Significance and Confidence in Evaluation

David M W Powers[1]

Abstract. Significance is becoming a matter of considerable concern for Machine Learning, but remains of little concern in other areas of Artificial Intelligence, particularly where Machine Learning paradigms are not rigorously applied. The Machine Learning paradigm of independent Validation and Evaluation, or more complex paradigms such as Cross-Validation or Bootstrapping, allows a quantifiable measure of confidence in the results of an evaluation.

However, over the last decade there has been increasing concern about the biases embodied in traditional evaluation methods, as well as questions about how to deal with prevalence and bias. This paper briefly introduces unbiased alternatives to Recall, Precision and Accuracy and shows how they can be used to directly estimate familiar estimates of significance and confidence.

In this paper we will develop significance and confidence estimates theoretically, as well as evaluating their performance empirically using a Monte Carlo simulation.

In relation to significance, we note the existence of measures used to estimate correlation from chi-squared sums, and relate our proposals to these estimates.

In relation to confidence, we discuss the advantages of confidence intervals over mere statistical significance. These advantages are particularly pertinent at a time when the Machine Learning community is increasingly concerned about the overuse of repositories of standard datasets – one in twenty experiments may be expected to be significantly better than chance or than any other specific result when significance is evaluated to the 0.05 level, but on the other hand standard correction techniques tend to be overly conservative and represent an explicit bias against later work.

1 INTRODUCTION

Recent theoretical development of unbiased evaluation measures [1,2] have been shown empirically to be excellent measures of human association [3,4] and to have considerable advantage over other common measures including Recall, Precision, Rand Accuracy and F-factor [5], and to have a strong relationship with Correlation [5] that makes them also preferable to Cohen Kappa [6-9].

We recapitulate both traditional and unbiased measures in section 2, then examine their relationship with a variety of standard significance measures before turning to consider an alternate approach to significance via confidence intervals defined directly from the measures themselves.

Finally we present empirical results based on binomial Monte Carlo simulation to clearly illustrate the power of both our significance and confidence measures and complement the results of [5]. We also recommend an approach to handling significance that specifically allows for multiple experiments, algorithms and parameterizations being tested against the same dataset, including in particular datasets stored in a Machine Learning repository.

[1] AILab, CSEM, Flinders University of South Australia, email:David.Powers@flinders.edu.au

2 THE DICHOTOMOUS MEASURES

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. `UPPER CASE` typewriter font is used where the values are counts, and `lower case` where the values are probabilities or proportions relative to `N` or the marginal probabilities; in addition will use Mixed Case 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 labelling. 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 no marginal sums or probabilities are 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 positive diagonal represents correct predictions, and the negative diagonal incorrect predictions. The predictions of the contingency table may be the predictions of a theory or grammar, of some computational rule or system (e.g. an Expert System or a Neural Network or a POS Tagger), 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.

Table 1. Systematic and traditional notations in a contingency table.

	+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

2.1 Recall & Precision, Sensitivity & Specificity

Recall or Sensitivity (as it is called in Psychology) is by equation (1) and while often deprecated in Information Retrieval is regarded as the primary statistic of relevance in the Medical and Social Sciences:

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

Recall is recognized to supply an incomplete picture, and in Artificial Intelligence, Precision or Confidence (as it is called in Data Mining) is its common counterpart, as defined in (2):

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

Inverse Recall or Specificity is the complementary measure most commonly used in Medical and Social Science, and is also known as the True Negative Rate (tnr). Conversely, the rarely used can also be called True Negative Accuracy (tna):

$$\begin{aligned} \text{Inverse Recall} &= tnr = tn/rn \\ &= TN/RN = D/(B+D) \end{aligned} \quad (3)$$

$$\begin{aligned} \text{Inverse Precision} &= tna = tn/pn \\ &= TN/PN = D/(C+D) \end{aligned} \quad (4)$$

Rand Accuracy 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. Cohen Kappa [6] is noteworthy as an approach to computing a debiased version of Accuracy, but its non-linearity makes it less desirable than conventional correlation [5,7,8,9].

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.

2.2 Prevalence, Bias, Cost & Skew

We 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 in general not under the control of the experimenter. By contrast, pp represents the Bias of the model [5], 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. Note that the normalized binary contingency table with unspecified margins has three degrees of freedom – setting three non−redundant ratios determines the rest.

