Statistics for Data Science Project

In the following presentation we will discuss the "Accurate Robust and Efficient Error Estimation for Decision Trees (by Lixin Fan)" paper implementation by using R programming language.

By Matteo Rofrano and Francesco Lanci Lanci

ABOUT THE PAPER

Aims:

This paper try to solve the problem of estimating the generalization error when the conditions that allow the decision tree classifier to reach the Bayes error (which bounds below the generalization error) are not satisfied. The decision tree is not able to reach the Bayes error because of the fact that the number of data samples is limited and even for huge amout of data the small sample problem may still be present in certain leaf nodes when they belong to long branches.

Abstract:

Extensive experimental results show that the proposed error estimate is superior to the K-fold cross K-fold cross validation methods in terms of robustness and accuracy. Moreover it is orders of orders of magnitudes more efficient than cross validation methods.

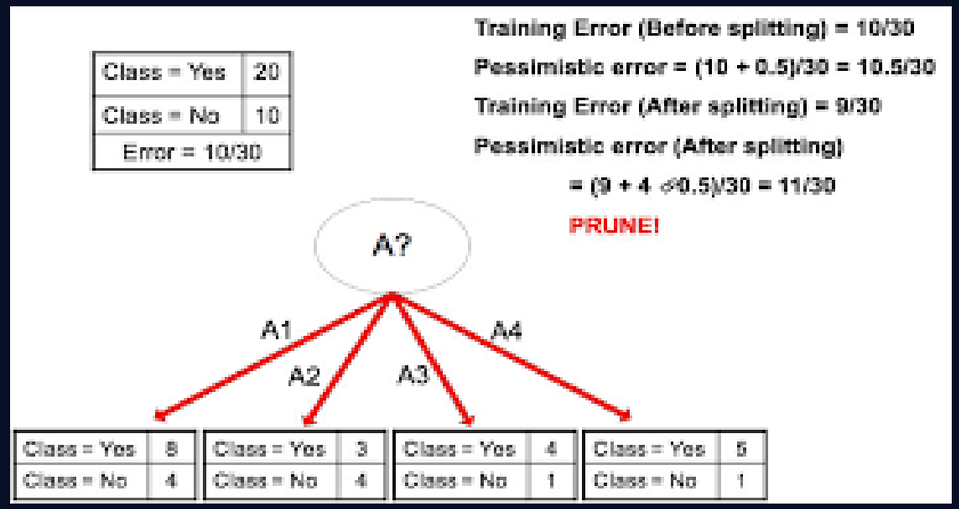
Related works and weekness of CV

Other methods in the literature have been used to estimate the generalization error for decision tree. Examples are the Pessimistic Error Pruning (PEP), Minimum Error Pruning (MEP) and the K-fold cross validation.

K-fold Cross validation

This method had demonstrated superior performances performances for various applications and in this paper this paper is used to estimate the generalization method by using the validation set. The main issues of issues of this method are:

1. The long estimation time due to the repeated repeated training and testing on validation sets. validation sets.
2. The high variance in error estimation due to the use of data.



PEP and MEP

This 2 methods are way more efficient in a computationally point of view with respect to the K-fold cross validation but they are inconsistent and they tend to under-estimate or over-estimate the generalization error when they are applied to small-sized dataset.

