On extending F-measure and G-mean metrics to multi-class problems

R. P. Espíndola & N. F. F. Ebecken
COPPE/Federal University of Rio de Janeiro, Brazil

Abstract

The evaluation of classifiers is not an easy task. There are various ways of testing them and measures to estimate their performance. The great majority of these measures were defined for two-class problems and there is not a consensus about how to generalize them to multiclass problems. This paper proposes the extension of the F-measure and G-mean in the same fashion as carried out with the AUC. Some datasets with diverse characteristics are used to generate fuzzy classifiers and C4.5 trees. The most common evaluation metrics are implemented and they are compared in terms of their output values: the greater the response the more optimistic the measure. The results suggest that there are two well-behaved measures in opposite roles: one is always optimistic and the other always pessimistic.
Keywords: classification, classifier evaluation, ROC graphs, AUC, F-measure, G-mean.

1 Introduction

Classification [1] is an important task in all knowledge fields. It consists of classifying elements described by a fixed set of attributes into one of a finite set of categories or classes. For example: to diagnose a disease of a person by his medical exams or to identify a potential customer of a product by his purchases. Several artificial intelligence approaches have been applied to this problem like artificial neural networks, decision trees and production rules systems.

In order to test a classifier or a methodology, a researcher may choose some techniques such as leave-one-out, hold-out, bootstrap and cross-validation. Kohavi [2] performed large-scale experiments to compare two of them, bootstrap and cross-validation, and concluded that 10-fold stratified cross-validation was

the best choice, even if computational power allows the use of more folds. This is the scheme employed on this research as detailed in the fourth section.

Along with the testing strategy, the performance evaluators play important role in classification task. The most popular is accuracy, which describes the ability of correctly classify new objects. It computes the ratio of correct decisions made by a classifier and it is easy to be obtained on all situations. Accuracy estimation assumes that all kinds of mistakes are of equal importance just as the benefits of the hits [3]. However, there are cases in which the accuracy estimation can be misled [4].

One of them occurs in problems with imbalanced class distribution [5], in which accuracy tends to favor classifiers with low performance in the rare classes [6]. In real problems, there are many situations in which the cost of this kind of error is very relevant and has to be minimized, such as fraud detection and diseases diagnostics.

Therefore, alternative evaluation metrics should be employed and they are presented in the next section. The third section presents the extensions of some metrics to multi-class problems. Later, the experiments performed are detailed and the results analysis is exposed. In the last section, some concluding remarks and suggestions of future research are done.

2 Classifier performance evaluators

Before presenting the metrics, it is relevant to point that they were defined to two-class problems and they are based on confusion matrix, a tool which informs the sorts of hits and errors made by a classifier. The classes are named positive and negative and the confusion matrix has four values computed in terms of real and predicted classes, namely:

- TP (true positives): the amount of positive elements predicted as positive;
- FP (false positives): the amount of negative elements predicted as positive;
- FN (false negatives): the amount of positive elements predicted as negative;
- TN (true negatives): the amount of negative elements predicted as negative;

The most common performance evaluators are:

1. **accuracy**: it is the ratio of correct decisions made by a classifier

$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}$$

2. **sensitivity**: also called hit rate or recall, it measures how much a classifier can recognize positive examples

$$\text{sens} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

3. **specificity**: it measures how much a classifier can recognize negative examples

$$\text{spec} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3}$$

4. **precision**: it is the ratio of predicted positive examples which really are positive

$$\text{prec} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

5. **F-measure**: it is the harmonic mean of sensitivity and precision [7]

$$\text{F.mea} = \frac{\left(\beta^2 + 1\right) \times \text{sens} \times \text{prec}}{\text{sens} + \beta^2 \times \text{prec}}, \ \beta \geq 0 \tag{5}$$

6. **G-mean1**: it is the geometric mean of sensitivity and precision [8]

$$\text{GSP} = \sqrt{\text{sens} \times \text{prec}} \tag{6}$$

7. **G-mean2**: it is the geometric mean of sensitivity and specificity [8]

$$\text{GSS} = \sqrt{\text{sens} \times \text{spec}} \tag{7}$$

In this study, the β parameter on F-measure is equal to 1, which means that sensitivity and precision have the same importance.

