Evaluating Probability Estimates from Decision Trees

Nitesh V. Chawla and **David A. Cieslak**
{nchawla,dcieslak}@cse.nd.edu
Department of Computer Science and Engineering
University of Notre Dame, IN 46556

Abstract

Decision trees, a popular choice for classification, have their limitation in providing good quality probability estimates. Typically, smoothing methods such as Laplace or m-estimate are applied at the decision tree leaves to overcome the systematic bias introduced by the frequency-based estimates. An ensemble of decision trees has also been shown to help in reducing the bias and variance in the leaf estimates, resulting in better calibrated probabilistic predictions. In this work, we evaluate the calibration or quality of these estimates using various loss measures. We also examine the relationship between the quality of such estimates and resulting rank-ordering of test instances. Our results quantify the impact of smoothing in terms of the loss measures, and the coupled relationship with the AUC measure.

Introduction

Decision trees typically produce crisp classifications; that is, the leaves carry decisions for individual classes. However, that is insufficient for various applications. One can require a score output from a supervised learning method to rank order the instances. For instance, consider the classification of pixels in mammogram images as possibly cancerous (Chawla *et al.* 2002). A typical mammography dataset might contain 98% normal pixels and 2% abnormal pixels. A simple default strategy of guessing the majority class would give a predictive accuracy of 98%. Ideally, a fairly high rate of correct cancerous predictions is required, while allowing for a small to moderate error rate in the majority class. It is more costly to predict a cancerous case as non-cancerous than otherwise. Thus, a probabilistic estimate or ranking of cancerous cases can be decisive for the practitioner. The cost of further tests can be decreased by thresholding the patients at a particular rank. Secondly, probabilistic estimates can allow one to threshold ranking for class membership at values < 0.5.

Thus, the classes assigned at the leaves of the decision trees have to be appropriately converted to reliable probabilistic estimates. However, the leaf frequencies can require smoothing to improve the "quality" of the estimates (Provost & Domingos 2003; Pazzani *et al.* 1994; Smyth, Gray, & Fayyad 1995; Bradley 1997; Ferri, Flach, & Hernandez-Orallo 2003; Margineantu & Dietterich 2001). A classifier is considered to be well-calibrated if the predicted probability approaches the empirical probability as the number of predictions goes to infinity (DeGroot & Fienberg 1983). Previous work has pointed out that probability estimates derived from leaf frequencies are not appropriate for ranking test examples (Zadrozny & Elkan 2001). This is mitigated by applying smoothing to the leaf estimates. On the other hand, Naive Bayes classifier can produce poor probability estimates but still result in more useful ranking than probabilistic decision trees (Domingos & Pazzani 1996). This begs a careful evaluation of the "quality of probability estimates".

The focus of our paper is to measure the probability estimates using different losses. We also want to quantify the improvement offered by the smoothing methods when applied to the leaf frequencies — *is there a significant decrement in the losses, thus implying an improvement in quality (or calibration) of the estimates?* Finally, we empirically motivate the relationship between the quality of estimates produced by decision trees and the rank-ordering of the test instances. That is, *does the AUC improve once the model is well-calibrated?* We believe it is important to study and quantify the relationship between the calibration of decision trees and the resulting rank-ordering. We implemented the following loss measures[1].

- *Negative Cross Entropy (NCE).* This measure was also utilized for evaluating losses in the NIPS 2003 Challenge on Evaluating Predictive Uncertainty (Candella 2004).
- *Quadratic Loss (QL).*
- *Error (0/1 Loss: 01L).*

We then correlate these with the ROC curve analysis (Bradley 1997; Provost & Fawcett 2001). We use ROC curves and the resulting AUC to demonstrate the sensitivity of ranking to the probabilistic estimates. Due to space constraints, while the ROC curves are not directly presented, the AUC as an indicator of the rank-ordering achieved on the test examples is included. This leads us to the question: *Is there an empirically justified correlation between the loss measures and AUC?*