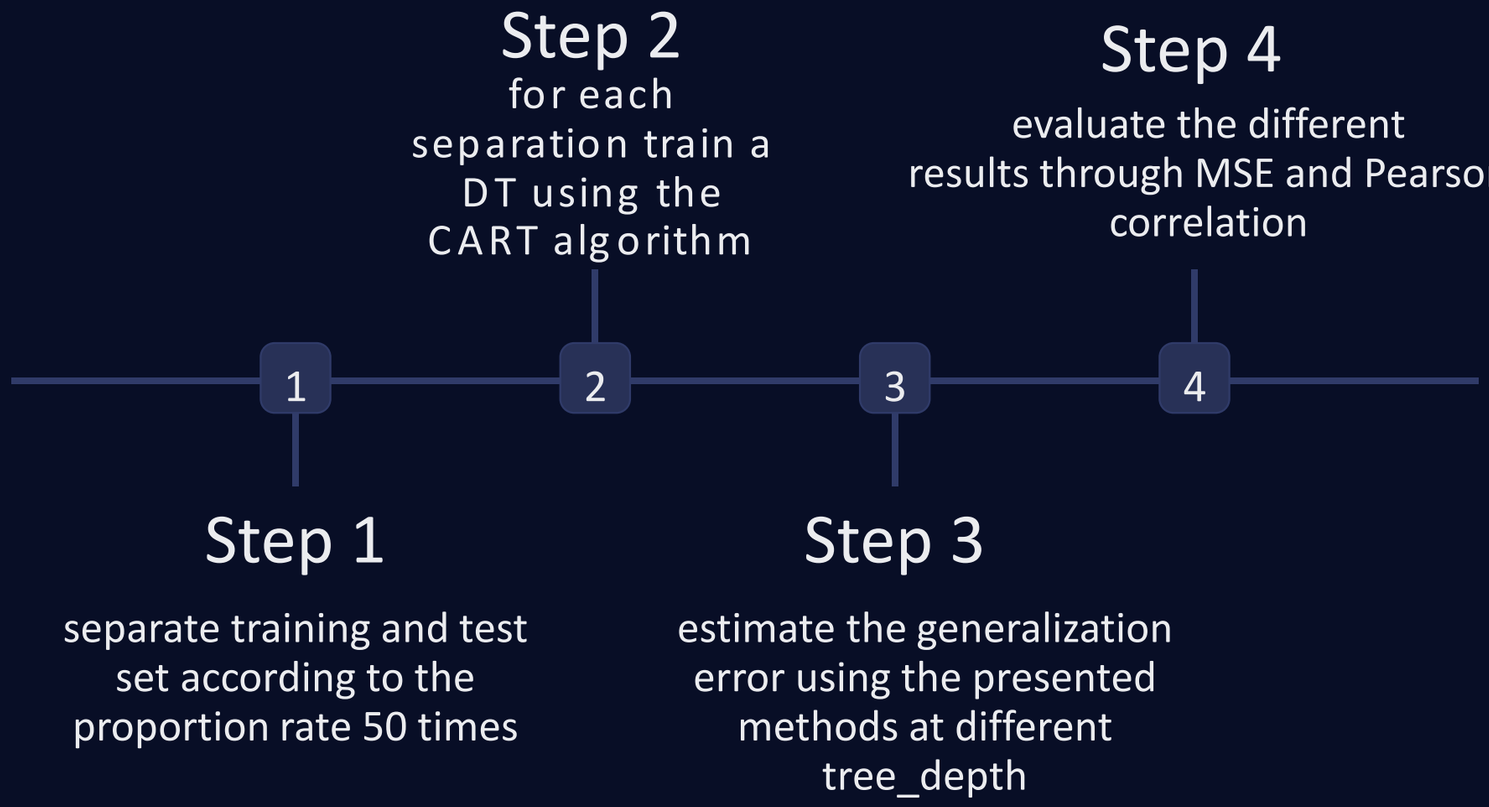
Theory behind the proposed method

First of all, for a plug in decision function, we let (X, Y) be a pair of random variables, with their values from R^d and {0, 1}. Through the true posterior probability function $\eta(x)$ we can calculate the Bayes decision function: $g^*(x) = \begin{cases} 1 & if\ \eta(x) > 1/2 \\ 0 & otherwise. \end{cases}$, but since the true posterior probability function is often unknown, we need to approximate it from the i.i.d. training samples D_n, so we obtain: $g_n(x) = \begin{cases} 1 & if\ \tilde{\eta}(x, D_n) > 1/2 \\ 0 & otherwise. \end{cases}$,with the probability of error: $L(g_n) = P\{g_n(X) \neq Y | D_n\}$ .Now, in the case of decision tree, we partition the space into N disjointed cells $\boldsymbol{A} = \{A_1, \dots, A_N\}$ corresponding to N leaf nodes. For multiclass data we have $y \in \{1,2, \dots, M\}$ classes, and for each terminal node A_i we have the probability of observing k_i^y samples out of n_l total samples in the node.

$$\tilde{L}(\tilde{\eta}(x, D_n)) \approx \mathbf{E}\{g_n(\tilde{\eta}(X, D_n)) \neq Y | D_n\} + \sum_{y=1}^{M} \sum_{i=1}^{N} \sqrt{\left(\mathrm{Var}(\tilde{\eta}_i^y) + \left(\mathrm{Bias}(\tilde{\eta}_i^y)\right)^2\right)} f(A_i)$$

