Accurate Robust and Efficient Error Estimation for Decision Trees

Lixin Fan LIXIN.FAN@NOKIA.COM
Nokia Technologies, Valtatie 30, Tampere, Finland

Abstract

This paper illustrates a novel approach to the estimation of generalization error of decision tree classifiers. We set out the study of decision tree errors in the context of *consistency analysis* theory, which proved that the *Bayes error* can be achieved only if when the number of data samples thrown into each leaf node goes to infinity. For the more challenging and practical case where the sample size is finite or small, a novel sampling error term is introduced in this paper to cope with the small sample problem effectively and efficiently. Extensive experimental results show that the proposed error estimate is superior to the well-known K-fold cross validation methods in terms of *robustness* and *accuracy*. Moreover it is orders of magnitudes more *efficient* than cross validation methods.

1. Introduction

The aim of decision tree classifier learning is to construct a *tree-based predictive model* which maps unseen data *observations* to designated class *labels*. Often the learning process is driven by fitting a given set of training data and corresponding class labels with recursively partitioned leaf nodes e.g. as in (Breiman et al., 1984). While in recent years a number of learning approaches have demonstrated more advanced features and superior performances than decision tree approaches, the study of fundamental learning mechanisms of decision trees still give deep insights into the driving force of advanced learning algorithms e.g. random forest (Breiman, 2001). In particular the estimation of the *probability of classification errors* on unseen data i.e. the *generalization error* of a learned decision tree is of the utmost importance.

It is well known that the generalization error of any learning rules is bounded below by the *Bayes error* or *Bayes risk*. It is also shown by Theorem 6.1 of (Devroye et al., 1996) that the Bayes error is achieved for decision tree classifiers if two conditions are satisfied: 1) the *diameter* of decision tree leaf node cells decreases to infinitesimal, and 2) the number of observations thrown into each leaf node goes to infinity. While the asymptotic performance of a learned decision tree is guaranteed by the Theorem, in this article we address the open issue of *how to estimate the generalization error when above two conditions are not fulfilled*. The goal to make estimation under such circumstances is motivated by the unfortunate facts that a) the number of data samples is often limited in practice, and b) even for large datasets with millions of samples the small sample problem may still persist in certain leaf nodes attached to exceedingly long branches.

This paper shows that by applying the consistency analysis theory, one is able to decompose the upper bound of generalization error into the *quantized Bayes error* and the *sampling error*. We prove that the *quantized Bayes error* is the minimal error bound that one can reach for a given set of partitioning leaf nodes, no matter how node-wise class posteriors are actually estimated. The upper bound of the *sampling error*, on the other hand, is related to *variances* and *biases* of estimated node-wise class posteriors, and inversely depends on the number of data samples thrown into each leaf nodes. The resultant generalization error estimate, without resorting to explicit complexity analysis of tree models, can be readily used to prevent overfitting of deep decision trees.

It was shown on ten benchmark datasets that the proposed generalization error estimate compared favourably with the well-known K-fold cross validation methods in terms of both *robustness* and *accuracy* (see **Figure 1** for a summarized comparison). Moreover the proposed method is orders of magnitudes more *efficient* than cross validation methods, since repeated training of decision trees on validation datasets is no longer needed.

The layout of this paper is as follows. Section 2 reviews related K-fold cross validation methods, and model complexity based error bounds. Section 3 lays down theoretic foundation for the proposed generalization error estimate from a consistency analysis point of view. Section 4 com-

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

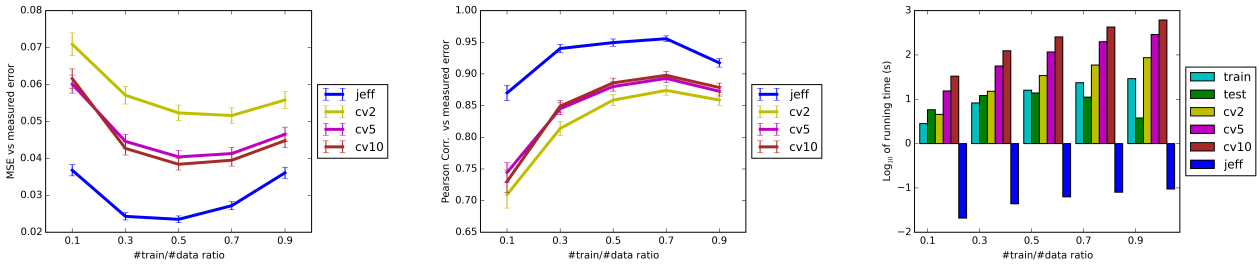

Figure 1. Summarized performance comparison of the proposed method ("jeff") vs K-fold cross validation methods ("cv?"). **X-axis** shows the varying ratio $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ of the number of training samples over the total number of dataset samples. **Left**: **Y-axis** shows Mean Squared Error between the *estimated errors* with respect to the *errors measured with ground truth labels*. Lower MSEs correspond to more *accurate* error estimation, and lower standard deviations of MSEs (denoted by error bars) manifest more *robust* estimation. **Center**: Y-axis shows Pearson Correlation (PE) between the *estimated errors* with respect to the measured errors. Higher PEs are more favourable for the sake of parameter selection. **Right**: Y-axis shows the logarithm of running time of different methods as well as that of training and testing stages. Note that above plots are averaged over ten benchmark datasets (see Table 4.1).

pares the proposed error estimation with the well known K-fold cross validation method, followed by conclusive remarks in Section 5 and proof of propositions in appendix.

