The problem of Area Under the Curve

Conference Paper · March 2012

DOI: 10.1109/ICIST.2012.6221710

1 author:

David Martin Ward Powers
Flinders University
358 PUBLICATIONS **11,190** CITATIONS

SEE PROFILE

The Problem of Area Under the Curve

David M W Powers

Abstract—**Receiver Operating Characterics curves have recently received considerable attention as mechanisms for comparing different algorithms or experimental results. However the common approach of comparing Area Under the Curve has come in for some criticism with alternatives such as Area Under Kappa and H-measure being proposed as alternative measure. However, these measures have their own idiosyncracies and neglect certain advantages of RoC analysis that do not carry over to the proposed approach. In addition they suffer from the general desire for a one fits all mentality as opposed to a pareto approach. That is we want a single number to optimize, rather than a suite of numbers to trade off. The starting point for all of this research is the inadequacy and bias of traditional measures such as Accuracy, Recall and Precision, and F-measure – these should *never* be used singly or as a group, and ROC analysis is a very good alternative to them if used correctly, but treating it as a graph rather than trying to boil it down to a single number. Other measures that seek to remove the bias in traditional measures include Krippendorf's Alpha, Scott's Pi, Powers' Informedness and Markedness, as well as a great many variant Kappa statistics. The original Kappa statistics were intrinsically dichotomous but the family has been well generalized to allow for multiple classes and multiple sources of labels. We discuss the proper and improper use of ROC curves, the issues with AUC and Kappa, and make a recommendation as to the approach to take in comparing experimental algorithms or other kinds of tests.**

I. Introduction

THIS paper addresses a fundamental problem in science that has come to a head recently in computer science [1],[2], namely how to compare research results objectively. Generally, the signal processing, artificial intelligence, machine learning and computational linguistics communities, the medical and psychology communities, and other areas of science and engineering, use a wide range of closely related measures for assessing experimental results.

Psychology and Medicine have traditionally employed Specificity and Sensitivity as two dimensions to characterize and experiment – how specifically do we predict the target (viz. how well do we avoid false positives – irrelevant hits in information retrieval terms) and how sensitively do we predict the target (viz. how well do we avoid false negatives – misses in other words). Given that various parameters may be tuned, it is possible to plot Sensitivity against Specificity (or equivalently True Positive Rate against False Positive Rate, or Recall against Fallout, or vice-versa). The resulting curve is known by its Engineering appelation of Receiver Operating Characteristics (ROC) and is designed to be used as a tuning curve to set the desired (or optimal) operation point. ROC is ubiquitous in some fields, and rare in others, but has also been extensively studyed from a theoretical standpoint [3]-[13]. In the traditional dichotomous ROC described, the denominators of the proportions plotted along each axis are the number of real positives and negatives. Valid variants reflect an axis according to whether correctly or incorrectly predicted instances are counted. This formulation corresponds to trading off Recall and Inverse Recall, whilst a dual formulation trades Precision and Inverse Recall with denominators being the numbers of positive and negative predictions rather than actual classes. We will define all these concepts formally in the next section.

In the related problem of comparing human judges, against each other or a gold standard, a family of Kappa statistics has been developed [13]-[22]. These aim to subtract off the chance level of accuracy and renormalize to the form of a probability. Two main families of Kappa exist, those that make the assumption that raters employ the same distribution (Cohen Kappa [14]), or that they use different distributions (Scott Pi or Fleiss Kappa [15]). The families have at core a two rater two class measure, and generalize in different ways to multiple rater multiclass measures, and often strict bounds separate the different measures due to their different assumptions [18]. Recently Kappas have fairly naturally been pressed into service for other kinds of comparison, with Cohen Kappa being the most commonly used in Machine Learning, but having its share of problems for this purpose [16] as well as the original purpose [21]. In particular Cohen's kappa penalizes judges who demonstrate a similar distribution of judgements [19], whilst Fleiss Kappa is explicitly biased to anticipate different distributions [15]. Different ways of weighting and combining produce different effects, with linear weighting allowing a reductive explanation in terms of dichotomous contingencies and quadratic weighting allowing interpretation as intraclass correlation coefficients [21].

This raises the question as to why not use a direct correlation coefficient based on the contingency table, such as the Matthews or tetrachoric correlation [22]-[24]. In fact the correlation reflects the geometric mean of the gradients of the lines of best fit, viz. of the regression coefficients, for the two possible directions of implication in a contingency table, with Matthews correlation corresponding to the geometric mean of DeltaP and DeltaP' [24], measures of human associative judgement empirically derived and theoretically justified within Psychology.