Now we have calculated estimation of the generalization error, as we can see we have two terms called: the quantized error and the sampling error. The quantized error can be approximated by the empirical loss using the training sample, while the sampling error is calculated using the variance and the squared bias of the estimator $\tilde{\eta}_i^y$ of the unknown poster probabilites $\overline{\eta^y}(A_i)$, so we have: $Var(\tilde{\eta}_i^y) = \frac{\hat{\eta}_i^y(1-\hat{\eta}_i^y)}{1+n_l}$ and $(Bias(\tilde{\eta}_i^y))^2 = \left(\hat{\eta}_i^y - \frac{k_i^y}{n_i}\right)^2$ where $\hat{\eta}_i^y = E(\tilde{\eta}_i^y) = \frac{k_i^y + n_s}{n_i + M \cdot n_s}$ with the total number of classes M, and $n_s \in \{0, 1/2, 1\}$, but in the work of the paper it turns out that the second one leads to the best performance.

PAPER METHODOLOGY



The accuracy and robustness of the proposed estimation method has been proven by using different proportion of training data (0.1, 0.3, 0.5, 0.7, 0.9). For each of the proportion there have been made 50 separation in training and test set and a DT has been constructed. Then for each DT the generalization errors are measured at different tree nodes depths. For the K-fold cross validation has been used a k = 2,5 and 10 folds.

PAPER METHODOLOGY

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

In the paper are used the datasets presented in the table. We have focused on the diabet and letter dataset in order to have an analysis on a binary classification problem and on a multinomial classification problem

To evaluate the performances of the different error estimation methods are used:

- the MSE to quantify the differences between the measured error (on test set) and those estimated ones
- the Pearson correlation between the estimation over the 50 splits and the measured error.
- The running time needed for providing an estimation.

OUR IMPLEMENTATION

We have structured our implementation as follow:

1) We have implemented the proposed method in a function called “Gen_error”.

2) We have implemented a cross validation function useful to get the plot of the tree_depth.

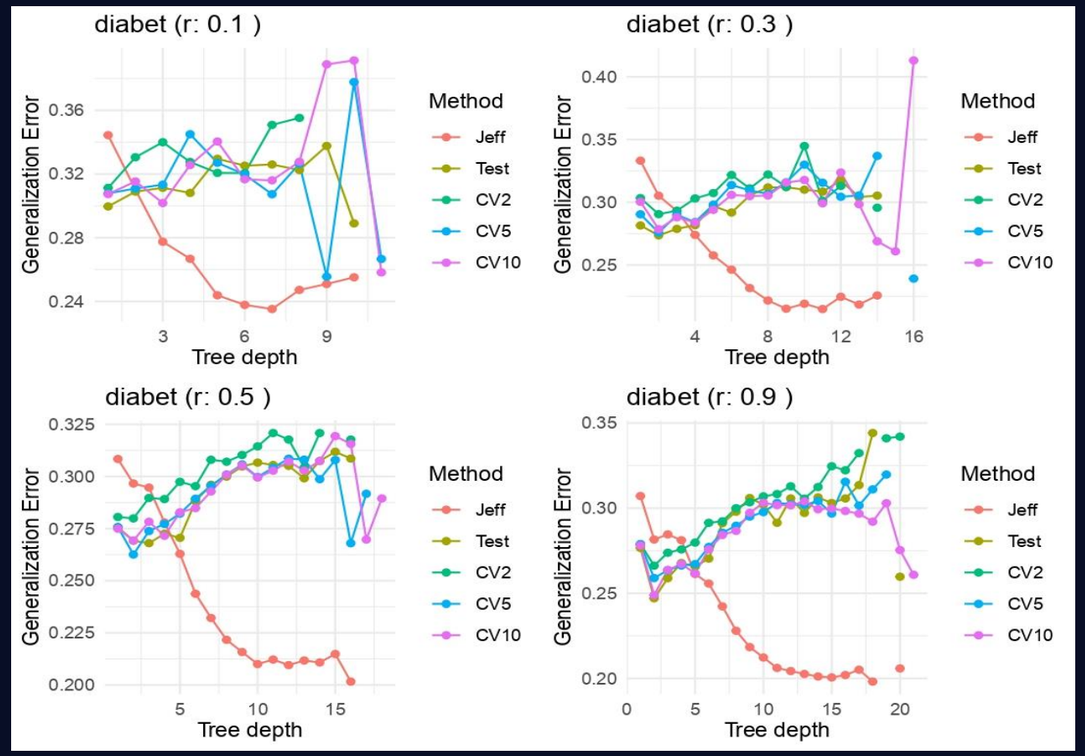
3) We have tested our implementation on 2 datasets: diabet (with a binary target variable) and letter with a (multinomial target variable) by using as estimation methods: the proposed method “jeff”, the CV with 2 folds, the CV with 5 folds and the CV with 10 folds.