[1] The equations are provided in the subsequent sections

Probabilistic Decision Trees with C4.5

The decision tree probability estimates, which are a natural calculation from the frequencies at the leaves, can be systematically skewed towards 0 and 1, as the leaves are essentially dominated by one class. For notational purposes, let us consider the confusion matrix given in Figure 1. TP is the number of true positives at the leaf, and FP is the number of false positives. Typically, the probabilistic (frequency-based) estimate at a decision tree leaf for a class y is:

	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Figure 1: Confusion Matrix

$$P(y|x) = TP/(TP + FP) \tag{1}$$

However, simply using the frequency derived from the correct counts of classes at a leaf might not give sound probabilistic estimates (Provost & Domingos 2003; Zadrozny & Elkan 2001). A small leaf can potentially give optimistic estimates for classification purposes. For instance, the frequency based estimate will give the same weights to leaves with the following (TP, FP) distributions: $(5, 0)$ and $(50, 0)$. The relative coverage of the leaves and the original class distribution is not taken into consideration. Given the evidence, a probabilistic estimate of 1 for the $(5, 0)$ leaf is not very sound. Smoothing the frequency-based estimates can mitigate the aforementioned problem (Provost & Domingos 2003).

Aiming to perfectly classify the given set of training examples, a decision tree may overfit the training set. Overfitting is typically circumvented by deploying various pruning methodologies. But pruning deploys methods that typically maximize accuracies. Pruning is equivalent to coalescing different decision regions obtained by thresholding at feature values. This can result in coarser probability estimates at the leaves. While pruning improves the decision tree generalization, it can give poorer estimates as all the examples belonging to a decision tree leaves are given the same estimate.

Smoothing Leaf Frequencies

One way of improving the probability estimates given by an unpruned decision tree is to smooth them to make them less extreme. One can smooth these estimated probabilities by using the Laplace estimate (Provost & Domingos 2003), which can be written as follows:

$$P(y|x) = (TP + 1)/(TP + FP + C) \tag{2}$$

Laplace estimate introduces a prior probability of $1/C$ for each class. Again considering the two pathological cases of $TP = 5$ and $TP = 50$, the Laplace estimates are 0.86 and 0.98, respectively, which are more reliable given the evidence.

However, Laplace estimates might not be very appropriate for highly unbalanced datasets (Zadrozny & Elkan 2001). In that scenario, it could be useful to incorporate the prior of positive class to smooth the probabilities so that the estimates are shifted towards the minority class base rate (b). The m-estimate (Cussens 1993) can be used as follows (Zadrozny & Elkan 2001):

$$P(y|x) = (TP + bm)/(TP + FP + m) \tag{3}$$

where b is the base rate or the prior of positive class, and m is the parameter for controlling the shift towards b. Zadrozny and Elkan (2001) suggest using m, given b, such that $bm = 10$.

Niculescu-Mizil & Caruana (2005) explore two smoothing methods not surveyed in this paper: Platt Calibration and Isotonic Regression. Both are powerful calibrating methods, which rely for minimizing loss by searching an argument space to find improved probability estimates. A comparison between these two and the other smoothing methods is part of our future work.

Bagged Decision Trees

We use ensemble methods to further "smooth" out the probability estimates at the leaves. Each leaf will potentially have a different $P(y|x)$ due to different training set composition. Averaging these estimates will improve the quality of the estimates, as it overcomes the bias introduced by the systematic error caused by having axis-parallel splits. The overfitting will also be countered as the variance component will be reduced by voting or averaging. Bagging (Breiman 1996), has been shown to improve classifier accuracy. Bagging basically aggregates predictions (by voting or averaging) from classifiers learned on multiple bootstraps of data.

