Multi-class AUC metrics and weighted alternatives

Ben Van Calster, Vanya Van Belle, George Condous,
Tom Bourne, Dirk Timmerman, Sabine Van Huffel, *Senior Member, IEEE*

***Abstract*— The area under the receiver operating characteristic curve (AUC) is a useful and widely used measure to evaluate the performance of binary and multi-class classification models. However, it does not take into account the exact numerical output of the models, but rather looks at how the output ranks the cases. AUC metrics that incorporate the exact numerical output have been developed for binary classification. In this paper, we try to extend such weighted metrics to the multi-class case. Several metrics are suggested. Using real world data, we investigate intercorrelations between these metrics and demonstrate their use.**

I. Introduction

Performance evaluation for classification models is crucial. An important aspect of this evaluation is the model's discriminative potential. Good binary classification models are well able to separate cases from both classes. For binary classification models with a numerical score output rather than a black-and-white class prediction, a useful metric is the area under the receiver operating characteristic (ROC) curve (AUC) [1], [2], [3], [4]. In many research domains, the AUC is well-known as a tool to evaluate performance of classification models. An ROC curve is constructed by plotting sensitivity (true positive rate) versus 1 minus specificity (false positive rate) by varying the decision threshold over its whole range. For a given threshold, class 1 (C_1) is predicted if the score is larger than or equal to the threshold, and class 2 (C_2) is predicted otherwise. This allows the computation of the true and false positive rates. The AUC summarizes model performance over all possible thresholds. The ROC curve and associated AUC can be inferred using parametric [5] or nonparametric methods [2], with the latter being very popular for example in medical research. In this method, the true ROC curve is approximated by connecting the (true positive rate, false positive rate) points obtained on the available data. The trapezoidal rule is used to determine the AUC. The resulting area is equivalent to the Mann-Whitney U statistic divided by $N_1 * N_2$, where N_1 and N_2 are the number of cases from C_1 and C_2, respectively. The AUC can be interpreted as the probability to correctly identify the C_1 case when confronted with a randomly selected case from each class.

Apart from its interpretation, the nonparametric AUC has the advantages to have no underlying assumptions regarding the distribution of the scores in both classes, and to be insensitive to varying class distributions. The AUC is superior to the often used misclassification rate for the evaluation of model performance [6], [7]. Moreover, the misclassification rate requires the use of a single threshold, which is often not desirable. Other performance measures such as the cross-entropy or Brier score, for example, evaluate the class probabilities by measuring how well they approach the true class. However, they do not evaluate discriminatory power. A drawback of the AUC is that, despite its representation of the degree of overlap between score distributions for both classes, it largely ignores the actual scores. The scores are merely used to rank the cases. Nevertheless, these scores may convey useful information. Let us mention an example given in [8], using a sample consisting of 3 C_1 and 3 C_2 cases. Assume that model A gives scores 1.0, 0.7, and 0.6 for the C_1 cases, and 0.5, 0.4, and 0.0 for the C_2 cases. Model B gives 1.0, 0.9, 0.5, and 0.6, 0.2, 0.0, respectively. Even though model A perfectly separated the cases from both classes, it is pointed out that model B may still be preferred as its ranking appears less sensitive to drift of the scores [8]. To our knowledge, there are two AUC metrics that incorporate actual scores [9], [10]. In the present work, we suggest extensions of these metrics to multi-class problems. This would provide us with weighted multi-class AUC-based metrics that go beyond traditional multi-class AUC extensions [11], [12].

Hereafter, we will assume that the model outputs the predicted probability for C_1 (binary classification), or the predicted probabilities for all K classes (multi-class classification). Methods exist to transform other scores (such as those obtained with support vector machines) to probability estimates [13], [14], [15], [16]. If well calibrated [17], [18], the actual probability values can be very interesting, for example in medical decision making problems where they can help to optimize clinicians' treatment decisions.

The paper is organized as follows. In the next section, we describe the two approaches to obtain weighted AUCs for binary classification. Section 3 presents our extensions of these metrics to the multi-class case. Sections 4 and 5 provide experimental results using these extensions. Section 6 concludes the paper.

