# An Experimental Comparison of Classification Performance Metrics*

**C. Ferri, J. Hernández-Orallo, R. Modroiu**
Departament de Sistemes Informàtics i Computació
Universitat Politècnica de València, Spain.
{cferri, jorallo, emodroiu}@dsic.upv.es

## ABSTRACT

Performance metrics in classification are fundamental to assess the quality of learning methods and learned models. However, many different measures have been defined and choices made by one metric are different from choices made by other metrics. In this work we analyse experimentally the behaviour of 16 different performance metrics in different scenarios, identifying different clusters and relationships between measures.

## 1 INTRODUCTION

The correct evaluation of learned models is one of the most important issues in machine learning. One perspective of this evaluation can be based on statistical significance and confidence intervals, when we want to claim that one model is better than another or that one method is better than another. A different perspective, however, relies on which *metric* is used to evaluate a learned model. It is certainly not the same to evaluate a regression model with absolute error that with squared error.

In this work we concentrate on metrics for evaluating classifiers, such as accuracy, F-measure, rank rate, Area Under the ROC Curve (AUC), squared error, log loss/entropy, etc. Some of these metrics have different applications and measure quite different things. More specifically, we will use 16 different metrics, which we classify in three families as follows:

- Metrics based on a threshold and a *qualitative* understanding of error: accuracy, macro-averaged accuracy (arithmetic, geometric and mixed), mean F-measure (F-score) and Kappa statistic. These measures are used when we want a model to minimise the number of errors and, hence, these metrics are usual in many direct applications of classifiers. Inside this family, some of these measures are more appropriate for balanced or imbalanced datasets, or for information retrieval tasks.

- Metrics based on a *probabilistic* understanding of error, i.e. measuring the deviation from the true probability: mean absolute error, mean squared error (Brier's score)

*This work has been partially supported by CICYT under grant TIC 2001-2705-C03-01, ICT for EU-India Cross Cultural Dissemination Project ALA/95/23/2003/077-054 and Generalitat Valenciana under grant GV04A/389.

and log loss (cross-entropy). These measures are especially useful when we want an assessment on the reliability of the classifiers, not only measuring when they fail but whether they have selected the wrong class with a high or low probability. This is crucial in committee models, machine ensembles, to properly perform a weighted fusion of the models.

- Metrics based on how well the model *ranks* the examples: AUC (Flach et al., 2003), which for two classes is equivalent to the Mann-Whitney-Wilcoxon statistic (we will use five different extensions for multiclass problems, following the extension introduced by Hand and Till (Hand and Till, 2001)) and probability (rank) rate (two variants). These are important for many applications, such as mailing campaign design, CRM, fraud detection, spam filtering, etc., where classifiers are used to select the best $n$ instances of a set of data.

In this paper we analyse how these 16 metrics correlate to each other, in order to analyse in which extent and in which situations the results obtained and the model choices performed with one metric are extensible to the other metrics. The results show that most of these metrics really measure different things and in many situations the choice made with one metric can be different to the choice made with another. These differences become higher for multiclass problems or problems with very imbalanced class distribution.

Although there are some previous works that compare theoretically some of these measures (see e.g. (Flach, 2003)), empirical studies have been scarce and limited in the literature. The only exception to this is (Caruana and Niculescu-Mizil, 2004), but it only uses seven datasets and it is restricted to binary (two-class) evaluation metrics. Additionally, some important measures such as macro-averaged accuracy, the AUC variants and probability rate are not included in their study.

To our knowledge, this is the first experimental work which thoroughly compares the most generally used classifier evaluation metrics for binary and multiclass problems drawing conclusions about the correlation and interdependence of these measures.

## 2 METHODOLOGY

As we have said before, for this study 16 classification performance metrics were chosen: Percent_Correct (PC), Kappa_statistic (KS), Mean_Absolute_Error (MAE), Mean_Squared_Error (MSE, also known as Brier score), Log_Loss (LLoss, also known as Cross Entropy), AUC_1vs1u (A11u), AUC_1vs1p (A11p), AUC_1vsNu (A1nu), AUC_1vsNp (A1np) (corresponding to one vs. one or one vs. the rest extensions to the Area Under the ROC Curve, using a uniform distribution or using class probabilities), MAVG_Acc_A (MAA, arithmetic macroaveraged accuracy), MAVG_Acc_G (MAG, geometric macroaveraged accuracy), MAVG_Acc_M (MAM, a mixture between MAVG_Acc_A and MAVG_Acc_G), MeanF-Measure (MFM, the mean of all the F-scores of one vs. the rest class combinations), MeanProbRate (MPR, the sum of the probabilities of the correct class), MAVG_MeanProbRate (MAMPR, the sum of the probabilities of the correct class, uniformly averaging all classes) and Multiclass Wilcoxon (MV). Proper definitions and references for all these metrics are included in (Ferri et al, 2004).

