T. Landgrebe and R.P.W. Duin,
A simplified volume under the ROC hypersurface,
SAIEE Africa Research Journal, vol. 98, no. 3, 2007, 94-100.

A SIMPLIFIED VOLUME UNDER THE ROC HYPERSURFACE.

T.C.W. Landgrebe and R.P.W. Duin

Elect. Eng., Maths and Comp. Sc., Delft University of Technology, The Netherlands

Abstract: The Receiver Operator Characteristic (ROC) plot allows a classifier to be evaluated and optimised over all possible operating points. The Area Under the ROC (AUC) has become a standard performance evaluation criterion in two-class pattern recognition problems, used to compare different classification algorithms independently of operating points, priors, and costs. Extending the AUC to the multiclass case is considered in this paper, called the volume under the ROC hypersurface (VUS). A simplified VUS measure is derived that ignores specific intra-class dimensions, and regards inter-class performances only. It is shown that the VUS measure generalises from the 2-class case, but the bounds between random and perfect classification differ, with the lower bound tending towards zero as the dimensionality increases. A number of experiments with known distributions are used to verify the bounds, and to investigate a numerical integration approach to estimating the VUS. Experiments on real data compare several competing classifiers in terms of both error-rate and VUS. It was found that some classifiers compete in terms of error-rate, but have significantly different VUS scores, illustrating the importance of the VUS approach.

Key Words: Pattern recognition, ROC analysis, Volume under the ROC, AUC, multiclass

1. INTRODUCTION

A very active area in pattern recognition has been the consideration of classifier design and evaluation in less well-defined environments e.g. undefined or varying prior probabilities [1], or poorly defined costs [2]. A primary analysis tool developed for this domain is Receiver Operator Characteristic (ROC) analysis [3], allowing a classifier to be inspected over a range of possible conditions. A popular scalar performance measure that has emerged is the Area Under the ROC (AUC) [4], allowing classifiers to be evaluated independent of priors, costs, and operating points. The AUC measure is however only applicable to the 2-class case. Considering the multiclass extension of this measure has become a topic of interest more recently, often referred to as the Volume Under the ROC hyper-Surface (VUS). Formalisation and computational aspects are more complex, but nevertheless a number of steps have been taken to generalise the AUC. In [5], a simplified VUS is estimated from a multiclass classifier by considering the AUC between each class, and all other classes (a one vs all approach), resulting in a computationally tractable algorithm $O(C)$, where there are C classes. This measure is however inherently dependent on class priors and costs, and ignores higher-order interactions. In [6], a similar estimation of the VUS is proposed that averages the AUC between all pairs of classes, which has a higher complexity of $O((C-1)(C-3)(C-5)\dots 1)$. The exact theoretical extension to the VUS in the 3-class case has been considered in [7] and [8]. In [9] the generalised VUS has been studied, providing calculations/estimations of the performance bounds of the VUS as a function of an increasing number of classes C. This involved comparing performance between perfect (separable) classifiers and random classifiers (random performance). This non-trivial study provides an important step in understanding the VUS performance measure. A related paper was presented in [10], which argued that since the VUS of a random classifier approaches that of a perfect classifier as C increases, the VUS may not in fact be a very useful performance measure.

Previous works have not gone into detail as to how the VUS can practically be applied to an arbitrary set of classifiers in realistic scenarios. In this paper we consider the practical implementation of the VUS, applied to the simplified scenario in which the overall class performances are considered, ignoring specific intra- and inter-class errors. This type of simplification restricts the VUS analysis, but nevertheless may be suitable for some problems e.g. where we are still interested in all operating points in terms of overall class performance, but the class to which an erroneous object is assigned is arbitrary (hand-written digit recognition/ face recognition are two possible applications). This simplification ensures that good classifiers tend to result in higher VUS scores than poorer ones, irrespective of C (as will be shown), resulting in an alternative measure in line with the argument in [10]. The approach presented here provides a practical methodology for computing the VUS for problems with low C^1, demonstrated via a number of experiments. In Section 2 the notation is presented, followed by a brief formalisation of multiclass ROC analysis, and the well-known AUC in Section 3. In Section 4 the simplified VUS is presented. First performance bounds are derived as a function of C. A numerical integration procedure is then proposed in order to resample the irregularly-spaced multiclass ROC, allowing for accurate estimations of the VUS. A number of problems involving known distributions are used to verify the bounds and the methodology. In Section 5 a number of experiments involving real data are presented, demonstrating practical usage of the VUS measure in 3- and 4-class problems. Finally conclusions are presented in Section 6.