These measures all implicitly put some cost on the types of errors, with false positives (irrelevant hits) and false negatives (misses) being equally weighted with traditional

Manuscript received November 1, 2011.

D M W Powers is with the School of Computer Science, Engineering & Mathematics, Flinders University, South Australia (phone: +61-414-8204307; fax: +61-8-8201-3663; e-mail: David.Powers@flinders.edu.au).

TABLE 1.

SYSTEMATIC CODINGS FOR A BINARY CONTINGENCY TABLE: UPPER CASE REPRESENTS COUNTS (CELLS AND MARGINS EACH SUM TO N) AND LOWER CASE REPRESENTS PROBABILITIES (CELLS AND BOTH MARGINS EACH SUM TO N=1), AND A THIRD LETTER A OR R MAY BE ADDED TO CELL CODES TO DENOTE ACCURACY OF PREDICTION OR RATE RELATIVE TO REAL CLASS: FPA. AVERAGES ACROSS CODES WITH COMMON LETTERS REPLACE DIFFERING LETTERS WITH TYPE OF MEAN: GEOM(FP,FN)→FG. *SHADING INDICATES CORRECT (LIGHTGREEN/T) AND INCORRECT (DARKPINK/F) PROBABILITIES OR COUNTS IN CONTINGENCY TABLE.*

	+R	−R	
+P	tp	fp	pp
−P	fn	tn	pn
	rp	rn	n=1

	+R	−R	
+P	TP	FP	PP
−P	FN	TN	RN
	RP	RN	N

Accuracy, Recall and Precision measures, irrespective of the relative Prevalence of positive and negative instances, the Bias of the prediction variable or classifier to positive or negative labels, or any difference in the cost of the two types of errors. Working on the basis of the financial gain a punter or speculator would make given fair odds by his bookmaker or broker, both dichotomous and multiclass Informedness statistics have been derived [26] that like Kappa discount for chance and renormalize, but turn out to be identical to DeltaP' as well as having a connection to ROC.

Dichotomous Informedness or DeltaP', and its dual Nakedness or DeltaP, turn out to be marginal weightings (renormalizations) of the determinant of the contingency matrix [26] as does Cohen Kappa [20].

II. NOTATION & BASIC FORMULAE

In the notation of Table I, we see that a Real class **R** has positive or negative labels, **+R** and **–R**, and Prediction labels **P** similarly have positive or negative labels, **+P** and **–P**. The Prevalence of the positive class is abbreviated as Prev and is also abbreviated as `rp` for real positive, whilst the Inverse Prevalence is IPrev or `rn` for real negative. Similarly the Bias of the classifier (or rater or predictor) towards positive labels is also represented as `pp` for predicted positive, and the Inverse Bias, IBias is `pn`. This completes the definition of the four marginal probabilities.

The four joint probabilities are named to represent the proportion of true positives (positives correctly labeled positive), false positives (negatives incorrectly labeled positive), false negatives (positives incorrectly labeled negative) and true negatives (negatives correctly labeled negative). In many works, these are respectively labeled by the letters a, b, c and d. By convention lower case letters represent probabilities (and both the four joint probabilities and the pairs of real or predicted marginal probabilities sum to 1). However, sometimes it is convenient to refer to counts that sum to N, and upper case letters of both the systematic and A/B/C/D nomenclature are also in common use for this purpose.

For rates, they are written lower case, but may be written in upper case in other work, and are measured in terms of the rates of our classifier turning up real positives (in the case of Recall=`tpr=tp/rp` and Fallout=`fnr-fn/rp`) or rejecting negatives (Inverse Recall=`tnr=tn/rn`, `fpr=fp/rn`). Conversely, accuracies are measured with respect to the predictions (so that Precision=`tpa=tp/pp`) and as a special case Rand Accuracy, Acc = `(TP+TN)/N=tp+tn`.

Expected values of these probabilities (and counts) are written with an initial e (or E) and for Cohen Kappa are calculated as a product of the corresponding marginal probabilities (e.g. `etp=pp*rp`) and Expected Accuracy is thus `etp+etn`. Similarly we can calculate the Absolute Error as `fp+fn` and Expected Absolute Error as `efp+efn`.

The table of four counts is called a contingency table, although this is also sometimes loosely applied to the normalized table of probabilities (although significance cannot be estimated from this without knowing `N`).