The experiments were performed using Weka, which we extended with several new metrics, not included in the current distribution (the last fourteen metrics in the above list). We used six well-known machine learning algorithms: C4.5, Naive Bayes, Logistic Regression, Multilayer Perceptron, K-Nearest Neighbour, AdaBoost and we performed the experiments with 20 medium-size datasets[1] included in the machine learning repository (Blake and Merz, 1998), 10 of them being two-class (binary) problems from which 5 balanced and 5 imbalanced and 10 of them being multiclass, 5 balanced, 5 imbalanced. The above mentioned models were evaluated using 20 × 5 fold cross-validation, each of the 6 models being applied to each of the 20 datasets, getting 600 results for each dataset, making 12,000 results in total. We set up five types of analysis: an overall analysis for all datasets, an analysis for binary problems, for multiclass problems, for balanced and for imbalanced problems. In each case we calculated the standard linear correlation and Spearman rank correlation between all sixteen metrics. In order to avoid negative values for correlation we work with 1-MAE, 1-MSE and 1-LLoss. We will use dendrograms for representation; where the link distance is defined as (1-correlation).

## 3 Analysis of Results

In this section we discuss some of the interesting outcomes we encountered from the analysis of the correlation between metrics. First we analyse the standard correlations and then the rank correlations. Table 1 shows the standard correlation between all metrics.

| - | pc | ks | mae | mse | lloss | a11u | a11p | a1nu | a1np | maa | mag | mam | mfm | mampr | mpr | mw | mc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| pc | 1.00 | 0.9 | 0.92 | 0.88 | 0.72 | 0.83 | 0.83 | 0.84 | 0.84 | 0.94 | 0.7 | 0.89 | 0.94 | 0.88 | 0.93 | 0.84 | 0.87 |
| ks | 0.90 | 1 | 0.95 | 0.94 | 0.64 | 0.94 | 0.94 | 0.94 | 0.94 | 0.94 | 0.75 | 0.91 | 0.94 | 0.86 | 0.82 | 0.94 | 0.9 |
| mae | 0.92 | 0.95 | 1 | 0.93 | 0.6 | 0.9 | 0.9 | 0.9 | 0.9 | 0.92 | 0.72 | 0.88 | 0.92 | 0.88 | 0.89 | 0.91 | 0.88 |
| mse | 0.88 | 0.94 | 0.93 | 1 | 0.79 | 0.9 | 0.91 | 0.91 | 0.91 | 0.88 | 0.66 | 0.83 | 0.88 | 0.8 | 0.79 | 0.9 | 0.87 |
| lloss | 0.72 | 0.64 | 0.6 | 0.79 | 1 | 0.63 | 0.63 | 0.64 | 0.64 | 0.67 | 0.51 | 0.64 | 0.66 | 0.62 | 0.63 | 0.63 | 0.67 |
| a11u | 0.83 | 0.94 | 0.9 | 0.9 | 0.63 | 1 | 0.999 | 0.996 | 0.99 | 0.89 | 0.75 | 0.87 | 0.89 | 0.81 | 0.76 | 1 | 0.89 |
| a11p | 0.83 | 0.94 | 0.9 | 0.91 | 0.63 | 0.999 | 1 | 0.997 | 0.99 | 0.89 | 0.73 | 0.86 | 0.89 | 0.81 | 0.76 | 0.99 | 0.88 |
| a1nu | 0.84 | 0.94 | 0.9 | 0.91 | 0.64 | 0.996 | 0.997 | 1 | 0.99 | 0.9 | 0.75 | 0.88 | 0.9 | 0.82 | 0.78 | 1 | 0.89 |
| a1np | 0.84 | 0.94 | 0.9 | 0.91 | 0.64 | 0.99 | 0.99 | 0.99 | 1 | 0.88 | 0.7 | 0.84 | 0.88 | 0.8 | 0.77 | 0.99 | 0.88 |
| maa | 0.94 | 0.94 | 0.92 | 0.88 | 0.67 | 0.89 | 0.89 | 0.9 | 0.88 | 1 | 0.86 | 0.98 | 1 | 0.93 | 0.88 | 0.9 | 0.9 |
| mag | 0.70 | 0.75 | 0.72 | 0.66 | 0.51 | 0.75 | 0.73 | 0.75 | 0.7 | 0.86 | 1 | 0.94 | 0.87 | 0.84 | 0.73 | 0.76 | 0.77 |
| mam | 0.89 | 0.91 | 0.88 | 0.83 | 0.64 | 0.87 | 0.86 | 0.88 | 0.84 | 0.98 | 0.94 | 1 | 0.99 | 0.93 | 0.86 | 0.88 | 0.89 |
| mfm | 0.94 | 0.94 | 0.92 | 0.88 | 0.66 | 0.89 | 0.89 | 0.9 | 0.88 | 1 | 0.87 | 0.99 | 1 | 0.93 | 0.89 | 0.9 | 0.91 |
| mampr | 0.88 | 0.86 | 0.88 | 0.8 | 0.62 | 0.81 | 0.81 | 0.82 | 0.8 | 0.93 | 0.84 | 0.93 | 0.93 | 1 | 0.96 | 0.83 | 0.86 |
| mpr | 0.93 | 0.82 | 0.89 | 0.79 | 0.63 | 0.76 | 0.76 | 0.78 | 0.77 | 0.88 | 0.73 | 0.86 | 0.89 | 0.96 | 1 | 0.79 | 0.83 |
| mw | 0.84 | 0.94 | 0.91 | 0.9 | 0.63 | 1 | 0.99 | 1 | 0.99 | 0.9 | 0.76 | 0.88 | 0.9 | 0.83 | 0.79 | 1 | 0.89 |

