DOI: https://doi.org/10.46354/i3m.2019.emss.046

TOWARDS MORE EFFICIENT MULTICLASS AUC COMPUTATIONS

S. Dreiseitl[(a)]

[(a)]University of Applied Sciences Upper Austria, Campus Hagenberg, Austria
[(a)]stephan.dreiseitl@fh-hagenberg.at

ABSTRACT

The area under the receiver operating characteristics curve (AUC) can be used to assess the discriminatory power of a dichotomous classifier model. Extending this measure to more than two classes is not obvious, and a number of variants have been proposed in the literature. We investigate a heuristic approximation to a method that generalizes the notion of probabilities being correctly ordered, which is equivalent to AUC, to an arbitrary number of classes. While the exact method is computationally complex, we propose a much simpler heuristic that is linear in the number of classes for every combination of data points. Using one artificial and one real-world data set, we demonstrate empirically that this simple heuristic can provide good approximations to the exact method, with Pearson correlation coefficients between 0.85 and 0.998 across all data sets.

Keywords: multiclass AUC, multiclass ROC, classifier performance assessment

1. INTRODUCTION

Receiver operating characteristics (*ROC*) *curves* have a long and storied history as tools for evaluating classification performance of predictive models, in particular in the application domain of biomedicine (Metz 1978; Lusted 1978; Lasko et al. 2005). At around the new millenium, machine learning researchers also discovered the usefulness of ROC curves for the analysis of their models (Flach 2003; Fürnkranz and Flach 2005; Fawcett 2006; Davis and Goadrich 2006).

ROC curves provide a graphical visualization of false positive rate (1− specificity) vs. true positive rate (sensitivity) across a spectrum of thresholds for any two-class discriminatory task based on a linearly ordered measurement. The *area under ROC curve* (*AUC*) is therefore indicative of how well two classes can be distinguished from one another, regardless of the chosen threshold (Hanley and McNeil 1982, 1983; Bradley 1997). AUC can be shown to be equivalent to the c-index $P(X < Y)$ (Bamber 1975), the probability that two randomly chosen measurements X and Y from two classes are correctly arranged on a linear scale. For a classifier that outputs posterior class membership probabilities, AUC is thus an alternative to accuracy, which is generally thresholded at 0.5. A classifier that is not well calibrated (possibly owing to changes in class distributions between training and test set) may therefore achieve an accuracy of only 50%, although its AUC may be close to 100%. This situation is shown graphically in Figure 1.

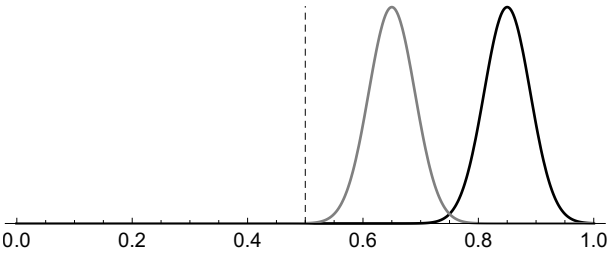


Figure 1: Hypothetical distribution of classifier outputs for a two-class problem. The left normal distribution represents outputs for true class 0, and the right normal distribution represents outputs for true class 1. If these outputs are thresholded at 0.5 (smaller values are considered to belong to class 0, larger values to belong to class 1), all the cases of true class 0 are misclassified. Nevertheless, the classifier achieves almost perfect discrimination between the two classes, with a corresponding AUC value of close to 1.

Deep learning problem settings, architectures and training techniques have brought a renewed interest in extending the use of AUC as a discriminatory measure to the multiclass case. There has been some work on constructing and interpreting ROC curves in multiclass settings (Edwards et al. 2004, 2005; He et al. 2006; He and Frey 2006). Previous work on extending the discriminatory measure AUC to N classes has focused either on combining multiple AUCs from one-vs-all classifiers (Hand and Till 2001; Landgrebe and Duin 2007), or on directly generalizing the equivalence of AUC to c-index and the underlying notion of what it means for class-membership probabilities to be "correctly ordered". Here, we will pursue this second direction of reasoning.

