School of Informatics and Engineering
Flinders University • Adelaide • Australia

Technical Report SIE-to be allocated
December 2007

Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation

David M W Powers
School of Informatics and Engineering
Flinders University of South Australia
PO Box 2100, Adelaide 5001, South Australia

powers@infoeng.flinders.edu.au

Abstract

Commonly used evaluation measures including Recall, Precision, F-Factor, Rand Accuracy and Cohen Kappa are biased and should not be used without clear understanding of the biases, and corresponding identification of chance or base case levels of the statistic. Using these measures a system that performs worse in the objective sense of Informedness, can appear to perform better under any of these commonly used measures. We discuss several concepts and measures that reflect the probability that prediction is informed versus chance. Informedness and introduce Markedness as a dual measure for the probability that prediction is marked versus chance. Finally we demonstrate elegant connections between the concepts of Informedness, Markedness, Correlation and Significance as well as their intuitive relationships with Recall and Precision, and outline the extension from the dichotomous case to the general multi-class case.

Keywords: Recall, Precision, F-Factor, Rand Accuracy, Cohen Kappa, Chi-Squared Significance, Log-Likelihood Significance, Matthews Correlation, Pearson Correlation, Evenness, Bookmaker Informedness and Markedness

1 Introduction

A common but poorly motivated way of evaluating results of Machine Learning experiments is using Recall, Precision and F-factor. These measures are named for their origin in Information Retrieval and present specific biases, namely that they ignore performance in correctly handling negative examples, they propagate the underlying marginal prevalences and biases, and they fail to take account the chance level performance. In the Medical Sciences, Receiver Operating Characteristics (ROC) analysis has been borrowed from Signal Processing to become a standard for evaluation and standard setting, comparing True Positive Rate and False Positive Rate. In the Behavioral Sciences, Specificity and Sensitivity, are commonly used. Alternate techniques, such as Rand Accuracy and Cohen Kappa, have some advantages but are nonetheless still biased measures. We will recapitulate some of the literature relating to the problems with these measures, as well as considering a number of other techniques that have been introduced and argued within each of these fields, aiming/claiming to address the problems with these simplistic measures.

This paper recapitulates and reexamines the relationships between these various measures, develops new insights into the problem of measuring the effectiveness of an empirical decision system or a scientific experiment, analyzing and introducing new probabilistic and information theoretic measures that overcome the problems with Recall, Precision and their derivatives.

2 The Binary Case

It is common to introduce the various measures in the context of a dichotomous binary classification problem, where the labels are by convention + and − and the predictions of a classifier are summarized in a four cell contingency table. This contingency table may be expressed using raw counts of the number of times each predicted label is associated with each real class, or may be expressed in relative terms. Cell and margin labels may be formal probability expressions, may derive cell expressions from margin labels or vice-versa, may use alphabetic constant labels `a, b, c, d` or `A, B, C, D`, or may use acronyms for the generic terms for True and False, Real and Predicted Positives and Negatives. Often UPPER CASE is used where the values are counts, and lower case letters where the values are probabilities or proportions relative to `N` or the marginal probabilities - we will adopt this convention throughout this paper (always

This is an extension of papers presented at the 2003 International Cognitive Science Conference and the 2007 Human Communication Science SummerFest. Both papers, along with scripts/spreadsheets to calculate many of the statistics discussed, may be found at http://david.wardpowers.info/BM/index.htm.

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

	+R	−R	
+P	A	B	A+B
−P	C	D	C+D
	A+C	B+D	N

Table 1. Systematic and traditional notations in a binary contingency table. ***Colour coding indicates correct (green) and incorrect (pink) rates or counts in the contingency table.***

written in `typewriter font`), and in addition will use Mixed Case (in the normal text font) for popular nomenclature that may or may not correspond directly to one of our formal systematic names. True and False Positives (`TP/FP`) refer to the number of Predicted Positives that were correct/incorrect, and similarly for True and False Negatives (`TN/FN`), and these four cells sum to `N`. On the other hand `tp, fp, fn, tn` and `rp, rn` and `pp, pn` refer to the joint and marginal probabilities, and the four contingency cells and the two pairs of marginal probabilities each sum to 1. We will attach other popular names to some of these probabilities in due course.

We thus make the specific assumptions that we are predicting and assessing a single condition that is either positive or negative (dichotomous), that we have one predicting model, and one gold standard labeling. Unless otherwise noted we will also for simplicity assume that the contingency is non-trivial in the sense that both positive and negative states of both predicted and real conditions occur, so that none of the marginal sums or probabilities is zero.

We illustrate in Table 1 the general form of a binary contingency table using both the traditional alphabetic notation and the directly interpretable systematic approach. Both definitions and derivations in this paper are made relative to these labellings, although English terms (e.g. from Information Retrieval) will also be introduced for various ratios and probabilities. The green positive diagonal represents correct predictions, and the pink negative diagonal incorrect predictions. The predictions of the contingency table may be the predictions of a theory, of some computational rule or system (e.g. an Expert System or a Neural Network), or may simply be a direct measurement, a calculated metric, or a latent condition, symptom or marker. We will refer generically to "the model" as the source of the predicted labels, and "the population" or "the world" as the source of the real conditions. We are interested in understanding to what extent the model "informs" predictions about the world/population, and the world/population "marks" conditions in the model.

2.1 Recall & Precision, Sensitivity & Specificity

Recall or Sensitivity (as it is called in Psychology) is the proportion of Real Positive cases that are correctly Predicted Positive. This measures the Coverage of the Real Positive cases by the **+P** (Predicted Positive) rule. Its desirable feature is that it reflects how many of the relevant cases the **+P** rule picks up. It tends not to be very highly valued in Information Retrieval (on the assumptions that there are many relevant documents, that it doesn't really matter which subset we find, that we can't know anything about the relevance of documents that aren't returned). Recall tends to be neglected or averaged away in Machine Learning and Computational Linguistics (where the focus is on how confident we can be in the rule or classifier). However, in a Computational Linguistics/Machine Translation context Recall has been shown to have a major weight in predicting the success of Word Alignment (Fraser & Marcu, 2007). In a Medical context Recall is moreover regarded as primary, as the aim is to identify all Real Positive cases, and it is also one of the legs on which ROC analysis stands. In this context it is referred to as True Positive Rate (`tpr`). Recall is defined, with its various common appellations, by equation (1):

$$\begin{aligned} \text{Recall} &= \text{Sensitivity} = tpr = tp/rp \\ &= TP\ /\ RP = A\ /(A+C) \end{aligned} \qquad (1)$$

Conversely, Precision or Confidence (as it is called in Data Mining) denotes the proportion of Predicted Positive cases that are correctly Real Positives. This is what Machine Learning, Data Mining and Information Retrieval focus on, but it is totally ignored in ROC analysis. It can however analogously be called True Positive Accuracy (`tpa`), being a measure of accuracy of Predicted Positives in contrast with the rate of discovery of Real Positives (`tpr`). Precision is defined in (2):

$$\begin{aligned} \text{Precision} &= \text{Confidence} = tpa = tp/pp \\ &= TP\ /\ PP = A\ /(A+B) \end{aligned} \qquad (2)$$

These two measures and their combinations focus only on the positive examples and predictions, although between them they capture some information about the rates and kinds of errors made. However, neither of them captures any information about how well the model handles negative cases. Recall relates only to the **+R** column and Precision only to the **+P** row. Neither of these takes into account the number of True Negatives. This also applies to their Arithmetic, Geometric and Harmonic Means: A, G and F=G2/A (the F-factor or F-measure). Note that the F-measure effectively references the True Positives to the Arithmetic Mean of Predicted Positives and Real Positives, being a constructed rate normalized to an idealized value. The Geometric Mean of Recall and Precision (G-measure) effectively normalizes TP to the Geometric Mean of Predicted Positives and Real Positives, and its Information content corresponds to the Arithmetic Mean of the Information represented by Recall and Precision.