2. Related Work

A number of error bounds were used to *prune decision trees* and these bounds often consist of two additive error terms: 1) the *empirical error* on the training set samples and 2) penalties on *model complexities*. A large variety of model complexity penalty terms have been proposed in the literature, such as structure complexities of subtrees in (Kearns & Mansour, 1998), tighter bound depending on the length of concept class encoding strings in (Mansour & McAllester, 2000), data-dependent complexities of hypothesis class in (Freund, 1998), tree structure based microchoice bounds in (Langford & Blum, 2000), Rademacher complexity based bounds in (Bartlett & Mendelson, 2003; Kääriäinen & Elomaa, 2003) and sample compression and Occam's Razor bounds in (Shah, 2007). The theoretic foundation for integrating these complexity terms with empirical training error is not fully understood though. In this paper we provide a novel and theoretically sound justification in light of reduced sampling errors.

In the context of *consistency analysis* theory, it was shown in (Devroye et al., 1996) that a variety of decision tree classifiers are in general consistent for all possible distributions of data points (see Chpt 6, 20 and 21). This consistency analysis also lays down theoretic foundation for us to investigate generalization error when only *finite number of data samples are used* in learning. Unfortunately it turns out no general classification rules can reach the Bayes error under this circumstance (see Chap 7 in (Devroye et al., 1996)). As shown in Section 3.2.2 of this paper, sampling errors ascribed to the small sample problem can be taken into account in a data-dependent manner.

The practical generalization error estimation techniques include the well-known bootstrap in (Efron & Tibshirani, 1993) and cross validation in (Stone, 1997). In particular K-fold cross validation had demonstrated superior empirical performances for various applications e.g. for linear regression feature selection in (Breiman & Spector, 1992) and for decision tree based model selection in (Kohavi, 1995). K-fold cross validation is often performed on a validation dataset, which is randomly selected from the training dataset in multiple runs, thus obviating the need for a separated test datasets. Nevertheless there are two main issues concerning cross validation methods: a) the exceedingly long estimation time involved in repeatedly training of decision trees on validation datasets; b) the high variance in error estimation caused by inefficient use of data samples. Other pitfalls concerning K-fold cross validation are also discussed in Section 4.

Representative tree-pruning methods include the Pessimistic Error Pruning (PEP) proposed by Quinlan and the Minimum Error Pruning (MEP) proposed by Niblett and Bratko following Cestnik and Bratko's *m-probability estimate* (Esposito et al., 1997). While these methods also share the same computational advantages over cross-validation methods, nevertheless, these error estimates are inconsistent and tend to either under-estimate or over-estimate when applied to small-sized datasets as shown in Section 4 of this paper.

3. Generalization error estimation

In this section we first briefly review Corollary 6.1 from (Devroye et al., 1996) and apply it to estimate the generalization error for decision tree classifiers. Our initial focus is

on the binary classification case, followed by the extension to multiple classes in Section 3.2.2.

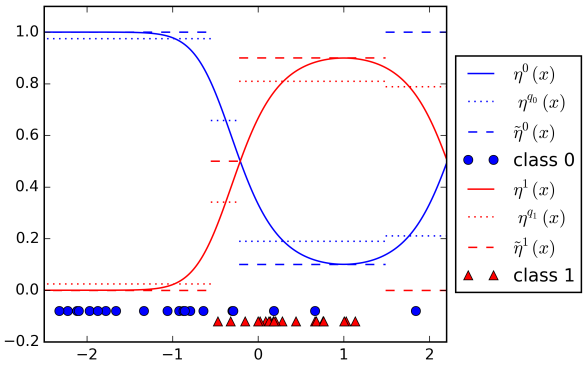

Figure 2. A toy example in which 1D sample points, denoted by ● and ▲, are randomly generated from gaussian distributions $\mathcal{N}(-1.0, 1.0)$ and $\mathcal{N}(0.5, 0.5)$. **X-axis** denotes the sample space and **Y-axis** the probability. Corresponding true posterior $\eta^i(x)$, piecewise quantized posterior $\eta^{q_i}(x)$ and piecewise estimated posterior $\tilde{\eta}^i(x)$ are shown for classes $i = \{0, 1\}$. In this example four leaf nodes are selected by an optimized CART algoirthm (Pedregosa et al., 2011). Note that overfiiting is incurred for the right-most cell to which only one blue sample point ● is assigned.

3.1. Generalization error for plug in decision function

Let (X, Y) be a pair of random variables representing data observations and corresponding labels which takes their respective values from $\mathcal{R}^d$ and $\{0, 1\}$. And $\eta(x) = \mathbf{P}\{Y = 1|X = x\} = E\{Y|X = x\}$ is the class conditional posterior probability that the class label $Y = 1$ given observation $X = x$. Thus the *Bayes error* refers to the probability of error $L^* = \mathbf{P}\{g^*(X) \neq Y\}$ for the Bayes decision function as such:

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

The true posterior probability function $\eta(x)$ is often unknown in reality and one has to approximate it by some function $\tilde{\eta}(x, D_n)$ estimated from the *i.i.d* training samples $D_n = \{(X_1, Y_1), (X_2, Y_2), ...(X_n, Y_n)\}$.