It is known that there is a decreasing hyperbolic relation between sensitivity and precision [9] and a way to deal with this employs ROC graphs. These graphs have been used as a tool for visualization, organization and selection of classifiers based on their performances [10]. A ROC graph is bidimensional in which the FP rate (1 – specificity) is plotted on the horizontal axis and the sensitivity on the vertical one. Fig. 1 shows some classifiers represented as dots in the ROC space. Fawcett [10] calls them discrete due to the lack of class membership information on the predictions, that is, a classifier only outputs the class and not the degree to which an object is a member of the class. The ones which provide these degrees are called by the author as scoring classifiers.

It is relevant to note that the nearer to the upper-left side of ROC space, the better a classifier is. Moreover all classifiers in the diagonal line have random behavior and the ones below this line should be discarded.

The focus of this study is on discrete classifiers and their ROC curves are the "curves" which connect the classifiers dots to the diagonal edges (fig. 2). It is easy to notice that classifiers A and B are better than the others but the comparison between them is difficult.

A way to solve this problem is to calculate the AUC, that is, the area under ROC curve (fig. 3). The greater the area, the better is the classifier.

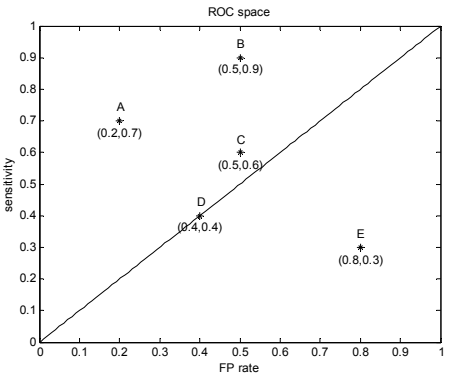


Figure 1: Discrete classifiers performance on ROC space.

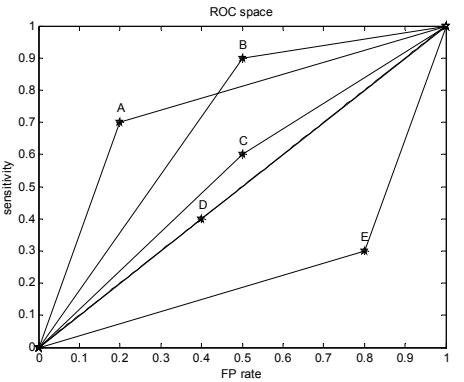


Figure 2: ROC curves of some discrete classifiers.

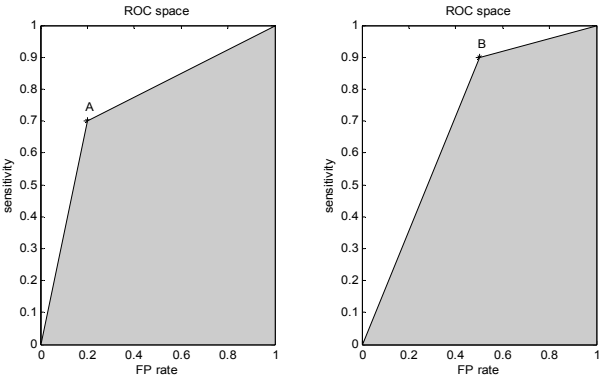


Figure 3: AUC of some discrete classifiers.

3 Generalizing some measures to multi-class problems

There is no consensus about how to act when problems with more than two classes are faced. Two strategies have been proposed for AUC and this work proposes to perform the same operations for F-measure and G-mean.

The first strategy [11] draws a ROC curve for each class of a problem in which each class is considered as the positive class and the remaining ones the negative class. Therefore, after the calculation of the AUC for each class, the final AUC is the weighted mean of them, in which the relative frequencies of the classes on the data are their weights:

$$AUC1_{total} = \sum_{i=1}^{K} AUC(c_i) \times f_r(c_i) \tag{8}$$

in which K is the amount of classes. It is relevant to point out that this procedure causes the imbalancing of classes, but Fawcett [10] defends it by noticing that the computations are very simple and the curves are easily visualized.

