to affect the accuracy measures: the classifier distribution of error, the dataset class proportions and the number of classes.

6.1 Methods

We first give an overview of our methods, and then focus on its different steps. We generate two series of matrices with various error distribution and fixed class proportions. We compute the accuracy according to every measure under study. For a given measure we then consider all possible pairs of matrices obtained by associating one matrix from the first series to one of the second series. For each pair we compute the difference between the two values of the measure. The line corresponding to a zero difference separates the plane between pairs for which the first matrix is preferred by the measure, and pairs for which the second one is. We call this line the *discrimination line*.

Our approach is related to the isometrics concept described in [13]. The main difference is our discrimination lines are function of the error distribution, while ROC curves uses TPR and FPR. Moreover, we focus on a single isometrics: the one associated with a zero difference. This allows us to represent several discrimination lines on the same plots. By repeating the same process for different class proportions, or number of classes, we can therefore study if and how the discrimination lines are affected by these parameters.

We now focus on the modeling of classification error distribution. Let us consider a confusion matrix corresponding to a perfect classification, as presented in Table 4. Applying a classifier with lower performance on the same dataset will lead to a matrix diverging only in the distribution of instances in columns taken independently. Indeed, since the dataset is fixed, the class proportions, and hence the distributions inside rows, cannot change (i.e. the π_i are constant).

Table 9: confusion matrix with controlled errors for a classifier imperfect in all classes.

	C_1	…	C_j	…	C_k
$\hat{C}_1$	$c_1\pi_1$	…	$\frac{1-c_j}{k-1}\pi_j$	…	$\frac{1-c_k}{k-1}\pi_k$
⋮	⋮	⋱	⋮	⋮	⋮
$\hat{C}_i$	$\frac{1-c_1}{k-1}\pi_1$	…	$c_j\pi_j$	…	$\frac{1-c_k}{k-1}\pi_k$
⋮	⋮	⋮	⋮	⋱	⋮
$\hat{C}_k$	$\frac{1-c_1}{k-1}\pi_1$	…	$\frac{1-c_j}{k-1}\pi_j$	…	$c_k\pi_k$

For simplicity purposes, we suppose the misclassified instances for some class C_i are uniformly distributed by the classifier on the other estimated classes $\hat{C}_{j\neq i}$. In other words, the perfect classifier correctly puts a proportion π_i of the dataset instances in $\hat{C}_i$ and none in $\hat{C}_{j\neq i}$, whereas our imperfect classifier correctly process only a proportion $c_i\pi_i$ $(0 \le c_i \le 1)$ and incorrectly puts a proportion $\frac{1-c_i}{k-1}\pi_{j\neq i}$ in each other class $\hat{C}_{j\neq i}$, where $1-c_i$ is the accuracy drop for this class. This allows us to control the error level in the confusion matrix, a perfect classification corresponding to $c_i = 1$ for all classes. Table 9 represents the confusion matrix obtained for a k-class problem in the case of a classifier undergoing an accuracy drop in all classes.

By using a range of values in $[0;1]$ for c, we can generate a series of matrices with decreasing error level. However, comparing pairs of matrices from the same series is fruitless, since it will lead by construction to the same discrimination lines for all measures, when we want to study their differences. We therefore considered two different series: in the first (represented on the x axis), the same accuracy drop c is applied to all classes, whereas in the second (y axis), it is applied only to the first class. In Table 9, the first series corresponds to $c_i = c$ $(\forall i)$, and the second to $c_1 = c$ and $c_{i\geq 2} = 1$. We thus expect the accuracy measures to favor the second series, since only its first class is subject to classification errors..