The Determinant and Odds Ratio [20] of the probability table is given by

`dtp = (tp · tn) − (fp · fn)` and
`OR = (tp · tn) / (fp · fn)`

Cohen's Kappa is defined and reformulated [20] as

Kappa = (Rand Acc – Expected Acc)/(1 – Expected Acc)
`= dtp / [(pp · rn + pn · rp)/2]`

Powers' Informedness and Markedness are defined for the general case, as is Matthews' Correlation, but for the dichotomous case they simplify [27] to:

Informedness = `tpr − fpr = dtp / (rp · rn)`
Markedness = `tpa − fpa = dtp / (pp · pn)`
Correlation = `dtp / √(pp · pn · rp · rn)`

Comparison of these formulae demonstrates that all four measures, Cohen Kappa, Powers Informedness and Markedness, and Correlation, are just different normalizations of the determinant by a function of the marginal coefficients (for either the probability matrix or the contingency matrix).

In fact in general different versions of Kappa are based on the above formulae but with different assumptions and thus different calculations for the Expected Accuracy. Fleiss Kappa in particular allows for k raters to between them make n ratings of each of a pool of N subjects and calculates the expectation relative to the number of possible pairings [15].

Note that in the dichotomous (binary or two-class) case, the sign of all four measures is determined by the determinant. Fleiss Kappa does not always agree in sign, and is regarded as beyond the scope of this paper but will sometimes be mentioned by way of comparison.

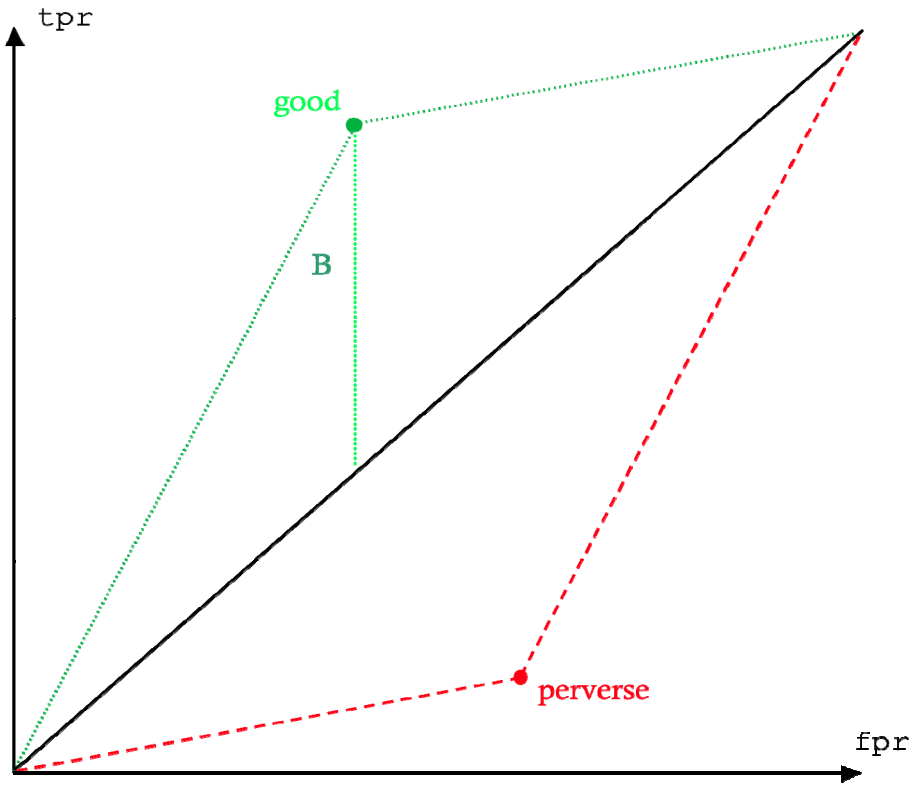

Fig. 1. Illustration of ROC Analysis. True positive rate is plotted against false positive rate, although other variants such as tpr-tnr and fnr-fpr can also be used and represent reflection of the axes. The main diagonal represents chance. Points above the diagonal are better than chance, those below are worse than chance. AUC is the area under the curve, here representing a single parameterization or system. Traditionally multiple parameterizations/systems are plotted and allow selecting the one that best reflects the cost gradient for the target population). Reversing the sense of the predictions solves the inverse problem. Here the green system uses the information to maximize performance, while the red inverse system perversely uses the same information to minimize performance.

III. Averaging over Classes or Labels

Statistics derived for individual actual classes or predicted labels need to be weighted with care, in particular noting that statistics that are rates of labelling members of a particular class (e.g. Recall) need to be weighted according to the prevalence of that class, whilst accuracy of labels need to be weighted according to the bias to that label (e.g. Prevalence). On this basis, dichotomous Recalls weighted by the prevalences of the positive and negative classes give Rand Accuracy, as does the dichotomous weighting of Prevalence by label bias.

