Bayesian Information Reward

Lucas R. Hope and Kevin B. Korb

School of Computer Science
and Software Engineering
Monash University
Clayton, VIC 3168, Australia

email: {lhope,korb}@csse.monash.edu.au

Abstract. We generalize an information-based reward function, introduced by Good (1952), for use with machine learners of classification functions. We discuss the advantages of our function over predictive accuracy and the metric of Kononenko and Bratko (1991). We examine the use of information reward to evaluate popular machine learning algorithms (e.g., C5.0, Naive Bayes, CaMML) using UCI archive datasets, finding that the assessment implied by predictive accuracy is often reversed when using information reward.

Keywords: Evaluation metrics, information reward, Bayesian evaluation, evaluation of machine learners, predictive accuracy.

1 Introduction

Predictive accuracy as an evaluation metric for machine learners has a number of notable weaknesses: it fails to differentiate between the value of correct classification and incorrect classification, a problem addressed by cost-sensitive classification (cf. Turney (1995)); it also fails to take into account the uncertainty of predictions, treating a fully certain binary prediction as the same as one with probability 0.51. Given these substantial drawbacks, it is somewhat surprising how many researchers use predictive accuracy (or its converse, error rate) as their one and only metric. Perhaps it is the extreme simplicity of its application which maintains its widespread use: computing accuracy requires only a simple yes/no answer to the question, does this instance belong to this class?

We believe a cost-sensitive assessment, namely which machine learner maximizes expected reward, is clearly the best one for evaluating learning algorithms. Unfortunately, finding an appropriate cost function may be difficult or impossible. No expert may be available to provide a suitable cost function; or the algorithms being assessed may be applied across an open-ended variety of domains. An evaluation method independent of cost function which has become popular recently uses ROC curves, as in Provost and Fawcett (1997). ROC curves, however, again ignore the probabilistic aspect of prediction, as does predictive accuracy simpliciter. Here we examine a metric which specifically attends to the estimated probability of a classification, but is also independent of cost, and so easier to apply than cost-sensitive metrics; in particular, we examine the *Information Reward* (IR) measure, its properties, requirements, and generalization. We also present some empirical results which show the surprising dominance of Naive Bayes when compared with other well known machine learning algorithms such as C5.0 (Quinlan, 1998).

We take the right model for computing reward in uncertain predictions to be that of gambling: a bettor is rewarded not just for identifying the ultimate winners and losers, but more importantly for identifying the appropriate *odds* — namely, those odds which give neither side to a bet an advantage over the other, that is *fair odds*. An agent, artificial or natural, which can consistently beat its opposition in making bets about outcomes in a domain, or across a range of domains, is clearly a superior predictor to is opposition. Predictive accuracy can never hope to assess this ability, since it is constrained to ignore probabilities and therefore odds. IR measures exactly this ability. IR reports an information-theoretic function of class predictions in comparison with their prior probabilities, rewarding domain understanding as reflected in the correctness of modal predictions, but also rewarding the *calibration* of predictions, penalizing over- and under-confidence while rewarding matches between probabilitistic predictions and the frequency with which those predictions are realized. IR is equivalent to the gambling reward over a series of fair bets (see Korb et al. (2001)).

2 Definitions

2.1 Original Information Reward

The original definition of IR was introduced by Good (1952) as *fair betting fees*, that is, the cost of buying a bet which makes the expected value of the purchase zero. Good's IR positively rewarded binary classifications which were informative relative to an uninformed, uniform prior. The score of a single classification is generated in terms of the generating machine learner's estimated probability p. IR is split into two cases: that where the classification is correct, indicated by a superscripted '+', and where the classification is incorrect, indicated by a superscripted '−'.