The hyperplanes constructed for each tree will be different, as each tree is essentially constructed from a bootstrap. The classification can either be done by taking the most popular class attached to the test example or by aggregating the probability estimate computed from each of the subspaces. Each tree has a potentially different representation of the original data set, thus resulting in a different function for $P(y|x)$ at each leaf. The classification assigned by the individual decision trees is effectively invariant for test examples.

We let the trees grow fully to get precise estimates, as the averaging will then reduce the overall variance in the estimates. Let $\hat{p}_k(y|x)$ indicate the probability assigned by a tree k to a test example x. $\hat{p}$ can either be the leaf frequency based estimate or smoothed by Laplace or m-estimate. Then, $g_y(x)$ averages over probabilities assigned by the different trees to a test example.

$$g_y(x) = \frac{1}{K}\sum_{k=1}^{K} \hat{p}(y_k|x) \tag{4}$$

Each leaf is, in essence, defining its own region of probability distribution. Since, the trees are constructed from bags, the regions can be of different shapes and sizes. The individual classifiers can be weaker than the aggregate or even the global classifier. An aggregation of the same can lead to a reduction in the variance component of the error term, thereby reducing the overall error (Breiman 1996).

Zadrozny & Elkan (2001) find that bagging doesn't always improve the probability estimates for large unbalanced datasets. However, we show that even for large and unbalanced datasets, there is an improvement in the quality of probabilistic estimates. Our findings are in agreement with the work of Provost & Domingos (2003) and Bauer & Kohavi (1999).

Loss Measures

As mentioned in the Introduction, we used the following three loss measures to evaluate the quality of the probability estimates. We will assume a two-class case ($y \in 0, 1$), x_i is a test instance, and c is the actual class of x_i.

- NCE: The NCE measure is the average Negative Cross Entropy of predicting the true labels of the testing set instances. Thus, it can be considered as the measure that must be minimized to obtain the maximum likelihood probability estimates. One word of caution with NCE is that it will be undefined for $log(0)$. Thus, the minimum loss is 0, but the maximum can be infinity if $p(y = 1|x_i) = 0$ or $1 - p(y = 1|x_i) = 0$. NCE essentially measures the proximity of the predicted values to the actual class values. That is, the class 1 predictions should have probabilities closer to 1.

$$NCE = -\frac{1}{n}\{(\sum_{i|y=1} \log(p(y = 1|x_i)) + \sum_{i|y=0} \log(1 - p(y = 1|x_i)))\}$$

- QL: The QL measure is the average Quadratic Loss occured on each instance in the test set. The QL indicates predictions that make the best estimates at the true probabilities. It not only accounts for the probability assigned to the actual class, but also the probabilities assigned for the other possible class. Thus, the more confidence we have in predicting the actual class, the lower the loss. The quadratic loss is averaged over all the test instances. In the subsequent equation, the squared term sums over all the possible probability values assigned to the test instance, which is two for our case. For instance, the worst case will be when $p(y = 1|x_i) = 0$, when true label $y = 1$. This will lead to $1 - 2 \times 0 + (1 + 0)^2 = 2$. The best case will be when $p(y = 1|x_i) = 1$, $y = 1$. This will lead to $1 - 2 \times 1 + (1 + 0)^2) = 0$.

$$QL = \frac{1}{n}\sum_i\{1 - 2p(y = c|x_i) + \sum_{j\in 0,1} p(y = j|x_i)^2\}$$

- O1L: This is the classification error, 0/1 loss, where the estimated probabilities are thresholded at 0.5 for generating the classification.