II. Weighted AUC metrics

A. Nonparametric AUC

The nonparametric AUC is interpreted as the probability that the model correctly identifies the C_1 case from a randomly selected (C_1, C_2)-pair of cases. Let p_{1n} and p_{2n}

Ben Van Calster, Vanya Van Belle, Sabine Van Huffel: Dept of Electrical Engineering (ESAT-SISTA), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. Phone: +32 (0) 16321925; fax: +32 (0) 16321970; email: bvancals@esat.kuleuven.be. George Condous: Nepean Hospital, University of Sydney, Australia. Tom Bourne: St Georges Hospital, University of London, UK. Dirk Timmerman: University Hospitals, Katholieke Universiteit Leuven, Belgium.

978-1-4244-1821-3/08/$25.00©2008 IEEE

represent the probability of C_1 and C_2 for members of C_1, $n = 1, \ldots, N_1$; p_{1m} and p_{2m} the probability of C_1 and C_2 for members of C_2, $m = 1, \ldots, N_2$; $\mathbf{p}_n = [p_{1n},\ p_{2n}]$ the estimated probability vector for members of C_1; and $\mathbf{p}_m = [p_{1m},\ p_{2m}]$ the estimated probability vector for members of C_2. Then, the AUC can be written as

$$\text{AUC} = \frac{1}{N_1 N_2} \sum_{n=1}^{N_1} \sum_{m=1}^{N_2} c(\mathbf{p}_n, \mathbf{p}_m), \tag{1}$$

where

$$c(\mathbf{p}_n, \mathbf{p}_m) = \begin{cases} 1 & \text{if } p_{1n} > p_{1m} \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

B. Probabilistic AUC (pAUC) [9]

Building upon the above-mentioned interpretation of the AUC, a probabilistic AUC is defined in [9] to estimate the score difference for a randomly selected (C_1, C_2)-pair of cases (pAUC). The unbiased estimator they use is the difference between the average probability of C_1 for C_1 cases and C_2 cases. To obtain a metric on the same scale as the AUC, this difference is linearly transformed:

$$\text{pAUC} = 0.5 + \frac{1}{2}\left(\frac{\sum_{n=1}^{N_1} p_{1n}}{N_1} - \frac{\sum_{m=1}^{N_2} p_{1m}}{N_2} \right). \tag{3}$$

Reformulating this expression to render it consistent with the expression of the standard AUC, one obtains

$$\text{pAUC} = 0.5 + \frac{1}{2N_1 N_2} \sum_{n=1}^{N_1} \sum_{m=1}^{N_2} (p_{1n} - p_{1m}). \tag{4}$$

Next, a probabilistic ROC curve (pROC) was derived for which the area under the curve equals pAUC. The pROC curve is obtained by constructing intervals of width d around the estimated probability of C_1, based on the idea that this probability is only approximately equal to the true probability. For a given threshold, the true positive rate is computed by checking for each C_1 case what part of the interval d around p_{1n} exceeds the cut-off. The sum over all N_1 cases divided by dN_1 yields the true positive rate. The false positive rate is computed analogously. The interval length d can be found numerically by running through a grid of possible values until the value is found for which the area under the pROC curve equals pAUC. Note that the solution may not be unique. In this case, the smallest interval length is chosen. The resulting pROC curve is a smoothed curve that evolves towards the original ROC curve for increasing sample size.

C. Scored AUC (sAUC) [10]

Another extension of the original AUC is the scored AUC [10] (sAUC). It is expressed as a weighted version of the original AUC, in which correctly classified pairs are weighted by $p_{1n} - p_{1m}$:

$$\text{sAUC} = \frac{1}{N_1 N_2} \sum_{n=1}^{N_1} \sum_{m=1}^{N_2} (p_{1n} - p_{1m}) c(\mathbf{p}_n, \mathbf{p}_m), \tag{5}$$

or, equivalently,

$$\text{sAUC} = \frac{1}{N_1 N_2} \left(\sum_{m=1}^{N_2} \sum_{p_{1n} > p_{1m}} p_{1n} - \sum_{n=1}^{N_1} \sum_{p_{1m} < p_{1n}} p_{1m} \right). \tag{6}$$

Recently, the sAUC was interpreted as the area under the scored ROC (sROC) curve [8]. The sROC curve plots a margin-based AUC measure versus τ, where τ is a number between 0 and 1 that is subtracted from the probability of C_+ cases. The sAUC reflects the stability of the standard AUC with increasing margin τ [8].