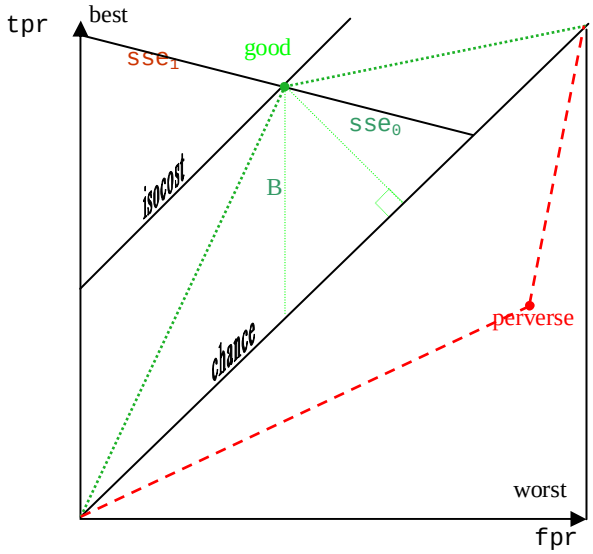

Figure 1. Illustration of ROC Analysis. The main diagonal represents chance with parallel isocost lines representing equal cost-performance. Points above the diagonal represent performance better than chance, those below worse than chance.

2.3 AUC, DeltaP, Informedness and Markedness

Powers [1] derived an unbiased accuracy measure, Bookmaker Informedness to avoid the bias of Recall, Precision and Accuracy due to population Prevalence and label bias. Optimizing Info. This is equivalent to unbiased WRAcc=2AUC-1 in ROC analysis [2].

An dual of Informedness, Markedness, is defined in [5]:

$$\begin{aligned} \text{Informedness} &= \text{Recall} + \text{Inverse Recall} - 1 \\ &= tpr - fpr = 1 - fnr - fpr \end{aligned} \quad (5)$$

$$\begin{aligned} \text{Markedness} &= \text{Precision} + \text{Inverse Precision} - 1 \\ &= tpa - fna = 1 - fpa - fna \end{aligned} \quad (6)$$

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 [3,4] and DeltaP' [3] corresponds to Informedness. These correspond to the regression coefficient for the dual directions of association [3,5], and their geometric mean is by definition the correlation [3,5].

2.4 Effect of Bias & Prev on Recall & Precision

We present some simple relationships between these biased and unbiased measures to make explicit the role of Prevalence and Bias:

$$\begin{aligned} \text{Recall} &= \text{Informedness}\ (1-\text{Prevalence}) + \text{Bias} \\ \text{Informedness} &= (\text{Recall}-\text{Bias})\ /\ (1-\text{Prevalence}) \end{aligned} \quad (7)$$

$$\begin{aligned} \text{Precision} &= \text{Markedness}\ (1-\text{Bias}) + \text{Prevalence} \\ \text{Markedness} &= (\text{Precision}-\text{Prev})\ /\ (1-\text{Bias}) \end{aligned} \quad (8)$$

Bookmaker and Markedness are unbiased estimators of above chance performance (relative to respectively the predicting conditions or the predicted markers). Recall = Precision and Informedness = Markedness if and only if Bias = Prevalence [5].

We can gain further insight into the nature of these regression and correlation coefficients expressing them as distinct normalization of

the determinant of the contingency matrix, dp, which is a common numerator common across all three coefficients. Viz. Informedness (B) and Markedness (M) and Correlation (C) may be re-expressed in terms of Precision (Prec) or Recall, along with the Geometric means of Bias or Prevalence and their respective inverses (IBias=1-Bias, IPrev=1-Prev), defining respective Evenness terms that are maximum for even Bias or Prevalence:

$$\begin{aligned} M &= dp \,/\, [\text{Bias} \cdot (1-\text{Bias})] \\ &= dp \,/\, \text{BiasG}^2 = dp \,/\, \text{Evenness}_{\mathbf{P}} \\ &= [\text{Precision} - \text{Prevalence}] \,/\, \text{IBias} \qquad (9) \\ B &= dp \,/\, [\text{Prevalence} \cdot (1-\text{Prevalence})] \\ &= dp \,/\, \text{PrevG}^2 = dp \,/\, \text{Evenness}_{\mathbf{R}} \\ &= [\text{Recall} - \text{Bias}] \,/\, \text{IPrev} \qquad (10) \\ C &= dp \,/\, [\text{PrevG} \cdot \text{BiasG}] = dp \,/\, \text{Evenness}_{\mathbf{G}} \\ &= \sqrt{[(\text{Recall}-\text{Bias})\cdot(\text{Prec}-\text{Prev})]/(\text{IPrev}\cdot\text{IBias})} \qquad (11) \end{aligned}$$

2.5 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?