To investigate the sensitivity of the measures to different class proportions values (i.e. π_i), we generated several pairs of series with controlled class imbalance. In the balanced case, each class represents a proportion $\pi_i = 1/k$ of the instances. We define the completely imbalanced case by defining the 1st class as having twice the number of instances in the 2nd one, which has itself twice the size of the 3rd one, and so on. In other words, $\pi_i = 2^{k-i}/\left(2^k - 1\right)$, where the denominator corresponds to the quantity $\sum_{m=0}^{k-1} 2^m$ and allows the π_i summing to unity. To control the amount of variation in the class proportion between the balanced and imbalanced cases, we use a multiplicative coefficient p $(0 < p < 1)$. Finally, the class proportions are defined as:

$$\pi_i(p) = (p-1)/k + p\,2^{k-i}/\left(2^k - 1\right) \qquad (10)$$

The classes are perfectly balanced for $p = 0$ and they become more and more imbalanced as p increases. For instance, a fully imbalanced 5-class datasets will have the following proportions: 0.52, 0.26, 0.13, 0.06 and 0.03, from the 1st to 5th classes, respectively. For each measure, we are now able to plot a discrimination line for each considered value of p. This allows us not only to compare several measures for a given p value but also the different discrimination lines of a single measure as a function of p.

6.2 Error Distribution Sensitivity

We generated matrices for 3 balanced classes ($p = 0$) using the methodology described above. Fig. 1 and 2 show the discrimination lines for class-specific and multiclass measures, respectively. Except for TPR, all discrimination lines are located under the $y = x$ line. So, as expected, the measures favor the case where the errors are uniformly distributed in one class (y series) against the case where the errors affect all the classes (x series).

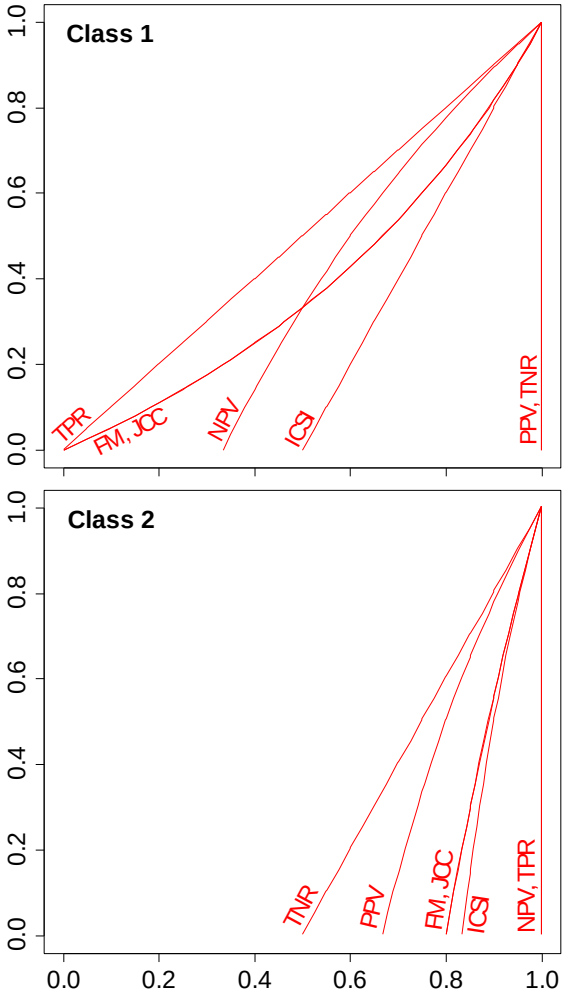

Figure 1: Discrimination lines of all class-specific measures for classes 1 (top) and 2 (bottom), for 3 balanced class ($p = 0$, $k = 3$).

For class-specific measures, we considered first class 1, which is affected by error distribution changes in both series. The discrimination lines are clearly different for all measures. TPR is affected by changes in the distribution of instances only inside the column associated to the considered class. In the case of the first class, these columns are similar on both axes: this explains the $y = x$ discrimination line. The F-measure additionally integrates the PPV value. This explains why it favors the y series matrices. Indeed, the PPV is always greater (or equal) for this series due to the fact errors are present in the first class only. The discrimination line of JCC is exactly similar. NPV does not consider TP, so it is constant for the y series, whereas it decreases for the x series. This is due to the fact more and more errors are added to classes 2 and 3 in these matrices when p increases. This explains why matrices of the y series are largely favored by this measure. PPV and TNR are represented as a vertical line on the extreme right of the plot. According to these measures, the y series matrices are always more accurate. This is due to the fact both measures decrease when the error level increases for the x series (p_{TN} decreases, p_{FP} increases) whereas TNR is constant and PPV decreases less for the y series. Finally, ICSI, which is a linear combination of PPV and TPR, lies in between those measures.