Their Harmonic Mean, individually or averaged, corresponds to F-Measure (which references the true positives to the arithmetic mean of Prevalence and Bias) which thus also has Rand Accuracy as the appropriate weighted mean.

Averaging over class prevalences is also appropriate for Informedness, whilst averaging over label biases is appropriate for Markedness, and Correlation can be calculated as a geometric means of these, although for the dichotomous case there is no difference between the Informedness and Markedness values for the different classes. Averaging Cohen Kappa is more difficult as like F-Measure and Correlation, it is measured with respect to an averaged denominator, in this case it is a harmonic mean of cross-biased variants of Informedness and Markedness.

Generally, all discrepancies between the families of measures go away if Prevalence = Bias, and further simplification can be made if Prevalence = Bias = ½, which standardization removes the uncertainty and variation that make Recall, Precision, Accuracy and F-Measure unconscionable when Bias and Prevalence are uncontrolled.

We have seen that Prevalence and Bias make a profound difference to these measures, and we have examined the effect of Bias both empirically and theoretically [29] on three competing measures, Scott Pi, Cohen Kappa and Powers Informedness, as they approach central and extreme values of Prevalence, for extreme values of Bias (remember all are similar for equal Prevalence and Bias so we don't show midrange Bias). These limits are shown in Table 2.

It will be noted that Cohen Kappa and Powers Informedness give no weight to chance for extreme mismatch between Bias and Prevalence, whilst Scott Pi or Fleiss Kappa regards a mismatch as evidence of 50% chance performance, and is thus not regarded as suitable for evaluation of classifiers. Powers Informedness like Cohen Kappa takes the margins (Prevalence and Bias) as givens from the same distribution, and compensates for them as such, but Informedness represents the probability that the results relate to an informed decision rather than chance (guessing), and is not influenced by Bias, whereas Cohen Kappa has no such straightforward interpretation and is influenced by a complex mix of Informedness and Bias as illustrated in the reformulation of its definition above.