The obvious starting point is to consider three classes. Mossman (1999) advocated a geometric argument to generalize the notion of a changing threshold; a more involved derivation of a similar argument is given by He and Frey (2008). Dreiseitl et al. (2000) argued for a broader interpretation of the notion of "correctly ordered" to the three-class case. This latter idea,

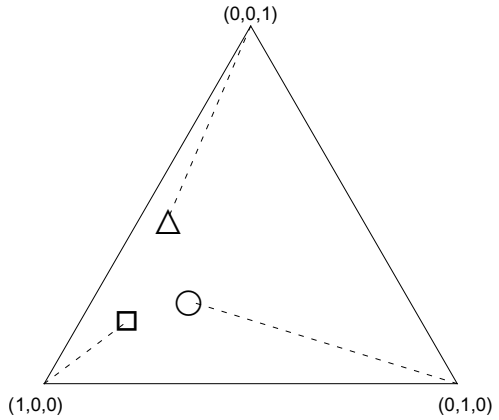


Figure 2: Three probability estimate triplets △ (for an element of class 0), ○ (for an element of class 1) and □ (for an element of class 2) on the triangle $\{(x, y, z) \in \mathbb{R}^3 \,|\, x + y + z = 1\}$. The three triplets are considered correctly ordered if the sum of distances between triplets and true-class corners is smaller than the sum of distances between triplets and all permutations of corners.

which will form the basis for the following argument, rests on viewing a traditional two-class pair of probabilities $p_0 = P(\text{class is } 1 \,|\, \text{case is class } 0)$ and $p_1 = P(\text{class is } 1 \,|\, \text{case is class } 1)$ as correctly ordered if $p_0 < p_1$. For the three-class case, a classifier output is a probability triplet (p_0, p_1, p_2), with the p_i being posterior class-membership probabilities for each of the three classes. One possible way of considering three probability triplets $p_0 = (p_{00}, p_{01}, p_{02})$, $p_1 = (p_{10}, p_{11}, p_{12})$ and $p_2 = (p_{20}, p_{21}, p_{22})$ to be correctly ordered is if the sum of distances of the p_i to their true-class corners of the convex set $\{(x, y, z) \in \mathbb{R}^3 \,|\, x + y + z = 1\}$ is smaller than all other possible distances of triplets to corners. This idea is expressed graphically in Figure 2. Note that this notion of "correctly ordered" depends only on the positioning of the three symbols relative to one another, and not on *where* on the triangle the three symbols are placed, as long as their relative positioning is the same. It is thus possible to transfer the property of AUCs of being invariant to monotonic transformations to the three-class case.
In the three-class case, the equivalent to AUC is the *volume under the surface*, which is calculated as the fraction of all triplet combinations, one each from the three classes c_0, c_1 and c_2, that are correctly ordered in the sense above:

$$\text{VUS}(c_0, c_1, c_2) := \frac{1}{|c_0|\,|c_1|\,|c_2|} \sum_{p_0 \in c_0} \sum_{p_1 \in c_1} \sum_{p_2 \in c_2} \text{co}(p_0, p_1, p_2)\,,$$