Experiments with Unbalanced Datasets

A dataset is unbalanced if the classes are not approximately equally represented (Chawla *et al.* 2002; Japkowicz & Stephen 2002). There have been attempts to deal with unbalanced datasets in domains such as fraudulent telephone calls (Fawcett & Provost 1996), telecommunications management (Ezawa, Singh, & Norton 1996), text classification (Dumais *et al.* 1998; Mladenić & Grobelnik 1999; Cohen 1995) and detection of oil spills in satellite images (Kubat, Holte, & Matwin 1998). Distribution/cost sensitive applications can require a ranking or a probabilistic estimate of the instances. Hence, the classes assigned at the leaves of the decision trees have to be appropriately converted to probabilistic estimates (Provost & Domingos 2003). This brings us to another question: *What is the right probabilistic estimate for unbalanced datasets?*

Table 1 summarizes the datasets. We divided our datasets into 70% and 30% splits for training and testing, respectively (hold-out method). We generated 30 bags for each of the datasets. As in the related work with probability estimation decision trees, we primarily focus on our results obtained from unpruned decision trees. We used C4.5 decision trees for our experiments (Quinlan 1992). However, we do include some of the results using pruned trees without bagging.

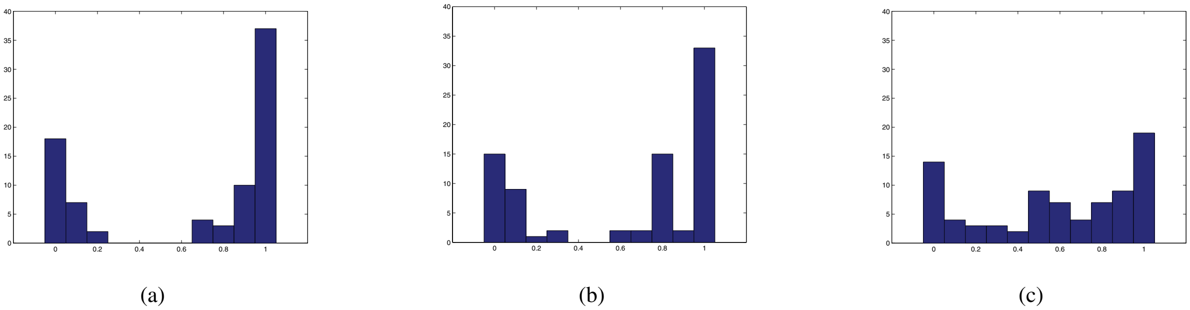
Results

The main goal of our evaluation is to understand and demonstrate a) the impact of smoothing on the probabilitiy and resulting losses and AUC; b) the relationship between the quality of probability estimates produced by decision trees and AUC. For this we utilize the NCE and QL losses to indicate the quality of the probability predictions; and c) the accuracy of the point predictions (error estimate or 0/1 loss). Figures 2 shows the distribution of $p(y_i = +1|x_i)$ for all the positive class (+1) examples, generated from different scenarios, for the mammography dataset. One would expect smoothing to overcome the skewness (bias) in the estimates at the leaves, resulting in broadly distributed estimates. Similar trends as Figure 2 were observed for the other datasets as well. Thus, there is clear evidence of the impact of applying smoothing to the leaf estimates, particularly using ensemble methods such as bagging.

We are interested in objectively quantifying the resulting improvement in the estimates, obtained by smoothing, with respect to the different losses. We also want to demonstrate the relationships between the quality of the estimates, defined by losses, and the rank-ordering, defined by AUC. To that end, we set the unpruned decision trees with leaf-frequency based estimates as our baseline. We then calculate the relative difference, on the losses and AUC, between the different smoothing techniques and this baseline. If $TreeMethod$ is the measure generated from either of {Laplace, M-estimate and Bagging}, and $UnprunedTree$ is the measure derived from the unsmoothed leaf-frequency estimate, then $RelativeDifference = \frac{TreeMethod - UnprunedTree}{UnprunedTree}$. We want to look for the following trends. If the difference for

Table 1: Datasets.

Dataset	Number of Features	Number of Examples	Proportion of Minority (positive) class examples
Phoneme	5	5,400	0.29
Adult	14	48,840	0.24
Satimage	36	6,430	0.097
Forest Cover	54	38,500	0.071
Mammography	6	11,183	0.023



(a)

(b)

(c)

Figure 2: a) Probability Distribution using the leaf frequencies as estimates. b) Probability distribution by smoothing leaf frequencies using Laplace estimates. c) Probability Distribution using bagging. The probabilities are $g_y(x)$ that are averaged from the (frequency-based) leaf estimates.