In fact, there is in principle nothing special about the Positive case, and we can define Inverse statistics in terms of the Inverse problem in which we interchange positive and negative and are predicting the opposite case. Inverse Recall or Specificity is thus the proportion of Real Negative cases that are correctly Predicted Negative (3), and is also known as the True Negative Rate (`tnr`). Conversely, Inverse Precision is the proportion of Predicted Negative cases that are indeed Real Negatives (4), and can also be called True Negative Accuracy (`tna`):

Inverse Recall	`= tnr`	`= tn/rn`	
	`= TN/RN`	`= D/(B+D)`	(3)
Inverse Precision	`= tna`	`= tn/pn`	
	`= TN/PN`	`= D/(C+D)`	(4)

Rand Accuracy is explicitly takes into account the classification of negatives, and is expressible (5) both as a weighted average of Precision and Inverse Precision and as a weighted average of Recall and Inverse Recall. Conversely, the Jaccard or Tanimoto similarity coefficient explicitly ignores the correct classification of negatives (TN):

Accuracy	`= tea = ter = tp+tn`		
	`= rp*tpr+rn*tnr`		
	`= pp*tpa+pn*tna`	`= (A+D)/N`	(5)
Jaccard	`= tp/(tp+fn+fp)`	`= TP/(N-TN)`	
	`= A/(A+B+C)`	`= A/(N-D)`	(6)

Each of the above also has a complementary form defining an error rate, of which some have specific names and importance: Fallout or False Positive Rate (`fpr`) are the proportion of Real Negatives that occur as Predicted Positive (ring-ins); Miss Rate or False Negative Rate (`fnr`) are the proportion of Real Positives that are Predicted Negatives (false-drops). False Positive Rate is the second of the legs on which ROC analysis is based.

Fallout	`= fpr`	`= fp/rp`	
	`= FP/RP`	`= B/(B+D)`	(7)
Miss Rate	`= fnr`	`= fn/rn`	
	`= FN/RN`	`= C/(A+C)`	(8)

Note that FN and FP are sometimes referred to as Type I and Type II Errors, and the rates `fn` and `fp` as alpha and beta, respectively - referring to falsely rejecting or accepting a hypothesis. More correctly, these terms apply specifically to the meta-level problem discussed later of whether the precise pattern of counts (not rates) in the contingency table fit the null hypothesis of random distribution rather than reflecting the effect of some alternative hypothesis (which is not in general the one represented by either **+P** -> **+R** or **−P** -> **−R** or both).

2.2 Prevalence, Bias, Cost & Skew

We now turn our attention to various forms of bias that detract from the utility of all of the above surface measures (Reeker, 2000). We will first note that `rp` represents the Prevalence of positive cases, `RP/N`, and is assumed to be a property of the population of interest - it may be constant, or it may vary across subpopulations, but is regarded here as not being under the control of the experimenter. By contrast, `pp` represents the (label) Bias of the model (Lafferty, McCallum and Pereira, 2002), the tendency of the model to output positive labels, `PP/N`, and is directly under the control of the experimenter, who can change the model by changing the theory or algorithm, or some parameter or threshold, to better fit the world/population being modeled. Note that F-factor effectively references `tp` (probability or proportion of True Positives) to the Arithmetic Mean of Bias and Prevalence. A common rule of thumb, or even a characteristic of some algorithms, is to parameterize a model so that Prevalence = Bias, viz. `rp = pp`. Corollaries of this setting are Recall = Precision (`= A = G = F`), Inverse Recall = Inverse Precision and Fallout = Miss Rate.

Alternate characterizations of Prevalence are in terms of Odds `= rp/rn` (Powers, 2003) and Skew or Class Ratio c^s `= rn/rp` (Flach, 2003), recalling that by definition `rp+rn = 1` and `RN+RP = N`. If the distribution is highly skewed, typically there are many more negative cases than positive, this means the number of errors due to poor Inverse Recall will be much greater than the number of errors due to poor Recall. Given the cost of both False Positives and False Negatives is equal, individually, the overall component of the total cost due to False Positives (as Negatives) will be much greater at any significant level of chance performance, due to the higher prevalence of Real Negatives.

Note that the normalized binary contingency table with unspecified margins has three degrees of freedom - setting any three non−Redundant ratios determines the rest (setting any count supplies the remaining information to recover the original table of counts with its four degrees of freedom). In particular, Recall, Inverse Recall and Prevalence, or equivalently tpr, fpr and c_s, suffice to determine all ratios and measures derivable from the normalized contingency table, but `N` is also required to determine significance. As another case of specific interest, Precision, Inverse Precision and Bias, in combination, suffice to determine all ratios or measures, although we will show later that an alternate characterization of Prevalence and Bias in terms of Evenness allows for even simpler relationships to be exposed.

We can also take into account a differential value for positives (`cp`) and negatives (`cn`) - this can be applied to errors as a cost (loss or debit) and/or to correct cases as a gain (profit or credit), and can be combined into a single Cost Ratio c_v `= cn/cp`. Note that the value and skew determined costs have similar effects, and may be multiplied to produce a single skew-like cost factor $c = c_v c_s$. Formulations of measures that are expressed using tpr, fpr and c_s may be made cost-sensitive by using $c = c_v c_s$ in place of $c = c_s$, or can be made skew/cost-insensitive by using `c = 1` (Flach, 2003).

2.3 ROC and PN Analyses

Flach (2003) has highlighted the utility of ROC analysis to the Machine Learning community, and characterized the skew sensitivity of many measures in that context,

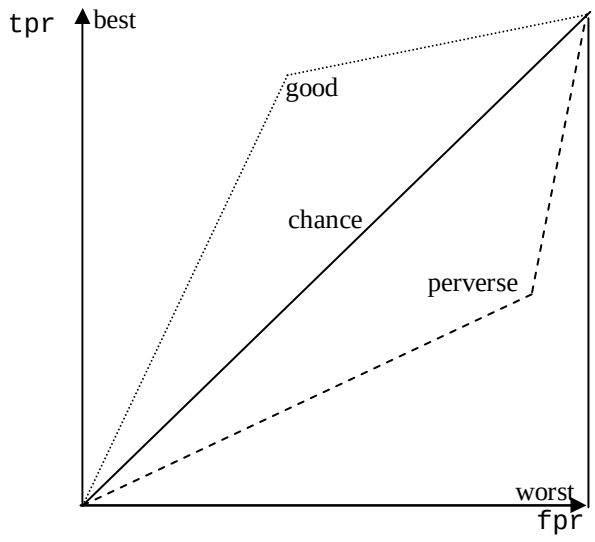

Figure 1. Illustration of ROC Analysis. *The diagonal line represents chance, points above the diagonal represent performance better than chance, those below worse than chance. For a single system, AUC is the area under the curve (trapezoid formed between the point and the x-axis). The perverse system shown is the same (good) system applied to a problem with class labels reversed.*

utilizing the ROC format to give geometric insights into the nature of the measures and their sensitivity to skew. Fürnkranz & Flach (2005) have further elaborated this analysis, extending it to the unnormalized PN variant of ROC, and targeting their analysis specifically to rule learning. We will not examine the advantages of ROC analysis here, but will briefly explain the principles and recapitulate some of the results.

ROC analysis plots the rate `tpr` against the rate `fpr`, whilst PN plots the unnormalized `TP` against `FP`. This difference in normalization only changes the scales and gradients, and we will deal only with the normalized form of ROC analysis. A perfect classifier will score in the top left hand corner (`fpr=0,tpr=100%`). A worst case classifier will score in the bottom right hand corner (`fpr=100%,tpr=0`). A random classifier would be expected to score somewhere along the positive diagonal (`tpr=fpr`) since the model will throw up positive and negative examples at the same rate (relative to their populations - these are Recall-like scales: `tpr` = Recall, `1-fpr` = Inverse Recall). For the negative diagonal (`tpr+c*fpr=1`) corresponds to matching Bias to Prevalence for a skew of `c`.

The ROC plot allows us to compare classifiers (models and/or parameterizations) and choose the one that is closest to (0,1) and furtherest from `tpr=fpr` in some sense. These conditions for choosing the optimal parameterization or model are not identical, and in fact the most common condition is to minimize the area under the curve (AUC), which for a single parameterization of a model is defined by a single point and the segments connecting it to (0,0) and (1,1). For a parameterized model it will be a monotonic function consisting of a sequence of segments from (0,0) to (1,1). A particular cost model and/or accuracy measure defines an isocost gradient, which for a skew and cost insensitive model will be `c=1`, and hence another common approach is to choose a tangent point on the highest isocost line that touches the curve. The simple condition of choosing the point on the curve nearest the optimum point (0,1) is not commonly used, but this distance to (0,1) is given by $\sqrt{[(1-tpr)^2+(1-tpr)^2]}$, and minimizing this amounts to minimizing fpr^2+fnr^2.