This raises the question of how our measures of association and accuracy, Informedness, Markedness and Correlation, relate to standard measures of significance.

This paper has been written in the context of a Prevailing methodology in Computational Linguistics and Information Retrieval that concentrates on target positive cases and ignores the negative case for the purpose of both measures of association and significance. A classic example is saying “water” can only be a noun because the system is inadequate to the task of Part of Speech identification, so this boosts Recall and hence F-factor, or otherwise setting the Bias to nouns close to 1, and the Inverse Bias to verbs close to 0. Of course, Bookmaker will then approach 0 and Markedness will be unstable (undefined, and very sensitive to any words that do actually get labelled verbs). We would clearly expect that significance would also be 0 (or approaching zero given a vanishingly small number of verb labels). We would like to be able to calculate significance based on the positive case alone, if either the full negative information is unavailable, or if it is not labelled.

Generally when dealing with contingency tables it is assumed that unused labels or unrepresented classes are dropped from the table, with corresponding reduction of degrees of freedom. For simplicity we have assumed that the margins are all non-zero (contra to the Bias=1 example), but the freedoms are there whether they are used or not, so we will not reduce them or reduce the table.

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 (r=1 given the marginal counts are known), depend on distributional assumptions, and focus only on Positives.

χ^2 captures the Total Squared Deviation relative to expectation (ETP=E(TP), etc) and is here calculated only in relation to positive predictions, as often only the overt prediction is considered, and the implicit prediction of negative case is ignored:

$$\begin{aligned} \chi^2_{+\mathbf{P}} &= (TP-ETP)^2/ETP + (FP-EFP)^2/EFP \\ &= DTP^2/ETP + DFP^2/EFP \\ &= 2DP^2/EHP,\ EHP = 2ETP\cdot EFP/[ETP+EFP] \\ &= 2N\cdot dp^2/ehp, ehp = 2etp\cdot efp/[etp+efp] \\ &= 2N\cdot dp^2/[rh\cdot pp] = N\cdot dp^2/\text{PrevG}^2/\text{Bias} \\ &= N\cdot B^2\cdot\text{Evenness}_{\mathbf{R}}/\text{Bias} = N\cdot r^2_{\mathbf{P}}\cdot\text{PrevG}^2/\text{Bias} \qquad (12) \end{aligned}$$

G^2 captures Total Information Gain, being N times the Average Information Gain in nats, otherwise known as Mutual Information. We deal with G^2 for positive predictions in the case of small effect, that is dp close to zero, where G^2 is twice as sensitive as χ^2.

$$\begin{aligned} G^2_{+\mathbf{P}}/2 &= TP\cdot\ln(TP/ETP) + FP\cdot\ln(FP/EFP) \\ &= TP\cdot\ln(1+DTP/ETP)+FP\cdot\ln(1+DFP/EFP) \\ &\approx TP\cdot(DTP/ETP) + FP\cdot(DFP/EFP) \\ &= 2N\cdot dp^2/ehp, ehp = 2etp\cdot efp/[etp+efp] \\ &= 2N\cdot dp^2/[rh\cdot pp] = N\cdot dp^2/\text{PrevG}^2/\text{Bias} \\ &= N\cdot B^2\cdot\text{Evenness}_{\mathbf{R}}/\text{Bias} = N\cdot r^2_{\mathbf{P}}\cdot\text{PrevG}^2/\text{Bias} \qquad (13) \end{aligned}$$

Our result (12-13) 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 ($\text{Evenness}_{\mathbf{R}} = \text{PrevG}^2 = \text{Prev}\cdot(1-\text{Prev})$) and the number of Predicted Negatives (viz. with Inverse Bias)! This is also 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 $KN\cdot B^2\cdot\text{PrevG}^2$ which gives us an estimate of the significance without focussing on either case in particular.

$$\begin{aligned} \chi^2_{\mathbf{KB}} &= 2N\cdot dtp^2/\text{PrevG}^2 = 2N\cdot r^2_{\mathbf{P}}\cdot\text{PrevG}^2 \\ &= 2N\cdot r^2_{\mathbf{P}}\cdot\text{Evenness}_{\mathbf{R}} \\ &= 2N\cdot B^2\cdot\text{Evenness}_{\mathbf{R}} \qquad (14) \end{aligned}$$