QL and NCE is negative, then smoothing the estimates is actually improving the quality of our estimates and reducing the loss. If the difference for error is positive, then the point predictions (accuracy) are deteriorating. If the difference for AUC is positive, then the rank-ordering of the test examples is improving. Figure 3 show the results.

There are various observations from this Figure. Smoothing invariably and consistently reduces the NCE loss. The Bagged-Laplace trees are the most consistent in their quality of probabilistic predictions based on both QL and NCE. The error rate shows an interesting trend. Laplace estimate results in no change in error for 4 out 5 datasets in no change in error. On the other hand, error rate is very sensitive to the m-estimate as the probabilities are shifted towards the base rate. As one would expect, bagging always reduces the error rate. However, m-estimate does not result in improved performance over Laplace for any of the datasets. In fact, for the two most skewed datasets — mammography and covtype — m-estimate leads to an increase in the quadratic loss. We also notice a very compelling trend from these Figures. The NCE measure is very tightly inversely correlated with the resulting AUC. Thus, for the probabilistic decision trees, the quality of the estimates directly impacts the resulting rank-ordering. Table 2 shows the correlation coffecient between the different loss measures and AUC. Notably, there is a high negative correlation between NCE and AUC, which adds evidence to our observation that as NCE decreases, AUC increases for the probability estimation decision trees. There were no such trends between O1L and AUC, and QL and AUC. On the other hand, the error estimate (01L) is tightly correlated with QL.

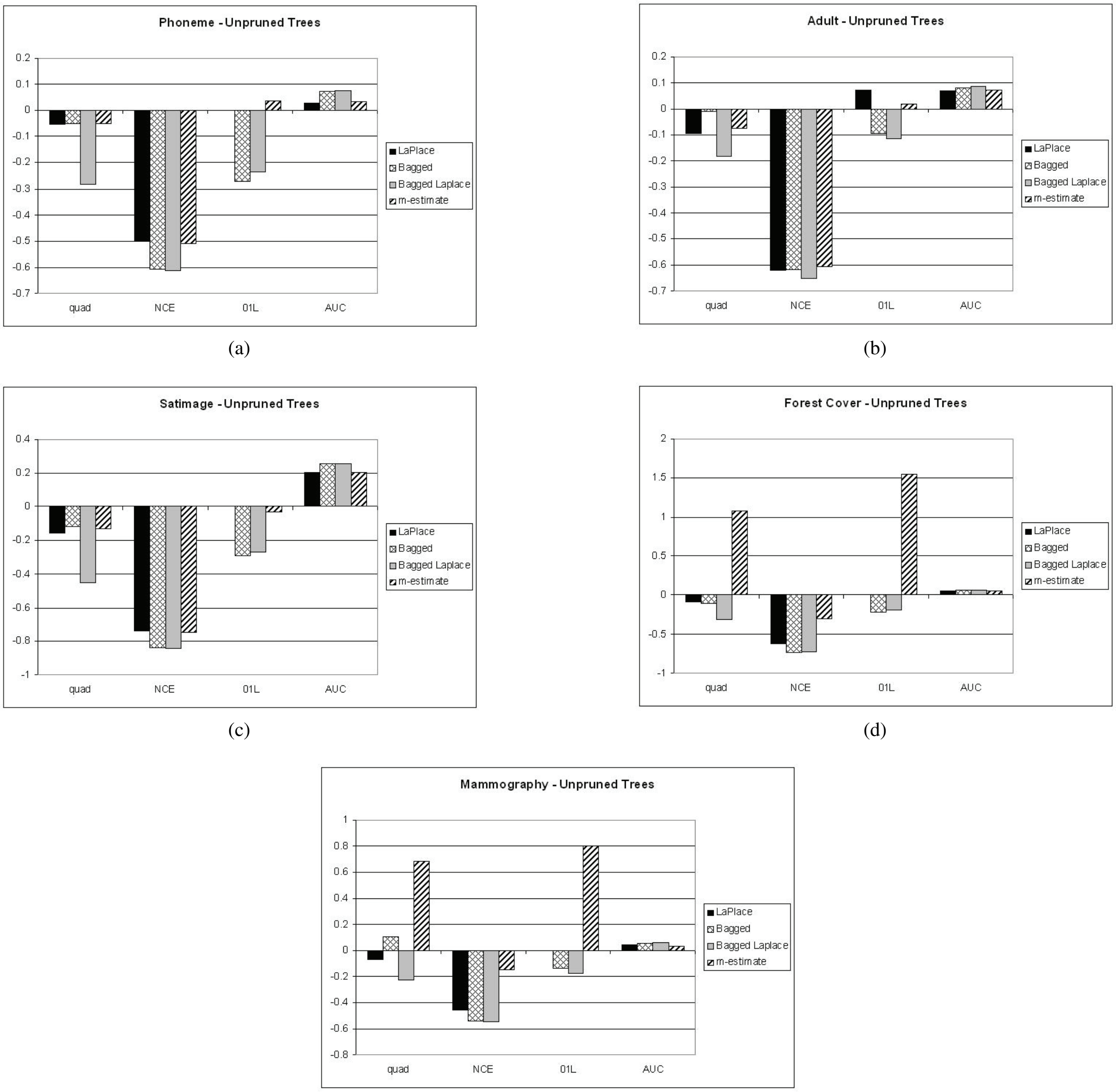
Table 2: Correlation Among the Losses and AUC