A ROC curve with concavities can also be locally interpolated to produce a smoothed model following the convex hull of the original ROC curve. It is even possible to locally invert across the convex hull to repair concavities, but this may overfit and thus not generalize to unseen data. Such repairs can lead to selecting an improved model, and the ROC curve can also be used to return a model to changing prevalence and costs. The area under such a multipoint curve is thus of some value, but the optimum in practice is the area under the simple trapezoid defined by the model:

$$\begin{aligned} AUC &= (tpr-fpr+1)/2 \\ &= (tpr+tnr)/2 \\ &= 1 - (fpr+fnr)/2 \end{aligned} \quad (9)$$

For the cost and skew insensitive case, with `c=1`, maximizing AUC is thus equivalent to maximizing `tpr-fpr` or minimizing `fpr+fnr`. The chance line corresponds to `tpr-fpr`=0, and parallel isocost lines for `c=1` have the form `tpr-fpr`=k. The highest isocost line also maximizes `tpr-fpr` and AUC so that these two approaches are equivalent. Minimizing fpr^2+fnr^2 instead corresponds to a distance-minimization heuristic.

We now summarize relationships between the various candidate accuracy measures as rewritten in terms of `tpr`, `fpr` and the skew, `c` (Flach,2003), as well in terms of Recall, Bias and Prevalence:

$$\begin{aligned} \text{Accuracy} &= [tpr+c\cdot(1-fpr)]/[1+c] \\ &= 2\cdot\text{Recall}\cdot\text{Prev}+1-\text{Bias}-\text{Prev} \end{aligned} \quad (10)$$

$$\begin{aligned} \text{Precision} &= tpr/[tpr+c\cdot fpr] \\ &= \text{Recall}\cdot\text{Prev}/\text{Bias} \end{aligned} \quad (11)$$

$$\begin{aligned} \text{F-Measure} &= 2\cdot tpr/[tpr+c\cdot fpr+1] \\ &= 2\cdot\text{Recall}\cdot\text{Prev}/[\text{Bias}+\text{Prev}] \end{aligned} \quad (12)$$

$$\begin{aligned} \text{WRacc} &= 4c\cdot[tpr-fpr]/[1+c]^2 \\ &= 4\cdot[\text{Recall}-\text{Bias}]\cdot\text{Prev} \end{aligned} \quad (13)$$

The last measure, Weighted Relative Accuracy, was defined by Lavrac, Flach & Zupan (1999) to subtract off the component of the True Positive score that is attributable to chance and rescale to the range ±1. Note that maximizing WRacc is equivalent to maximizing AUC or `tpr-fpr` = 2·AUC−1, as `c` is constant. Thus WRAcc is an unbiased accuracy measure, and the skew-insensitive form of WRAcc, with `c=1`, is precisely `tpr-fpr`. Each of the other measures (10−12) shows a bias in that it can not be maximized independent of skew, although skew-insensitive versions can be defined by setting `c=1`. The recasting of Accuracy, Precision and F-Measure in terms of Recall makes clear how all of these vary only in terms of the way they are affected by Prevalence and Bias.

Prevalence is regarded as a constant of the target condition or data set (and c = [1−Prev]/Prev), whilst parameterizing or selecting a model can be viewed in terms of trading of tpr and fpr as in ROC analysis, or equivalently as controlling the relative number of positive and negative predictions, namely the Bias, in order to maximize a particular accuracy measure (Recall, Precision, F-Measure, Rand Accuracy and AUC). Note that for a particular Recall level, the other measures (11−14) all decrease with increasing Bias towards positive predictions.

2.4 DeltaP, Informedness and Markedness

Powers (2003) also derived an unbiased accuracy measure to avoid the bias of Recall, Precision and Accuracy due to population prevalence and label bias. The Bookmaker algorithm costs wins and losses in the same way a fair bookmaker would set prices based on the odds. Powers then defines the concept of Informedness which represents the 'edge' a punter has in making his bet, as evidenced and quantified by his winnings. Fair pricing based on correct odds should be zero sum - that is, guessing will leave you with nothing in the long run, whilst a punter with certain knowledge will win every time. Informedness is the probability that a punter is making an informed bet and is explained in terms of the proportion of the time the edge works out versus ends up being pure guesswork. Powers defined Bookmaker Informedness for the general, K-label, case, but we will defer discussion of the general case for now and present a simplified formulation of Informedness, as well as the complementary concept of Markedness.

Definition 1

Informedness quantifies how informed a predictor is for the specified condition, and specifies the probability that a prediction is informed in relation to the condition (versus chance).

Definition 2

Markedness quantifies how marked a condition is for the specified predictor, and specifies the probability that a condition is marked by the predictor (versus chance).

These definitions are aligned with the psychological and linguistic uses of the terms condition and marker. The condition represents the experimental outcome we are trying to determine by indirect means. A marker or predictor (cf. biomarker or neuromarker) represents the indicator we are using to determine the outcome. There is no implication of causality - that is something we will address later. However there are two possible directions of implication we will address now. Detection of the predictor may reliably predict the outcome, with or without the occurrence of a specific outcome condition reliably triggering the predictor.

For the binary case we have

$$\begin{aligned}\text{Informedness} &= \text{Recall} + \text{Inverse Recall} - 1 \\ &= \text{tpr} - \text{fpr} \qquad (14)\\ \text{Markedness} &= \text{Precision} + \text{Inverse Precision} - 1 \\ &= \text{tpa} - \text{fna} \qquad (15)\end{aligned}$$

We noted above that maximizing AUC or the unbiased WRAcc measure effectively maximized tpr-fpr and indeed WRAcc reduced to this in the skew independent case. This is not surprising given both Powers and Flach set out to produce an unbiased measure, and the linear definition of Informedness will define a unique linear form. Note that while Informedness is a deep measure of how consistently the Predictor predicts the Outcome by combining surface measures about what proportion of Outcomes are correctly predicted, Markedness is a deep measure of how consistently the Outcome has the Predictor as a Marker by combining surface measures about what proportion of Predictions are correct.

In the Psychology literature, Markedness is known as DeltaP and is empirically a good predictor of human associative judgements - that is it seems we develop associative relationships between a predictor and an outcome when DeltaP is high, and this is true even when multiple predictors are in competition (Shanks, 1995). Perruchet and Peereman (2004), in the context of experiments on information use in syllable processing, note that Schanks sees DeltaP as "the normative measure of contingency", but propose a complementary, backward, additional measure of strength of association, DeltaP' aka Informedness. Perruchet and Peeremant also note the analog of DeltaP to regression coefficient, and that the Geometric Mean of the two measures is a dichotomous form of the Pearson correlation coefficient, the Matthews' Correlation Coefficient, which is appropriate unless a continuous scale is being measured dicotomously in which case a Tetrachoric Correlation estimate would be appropriate, as discussed by Bonnet and Price (2005).

2.5 Causality, Correlation and Regression

In a linear regression of two variables, we seek to predict one variable, y, as a linear combination of the other, x, finding a line of best fit in the sense of minimizing the sum of squared error (in y). The equation of fit has the form

$$\begin{aligned}y &= y_0 + r_x \cdot x \qquad \text{where}\\ r_x &= [n\textstyle\sum x\cdot y - \sum x \cdot \sum y]/[n\sum x^2 - \sum x \cdot \sum x] \qquad (16)\end{aligned}$$

Substituting in counts from the contingency table, for the regression of predicting **+R** (1) versus **−R** (0) given **+P** (1) versus **−P** (0), we obtain this gradient of best fit (minimizing the error in the real values R)−

$$\begin{aligned}r_P &= [AD - BC] / [(A+B)(C+D)]\\ &= A/(A+B) - C/(C+D)\\ &= \text{DeltaP'} = \text{Markedness} \qquad (17)\end{aligned}$$

Conversely, we can find the regression coefficient for predicting **P** from **R** (minimizing the error in the predictions P):

$r_{\mathbf{R}}$ = [AD - BC] / [(A+C)(B+D)]
= A/(A+C) - B/(B+D)
= DeltaP = Informedness (19)

Finally we see that the Matthews correlation is defined b8y

$r_{\mathbf{G}}$ = [AD - BC] / √[(A+C)(B+D)(A+B)(C+D)]
= Correlation = √[Informedness·Markedness] (19)

Given the regressions find the same line of best fit, these gradients should be reciprocal, defining a perfect Correlation of 1. However, both Informedness and Markedness are probabilities with an upper bound of 1, so perfect correlation requires perfect regression. The squared correlation is a coefficient of proportionality indicating the proportion of the variance in R that is explained by P, and is traditionally also interpreted as a probability. We can now interpret it either as the joint probability that P informs R and R marks P, given that the two directions of predictability are independent, or as the probability that the variance is (causally) explained reciprocally. Psychologists traditionally explain DeltaP in terms of causal prediction, but it is important to note that the direction of stronger prediction is not necessarily the direction of causality, and the fallacy of abductive reasoning is that the truth of A → B does not in general have any bearing on the truth of B → A.

If **Pi** is one of several independent possible causes of **R**, **Pi** -> **R** is strong, but **R** -> **Pi** is in general weak for any specific **Pi**. If **Pi** is one of several necessary contributing factors to **R**, **Pi** -> **R** is weak for any single **Pi**, but **R** -> **Pi** is strong. The directions of the implication are thus not in general dependent.

