ECAI 2008
M. Ghallab et al. (Eds.)
IOS Press, 2008

doi:10.3233/978-1-58603-891-5-843

Evaluation Evaluation

David M W Powers[1]

Abstract. Over the last decade there has been increasing concern about the biases embodied in traditional evaluation methods for Natural Language Processing/Learning, particularly methods borrowed from Information Retrieval. Without knowledge of the Bias and Prevalence of the contingency being tested, or equivalently the expectation due to chance, the simple conditional probabilities Recall, Precision and Accuracy are not meaningful as evaluation measures, either individually or in combinations such as F-factor.

The existence of bias in NLP measures leads to the 'improvement' of systems by increasing their bias, such as the practice of improving tagging and parsing scores by using most common value (e.g. water is always a Noun) rather than the attempting to discover the correct one. In this paper, we will analyze both biased and unbiased measures theoretically, characterizing the precise relationship between all these measures.

1 INTRODUCTION

A common but poorly motivated way of evaluating results of Language and 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 the Recall-like True Positive Rate and False Positive Rate. In the Behavioural Sciences, the related concepts of Specificity and Sensitivity, are commonly used. Alternate techniques, such as Rand Accuracy, have some advantages but are nonetheless still biased measures unless explicitly debiased.

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, `A, B, C, D`, summing to `N`, or we may use acronyms for the generic terms for True and False, Real and Predicted Positives and Negatives, or else relative versions of these, e.g: `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. These systems are both illustrated in Table 1.

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.

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, Recall has been shown to have a major weight in predicting success in several context including these areas, and in a Medical context Recall is primary but it is referred to as True Positive Rate (`tpr`). Recall is defined, with its various common appellations, by equation (1):

$$\text{Recall} = \text{Sensitivity} = \mathtt{tpr} = \mathtt{tp/rp} \tag{1}$$

Conversely, Precision or Confidence (as it is called in Data Mining) denotes the proportion of Predicted Positive cases that are correctly Real Positives. It can also be called True Positive Accuracy (`tpa`), as a measure of accuracy of Predicted Positives in contrast with rate of discovery of Real Positives (`tpr`). Precision is defined in (2):

$$\text{Precision} = \text{Confidence} = \mathtt{tpa} = \mathtt{tp/pp} \tag{2}$$

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 $\mathtt{F=G^2/A}$ (the F-factor or F-measure).

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

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

Usually, 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. Rand Accuracy explicitly takes into account the classification of negatives, and is expressible 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 correctly classified negatives (TN). Each of these measures also has a complementary form defining an error rate, of which some have specific names and importance: Fallout or False Positive Rate (fpr) is the proportion of Real Negatives that occur as Predicted Positive (ring-ins); Miss Rate or False Negative Rate (fnr) is the proportion of Real Positives that are Predicted Negatives (false-drops).

2.2 Prevalence, Bias, Cost & Skew

We now turn our attention to various forms of bias or skew that detract from the utility of all of the above surface measures [1,2]. We will first note that rp represents the Prevalence of positive cases, RP/N – it is not usually under the control of the experimenter. By contrast, pp represents the (label) Bias of the model [1], 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. A common rule of thumb, and 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.

2.3 ROC and PN Analyses

Flach [4] has highlighted the utility of ROC analysis to the Machine Learning community, and characterized the skew sensitivity of many measures in that context, utilizing the ROC format to give geometric insights into the nature of the measures and their sensitivity to skew. ROC analysis plots the rate tpr against the rate fpr. 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). 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 area under the simple trapezoid is: AUC = $1 - (fpr+fnr)/2$.

2.4 DeltaP, Informedness and Markedness

Powers [2] 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 present only the dichotomous formulation of Powers Informedness, as well as the complementary concept of Markedness. In fact, Bookmaker Informedness-based formulae may be averaged over all labels according to the label bias, and Markedness-based formulae over all classes by prevalence.

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).

$$\begin{aligned}\text{Informedness} &= \text{Recall} + \text{Inverse Recall} - 1\\ &= tpr-fpr = 1-fnr-fpr = 2\text{AUC}-1 \qquad (3)\\ &= (\text{Recall-Bias}) / (1-\text{Prevalence})\end{aligned}$$

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).

$$\begin{aligned}\text{Markedness} &= \text{Precision} + \text{Inverse Precision} - 1\\ &= tpa-fna = 1-fpa-fna \qquad (4)\\ &= (\text{Precision}-\text{Prevalence}) / (1\text{-Bias})\end{aligned}$$

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, however there are two possible directions of implication. Detection of the predictor may reliably predict the outcome, with or without the occurrence of a specific outcome condition reliably evincing the predictor.

In the Psychology literature, Markedness is known as DeltaP and is empirically a good (normative) 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. Conversely a complementary, backward, additional measure of strength of association, DeltaP' aka Informedness has been proposed [5].

Note that we can also estimate significance and confidence [3]:

$$\chi^2 = N \cdot \text{Informedness} \cdot \text{Markedness} \qquad (5)$$

$$\text{CI} = 1\text{-}|\text{Informedness}|/\sqrt{[N-1]}; \quad \text{CM} = 1\text{-}|\text{Markedness}|/\sqrt{[N-1]}$$

REFERENCES

[1] 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* (**ICML-2001**), CA: Morgan Kaufmann, pp. 282-289.

[2] 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. http://david.wardpowers.info/BM/index.htm accessed 22 December 2007

[3] Powers, David M. W. (2007) *Evaluation*, Flinders InfoEng Tech Rept SIE07001 http://www.infoeng.flinders.edu.au/research/techreps/SIE07001.pdf

[4] 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.

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