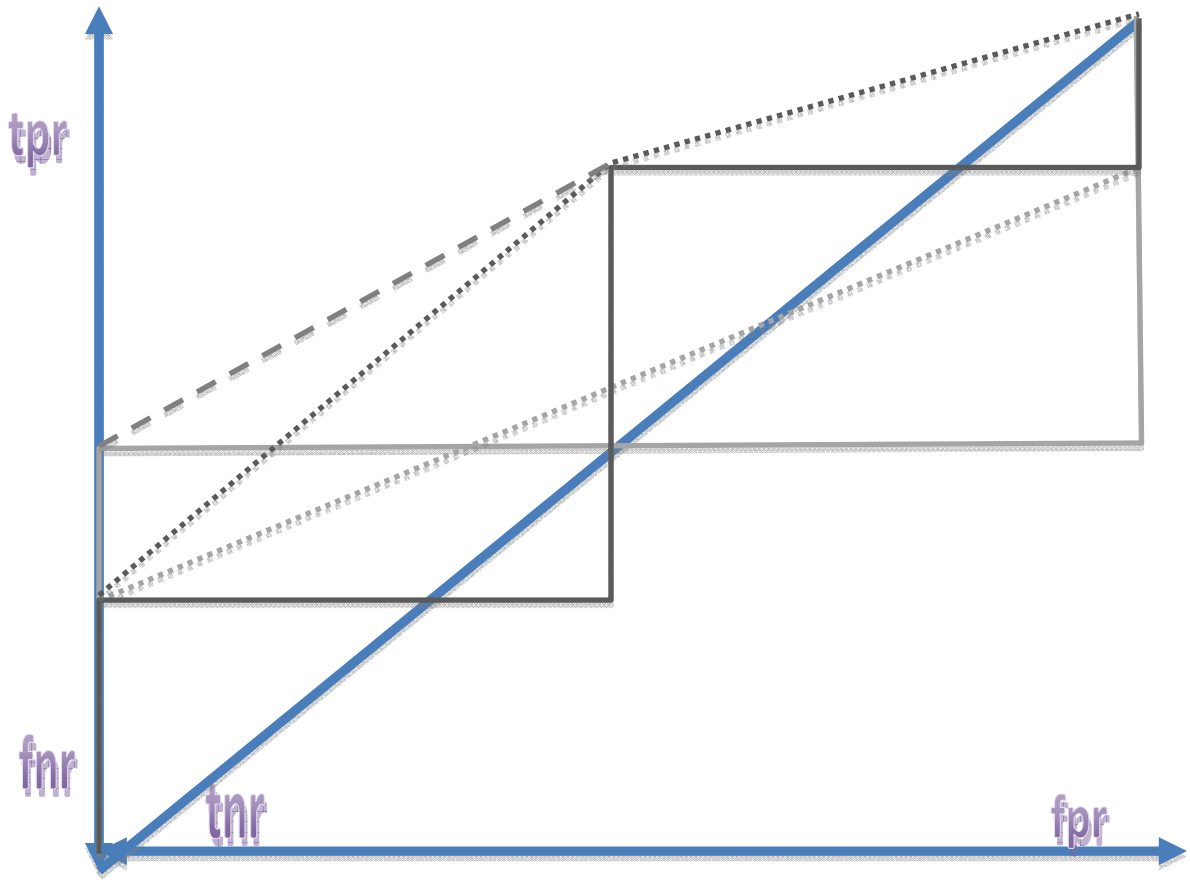

Fig. 2. Illustration of two systems (dark and light) in relation to chance line (thick diagonal), with ROC convex hulls for the individual systems, as well as a composite system formed from the two systems.

TABLE 2.
THE CHANCE ACCURACY LEVELS ASSUMED BY THREE KAPPA MEASURES AND THEIR RESPONSE TO EXTREME BIAS FOR DIFFERENT BALANCED AND EXTREME LEVELS OF PREVALENCE. [29]

Prev	→ 1		→ .5		→ 0	
Bias	→ 0	→ 1	→ 0	→ 1	→ 0	→ 1
Scott	½	1	⅝	⅝	1	½
Cohen	0	1	½	½	1	0
Powers	0	1	½	½	1	0

IV. ROC AND AUC AND AUCH

As is illustrated in Fig. 1, ROC Analysis simply plots either `tpr` against `fpr` (most common) or Sensitivity = Recall = `tpr` against Specificity = Inverse Recall = `tnr` (flipping the x-axis). In general multiple points or 'systems' are plotted corresponding classically to different thresholds against some numeric decision criterion. However, results from different algorithms may be usefully plotted, and WEKA [28] attempts to induce a curve probabilistically even in the absence of a numerical decision value – viz. where only a yes/no decision is available for each case rather than a real or multivalued rating. Such interpolation has to be done with great care and in general WEKA's probabilistic interpolations, and related AUC or error estimates, are not realistic, as will be illustrated below.

In Fig. 1 a single system is plotted along with the ever present points corresponding to always guessing positive (so Bias, `tpr` and `fpr` are all 1) or negative (so Bias, `tpr` and `fpr` are all 0). In the absence of a system, the line between the 0% and 100% Bias points corresponds to Chance and a point K% of the way along this line can be achieved by guessing with a K% Bias. With a single system that is better than chance a triangle is formed above the Chance diagonal. If the information was available but used the wrong way round (reversing the labels) a similar triangle will be formed below the diagonal as shown in Fig. 1. The area between the actual and inverted curve is often called the Gini coefficient, and for the single system case corresponds to the Dichotomous Informedness measure, B. Since the diagonal is `tpr=fpr`, the distance of a point from the triangle in the vertical dimension is B=`tpr-fpr`, from which the area of the triangles can be calculated as B/2.

Two multipoint ROC curves representing two distinct systems are illustrated in Fig. 2, one with two non-trivial points (lighter) and one with three points (darker). Each of

these points represents a system configuration that can be achieved with appropriate parameterization. The dotted lines represent the convex hull of each system, and each point along this line represents a virtual parameterization that can be achieved by probabilistic selection – the probability of a point represents its proportion of the way along the segment. The dashed line that joins the two systems completes a convex hull for a fusion of the two systems. All operating points along this convex hull are achievable using the same probabilistic techniques [2].

However not only may concavities be repaired, they may also be inverted [30], and given all decisions have unique thresholds this may be repeated until a perfect classifier is achieved, in theory. Unfortunately, we need information rather than noise, and this boosting technique will only work to a limited degree for new data, to the extent information is really used the wrong way in the original model.

The convex hull thus represents a series of effective and achievable models that can be expected to achieve a generalizable improvement in performance. The final convex hull clearly dominates the individual systems or classifier in that it is at least as good for any given `tpr` or `fpr` value.

A curve that strictly dominates another curve will clearly have greater area, and thus Area Under the Curve (AUC) has become a common heuristic for deciding between two classifiers. However, it is not particularly useful in practice for multipoint curves, for several reasons. The first of these we have seen already – it is the convex hull that really describes what can be guaranteed from the classifier, and it may even be possible to boost beyond this. Thus Area Under the Convex Hull (AUCH) is a more useful variant.