In terms of the regression to fit **R** from **P**, since there are only two correct points and two error points, and errors are calculated in the vertical (**R**) direction only, all errors contribute equally to tilting the regression down from the ideal line of fit. This Markedness regression thus provides information about the consistency of the Outcome in terms of having the Predictor as a Marker - the errors measured from the Outcome **R** relate to the failure of the Marker **P** to be present.

We can gain further insight into the nature of these regression and correlation coefficients by reducing the top and bottom of each expression to probabilities (dividing by N^2, noting that the original contingency counts sum to N, and the joint probabilities after reduction sum to 1). The numerator is the determinant of the contingency matrix, and common across all three coefficients, reducing to dtp, whilst the reduced denominator of the regression coefficients depends only on the Prevalence or Bias of the base variates. The regression coefficients, Bookmaker Informedness (B) and Markedness (M), may thus be re-expressed in terms of Precision (Prec) or Recall, along with Bias and Prevalence (Prev):

M = dtp / [Bias · (1-Bias)]
= dtp / BiasG^2 = dtp / $\text{Evenness}_{\mathbf{P}}$
= [Precision - Prevalence] / IBias (20)

B = dtp / [Prevalence · (1−Prevalence)]
= dtp / PrevG^2 = dtp / $\text{Evenness}_{\mathbf{R}}$
= [Recall - Bias] / IPrev (21)

Simlarly the Matthews/Pearson correlation is expressed in reduced form as the Geometric Mean of Bookmaker Informedness and Markedness, abbreviating their product as BookMark (BM) and recalling that it is BookMark that acts as a probability-like coefficient of determination, not its root, the Geometric Mean (BookMarkG or BMG):

BMG = dtp / √[Prev · (1−Prev) · Bias · (1-Bias)]
= dtp / [PrevG · BiasG]
= dtp / $\text{Evenness}_{\mathbf{G}}$
= √[(Recall−Bias)·(Prec−Prev)]/(IPrev·IBias) (22)

These equations clearly indicate how the Bookmaker coefficients of regression and correlation depend only on the proportion of True Positives and the Prevalence and Bias applicable to the respective variables. Furthermore, Prev · Bias represents the Expected proportion of True Positives (etp) relative to N, showing that the coefficients each represent the proportion of Delta True Positives (the deviation from expectation, dtp=tp-etp) renormalized in different ways to give different probabilities. Equations (20-22) illustrate this, showing that these coefficients depend only on dtp and either Prevalence, Bias or their combination. Note that for a particular dtp these coefficients are minimized when the Prevalence and/or Bias are at the evenly biased 0.5 level, however in a learning or parameterization context changing the Prevalence or Bias will in general change both tp and etp, and hence can change dtp.

It is also worth considering further the relationship of the denominators to the Geometric Means, PrevG of Prevalence and Inverse Prevalence (IPrev = 1−Prev is prevalence of Real Negatives) and BiasG of Bias and Inverse Bias (IBias = 1−Bias is bias to Predicted Negatives). These Geometric Means represent the Evenness of Real classes ($\text{Evenness}_{\mathbf{R}}$ = PrevG^2) and Predicted labels ($\text{Evenness}_{\mathbf{P}}$ = BiasG^2). We also introduce the concept of Global Evenness as the Geometric Mean of these two natural kinds of Evenness, $\text{Evenness}_{\mathbf{G}}$. From this formulation we can see that for a given relative delta of true positive prediction above expectation (dtp), the correlation is at minimum when predictions and outcomes are both evenly distributed (√$\text{Evenness}_{\mathbf{G}}$ = √$\text{Evenness}_{\mathbf{R}}$ = √$\text{Evenness}_{\mathbf{P}}$ = Prev = Bias = 0.5), and Markedness and Bookmaker are individually minimal when Bias resp. Prevalence are evenly distributed (viz. Bias resp. Prev = 0.5). This suggests that setting Learner Bias (and regularized, cost-weighted or subsampled Prevalence) to 0.5, as sometimes performed in Artificial Neural Network training is in fact inappropriate on theoretical grounds, as has previously been shown both empirically and based on Bayesian principles - rather it is best to use Learner Bias = Natural Prevalence which is in general much less than 0.5 (Lisboa and Wong, 2000).

Note that in the above equations (20-22) the denominator is always strictly positive since we have occurrences and predictions of both Positives and Negatives by earlier

assumption, but we note that if in violation of this constraint we have a degenerate case in which there is nothing to predict or we make no effective prediction, then tp=etp and dtp=0, and all the above regression and correlation coefficients are defined in the limit approaching zero. Thus the coefficients are zero if and only if dtp is zero, and they have the same sign as dtp otherwise. Assuming that we are using the model the right way round, then dtp, B and M are non-negative, and BMG is similarly non-negative as expected. If the model is the wrong way round, then dtp, B, M and BMG can indicate this by expressing below chance performance, negative regressions and negative correlation, and we can reverse the sense of **P** to correct this.

The unnormalized determinant of the contingency matrix, dp = dtp, in these probability formulae (20-22) also has a geometric interpretation as the area of a trapezoid in PN-space, the unnormalized variant of ROC (Fürnkranz & Flach, 2005). We already observed that in (normalized) ROC analysis, (normalized) Informedness is twice the triangular between a positively informed system and the chance line, and it thus corresponds to the area of the trapezoid defined by a system (assumed to perform no worse than chance), and its perversions (interchanging prediction labels but not the real classes, or vice-versa, so as to derive a system that performs no better than chance), and the endpoints of the chance line (the trivial cases in which the system labels all cases true or conversely all are labeled false). This kite-shaped area is delimited by the dotted (system) and dashed (perversion) lines in Figure 1. The Informedness of the perversion is the negation of the Informedness of the correctly polarized system.

We can now also express the Informedness and Markedness forms of DeltaP in terms of deviations from expected values along with the Harmonic mean of the marginal cardinalities of the Real classes or Predicted labels respectively, defining DP, DELTAP, RH, PH and related forms in terms of their N−Relative probabilistic forms defined as follows:

$$etp = rp \cdot pp;\ etn = rn \cdot pn \tag{23}$$

$$\begin{aligned} dp &= tp - etp = dtp \\ &= -dtn = -(tn - etn) \\ deltap &= dtp - dtn = 2dp \end{aligned} \tag{24}$$

$$\begin{aligned} rh &= 2rp \cdot rn \,/\, [rp+rn] \\ ph &= 2pp \cdot pn \,/\, [pp+pn] \end{aligned} \tag{25}$$

DeltaP' or Bookmaker Informedness may now be expressed in terms of deltap and rh, and DeltaP or Markedness analogously in terms of deltap and ph:

$$\begin{aligned} B = DeltaP' &= [etp+dtp]/rp - [efp-dtp]/rn \\ &= etp/rp - efp/rn + 2dtp/rh \\ &= 2dp/rh = deltap/rh \end{aligned} \tag{26}$$

$$M = DeltaP = 2dp/ph = deltap/ph \tag{27}$$

These Harmonic relationships connect directly with the previous Geometric relationships by observing that ArithmeticMean = GeometricMean^2/HarmonicMean (0.5 for marginal rates and N/2 for marginal counts). The interpretation in terms of GeometricMean is preferred as an estimate of central tendency that more accurately estimates the mode for well distributed (e.g. Poisson) data, and as the central limit of the family of Lp based averages (note that the Geometric Mean is the Geometric Mean of the Harmonic and Arithmetic Means).

2.6 Effect of Bias & Prev on Recall & Precision

The final form of the equations (20-22) cancels out the common Bias and Prevalence (Prev) terms, converting tp to tpr (Recall) or tpa (Precision). We now recast the Bookmaker Informedness and Markedness equations to show Recall and Precision as subject (23-24), in order to explore the affect of Bias and Prevalence on Recall and Precision, as well as clarify the relationship of Bookmaker and Markedness to these ubiquitous and iniquitous measures.

$$\begin{aligned} Recall &= Bookmaker\ (1-Prevalence) + Bias \\ Bookmaker &= (Recall-Bias) / (1-Prevalence) \end{aligned} \tag{28}$$

$$\begin{aligned} Precision &= Markedness\ (1-Bias) + Prevalence \\ Markedness &= (Precision-Prev) / (1-Bias) \end{aligned} \tag{29}$$

Bookmaker and Markedness are unbiased estimators of above chance performance (relative to respectively the predicting conditions or the predicted markers). Equations (23-24) clearly show the nature of the bias introduced by both Label Bias and Class Prevalence. If operating at chance level, both Bookmaker and Markedness will be zero, and Recall, Precision, and derivatives such as the F-measure, will merely reflect the biases. Note that increasing Bias or decreasing Prevalence increases Recall and decreases Precision, for a constant level of unbiased performance. We can more specifically see that the regression coefficient for the prediction of Recall from Prevalence is −Bookmaker and from Bias is +1, and similarly the regression coefficient for the prediction of Precision from Bias is −Markedness and from Prevalence is +1.

