences for measures due to different goals. Over the time, sets of measures, preferred within each particular area, have been naturally formed. A new measure, which might be succeeding in a particular area, must naturally fit to the appropriate set of widely accepted measures. In other words, outputs of such a measure should enable a comparison with other published results. This basic requirement will be further called **comparability requirement**. However, there are also other aspects influencing the probability of acceptance of a new measure in an application area. One important aspect is behaviour of the measure on data with a highly imbalanced class distribution.

Data with a highly imbalanced class distribution are said to suffer a **class imbalance problem**. Since class imbalanced datasets occur in many real-world applications, the class imbalance problem is a hot issue. This is evidenced by the long list of publications dealing with this topic. A large proportion of them deal with training of classifiers on imbalanced data; however, the class imbalance problem can also adversely affect the evaluation of classifiers.

The adverse influence of class distribution on some performance measures has been known for a long time and some works dealing with the evaluation on imbalanced data have been already published. Nevertheless, none of them brings an answer to one fundamental question: Which measures should be chosen so that a desired information value would be kept and the comparability requirement would be met? This question is opened and analysed in context of binary classifiers in this contribution. On the basis of the analysis, a suggestion for a fully-valued reporting performance of binary classifiers, which reflects all the above stated facts, is given. For this purpose, so called **class balanced measures** are presented as the appropriate means.

The rest of the article is organized in the following way. The class imbalance problem, and its impact on classification tasks, is considered in section 1. Basic variables used by evaluation of binary classifiers, and the most popular performance measures, are stated in section 2. The influence of imbalanced data on the performance measures is analysed in section 3. The class balanced measures are introduced in section 4. The opened question of reporting performance of binary classifiers is discussed in section 5. Finally, a conclusion is stated in section 6.

1 Impact of the class imbalance problem on classification related tasks

It is well known that a classifier trained on imbalanced data might be biased in favour of a major class. This issue has been widely studied and many related works have been published. A short summary about this issue is given in subsection 1.1. However, the class imbalance problem may become apparent also within an evaluation process. This issue is discussed in subsection 1.2.

1.1 Training on imbalanced data

As pointed out by (Garcia et al., 2007), two groups of approaches can be used to handle a class imbalance in data, by training of binary classifiers. Namely, a re-sampling method can be used, or measuring of a classifier's performance in imbalanced domains can be utilized within a classification method.

Generally, the re-sampling methods aim to form balanced datasets. Many different approaches belonging to this group have been presented, such as, random or focused over-sampling (Japkowicz et al., 2002); over-sampling with informed generation of new samples (Chawla at al., 2002); random under-sampling (Kotsiantis et al., 2003); or direct under-sampling (Mani et al, 2003).

The second group of approaches is aimed to deal with imbalanced datasets directly. At first, it might be pointed out that some performance measures are not influenced by the distribution of classes in datasets. (Garcia et al., 2007) pointed to this fact and they have logically inferred that these measures can be safely used on imbalanced data. It is worth mentioning that new measures, which are resistant to imbalances in data, are constantly developed (Huang et al., 2007), (Brodersen et al., 2010), (Garcia et al., 2010), (Koyejo et al., 2014). The essence of such measures was the inspiration of many classification methods which are designed for direct application on imbalanced datasets, e.g. (Barandela et al., 2003), (Rosenberg, 2012), (Koyejo et al., 2014).

1.2 Evaluation on imbalanced data

Although composition of datasets affects outcomes of some performance measures (Daskalaki et al., 2006), these measures are widely used due to their information value and comprehensibility (Hand, 2012). In order to keep the comparability, researchers usually report about performances of classifiers on datasets with nearly uniform distribution of classes. However, achievement of this precondition may not be always possible or advisable, such in the case of fraud detection (Phua et al., 2004), mining data streams (Zhao et al., 2012), or object detection (Škrabánek et al., 2016).

As has been already mentioned, there are a number of performance measures resistant to distribution of classes in datasets. These measures can be safely used on

imbalanced datasets. This fact is commonly used when facing the class imbalance problem within the evaluation process (Daskalaki et al., 2006), (Jeni et al., 2013), but this is not the adequate solution in each situation.

In summary, despite general awareness about this issue, one fundamental question has not yet been opened nor answered. The question is, which measures should be chosen so that a desired information value would be kept and the comparability requirement would be met? Search for a generally valid answer to this question is the scope of interest in this article.