But there is a second problem, as it is the point on the curve that best matches the required cost for the real world application. AUC or AUCH is really just a heuristic for ranking classifiers, but says nothing about whether that classifier on the target problem is better or not. Thus neither AUC or AUCH should be relied on as a single definitive measure, and from this point we will only consider AUCH, and note that AUC=AUCH for a single point ROC curve, however we only recommend it for this case when it represents Informedness=2AUC-1.

It should also be noted that the gradient of the ROC curve tells us that rate at which our positive predictions are picking up positives rather than negatives, in a prevalence weighted sense – the chance diagonal, or a segment parallel to it, corresponds to picking up the same percentage of both positive and negative instances. This effectively give weight to classes in inverse proportion to their prevalence – this is how betting odds are worked out and bets laid, and winnings correspond to how good a system is in terms of single point ROC AUC or Informedness.

As is made explicit in a different way in Information Theory, the rarer cases, usually positives, are worth more than their more common negative counterparts, and the Informedness and Correlation measures are closely related to the Mutual Information and the Chi-Squared Significance of the predictions [27]. On the other hand, Rand Accuracy gives the same weight to positives and negatives irrespective of their relative frequency, viz. the cost ratio `cp/cn = 1`, where `cp` represents the value of correctly identifying a positive (or the cost of missing one), and `cn` the value of identifying a negative (or the cost of missing one).

In ROC, the cost ratio is given by `cp/cn = rn/rp` along segments parallel to the chance line, and this is reflected in Informedness. If we would like to double the cost ratio, we can simply look for a segment of the curve with double the gradient. For a system that has even one point off the diagonal, it's convex hull can be constructed so it need never dip below – we can just ignore or repair those points. It is then clear that gradient is monotonically decreasing from infinite (vertical as it leaves the origin) to zero (horizontal as it approaches (1,1)). If the curve were smooth, any gradient would correspond to at least one point on the convex hull, and thus one specific parameterized or interpolated system. Given there are discontinuities in the gradient (as in the both Figs 1 and 2), then the target gradient may be skipped, but this is in fact the point on the hull that is optimal for the specified cost.

Some authors have been concerned that AUC is hard to interpret and is integrated over an unknown cost function with a classifier dependent parameter, and thus some have argued for alternatives including AUK, based on Kappa [13,31], or H-Measure, which assumes a centrally biased distribution of costs [11].

H-Measure makes the assumption that the centre of a range is more important, but many optima are reached at boundary conditions, and many natural distributions are highly skewed, or even multimodal. AUK was developed in specific objection to this assumption [31].

To the extent Cohen Kappa is similar to Informedness, the Kappa curve is like rotating so the chance line becomes the origin, and to the extent it deviates from Informedness it is reflecting a mismatch between Prevalence and Bias. It does not remove the issues in dealing with an area integrated across a range of costs and parameters, and does not reflect a specific system, but rather the capabilities of a complex ensemble of systems. And it continues to overlook the fact that the important feature of the ROC curve or hull is not its area but its shape.

Note that all the considerations discussed here for the Recall trading ROC curves apply equally well to the Precision trading versions. Recall-ROC is used for adapting a system to the actual costs and prevalences [8] of the deployment situation, which may differ from the training situation/dataset. Precision-Inverse-precision curves are useful for examining the effect of Bias on the Markedness [27] of the predictor (how well the predictor is predicted by the real variable under different probabilities of the prediction labels, which will be prevalences where these are naturally occuring variables, and biases where they result from a classifier). On the other hand Precision-Recall curves do not have these properties and should not be used since they only address three degrees of freedom (they ignore true negatives, N, bias and precision, and are thus manipulable arbitrarily by varying any of these. These issues are discussed in [31] and [26].

V. Recommendations

In order to make effective use of a ROC curve it is easiest to look at it in relation to the costs. Isocost lines with gradients representing the extremes of the cost parameters will show where the optima for those specific costs, and those in between.

To the extent we have sharp points in the curve, representing discontinuities in the gradient, those points represent a considerable range of cost ratio.

To the extent we have concavities (deconvexities) in the curve, we have the opportunity to interpolate in that range, although note that all points along that line, including the original endpoints, will have the same gradient and thus represent the same cost ratio. Thus cost does not dictate forming a convex hull – this would be a matter of preference. For example one may prefer to be as close to (1,0) as possible, but normally one achieves this by using the isocost line that passes closest to that target, by default utilizing the prevalence weight isocost gradient of 1, viz. lines parallel to the chance diagonal. Some concavities may provide opportunity for boosting, but this is most likely if weak classifiers were used initially.