III. Weighted multi-class AUC metrics

In K-class classification, the model outputs K estimated class probabilities p_k for each data point, with $0 \leq p_k \leq 1$, $\sum_{k=1}^{K} p_k = 1$. A data point can thus be represented in K-dimensional space using its estimated class probability vector $\mathbf{p} = (p_1, \ldots, p_K)$. In fact, a $K-1$-dimensional space would suffice because the K estimated class probabilities sum to one. For example, in a 3-class problem, a data point i with vector (p_{1i}, p_{2i}, p_{3i}) can be represented in 3-dimensional space (Figure 1(a)). Because the class probabilities sum to one, all data points are situated on a 2-dimensional subplane defined by the coordinates $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ (i.e. the class corners). Further, because probabilities lie between 0 and 1, all data points are located on the equilateral triangle defined by the class corners (Figure 1(b)). Alternatively, the estimated probabilities for one of the classes can be ignored, and the data points (p_{1i}, p_{2i}) can be plotted in two-dimensional space (as a projection of the three-dimensional space). This time, all points lie within an isosceles right triangle defined by the origin, $(1,0)$, and $(0,1)$ (Figure 1(c)).

A. Standard multi-class AUC metrics

Mossman [11], [19] has extended the ROC curve to three-class problems. The three-class extension of the ROC curve plots the true positive rates for the three classes (TPR triple). By varying the decision function to map the probabilistic output (p_{1i}, p_{2i}, p_{3i}) into a crisp class prediction, many TPR triples are obtained. Plotting these triples in three-dimensional space results in a ROC surface. The volume under the surface (VUS) can be interpreted as the probability that a randomly selected (C_1, C_2, C_3)-triplet of cases is correctly classified by the model. It is not straightforward, however, to determine what it means to correctly classify a triplet. One rule that is also used in [19] states that a triplet

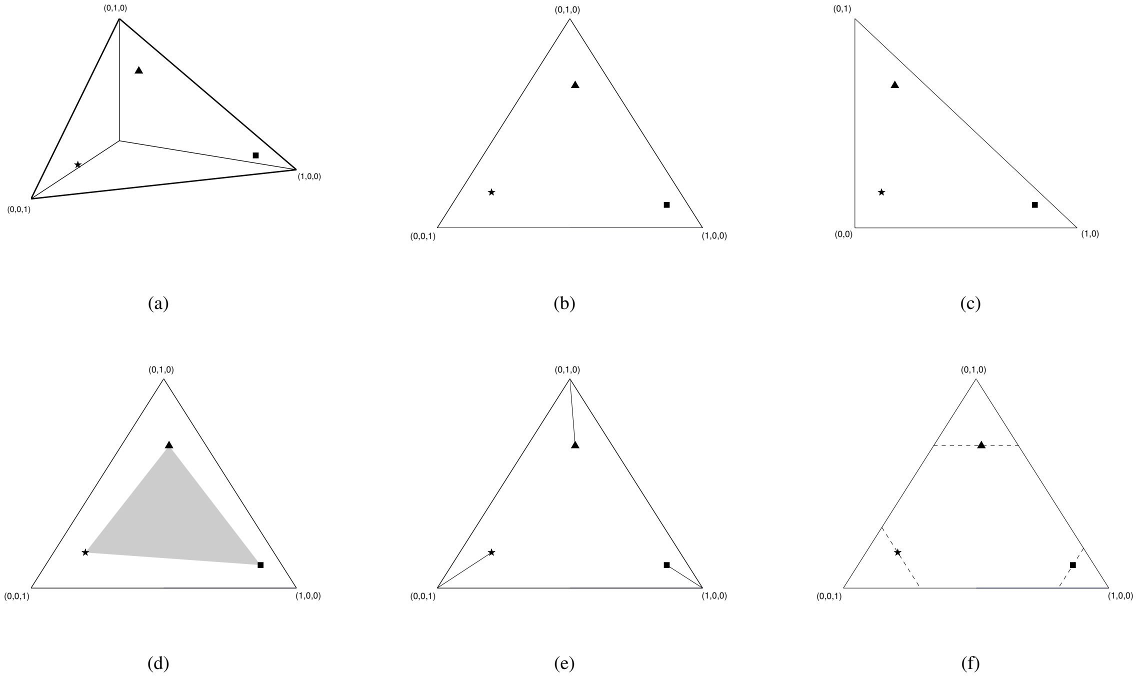


Fig. 1. Representation of predicted probabilities for a three-class problem. The square refers to a C_1 case, the triangle to a C_2 case, and the star to a C_3 case.