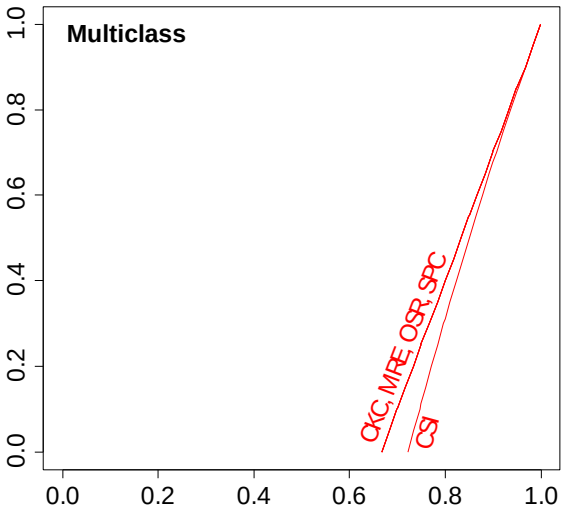

Figure 2: Discrimination lines of all multiclass measures, with $p = 0$ and $k = 3$.

The two other classes undergo similar changes,

so we only report the results for class 2. Unlike class 1, both classes 2 and 3 are affected by errors only in the x series matrices. Consequently, all measures clearly favor the y series matrices, even more than for class 1. The discrimination lines for NPV and TPR take the form of a vertical line on the right of the plot. This is due to the fact both measures decrease only for x series matrices (because of the increase in p_{FN} and p_{TP}). The F-measure and JCC are still not discernable. TNR and PPV favor the y series less than the other measures. This is due to the fact that on the one hand p_{TP} decreases only for the x series matrices, but on the other hand p_{FP} and p_{TN} decrease for both series. Finally, ICSI still lies in between PPV and TPR.

Except for CSI the discrimination lines of multiclass measures are identical. We can conclude that for balanced classes ($p=0$) these measures are equivalent. CSI is more sensitive to the type of error we introduced. Indeed it clearly favors the y series more than the other measures.

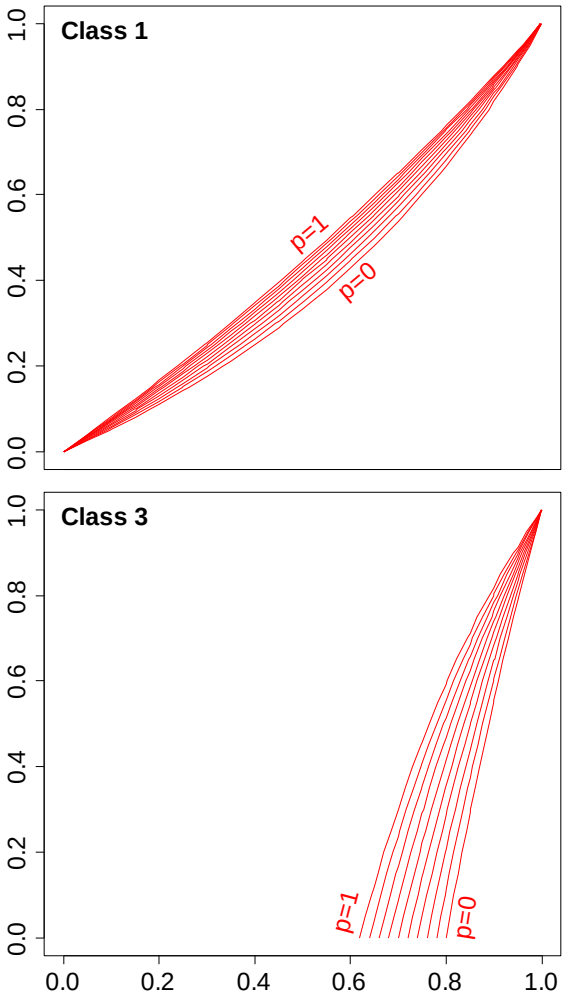

Figure 3: Discrimination lines of the F-measure for classes 1 (top) and 3 (bottom), with $k=3$.