where co is a Boolean indicator function that returns 1 iff its arguments are correctly ordered.
Due to the combinatorial nature of the problem, an exact VUS calculation is possible only for small cardinalities of the data sets involved. Approximations, however, are feasible, as sampling theory ensures that the error of an approximation will decrease with the square root of the data set cardinalities.
The derivations above are sufficiently general to be applicable in the $N > 3$-class case. We will use the term *probability vector* to denote the generalization of class membership probability triplets to an arbitrary number of classes. Extending the VUS calculations from three classes to the general N-class case, however, is hampered not only by the exponential growth of having to compare all $n_1 n_2 \cdots n_N$ possible combinations of N-class probability vectors (with the n_i denoting the cardinalities of the data sets from class i), but by the $O(N!)$ complexity of computing all possible distances of probability vectors to corners for every combination of probability vectors. This latter calculation cannot be approximated by sampling, because *all* of the $N! - 1$ other sums of distances have to be smaller than the sums of distances to the "true" corners for a combination of probability vectors to be correctly ordered.
This paper thus addresses the question of how to efficiently substitute for the factorial problem of computing all sums of distances between probability vectors and corners. The approach considered here is based on computing angles to the true-class corners of the estimate space, and only grows as $O(N)$ instead of $O(N!)$. Details of this approach are given in Section 2; experimental results in Section 3 demonstrate its feasibility.

2. METHODS

The approach presented here was inspired by the notion of varying thresholds in ROC curve construction first presented by Mossman (1999), although no explicit thresholds are required for AUC and VUS computations. As a heuristic approximation to N-dimensional probability vectors being correctly ordered (which would require $N!$ distance computations for an N-class problem), we compute only N angles between the lines formed by the centers of mass of N probability vectors, the individual probability vectors, and their corresponding "true" corners. A graphical representation of this idea is shown in Figure 3. As a visual aid, the figure also shows the grey lines obtained by orthogonal projections of the center of mass onto the triangle sides. These lines illustrate that for three classes, an arrangement of points can be considered to be correctly ordered if every point is in its own (correct) portion of the triangle; this translates to a restriction of $|\cos\alpha| \leq \frac{1}{2}$ for the angle α between center of mass, point and corresponding corner. One can show that for the general situation with N classes, this restriction is $|\cos\alpha| \leq \frac{1}{N-1}$, thus approaching an angle of $\alpha = 90°$ in the limit $N \to \infty$.
Preliminary experiments, however, revealed this limit to be too restrictive, as some situations that were correctly ordered in the computationally expensive distances-to-corners sense were not identified as such in the angles sense introduced here. The limit was therefore set to a constant $\alpha \leq 90°$, which gave reasonable results that were computationally cheap approximations to the much

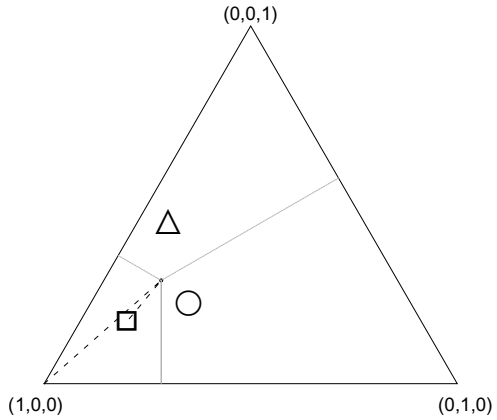


Figure 3: Three probability estimate triplets △ (for an element of class 0), ○ (for an element of class 1) and □ (for an element of class 2) on the triangle $\{(x, y, z) \in \mathbb{R}^3 \mid x + y + z = 1\}$. For clarity, only one line connecting the center of mass of these three triplets to one of the triplets and its corresponding corner are shown. The grey lines connect the center of mass to the triangle sides at right angles.

more expensive distances-to-corners calculations.

3. EXPERIMENTS

We investigated the effect of substituting a heuristic approximation for an exact computation using two data sets, one artificial, and one from the wide range of publically available machine learning data sets.

3.1. Artificial data set

For the artificial data set, we used random variates from multivariate Gaussians, one for each class, with spherical covariance matrices $C = \sigma I$, with I the identity matrix. To obtain probability vectors, we passed these random variates through the softmax function. Changing the distances between the distribution means relative to the σ values allows to increase or decrease VUS values as desired.

As there are two sources of computational complexity in N-class VUS calculation, we initially isolated only the first source, i.e., sampling from the data sets, while keeping the more complex distances-to-corners definition of "correctly ordered". For a three-class VUS value in the intermediate range of around 0.85, Figure 4 shows how increasing the data set size leads to more and more accurate estimates of the VUS value. The distribution of data points in this artificial data set is shown in Figure 5.