2. NOTATION

We use a framework similar to [11], in which observations $\mathbf{x}$ are to be classified into one of C classes, $\omega_1, \omega_2, \ldots, \omega_C$. Each class ω_i has a class-conditional distribution $p(\mathbf{x}|\omega_i)$, and prior probability $P(\omega_i)$. Class assignment is based on Bayes rule, which assigns membership to the highest posterior output:

$$P(\omega_i|\mathbf{x}) = \frac{P(\omega_i)p(\mathbf{x}|\omega_i)}{P(\omega_1)p(\mathbf{x}|\omega_1)+P(\omega_2)p(\mathbf{x}|\omega_2)+\ldots P(\omega_C)p(\mathbf{x}|\omega_C)} \qquad (1)$$

Thus $\mathbf{x}$ is assigned according to:

$$argmax_{i=1}^{C} P(\omega_i|\mathbf{x}) \qquad (2)$$

In the practical case in which class conditional distributions are usually unknown, these are typically estimated from representative examples that are assumed to be randomly drawn from the true distribution, and the same framework can be used. A given classifier is analysed in detail via the $C \times C$ dimensional normalised confusion matrix Ξ, in which diagonal elements represent the overall performance of each class, and off-diagonal elements the errors related to each class. Each element (i,j) of Ξ is denoted $\xi_{i,j}$. Ξ can be written as:

		estimated			
		ω_1	ω_2	$\cdots$	ω_C
	ω_1	$\xi_{1,1}$	$\xi_{1,2}$	$\cdots$	$\xi_{1,C}$
true	ω_2	$\xi_{2,1}$	$\xi_{2,2}$	$\cdots$	$\xi_{2,C}$
	$\vdots$	$\vdots$		$\ddots$	
	ω_C	$\xi_{C,1}$	$\xi_{C,2}$	$\cdots$	$\xi_{C,C}$

Table 1: Defining the multi-class normalised confusion matrix Ξ.

Each element $\xi_{i,j}$ is computed as follows:

$$\xi_{i,j} = p(\omega_i)\int p(\mathbf{x}|\omega_i)I_{ij}(\mathbf{x})dx \qquad (3)$$

The indicator function $I_{ij}(\mathbf{x})$ specifies the relevant domain (with the second line specifying performances on the diagonal elements):

$$I_{ij}(\mathbf{x}) = \begin{cases} 1 \text{ if } p(\omega_j|\mathbf{x}) > p(\omega_k|\mathbf{x})\ \forall k, k \neq j,\ i \neq j \\ 1 \text{ if } p(\omega_i|\mathbf{x}) > p(\omega_k|\mathbf{x})\ \forall k, k \neq i,\ i = j \\ 0 \text{ otherwise} \end{cases} \qquad (4)$$

In the practical case, $\xi_{i,j}$ is estimated via representative test sets, counting the number of objects classified to each element, normalised by the number of objects in that class.

3. MULTI-CLASS ROC ANALYSIS

It is important to understand that the confusion matrix actually only indicates the performance of a trained classifier at a single operating point i.e. different operating points result in different confusion matrices. The operating point is varied by weighting the posterior output of the classifier by the vector $\Phi = [\phi_1, \phi_2, \ldots, \phi_C], \phi_i > 0, \forall i$, which is analogous to classifier thresholds. Thus Equation 2 is modified as $argmax_{i=1}^{C} \phi_i P(\omega_i|\mathbf{x})$. All combinations of Φ result in all possible operating points of the classifier, which is the multiclass ROC. Note that there are in fact only $(C-1)$ degrees of freedom for a trained classifier, so one weight can be held constant, or normalised by the others. After applying all combinations of Φ, a C^2-dimensional operating characteristic results, with each confusion matrix element attributed to a new dimension. Note that only $(C^2 - C)$ dimensions are required, since:

$$\epsilon_{i,i} = 1 - \sum_{j=1, j\neq i}^{j=C} \epsilon_{i,j} \qquad (5)$$

The two class case is very well known, with two off-diagonal elements resulting ($\xi_{1,2}$ and $\xi_{2,1}$, popularly known as the false negative- and false positive-rates), and two diagonal elements ($\xi_{1,1}$ and $\xi_{2,2}$, the true positive and true negative-rates). This operating characteristic has well understood characteristics and bounds [4], [1]. Varying the classifier threshold results in a $1D$ ROC curve. Figure 1 show ROC plots for three different scenarios, ranging from a perfect/separable classifier (A), to a classifier with some overlap (B), and finally to the random classification case (C). Considering the area consumed by each classifier allows performance to be inspected independent of priors, costs, and operating points. In this 2-class case, perfect classification results in a larger area, bounded by 1, and poor classification in a smaller area, bounded by 0.5 (since the random classifier bisects the unit square). This area is known as the Area Under the ROC (AUC). Note that traditionally the ROC is plotted between $\xi_{1,1}$ and $\xi_{2,1}$, but Figure 1 results in an equivalent performance

[1] Extension to the high C case remains computationally infeasible, and thus our approach is restricted to low C problems e.g. $C = 3$ to 6. Simpler approaches such as [6] are the only candidates for high C.

measure, and is extensible to the multiclass case.

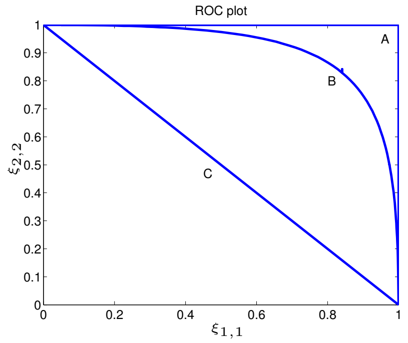

Figure 1: Comparing 3 different 2-class ROC plots. 'A' depicts perfect classification, 'B' is a classifier with some overlap, and 'C' is a random classifier.

The AUC can be written as:

$$AUC = \int \xi_{2,2} d\xi_{1,1} \tag{6}$$

The AUC can be applied to the realistic scenario by a numerical integration scheme. This work uses the trapezoidal integration rule. The AUC can also be estimated by counting the number of times two arbitrary objects in the test set from both classes are correctly ranked by the classifier, and normalising.

4. SIMPLIFIED VOLUME UNDER THE ROC

Extending the AUC to the multiclass case, i.e. the volume under the ROC hypersurface, can be achieved by measuring the volume bounded by the operating characteristic. In this case we consider only ROC dimensions pertaining to diagonal elements of the confusion matrix. The simplified VUS can be written as:

$$VUS = \int \ldots \int \int \xi_{C,C} d\xi_{C-1,C-1} d\xi_{C-2,C-2} \ldots d\xi_{1,1} \tag{7}$$

Thus the simplified measure considers the $C-$dimensional operating characteristic of a $C-$dimensional problem. This measure allows a classifier to be evaluated over all operating points responsible for the ROC dimensions corresponding to the diagonal confusion matrix elements. If these performances only are considered, the VUS is similar to the AUC in that better classifiers will result in a high VUS, and poorer classifiers in a lower score. However, before the VUS is blindly applied, it is important to characterise and understand the performance bounds between random and perfect classifiers.

4.1. *Bounds as a function of dimensionality*

Considering the 3-class case first, the simplified ROC dimensionality is 3, between the dimensions $\xi_{1,1}, \xi_{2,2}$, and $\xi_{3,3}$. A random classifier produces the ROC depicted in Figure 2. A more effective classifier (or more separable problem) is depicted in Figure 3, showing how the VUS increases.