In summary, Recall reflects the Bias plus a discounted estimation of Informedness and Precision reflects the Prevalence plus a discounted estimation of Markedness. Given usually Prevalence << ½ and Bias << ½, their complements Inverse Prevalence >> ½ and Inverse Bias >> ½ represent substantial weighting up of the true unbiased performance in both these measures, and hence also in F-factor. High Bias drives Recall up strongly and Precision down according to the strength of Informedness; high Prevalence drives Precision up and Recall down according to the strength of Markedness.

Alternately, Informedness can be viewed as a renormalization of Recall after subtracting off the Bias, and Markedness can be seen as a renormalization of Precision after subtracting off the Prevalence (and Flach's WRAcc, the unbiased form being equivalent to Bookmaker Informedness, was defined in this way as discussed in §2.3). The Kappa measure (Cohen, 1960/1968; Carletta, 1996) commonly used in assessor agreement evaluation was similarly defined as a

renormalization of Accuracy after subtracting off the expected Accuracy as estimated by the dot product of the Biases and Prevalences, and is expressible as a normalization of the discriminant of contingency, `deltap`, by the mean error rate (viz. Kappa is `deltap/[deltap`+mean(`fp`,`fn`)]). All three measures are invariant in the sense that they are properties of the contingency tables that remain unchanged when we flip to the Inverse problem (interchange positive and negative for both conditions and predictions). That is we observe:

Inverse Informedness = Informedness,
Inverse Markedness = Markedness,
Inverse Kappa = Kappa.

The Dual problem (interchange antecedent and consequent) reverses which condition is the predictor and the predicted condition, and hence interchanges Precision and Recall, Prevalence and Bias, as well as Markedness and Informedness. For cross-evaluator agreement, both Informedness and Markedness are meaningful although the polarity and orientation of the contingency is arbitrary. Similarly when examining causal relationships (conventionally DeltaP vs DeltaP'), it is useful to evaluate both deductive and abductive directions in determining the strength of association. For example, the connection between cloud and rain involves cloud as *one* causal antecedent of rain (but sunshowers occur occasionally), and rain as *one* causal consequent of cloud (but cloudy days aren't always wet) – only once we have identified the full causal chain can we reduce to equivalence, and lack of equivalence may be a result of unidentified causes, alternate outcomes or both.

Note that the effect of Prevalence on Accuracy, Recall and Precision has also been characterized above (§2.3) in terms of Flach's demonstration of how skew enters into their characterization in ROC analysis, and effectively assigns different costs to (False) Positives and (False) Negatives. This can be controlled for by setting the parameter `c` appropriately to reflect the desired skew and cost tradeoff, with `c=1` defining skew and cost insensitive versions. However, only Informedness (or equivalents such as DeltaP' and skew-insensitive WRAcc) precisely characterizes the probability with which a model informs the condition, and conversely only Markedness (or DeltaP) precisely characterizes the probability that a condition marks (informs) the predictor. Similarly, only the Correlation (aka Coefficient of Proportionality aka Coefficient of Determination aka Squared Matthews Correlation Coefficient) precisely characterizes the probability that condition and predictor inform/mark each other, under our dichotomous assumptions. Note the Tetrachoric Correlation is another estimate of the Pearson Correlation made under the alternate assumption of an underlying continuous variable (assumed normally distributed), and is appropriate if we instead assume that we are dichotomizing a normal continuous variable (Hutchison, 1993). But in this article we are making the explicit assumption that we are dealing with a right/wrong dichotomy that is intrinsically discontinuous.

Although Kappa does attempt to renormalize a debiased estimate of Accuracy, and is thus much more meaningful than Recall, Precision, Accuracy, and their biased derivatives, it is intrinsically non-linear, doesn't account for error well, and retains an influence of bias, so that there does not seem that there is any situation when Kappa would be preferable to Correlation as a standard independent measure of agreement (Uebersax, 1987; Bonett & Price, 2005). As we have seen, Bookmaker Informedness, Markedness and Correlation reflect the discriminant of relative contingency normalized according to different Evenness functions of the marginal Biases and Prevalences, and reflect probabilities relative to the corresponding marginal cases. However, we have seen that Kappa scales the discriminant in a way that reflects the actual error without taking into account expected error due to chance, and in effect it is really just using the discriminant to scale the actual mean error: Kappa is `dp`/[`dp`+mean(`fp`,`fn`)] = 1/[1+mean(`fp`,`fn`)/`dp`] which approximates for small error to `1-`mean(`fp`,`fn`)/`dp`.

2.7 Significance and Information Gain

The ability to calculate various probabilities from a contingency table says nothing about the significance of those numbers – is the effect real, or is it within the expected range of variation around the values expected by chance? Usually this is explored by considering deviation from the expected values (`ETP` and its relatives) implied by the marginal counts (`RP`, `PP` and relatives) – or from expected rates implied by the biases (Class Prevalence and Label Bias). In the case of Machine Learning, Data Mining, or other artificially derived models and rules, there is the further question of whether the training and parameterization of the model has set the 'correct' or 'best' Prevalence and Bias (or Cost) levels. Furthermore, should this determination be undertaken by reference to the model evaluation measures (Recall, Precision, Informedness, Markedness and their derivatives), or should the model be set to maximize the significance of the results?

There are several schools of thought about significance testing, but all agree on the utility of calculating a p-value (see e.g. Berger, 1985), by specifying some statistic or exact test T(X) and setting p = Prob(T(X) ≥ T(Data)). In our case, the Observed Data is summarized in a contingency table and there are a number of tests which can be used to evaluate the significance of the contingency table. For example, Fisher's exact test calculates the proportion of contingency tables that are at least as favorable to the Prediction/Marking hypothesis, rather than the Null hypotheis, and provides an accurate estimate of the significance of the entire contingency table without any constraints on the values or distribution. The log-likelihood-based G^2 test and Pearson's approximating χ^2 tests are compared against a Chi-Squared Distribution of appropriate degree of freedom (1 for the binary contingency table given the marginal counts are known), and depend on assumptions about the distribution, and typically focus only on the Predicted Positives.

χ^2 captures the Total Squared Deviation relative to expectation, is usually calculated only in relation to positive predictions, and is valid only for reasonably sized contigencies (one rule of thumb is that the smallest cell is at least 5, see e.g. Lowry, 2000):

$$\begin{aligned}\chi^2_{+P} &= (TP-ETP)^2/ETP + (FP-EFP)^2/EFP \\ &= DTP^2/ETP + DFP^2/EFP \\ &= 2DP^2/EHP,\ EHP = 2ETP\cdot EFP/[ETP+EFP] \\ &= 2N\cdot dp^2/ehp, ehp = 2etp\cdot efp/[etp+efp] \\ &= 2N\cdot dp^2/[rh\cdot pp] = N\cdot r^2_P\cdot PrevG^2/Bias \\ &= N\cdot B^2\cdot Evenness_R/Bias \\ &\approx (N+PN)\cdot r^2_P\cdot PrevG^2 \qquad (Bias \to 1) \\ &= (N+PN)\cdot B^2\cdot Evenness_R \qquad (30)\end{aligned}$$

G^2 captures Total Information Gain, being N times the Average Information Gain in nats, otherwise known as Mutual Information, which however is normally expressed in bits. We will discuss this separately under the General Case. We deal with G^2 for positive predictions in the case of small effect, that is dp close to zero, showing that G^2 is twice as sensitive as χ^2 in this range.

$$\begin{aligned}G^2_{+P}/_2 &= TP\cdot ln(TP/ETP) + FP\cdot ln(FP/EFP) \\ &= TP\cdot ln(1+DTP/ETP)+FP\cdot ln(1+DFP/EFP) \\ &\approx TP\cdot(DTP/ETP) + FP\cdot(DFP/EFP) \\ &= 2N\cdot dp^2/ehp \\ &= 2N\cdot dp^2/[rh\cdot pp] = N\cdot r^2_P\cdot PrevG^2/Bias \\ &= N\cdot B^2\cdot Evenness_R/Bias \\ &\approx (N+PN)\cdot r^2_P\cdot PrevG^2 \qquad (Bias \to 1) \\ &= (N+PN)\cdot B^2\cdot Evenness_R \qquad (31)\end{aligned}$$