The second approach [12] tries to avoid the class imbalancing by computing the final AUC based on each pair of classes. In other words, at a given time, a pair of classes is selected and one is defined as the positive class and the other as the negative class. The AUC of this setting is calculated and the process is repeated with these same classes, but now in changed roles. This scheme is performed to each pair of classes and the final AUC is defined by the following expression:

$$AUC2_{total} = \frac{2}{K \times (K-1)} \times \sum_{1 \le i,j \le K} AUC(c_i, c_j) \tag{9}$$

This research extends F-measure and G-means on the same fashion as above.

4 Experimental results and analysis

4.1 Experiments performed

In order to observe the metrics behavior, a genetic fuzzy system [13] and a C4.5 decision tree tool [14] were used to produce classifiers on seven well- known datasets obtained in UCI repository, besides a meteorological dataset from International Airport of Rio de Janeiro. Table 1 shows the datasets, their dimensions, the amount of rules generated and their alias to future reference in this text.

The genetic fuzzy system is a genetic algorithm which optimizes zero-order TSK fuzzy rule bases in order to selects the shortest subset of rules with maximum accuracy and minimum amount of features possible. It has some special features like population initiation by fuzzy trees and two schemes for

Boolean recombination [15]. Table 2 shows some settings employed and each one was performed 10 times to obtain the mean results.

Before working with the datasets, some changes were made to allow the analysis of the method. Repeated records or records with incomplete information were eliminated and qualitative features were converted to discrete quantitative features. The employed scheme of testing was ten-fold stratified cross-validation.

Table 1: Summary of datasets' characteristics.

Dataset	Valid features	Classes	Valid records	Reference
balance scale	4	3	625	bala
car evaluation	6	4	1728	car
credit card approval	15	2	653	cred
fog classification	18	7	26482	fog
glass identification	9	7	143	glass
ionosphere	33	2	351	iono
pima indian diabetes	8	2	768	pima
yeast protein localization	8	10	1484	yeast

Table 2: Genetic fuzzy system settings.

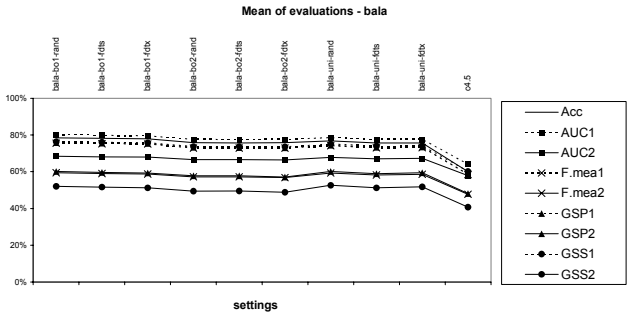
Recombination	Reference	Initialization	Reference
boolean-1	bo1	random	rand
boolean-2	bo2	fuzzy tree	fdts
uniform	uni	fuzzy tree with rule exclusion	fdtx

It is relevant to notice that the number beside the measures names represents the strategy of extension to multiclass problems employed: 1 for the first scheme – which considers one class against all – and 2 for the second – which deals with each pair of classes.

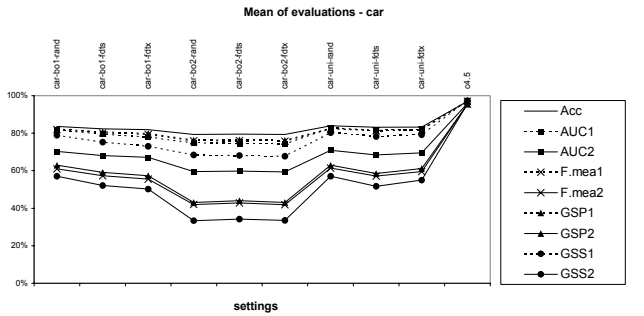
4.2 Results analysis

Observing the results from problems with two classes in figs. 4-5 – *cred*, *iono* and *pima* – the measures had practically the same output. On multi-class problems it is possible to notice the differences between them. Considering measures with higher values as optimistic and those with lower values as pessimistic, it is clearly shown that AUC1 is the most optimistic evaluation and GSS2 the most pessimistic.