is correctly rated if the total length of the lines connecting each case with its true class corner in the equilateral triangle is smaller than any other total length obtained by connecting the three cases to the three class corners. Let N_k denote the number of cases in class k, and let $\mathbf{p}_n$, $\mathbf{p}_m$, and $\mathbf{p}_q$ represent the class probability vectors for cases from C_1, C_2, and C_3, respectively. Then, the VUS is estimated as

$$\text{VUS} = \frac{1}{N_1 N_2 N_3} \sum_{n=1}^{N_1} \sum_{m=1}^{N_2} \sum_{q=1}^{N_3} C(\mathbf{p}_n, \mathbf{p}_m, \mathbf{p}_q), \tag{7}$$

where

$$C(\mathbf{p}_n, \mathbf{p}_m, \mathbf{p}_q) = \begin{cases} 1 & \text{if correctly classified} \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

This approach to the computation of the VUS can be used for any K-class problem. The rule that is used to define correct classification is rather technical, however, and in our opinion not always optimal. Another rule is suggested in section III-D.

Hand and Till [12] have suggested a simple extension of binary AUC values to K-class AUCs for any value of K. They compute AUC$_{kr}$, the AUC for each pair of classes k and r ($k \neq r$) using p_k. Note that AUC$_{kr} \neq$ AUC$_{rk}$, except for binary problems. Their index M is computed as

$$M = \frac{1}{K(K-1)} \sum_{k \neq r} AUC_{kr}. \tag{9}$$

Provost and Domingos [20] applied a similar metric based on the K 1-versus-all AUCs (AUCs for each class versus all other classes). Their index equals the average of the 1-versus-all AUCs weighted by the prevalence of a class in the data set.

B. Area of the triangle

An attractive measure for three-class problems starts with the computation of the average class probability vector for each class k, $\mathbf{p}_k$:

$$\mathbf{p}_k = \frac{1}{N_k} \sum_{n_k=1}^{N_k} \mathbf{p}_{n_k}, \; k = 1, \; 2, \; 3. \tag{10}$$

The average probability vectors for each class can be plotted in three-dimensional space, and the area of the triangle spanned by the three points is computed and rescaled to yield a number between 0 and 1 (AOT). Since the three class probabilities sum to one, this triangle is located within the equilateral triangle mentioned previously (Figure 1(b)). The more a model tears the classes apart in terms of the predicted probabilities, the larger AOT. The area of the triangle cannot be seen as the average area over all possible triplets, since the area does not contain information about the exact location of the data points in three-dimensional space. A combination in which the probability vector of the C_k case lies close to the C_r class corner and vice versa will yield a similar area compared to a similar combination where the C_k and C_r cases lie close to their own class corner. Thus, a high AOT is obtained when the model tears the classes apart, regardless of the class corner to which each average probability vector is pulled. In practice, however, useful models will on average pull classes toward the correct class corner.

Note that the area of the triangle is proportional to the area of the triangle when projecting it on to any possible two-dimensional space defined by two of the three classes (as in Figure 1(d)). It does not matter which projection is chosen. The maximum area of the original triangle within the bounds of the equilateral triangle is $\frac{\sqrt{3}}{2}$, so the AOT is obtained by dividing the area of the triangle by $\frac{\sqrt{3}}{2}$.

C. Total length between probability coordinates and class corners

The rule to determine whether K randomly chosen cases from the K classes are correctly classified, presented in section III-A, can also be exploited to incorporate probabilities. A possibility is to compute the total length of the lines connecting the K average probability vectors with their class corners (TL_0, cf. Figure 1(e)). If a classifier is perfect, TL_0 is 0. If we want the index to vary between 0 to 1 (with 1 indicating a perfect classifier), we could transform TL_0 to obtain $\text{TL} = 1 - \frac{\text{TL}_0}{z}$ with z as the maximum of TL_0. The maximum of TL_0 for a K-class problem is $K\sqrt{2}$, and is obtained when all $\mathbf{p}_k$ are located exactly on a wrong class corner.