2 Standard performance measure used by evaluation of binary classifiers

Two classes, **positive** or simply *P*, and **negative** or simply *N*, are considered by the binary classification. The aim of a classifier is to correctly assign a class label to each judged sample. For each sample, the decision-making process falls into one of four possible scenarios: the sample is positive and the classifier correctly recognizes it as such (**true positive** or simply *TP*); the sample is negative and the classifier correctly recognizes it as such (**true negative** or simply *TN*); the sample is positive but the classifier labels it as negative (**false negative** or simply *FN*); or the sample is negative but the classifier labels it as positive (**false positive** or simply *FP*).

On the basis of the presented scenarios, four fundamental quantities for performance measure are formulized: number of true positive $|TP|$; number of true negative $|TN|$; number of false negative $|FN|$; and number of false positive $|FP|$ samples. The quantities are usually summarized into a 2×2 matrix. The matrix is known as **confusion matrix** and it is traditionally expressed as in Tab. 1.

Tab. 1: The confusion matrix

		Assigned label	
		positive	negative
True label	positive	$\|TP\|$	$\|FN\|$
	negative	$\|FP\|$	$\|TN\|$

Source: Authors

A number of performance measures derived from the confusion matrix have been introduced up to the present (Choi et al., 2010). However, not all of them have been widely accepted. Moreover, different measures are preferred in various scientific fields. Thus, only the most frequently used measures are considered further. Namely, the following measures are considered: accuracy (acc), error rate (er), precision (pr), recall (re), specificity (sp), false negative rate (fnr), false positive rate (fpr), harmonic mean of precision and recall (Fscore), geometric mean of precision and recall (Gmean), and area under the ROC curve (AUC).

3 Influence of imbalanced data on performance measures

As was already mentioned, the class imbalance problem is caused by highly imbalanced distribution of classes in datasets. The unfavourable properties of some performance measures on imbalanced data are well known; however, the core of this issue is not visible at first glance. In this section, the relation between the measures

and proportions of the classes in datasets is analysed. On the basis of the analysis, all the considered measures are expressed in the terms used within the analysis.

3.1 Preliminary

Let us consider a dataset of M labelled samples where each sample belongs either to the class P or N then

$$M = |P| + |N|, \tag{1}$$

where $|P|$ is number of positive samples, and $|N|$ is number of negative samples in the dataset. Supposing the confusion matrix stated in Tab. 1, the numbers of samples belonging to the classes can be expressed as

$$|P| = |TP| + |FN|, \quad |N| = |TN| + |FP|, \tag{2}$$

which allow us to express (1) as

$$M = |TP| + |FN| + |TN| + |FP|. \tag{3}$$

Let us express the numbers of samples in the classes as

$$|P| = \nu_P M, \quad |N| = \nu_N M, \tag{4}$$

where ν_P is the proportion of the positive samples in the dataset, and ν_N is the proportion of the negative samples in the dataset. Furthermore, it holds that $\nu_P, \nu_N \in [0,1]$ and $\nu_P + \nu_N = 1$.

3.2 Analysis

Let us consider the objective of a binary classifier now. As was already stated, the aim of a binary classifier is to correctly assign a sample to one of the two classes, P or N, if possible. A well working classifier will correctly assign all the samples, i.e. $|TP| = |P|$, $|TN| = |N|$, $|FP| = 0$, and $|FN| = 0$. A classifier with a worse performance will correctly classify a smaller proportion of the samples. Thus, let us express the number of correctly classified samples as

$$|TP| = \xi_{TP} |P|, \quad |TN| = \xi_{TN} |N|, \tag{5}$$

where ξ_{TP} is the proportion of correctly classified samples from all positive samples in the dataset, ξ_{TN} is the proportion of correctly classified samples from all negative samples in the dataset, and $\xi_{TP}, \xi_{TN} \in [0,1]$.

On the basis of formulae (2) and (5), the numbers of miss-classified samples can be expressed as

$$|FN| = (1 - \xi_{TP})|P|, \quad |FP| = (1 - \xi_{TN})|N|. \tag{6}$$

It is obvious that performance of a binary classifier can be positively determined using just two quantities, ξ_{TP} and ξ_{TN}.