6.3 Class Proportions Distribution Sensitivity

We now study the sensitivity of the on measures on variation in the class proportions. Roughly speaking we observe two different type of behaviors: measures are either sensitive or insensitive to variations in the class proportions distribution. In the first case, the discrimination lines for the different values of p are identical. In the second case, increasing the imbalance leads to lines located on the left of the $p=0$ line. The stronger the imbalance and the more the line is located on the left. This can be explained by the fact the more imbalanced the classes and the more similar the two series of matrices become, dragging the discrimination line closer to the $y=x$ line. Fig. 3 is a typical example of this behavior. It represents the results obtained for the F-measure applied to classes 1 and 3. Note that, like before, JCC and the F-measure have similar discrimination lines.

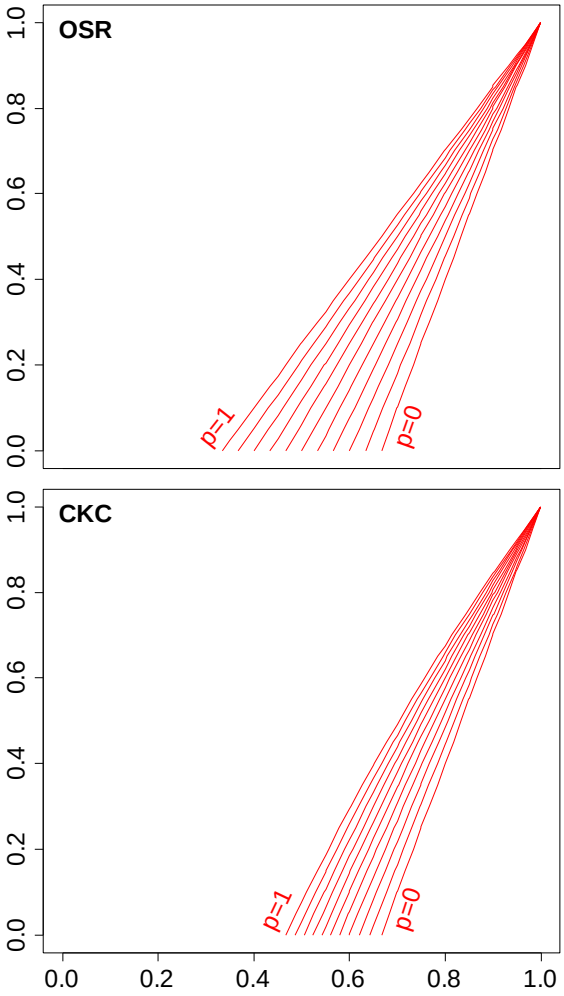

Figure 4: Discrimination lines of OSR (top) and CKC (bottom), with $k=3$.

Other than the F-measure, measures combining two marginal rates (ICSI, JCC) are sensitive to class proportions changes for all classes. This is not

the case for simple marginal rates. TPR and NPV are not sensitive at all for any classes. TNR and PPV present the behavior of measures sensitive to this parameter, but only for classes 2 and 3. As mentioned before, by construction of the considered matrices (the y series has errors only in class 1) they are always higher for the y than the x series, independently of the class proportions. Fig. 4 represents results obtained for the multiclass measures. As previously observed in the balanced case ($p=0$), OSR, SPC and MRE share the same discrimination lines, and this independently of p. CKC was matching them for $p=0$, but this is no more the case for imbalanced classes. The plot for CSI (not represented here) is similar but with tighter discrimination lines, indicating it is less sensitive to proportion changes.