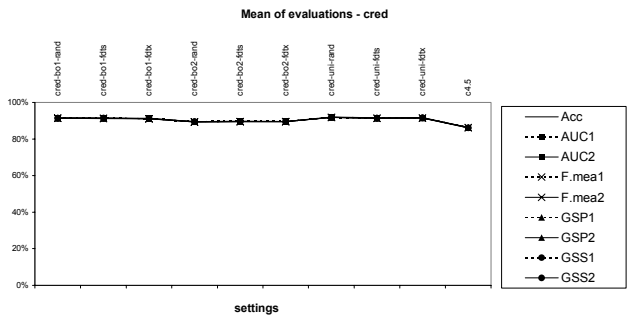
Following this concept, on comparing the two ways of extending evaluation metrics to multi-class problems, it can be seen that the first strategy is more optimistic than the second one irrespective of the measure employed.



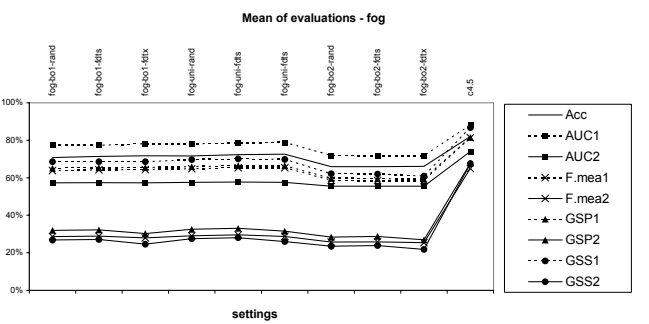
(a)



(b)

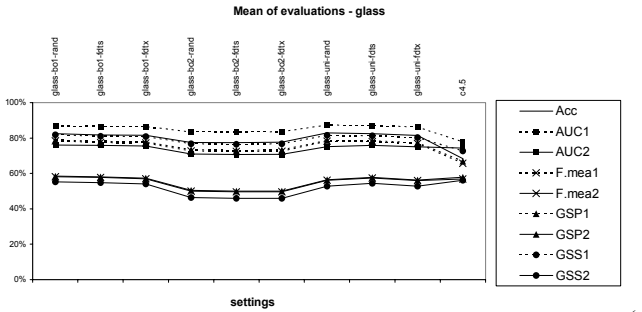


(c)

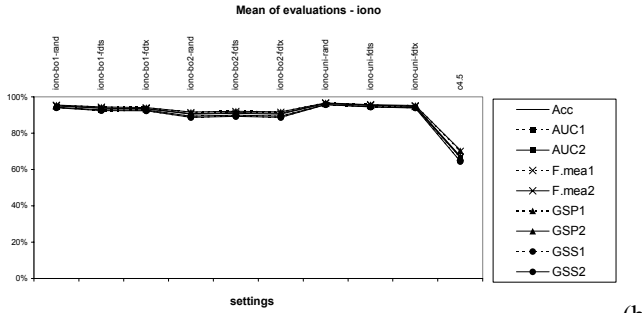


(d)

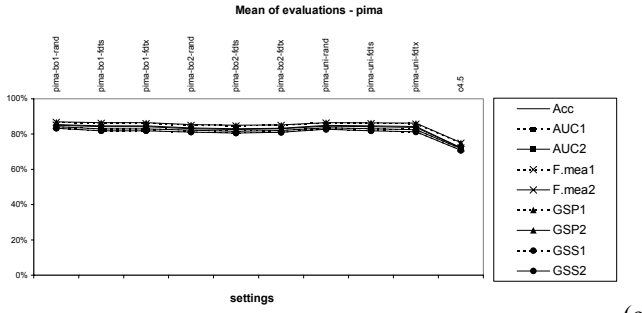
Figure 4: Mean of evaluations on *bala*, *car*, *cred* and *fog* datasets.