According to Theorem 2.1 of (Devroye et al., 1996), the corresponding plug-in function

$$g_n(x) = \begin{cases} 1 & \text{if } \tilde{\eta}(x, D_n) > 1/2 \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

has the probability of error $L(g_n) = \mathbf{P}\{g_n(X) \neq Y|D_n\}$ that is always *no less than* the Bayes error. In fact the extra error incurred by this plug-in function (on top of the Bayes error) is bounded by the expectation of L_1 distance between the posterior probability $\eta(x)$ and its estimate $\tilde{\eta}(x, D_n)$ according to Corollary 6.1 of (Devroye et al., 1996):

$$L^* \leq L(g_n) \leq L^* + 2\int_{\mathcal{R}^d} |\eta(x) - \tilde{\eta}(x, D_n)|\mu(dx) \tag{3}$$

where μ is the probability measure for X.

3.2. Generalization error for decision trees

Suppose a decision tree partitions observation space $S \subset \mathcal{R}^d$ into N disjointed cells $\mathbb{A} = \{A_1, ..., A_N\}$ corresponding to N leaf nodes, the posterior probability over S is thus approximated by a *piecewise constant* function

$$\tilde{\eta}(x, D_n) = \sum_{i=1}^{N} I_{A_i} \tilde{\eta}_i \tag{4}$$

where I_{A_i} denotes the indicator of the set $\{x \in A_i\}$ and the constant $\tilde{\eta}_i = k_i/n_i$ is the ratio of k_i positive samples observed out of n_i samples thrown into A_i.

For each partitioning cell $A_i, i \in \{1, 2, ..., N\}$, let $\bar{\eta}(A_i) = \frac{1}{\int_{A_i} \mu(dx)} \int_{A_i} \eta(x)\mu(dx)$ denote the posterior probability associated to A_i, and the piecewise constant function $\eta^q(A)$ denote the so called *quantized posterior*:

$$\eta^q(A) = \sum_{i=1}^{N} I_{A_i} \bar{\eta}(A_i). \tag{5}$$

Note that for any given cell A_i, $\bar{\eta}(A_i)$ is a *fixed constant* independent of training samples D_n albeit the actual value of $\bar{\eta}(A_i)$ still unknown. On the other hand, $\tilde{\eta}_i = k_i/n_i$ is an estimation of $\bar{\eta}(A_i)$ depending on training samples thrown into the cell A_i. This estimate might be seriously biased in case that tree nodes are overfitted. See Figure 2 for a toy example of $\eta(x), \tilde{\eta}(x, D_n)$ and $\eta^q(A)$ respectively.

The introduction of $\bar{\eta}(A_i)$ allows us to decompose the upper bound of generalization errors in (3) into two error terms:

$$\begin{aligned} L\big(\tilde{\eta}(x, D_n)\big) &\leq L^* + 2\int_{\mathcal{R}^d} |\big(\eta(x) - \tilde{\eta}(x, D_n)\big)|\mu(dx) \\ &= L^* + 2\int_{\mathcal{R}^d} |\big(\eta(x) - \eta^q(A)\big) + \big(\eta^q(A) - \tilde{\eta}(x, D_n)\big)|\mu(dx) \\ &\leq \underbrace{L^* + 2\sum_{i=1}^{N} \int_{A_i} |\eta(x) - \bar{\eta}(A_i)|\mu(dx)}_{\text{quantized error } L(\eta^q(A))} + \underbrace{\big(2\sum_{i=1}^{N} |\bar{\eta}(A_i) - \tilde{\eta}_i| f(A_i)\big)}_{\text{sampling error } L_s} \end{aligned} \tag{6}$$

in which $f(A_i) = \int_{A_i} \mu(dx)$ is the probability mass function of data distributed over A_i and $\sum_{i=1}^{N} f(A_i) = 1$. For reasons to be clear shortly, these two error terms are denoted as "*quantized error*" and "*sampling error*" respectively, and following subsections illustrate how to estimate these two terms.

3.2.1. QUANTIZED BAYES ERROR

A quantized Bayes decision function is based on the *quantized posterior* $\eta^q(A)$

$$g_n\big(\eta^q(A)\big) = \begin{cases} 1 & \text{if } \bar{\eta}(A_i) > 1/2 \\ 0 & \text{otherwise.} \end{cases} \qquad i \in 1, ..., N. \tag{7}$$

Clearly its error probability $L\big(\eta^q(A)\big)$ satisfies the following inequality which is a special case of (3)

$$L\big(\eta^q(A)\big) \leq L^* + 2\sum_{i=1}^{N}\int_{A_i} |\eta(x) - \bar{\eta}(A_i)|\mu(dx) \overset{\text{def}}{=\!=} \overline{L}\big(\eta^q(A)\big) \tag{8}$$

The upper bound $\overline{L}\big(\eta^q(A)\big)$ is referred to as the *quantized Bayes error* throughout this paper since it plays a crucial role similar to the Bayes error in the analysis of the generalization error. Firstly, the *quantized Bayes error* is only determined by the posterior $\eta(x)$ and partitioning set A, and as shown by Proposition 5.1 in appendix, it monotonically decreases and eventually approaches the Bayes error L^* when the learning algorithms recursively partitions leaf nodes into smaller children nodes.

Secondly, as shown by Proposition 5.2 the *quantized Bayes error* itself is the minimal error bound that one can reach given a *fixed* set of partitioning nodes, no matter how node-wise posterior probabilities are estimated.

Thirdly, one can approximate the unknown *quantized Bayes error* $\overline{L}\big(\eta^q(A)\big)$ by the empirical loss of $g_n\big(\tilde{\eta}\big(x, D_n\big)\big)$

$$\overline{L}\big(\eta^q(A)\big) \approx \mathbf{E}\big\{g_n\big(\tilde{\eta}\big(X, D_n\big)\big) \neq Y | D_n\big\}. \tag{9}$$

As shown by Proposition 5.3, the approximation error incurred by (9) approaches zero when the number of training samples thrown into each cell goes to infinity. However the sampling errors ascribed to *small* sample size is often non-negligible in practice, and has to be taken into account in a data-dependent manner as follows.