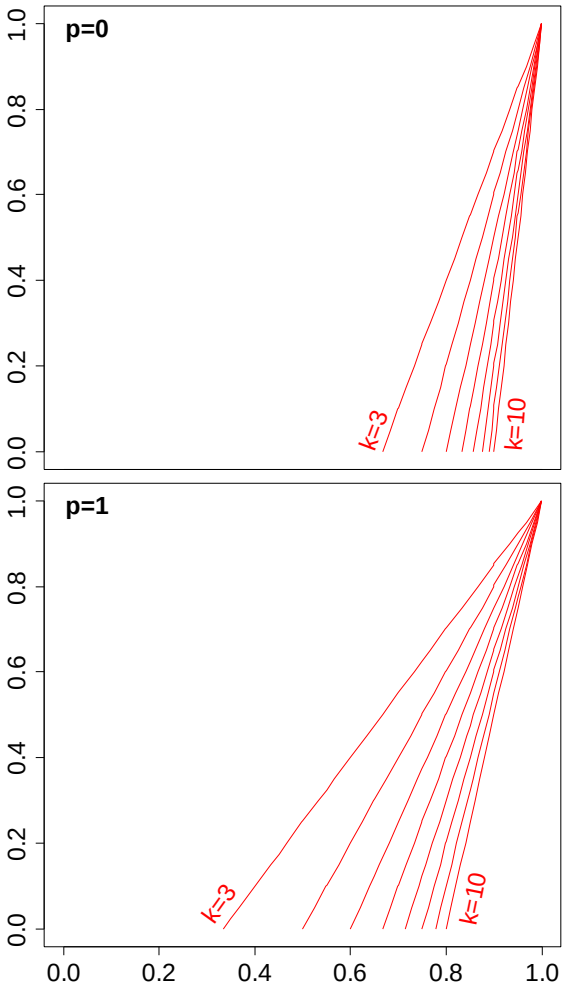

Figure 5: Discrimination lines of OSR for $p=0$ (top) and $p=1$ (bottom), with $k \in [3;10]$.

6.4 Class Number Sensitivity

We finally focus on the effect of the number of classes on the multiclass measures. Fig. 5 shows the results for OSR applied on matrices with size ranging from 3 to 10, for balanced ($p=0$) and imbalanced ($p=1$) cases. All the measures follow the same behavior. Increasing the number of classes strengthens the preference towards the y series matrices. In other words, having more classes gives more importance to the additional errors contained in the x series matrices. The effect is stronger on the imbalanced matrices. In this case, most of the instances are in the first class, which is the only one similar between the two models, so its dilution has a stronger impact on the measured accuracy.

7 DISCUSSION

As shown in the previous sections, measures differ in the way they discriminate different classifiers. However, besides this important aspect, they must also be compared according to several more theoretical traits.

7.1 Class Focus

As illustrated in the previous sections, a measure can assess the accuracy for a specific class or over all classes. The former is adapted to situations where one is interested in a given class, or wants to conduct a class-by-class analysis of the classification results.

It is possible to define an overall measure by combining class-specific values measured for all classes, for example by averaging them, like in CSI. However, even if the considered class-specific measure has a clear meaning, it is difficult to give a straightforward interpretation to the resulting overall measure, other than in terms of combination of the class-specific values. Inversely, it is possible to use an overall measure to assess a given class accuracy, by merging all classes except the considered one [2]. In this case, the interpretation is straightforward though, and depends directly on the overall measure.

One generally uses a class-specific measure in order to distinguish classes in terms of importance. This is not possible with most basic overall measures, because they consider all classes to be equally important. Certain more sophisticated measures allow associating a weight to each class, though [7]. However, a more flexible method makes this built-in feature redundant. It consists in associating a weight to each cell in the confusion matrix, and then using a regular (unweighted) overall measure [27]. This method allows distinguishing, in terms of importance, not only classes, but also any possible case of classification error.