If one curve completely encloses or dominates another, the inner system can be discarded without further consideration. If system curves overlap, then there is opportunity to form a convex hull representing a fusion of the ensemble of systems. The more different the curves are, the more effective ensemble fusion or boosting techniques are likely to be in taking you beyond the convex hull. Rather than using probabilistic techniques to interpolate onto a convex hull, it is possible to use arbitrary fusion or boosting techniques. For example three or more systems could vote, with votes only counting when in particular segments of their ROC curves. For linear combination or voting systems to make an improvement there need to be substantial differences between the systems – they may be equally good overall, or dramatically different in performance, that doesn't matter. What is important is that they make different kinds of errors. If the systems are mostly correct (they are on the positive side of the chance line), then chances of error will be reduced optimally if the errors are totally independent. If three systems produce 10% totally different and independent errors, then the voting system will produce only 0.1% error. Radically different ROC curves are good evidence of independence, but even identical ROC curves could in fact represent systems with totally different errors.

For most purposes, Precision or Recall-Precision curves will not useful for comparison as bias is not corrected for.

It is in general recommended that ROC analysis be used to compare different systems, parameters versus different cost and prevalence environments possible during deployment. However to evaluate a specific system, the single point ROC should be used, or equivalently Informedness [27]. ROC-ConCert [32] actually splits the AUC into two components, Consistency reflecting how well the system ports to other situations, and Certainty reflecting Informedness (where Informedness = 2•Certainty – 1 and Certainty + Consistency = AUC).

VI. Conclusions

The idea of looking for a single best measure to choose between classifiers is wrongheaded, whether we are looking at a results of an experiment, or a particular application or dataset or simply trying to develop or optimize learning algorithms. If we have information about our application, we should use it, and in particular we should make use of any information we have about the relative cost of positive or negative cases (we assume the value and the missed opportunity cost are the same).

In the absence of cost information, the prevalence weighted or "Bookmaker" cost [26] is a good default as enshrined in both ROC and Informedness. Matching Bias to Prevalence unifies the families of measures, both unbiased and biased and thus optimizes in a broader sense. This will correspond to a unique point on the ROC curve or hull, as it moves from Bias = 1 (everything is predicted positive) to Bias = 0 (predict everything as negative).

Another interesting constraint corresponds to the other diagonal given by `tpr = tnr`. This implies that the Prevalence within the correctly classified cases matches that overall. Where a cost optimal segment crosses the cross diagonal this thus ensures an even handed system.

The best system in terms of cost is the one indicated by the highest corresponding isocost line, with gradient 1 by default like the chance line. If there is a segment parallel to the chance line or with the desired cost gradient, any system on that segment will optimize cost.

The nearest system to (1,0) for the default cost ratio will be a point on the alternate diagonal as this is perpendicular to the main diagonal and its parallel isocosts lines.

Optimizing AUC or AUCH does not say anything about whether you will optimize for your unknown costs except in the weak sense that if costs truly are unknown and expected to be different from the default cost, it allows fielding a system that has the highest probability in some sense of optimizing the unknown cost, since on average better `tpr` is reached for a given `fpr`. But fpr is a function of the classifier, and for consistency H-Measure recommends a using a consistent but fictional cost curve – we want to average over the different costs, not the different `fpr` values achievable by the classifier. But there is not reason to think that the Beta distribution used by the H-Measure is better than the classifier's `fpr` distribution.

Furthermore, we have seen that it is usually better to understand the distribution and fuse multiple undominated and thus partially independent systems to achieve a better system, potentially better even than the convex hull [30], and potentially gaining the ability to tune the system in the field once as costs become known or as they vary over time. Of course the cost of running and fusing multiple classifiers must also be taken into account, although generally training classifiers is much more expensive than using them.

Although this paper is written in a computer science context where artificial intelligence systems are assumed to be learned, there is no requirement for the systems compared or fused to be developed using machine learning techniques. They could be knowledge engineered systems, or they could

be a handbook derived from years of experience, or they could be ecological manipulations determined or parameterized from a series of experiments or simulations, or we could be analyzing the results of psychological or human factors experiments. However, generally in other applications a single best model is required, and this will easiest be derived using a single point ROC curve with the default threshold, and thus determined simply by Informedness, or it will involve use of a specific range of costs, in which case ROC analysis is a straightforward way to optimize cost. One trick that can be used to allow standard ROC analysis, as well as techniques like Informedness and Kappa to be used for optimization, is to artificially modify (multiply) the prevalence counts in the contingency table so that `rn/pn` does reflect the desired cost [8]. Another trick is to factor of the Informedness component of AUC as Certainty and disclose the variability due to situation dependence as Consistency [32].