Definition 1. *The IR of a binary classification with probability p is*

$$I^+ = 1 + \log_2 p \qquad \textit{(for correct classification)} \tag{1a}$$

$$I^- = 1 + \log_2(1-p) \qquad \textit{(for misclassification)} \tag{1b}$$

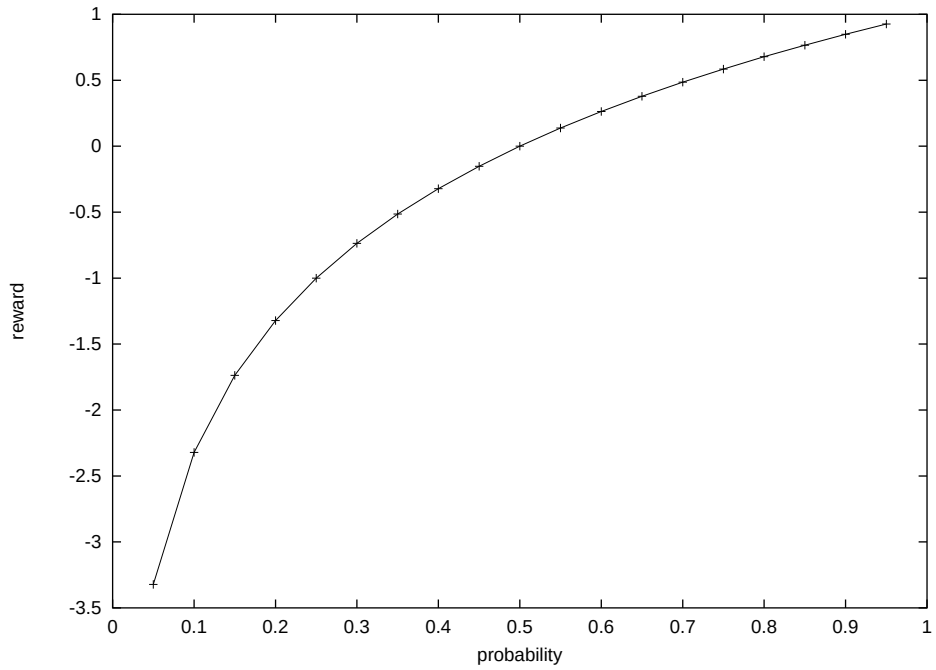

Fig. 1. Good's Information Reward

IR has the range $(-\infty, 1)$ (see Figure 1). For successful classification, it increases monotonically with p, and thus is maximized as p approaches 1. For misclassification, IR decreases monotonically from the value 0 when $p = 0.5$.

While the constant 1 in (1a) and (1b) is unnecessary for simply ranking machine learners, it makes sense in terms of fair fees. When the learner reports a probability of 0.5, it is not communicating any *information* (given a uniform prior), and thus receives a zero reward. Ignoring the constant 1, IR has a clear information-theoretic basis: it reports (the negation of) the number of bits required in a message reporting an outcome of the indicated probability. Thus, a certain message requires no bits at all, whereas a certainly false message can never be communicated successfully, requiring an infinitely long message.

Kononenko and Bratko (1991), when introducing a related metric (more about which below), have expressed the intuition that when such a reward is applied to a correct prediction with probability 1 and an incorrect prediction also with probability 1, the correct and incorrect predictions ought precisely to counterbalance, resulting in a total reward of 0. This intuition, however, is at variance with the supposed information-theoretic basis for their reward: on any account in accord with Shannon's information measure, a reward for a certain prediction coming true can only be finite, while a penalty for such a *certain* prediction coming false must always be infinite. Putting these into balance guarantees there will be no proper information-theoretic interpretation of a reward function.

2.2 Generalized Bayesian Information Reward

Our search for a definition of IR that generalizes Good's began with an attempt to apply Good's measure to multiclass datasets (Korb et al., 2001). Good's measure 'hardcodes' a zero score to one where the confidence p is 0.5. So, for example, if a machine learner correctly identifies class A in a 3-class problem with confidence of 0.4, the machine learner will receive a *negative score*. Following Good's treatment of binary variables, it seems that correct classification with probability greater than $\frac{1}{3}$ should be given a positive score. So one possible generalization simply replaces the relativization to a uniform prior over two cases with a uniform prior over n cases. But what if prior information about class A in the 3-class problem indicates that without any further information, its probability is 0.8? Should the machine learner be rewarded for its correct prediction at 0.4, or should it be penalized? We believe it should be penalized for underconfidence, and hence introduce a Bayesian prior p' to the IR calculation.

The idea behind fair fees, that you should only be paid for an *informative* prediction, is simply not adequately addressed by Definition 1. Suppose an expert has diagnosed patients with a disease that is carried by 10% of some population. This particular expert is lazy and simply reports that each patient does not have the disease, with 0.9 confidence. The expected reward per patient for this strategy, under Definition 1 is

$$0.9(1 + \log_2 0.9) + 0.1(1 + \log_2 0.1) = \ 0.531$$

So the expert is rewarded substantially for the uninformed strategy! The expected reward per patient we should like to see is 0, which our generalization below provides. Definition 1 breaks down in its application to multinomial classification: any *successful* prediction with confidence less than 0.5 is penalized, even when the confidence is greater than the prior. Good's fair fees are actually fair only when both the prior is uniform and the task binary.

Hence, we now define the IR of a single classification in terms of the estimated probability p and the class's prior probability p'. Henceforth, Definition 1 will be referred to as IR_G; Definition 2, immediately below, replacing it pro tem. IR is again split into two cases: that where the classification is correct, and that where the classification is incorrect.