Table 1: **Standard correlation results for all datasets.**

In the left part of Figure 1 we show a dendrogram built from the obtained standard correlations using all the available results. This figure represents the relations between the measures in an abridged and more comprehensible way.

[1] credit-ranking, pima-diabetes, heart-statlog, hepatitis, ionosphere, kr-vs-kp, post-operativeW, german-credit, spect, breast-cancer, balance-scale, iris, soybean, cmcW, dermatology, new-thyroidW, segment, taeW, wine, waveform.

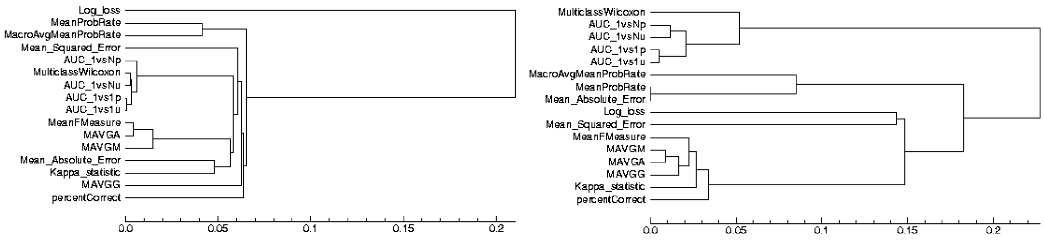

Figure 1: **Dendrograms of standard correlations (left) and rank correlations (right) between the metrics for all datasets.**

The first observation can be made with the AUC measures. The 5 variants of AUC behave quite similarly, so they can even be used interchangeably. This means that previous works in the literature using these different variants for evaluating rankers can be contrasted safely, independently of which variant they have used. Additionally, it is interesting to note that no other measure correlates to AUC more than 0.94, justifying the use of the AUC as a genuinely different measure. The results for Kappa are quite surprising, because this metric is just a modification of error (accuracy), which just corrects the right predictions occurred by chance. It should be close to accuracy but, unexpectedly, it is closer to Minimum Absolute Error and other measures. An expected result is the close relationship between MAVG_Acc_A and MeanFMeasure. Following their definitions, it seems that they are just measuring the same thing. Finally, Logloss seems to be a quite different metric with respect to the rest.