This result (31-32) shows that χ^2 and G^2 significance of the Informedness effect increases with N as expected, but also with the square of Bookmaker, the Evenness of Prevalence ($Evenness_R = PrevG^2 = Prev\cdot(1-Prev)$) and the number of Predicted Negatives (viz. with Inverse Bias)! This is as expected. The more Informed the contingency regarding positives, the less data will be needed to reach significance. The more Biased the contingency towards positives, the less significant each positive is and the more data is needed to ensure significance. The Bias-weighted average over all Predictions (here for K=2 case: Positive and Negative) is simply $KN\cdot B^2\cdot PrevG^2$ which gives us an estimate of the significance without focussing on either case in particular.

$$\begin{aligned}\chi^2_{KB} &= 2N\cdot dtp/Evenness_R = 2N\cdot r_P^2\cdot PrevG^2 \\ &= 2N\cdot r_P^2\cdot Evenness_R \\ &= 2N\cdot B^2\cdot Evenness_R \qquad (32)\end{aligned}$$

Analogous formulae can be derived for the significance of the Markedness effect for positive real classes, noting that $Evenness_P = BiasG^2$.

$$\begin{aligned}\chi^2_{+R} &= 2N\cdot dp^2/[ph\cdot rp] = N\cdot r_R^2\cdot BiasG/Prev \\ &= N\cdot M^2\cdot Evenness_P/Prev \\ &\approx (N+RN)\cdot M^2\cdot BiasG^2 \qquad (Bias \to 1) \\ &= (N+RN)\cdot r_R^2\cdot Evenness_P \qquad (33)\end{aligned}$$

$$\begin{aligned}\chi^2_{KM} &= 2N\cdot dtp/Evenness_G = 2N\cdot r_R^2\cdot BiasG^2 \\ &= 2N\cdot r_R^2\cdot BiasG^2 \\ &= 2N\cdot M^2\cdot Evenness_P \qquad (34)\end{aligned}$$

The Geometric Mean of these two overall estimates for the full contingency table is

$$\begin{aligned}\chi^2_{KBM} &= 2N\cdot dtp/Evenness_G = 2N\cdot r_P\cdot r_R\cdot PrevG\cdot BiasG \\ &= 2N\cdot r^2_G\cdot Evenness_G \\ &= 2N\cdot B\cdot M\cdot Evenness_G \qquad (35)\end{aligned}$$

This is simply the total sum of squares variance accounted for by the correlation coefficient BMG (22) over the N data points discounted by the Global Evenness factor, being the squared Geometric Mean of all four Positive and Negative Bias and Prevalence terms ($Evenness_G$ = $PrevG\cdot BiasG$). The less even the Bias and Prevalence, the more data will be required to achieve significance, the maximum evenness value of 0.5 being achieved with both even bias and even prevalence. Note that for even Bias or Prevalence, the corresponding positive and negative significance estimates match the global estimate.

Note that this is comparable to full contingency table estimation of p by the Fisher Exact Test (except for the distributional assumption) and is independent of any alternate hypothesis. Based on a Bayesian equal probability prior for the null hypothesis (H_0) and an unspecific one-tailed alternate hypothesis (H_A, e.g. that the current effect represents the correct estimate of accuracy), new posterior probability estimates for Type I (H_0 rejection, Alpha(p)) and Type II (H_A rejection, Beta(p)) errors can be estimated from the posthoc likelihood estimation (Sellke, Bayari and Berger, 1999):

$$\begin{aligned}L(p) &= Alpha(p)/Beta(p) \\ &\approx -\,e\,p\,log(p) \qquad (36)\end{aligned}$$

$$Alpha(p) = 1/[1+1/L(p)] \qquad (37)$$

$$Beta(p) = 1/[1+L(p)] \qquad (38)$$

3 Simple Examples

Bookmaker Informedness has been defined as the Probability of an informed decision, and we have shown identity with DeltaP' and WRAcc, and the close relationship (10, 15) with ROC AUC. A system that makes an informed (correct) decision for a target condition with probability B, and guesses the remainder of the time, will exhibit a Bookmaker Informedness (DeltaP') of B and a Recall of $B\cdot(1-Prev)$ + Bias. Conversely a proposed marker which is marked (correctly) for a target condition with probability M, and according to chance the remainder of the time, will exhibit a Markedness (DeltaP) of M and a Precision of $M\cdot(1-Bias)$ + Prev. Precision and Recall are thus biased by Prevalence and Bias, and variation of system parameters can make them rise or fall independently of Informedness and Markedness. Accuracy is similarly dependent on Prevalence and Bias:

$2\cdot(B\cdot(1-Prev)\cdot Prev+Bias\cdot Prev)+1-(Bias+Prev)$,

and Kappa has an additional problem of non-linearity due to its complex denominator:

	60.0%	40.0%												
42.0%	30	12	42	B	20.00%	Rec	50.00%	F	58.82%	α=0.05	3.85		
58.0%	30	28	58	M	19.70%	Prec	71.43%	G	59.76%	χ^2_{+P}	2.29	χ^2_{KB}	1.92
	60	40	100	C	19.85%	Rac	58.00%	κ	18.60%	χ^2_{+R}	2.22	χ^2_{KM}	1.89
										χ^2	2.29	χ^2_{KBM}	1.91

	68.0%	32.0%												
76.0%	56	20	76	B	19.85%	Rec	82.35%	F	77.78%	α=0.05	3.85		
24.0%	12	12	24	M	23.68%	Prec	73.68%	G	77.90%	χ^2_{+P}	1.13	χ^2_{KB}	1.72
	68	32	100	C	21.68%	Rac	68.00%	κ	21.26%	χ^2_{+R}	1.61	χ^2_{KM}	2.05
										χ^2	1.13	χ^2_{KBM}	1.87

Table 2. Binary contingency tables. *Colour coding is as in Table 1, showing example counts of correct (green) and incorrect (pink) decisions and the resulting Bookmaker Informedness* (B=WRacc=DeltaP'), *Markedness* (C=DeltaP), *Matthews Correlation* (C), *Recall* (Rec), *Precision* (Prec), *Rand Accuracy* (Rac), *Harmonic Mean of Recall and Precision* (F), *Geometric Mean of Recall and Precision* (G), *Cohen Kappa* (κ),*and* χ^2 *calculated using Bookmaker* (χ^2_{+P}), *Markedness* (χ^2_{+R}) *and standard* (χ^2) *methods across the positive prediction or condition only, as well as calculated across the entire* K*=2 class contingency using the newly proposed methods, all of which are designed to be referenced to alpha* (α) *according to the* χ^2 *distribution, and are more reliable due to taking into account all contingencies.. The single-tailed threshold is shown for* α=0.05.

B·(1−Prev)·Prev / (1-Bias·Prev-(Bias+Prev)/2).

It is thus useful to illustrate how each of these other measures can run counter to an improvement in overall system performance as captured by Informedness. For the examples in Table 2 (for N=100) all the other measure rise, some quite considerably, but Bookmaker actually falls. Table 2 also illustrates the usage of the Bookmaker and Markedness variants of the χ^2 statistic versus the standard formulation for the positive case, showing also the full K class contingency version (for K=2 in this case).

Note that under the distributional and approximative assumptions for χ^2 neither of these contingencies differ sufficiently from chance at N=100 to be significant to the 0.05 level due to the low Informedness Markedness and Correlation, however doubling the performance of the system would suffice to achieve significance at N=100 given the Evenness specified by the Prevalences and/or Biases). Moreover, even at the current performance levels the Inverse (Negative) and Dual (Marking) Problems show higher χ^2 significance, approaching the 0.05 level in some instances (and far exceeding it for the Inverse Dual). The KB variant gives a single conservative significance level for the entire table, sensitive only to the direction of proposed implication, and is thus to be preferred over the standard versions that depend on choice of condition.

Incidentally, the Fisher Exact Test shows significance to the 0.05 level for both the examples in Table 2, corresponding to an assumption of a hypergeometric distribution rather than normality - viz. all assignments of events to the cells of the contingency tables are assumed to be equally likely irrespective of the true means and standard deviations.

4 Practical Considerations

If we have a fixed size dataset, then it is arguably sufficient to maximize the determinant of the unnormalized contingency matrix, DT. However this is not comparable across datasets of different sizes, and we thus need to normalize for N, and hence consider the determinant of the normalized contingency matrix, dt. However, this value is still influenced by both Bias and Prevalence.

In the case where two evaluators or systems are being compared with no a priori preference, the Correlation gives the correct normalization by their respective Biases, and is to be preferred to Kappa.