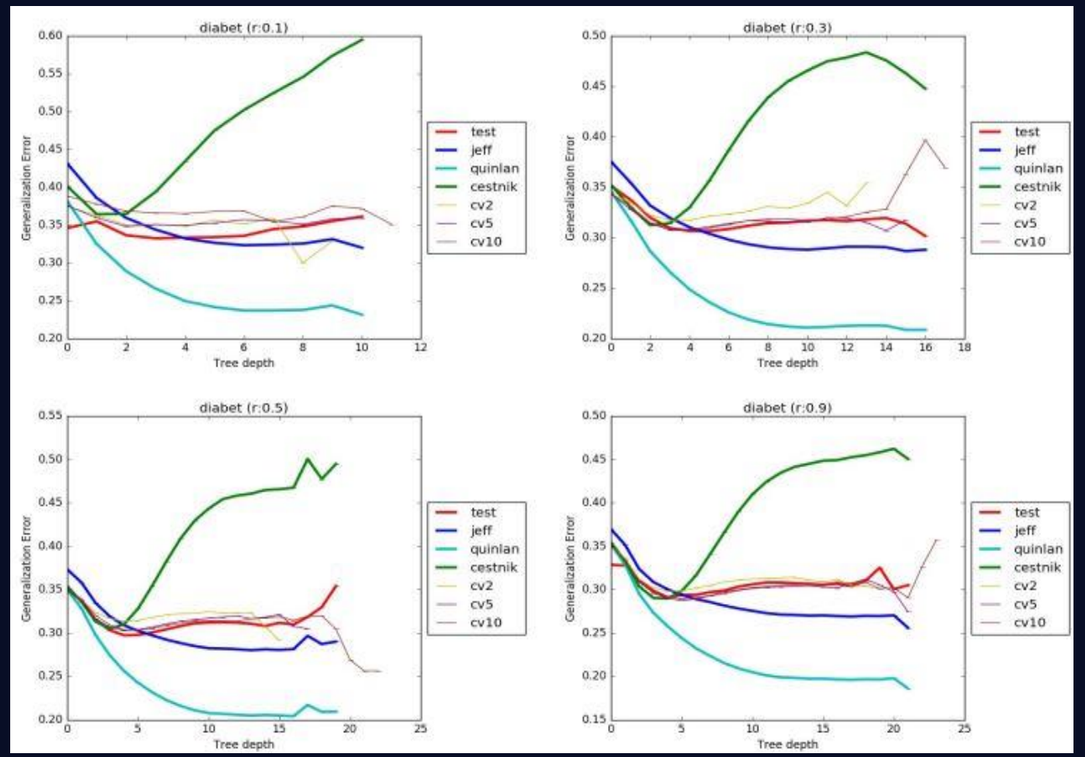
4) We have computed the generalization error estimations over different possible tree depths

5) We have computed the MSE and Pearson correlation to asses and compare the accuracy of the methods.

6) We have assessed the computational efficiency of the method by computing the relative running time of the estimation methods