Let us express all the performance measures using the quantities ξ_{TP} and ξ_{TN}. The modification will be demonstrated on the accuracy. The accuracy is given by

$$\text{acc} = \frac{|TP| + |TN|}{|TP| + |FN| + |TN| + |FP|}. \tag{7}$$

Using formulae (3) and (5), the original formulation (7) can be expressed as

$$\text{acc} = \frac{\xi_{TP}|P| + \xi_{TN}|N|}{M}. \tag{8}$$

This formula can be further modified using (4), i.e.

$$\text{acc} = \frac{\xi_{TP}\nu_P M + \xi_{TN}\nu_N M}{M} = \xi_{TP}\nu_P + \xi_{TN}\nu_N. \tag{9}$$

Formula (9) clearly shows that the accuracy (7) does not depend only on the performance of a classifier (ξ_{TP} and ξ_{TN}); however, composition of dataset is also reflected in this measure (ν_P and ν_N). The same procedure, which has been used for the accuracy, can be applied on other measures. The most common measures are summarized in Tab. 2 (AUC is expressed for a threshold value 0.5). Acronyms of the measures are stated in the first column. Their usual expressions are listed in the second one. The last column contains their modified expressions. It is apparent from the modified expressions that accuracy, error rate, precision and Fscore are biased by the class distribution in datasets while the other measures are invariant.

4 Class balanced performance measures

The previous analysis has clearly indicated the biased measures in Tab. 2, as well as the underlying problem. Simultaneously, a way of dealing with the problem has been outlined by the analysis. Specifically, the biased measures can be extended by **class weights**. Once the weights are properly set, measures resistant to the class distribution in a dataset are acquired. Since distributions of classes in datasets are known within the evaluation process, the weights can be set according to the proportion of the classes ν_P and ν_N. This idea was used when developing the class balanced measures. For simplicity, let us call them balanced measures, but do not confuse them with already published metrics such as a balanced accuracy (Brodersen et al., 2010) or a balanced error rate (Chi-Yuan, 2011).

In the case of the balanced measures, the setting of the weights was based on common practice. As already mentioned, it is usual to report performance of classifiers on datasets with nearly symmetrical prior probabilities of classes, i.e. magnitude of the classes in the biased measures is nearly uniform. Thus, the magnitude of classes in the balanced measures should be also uniform in order to get comparable results. It means that the basic quantities related to the positive class, $|TP|$ and $|FN|$, have to be multiplied by the proportion of the negative class in the dataset ν_N; and similarly, the basic quantities related to the negative class, $|TN|$ and $|FP|$, have to be multiplied by the proportion of the positive class ν_P. Following this idea, a class balanced complement can be developed for each biased measure.

Tab. 2: The most popular performance measures in the binary classification

Acronym	Standard expression	Modified expression
acc	$\frac{\lvert TP\rvert+\lvert TN\rvert}{\lvert TP\rvert+\lvert FN\rvert+\lvert TN\rvert+\lvert FP\rvert}$	$\xi_{TP}\nu_P+\xi_{TN}\nu_N$
er	$\frac{\lvert FP\rvert+\lvert FN\rvert}{\lvert TP\rvert+\lvert FN\rvert+\lvert TN\rvert+\lvert FP\rvert}$	$(1-\xi_{TP})\nu_P+(1-\xi_{TN})\nu_N$
pr	$\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FP\rvert}$	$\frac{\xi_{TP}\nu_P}{\xi_{TP}\nu_P+(1-\xi_{TN})\nu_N}$
re	$\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FN\rvert}$	ξ_{TP}
sp	$\frac{\lvert TN\rvert}{\lvert TN\rvert+\lvert FP\rvert}$	ξ_{TN}
fnr	$\frac{\lvert FN\rvert}{\lvert TP\rvert+\lvert FN\rvert}$	$1-\xi_{TP}$
fpr	$\frac{\lvert FP\rvert}{\lvert TN\rvert+\lvert FP\rvert}$	$1-\xi_{TN}$
Fscore	$\frac{(\beta^2+1)\lvert TP\rvert}{(\beta^2+1)\lvert TP\rvert+\beta^2\lvert FN\rvert+\lvert FP\rvert}$	$\frac{(\beta^2+1)\xi_{TP}\nu_P}{\left[(\beta^2+1)\xi_{TP}+\beta^2(1-\xi_{TP})\right]\nu_P+(1-\xi_{TN})\nu_N}$
Gmean	$\sqrt{\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FN\rvert}\times\frac{\lvert TN\rvert}{\lvert TN\rvert+\lvert FP\rvert}}$	$\sqrt{\xi_{TP}\times\xi_{TN}}$
AUC	$\frac{1}{2}\left(\frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FN\rvert}+\frac{\lvert TN\rvert}{\lvert TN\rvert+\lvert FP\rvert}\right)$	$\frac{1}{2}(\xi_{TP}+\xi_{TN})$

Source: Authors

In such a way, balanced accuracy (acc_B), balanced error rate (er_B), balanced precision (pr_B), and balanced harmonic mean of precision and recall ($Fscore_B$) have been established. All the class balanced complements of the biased measures in Tab. 2 are summarized in Tab. 3. Their acronyms are stated in the first column. Their usual expressions are listed in the second one while the modified expressions of these measures are stated in the last column.