D. Weighted VUS metrics

In the sense of sAUC, we could define a weighted version of the VUS (wVUS). For a three-class problem, this metric weights correctly classified triplets using the total length of the lines connecting each case with its class corner. Let l_n, l_m, and l_q represent the length of the lines for cases from C_1, C_2, and C_3, respectively. The wVUS can be written as

$$\text{wVUS} = \frac{1}{N_1 N_2 N_3} \sum_{n=1}^{N_1} \sum_{m=1}^{N_2} \sum_{q=1}^{N_3} W_{nmq} C(\mathbf{p}_n, \mathbf{p}_m, \mathbf{p}_q), \quad (11)$$

where

$$W_{nmq} = \left(1 - \frac{l_n + l_m + l_q}{z}\right). \quad (12)$$

Using the idea of the AOT, we suggest another rule to define correct classification when confronted with one case from each class. This results in an alternative multi-class AUC index (VUS_2), for which a weighted version can be derived. The easier to understand rule for correct classification is that the highest probability for C_k should be obtained for the C_k case. It is natural to think that the case with highest p_k will probably be the C_k case. Using the suggested rule, it is easy to see whether the classifier works well when given one case from each class. Figure 1(f) illustrates the new rule for a three-class problem. The dashed lines through each case represent the line of equal probability for the true class of that case. If any case crosses the line of another case, the triplet is not correctly classified. For a three-class problem, the weighted version (wVUS_2) weights a correctly classified triplet with the AOT. This is justified, since correct classification now means that each probability vector is closest to its own class corner such that no two cases are switched.

It is always true that $\text{VUS}_2 \leq \text{VUS}$. Since each probability vector is closest to its own class corner if the newly suggested rule applies, the sum of lines connecting the K cases with their class corner is always minimal. Thus, there cannot be situations where the VUS_2 rule applies but the VUS rule does not.

E. Combination of M and weighted AUC metrics

A simple generalization of binary weighted AUC metrics [9], [10] for any K-class problem could be to combine them in in the sense of Hand and Till's M index [12]:

$$M_p = \frac{1}{K(K-1)} \sum_{k \neq r} pAUC_{kr}, \quad (13)$$

or

$$M_s = \frac{1}{K(K-1)} \sum_{k \neq r} sAUC_{kr}. \quad (14)$$

IV. A REAL WORLD APPLICATION

We will now demonstrate the indices to a three-class problem dealing with pregnancies of unknown location (PULs). In such pregnancies, there is a positive pregnancy test with no signs of intra- or extra-uterine pregnancy on ultrasound, and no remnants of miscarriage [21]. Such pregnancies can evolve in three ways. The majority represent failing PULs, which represent spontaneously resolving ectopic or intra-uterine pregnancies that are never seen on ultrasound. Another large group of PULs are early intra-uterine pregnancies (IUPs). The third and smallest group of PULs represent ectopic pregnancies. The early prediction of the PUL outcome is important due to the risk of rupture for ectopic pregnancies. Ectopic pregnancy still is an important cause of maternal death, in particular during the first trimester of pregnancy. Our PUL data set consists of 856 pregnancies collected at St Georges Hospital (London, UK), for which 12 measurements were recorded.

We performed input selection and used six algorithms to derive a multi-class model for PULs: a multi-class logistic regression model (MLR; with checks for linearity in the logit and interactions), a Bayesian multi-layer perceptron (BMLP), a Bayesian perceptron (BPER; similar to a regularized MLR, but then without the checks mentioned for the MLR model), 1-versus-1 Bayesian least squares support vector machines using a linear or an RBF kernel (BLSSVMl and BLSSVMr;

the binary models are combined using pairwise coupling to yield multi-class probabilities), and a multi-class kernel logistic regression algorithm using an RBF kernel (MKLR). To compare the algorithms, they were applied to 100 random stratified splits of the data into a training (60%) and a test set (40%). Each algorithm was trained on the training set and was evaluated on the test set using the metrics described above. We summarized the 100 test set values for each metric by taking the mean. The performance of the six algorithms is compared by computing their average rank over the 100 test set performances (AR), and by computing the proportion of times they were ranked on top (P_1, along the ideas of [22]). More important at this moment, however, was our aim to compute correlations between the various metrics using the 100 test set values. The correlations are computed for all algorithms and were averaged. This allowed us to check how well different metrics are interrelated. It is interesting to investigate the associations between weighted and unweighted metrics, and between different types of weighted metrics.