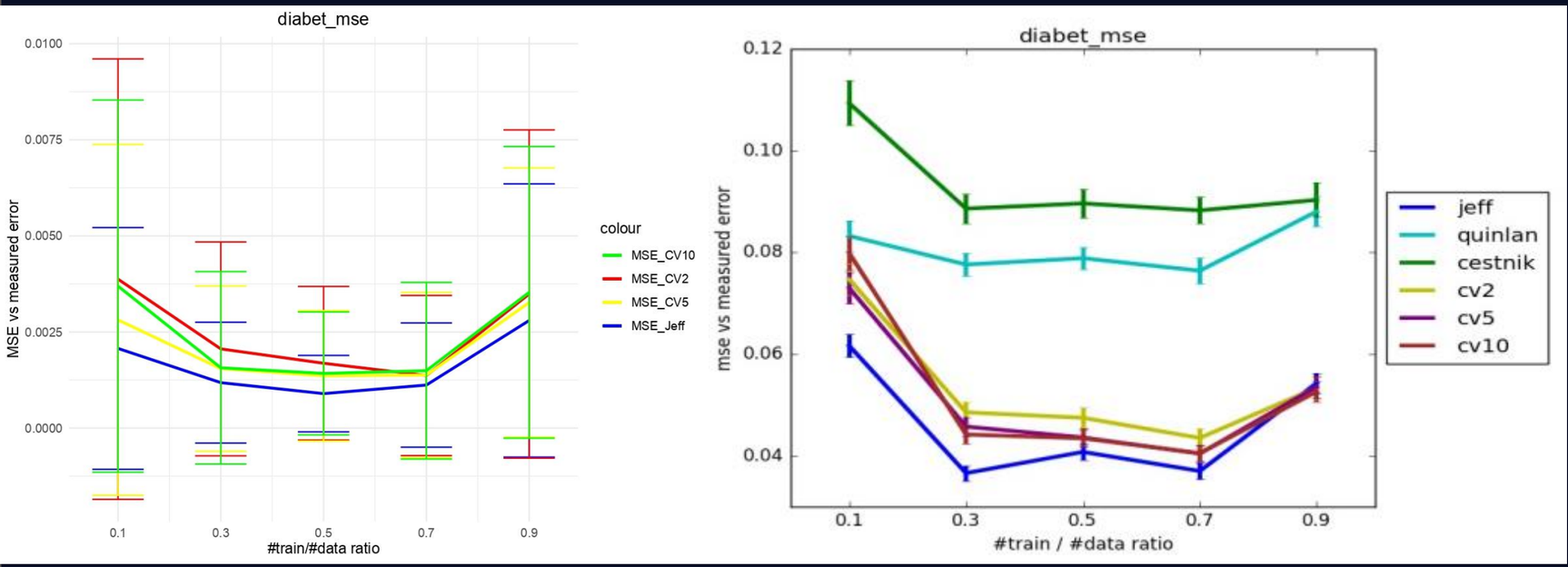
DIABET TREE_DEPTH



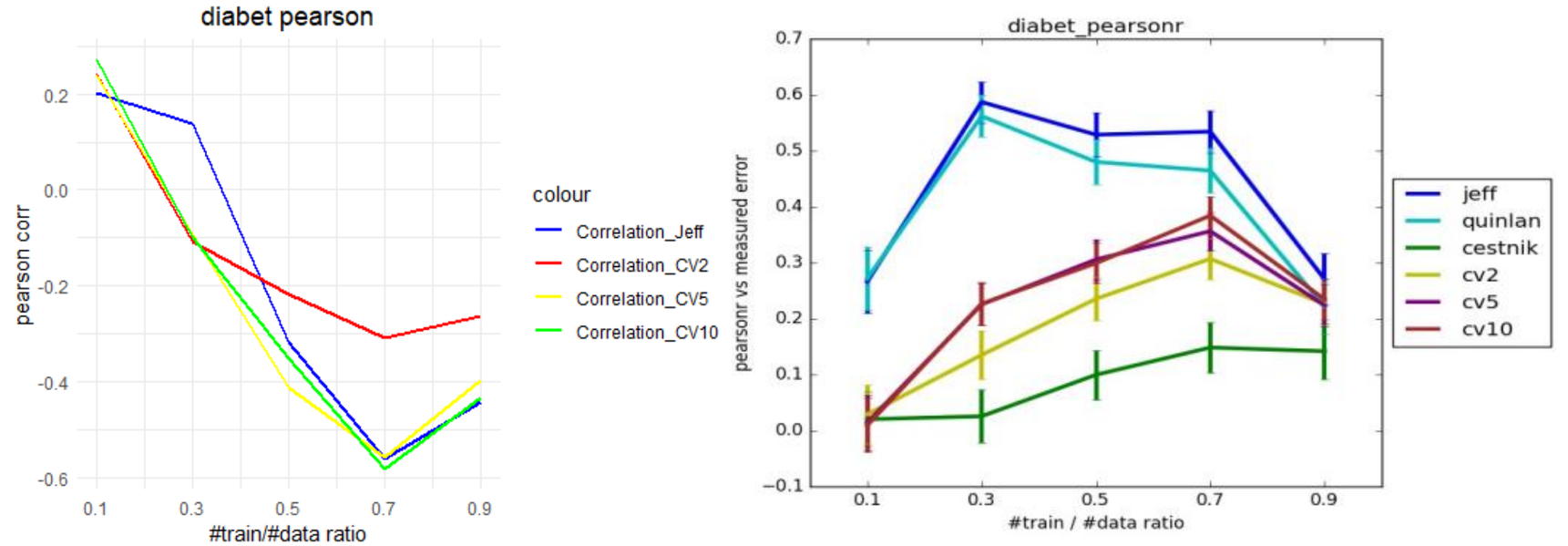


The results show the errors for each tree depths. As we can see, both on the paper (right) and on our implementation (left), the proposed generalization error starts higher than the others when the tree depth is low, and then drops lower than the other errors. In our case, the descent of the generalization error "Jeff" is more pronounced, probably because of the different parameters used in the construction of the trees. In addition, when depths are high, there is a marked variability of errors, especially for those from cross validation.