Tab. 3: The class balanced performance measures

Acronym	Standard expression	Modified expression
acc_B	$\frac{\nu_N\lvert TP\rvert+\nu_P\lvert TN\rvert}{\nu_N(\lvert TP\rvert+\lvert FN\rvert)+\nu_P(\lvert TN\rvert+\lvert FP\rvert)}$	$\frac{1}{2}(\xi_{TP}+\xi_{TN})$
er_B	$\frac{\nu_P\lvert FP\rvert+\nu_N\lvert FN\rvert}{\nu_N(\lvert TP\rvert+\lvert FN\rvert)+\nu_P(\lvert TN\rvert+\lvert FP\rvert)}$	$\frac{1}{2}\left[(1-\xi_{TP})+(1-\xi_{TN})\right]$
pr_B	$\frac{\nu_N\lvert TP\rvert}{\nu_N\lvert TP\rvert+\nu_P\lvert FP\rvert}$	$\frac{\xi_{TP}}{\xi_{TP}+1-\xi_{TN}}$
$Fscore_B$	$\frac{(\beta^2+1)\nu_N\lvert TP\rvert}{\nu_N\left[(\beta^2+1)\lvert TP\rvert+\beta^2\lvert FN\rvert\right]+\nu_P\lvert FP\rvert}$	$\frac{(\beta^2+1)\xi_{TP}}{(\beta^2+1)\xi_{TP}+\beta^2(1-\xi_{TP})+(1-\xi_{TN})}$

Source: Authors

As follows from the previous text, the class balanced measures should provide results corresponding to evaluation on balanced datasets. Fulfilment of this requirement can be simply verified on benchmark datasets or analytically. The analytical approach takes into account the uniform distribution of classes in balanced datasets ($\nu_P = \nu_N = 0.5$). Inserting these values into the modified expressions of original measures, leads to the class balanced measures which confirms the above stated requirement (third columns of Tab. 2 and Tab. 3).

The verification using benchmark datasets consists of a systematic application of original and balanced measures in different scenarios, followed by a comparison of the results. The results obtained using the original measures on a balanced dataset are taken as the reference values. However, this approach leads to a huge amount of data which could not be summarized in this article due to limited space.

In order to offer an alternative verification approach, we developed a specialized visualization method. It is based on the idea to use gradients of a performance measure m in order to show its dependence on the real performance of a classifier (ξ_{TP}, ξ_{TN}) and the proportion of samples in datasets. Just as a reminder, the composition of a dataset can be expressed either as the proportion of positive ν_P or negative ν_N samples, where $\nu_P + \nu_N = 1$. In our approach, the gradient is defined as

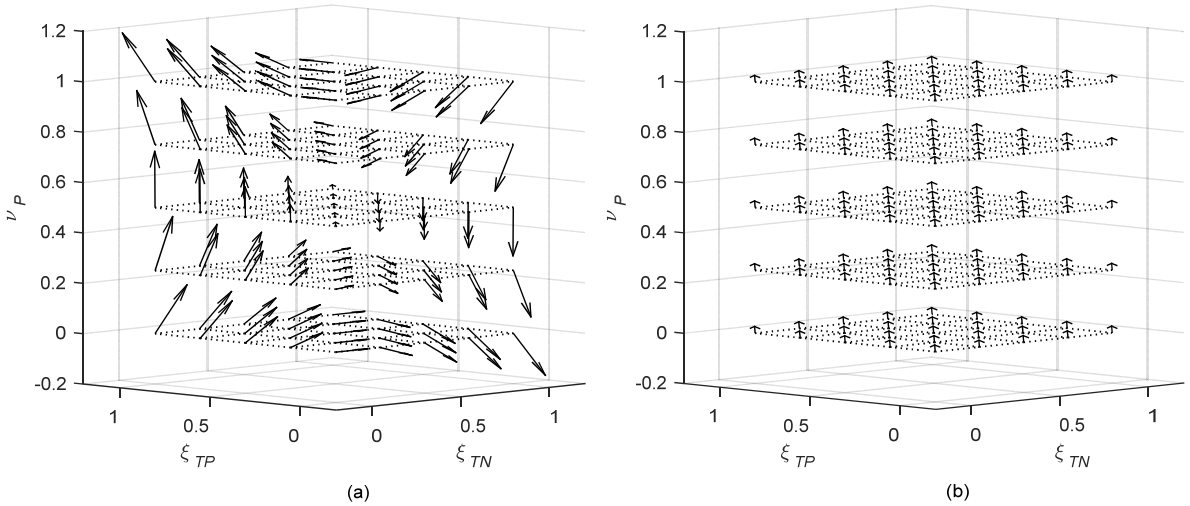
$$\nabla m = \left(\frac{\partial m}{\partial \xi_{TN}}, \frac{\partial m}{\partial \xi_{TP}}, \frac{\partial m}{\partial \nu_P} \right). \tag{9}$$