If we use the rank correlation, things are slightly different and more insightful, since this correlation tells us whether the decisions would be the same, independently of the absolute values. Below we show the table and in the right part of Figure 1 the dendograms for all the measures.

| - | pc | ks | mae | mse | llos | a11u | a11p | a1nu | a1np | maa | mag | mam | mfm | mampr | mpr | mw | mc |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| acc | 1 | 0.97 | 0.77 | 0.85 | 0.62 | 0.71 | 0.71 | 0.69 | 0.7 | 0.91 | 0.83 | 0.88 | 0.94 | 0.73 | 0.77 | 0.7 | 0.8 |
| ks | 0.97 | 1 | 0.76 | 0.82 | 0.6 | 0.72 | 0.73 | 0.71 | 0.72 | 0.97 | 0.91 | 0.95 | 0.97 | 0.78 | 0.76 | 0.72 | 0.82 |
| mae | 0.77 | 0.76 | 1 | 0.62 | 0.32 | 0.57 | 0.57 | 0.55 | 0.56 | 0.73 | 0.69 | 0.71 | 0.75 | 0.91 | 0.999 | 0.61 | 0.7 |
| mse | 0.85 | 0.82 | 0.62 | 1 | 0.86 | 0.77 | 0.77 | 0.76 | 0.76 | 0.78 | 0.72 | 0.76 | 0.81 | 0.62 | 0.62 | 0.76 | 0.77 |
| log | 0.62 | 0.6 | 0.32 | 0.86 | 1 | 0.72 | 0.72 | 0.71 | 0.71 | 0.58 | 0.53 | 0.56 | 0.59 | 0.36 | 0.32 | 0.65 | 0.61 |
| 11u | 0.7 | 0.72 | 0.57 | 0.77 | 0.72 | 1 | 0.995 | 0.98 | 0.97 | 0.73 | 0.7 | 0.72 | 0.72 | 0.62 | 0.57 | 0.95 | 0.78 |
| 11p | 0.71 | 0.73 | 0.57 | 0.77 | 0.72 | 0.995 | 1 | 0.98 | 0.98 | 0.72 | 0.69 | 0.71 | 0.72 | 0.62 | 0.57 | 0.94 | 0.78 |
| 1nu | 0.69 | 0.71 | 0.55 | 0.76 | 0.71 | 0.98 | 0.98 | 1 | 0.99 | 0.71 | 0.68 | 0.7 | 0.71 | 0.61 | 0.55 | 0.95 | 0.77 |
| 1np | 0.7 | 0.72 | 0.56 | 0.76 | 0.71 | 0.97 | 0.98 | 0.99 | 1 | 0.71 | 0.67 | 0.7 | 0.7 | 0.6 | 0.56 | 0.94 | 0.77 |
| maa | 0.91 | 0.97 | 0.73 | 0.78 | 0.58 | 0.73 | 0.72 | 0.71 | 0.71 | 1 | 0.96 | 0.99 | 0.98 | 0.82 | 0.73 | 0.73 | 0.82 |
| mag | 0.83 | 0.91 | 0.69 | 0.72 | 0.53 | 0.7 | 0.69 | 0.68 | 0.67 | 0.96 | 1 | 0.98 | 0.95 | 0.81 | 0.69 | 0.7 | 0.78 |
| mam | 0.88 | 0.95 | 0.71 | 0.76 | 0.56 | 0.72 | 0.71 | 0.7 | 0.7 | 0.99 | 0.98 | 1 | 0.98 | 0.82 | 0.71 | 0.72 | 0.81 |
| mfm | 0.94 | 0.97 | 0.75 | 0.81 | 0.59 | 0.72 | 0.72 | 0.71 | 0.7 | 0.98 | 0.95 | 0.98 | 1 | 0.8 | 0.75 | 0.72 | 0.82 |
| mampr | 0.73 | 0.78 | 0.91 | 0.62 | 0.36 | 0.62 | 0.62 | 0.61 | 0.6 | 0.82 | 0.81 | 0.82 | 0.8 | 1 | 0.91 | 0.66 | 0.73 |
| mpr | 0.77 | 0.76 | 0.999 | 0.62 | 0.32 | 0.57 | 0.57 | 0.55 | 0.56 | 0.73 | 0.69 | 0.71 | 0.75 | 0.91 | 1 | 0.61 | 0.7 |
| mwi | 0.7 | 0.72 | 0.61 | 0.76 | 0.65 | 0.95 | 0.94 | 0.95 | 0.94 | 0.73 | 0.7 | 0.72 | 0.72 | 0.66 | 0.61 | 1 | 0.77 |

Table 2: **Rank correlation results for all datasets.**

We can see several things from the dendrogram. First, there is an equivalence between

MeanProbRate and MAE, that is not surprising if we take a look at their definitions (for standard correlation they might differ but they always make the same decisions). Secondly, we have 3 clusters. The first includes the AUC measures, being, in some way, outliers from the rest, and we can even see more clearly that AUC is a completely different measure and no substitute can be used safely. Secondly, MSE and Logloss are now cluster and Logloss (if we use rank correlation) is not so different as it was with standard correlation. Finally, there is a third cluster with many measures based on counting hints and errors and ignoring the ranks or the probabilities.