Analogous formulae can be derived for significance of Markedness for positive real classes, noting that $\text{Evenness}_{\mathbf{P}} = \text{BiasG}^2$.

$$\begin{aligned} \chi^2_{\mathbf{KM}} &= 2N\cdot dtp^2/\text{BiasG}^2 = 2N\cdot r^2_{\mathbf{R}}\cdot\text{BiasG}^2 \\ &= 2N\cdot r^2_{\mathbf{R}}\cdot\text{BiasG}^2 \\ &= 2N\cdot M^2\cdot\text{Evenness}_{\mathbf{P}} \qquad (15) \end{aligned}$$

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

$$\begin{aligned} \chi^2_{\mathbf{KC}} &= 2N\cdot dtp^2/\text{PrevG}\cdot\text{BiasG} \\ &= 2N\cdot r_{\mathbf{P}}\cdot r_{\mathbf{R}}\cdot\text{PrevG}\cdot\text{BiasG} \\ &= 2N\cdot r^2_{\mathbf{G}}\cdot\text{Evenness}_{\mathbf{G}} = 2NC^2\cdot\text{Evenness}_{\mathbf{G}} \\ &= 2N\cdot B\cdot M\cdot\text{Evenness}_{\mathbf{G}} \qquad (16) \end{aligned}$$

This is simply the total Sum of Squares Deviance (SSD) accounted for by the correlation coefficient C (11) 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·BiasG). The less even the Bias and Prevalence, the more data will be required to achieve significance, the maximum evenness value of 0.25 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.

When χ^2_{+P} or G^2_{+P} is calculated for a specific label in a dichotomous contingency table, it has one degree of freedom for the purposes of assessment of significance. The full table also has one degree of freedom, and summing for goodness of fit over only the positive prediction label will clearly lead to a lower χ^2 estimate than summing across the full table, and while summing for only the negative label will often give a similar result it will in general be different. Thus the weighted arithmetic mean calculated by χ^2_{KB} is an expected value independent of the arbitrary choice of which predictive variate is investigated. This is used to see whether a hypothesized main effect (the alternate hypothesis, H_A) is borne out by a significant difference from the usual distribution (the null hypothesis, H_0). Summing over the entire table (rather than averaging of labels), is used for χ^2 or G^2 independence testing independent of any specific alternate hypothesis, and can be expected to achieve a χ^2 estimate approximately twice that achieved by the above estimates, effectively cancelling out the Evenness term, but is thus far less conservative (viz. it is more likely to satisfy $p<\alpha$):

$$\chi^2_C = N \cdot r^2_G = N \cdot \rho^2 = N \cdot \varphi^2 = N \cdot B \cdot M = N \cdot C^2 \qquad (17)$$

Note that this equates C corresponding to Pearson's Rho, ρ, with the Phi Correlation Coefficient, φ, which is defined in terms of the Inertia $\varphi^2=\chi^2/N$. We now have confirmed that not only does a factor of N connects the full contingency G^2 to Mutual Information (MI), but it also normalizes the full approximate χ^2 contingency to Matthews/Pearson Correlation (=√BM=C=Phi), for the dichotomous case. This tells us moreover, that MI and Correlation are measuring essentially the same thing, but MI and Phi do not tell us anything about the direction of the correlation, whilst the sign of Matthews or Pearson or √BM Correlation does (since it is the Biases and Prevalences that are multiplied and squarerooted).

2.6 Confidence Intervals and Deviations

An alternative to significance estimation is confidence estimation in the statistical rather than the data mining sense. We noted earlier that selecting the highest isocost line or maximizing AUC or Bookmaker Informedness, B, is equivalent to minimizing fpr+fnr=(1-B) or maximizing tpr+tnr=(1+B), which maximizes the sum of normalized squared deviations of B from chance, $sse_B=B^2$ (as is seen geometrically from Fig. 1). Note that this contrasts with minimizing the sum of squares distance from the optimum which minimizes the relative sum of squared normalized error of the aggregated contingency, $sse_B=fpr^2+fnr^2$. However, an alternate definition calculating the sum of squared deviation from *optimum* is as a normalization the square of the minimum distance to the isocost of contingency, $sse_B=(1-B)^2$.