7.2 Functional Relationships

It is interesting to notice that various combinations of two quantities can be sorted by increasing order, independently from the considered

quantities: minimum, harmonic mean, geometric mean, arithmetic mean, quadratic mean, maximum [32]. If the quantities belong to $[0;1]$, we can even put their product at the beginning of the previous list, as the smallest combination. If we consider the presented measures, this means combinations of the same marginal rates have a predefined order for a given classifier. For instance, the sensitivity-precision product will always be smaller than the F-measure (harmonic mean), which in turn will always be smaller than Kulczynski's measure (arithmetic mean). Besides these combinations of TPR and PPV, this also holds for various measures corresponding to combinations of TPR and TNR, not presented here because they are not very popular [20, 33].

More importantly, some of the measures we presented are monotonically related, and this property takes a particular importance in our situation. Indeed, our goal is to sort classifiers depending on their performance on a given data set. If two measures are monotonically related, then the order will be the same for both measures. This makes the F-measure and Jaccard's coefficient similar for classifier comparison, and so are the ICSI and Kulczynski's measure, and of course all measures defined as complements of other measures, such as the FNR. This confirms some of our observations from the previous section: it explains the systematic matching between JCC and the F-measure discrimination lines.

7.3 Range

In the classification context, one can consider two extreme situations: perfect classification (i.e. diagonal confusion matrix) and perfect misclassification (i.e. all diagonal elements are zeros). The former should be associated to the upper bound of the accuracy measure, and the latter to its lower bound.

Measure bounds can either be fixed or depend on the processed data. The former is generally considered as a favorable trait , because it allows comparing values measured on different data sets without having to normalize them for scale matters. Moreover, having fixed bounds makes it easier to give an absolute interpretation of the measured features.

In our case, we want to compare classifiers evaluated on the same data. Furthermore, we are interested in their relative accuracies, i.e. we focus only on their relative differences. Consequently, this trait is not necessary. But it turns out most authors normalized their measures in order to give them fixed bounds (usually $[-1;1]$ or $[0;1]$). Note their exact values are of little importance, since any measure defined on a given interval can easily be rescaled to fit another one. Thus, several supposedly different measures are actually the same, but transposed to different scales [34].

7.4 Interpretation

Our goal is to compare classifiers on a given dataset, for which all we need is the measured accuracies. In other words, numerical values are enough to assess which classifier is the best on the considered data. But identifying the best classifier is useless if we do not know the criteria underlying this discrimination, i.e. if we are not able to interpret the measure. For instance, being the best in terms of PPV or TPR has a totally different meaning, since these measures focus on type I and II errors, respectively.

Among the measures used in the literature to assess classifiers accuracy, some have been designed analytically, in order to have a clear interpretation (e.g. Jaccard's coefficient [4]). Sometimes, this interpretation is questioned, or different alternatives exist, leading to several related measures (e.g. agreement coefficients). In some other cases, the measure is an *ad hoc* construct, which can be justified by practical constraints or observation, but may lack an actual interpretation (e.g. CSI). Finally, some measures are heterogeneous mixes of other measures, and have no direct meaning (e.g. the combination of OSR and marginal rates described in [35]). They can only be interpreted in terms of the measures forming them, and this is generally considered to be a difficult task.

7.5 Correction for Chance

Correcting measures for chance is still an open debate. First, authors disagree on the necessity of this correction, depending on the application context [7, 27]. In our case, we want to generalize the accuracy measured on a sample to the whole population. In other terms, we want to distinguish the proportion of success the algorithm will be able to reproduce on different data from the lucky guesses made on the testing sample, so this correction seems necessary.

Second, authors disagree on the nature of the correction term, as illustrated in our description of agreement coefficients. We can distinguish two kinds of corrections: those depending only on the true class distribution (e.g. Scott's and Maxwell's) and those depending also on the estimated class distribution (e.g. Cohen's and Türk's). The former is of little practical interest for us, because such a measure is linearly related to the OSR (the correction value being the same for every tested algorithm), and would therefore lead to the same ordering of algorithms. This explains the systematic matching observed between the discrimination lines of these measures in the previous section. The latter correction is more relevant, but there is still concern regarding how chance should be modeled. Indeed, lucky guesses depend completely on the algorithm

behind the considered classifier. In other words, a very specific model would have to be designed for each algorithm in order to efficiently account for chance, which seems difficult or even impossible.