Now, if we compare the correlation results of 2-class problems with multiclass problems (see Figure 2), we have some expected results. All the AUC variants collapse, since they are all extensions for multiclass problems but equivalent for 2-class problems. The rest of the correlations are similar in both cases.

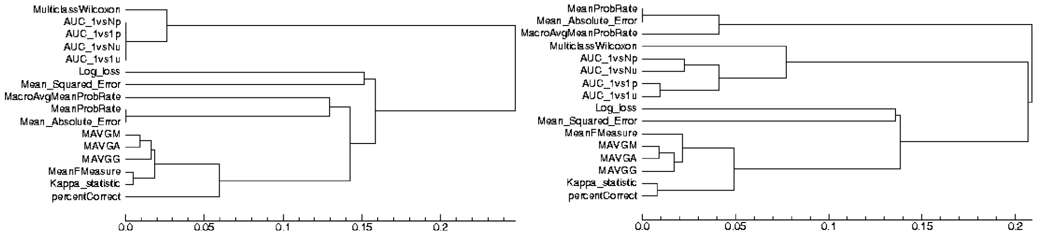

Figure 2: **Dendrograms of rank correlations between the metrics for two-class datasets(left) and multi-class datasets (right).**

Finally, if we compare the correlations for the datasets with balanced class distribution against the correlations for the datasets with imbalanced class distribution (see Figure 3), the results are much more interesting.

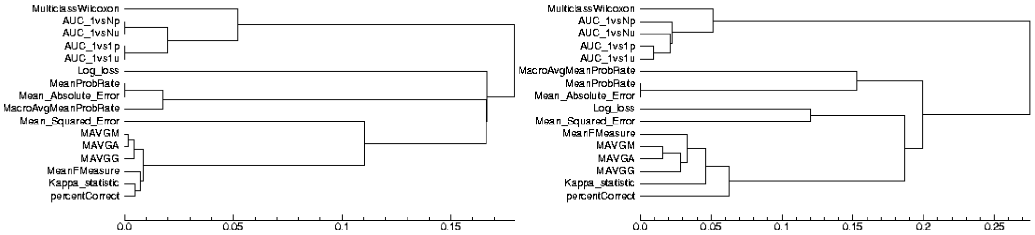

Figure 3: **Dendrogram of rank correlations between the metrics for balanced datasets (left) and imbalanced datasets (right).**

Overall results are significantly different and measures diverge very significantly (correlations are much lower) for imbalanced datasets.

## 4 Conclusions

Summing up, the previous analysis shows that most of the measures used in machine learning for evaluating classifiers really measure different things, especially for multiclass problems and problems with imbalanced class distribution. One of the most surprising results from the study is that the correlations between metrics inside the same family (of the three families: qualitative understanding of error, probabilistic understanding of error and ranking understanding of error) are not very high, showing that even with a qualitative understanding of error, it is significantly different to use accuracy or Kappa statistic, with a probabilistic understanding of error, it is significantly different to use MSE or Logloss, and, when we want to rank predictions, it is significantly different to use AUC or to use ProbRate. Consequently, the analyses of machine learning methods (stating, e.g., that one method is better than other) using different metrics (even inside the same family) could not be comparable and extensible to the other metrics, since, usually, the differences in performance between modern machine learning methods are tight.

## References

Blake, C. and Merz, C. (1998). UCI repository of machine learning databases.

Caruana, R. and Niculescu-Mizil, A. (2004). An empirical analysis of the relationship between auc and eight standard supervised learning performance criteria. In *1st First Workshop on ROC Analysis in AI.*

Flach, P. and Blockeel, H. and Ferri, C. and Hernández, J. and and Struyf, J. (2003). Decision support for data mining: introduction to ROC analysis and its applications. In Bohanec, M., editor, *Data Mining and Decision Support: Integration and Collaboration.* Kluwer Academic Publishers, Boston.

Flach, P. A. (2003). The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In *Proc. 20th International Conference on Machine Learning (ICML'03)*, pages 194–201. AAAI Press.

Ferri, C. and Hernández-Orallo, J. and Modroiu, R. (2004). An Experimental Comparison of Performance Measures for Classification . In Technical Report, DSIC, pages 1–10. In *http://www.dsic.upv.es/users/elp/cferri/measures.pdf*

Hand, D. and Till, R. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2):171–186.