3.2.2. INFLUENCE OF SMALL SAMPLES

Theorem 3.1 *The expectation of sampling error of decision trees* $\mathbf{E}[L_s] = \mathbf{E}\Big[\sum_{y=1}^{M}\sum_{i=1}^{N} |\bar{\eta}^y(A_i) - \tilde{\eta}_i^y| f(A_i)\Big]$ *is upper bounded by*

$$\mathbf{E}[L_s] \leq \sum_{y=1}^{M}\sum_{i=1}^{N} \sqrt{\Big(\mathrm{Var}(\tilde{\eta}_i^y) + \big(\mathrm{Bias}(\tilde{\eta}_i^y)\big)^2\Big)} f(A_i) \tag{10}$$

where $\mathrm{Var}(\tilde{\eta}_i^y)$ *and* $\mathrm{Bias}(\tilde{\eta}_i^y)$ *are the* ***variance*** *and* ***bias*** *of the estimator* $\tilde{\eta}_i^y$ *of the unknown* $\bar{\eta}^y(A_i)$ *for classes* $y = \{1, ..., M\}$.

Proof

$$\begin{aligned}
\mathbf{E}[L_s] &= \sum_{y=1}^{M}\sum_{i=1}^{N} \mathbf{E}\Big[|\bar{\eta}^y(A_i) - \tilde{\eta}_i^y|\Big] f(A_i) \\
&= \sum_{y=1}^{M}\sum_{i=1}^{N} \mathbf{E}\Big[\sqrt{\big(|\bar{\eta}^y(A_i) - \tilde{\eta}_i^y|\big)^2}\Big] f(A_i) \\
&\leq \sum_{y=1}^{M}\sum_{i=1}^{N} \sqrt{\mathbf{E}\Big[\big(|\bar{\eta}^y(A_i) - \tilde{\eta}_i^y|\big)^2\Big]} f(A_i) \qquad (11) \\
&= \sum_{y=1}^{M}\sum_{i=1}^{N} \sqrt{\Big(\mathrm{Var}(\tilde{\eta}_i^y) + \big(\mathrm{Bias}(\tilde{\eta}_i^y)\big)^2\Big)} f(A_i) \qquad (12)
\end{aligned}$$

where (11) follows Jensen's inequality $\mathbf{E}\big[\sqrt{x}\big] \leq \sqrt{\mathbf{E}[x]}$, *and (12) by the decomposition of Mean Squared Error* $\mathbf{E}\big[\big(\bar{\eta}^y(A_i) - \tilde{\eta}_i^y\big)^2\big]$.

For multiclass data with $y \in \{1, 2, ...M\}$ classes, the probability of observing k_i^y samples out of n_i samples in node A_i follows *multinomial* distribution with parameters $\big(n_i, \big[\bar{\eta}^1(A_i), ..., \bar{\eta}^M(A_i)\big]\big)$. For large n_i the ratio k_i^y/n_i converges to $\bar{\eta}^y(A_i)$ following the law of large numbers, thus the sampling error term $|\bar{\eta}^y(A_i) - \tilde{\eta}_i^y|$ essentially approaches zero.

When n_i is relatively small, following theorem 3.1, one can estimate the upper bound of the expectation of sampling error by working out the *variance* and *bias* of estimator $\tilde{\eta}_i^y$. For multiclass datasets we can take *Dirichlet* distribution as the conjugate prior probability for multinomial distribution of posterior probabilities $\bar{\eta}^y(A_i)$ and estimate the *mean*, *variance* and squared *bias* of estimator $\tilde{\eta}_i^y$ as follows (Murphy, 2006):

$$\begin{aligned}
&\hat{\eta}_i^y = E(\tilde{\eta}_i^y) = \frac{k_i^y + n_s}{n_i + M \cdot n_s}, \qquad \mathrm{Var}(\tilde{\eta}_i^y) = \frac{\hat{\eta}_i^y(1 - \hat{\eta}_i^y)}{1 + n_i}, \\
&\Big(\mathit{Bias}(\tilde{\eta}_i^y)\Big)^2 = \Big(\hat{\eta}_i^y - \frac{k_i^y}{n_i}\Big)^2
\end{aligned} \tag{13}$$

where $n_s \in \{0, 1/2, 1\}$ is the prior number of pseudo samples following Haldone, Jeffreys and Bayes prior probabilities respectively. In our work, it turns out Jeffreys prior leads to the best performance for all experiment tests.

3.2.3. ESTIMATION OF GENERALIZATION ERROR

Putting together (6), (9) and (10) we have the generalization error estimation as follows:

$$\boxed{\begin{aligned}
\tilde{L}\big(\tilde{\eta}(x, D_n)\big) \approx\ & \mathbf{E}\big\{g_n\big(\tilde{\eta}\big(X, D_n\big)\big) \neq Y | D_n\big\} \\
&+ \sum_{y=1}^{M}\sum_{i=1}^{N} \sqrt{\Big(\mathrm{Var}(\tilde{\eta}_i^y) + \big(\mathrm{Bias}(\tilde{\eta}_i^y)\big)^2\Big)} f(A_i)
\end{aligned}} \tag{14}$$

in which the empirical loss is directly measured using training samples D_n, while the variance $\mathrm{Var}(\tilde{\eta}_i^y)$ and squared

bias $\left(\text{Bias}(\tilde{\eta}_i^y)\right)^2$ terms are estimated using (13) for both binary class and multi-class datasets.