This approach contrasts with the approach of considering the error versus a specific null hypothesis representing the expectation from margins. Normalization is to the range [0,1] like |B| and normalizes (due to similar triangles) all orientations of the distance between isocosts (Fig. 1). With these estimates the relative error is constant and the relative size of confidence intervals around the null and full hypotheses only depend on N as |B| and |1-B| are already standardized measures of deviation from null or full correlation respectively (σ/μ=1). Note however that if the *empirical* value is 0 or 1, these measures admit no error versus no information or full information resp. If the *theoretical* value is B=0, then a full ±1 error is possible, particularly in the discrete low N case where it can be equilikely and will be more likely than expected values that are fractional and thus likely to become zeros. If the theoretical value is B=1, then no variation is expected unless due to measurement error. Thus |1-B| reflects the maximum (low N) deviation in the absence of measurement error.

The standard Confidence Interval is defined in terms of the Standard Error, SE =√[SSE/(N•(N-1))] =√[sse/(N-1)]. It is usual to use a multiplier X of around X=2 as, given the central limit theorem applies and the distribution can be regarded as normal, a multiplier of 1.96 corresponds to a confidence of 95% that the true mean lies in the specified interval around the estimated mean, viz. the probability that the derived confidence interval will bound the true mean is 0.95 and the test thus corresponds approximately to a significance test with alpha=0.05 as the probability of rejecting a correct null hypothesis, or a power test with beta=0.05 as the probability of rejecting a true full or partial correlation hypothesis. A number of other distributions also approximate 95% confidence at 2SE.

We specifically reject the more traditional approach which assumes that both Prevalence and Bias are fixed, defining margins which in turn define a specific chance case rather than an isocost line representing all chance cases – we cannot assume that any solution on an isocost line has greater error than any other since all are by definition equivalent. The above approach is thus argued to be appropriate for Bookmaker and ROC statistics which are based on the isocost concept, and reflects the fact that most practical systems do not in fact preset the Bias or match it to Prevalence, and indeed Prevalences in early trials may be different from those in the field.

he specific estimate of sse that we present for alpha, the probability of the current estimate for B occurring if the true Informedness is B=0, is $\sqrt{sse_{B0}}=|1-B|=1$, which is appropriate for testing the null hypothesis, and thus for defining unconventional error bars on B=0. Conversely, $\sqrt{sse_{B2}}=|B|=0$, is appropriate for testing deviation from the full hypothesis in the absence of measurement error, whilst $\sqrt{sse_{B2}}=|B|=1$ conservatively allows for full range measurement error, and thus defines unconventional error bars on B=M=C=1.

In view of the fact that there is confusion between the use of beta in relation to a specific full dependency hypothesis, B=1 as we have just considered, and the conventional definition of an arbitrary and unspecific alternate contingent hypothesis, B≠0, we designate the probability of incorrectly excluding the full hypothesis by gamma, and propose possible kinds of heuristic for the √sse for beta (which will typically be assumed to relate to the empirical estimate as the true value). We can use a mean of |B| and 1-|B| (the unweighted arithmetic mean is 0.5, the geometric mean is less conservative and the harmonic mean even less conservative, the maximum being extremely conservative, and the minimum too low an underestimate in general. Note that we allow an asymmetric interval that has one value on the null side, another on the full side.

The √sse means may be weighted or unweighted and in particular a self-weighted arithmetic mean gives our recommended definition, $\sqrt{sse_{B1}}=1-2|B|+2B^2$, with a minimum of 0.5 at B=±0.5 and a maximum of 1 at both B=0 and B=±1.

Using Monte Carlo simulations (Fig. 2), we have observed that setting $sse_{B1}=\sqrt{sse_{B2}}=1-|B|$ as per the usual convention is appropriately conservative on the upside but a little broad on the