	NCE	QL	01L	AUC
NCE	1	0.5044	0.6217	-0.8663
QL	0.5044	1	0.8245	-0.4668
O1L	0.6217	0.8245	1	-0.4931

Table 3 shows the results from applying default pruning. The results are quite compelling. Without smoothing the NCE of pruned trees is consistently lower than unpruned trees. Moreover, the AUC's for pruned trees are also better than unpruned trees. Thus, the coarser unsmoothed leafs are resulting in better quality estimates. However, once the leaves are smoothed by Laplace, there is definitely an advantage in using unpruned trees.

Summary

We show that decision trees are a viable strategy for probability estimates that can be used for rank-ordering the test examples. Laplace estimate, m-estimate, and ensembles are able to overcome the bias in estimates arising from the axis-parallel splits of decision trees, resulting in smoother estimates. We demonstrated that the rank-ordering of the test instances is related to the quality of the probability estimates. For most of the applications requiring unbalanced datasets, the resulting rank-order of examples or $P(Xp > Xn)$ can be very important, where Xp is the positive class example. Thus, having reliable probability estimates is important for an improved rank-ordering. Based on our results, we would



(a)

(b)

(c)

(d)

(e)

Figure 3: Relative differences of the different smoothing methods from the unsmoothed leaf frequency based estimates. The convention in the Figures is as follows: Laplace is the leaf frequency based estimate smoothed by laplace method; Bagged is for the averaged leaf frequency estimates over the 30 bags; Bagged-Laplace is for the averaged Laplace smoothed estimates over the 30 bags; and m-estimate is the leaf frequencie based estimate smoothed by m-estimate.

Table 3: Comparison of probability estimates produced with unpruned and pruned trees.

	NCE				AUC			
	Frequency		LaPlace		Frequency		LaPlace	
Dataset	Unpruned	Pruned	Unpruned	Pruned	Unpruned	Pruned	Unpruned	Pruned
Phoneme	.312174	.293877	.156307	.159342	.749640	.746100	.798680	.782070
Adult	.385605	.157325	.146401	.141756	.668420	.767510	.783980	.777330
Satimage	.451316	.354893	.117407	.112777	.514000	.397530	.822590	.795210
Forest Cover	.069245	.057237	.026180	.026670	.877140	.881790	.974840	.966270
Mammography	.041607	.039046	.022569	.023136	.777647	.781200	.849300	.808600

recommend bagging or other ensemble generation methods with decision trees for improving the calibration of the probability estimates from the decision trees. Bagging effectively reduces the variance and bias in estimation.