Remark 1: from (13) it is clear that the upper bound of sampling error L_s increases as the number samples n_i thrown in A_i decreases. When the generalization error estimate (14) is minimized in decision tree learning, the sampling error term therefore effectively penalizes complex decision tree models and prevents leaf node cells from becoming exceedingly small. Thus the overfitting is automatically avoided.

Remark 2: the sampling error term L_s is derived within an unified consistency analysis framework and its integration with empirical error is theoretically justified. Sampling errors can be directly estimated without resorting to model complexity analysis. We would argue that penalizing complex models in data-independent manners is merely a disguise of reducing sampling errors, although a close inspection of existing methods from this point of view remains to be done.

Remark 3: since the quantized Bayes error $\overline{L}\big(\eta^q(A)\big)$ is approximated by the training error and the sampling error L_s by its expectation $\mathbf{E}[L_s]$, so the generalization error estimate in (14) is not strictly bounded above. Nevertheless, as observed in our extensive experimental results this estimation appears to be consistently more robust and accurate than K-fold cross validation estimate.

4. Experimental results

4.1. Datasets and evaluation protocols

Table 4.1 below summarizes ten benchmark datasets from UCI Machine Learning Repository[1]. Each dataset is used to train a decision tree classifier using an optimised version of the CART algorithm (Pedregosa et al., 2011). Corresponding generalization errors are measured/estimated using four benchmarked approaches, namely, 1) the empirical measurement based on ground truth labels; 2) K-fold cross validation method; 3) the proposed generalization error estimation (denoted as "jeff"); 4) two pessimistic error estimation methods i.e. the Pessimistic Error Pruning (denoted as "quinlan") and the Minimum Error Pruning (denoted as "cestnik") that are reviewed in (Esposito et al., 1997). Errors by method 1) is then used as a "ground truth" to evaluate error estimates by methods 2), 3) and 4).

Note that a subset of data samples are used for the direct measurement of generalization errors. Therefore the whole set of data samples is randomly separated into *training* and *testing* subsets and the separation repeats multiple times to average out generalization errors. More specifically let separation ratio r equal $\frac{\#\text{training samples}}{\#\text{dataset samples}}$ and for each fixed ratio $r = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, a dataset is randomly separated 50 times and thus all together there are 250 randomly separations. Then for each separation a decision tree is constructed using *training* samples (see below explanation concerning cross validation methods). Finally for each constructed tree, generalization errors are measured (or estimated) at tree nodes with different depths ranging from 0 to maximal depth where depth 0 corresponds to root node.

[1] http://archive.ics.uci.edu/ml/

Datasets	#data samples	#attributes	#classes
diabet	768	8	2
german	1000	24	2
wine	6497	11	2
ecoli	336	7	8
imgseg	2310	19	7
letter	20000	16	26
satimg	6436	36	6
usps	9298	256	10
vehicle	964	18	4
vowel	990	10	11

Table 1. Summary of benchmark datasets

Indeed for K-fold cross validation the subset of training samples is randomly separated into *CV_training* and *CV_validation* subsets. Decision trees are constructed using *CV_training* samples and generalization errors are estimated using *CV_validation* samples. The process repeats K times to average out the generalization error where $K = \{2, 5, 10\}$ in our experiments.

4.2. Qualitative study of generalization error

Figures 3,4, 5 and 6 plot *generalization errors* against the depths of constructed trees for four example datasets *vehicle, diabet, wine, ecoli*. First of all, the empirical measurement of generalization error decreases when the tree depth increases from 0. However when tree depths are getting exceedingly large, the generalization error often flattens out or even increases due to large overfitting errors (e.g. see upper-left plot in Figure 3). On the one hand, this trend is well reflected in empirically measured generalization error as well as cross validation and the proposed method. Differences between the empirically measured and estimated generalization errors are relatively small in these Figures, and thus are scrutinized quantitatively in following subsection. On the other hand, two pessimistic error estimates tend to either under-estimate (for "quinlan") or over-estimate (for "cestnik") generalization errors. Moreover, the mean errors and standard deviations from the test errors become more pronounced for small-sized datasets e.g. *ecoli* and *diabet*.

Secondly, when more data samples are used for training, e.g. with larger ratio $r >= 0.5$, both empirically measured and estimated generalization errors are getting

smaller since leaf nodes are in general partitioned into smaller cells and deeper trees are constructed. This observation is in accordance with the consistency analysis of *quantized Bayes error* in Proposition 5.1, i.e. $\overline{L}\big(\eta^q(\mathbb{A})\big)$ approaches L^* when diameters of leaf nodes become infinitesimally small.

Thirdly, one pitfall of the cross validation method is that it often does not explore the whole range of depth parameter, since the number of samples in *CV_training* is smaller than those in the whole training subset thus more shallow trees have to be constructed for cross validation methods. See Figure 6 for an example of discussed observations.

4.3. Quantitative study of generalization error

We use Mean Squared Error (MSE) to quantify the differences between the measured error and those estimated ones. Figure 1 (left) shows that the proposed method *accurately* estimates the errors with averaged MSEs at 0.03, about 35% to 50% lower than that of cross validation methods. For the sake of parameter selection, Pearson Correlation turns out to be more informative than MSE since only the shape of error curve matters and adding an arbitrary constant to the error curve is irrelevant. Figure 1 (middle) shows that the proposed method also compares favourably with averaged PEs at about 0.93 while cross validation methods merely obtain Pearson Correlation at about 0.83.