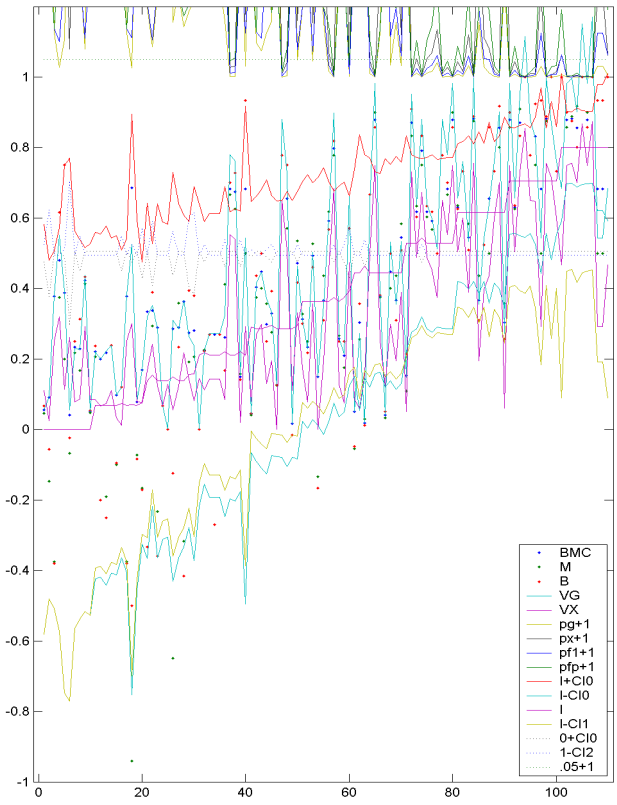

Figure 2. Accuracy of significance and confidence measures. 110 Monte Carlo simulations with 11 stepped expected Informedness levels (probability of correct decision versus random binomial decision and random margins) with calculated Informedness, Markedness and Correlation versus Phi calculated from G^2 and χ^2 and confidence intervals based on `beta` (power: $\sqrt{sse_{B1}}=1-2|B|+2B^2$), `alpha` and `gamma` (null & full significance: $\sqrt{sse_{B0}}=\sqrt{sse_{B1}}=1$). p-values based on G^2 and χ^2 and point and cumulative Fisher Test are shown +1.

downside, whilst the weighted arithmetic mean, $\sqrt{sse_{B1}}=1-2|B|+2B^2$, is sufficiently conservative on the downside, but too conservative for high B.

Note that these two-tailed ranges are valid for Bookmaker Informedness and Markedness that can go positive or negative, but a one tailed test would be appropriate for unsigned statistics or where a particular direction of prediction is assumed as we have for our contingency tables. In these cases a smaller multiplier of X=1.65 would suffice, however the more conservative convention is to use the overlapping of the confidence bars around the various hypotheses (although usually the null is not explicitly represented).

3 DISCUSSION AND CONCLUSIONS

In Machine Learning it is usual to try many different algorithms on a problem, whether from a repository or for a challenging application. This leads to a strong probability that a spurious improvement will be found if o(1/α) approaches are tested. The Bonferonni approach is overly conservative and even the Benjamini-Hochberg approach of reducing alpha progressively disadvantages later researchers if natural order rather than p order is used [10], and is impossible to apply properly between two systems given only p-values versus H_0. However, the publication of confidence intervals allows the calculation of p-values between systems and hence the proper application of Benjamini-Hochberg if required.

For any two hypotheses (including the null hypothesis, or one from a different contingency table or other experiment deriving from a different theory or system) the traditional approach of checking that 1.95SE (or 2SE) error bars don't overlap is too conservative: it is enough for the value to be outside the range for a two-sided test as between competing systems, whilst checking overlap of 1SE error bars is usually insufficiently conservative given that the upper represents `beta<alpha`. Where it is predicted that a given system will be better than the other, a 1.65SE error bar including the mean for the other hypothesis is enough to indicate significance (and power=`1-beta`) corresponding to `alpha` (resp. `beta`) as desired.

The traditional calculation of error bars based on Sum of Squared Error is closely related to the calculation of Chi-Squared significance based on Total Squared Deviation, and like it are not reliable when the assumptions of normality are not approximated, and in particular when the conditions for the central limit theorem are not satisfied (e.g. N<12 or cell-count<5). They are not appropriate for application to probabilistic measures of association or error. This is captured by the meeting of the X=2 error bars for the full (sse_{B2}) and null (sse_{B0}) hypotheses at N=16 (expected count of only 4 per cell), as shown in Fig. 2.

The proposed direct calculation of significance from the 'Bookmaker' measures, and the more robust approach using confidence intervals as error bars, gives a direct indication of significance without the need for expensive cross-validation. Of course, a one-fits-all generic approach does not take into account the specific problem, the priors, or the actual theoretical and empirical distributions, and where marginal significance is indicated a more accurately targeted methodology would be indicated.

REFERENCES

[1] 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. david.wardpowers.info/BM/index.htm. accessed 22 December 2007

[2] 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.

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

[4] Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology*, 48A, 257-279.

[5] Anonymous (submitted to ECAI08).

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

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

[8] 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.

[9] Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin* **101**, 140−146. ourworld.compuserve.com/homepages/jsuebersax/agree.htm

[10] Benjamini, Yoav and Hochbert, Yosef (1995), Controlling the False Discovery Rate. *J. Roy. Stat. Soc. Ser. B*, **57:**289-300.