V. Results

Table 1 shows the Spearman correlations between the multi-class AUC metrics. The Pearson correlations were nearly identical. The unweighted metrics have intercorrelations between 0.90 and 0.97, with the highest correlation between the M index and VUS$_2$. Except for wVUS, the weighted metrics are also highly related, with intercorrelations between 0.92 and 0.99. The intercorrelations between these weighted metrics and the unweighted metrics vary between 0.41 and 0.65, indicating that incorporating the predicted probabilities can give a different picture when performing model selection or evaluation. The correlations obtained for wVUS deserve special attention. This weighted metric correlates between 0.74 and 0.82 with the other weighted metrics, but correlates between 0.88 and 0.89 with the unweighted metrics. The problem with wVUS may be either the rule it uses to determine correct classification, or the weights it uses. Extra analyses showed that another measure, in which correct classification was determined according to (w)VUS$_2$ but weighting was according to wVUS, behaved like wVUS with respect to its correlations with other metrics. Thus, the problem of wVUS appears to be the weights it uses.

TABLE I

Spearman correlation matrix

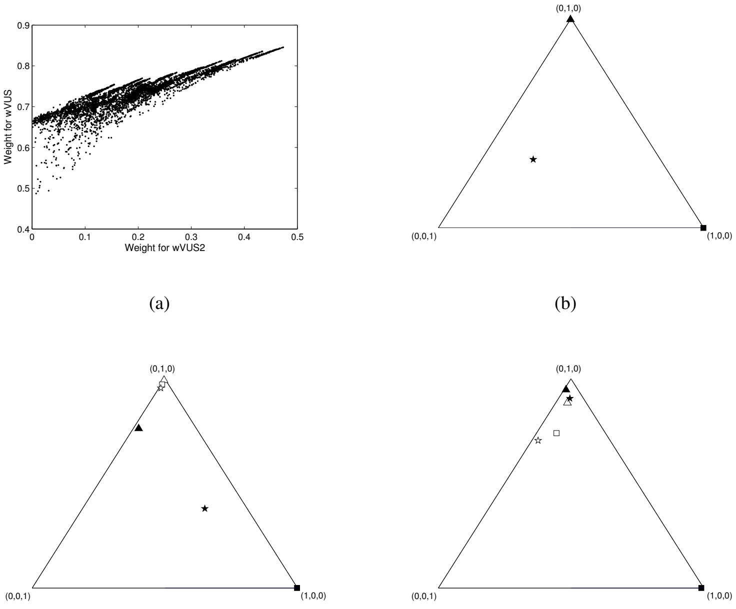
	M	VUS	VUS$_2$	AOT	TL	M_p	M_s	wVUS
M	1							
VUS	.94	1						
VUS$_2$	.97	.90	1					
AOT	.53	.48	.60	1				
TL	.56	.50	.65	.93	1			
M_p	.56	.51	.65	.94	.99	1		
M_s	.49	.44	.58	.95	.98	.99	1	
wVUS	.88	.88	.89	.79	.81	.82	.77	1
wVUS$_2$	.47	.41	.54	.99	.92	.93	.95	.74