8 CONCLUSION

In this work, we reviewed the main measures used for accuracy assessment, from a specific classification perspective. We consider the case where one wants to compare different classification algorithms by testing them on a given data sample, in order to determine which one will be the best on the sampled population.

We first reviewed and described the most widespread measures, and introduced the notion of discrimination plot to compare their behavior in the context of our specific situation. We considered three factors: changes in the error level, in the class proportions, and in the number of classes. As expected, most measures have a proper way to handle the error factor, although some similarities exist between some of them. The effect of the other factors is more homogeneous: decreasing the number of classes and/or increasing their imbalance tend to lower the importance of the error level for all measures.

We then compared the measure from a more theoretical point of view. In the situation studied here, it turns out several traits of the measures are not relevant to discriminate them. First, all monotonically related measures are similar to us, because they all lead to the same ordering of algorithms. This notably discards a type of chance correction. Second, their range is of little importance, because we are considering relative values. Moreover, a whole subset of measures associating weights to classes can be discarded, because a simpler method allows distinguishing classes in terms of importance while using an unweighted multiclass measure. Concerning chance-correction, it appears it is needed for our purpose; however no existing estimation for chance seems relevant. Finally, complex measures based on the combination of other measures are difficult or impossible to interpret correctly.

Under these conditions, we advise the user to choose the simplest measures, whose interpretation is straightforward. For overall accuracy assessment, the OSR seems to be the most adapted. If the focus has to be made on a specific class, we recommend using both the TPR and PPV, or a meaningful combination such as the F-measure. A weight matrix can be used to specify differences between classes or errors.

We plan to complete this work by focusing on the slightly different case of classifiers with real-valued output. This property allows using additional measures such as the area under the ROC curve and various error measures [20].

9 REFERENCES

[1] S. Koukoulas and G. A. Blackburn: Introducing new indices for accuracy evaluation of classified images representing semi-natural woodland environments, Photogramm Eng Rem S, vol. 67, pp. 499-510 (2001).

[2] L. A. Goodman and W. H. Kruskal: Measures of Association for Cross Classification, J Am Stat Assoc, vol. 49, pp. 732-64 (1954).

[3] J. Cohen: A Coefficient of Agreement for Nominal Scales, Educ Psychol Meas, vol. 20, pp. 37-46 (1960).

[4] P. Jaccard: The distribution of the flora in the alpine zone, New Phytol, vol. 11, pp. 37-50 (1912).

[5] G. M. Foody: Status of land cover classification accuracy assessment, Remote Sens Environ, vol. 80, pp. 185-201 (2002).

[6] M. Sokolova and G. Lapalme: A systematic analysis of performance measures for classification tasks, Information Processing & Management, vol. 45, pp. 427-437 (2009).

[7] S. V. Stehman: Comparing thematic maps based on map value, Int J Remote Sens, vol. 20, pp. 2347-2366 (1999).

[8] S. V. Stehman: Selecting and interpreting measures of thematic classification accuracy, Remote Sens Environ, vol. 62, pp. 77-89 (1997).

[9] I. Guggenmoos-Holzmann: How Reliable Are Chance-Corrected Measures of Agreement, Stat Med, vol. 12, pp. 2191-2205 (1993).

[10] V. Labatut and H. Cherifi: Accuracy Measures for the Comparison of Classifiers, in International Conference on Information Technology Amman, JO (2011).

[11] R. G. Congalton: A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data, Remote Sens Environ, vol. 37, pp. 35-46 (1991).

[12] C. X. Ling, J. Huang, and H. Zhang: AUC: a statistically consistent and more discriminating measure than accuracy, in 18th International Conference on Artificial Intelligence (2003).

[13] P. A. Flach: The geometry of ROC space: understanding machine learning metrics through ROC isometrics, in Twentieth International Conference on Machine Learning (ICML) Washington DC (2003).