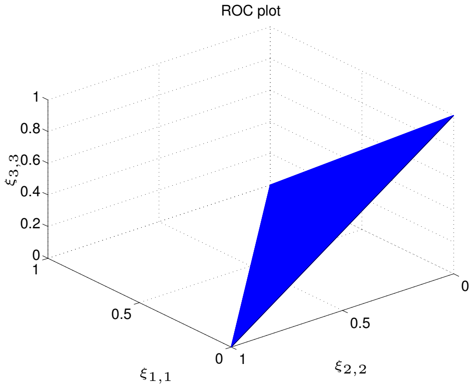

Figure 2: Random classification performance of the simplified 3-class ROC.

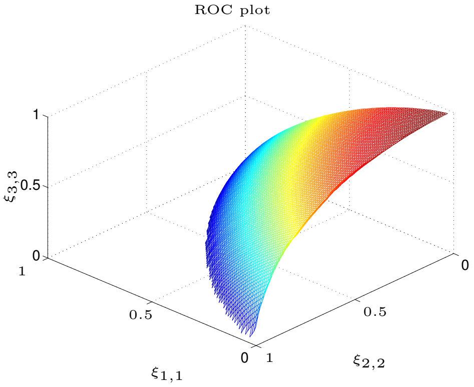

Figure 3: ROC plot for a 3-class problem with partially overlapping distributions.

In fact, the VUS approaches 1.0 as the classification becomes perfect. The VUS occupied by the random classifier can be found geometrically by computing the volume of the tri-rectangular tetrahedron formed under the surface, which is simply $\frac{1}{6}\xi_{1,1}\xi_{2,2}\xi_{3,3} = \frac{1}{6}$. Thus the bound has altered from $\frac{1}{2}$ in the two-class case, to $\frac{1}{6} = 0.16666$ in the 3-class case. Generalising the bounds to C classes is more difficult geometrically. A more extensible approach is to formalise the random ROC as a hyper-polyhedron, as proposed in [9]. Each vertex v_i of the hyper-polyhedron can easily be defined as (note that the origin is always included as a vertex, and there are C points per vertex):

$$\begin{array}{cccccc} v_1 & 0 & 0 & 0 & \dots & 0 \\ v_2 & 1 & 0 & 0 & \dots & 0 \\ v_3 & 0 & 1 & 0 & \dots & 0 \\ v_4 & 0 & 0 & 1 & \dots & 0 \\ & & \vdots & & & \\ v_{C+1} & 0 & 0 & 0 & \dots & 1 \end{array} \tag{8}$$

As in [9], the optimised QHull [12] algorithm is used to estimate the volume occupied by the hyper-polyhedron. The following lower bounds result, up to $C = 12$, showing how the lower bound approaches zero with an increasing C. In fact, it can be seen that the lower bound is $\frac{1}{C!}$, which is proven in Appendix A[2]:

C	**Estimated VUS**
2	0.50000000001826
3	0.16666666668765
4	0.04166666667598
5	0.00833333333563
6	0.00138888888932
7	0.00019841269853
8	0.00002480158732
9	0.00000275573193
10	0.00000027557319
11	0.00000002505211
12	0.00000000208768

(9)

4.2. *Estimating the VUS for general classifiers*

In the practical situation in which a sparse set of points are given, representing the multiclass ROC, a different approach is required. Since the ROC surface is derived by the nature of the problem and classifier, it cannot be computed analytically. A more appropriate approach to estimating the VUS is to use a numerical integration approach. The inherent uneven sampling of the ROC is converted to an even form via linear resampling and interpolation. The trapezoidal rule is then used to estimate the volume (in $C-$dimensions), with the following results as a function of r, the number of ROC steps used:

C	**r**	VUS estimation	Actual VUS
3	50	0.1667014	0.1666666
3	100	0.1666752	0.1666666
4	50	0.0417014	0.0416667
4	100	0.0416752	0.0416667
5	50	0.0083507	0.0083333
5	100	0.0083376	0.0083333
6	20	0.0014275	0.0013889
6	40	0.0013980	0.0013889

These results show that the numerical integration approach provides a good approximation of the true VUS, and that as expected a higher step size results in higher accuracy.