We compared wVUS and wVUS$_2$ using a few real examples. We chose one of the 100 runs for the MKLR model. For all $N_1N_2N_3 = 184 * 132 * 26 = 631,488$ triplets in the test set, we determined whether they were correctly classified or not according to both metrics, and computed the weight that would be assigned if correctly classified. Combinations were correctly classified for both metrics in 77.9% of the occasions, and both types of weights were correlated 0.90 (Figure 2(a) presents a scatter plot for 1% of these combinations). Combinations were correctly classified only for wVUS in 9.5% of the occasions, and the correlation between weight types was only 0.25. As a reference, Figure 2(b) shows one of the best triplets. The instance from class 1 and 2 had near perfect predicted probability for the true class and near zero probability for the other classes. Given the low prevalence of class 3, the predicted class probabilities for the C_3 case were very good as well. Figure 2(c) presents two triplets that were correctly classified for wVUS only, with one triplet (black triplet) having markedly better weight for wVUS than the other (white triplet). Even though the rule for correct classification does not seem to be problematic as such, both triplets show that the limit for correct classification is less strict for wVUS (or VUS) than for wVUS$_2$ or (VUS$_2$). The black triplet does not look that bad, but there is no clear agreement as to what case should be the C_3 case. Actually, the C_1 case has a higher probability to belong to C_3 than the C_3 case. The white triplet is not good at all. Figure 2(a) shows that triplets with high weight for wVUS$_2$ also have a high weight for wVUS. When wVUS$_2$ gives a low weight, however, the weight for wVUS can vary. Figure 2(d) shows two triplets that are correctly classified using both rules, but for which wVUS$_2$ gives low weight while wVUS gives clearly lower weight for one triplet (white) relative to the other (black). To us, wVUS$_2$ appears more correct because the three cases are not clearly separated from each other in both triplets.

Table 2 presents the results for five metrics, two unweighted (M and VUS$_2$) and three weighted (AOT, M_p, and wVUS$_2$) ones. The unweighted metrics suggest that MLR is the best algorithm, followed by MKLR in second and BLSSVMr in third. The probability that MLR is the best of the algorithms is estimated to be more than 1 in 2. The multi-layer and single-layer perceptron models perform worst. When looking at the weighted metrics, however, we observe a stronger separation between algorithms that perform well and those that perform poor. This time, both MLR and BLSSVMr perform best, MKLR is less outstanding. Apparently BLSSVMr produced better probabilities than MKLR. Note that the AOT is rather small. This reflects the low prevalence of ectopics.

VI. Discussion and conclusions

Inspired by the work on weighted AUC metrics for binary classification, weighted AUC metrics for the multi-class case were suggested in this paper. These metrics are based on the average predicted class probabilities for cases belonging



(a)

(b)

(c)

(d)

Fig. 2. Comparison of wVUS and wVUS$_2$. In the triangle plots, the square refers to a C_1 case, the triangle to a C_2 case, and the star to a C_3 case.

TABLE II

AVERAGE TEST SET PERFORMANCE OF THE SIX ALGORITHMS: THE AVERAGE PERFORMANCE, AND AR AND P_1 (SMALL FONT).

	M	VUS$_2$	AOT	M_p	wVUS$_2$
MLR	.954	.829	.288	.801	.296
	1.9 − .55	1.8 − .65	1.5 − .57	1.6 − .45	1.5 − .55
BMLP	.943	.779	.123	.706	.125
	4.2 − .08	5.0 − .01	5.7 − .00	6.0 − .00	5.7 − .00
BPER	.941	.781	.137	.744	.140
	5.0 − .00	5.2 − .00	5.3 − .00	5.0 − .00	5.3 − .00
BLSSVMl	.946	.808	.233	.787	.241
	4.1 − .02	3.5 − .07	3.4 − .00	3.2 − .00	3.4 − .00
BLSSVMr	.948	.815	.280	.802	.290
	3.5 − .07	2.9 − .16	1.7 − .43	1.5 − .55	1.6 − .45
MKLR	.952	.818	.227	.781	.233
	2.4 − .29	2.6 − .11	3.5 − .00	3.7 − .00	3.5 − .00

to the same class (AOT, TL), on weighting only correctly classified combinations of cases from each class (wVUS, wVUS$_2$), or on the simple combination of binary weighted AUC metrics (M_p, M_s). We presented correlations between these metrics, obtained from the application of six algorithms to 100 random train-test splits of a data set on pregnancies of unknown location. The correlations suggest that unweighted metrics are highly correlated, and that all but one weighted metrics are highly correlated. The wVUS metric appeared to correlate less strong with other weighted metrics than with unweighted metrics. Unweighted and weighted metrics are moderately correlated. Since metrics such as TL, M_p and M_s are easy to compute, they are attractive for problems where $K >> 2$. Metrics such as AOT or wVUS$_2$ are interesting and visually attractive when $K = 3$.