Definition 2. *The Bayesian IR of a single classification with estimated probability p and prior probability p', is*

$$I^+ = 1 - \frac{\log p}{\log p'} \qquad \textit{(for correct classification)} \tag{2a}$$

$$I^- = 1 - \frac{\log (1-p)}{\log (1-p')} \qquad \textit{(for misclassification)} \tag{2b}$$

This IR also has the range $(-\infty, 1)$. For successful classification, it increases monotonically with p, and thus is maximized as p approaches 1, and approaches negative infinity as p approaches 0. IR is 0 precisely when $p = p'$. So, increased certainty ($p > p'$) is rewarded, while decreased certainty ($p < p'$) is punished. For misclassification, IR decreases monotonically as p increases, taking the value 0 when $p = p'$. Thus, misplaced increased certainty ($p > p'$) is punished, while a decreased certainty ($p < p'$) when misclassifying is rewarded.

The prior probability p' can be obtained any number of ways, including being set arbitrarily (or subjectively). We use frequency from the training set given to the machine learner to calculate the prior, for two reasons. First, we are obtaining the prior from a source that the machine learner has full access to, and thus there is no 'unfair' bias in the measure. Second, this means that the simplest algorithm, one which translates observed prior frequencies into posterior probabilities of future occurrence, will receive a score of zero, acting as a baseline for assessing more intelligent algorithms.

Our Definition 2 subsumes Definition 1: given a uniform prior and binary classification, IR_G and IR are identical.

There are, however, some difficulties with Definition 2 since it assesses the machine learner's probability distribution over classes only on the basis of the modal class, that is, that class which has the greatest probability according to the learner. Since the posterior distribution is only being assessed against a single class, its potential to inform us about the quality of its learning by examining *other* classes is being wasted. This also has the effect of producing the "kink" reported for information reward in Figure 2: since the true class in that figure is no longer the modal class below its prior probability of $1/3$, IR is computed relative to a different class; and the penalty for that modal class changes at a different rate than the reward for the true class when it is modal. Even worse than these points is the

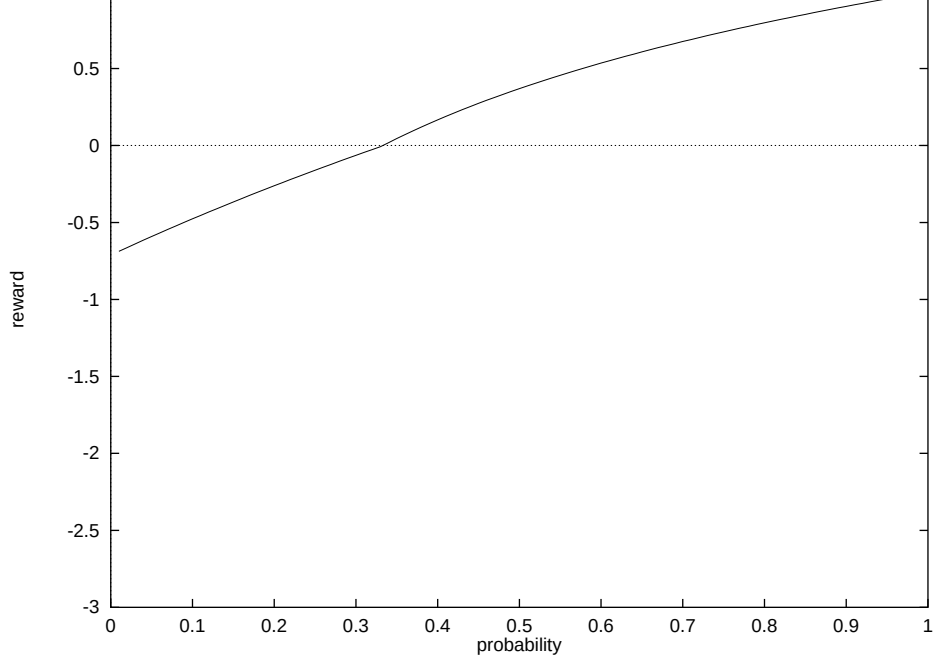

Fig. 2. Bayesian Information Reward. This is computed assuming three possible classes, with the x axis indicating the posterior probability given to the true class, and assuming a prior of 1/3.

fact that should the learner incorrectly assign probability zero to some class, and thus be potentially deserving of an infinitely negative reward (as we argued above), the learner will escape its due punishment, since that class will never be modal. These difficulties are all easily rectified by summing the reward function over all the classes:

Definition 3. *The Generalized Bayesian IR for a classification into classes* $\{C_1, \ldots, C_k\}$ *with estimated probabilites* p_i *and prior probabilities* p'_i*, where* $i \in \{1, \ldots, k\}$*, is*

$$IR = \frac{\sum_i I_i}{k} \tag{3}$$

where $I_i = I_i^+$ *below for correct classes and* I_i^- *for incorrect classes:*

$$I_i^+ = 1 - \frac{\log p_i}{\log p'_i} \qquad \textit{(for correct classification)} \tag{3a}$$

$$I_i^- = 1 - \frac{\log (1 - p_i)}{\log (1 - p'_i)} \qquad \textit{(for misclassification)} \tag{3b}$$

Generalized Bayesian information reward reflects the gambling metaphor more adequately than does Definition 2. Book makers are required to take bets for and against whatever events are in their books, with their earnings depending on the spread between bets for and against particular outcomes. They are, in effect, being rated on the quality of the odds they generate for all outcomes simultaneously. Generalized IR does the same for machine learning algorithms: the odds (probabilities) they offer on all the possible classes are simultaneously assessed, extracting the maximum information from each probabilistic classification.

We illustrate generalized Bayesian information reward in Figure 3, which also displays Kononenko and Bratko's measure (discussed immediately below). Some differences will be observed with Figure 2, where generalized IR modifies the assessment of Definition 2 by incorporating non-modal class probabilities. This is most noticeable when the low probability accorded the true class (in the range $(0, 1/3)$) keeps it out of the assessment in Figure 2. This final version of Bayesian information reward again subsumes the original one of Good: since for classification into two classes $\{C_0, C_1\}$, where C_0 is for

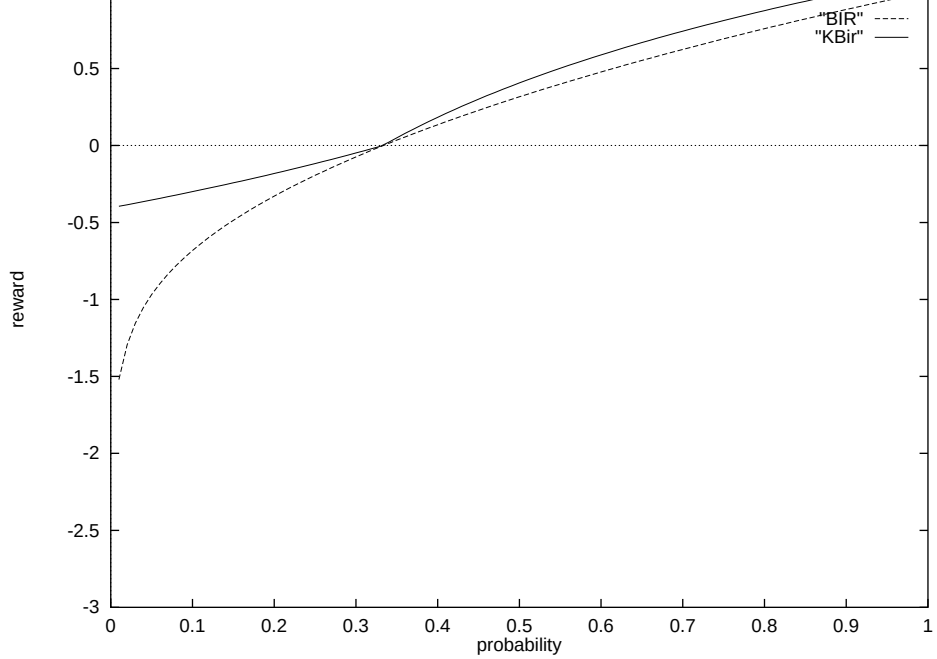

Fig. 3. Generalized Bayesian information reward (BIR) together with Kononenko & Bratko's information reward (KBir). These are computed based upon three classes and a uniform prior distribution, with the probability for the true class given on the x axis.

example correct, $IR_G = 1 + \log_2 p_0$, whereas (taking logs to base 2)

$$IR = \frac{2 + \log_2 p_0 + \log_2(1 - p_1)}{2} = IR_G$$

on the assumption of a uniform prior.

3 Kononenko and Bratko's Measure

A related measure introduced by Kononenko and Bratko (1991) also relativizes reward to prior probabilities. Furthermore, it too is nominally based upon information theory, although as we pointed out above, that interpretation is undermined by the introduction of an inappropriate symmetry in the reward for correct and incorrect classifications.

Another dubious aspect of Kononenko and Bratko's analysis is their claim that costs can be computed from prior probabilities. Thus, they assert that when $P(C_1) > P(C_2)$, "if we denote the credit for correct classification into class C with $V_c(C)$, and the penalty for misclassification with $V_m(C)$, then the following should hold: $V_c(C_1) < V_c(C_2)$ and $V_m(C_1) > V_m(C_2)$" (p. 70). Cost and probability functions are, in fact, orthogonal: any combination of high and low cost with high and low probability is possible. For example, the cost of misclassifying a disease might be very high (e.g., leading to death), even when the frequency of the disease is also very high, as might be the case for patients referred to a specialty clinic. We nevertheless agree with Kononenko and Bratko that the kind of cost-neutral reward we are attempting to identify here needs to be relativized to prior probability: otherwise there is no way to avoid rewarding a learner which slavishly mimicks frequencies in a training set and no way to penalize algorithms which simply fail to learn from such frequencies.