References

Bauer, E., and Kohavi, R. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. *Machine Learning* 36(1,2).

Bradley, A. P. 1997. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition* 30(6):1145–1159.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

Candella, J. 2004. Evaluating Predictive Uncertainty Challenge, NIPS 2004.

Chawla, N.; Hall, L.; Bowyer, K.; and Kegelmeyer, W. 2002. SMOTE: Synthetic Minority Oversampling TEchnique. *Journal of Artificial Intelligence Research* 16:321–357.

Cohen, W. 1995. Learning to Classify English Text with ILP Methods. In *Proceedings of the 5th International Workshop on Inductive Logic Programming*, 3–24. Department of Computer Science, Katholieke Universiteit Leuven.

Cussens, J. 1993. Bayes and pseudo-bayes estimates of conditional probabilities and their reliabilities. In *Proceedings of European Conference on Machine Learning*.

DeGroot, M., and Fienberg, S. 1983. The Comparison and Evaluation of Forecasters. *Statistician* 32:12 – 22.

Domingos, P., and Pazzani, M. J. 1996. Beyond independence: Conditions for the optimality of the simple bayesian classifier. In *International Conference on Machine Learning*, 105–112.

Dumais, S.; Platt, J.; Heckerman, D.; and Sahami, M. 1998. Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the Seventh International Conference on Information and Knowledge Management.*, 148–155.

Ezawa, K., J.; Singh, M.; and Norton, S., W. 1996. Learning Goal Oriented Bayesian Networks for Telecommunications Risk Management. In *Proceedings of the International Conference on Machine Learning, ICML-96*, 139–147. Bari, Italy: Morgan Kauffman.

Fawcett, T., and Provost, F. 1996. Combining Data Mining and Machine Learning for Effective User Profile. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 8–13. Portland, OR: AAAI.

Ferri, C.; Flach, P. A.; and Hernandez-Orallo, J. 2003. Improving the auc of probabilistic estimation trees. In *European Conference on Machine Learning*, 121–132.

Japkowicz, N., and Stephen, S. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis* 6(5).

Kubat, M.; Holte, R.; and Matwin, S. 1998. Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Machine Learning* 30:195–215.

Margineantu, D., and Dietterich, T. 2001. Improved class probability estimates from decision tree models. In *Nonlinear Estimation and Classification*, 169–184.

Mladenić, D., and Grobelnik, M. 1999. Feature Selection for Unbalanced Class Distribution and Naive Bayes. In *Proceedings of the 16th International Conference on Machine Learning.*, 258–267. Morgan Kaufmann.

Niculescu-Mizil, A., and Caruana, R. 2005. Predicting good probabilities with supervised learning. In *International Conference on Machine Learning*, 625–632.

Pazzani, M.; Merz, C.; Murphy, P.; Ali, K.; Hume, T.; and Brunk, C. 1994. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on Machine Learning*, 217–215.

Provost, F., and Domingos, P. 2003. Tree induction for probability-based rankings. *Machine Learning* 52(3).

Provost, F., and Fawcett, T. 2001. Robust Classification for Imprecise Environments. *Machine Learning* 42/3:203–231.

Quinlan, J. 1992. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

Smyth, P.; Gray, A.; and Fayyad, U. 1995. Retrofitting decision tree classifiers using kernel density estimation. In *Proceedings of the Twelth International Conference on Machine Learning*, 506–514.

Zadrozny, B., and Elkan, C. 2001. Learning and making decisions when costs and probabilities are both unknown. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*.