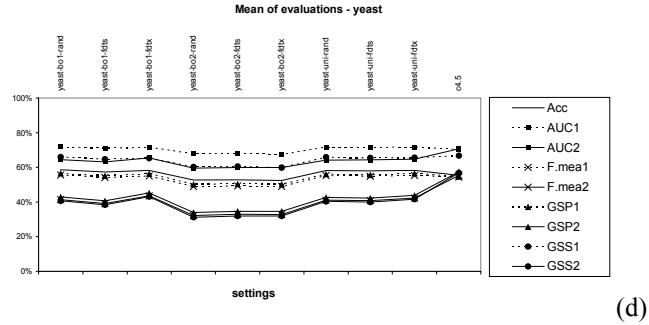
(a)



(b)



(c)



(d)

Figure 5: Mean of evaluations on *glass*, *iono*, *pima* and *yeast* datasets.

5 Final considerations

This study aimed to contribute to the discussion of how to evaluate a classifier performance by extending F-measure and G-mean metrics to multi-class problems as done with the area under ROC curve. Some well-known problems were approached by a genetic fuzzy system and by a decision tree tool. The results showed that on two-class problems the metrics have similar behaviour. This situation may be justified by the fact that these problems do not have imbalanced classes. On multi-class problems, two metrics were well-behaved: AUC1 produced the highest evaluations and GSS2 the lowest ones, being considered optimistic and pessimistic, respectively.

The results obtained from the eight problems suggest that the second strategy of metrics extension to multi-class problems is more rigorous than the first, mainly when there are rare classes. Future studies will consider other datasets with two classes, one being rare, or more classes. Moreover other classification models will be employed in order to verify whether these observations will be repeated.

Acknowledgements

This research was supported by CNPQ and the Petroleum National Agency under the program PRH-ANP/MME/MCT.

References

[1] Gordon, A.D., *Classification*, Chapman and Hall: London, 1981.

[2] Kohavi, R., A Study Of Cross-Validation And Bootstrap For Accuracy Estimation and Model Selection. *Proc. of Int. Joint Conf. on Artificial Intelligence*, pp. 1137-1145, Quebec, Canada, 1995.

[3] Pietersma, D., Lacroix, R., Lefebvre, D., Wade, K.M., Performance analysis for machine-learning experiments using small data sets. *Computers and Electronics in Agriculture*, **38(1)**, pp. 1-17, 2003.

[4] Provost, F., Fawcett, T., Kohavi, R., The Case Against Accuracy Estimation for Comparing Classifiers. *Proc. of 15th Int. Conf. of Machine Learning*, pp. 445-553, Wisconsin, USA, 1998.

[5] Weiss, G.M., Mining with Rarity: A Unifying Framework. *ACM SIGKDD Explorations*, **6(1)**, pp. 7-19, 2004.

[6] Weiss, G.M., Provost, F., Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*, **19**, pp. 315-354, 2003.

[7] Lewis, D., Gale, W., Training text classifiers by uncertainty sampling. *Proc. of 7th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 3-12, Dublin, Ireland, 1994.

[8] Kubat, M., Holte, R.C., Matwin, S., Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning*, **30**, pp. 195–215, 1998.

[9] Egghe, L., Rousseau, R., A theoretical study of recall and precision using a topological approach to information retrieval. *Information Processing & Management*, **34(2/3)**, pp. 191-218, 1998.
[10] Fawcett, T., ROC Graphs - Notes and Practical Considerations, *Machine Learning*, submitted, 2004.
[11] Provost, F., Domingos, P., Well-trained PETs - Improving Probability Estimation Trees, New York University CeDER Working Paper #IS-00-04. 2001.
[12] Hand, D.J., Till, R.J., A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, **45**, pp. 171-186, 2001.
[13] Espíndola, R.P., Ebecken, N.F.F., Population Initiation by a Fuzzy Decision Tree. *Proc. of 5th Int. Conf. on Data Mining*, Malaga, Spain, 2004.
[14] Witten, I.H., Frank, E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann: California, 1999
[15] Espíndola, R.P., Ebecken, N.F.F., Boolean Recombination In A Fuzzy Genetic System. *Proc. of 25th Iberian Latin American Congress on Computational Methods in Engineering*, Pernambuco, Brazil, 2004.