Error bars in Figure 1 represent *standard deviations* of MSE and Pearson Correlation measured over 50 random separations of datasets. Standard deviations for the proposed method are about 40% and 30% lower in MSEs and PEs respectively. We view this statistically meaningful margin as a clear indication of the *robustness* of the proposed method in general. Figures 3, 4, 5, 6 illustrate detailed results measured on individual datasets.

4.4. Comparison of estimation time

Since evaluating equation (14) is the only computational cost incurred for the proposed error estimation, a substantial speed up is obtained as compared to cross validation methods which involve the costly construction of decision trees for each fold of training data. The comparison of estimation time in Figure 1 (right) showcases that the proposed method is at least **2 orders of magnitudes more efficient** than K-fold cross validation methods. The estimation time for the proposed method is actually negligible as compared to the training time for constructing decision trees. In contrast, the exceedingly long estimation time of K-fold cross validation methods makes it cumbersome to be used in practice for large datasets such as *letter* and *usps*.

5. Conclusion

We proposed a novel estimation of generalization error for decision tree learning methods. The estimation is based on the consistency analysis of two error terms i.e. the *quantized Bayes error* and *sampling error*, which essentially depends on the number of data samples thrown into each leaf node. As compared with the popular cross validation methods, the proposed generalization error estimation is *statistically more accurate, robust and substantially more efficient*.

While the proposed error estimation method is dedicated to decision tree type of classifiers, extensions to other histogram based classifiers such as *random forest* will be explored in future work.

Appendix

Proposition 5.1 *Let $diam(A_u) = \sup_{x,y\in A_u} |x-y|$ the diameter of an arbitrary set $A_u \subset \mathcal{R}^d$ and assume there is a sequence of sets $\mathbb{A}_j := \{A_{1j}, \ldots, A_{N_j}\}$ recursively constructed by a learning algorithm such that $diam(A_{ij}) \to 0$ as the number of sets in the sequence $N_j \to \infty$,*

then the quantized Bayes error $\overline{L}\big(\eta^q(\mathbb{A})\big)$ of the decision function in (7) approaches the Bayes error L^*.

Proof *Since $\eta(x)$ is absolute continuous, then for $x \in \bigcap_{i=1}^{N_j} A_{ij}$, $\bar{\eta}(A_{ij}) \to \eta(x)$ as $diam(A_{ij} \to 0)$, followed by $|\eta(x) - \eta^q(\mathbb{A}_j)| \to 0$ for every x and corresponding A_{ij}.*

Proposition 5.2 *Given a fixed set of partitioning cells $A = \{A_1, ..., A_N\}$, the quantized Bayes error $\overline{L}\big(\eta^q(A)\big)$ defined in (8) is minimal in the sense that $\overline{L}\big(\eta^q(A)\big) \leq \overline{L}\big(\hat{\eta}(A)\big)$ for any piecewise constant functions $\hat{\eta}(A) = \sum_{i=1}^{N} I_{A_i} c_i$ where $c_i \in \mathcal{R}$ are arbitrary constants and $\overline{L}(\hat{\eta}(A))$ the upper bound of generalization error of the decision function $g_n(\hat{\eta}(A)) = \begin{cases} 1 & \text{if } c_i > 1/2, \\ 0 & \text{otherwise.} \end{cases}$*

Proof *It suffices to show that for every cell A_i, $\mathbf{P}\{g_n\big(\hat{\eta}(A_i)\big) \neq Y\} \geq \mathbf{P}\{g_n\big(\bar{\eta}(A_i)\big) \neq Y\}$. By definition of $g_n\big(\bar{\eta}(A_i)\big)$ its error probability is $\min\big(\bar{\eta}(A_i), 1-\bar{\eta}(A_i)\big)$, and in case that $g_n\big(\hat{\eta}(A_i)\big) \neq g_n\big(\bar{\eta}(A_i)\big)$, the error probability for $g_n\big(\hat{\eta}(A_i)\big)$ is $1 - \min\big(\bar{\eta}(A_i), 1-\bar{\eta}(A_i)\big) = \max\big(\bar{\eta}(A_i), 1-\bar{\eta}(A_i)\big) \geq \min\big(\bar{\eta}(A_i), 1-\bar{\eta}(A_i)\big)$.*

Proposition 5.3 *Given a fixed set of partitioning cells $A = \{A_1, ..., A_N\}$, considering the error probability $L\big(\tilde{\eta}(x, D_n)\big) = \mathbf{P}\{g_n\big(\tilde{\eta}(x, D_n)\big) \neq Y\}$ of the plug-in function $g_n\big(\tilde{\eta}(x, D_n)\big)$ as defined in (6). If $n_i \to \infty$ for $i \in \{1, 2, ...N\}$ then the error bound $\overline{L}\big(\tilde{\eta}(x, D_n)\big) \to \overline{L}\big(\eta^q(A)\big)$ i.e. approaches the upper bound of $L\big(\tilde{\eta}(x, D_n)\big)$.*

Proof *It suffices to show that the "sampling error" term in (6) i.e. $|\bar{\eta}(A_i) - \tilde{\eta}_i|$ approaches zero as $n_i \to \infty$ for $i \in \{1, 2, ..., N\}$. Clearly this condition is fulfilled since $k_i/n_i \to \bar{\eta}(A_i)$ due to the law of large numbers.*