Because the proposed approach of substituting the angles calculation introduced in Section 2 for the distances-to-corners definition is only a heuristic approximation, it is of primary interest to evaluate how much results obtained using these two approaches match. For the three-class case, we placed the means of three Gaussians at equal distances from one another, and varied the values of σ to create more or less overlap between the classes. A scatterplot of the VUS values obtained from exact and approximate approaches is shown in Figure 6. One can observe that the heuristic calculation using angles provides a good approximation to the exact, but more computationally complex distances-to-angles VUS values. The Pearson correlation coefficient between both sets of values is 0.998.

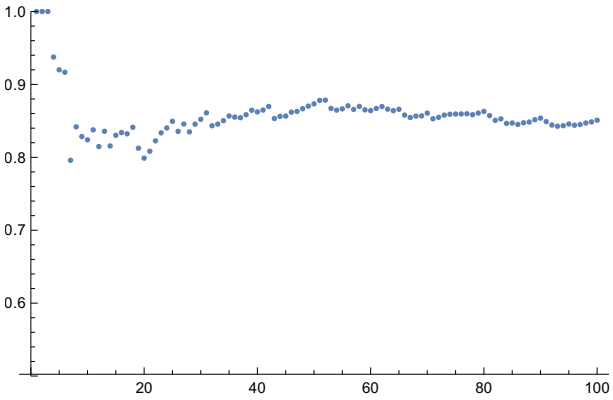


Figure 4: VUS values as function of data set size, for sizes from 1 to 100, on the artificial three-dimensional set of probability vectors shown in Figure 5.

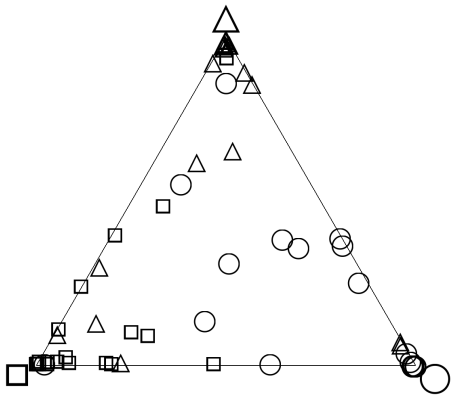


Figure 5: The distribution of the three-dimensional probability vectors used for generating the VUS values shown in Figure 4. For clarity of presentation, only 20 points from each data set are shown. The "true corners" of each class are marked with the corresponding symbol.

Moving from three to four classes, we observed the same agreement between exact calculation and heuristic approximation (Pearson correlation coefficient again at 0.995), leading us to believe that the agreement might extend to $N > 4$, for which the exact calculation is all but infeasible. The probability vectors were generated as described above; the result of a scatterplot between the two sets of results is given in Figure 6.

3.2. CIFAR-10 data set

The CIFAR-10 data set (Krizhevsky 2009) is one of the most widely-used data sets in machine learning. It consists of 50 000 training and 10 000 test color images at a

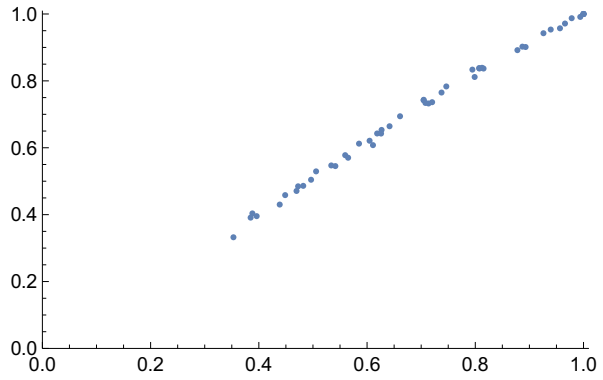


Figure 6: Scatterplot of the 50 VUS values obtained by the traditional distances-to-corner method (on the x-axis) vs. the proposed angles heuristic (on the y-axis), for $N = 3$ classes on the artificial data set. All values were obtained with 50 probability vectors in each of the three classes.