REFERENCES

[1] D. Janez. Statistical comparisons of classifiers over multiple datasets. Journal of Machine Learning Research, 7:1–30, 2006.

[2] F. Provost and T. Fawcett. Robust classification for imprecise environments. Machine Learning, 44:203–231, 2001.

[3] J. S. Uebersax. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin* 101, 140• 146.

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

[5] J. Carletta (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* 22(2):249-254

[6] J. Cohen (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement***,** 1960:37-46.

[7] J. Cohen (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70:213-20.

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

[9] P.A. Flach (2003). The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics, *Proc. Twentieth International Conference on Machine Learning* (ICML-2003), Washington DC, 2003, pp. 226-233.

[10] J. Fürnkranz & P. A. Flach (2005). ROC 'n' Rule Learning – Towards a Better Understanding of Covering Algorithms, *Machine Learning* 58(1):39-77.

[11] D. J. Hand. Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, 77:103–123, 2009.

[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*, 45:171–186, 2001.

[13] A. Ben-David. About the relationship between ROC curves and Cohen's kappa. *Engineering Applications of AI*, 21:874–882, 2008.

[14] J.A. Cohen. A coefficient of agreement for nominal scales, educational and psychological measurement. *Psychological Measurement*, 20:37–46, 1960.

[15] J. L. Fleiss. *Statistical Methods for Rates and Proportions*. John Wiley and Sons, 2nd edition, 1981.

[16] B. Di Eugenio and M. Glass (2004), The Kappa Statistic: A Second Look., *Computational Linguistics* 30:1 95-101.

[17] A. Ben-David (2008), Comparison of classification accuracy using Cohen's Weighted Kappa, *Expert Systems with Applications* 34:825-832.

[18] M. J. Warrens (2010), Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification* 4:271-286.

[19] M. J. Warrens (2010). A formal proof of a paradox associated with Cohen's kappa. *Journal of Classificaiton* 27:322-332.

[20] M. J. Warrens (2010). A Kraemer-type rescaling that transforms the Odds Ratio into the Weighted Kappa Coefficient. *Psychometrika* 75:2 328-330.

[21] M. J. Warrens (2011). Cohen's linearly wieghted Kappa is a weighted average of 2x2 Kappas. *Psychometrika* 76:3, 471-486.

[22] T.P. Hutchinson. (1993). Focus on Psychometrics. Kappa muddles together two sources of disagreement: tetrachoric correlation is preferable. Research in Nursing & Health 16(4):313-6, 1993 Aug.

[23] D.G. Bonett & R.M. Price, (2005). Inferential Methods for the Tetrachoric Correlation Coefficient, *Journal of Educational and Behavioral Statistics* 30:2, 213-225

[24] P. Perruchet, and R.Peereman. (2004). The exploitation of distributional information in syllable processing, *J. Neurolinguistics* 17:97−119.

[25] D. R. Shanks (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology*, 48A, 57-279.

[26] D. M. W. Powers (2003), Recall and Precision versus the Bookmaker, *Proc. of the International Conference on Cognitive Science* (ICSC-2003), Sydney Australia, 2003, pp.529-534.

[27] D. M. W.Powers, (2011). Evaluation: From Precision, Recall and F-Measure to Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2:1-37-63

[28] I. H.Witten, E.Frank, M.A.Hall (2011). – *Data Mining: practical Machine Learning Tools and Techniques*, 3rd Edition, Morgan Kaufmann

[29] D. M. W. Powers (2012) The Problem of Kappa. *European Chapter of the Association for Computational Linguistics*, EACL2012, to be published.

[30] P.A. Flach & S. Wu (2005), Repairing concavities in ROC curves. *Proc. 19th International Joint Conference on Artificial Intelligence* (IJCAI'05), pp. 702–707.

[31] J. Entwisle and D. M. W. Powers (1998). "The Present Use of Statistics in the Evaluation of NLP Parsers", pp215-224, *Proc. NeMLaP3/CoNLL98 Joint Conference*, Sydney, January 1998.

[32] D. M. W. Powers (2012). "ROC-ConCert ROC-based measurement of Consistency and Certainty", *Spring World Congress on Science and Technology*, SCET2012, IEEE, to be published.