We demonstrated the metrics by applying them for the sake of model comparison, and observed that both types of metrics did not favor the same set of models. Regarding model comparison, it is interesting to check for differences in results based on unweighted versus weighted metrics. We acknowledge that another highly useful application area for weighted multi-class AUC metrics is that of model selection. An experimental study on this issue would form an important topic for further research. At this moment, we can point at experimental studies where the use of the sAUC for model selection in binary classification problems yielded promising results [8], [10]. Increasing attention is given to methods that construct classifiers that try to maximize AUC rather than minimize some error function. Weighted AUC metrics can build upon this.

ACKNOWLEDGMENT

Research supported by Research Council KUL: GOA-AMBioRICS, CoE EF/05/006 Optimization in Engineering; Belgian Federal Science Policy Office IUAP P6/04 ('Dynamical systems, control and optimization', 2007-2011); EU: BIOPATTERN (FP6-2002-IST 508803), ETUMOUR (FP6-2002- LIFESCIHEALTH 503094).

REFERENCES

[1] T.A. Lasko, J.G. Bhagwat, K.H. Zou, and L. Ohno-Machado. The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics* 38 (2005) 404–415.

[2] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143 (1982) 29–36.

[3] F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning* 42 (2001) 203–231.

[4] P.A. Flach and S. Wu. Repairing concavities in ROC curves. In L.P. Kaelbling and A. Saffiotti (eds), *Proc. 19th International Joint Conference on Artificial Intelligence* (2005) 702–707. Springer, Berlin.

[5] C.E. Metz, B.A. Herman, and J.-H. Shen. Maximum likelihood estimation of receiver operating characteristic ROC curves from continuously-distributed data. *Statistics in Medicine* 17 (1998) 1033–1053.

[6] J. Huang and C.X. Ling. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 17 (2005) 299–310.

[7] F. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *Proc. of the 15th International Conference on Machine Learning* (1998) 445–453.

[8] S. Wu, P. Flach, and C. Ferri. An improved model selection heuristic for AUC. In J.N. Kok, J. Koronacki, R. López de Mántaras, S. Matwin, D. Mladenic, and A. Skowron (eds), *Machine Learning: ECML2007* (2007) 478-489. Lecture Notes in Computer Science 4701, Springer, Berlin.

[9] C. Ferri, P. Flach, J. Hernández-Orallo, and A. Senad. Modifying ROC curves to incorporate predicted probabilities. In *Proc. of the ICML2005 Workshop on ROC analysis in Machine Learning* (2005).

[10] S. Wu and P. Flach. A scored AUC metric for classifier evaluation and selection. In *Proc. of the ICML2005 Workshop on ROC analysis in Machine Learning* (2005).

[11] D. Mossman. Three-way ROCs. *Medical Decision Making* 19 (1999) 78–89.

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

[13] H.-T. Lin, C.-J. Lin, and R.C. Weng. A note on Platt's probabilistic outputs for support vector machines. Technical Report, Department of Computer Science, National Taiwan University (2003).

[14] B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2002) 694–699.

[15] S. Rüping. Robust probabilistic calibration. In J. Fürnkranz (ed), *Machine Learning: ECML2006* (2006) 743–750. Lecture Notes in Computer Science 4212, Springer, Berlin.

[16] A. Niculescu-Mizil and R. Caruana. Predicting good probabilities with supervised learning. In *Proc. 21st International Conference on Machine Learning* (2005) 625–632. Omnipress, Madison.

[17] S. Dreiseitl and L. Ohno-Machado. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics* 35 (2002) 352–359.

[18] A.F.G. Taktak, A. Eleuteri, S.P. Lake, and A.C. Fisher. Evaluation of prognostic models: discrimination and calibration performance. In *Proc. 3rd International Conference on Computational Intelligence in Medicine and Healthcare* (2007).

[19] S. Dreiseitl, L. Ohno-Machado, and M. Binder. Comparing three-class diagnostic tests by three-way ROC analysis. *Medical Decision Making* 20 (2000) 323–331.

[20] F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning* 52 (2003) 199–215.

[21] G. Condous, D. Timmerman, S. Goldstein, L. Valentin, D. Jurkovic, and T. Bourne. Pregnancies of unknown location: consensus statement. *Ultrasound in Obstetrics and Gynecology* 28 (2006) 121–122.

[22] M.S. Pepe, G. Longton, G.L. Anderson, and M. Schummer. Selecting differentially expressed genes from microarray experiments. *Biometrics* 59 (2003) 133–142.