Kononenko and Bratko specifically introduced the following reward function, which is assessed for each instance against the true class only:

$$I^+_{KB} = \log p - \log p' \qquad \text{(for correct classification)} \tag{4a}$$

$$I^-_{KB} = -\log(1 - p) + \log(1 - p') \qquad \text{(for misclassification)} \tag{4b}$$

This function is mapped for the simple three-class case with a uniform prior probability and varying probabilities for the true class in Figure 3, while being compared with IR. There are two substantive

differences between the two metrics, both of which are unfavorable to Kononenko and Bratko's reward. (1) Their reward function has a kink located at the true class's prior probability, as did our intermediate IR; this reflects their inappropriate concern to even out rewards and punishments, so that their reward function no longer has a suitable information-theoretic interpretation. (2) Their reward function is assessed only against the prior probability of the *true* class. This is a failing again with some analogy to that of our intermediate Definition 2: since the probabilities of false classes are not considered, an overconfident assessment of what is false will go unpunished. For these reasons we do not consider the Kononenko and Bratko function to be adequate; however, we will examine their measure empirically below.

4 Information Reward for a Test Set

Ideally, we would like to know the expected information reward for each machine learner in a domain, or across a set of domains over which we anticipate they will be used. Since we often don't know enough about the domain(s) of application, we may sample from the domain(s), obtaining a test set with which the learners can be evaluated. The cumulative reward, divided by the number of test cases, then serves as a best estimator for expected reward.

	Predicted	
Actual	+ve	-ve
+ve	TP	FN
-ve	FP	TN

Table 1. The four types of possible classification.

For a binary task, where one class is denoted as *Positive (+ve)* and the other as *Negative (-ve)*, there are four different types of classification that can be made: *True Positive (TP)*, where the learner correctly classifies a positive instance, *False Positive (FP)*, where the learner misclassifies an instance as positive (the instance was actually negative), *True Negative (TN)*, where the learner classifies a negative instance correctly, and *False Negative (FN)*, where the learner incorrectly classifies a positive instance as negative, as in Table 1.

Lemma 1. *The cumulative IR for the machine learner ML on a binary classification task, where ML generates n classifications, each classification* $i \in \{1 \ldots n\}$ *having an associated probability* p_i*, is given by:*

$$I(ML) = n - \left(\frac{\log wz}{\log p'} + \frac{\log xy}{\log(1-p')} \right) \tag{5}$$

where,

$$w = \prod_{i \in TP} p_i, \qquad z = \prod_{i \in FN} (1 - p_i),$$
$$x = \prod_{i \in FP} (1 - p_i), \qquad y = \prod_{i \in TN} p_i,$$

and p' *represents the prior probability of the +ve class.*

Proof. The IR for an entire test set is the sum of each IR for individual classifications, thus:

$$\begin{aligned} I(ML) = n &- \sum_{i \in TP} \frac{\log p_i}{\log p'} \\ &- \sum_{i \in FP} \frac{\log(1-p_i)}{\log(1-p')} \\ &- \sum_{i \in TN} \frac{\log p_i}{\log(1-p')} \\ &- \sum_{i \in FN} \frac{\log(1-p_i)}{\log p'} \\ = n &- \frac{\log \prod_{i \in TP} p_i}{\log p'} - \frac{\log \prod_{i \in FP}(1-p_i)}{\log(1-p')} \\ &- \frac{\log \prod_{i \in TN} p_i}{\log(1-p')} - \frac{\log \prod_{i \in FN}(1-p_i)}{\log p'} \end{aligned}$$

And with w, x, y and z as defined in (5),

$$= n - \left(\frac{\log wz}{\log p'} + \frac{\log xy}{\log(1-p')} \right), \text{ as required.}$$

The uniform IR for a test set (that is, where each class has equal prior) is simplified substantially. For the binary case, this is precisely what applying Definition 1 to each case in a test set would yield.

Lemma 2. *The uniform IR on a binary task, corresponding to Definition 1, is denoted by $I_u(ML)$ and simplifies as follows:*

$$I_u(ML) = n + \frac{\log(wxyz)}{\log 2} \tag{6}$$

with w, x, y, z and n as defined in (5).

Proof. Lemma 2 is obtained from Lemma 1 by setting $p' = 1 - p'$ and applying the log laws.