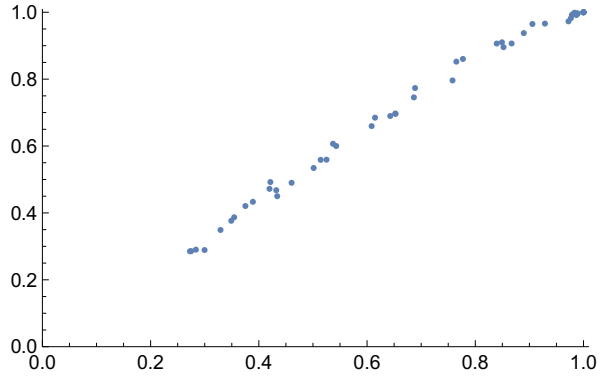


Figure 7: Scatterplot of the 50 VUS values obtained by the traditional distances-to-corner method (on the x-axis) vs. the proposed angles heuristic (on the y-axis), for $N = 4$ classes on the artificial data set. All values were obtained with 50 probability vectors in each of the four classes.

resolution of 32×32 pixels from ten different classes, ranging from airplanes and automobiles to cats and dogs. We used a Tensorflow implementation of a deep neural network with four convolutional and two fully connected layers to train a classifier for distinguishing between all ten classes. To obtain results across a more generally representative range of VUS values (and not just close to 1), we trained the neural network model for only one iteration through all 50 000 training instances. Already in this one iteration, we achieved an accuracy value of 61.65% on the test set.

For the experiments reported here, we picked arbitrary three or four classes and renormalized the corresponding vector components to turn them into probability vectors. This choice of classes was repeated 10 times; for each such data set, we sampled 5 instances of 50 probability vectors in each class. The scatterplot comparing the angles heuristics with the distances-to-corners method for three classes is shown in Figure 8. One can observe

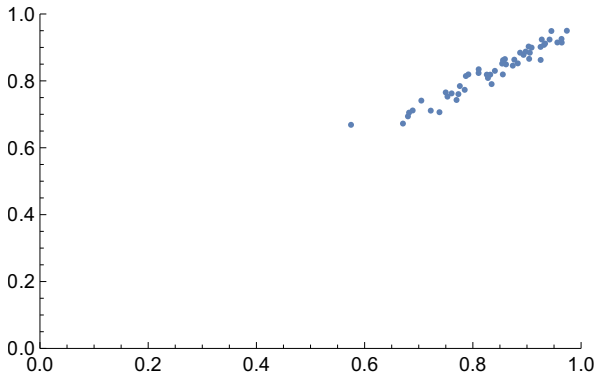


Figure 8: Scatterplot of the 50 VUS values obtained by the traditional distances-to-corner method (on the x-axis) vs. the proposed angles heuristic (on the y-axis), for $N = 3$ classes on the CIFAR-10 set. All values were obtained with 50 probability vectors in each of the three classes.

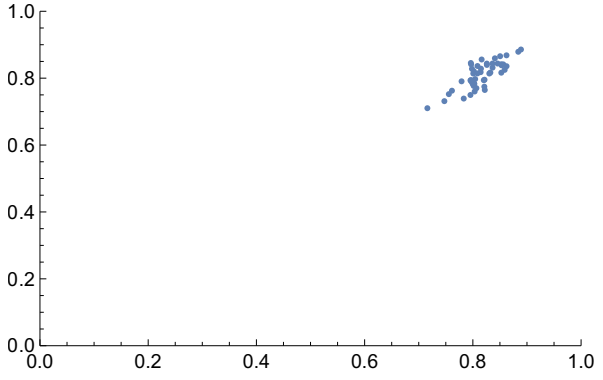


Figure 9: Scatterplot of the 50 VUS values obtained by the traditional distances-to-corner method (on the x-axis) vs. the proposed angles heuristic (on the y-axis), for $N = 4$ classes on the CIFAR-10 set. All values were obtained with 15 probability vectors in each of the four classes.