The gradients are then represented as a 3D graph, where ξ_{TN}, ξ_{TP} and ν_P are shown on x, y and z axes, respectively. For each explored scenario, the gradient ∇m is symbolized using an arrow. While size of the gradient ∇m determines length of the arrow, orientation of the arrow reflects degree of influence of each independent variable. Thus, a non-zero angle, determined by the arrow and its base, indicates influence of the composition on the measure. The base of the arrow is a plane parallel with x and y axes, passing through the z axis at the level given by ν_P. The size of the angle is proportional to the degree of influence of the proportion of positive samples on the measure.

Application of this method on the original and balanced accuracy is shown in Fig. 1. Evaluation of the measure for $\xi_{TN}, \xi_{TP}, \nu_P \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$ was carried out for the purpose of this article. While the original accuracy (acc) is strongly influenced by the proportion of positive samples ν_P (non-zero angles for a majority of the expected settings in Fig. 1 (a)), the balanced accuracy (acc_B) is resistant to the composition (zero angles for all the settings in Fig. 1 (b)). Similar results were achieved for all other biased measures and their balanced counterparts.

From this perspective, the class balanced measures seem to be the perfect solution. Unfortunately, it is not true. A miss-classified sample, which belongs to a minority class, has a different impact on the final value of a balanced measure, comparing with a miss-classified sample which belongs to a majority class. Naturally, the same holds for correctly classified samples. Impact of this issue will be discussed in the following section.

Fig. 1: Visualization of a dependence of the original accuracy acc (a), and the balanced accuracy acc_B (b) on the performance of a classifier and on the composition of a dataset. Non-zero angles, determined by the arrows and their bases in the left panel, indicate the influence of the dataset composition on acc



Source: Authors

5 Discussion

The question of reporting performance of binary classifiers has been opened in the context of the class imbalance problem. As follows from the above stated facts, realization of an expressive evaluation of a binary classifier on a highly imbalanced dataset is not a simple task. Let us briefly summarize the nature of this issue.

Outcomes of an evaluation process are required to be comparable with results published by other researchers. In order to preserve the comparability, the evaluation should be accomplished using widely accepted performance measures. Nevertheless, some of the widely used measures are biased by class distribution in datasets. Once evaluation on imbalanced dataset has to be done, only unbiased measures should be used in order to obtain meaningful results. This poses the question of, which measures should be chosen so that a desired information value would be kept and the comparability requirement would be met?

In our opinion, requirements on eligible measures stated in the question, i.e. obtaining of the desired information value while keeping the comparability, are equally important. Of course, all three requirements on a successful measure, which were stated in the introduction, should be met in order to obtain eligible measures. As already said, eligible measures should be also unbiased. Keeping in mind all these requests, use of the class balanced performance measures seems to be the appropriate solution.

The class balanced measures, which are basically extensions of the widely used biased measures, are resistant to class distributions in datasets, i.e. they are unbiased. They are aimed to provide comparable results on datasets with an arbitrary distribution of classes. The balanced measures capture the same aspects of performance as their

original counterparts, i.e. meaning of the balanced measures is easy to understand for the majority of researchers.