In the case where an unimpeachable Gold Standard is employed for evaluation of a system, the appropriate normalization is for Prevalence or Evenness of the real gold standard values, giving Informedness. Since this is constant, optimizing Informedness and optimizing dt are equivalent.

More generally, we can look not only at what proposed solution best solves a problem, by comparing Informedness, but which problem is most usefully solved by a proposed system. In a medical context, for example, it is usual to come up with potentially useful medications or tests, and then explore their effectiveness across a wide range of complaints. In this case Markedness may be appropriate for the comparison of performance across different conditions.

Recall and Informedness, as biased and unbiased variants of the same measure, are appropriate for testing effectiveness relative to a set of conditions, and the importance of Recall is being increasingly recognized as having an important role in matching human performance, for example in Word Alignment for Machine Translation (Fraser and Marcu, 2007). Precision and Markedness, as biased and unbiased variants of the same measure, are appropriate for testing effectiveness relative to a set of predictions. This is particularly appropriate where we do not have an appropriate gold standard giving correct labels for every case, and is the primary measure used in Information Retrieval for this reason, as we cannot know the full set of relevant documents for a query and thus cannot calculate Recall.

However, in this latter case of an incompletely characterized test set, we do not have a fully specified contingency matrix and cannot apply any of the other measures we have introduced. Rather, whether for

Information Retrieval or Medical Trials, it is assumed that a test set is developed in which all real labels are reliably (but not necessarily perfectly) assigned. Note that in some domains, labels are assigned reflecting different levels of assurance, but this has lead to further confusion in relation to possible measures and the effectiveness of the techniques evaluated (Fraser and Marcu, 2007). In Information Retrieval, the labelling of a subset of relevant documents selected by an initial collection of systems can lead to relevant documents being labelled as irrelevant because they were missed by the first generation systems - so for example systems are actually penalized for improvements that lead to discovery of relevant documents that do not contain all specified query words. Thus here too, it is important to develop test sets that of appropriate size, fully labelled, and appropriate for the correct application of both Informedness and Markedness, as unbiased versions of Recall and Precision.

This Information Retrieval paradigm indeed provides a good example for the understanding of the Informedness and Markedness measures. Not only can documents retrieved be assessed in terms of prediction of relevance labels for a query using Informedness, but queries can be assessed in terms of their appropriateness for the desired documents using Markedness, and the different kinds of search tasks can be evaluated with the combination of the two measures. The standard Information Retrieval mantra that we do not need to find *all* relevant documents (so that Recall or Informedness is not so relevant) applies only where there are huge numbers of documents containing the required information and a small number can be expected to provide that information with confidence. However another kind of Document Retrieval task involves a specific and rather small set of documents for which we need to be confident that all or most of them have been found (and so Recall or Informedness are especially relevant). This is quite typical of literature review in a specialized area, and may be complicated by new developments being presented in quite different forms by researchers who are coming at it from different directions, if not different disciplinary backgrounds. A good example of this is the decade it has taken to find the literature that discusses the concept variously known as Edge, Informedness, Regression, DeltaP' and ROC AUC - and perhaps this wheel has been invented in yet other contexts as well.

5 The General Case

So far we have examined only the binary case with dichotomous Positive versus Negative classes and labels.

It is beyond the scope of this article to consider the continuous or multi-valued cases, although the Matthews Correlation is a discretization of the Pearson Correlation with its continuous-valued assumption, and the Spearman Rank Correlation is an alternate form applicable to arbitrary discrete value (Likert) scales, and Tetrachoric Correlation is available to estimate the correlation of an underlying continuous scale. If continuous measures corresponding to Informedness and Markedness are required due to the canonical nature of one of the scales, the corresponding Regression Coefficients are available.

It is however, useful in concluding this article to consider briefly the generalization to the multi-class case, and we will assume that both real classes and predicted classes are categorized with K labels, and again we will assume that each class is non-empty unless explicitly allowed (this is because Precision is ill-defined where there are no predictions of a label, and Recall is ill-defined where there are no members of a class).

Powers (2003) derives Bookmaker Informedness (41) analogously to Mutual Information & Conditional Entropy (39-40) as a pointwise average across the contingency cells, expressed in terms of label probabilities $P_{\mathbf{P}}(l)$, where $P_{\mathbf{P}}(l)$ is the probability of Prediction l, and label-conditioned class probabilities $P_{\mathbf{R}}(c|l)$, where $P_{\mathbf{R}}(c|l)$ is the probability that the Prediction labeled l is actually of Real class c, and in particular $P_{\mathbf{R}}(l|l)$= Precision(l), and where the delta functions are mathematical shorthands for Boolean expressions interpreted algorithmically as in C, with true expressions taking the value 1 and false expressions 0, so that $\delta_{cl} = (c = l)$ represents the standard Dirac delta function and $\partial_{cl} = (c \neq l)$ its complement.

$$\mathrm{MI}(\mathbf{R}||\mathbf{P}) = \sum_l P_{\mathbf{P}}(l) \sum_c P_{\mathbf{R}}(c|l) [-\log(P_{\mathbf{R}}(c|l))/P_{\mathbf{R}}(c)] \quad (39)$$

$$\mathrm{H}(\mathbf{R}|\mathbf{P}) = \sum_l P_{\mathbf{P}}(l) \sum_c P_{\mathbf{R}}(c|l) [-\log(P_{\mathbf{R}}(c|l))] \quad (40)$$

$$\mathrm{B}(\mathbf{R}|\mathbf{P}) = \sum_l P_{\mathbf{P}}(l) \sum_c P_{\mathbf{R}}(c|l) [P_{\mathbf{P}}(l)/(P_{\mathbf{R}}(l) - \partial_{cl})] \quad (41)$$

We now define a binary dichotomy for each label l with l and the corresponding c as the Positive cases (and all other labels/classes grouped as the Negative case). We now denote its prevalence Prev(l) and its dichotomous Bookmaker Informedness B(l), and thus can simplify (41) to

$$\mathrm{B}(\mathbf{R}|\mathbf{P}) = \sum_l \mathrm{Prev}(l)\, \mathrm{B}(l) \quad (42)$$

Analogously we define dichotomous Bias(c) and Markedness(c) and derive

$$\mathrm{M}(\mathbf{P}|\mathbf{R}) = \sum_c \mathrm{Bias}(c)\, \mathrm{M}(c) \quad (43)$$

These formulations remain consistent with the definition of Informedness as the probability of an informed decision versus chance, and Markedness as its dual. The Geometric Mean of multi-class Informedness and Markedness would appear to give us a new definition of Correlation, whose square provides a well defined Coefficient of Determination. Recall that the dichotomous forms of Markedness (20) and Informedness (21) have the determinant of the contingency matrix as common numerators, and have denominators that relate only to the margins, to Prevalence and Bias respectively. Correlation, Markedness and Informedness are thus equal when Prevalence = Bias. The dichotomous Correlation Coefficient would thus appear to have three factors, a common factor across Markedness and Informedness, representing their conditional dependence, and factors representing Evenness of Bias (cancelled in Markedness) and Evenness of Prevalence (cancelled in Informedness), each representing a marginal independence.

In fact, Bookmaker Informedness can be driven arbitrarily close to 0 whilst Markedness is driven arbitrarily close to 1, demonstrating their independence - in this case Recall and Precision will be driven to or close to 1. The arbitrarily close hedge relates to our assumption that all predicted and real classes are non-empty, although appropriate limits could be defined to deal with the divide by zero problems associated with these extreme cases. Technically, Informedness and Markedness are conditionally independent - once the determinant numerator is fixed, their values depend only on their respective marginal denominators which can vary independently. To the extent that they are independent, the Coefficient of Determination acts as the joint probability of mutual determination, but to the extent that they are dependent, the Correlation Coefficient itself acts as the joint probability of mutual determination.

These conditions carry over to the definition of Correlation in the multi-class case as the Geometric Mean of Markedness and Informedness - once all numerators are fixed, the denominators demonstrate marginal independence.