that the VUS values are rather high, with most in the range of 0.9 to 1. The Pearson correlation coefficient between both the angles heuristic and the distances-to-corners method was 0.968. The corresponding plot for four classes is shown in Figure 9, with a similar appearance, but a more narrow range of VUS values, and more dispersion of points and a correspondingly lower Pearson correlation coefficient of 0.797. We speculate, however, that this lower value may be due to the more narrow range of VUS values.

4. CONCLUSION

The area under the ROC curve, and its multiclass variant, the volume under the ROC surface, provide alternatives to accuracy as a measure of classifier performance. We demonstrated empirically that a simple $O(N)$ heuristic approximation to a prohibitively computationally expensive $O(N!)$ calculation for N classes is able to achieve similar results.

REFERENCES

Bamber D., 1975. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12:387–415.

Bradley A., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159.

Davis J. and Goadrich M., 2006. The relationship between Precision-Recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML-2006)*, pages 233–240.

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

Edwards D., Metz C., and Kupinski M., 2004. Ideal observers and optimal ROC hypersurfaces in N-class classification. *IEEE Transactions on Medical Imaging*, 23:891–895.

Edwards D., Metz C., and Naishikawa R., 2005. The hypervolume under the ROC hypersurface of "near-guessing" and "near-perfect" observers in N-class classification tasks. *IEEE Transactions on Medical Imaging*, 24:293–299.

Fawcett T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Flach P., 2003. The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. In *Proceedings of the 20th International Conference on Machine Learning (ICML-2003)*, pages 226–233.

Fürnkranz J. and Flach P., 2005. ROC 'n' rule learning—towards a better understanding of covering algorithms. *Machine Learning*, 58:39–77.

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

Hanley J. and McNeil B., 1982. The meaning and use of the area under the receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36.

Hanley J. and McNeil B., 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843.

He X. and Frey E., 2006. Three-class ROC analysis—the equal error utility assumption and the optimality of three-class roc surface using the ideal observer. *IEEE Transactions on Medical Imaging*, 25:979–986.

He X. and Frey E., 2008. The meaning and use of the volume under a three-class ROC surface (VUS). *IEEE Transactions on Medical Imaging*, 28:577–588.

He X., Metz C., Tsui B., Links J., and Frey E., 2006. Three-class ROC analysis—a decision theoretic approach under the ideal observer framework. *IEEE Transactions on Medical Imaging*, 25:571–581.

Krizhevsky A., 2009. Learning multiple layers of features from tiny images. Technical report, Computer Science Department, University of Toronto.

Landgrebe T. and Duin R., 2007. Approximating the multiclass roc by pairwise analysis. *Pattern Recognition Letters*, 28:1747–1758.

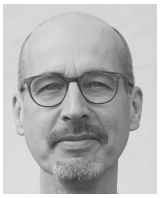
Lasko T., Bhagwat J., Zhou K., and Ohno-Machado L., 2005. The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, 38(5):404–415.

Lusted L., 1978. General problems in medical decision making with comments on ROC analysis. *Seminars in Nuclear Medicine*, 8:299–306.

Metz C., 1978. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8:283–298.

Mossman D., 1999. Three-way ROCs. *Medical Decision Making*, 19:78–89.

AUTHOR BIOGRAPHIES



STEPHAN DREISEITL received his MSc and PhD degrees from the University of Linz, Austria, in 1993 and 1997, respectively. He worked as a visiting researcher at the Decision Systems Group/Harvard Medical School before accepting a post as professor at the Upper Austria University of Applied Sciences in Hagenberg, Austria, in 2000. His research interests lie in the development of machine learning models and their application as decision support tools in biomedicine.