As follows from the essence of the balanced measures, their outcomes should fully correspond to results, which would be obtained by evaluation of classifiers on balanced datasets, using appropriate original biased measures. The balanced measures emulate their originals, where the originals coherently capture aspects of performance of interest. From this perspective, the balanced measures coherently capture the aspect of performance of interest as well. On the other hand, a miss-classified sample belonging to a minority class has a different impact on a balanced measure compared to a miss-classified sample which belongs to a majority class. In this light, fulfilment of the first requirement of a successful measure (see introduction) is disputable.

Despite this drawback, the information value mediated by the class balanced measures is very high, as can be shown in real world data. For example, object detection in large-scale images using the sliding window, inherently leads to highly imbalanced datasets. This issue was demonstrated on a grape detector which was evaluated on real-life images (Škrabánek et al., 2016) where proportions of classes in datasets generated by the sliding window were $\nu_P = 0.001$ and $\nu_N = 0.999$.

Let us compare performance of the grape detector evaluated on these datasets using the balanced (Škrabánek et al., 2016) and imbalanced measures (Škrabánek et al., 2015). The average accuracy of the detector was 0.963 but its average balanced accuracy was 0.936. Its average precision was 0.027 but its average balanced precision was 0.966. For comparison, its average accuracy by a 10-fold cross-validation on balanced datasets was 0.982 and its average precision was 0.980 (Škrabánek et al., 2015). It is evident that the balanced measures provide meaningful results with a high information value, even on highly imbalanced datasets. Thus, at least, the class balanced measures allow a rough comparison with other results.

Moreover, there are two solid facts which strongly support using of the class balanced measures by evaluation of classifiers on imbalanced data. First, the meaning of these measures is evident to a broad community of researchers. Second, the balanced measures can be safely used both on balanced and on imbalanced datasets. In short, the class balanced measures are universal, intuitive, and simple to report. Thus, the second and the third requirement on a successful measure are fully met. Considering all these facts, the class balanced measures have a high probability of the broad acceptance by researchers across the majority of application areas.

The above stated facts lead us to a following conclusion: "When reporting performance of binary classifiers, the balanced measures should be used primarily on balanced as well as on imbalanced datasets." A merit of the class balanced measures is the fact that the selection of appropriate measures does not differ from the current practice. Moreover, once the balanced measures are applied on datasets with a balanced distribution of classes, they provide identical results to the original biased measures. Thus, the balanced measures can be also safely used when training or tuning classifiers on balanced datasets which we also positively recommend as the best practice. This suggestion is aimed to keep uniformity and clarity within every single report, paper or article.

6 Conclusion

In this article, the question of reporting performance of binary classifiers has been discussed. This issue is fundamental in many areas of economics; especially when machine learning or data mining methods are applied on datasets with imbalance distribution of classes. However, the discussed topic is not limited on the economics. It is relevant in many other fields such as medicine or computer vision.

In order to solve the discussed issue, the class balanced measures were suggested as a new standard while reporting performance of binary classifiers. The class balanced measures are basically extensions of the widely used biased measures. They capture the same aspects of performance as their counterparts. Results provided by these measures on balanced datasets do not differ from the original biased measures. The class balanced measures provide also meaningful results with high information value on imbalanced datasets. Since the meaning of these measures does not differ from the original ones, their acceptation by a broad professional community is expected.

As the next step, generalization of the introduced concept for multiclass classification problem is considered. However, less emphasis should not be placed on the presented class balanced measures. In the work (Brodersen et al., 2010), posterior distribution of a measure within the cross-validation has been considered. The way of looking at measures provided by Brodersen et al. might be applied to all the class balanced measures in order to obtain more detail information about their features.

References

Barandela, R. et al. (2003). Strategies for learning in class imbalance problems. *Pattern Recognition*, 36 (3), pp. 849-851.

Brodersen, K. H. et al. (2010). The Balanced Accuracy and Its Posterior Distribution. In: *Pattern Recognition (ICPR), 2010 20th International Conference on*. Istanbul: IEEE, pp. 3121-3124.

Chawla, N. V. et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16 (1), pp. 321-357.

Chi-Yuan Yeh et al. (2011) Employing multiple-kernel support vector machines for counterfeit banknote recognition, *Applied Soft Computing*, 11 (1), pp. 1439-1447.

Choi, S., Cha, S., Tappert, C. C. (2010). A survey of Binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8 (1), pp. 43-48.

Daskalaki, S., Kopanas, I., Avouris, N. (2006). Evaluation of classifiers for an uneven class distribution problem. *Applied Artificial Intelligence*, 20 (5), pp. 381-417.