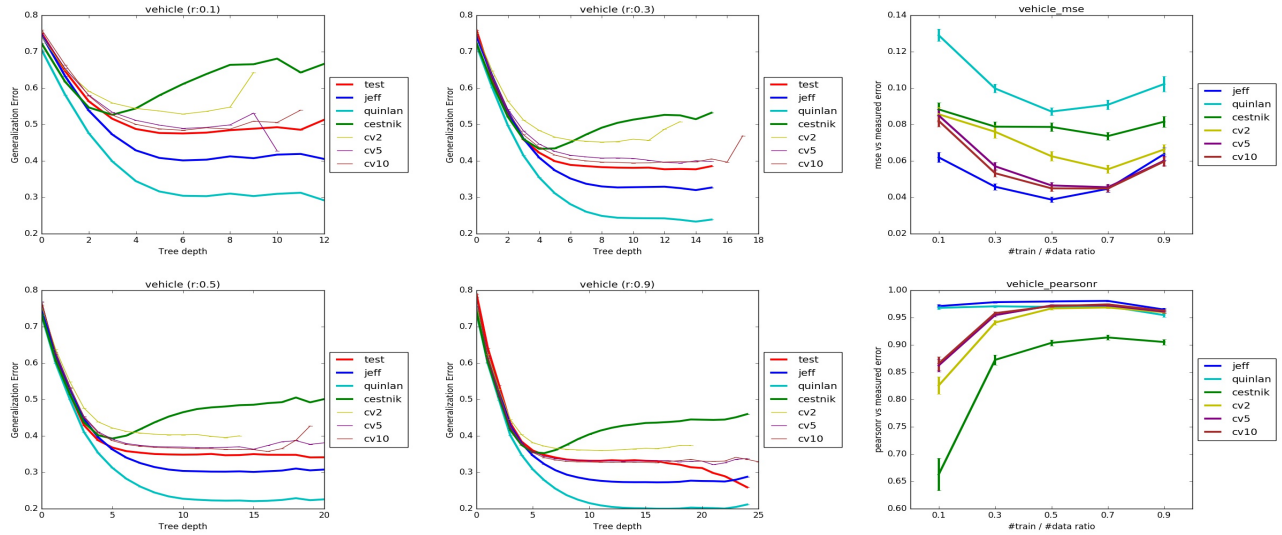

Figure 3. **Columns 1 and 2**: generalization errors for *vehicle* dataset.**Column 3**: Mean Squared Error (mse) and Pearson Correlation of error curves w.r.t. error curves measured with ground truth labels. Err bars represent *standard deviations* over 50 random runs.

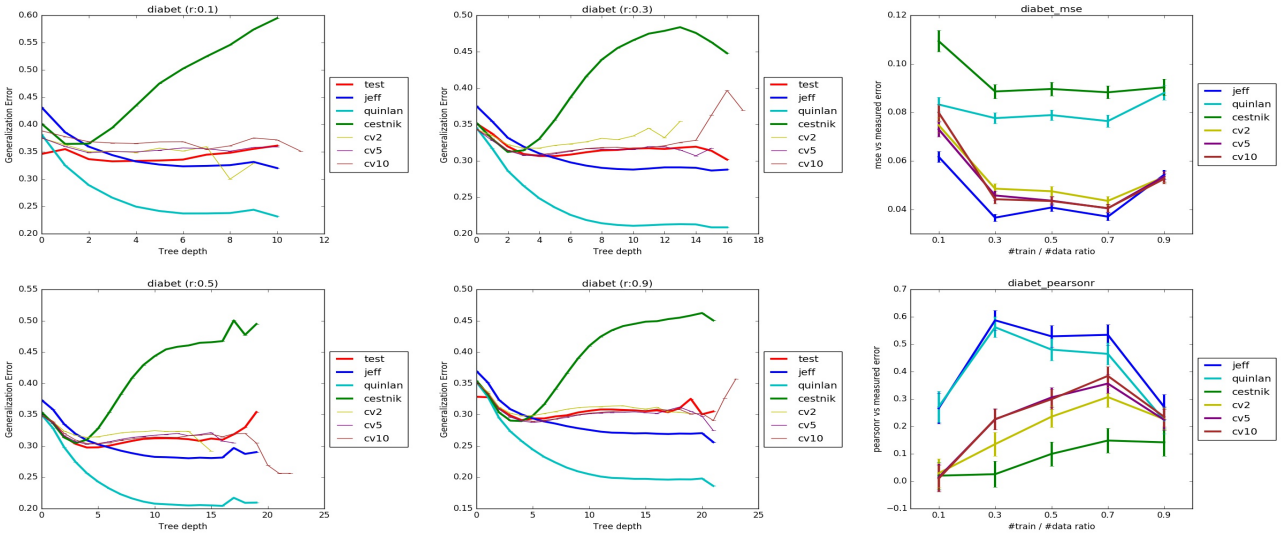

Figure 4. Generalization error of different methods for *diabet* dataset (see Figure 3 for elaborated explanations).