5 Information Reward and Evaluation

With Lemma 1 and Lemma 2 in hand, we may now investigate the application of Bayesian IR to evaluation. We first consider whether IR can make a difference between the relative rankings of two machine learners, ML_1 and ML_2, compared to IR_G. If our generalization cannot make any difference to the relative rankings of machine learners, then it cannot represent any very important improvement upon IR_G.

Thesis 1 *There exists a binary test set, machine learners ML_1 and ML_2 and prior probability p' such that:*

$$I_u(ML_1) < I_u(ML_2) \tag{7a}$$

and,

$$I(ML_1) > I(ML_2) \text{ for some } p' \neq 0.5. \tag{7b}$$

Proof. Substituting (6) into (7a):

$$n + \frac{\log w_1 x_1 y_1 z_1}{\log 2} < n + \frac{\log w_2 x_2 y_2 z_2}{\log 2} \tag{8}$$

$$\iff \quad w_1 x_1 y_1 z_1 < w_2 x_2 y_2 z_2 \tag{9}$$

likelihood ranking.)

$$\iff \qquad \frac{w_1 z_1}{w_2 z_2} < \frac{x_2 y_2}{x_1 y_1} \tag{10}$$

Substituting (5) into (7b):

$$n - \left(\frac{\log w_1 z_1}{\log p'} + \frac{\log x_1 y_1}{\log(1-p')} \right) > \quad n - \left(\frac{\log w_2 z_2}{\log p'} + \frac{\log x_2 y_2}{\log(1-p')} \right) \tag{11}$$

$$\iff \frac{\log w_1 z_1}{\log p'} + \frac{\log x_1 y_1}{\log(1-p')} < \frac{\log w_2 z_2}{\log p'} + \frac{\log x_2 y_2}{\log(1-p')} \tag{12}$$

$$\iff \frac{\log w_1 z_1 - \log w_2 z_2}{\log p'} < \frac{\log x_2 y_2 - \log x_1 y_1}{\log(1-p')} \tag{13}$$

$$\iff \log(1-p') \log \left(\frac{w_1 z_1}{w_2 z_2} \right) < \log p' \log \left(\frac{x_2 y_2}{x_1 y_1} \right) \tag{14}$$

To finish the proof, we simply produce a set of numbers simultaneously satisfying inequalities (10) and (14). Setting $w_1 z_1 = 0.6$, $w_2 z_2 = .5$, $x_2 y_2 = .7$, $x_1 y_1 = .3$, together with a prior $p' = .8$ suffices. Thus the thesis is proven.

6 Results

For this study we tested a number of well-known machine learning algorithms, using the same datasets employed by Holte (1993) and Korb et al. (2001):

- **C5.0** (Quinlan, 1998): C5.0 is an improvement over C4.5, and comes with the option of *boosting*. C5.0 was run with both boosting enabled (cB) and disabled (c5).
- **Causal MML (ca)** (Wallace et al., 1996): This learns Bayesian Networks from data using the Minimum Message Length (MML) principle (Wallace & Boulton, 1968).
- **Naive Bayes (nb)**: These simple Bayesian net models split on class membership, with the leaves representing the different available attributes. They are "naive" because they assume that the attributes, given knowledge of class membership, are independent of each other. Observed attribute values are filled in and a simple Bayesian net propagation provides a posterior probability of class membership. We implemented the algorithm as described by Mitchell (1997).

Each of these learning algorithms allowed classification probabilities to be read.

The empirical evaluation in this study was performed using Dietterich's 5x2cv paired t test, which has been shown in his empirical work to be superior to standardly used tests (Dietterich, 1998). Briefly, this method requires 5 replications of 2-fold cross-validation and approximates a t test with 5 degrees of freedom. The method is used because it more closely approximates the t distribution by better supporting the independence assumptions required than more common tests such as the resampled paired t test and other cross-validation techniques (Dietterich, 1998).

Three different evaluation measures are shown. Predictive accuracy is calculated by giving each classification a score of 1 if the true class is given the highest probability by the machine learner, and 0 otherwise. Kononenko and Bratko's measure is calculated as in Section 3. Information reward is calculated as in Definition 3. All the measures are normalised by dividing by the number of items in the test set.

Since both IR and Kononenko and Bratko's measure can penalize wrong predictions without limit — for example, probabilities of 0 and 1 correspond to offering infinite odds, and so when wrong are penalized infinitely — we applied a cutoff to extreme probability estimates supported by MML theory (Dowe, 2000), enforcing the range $\left[\frac{0.5}{n+(0.5k)}, \frac{n+0.5}{n+(0.5k)} \right]$ where n is the sample size and k is the number of classes.[1]

[1] C5.0, and to a lesser extent CaMML, reported numerous classes with extreme probabilities; thus, this adjustment gave them an important advantage in estimating these reward metrics. Only Naive Bayes avoided extreme probability estimates on its own.