4.3. *Experiments with known distributions*

In order to judge the numerical VUS approach and verify the bounds, a number of controlled experiments are conducted, consisting of generated Gaussian classes with known parameters. The first set of experiments consist of 3-class Gaussian problems with classes ω_1, ω_2, and ω_3, in which the means are varied, and the variances held at unity. The means are varied such that the problems range from near-separable problems, to near-random. Similarly the second set of experiments involve varying the means of 4 Gaussian classes. Tables 2 and 3 depict the results for the 3- and 4-class cases respectively, also showing r (a higher resolution was required as the distributions approached complete overlap). In Figure 4, the distributions used in the 2nd and 4th 4-class experiments are shown, demonstrating how class overlap was increased.

Means	**r**	**VUS est.**
$-0.05; 0.0; 0.05$	200	0.16876
$-0.3; 0.0; 0.3$	100	0.24140
$-0.5; 0.0; 0.5$	100	0.31428
$-1.0; 0.0; 1.0$	100	0.51214
$-1.5; 0.0; 1.5$	100	0.70597
$-4.0; 0.0; 4.0$	100	0.98582

(10)

Table 2: Results for 3-class experiments with known distributions.

Means	**r**	**VUS est.**
$-0.15; -0.05; 0.05; 0.15$	70	0.05688
$-0.75; -0.25; 0.25; 0.75$	50	0.07782
$-1.00; -0.33; 0.33; 1.00$	50	0.19972
$-1.50; -0.50; 0.50; 1.5$	50	0.33097
$-2.25; -0.75; 0.75; 2.25$	50	0.57990
$-3.00; -1.00; 1.00; 3.0$	50	0.75451

(11)

Table 3: Results for 4-class experiments with known distributions.

These experiments verify that the VUS approach used does make intuitive sense, since it can be seen that as the problems vary from the separable to the random

[2] The bounds of the simplified VUS suggest this method is a good alternative to the true unsimplified VUS (regarding the argument given in [10] pertaining to poor resolution between perfect and random classifiers for high dimensions, bringing the validity of the VUS into question). This is because in the simplified case for high C, good classifiers tend to 1, and poor ones tend to 0.

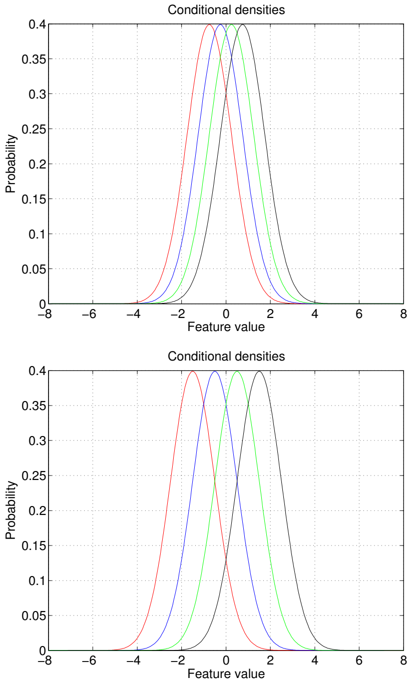

Figure 4: Demonstrating the 2nd and 4th experiment in the 4-class case.

case, the VUS decreases accordingly. For highly overlapping cases, the two sets of experiments demonstrate a VUS that approaches the predicted lower bounds.