[14] A. N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko: On Similarity Indices and Correction for Chance Agreement J Classif, vol. 23, pp. 301-313 (2006).

[15] R. Caruana and A. Niculescu-Mizil: Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria, in International Conference on

Knowledge Discovery and Data Mining Seattle, US-WA (2004).
[16]C. R. Liu, P. Frazier, and L. Kumar: Comparative assessment of the measures of thematic classification accuracy, Remote Sens Environ, vol. 107, pp. 606-616 (2007).
[17]C. Ferri, J. Hernández-Orallo, and R. Modroiu: An experimental comparison of performance measures for classification, Pattern Recognition Letters, vol. 30, pp. 27-38 (2009).
[18]G. Türk: GT index: A measure of the success of predictions, Remote Sens Environ, vol. 8, p. 65−75 (1979).
[19]J. T. Finn: Use of the average mutual information index in evaluating classification error and consistency, Int J Geogr Inf Syst, vol. 7, p. 349−366 (1993).
[20]I. H. Witten and E. Frank: Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed.: Morgan Kaufmann (2005).
[21]L. R. Dice: Measures of the amount of ecologic association between species, Ecology, vol. 26, pp. 297-302 (1945).
[22]P. Villegas, E. Bru, B. Mayayo, L. Carpio, E. Alonso, and V. J. Ruiz: Visual scene classification for image and video home content, in International Workshop on Content-Based Multimedia Indexing (CBMI), pp. 77-84 (2008).
[23]J.-P. Linnartz, T. Kalker, and G. Depovere: Modelling the False Alarm and Missed Detection Rate for Electronic Watermarks, Lecture Notes in Computer Science, vol. 1525/1998, pp. 329-343 (1998).
[24]T. Sørensen: A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons, Biologiske Skrifter / Kongelige Danske Videnskabernes Selskab, vol. 5, pp. 1-34 (1948).
[25]U. Hellden: A test of landsat-2 imagery and digital data for thematic mapping illustrated by an environmental study in northern Kenya, Lund Univ. Nat. Geog. Inst, Lund, Sweden 47 (1980).
[26]N. M. Short: The landsat tutorical workbook—Basics of satellite remote sensing, Goddard Space Flight Center, Greenbelt, MD NASA ref. pub. 1078 (1982).
[27]G. Türk: Map Evaluation and "Chance Correction", Photogramm Eng Rem S, vol. 68, pp. 123-129 (2002).
[28]S. Kulczynski: Die Pflanzenassociationen der Pienenen, Bulletin International de L'Académie Polonaise des Sciences et des lettres, Classe des sciences mathématiques et naturelles, Série B, Supplément II, vol. 2, pp. 57-203 (1927).
[29]W. A. Scott: Reliability of Content Analysis: The Case of Nominal Scale Coding, Public Opin Quart, vol. 19, pp. 321-325 (1955).
[30]A. E. Maxwell: Coefficients of agreement between observers and their interpretation, British Journal of Psychiatry, vol. 130, pp. 79-83 (1977).
[31]L. A. Goodman: The analysis of cross-classified data: independence, quasi-independence, and interaction in contingency tables with and without missing entries, J Am Stat Assoc, vol. 63, pp. 1091-1131 (1968).
[32]P. S. Bullen: Handbook of Means and Their Inequalities, 2nd ed. Dordrecht, NL: Kluwer (2003).
[33]D. V. Cicchetti and A. R. Feinstein: High agreement but low kappa: II. Resolving the paradoxes, J Clin Epidemiol, vol. 43, pp. 551-8 (1990).
[34]F. E. Zegers and J. M. F. ten Berge: A Family of Association Coefficients for Metric Scales, Psychometrika, vol. 50, pp. 17-24 (1985).
[35]T. Fung and E. LeDrew: The determination of optimal threshold levels for change detection using various accuracy indices, Photogramm Eng Rem S, vol. 54, p. 1449−1454 (1988).