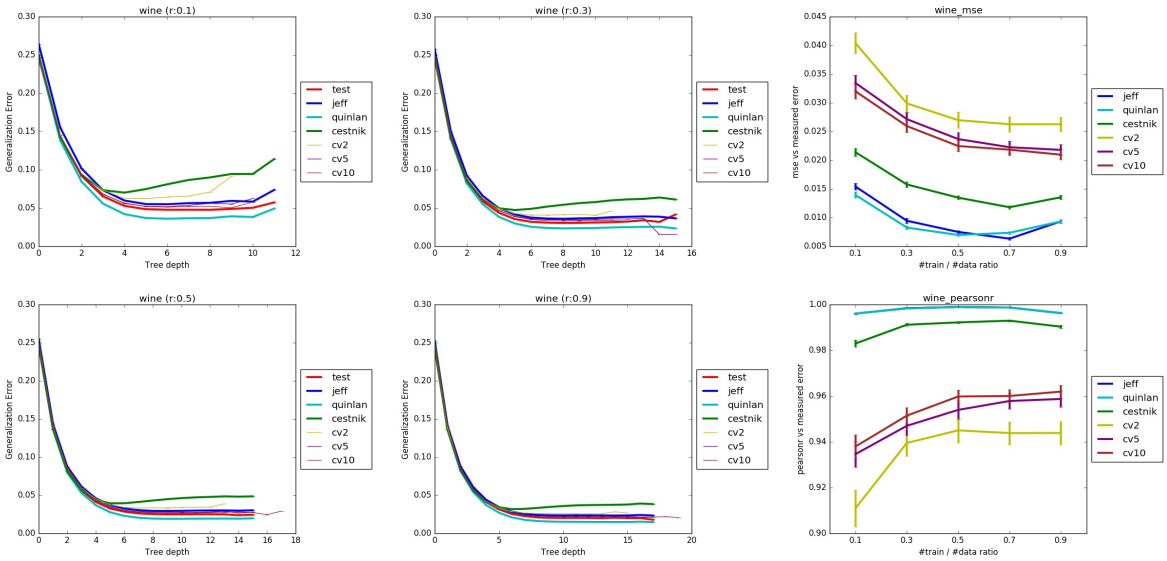

Figure 5. Generalization error of different methods for *wine* dataset (see Figure 3 for elaborated explanations).

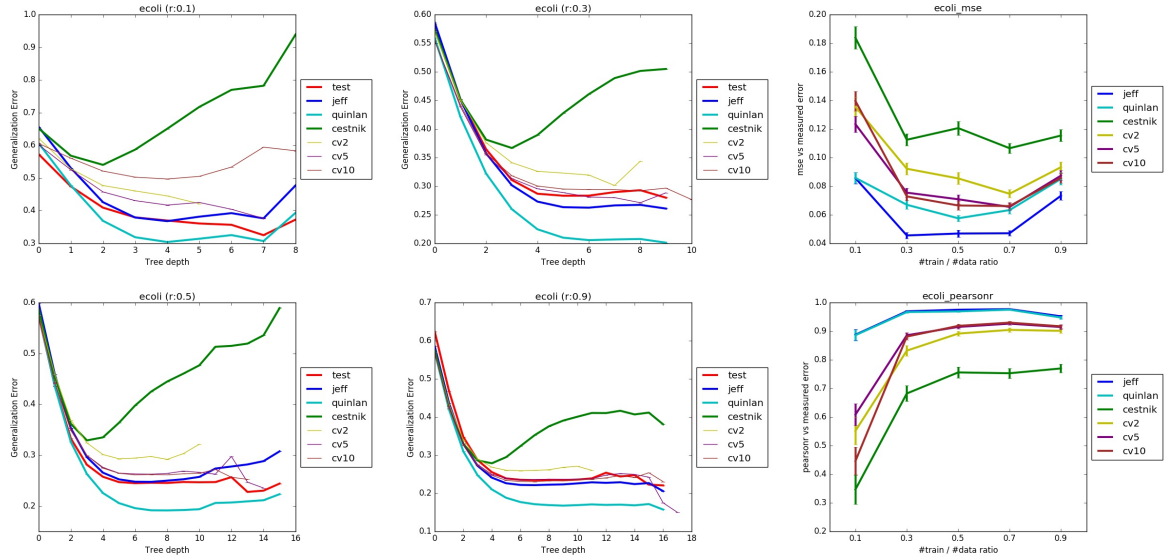

Figure 6. Generalization error of different methods for *ecoli* dataset (see Figure 3 for elaborated explanations).

References

Bartlett, Peter L. and Mendelson, Shahar. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(3):463–482, 2003.

Breiman, L. Random forests. *Machine Learning*, 45(1): 5–32, 2001.

Breiman, L. and Spector, P. Submodel selection and evaluation in regression: The x-random case. *International Statistical Review*, 60:291 – 319, 1992.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Wadsworth International Group, Belmon, CA, 1984.

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

Efron, B. and Tibshirani, R. *An introduction to the bootstrap*. Chapman and Hall, 1993.

Esposito, F., Malerba, D., and Semeraro, G. A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(5):476–491, May 1997.

Freund, Y. Self bounding learning algorithms. In *Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann, 1998.

Kääriäinen, M. and Elomaa, T. Rademacher penalization over decision tree prunings. In *Machine Learning: ECML 2003, 14th European Conference on Machine Learning*, pp. 193–204, Sept. 22-26 2003.

Kearns, M. and Mansour, Y. A fast bottom-up decision tree pruning algorithm with near-optimal generalization. In *In Proceedings of the 15th International Conference on Machine Learning*, pp. 269–277. Morgan Kaufmann, 1998.

Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence IJCAI*, pp. 1137 – 1143, 1995.

Langford, J. and Blum, A. Microchoice bounds and self bounding learning algorithms. In *Machine Learning*, pp. 209–214. Morgan Kaufmann, 2000.

Mansour, Y. and McAllester, D. Generalization bounds for decision trees. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pp. 69 – 74, 2000.

Murphy, K.P. Binomial and multinomial distributions. 2006.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Shah, M. Sample compression bounds for decision trees. In *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, pp. 799–806, 2007.

Stone, M. Asymptotics For and Against Cross-Validationn. *Biometrika*, 64:29–35, 1997.