In relation to Significance, the single class $\chi_{+\mathbf{P}}^2$ and $G_{+\mathbf{P}}^2$ definitions both can be formulated in terms of cell counts and a function of ratios, and would normally be summed over cells of $(K-1)^2$ a K-class contingency table with $(K-1)^2$ degrees of freedom to produce a statistic for the table as a whole. However, these statistics are not independent of which variables are selected for evaluation or summation, and the p-values obtained are thus quite misleading, and for highly skewed distributions (in terms of Bias or Prevalence) can be outlandishly incorrect. If we sum log-likelihood (31) over all K^2 cells we get $N \cdot MI(\mathbf{R}||\mathbf{P})$ which is invariant over Inverses and Duals. The analogous Prevalence-weighted multi-class statistic derived from the Bookmaker Informedness form of the Significance statistic, and the Bias-weighted statistic derived from the Markedness form, extend Eqns 32, 34 & 35 to the K>2 case by appropriate summation across all K^2 cells:

$$\chi^2_{\mathbf{KB}} = KN \cdot B^2 \cdot Evenness_{\mathbf{R}} = KN \cdot r_{\mathbf{P}}^2 \cdot Evenness_{\mathbf{R}} \quad (39)$$

$$\chi^2_{\mathbf{KM}} = KN \cdot M^2 \cdot Evenness_{\mathbf{P}} = KN \cdot r_{\mathbf{R}}^2 \cdot Evenness_{\mathbf{P}} \quad (40)$$

$$\chi^2_{\mathbf{KMB}} = KN \cdot B \cdot M \cdot Evenness_{\mathbf{G}} = KN \cdot r_{\mathbf{P}} \cdot r_{\mathbf{R}} \cdot Evenness_{\mathbf{G}} \quad (41)$$

A further problem with the standard approach applied to K-class contingency tables is the $(K-1)^2$ degrees of freedom which assumes independence of the counts in $(K-1)^2$ of the cells. This is appropriate for the null hypothesis and the calculation of `alpha`, but is patently not the case when the cells are generated by K condition variables and K prediction variables that mirror them. Thus some kind of correction is in order for the calculation of `beta`. Whilst many corrections are possible, in this case correcting the degrees of freedom directly seems appropriate and whilst using $r = (K-1)^2$ degrees of freedom is appropriate for `alpha`, using $r = K-1$ degrees of freedom is suggested for `beta` under the conditions where significance is worth testing, given the association (mirroring) between the variables is almost complete.

At this stage we still have to define the multi-class extension of the Evenness terms, but considering the way (21) is substituted into (30) in the formulation of the Informedness significance statistics, it seems clear that the Harmonic mean should be used. Since due to the marginal constraints on Prevalence and Inverse Prevalence, as noted in relation to (26) and (27), the square of their Geometric Mean is simply half the Harmonic Mean. Thus $Evenness_{\mathbf{R}}$ is simply half the Harmonic Mean of all the Prevalences and Inverse Prevalences, $Evenness_{\mathbf{P}}$ is similarly half the Harmonic Mean of all the Biases and Inverse Biases, but $Evenness_{\mathbf{G}}$ is still defined as the Geometric Mean of. $Evenness_{\mathbf{R}}$ and $Evenness_{\mathbf{P}}$.

6 Exploratory Work and Future Work

The Bookmaker Informedness measure has been used extensively by the AI Group at Flinders over the last 5 years, in particular in the PhD Theses and other publications of Trent Lewis (2003ab) relating to AudioVisual Speech Recognition, and the publications of Sean Fitzgibbon (2007ab) relating to EEG/Brain Computer Interface. Sean was also the original author of the Matlab scripts that are available for calculating both the standard and Bookmaker statistics (see footnote on first page). The connection with DeltaP was discovered by Richard Leibbrandt in the course of his PhD research in Syntactic and Semantic Language Learning. We have also referred extensively to the equivalence of Bookmaker Informedness to ROC AUC, as used standardly in Medicine, although AUC has the form of an undemeaned probability, and B is a demeaned renormalized form.

The Informedness measure has thus proven its worth across a wide range of disciplines, at least in its dichotomous form. A particular feature of the Lewis and Fitzgibbon studies, is that they covered different numbers of classes (exercising the multi-class form of Bookmaker), as well as a number of different noise and artefact conditions. Both of these aspects of their work meant that the traditional measures and derivatives of Recall, Precision and Accuracy were useless for comparing the different runs and the different conditions, whilst Bookmaker gave clear unambiguous, easily interpretable results which were contrasted with the traditional measures in these studies.

The new $\chi^2_{\mathbf{KB}}$, $\chi^2_{\mathbf{KM}}$ and $\chi^2_{\mathbf{KBM}}$ correlation statistics have only been investigated to date in toy contrived situations, and whilst they work well there and demonstrate a clear advantage over χ^2 traditional approaches, there has been no systematic Monte Carlo analysis, and no major body of work comparing new and conventional approaches to significance. Just as Bookmaker (or DeltaP') is the normative measure of accuracy for a system against a Gold Standard, so is $\chi^2_{\mathbf{KB}}$ the proposed χ^2 significance statistic for this most common situation. For the cross-rater or cross-system comparison, where neither is normative, the BMG Correlation is the appropriate measure, and correspondingly we propose that $\chi^2_{\mathbf{KBM}}$ is the appropriate χ^2 significance statistic. To explore these thoroughly is a matter for future research.

Thus whilst our understanding of Bookmaker and Markedness as performance measure is now quite mature, particularly in view of the clear relationships with existing measures exposed in this article, a better understanding of the significance measures remains a matter for further work, including in particular, research into the multi-class application of the technique, and exploration of the asymmetry in degrees of freedom appropriate to `alpha` and `beta`, which does not seem to have been explored hitherto. Nonetheless, based on pilot experiments, these statistics seem far more reliable and well-founded than the traditional χ^2 and G^2 statistics.

7 Acknowledgements

This work has benefited from invaluable discussions with a great many members of the Flinders AILab, as well as diverse others elsewhere at Flinders, and at conferences and summer schools. I would particularly highlight the valuable contributions made by Sean Fitzgibbon, in writing the Matlab scripts and finding the determinant connection, by Trent Lewis in the first comprehensive comparative studies performed with Bookmaker and conventional measures, and Richard Leibbrandt for drawing to my attention the connection with DeltaP.

References

Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. New York: Springer-Verlag.

Bonett DG & RM Price, (2005). Inferential Methods for the Tetrachoric Correlation Coefficient, **Journal of Educational and Behavioral Statistics 30**:2, 213-225

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

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

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

Flach, PA. (2003). The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics, Proceedings of the Twentieth International Conference on Machine Learning (**ICML-2003**), Washington DC, 2003, pp. 226-233.

Fitzgibbon, Sean P., David M. W. **Powers**, Kenneth Pope, and C. Richard Clark (2007). Removal of EEG noise and artefact using blind source separation. **Journal of Clinical Neurophysiology 24**(3):232-243, June 2007

Fitzgibbon, Sean P (2007) A Machine Learning Approach to Brain-Computer Interfacing, PhD Thesis, School of Psychology, Flinders University, Adelaide.

Fraser, Alexander & Daniel Marcu (2007). Measuring Word Alignment Quality for Statistical Machine Translation, **Computational Linguistics 33**(3):293-303.

Fürnkranz Johannes & Peter A. Flach (2005). ROC 'n' Rule Learning - Towards a Better Understanding of Covering Algorithms, **Machine Learning 58**(1):39-77.

Hutchinson TP. (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.

Lafferty, J., McCallum, A. & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proceedings of the 18th International Conference on Machine Learning (I**CML-2001**), San Francisco, CA: **Morgan Kaufmann**, pp. 282-289.

Lavrac, N., Flach, P., & Zupan, B. (1999). Rule evaluation measures: A unifying view. Proceedings of the 9th International Workshop on Inductive Logic Programming (**ILP-99**). **Springer-Verlag**, pp. 174–185.

Lewis, T. W. and D. M. W. **Powers** (2003). Audio-Visual Speech Recognition using Red Exclusion and Neural Networks. **Journal of Research and Practice in Information Technology 35**#1:41-64

Lewis, Trent W. (2003), Noise-robust Audio Visual Phoneme Recognition, PhD Thesis, School of Informatics and Engineering, Flinders University, Adelaide

Lisboa, P.J.G., A. Vellido & H. Wong (2000). Bias reduction in skewed binary classfication with Bayesian neural networks**. Neural Networks 13**:407-410.

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

Powers, David M. W. (2003), Recall and Precision versus the Bookmaker, Proceedings of the International Conference on Cognitive Science (**ICSC-2003**), Sydney Australia, 2003, pp. 529-534. (See http://david.wardpowers.info/BM/index.htm.)

Reeker, L.H. (2000), Theoretic Constructs and Measurement of Performance and Intelligence in Intelligent Systems, **PerMIS 2000**. (See http://www.isd.mel.nist.gov/research_areas/research_engineering/PerMIS_Workshop/ accessed 22 December 2007.)

Sellke, T., Bayarri, M.J. and Berger, J. (2001), Calibration of P-values for testing precise null hypotheses, **American Statistician 55**, 62-71. (See http://www.stat.duke.edu/%7Eberger/papers.html#p-value accessed 22 December 2007.)

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. **Psychological Bulletin 101**, 140−146. (See http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm accessed 19 December 2007.)