The results for predictive accuracy are shown in Table 2 and Table 3. Notice that Naive Bayes is ranked worse than all other learners five times (those being datasets ch, hy, mu, se and vo).

Kononenko and Bratko's measure is shown in Table 4 and Table 5. Under this measure, Naive Bayes does even worse than on accuracy. It is deemed worse on the ir dataset, as well as those found using the accuracy measure.

Information reward is shown in Table 6 and Table 7. Using information reward, Naive Bayes is shown to be better than the other machine learners, contrary to both the accuracy and Kononenko and Bratko's measures. Although Naive Bayes appears to be soundly beaten by the alternatives on the ch and mu datasets, on the whole it is substantially better, with 10 statistically significant results outperforming them. For example, in the hy data set C5.0 with boosting outperforms Naive Bayes to statistical significance in both accuracy and KB reward, but this verdict is reversed with IR.

7 Discussion and Conclusion

The interpretation of IR presented in this paper has numerous advantages over that presented by Good (1952), used to evaluate football tipsters (Dowe et al., 1996) and recently machine learners (Korb et al., 2001).

The key argument for generalized IR rests on the interpretation of information. Information reduces uncertainty about the world. Thus when a machine learner correctly classifies an instance with probability p, p must be greater than the prior probability p' to inform, or reduce uncertainty. This is reflected in the definition of generalized IR; thus $p > p'$ is rewarded and $p < p'$ is penalized, given correct classification. Given misclassification, $p < p'$ is rewarded and $p > p'$ penalized. This can be interpreted as the following: the machine learner indicated that the probability p of the event was less than what you had expected (p'). That event did not occur, so the learner should be rewarded for reducing the expectation in the event, while if p was increased, the expectation was increased, and thus the learner should be penalized for its estimation.

Information reward is a good objective measure of classification performance. The constant 1 is added so that good machine learners are rewarded, that is $I(ML) > 0$, and bad ones penalized: $I(ML) < 0$. Bad machine learners are actually *misinforming*, relative to the prior! That is, they perform worse than a machine learner who just reports the modal class and its prior for each instance. The average information reward ($I(ML)/n$, where n is the test set size), also has the advantage of being bounded by 1, the value only a perfect predictor could obtain.

Where do the priors come from? In most machine learning tasks, the samples are split into two sets, a training set and a data set. A straightforward prior is to use the relative frequencies of the classes from the training set. Thus the prior is obtained from a source that is accessible by the machine learner. IR also allows us to measure machine learners against any prior standard if we wish, for example one derived from a human expert.

Dataset	c5	cB	ca	nb
bc	0.7622378	0.7622378	0.73426574	0.7692308
ch	0.99248594	0.99686915	0.95241076	0.77332497
g2	0.7160494	0.8518519	0.91358024	0.86419755
gl	0.69158876	0.69158876	0.8224299	0.6168224
hd	0.7086093	0.78807944	0.7549669	0.80794704
he	0.7922078	0.8181818	0.8051948	0.8961039
ho	0.8369565	0.8152174	0.8097826	0.8043478
hy	0.9886148	0.98987985	0.9892473	0.9550917
ir	0.94666666	0.96	0.96	0.97333336
la	0.8214286	0.8214286	0.89285713	0.8214286
ly	0.7702703	0.7027027	0.7432432	0.7702703
mu	1.0	1.0	1.0	0.9054653
se	0.97975963	0.97975963	0.97406703	0.9089184
so	1.0	1.0	1.0	1.0
v2	0.85714287	0.9078341	0.92626727	0.875576
vo	0.95852536	0.95852536	0.9447005	0.89400923

Table 2. Predictive accuracy reported by each machine learner, for each dataset.

Dataset	c5	cB	ca	nb
bc				
ch	ca nb	ca nb	nb	
g2		c5	c5	c5
gl			c5 cB nb	
hd		ca		c5 ca
he				
ho				
hy	nb	nb	nb	
ir				
la				
ly				
mu	nb	nb	nb	
se	ca nb	ca nb	nb	
so				
v2		c5	c5 nb	
vo	nb	nb	nb	

Table 3. Significant differences between machine learners, using the accuracy measure. Each cell records which machine learners were judged inferior to that particular machine learner, on each particular dataset.