Feelders, A. J. (2002). Prior Knowledge in Economic Applications of Data Mining. In: *Principles of Data Mining and Knowledge Discovery*. Heidelberg: Springer, pp. 395-400.

Garcia, V. et al. (2007). The class imbalance problem in pattern classification and learning. In: *II Congreso Español de Informática*. Zaragoza: Thomson, pp. 283-291.

Garcia, V. et al. (2010). Theoretical Analysis of a Performance Measure for Imbalanced Data. In: *2010 20th International Conference on Pattern Recognition*. Istanbul: IEEE, pp. 617-620.

Goeuriot, L. et al. (2016). Medical information retrieval: introduction to the special issue. *Information Retrieval Journal*, 19 (1), pp. 1-5.

Hand, D. J. (2012). Assessing the Performance of Classification Methods. *International Statistical Review*, 80 (3), pp. 400-414.

Huang, J., Ling, C. X. (2007). Constructing new and better evaluation measures for machine learning. In: *IJCAI International Joint Conference on Artificial Intelligence*. Hyderabad: IJCAI, pp. 859-864.

Japkowicz, N., Stephen, S. (2002). The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis*, 6 (5), pp. 429-449.

Jeni, L. A., Jeffrey, F. C., Fernando, D. T. (2013). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. In: *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*. Geneva: IEEE, pp. 245-251.

Kotsiantis, S. B., Pintelas, P. E. (2003). Mixture of expert agents for handling imbalanced data sets. *Annals of Mathematics, Computing & Teleinformatics*, 1 (1), pp. 46-55.

Koyejo, O. et al. (2014). Consistent binary classification with generalized performance metrics. In: *Advances in Neural Information Processing Systems (NIPS)*. Montréal: Caltech, pp. 2744-2752.

Lew, M. S. (Ed.) (2001). *Principles of Visual Information Retrieval*. London: Springer-Verlag.

Mani, I., Zhang, I. (2003). kNN approach to unbalanced data distributions: a case study involving information extraction. In: *Proceedings of Workshop on Learning from Imbalanced Datasets.* [online]. Washington: International Machine Learning Society. Available at: http://www.elg.uottawa.ca/~nat/Workshop2003/jzhang.pdf [Accessed 5. 4. 2017].

Murphy, K. P. (2012). *Machine Learning - A Probabilistic Perspective*. Cambridge: MIT Press.

Nigam, K. et al. (2000). Text Classification from Labeled and Unlabeled Documents using EM, *Machine Learning*, 39 (2), pp. 103-134. DOI 10.1023/A:1007692713085

Phua, C. et al. (2004). Minority report in fraud detection: classification of skewed data, A*CM SIGKDD Explorations Newsletter*, 6 (1), pp. 50-59.

Qiao, Q., Beling, P. A. (2016). Decision analytics and machine learning in economic and financial systems, *Environment Systems and Decisions*, 36 (2), pp. 109-113. DOI 10.1007/s10669-016-9601-x

Rosenberg, A. (2012). Classifying skewed data: Importance weighting to optimize Average Recall. In: *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*. Oregon: International Speech Communication Association, pp. 245-251.

Škrabánek, P., Runarsson, T. F. (2015). Detection of grapes in natural environment using support vector machine classifier. In: *Proceedings of the 21st International Conference on Soft Computing MENDEL 2015*. Brno: Brno University of Technology, pp. 143-150.

Škrabánek, P., Majerík F. (2016). Evaluation of performance of grape berry detectors on real-life images. In: *Proceedings of the 22st International Conference on Soft Computing MENDEL 2016*. Brno: Brno University of Technology, pp. 217-224.

Zhao, L. et al. (2012). Data stream classification with artificial endocrine system, *Applied Intelligence*, 37 (3), pp. 390-404. DOI 10.1007/s10489-011-0334-8

Contact Address

Ing. Pavel Škrabánek, Ph.D.; Ing. Petr Doležel, Ph.D.
University of Pardubice, Faculty of Electrical Engineering and Informatics, Department of Process Control
Studentská 95, 532 10, Pardubice, Czech Republic
Email: pavel.skrabanek@upce.cz; petr.dolezel@upce.cz
Phone number: +420 466 037 124; +420 466 037 450

Received: 16. 12. 2016, reviewed: 07. 02. 2017, 30. 03. 2017

Approved for publication: 23. 10. 2017