5. EXPERIMENTS

The VUS methodology is demonstrated in real settings by comparing a number of competing classifiers over a number of different problems. The first group of experiments consist of 3-class problems, with the following datasets used: *Banana* is a 2-dimensional dataset consisting of a Banana-shaped class [13], a Gaussian distributed class, and a bimodal Gaussian class, which are all overlapping, with 5073 objects generated in total. The *Sign* dataset [14] consists of images of 3-classes of road-signs, with a total of 381 objects. The *Sat* dataset [15] consists of 6435 multi-spectral values of a satellite image, with 36 dimensions (4 spectral bands in a 9 pixel neighbourhood). Classes 1, 3, 5 and 6 have been grouped together into a single class, forming a 3-class problem together with classes 2 and 4. The second group of experiments consist of 4-class problems. The *Vehicle* dataset [16] consists of 846 objects of vehicle silhouettes from 4 vehicle types, and the *Digits* dataset consists of examples of ten handwritten digits, originating from Dutch utility maps (available from [16]). In this dataset, Fourier components have been extracted from the original images, resulting in a 76-dimensional representation of each digit. Digits '3', '6', and '9' have been extracted, and the remaining digits grouped into a single class. The experimental methodology involves rotation of the data using a randomised hold-out method in which 80% of the data is used in training, and the remainder for testing, repeated 10 times. Two performance measures are compared, namely the well-known equal-error rate (priors inherent to dataset used), and the simplified VUS measure. Results are compared statistically via a 2-way ANOVA (ANalysis Of VAriance) scheme, with significance judged via a $p-$ value of 0.995. In each experiment, a number of classifiers are compared, with the following abbreviations: *sc* is where unit-variance scaling of the data is used; *pca* is a principal component feature extraction followed by the number of components used; *fisher* and *nlfisher* are the Fisher and non-linear Fisher projections; *nmc*, *ldc*, and *qdc* are nearest-mean, Bayes-linear, and Bayes-quadratic classifiers respectively; *mogc* is a Bayes mixture of Gaussians classifier followed by the number of mixtures used per class; *knn3* is a 3-nearest neighbour classifier; *svc p* is a support vector classifier with a polynomial kernel, followed by the order of the polynomial.

The 3-class table presents the first set of results. The *Banana* dataset shows that the VUS scores tend to track the equal error scores, for example the *nmc* classifier has a high error, and significantly lower VUS than the other classifiers. An interesting result can be seen for the *Sat* case, comparing the second and third models. In this case both classifiers have the same (statistical) error-rate, but significantly different VUS scores (F-value of 275), showing that the third model is a better choice on average over all operating points. In the *Sign* experiments, similar VUS scores result for all classifiers.

3-class	**Classifier**	**Error**	**VUS**
Banana	sc-pca1 nmc	0.329(0.004)	0.667(0.083)
	sc-mogc2,2,1	0.058(0.003)	0.990(0.002)
	sc-qdc	0.077(0.004)	0.970(0.006)
	sc-ldc	0.091(0.004)	0.964(0.007)
Sat	knn3	0.064(0.002)	0.911(0.020)
	ldc	0.111(0.001)	0.729(0.015)
	qdc	0.108(0.002)	0.862(0.012)
	mogc2,1,2	0.099(0.002)	0.866(0.012)
Sign	sc pca8 svc p2	0.115(0.018)	0.948(0.023)
	sc-pca10 mog2,2,2	0.075(0.003)	0.946(0.025)
	pca5 mog2,2,2	0.099(0.011)	0.954(0.019)
	pca5 qdc	0.179(0.020)	0.945(0.023)

Table 4: Experimental results on 3-class problems.

Next the 4-class experiments are considered. A few interesting observations can again be made, for example the first and second classifiers have competing error-rates, but significantly different VUS scores. It appears the *linear* classifier was a far better fit to the data than the *fisher-nmc* model, which only performed well for some operating points. Finally in the *Digits* case, the

VUS tended to track the error-rates. It can be seen that some classifiers perform very well, approaching a VUS of 1, whereas others are poor.

4-class	**Classifier**	**Error**	**VUS**
Vehicle	fisher nmc	0.218(0.007)	0.512(0.037)
	ldc	0.219(0.006)	0.714(0.035)
	qdc	0.150(0.010)	0.834(0.036)
	sc-svc p2	0.164(0.007)	0.794(0.022)
	sc-svc p3	0.187(0.009)	0.727(0.039)
	nlfisher qdc	0.208(0.005)	0.724(0.041)
Digits	pca10 mog1,1,1,3	0.119(0.004)	0.985(0.008)
	pca15 mog1,1,1,3	0.114(0.003)	0.955(0.007)
	pca5 mog1,1,1,3	0.133(0.003)	0.956(0.008)
	pca10 qdc	0.127(0.004)	0.978(0.006)
	pca10 ldc	0.211(0.005)	0.704(0.041)
	nlfisher mogc1,1,1,3	0.158(0.003)	0.857(0.024)

Table 5: Experimental results on 4-class problems.