Dataset	c5	cB	ca	nb
bc	0.069981635	0.069981635	−0.0018902452	0.13689728
ch	0.6723401	0.64563507	0.5885489	0.16337916
g2	0.2591113	0.35606107	0.5546989	0.37491697
gl	0.81675977	0.86755884	1.130047	0.65977937
hd	0.2506573	0.31211603	0.3526309	0.32222697
he	0.1018493	0.13039692	0.09669634	0.23335941
ho	0.3245233	0.31426513	0.3549378	0.28417683
hy	0.1355615	0.07641858	0.13847339	−0.014177129
ir	1.0032248	0.9701727	1.0044801	0.7756945
la	0.24943754	0.24943754	0.3933418	0.3677128
ly	0.42263785	0.39556766	0.43005717	0.3594238
mu	0.69259095	0.69259095	0.69259053	0.55018294
se	0.22789803	0.22789803	0.21835881	0.049714945
so	1.3987643	1.3987643	1.4630065	1.2994881
v2	0.43209726	0.48362544	0.52688074	0.47609127
vo	0.5957392	0.5655813	0.59266555	0.5302857

Table 4. Kononenko and Bratko reward reported by each machine learner, for each dataset.

Dataset	c5	cB	ca	nb
bc	ca	ca		ca
ch	ca cB nb	ca nb	nb	
g2			c5 cB nb	c5
gl		nb	c5 cB nb	
hd				
he				c5 ca
ho			nb	
hy	nb	nb	nb	
ir	nb	nb	nb	
la				
ly				
mu	nb	nb	nb	
se	nb	nb	nb	
so				
v2		c5	c5 cB nb	
vo	cB nb	nb	cB nb	

Table 5. Significant differences between machine learners, using the Kononenko and Bratko reward. Each cell records which machine learners were judged inferior to that particular machine learner, on each particular dataset.

Dataset	c5	cB	ca	nb
bc	−0.23610511	−0.23610511	−0.002686911	0.2808528
ch	0.9522189	0.93266416	0.79834837	0.2326152
g2	−0.49246055	0.45041668	0.7281166	0.5455182
gl	0.07175887	0.2923216	0.5108723	0.44003078
hd	−0.22732624	0.20924993	0.22064793	0.37742698
he	−0.15546304	0.0129868025	−0.4243838	0.5606905
ho	0.18764098	0.28545344	0.20742324	0.3494705
hy	0.5477152	0.46073914	0.60719365	0.6828912
ir	0.79086244	0.82428205	0.81571364	0.64188224
la	0.2903297	0.2903297	0.3076863	0.5979209
ly	−0.40623546	0.1889826	0.17918634	0.61388963
mu	0.9998209	0.9998209	0.99982065	0.6382106
se	0.58898747	0.58898747	0.4046343	0.62509626
so	0.9015267	0.9015267	0.9772743	0.8461196
v2	0.4839226	0.5787031	0.58912724	0.35221466
vo	0.8161831	0.7789633	0.81813234	0.62418294

Table 6. Information reward reported by each machine learner, for each dataset.

Dataset	c5	cB	ca	nb
bc				ca cB
ch	ca nb	ca nb	nb	
g2		c5	c5 cB nb	c5
gl			cB	
hd		c5	c5	c5 ca cB
he		ca		ca cB
ho				
hy				cB
ir				
la				
ly				cB
mu	nb	nb	nb	
se				
so				
v2		nb		
vo				

Table 7. Significant differences between machine learners, using information reward. Each cell records which machine learners were judged inferior to that particular machine learner, on each particular dataset.

Bibliography

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *7*, 1895–1924.

Dowe, D. L. (2000). *Learning and prediction notes.* School of Computer Science and Software Engineering, Monash University.

Dowe, D. L., Farr, G. E., Hurst, A. J., & Lentin, K. L. (1996). *Information-theoretic football tipping* (Technical Report 96/297). Dept. of Computer Science, Monash University.

Good, I. J. (1952). Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *14*, 107–114.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*, 63–91.

Kononenko, I., & Bratko, I. (1991). Information-based evaluation criterion for classifier's performance. *Machine Learning*, *6*, 67–80.

Korb, K. B., Hope, L. R., & Hughes, M. J. (2001). The evaluation of predictive learners: Some theoretical and empirical results. *European Conference on Machine Learning (ECML'01)* (pp. 276–287).

Mitchell, T. (1997). *Machine learning.* McGraw-Hill.

Provost, F., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97).* AAAI Press.

Quinlan, R. (1998). *Data mining tools See5 and C5.0* (Technical Report). RuleQuest Research.

Turney, P. (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, *2*, 369–409.

Wallace, C., & Boulton, D. (1968). An information measure for classification. *The Computer Journal*, *11*, 185–194.

Wallace, C. S., Korb, K. B., & Dai, H. (1996). Causal discovery via MML. *Internatinal Conference on Machine Learning (ICML'96)* (pp. 516–524). Morgan Kaufmann Publishers.