The experiments showed the usefulness of the VUS approach in the multiclass case, clearly showing examples where the VUS was required to perform better model selection for classifiers that competed from an equal-error perspective.

6. CONCLUSIONS

This paper considered the extension of the AUC measure to the multiclass case, termed the *volume under the ROC hypersurface*. A simplified extension was considered that evaluates the VUS over the $C-$dimensional ROC surface pertaining to diagonal elements of the confusion matrix only, thus ignoring specific inter- and intra-class performances. This allows for a measure that generalises from the 2-class case, in which high scores result for good classifiers, and low ones for poor ones. It was seen that the VUS bounds vary as a function of the ROC dimensionality, with the lower bound tending to 0 with high dimensionality. A few experiments using known distributions verified the bounds, as well as a proposed numerical integration approach to estimating the hyper-volumes. Finally a set of real experiments were performed that compared equal-errors to VUS scores for a number of competing classifiers. It was found that poor error rates often lead to poor VUS scores, but in some cases competing classifiers in terms of error-rate are not competing in terms of VUS, implying that some classifiers perform better on average over all operating points than others. This work is considered useful to problems involving a low number of classes, restricted by the computational complexity of the ROC generation, but may nevertheless be useful for many real problems.

Acknowledgements: This research is/was supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs, The Netherlands. Special thanks go to Etienne Barnard for his assistance with the analytical bound proof.

[1] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine Learning*, vol. 42, pp. 203–231, 2001.

[2] N. Adams and D. Hand, "Comparing classifiers when misallocation costs are uncertain," *Pattern Recognition*, vol. 32, no. 7, pp. 1139–1147, 1999.

[3] C. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 3, no. 4, 1978.

[4] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[5] F. Provost and P. Domingos, "Well trained PETs: Improving probability estimation trees," *CeDER Working Paper IS-00-04, Stern School of Business, New York University NY 10012*, 2001.

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

[7] D. Mossman, "Three-way roc's," *Medical Decision Making*, vol. 19, pp. 78–89, 1999.

[8] S. Dreisetl, S. Ohno-Machado, and M. Binder, "Comparing trichotomous tests by three-way ROC analysis," *Medical Decision Making*, vol. 20, no. 3, pp. 323–331, 2000.

[9] C. Ferri, J. Hernandez-Orallo, and M. Salido, "Volume under the roc surface for multi-class problems," *Proc. of 14th European Conference on Machine Learning*, pp. 108–120, 2003.

[10] D. Edwards, C. Metz, and R. Nishikawa, "The hypervolume under the ROC hypersurface of "near-guessing" and "near-perfect" observers in N-class classification tasks," *IEEE Trans. on Medical Imaging*, vol. 24, pp. 293–299, March 2005.

[11] T. Landgrebe and R. Duin, "On Neyman-Pearson optimisation for multiclass classifiers," *Sixteenth Annual Symposium of the Pattern Recognition Assoc. of South Africa*, November 2005.

[12] C. Barber and H. Huhdanpaa, "Qhull," *The Geometry Center, University of Minnesota, http://www.geom.umn.edu/software/qhull.*

[13] R. Duin, P. Juszcak, P. Paclik, E. Pekalska, D. de Ridder, and D. Tax, "Prtools, a matlab toolbox for pattern recognition," January 2004.

[14] P. Paclík, "Building road sign classifiers," *PhD thesis, CTU Prague, Czech Republic*, December 2004.

[15] ELENA project, "European ESPRIT 5516 project," *Satimage dataset*, 2004.

[16] P. Murphy and D. Aha, "UCI repository of machine learning databases,

ftp://ftp.ics.uci.edu/pub/machine-learning-databases," *University of California, Department of Information and Computer Science*, 1992.

A. APPENDIX: PROOF OF LOWER SIMPLIFIED VUS BOUND

Figures 1 and 2, graphically depicted a random 2-class and 3-class ROC plot. We refer to the volume consumed by a random classifier as the lower bound. In the 2-class case, the lower bound can be written as in Equation 12, since $\xi_{2,2}$ is the straight line $1-\xi_{1,1}$.

$$\begin{aligned} AUC_{random} &= \int_0^1 (1-\xi_{1,1})d\xi_{1,1} \\ &= \Big[\xi_{1,1} - \tfrac{1}{2}\xi_{1,1}^2\Big]_0^1 \\ &= \tfrac{1}{2} \end{aligned} \tag{12}$$

In the 3-class case, the volume can be computed analytically by considering the volume under the plane $1-\xi_{1,1}-\xi_{2,2}$:

$$\begin{aligned} VUS_{random} &= \int_0^1 \int_0^{1-\xi_{1,1}} (1-\xi_{1,1}-\xi_{2,2})d\xi_{2,2}d\xi_{1,1} \\ &= \tfrac{1}{2}\int_0^1 (1-\xi_{1,1})^2 d\xi_{1,1} \\ &= \tfrac{1}{2}\tfrac{1}{3} \\ &= \tfrac{1}{6} \end{aligned} \tag{13}$$

Similarly, the 4-class case considers the volume under the hyperplane $1-\xi_{1,1}-\xi_{2,2}-\xi_{3,3}$:

$$\begin{aligned} &VUS_{random} = \\ &\int_0^1 \int_0^{1-\xi_{1,1}} \int_0^{1-\xi_{1,1}-\xi_{2,2}} (1-\xi_{1,1}-\xi_{2,2}-\xi_{3,3})d\xi_{3,3}d\xi_{2,2}d\xi_{1,1} \\ &= \tfrac{1}{2}\int_0^1 \int_0^{1-\xi_{1,1}} (1-\xi_{1,1}-\xi_{2,2})^2 d\xi_{2,2}d\xi_{1,1} \\ &= -\tfrac{1}{2}\tfrac{1}{3}\int_0^1 \Big[(1-\xi_{1,1}-\xi_{2,2})^3\Big]_0^{1-\xi_{1,1}} d\xi_{1,1} \\ &= \tfrac{1}{2}\tfrac{1}{3}\tfrac{1}{4} = \tfrac{1}{24} \end{aligned} \tag{14}$$

As C increases, it can be seen that the VUS calculation can be simplified by using the following well-known integration rule recursively:

$$\int (ax+b)^n = \frac{(ax+b)^{n+1}}{a(n+1)}, n \neq -1 \tag{15}$$

The bound for any C can then be computed as follows:

$$\begin{aligned} &VUS_{random} = \\ &\int_0^1 \int_0^{1-\xi_{1,1}} \int_0^{1-\xi_{1,1}-\xi_{2,2}} \cdots \int_0^{1-\xi_{1,1}-\xi_{2,2}-\dots\xi_{C-2,C-2}} \\ &(1-\xi_{1,1}-\xi_{2,2}-\dots\xi_{C-1,C-1})d\xi_{C-1,C-1}\dots d\xi_{2,2}d\xi_{1,1} \\ &= \tfrac{1}{2}\int_0^1 \int_0^{1-\xi_{1,1}} \cdots \int_0^{1-\xi_{1,1}-\xi_{2,2}-\dots\xi_{C-3,C-3}} \\ &(1-\xi_{1,1}-\xi_{2,2}-\dots\xi_{C-2,C-2})^2 d\xi_{C-2,C-2}\dots d\xi_{2,2}d\xi_{1,1} \\ &= \tfrac{1}{2}\tfrac{1}{3}\int_0^1 \int_0^{1-\xi_{1,1}} \cdots \int_0^{1-\xi_{1,1}-\xi_{2,2}-\dots\xi_{C-4,C-4}} \\ &(1-\xi_{1,1}-\xi_{2,2}-\dots\xi_{C-3,C-3})^3 d\xi_{C-3,C-3}\dots d\xi_{2,2}d\xi_{1,1} \\ &= \tfrac{1}{2}\tfrac{1}{3}\tfrac{1}{4}\cdots\tfrac{1}{C} = \tfrac{1}{C!} \end{aligned} \tag{16}$$