DIABET MSE AND PEARSON

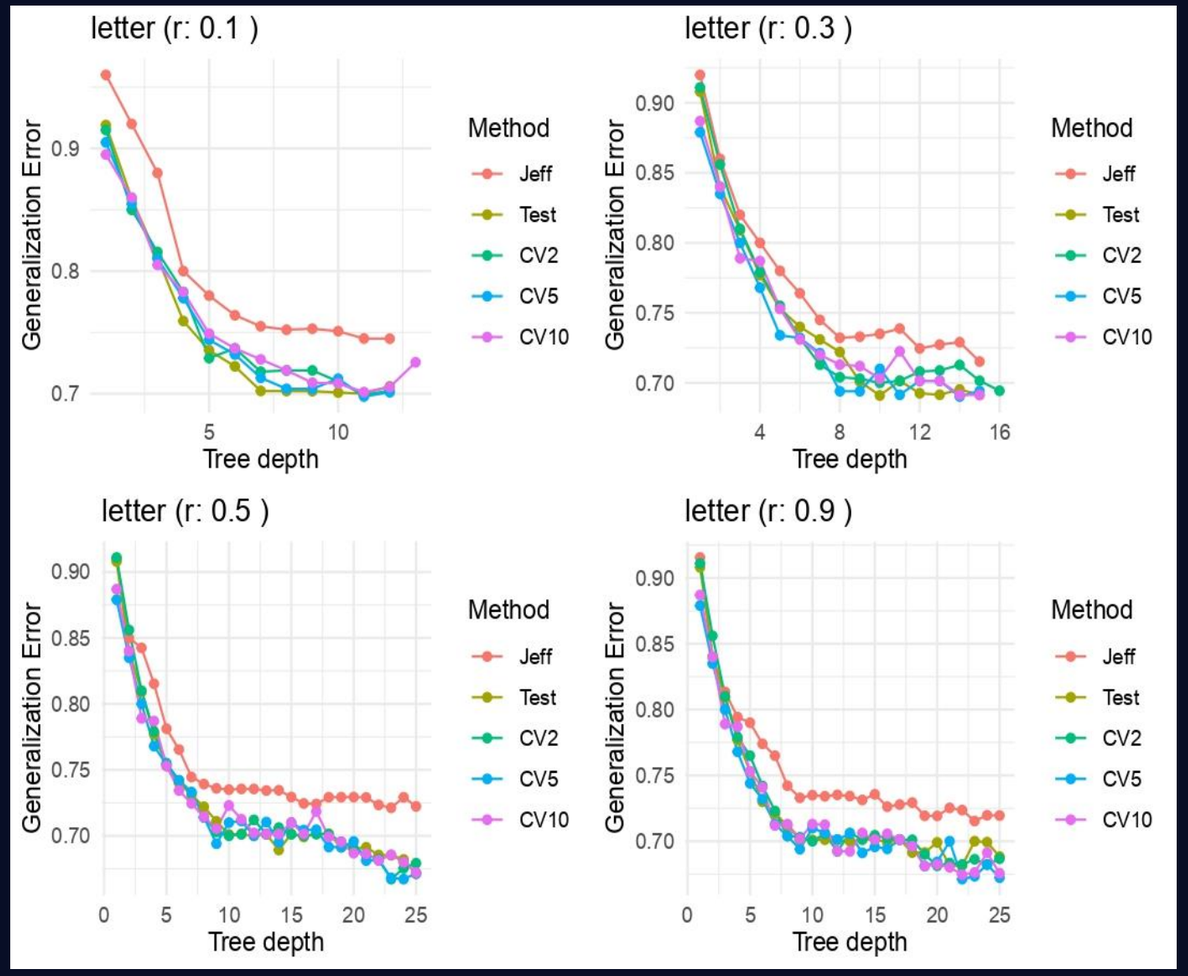


From the MSE plot we can see that the proposed method is able to provide an estimation of the generalization error that is closer to the estimation of the error obtained by using the test set. Moreover, we can see how the proposed method obtain the smallest standard deviation with respect the other methods and this leads to less variability in our results.



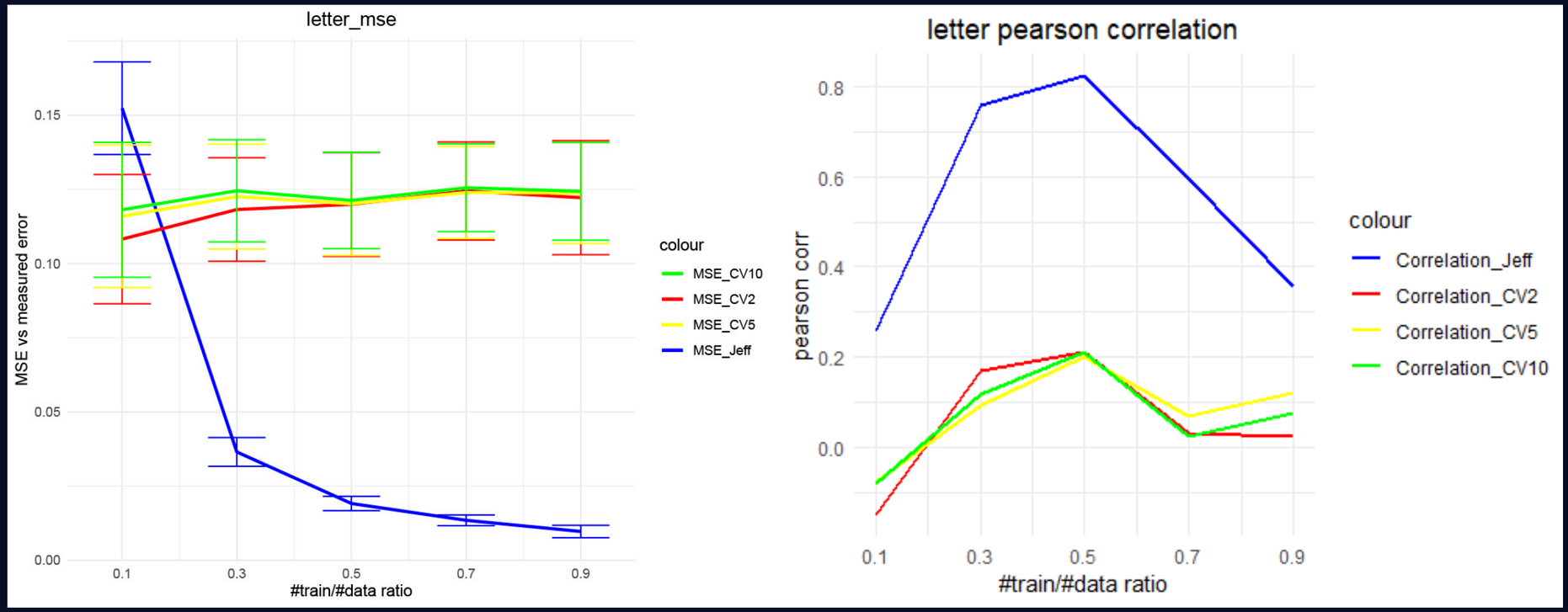
The plot of the Pearson correlation tend to be different from the one of the paper. This could be caused probably by the different decision tree paramenters used. In our experiments we have used the default parameters provided by the rpart library of R. However the proposed method still has higher correlation than CV5 and CV10.

LETTER TREE_DEPTH



Here the results show that for three out of four training set proportions, the generalization error proposed by the paper starts out fairly close to the other errors, and then, unlike the graph seen on the diabet dataset, moves away from the others errors by decreasing less. When the portion of the training set is 0.1, the generalization error "Jeff" is far away at both the beginning and the end.

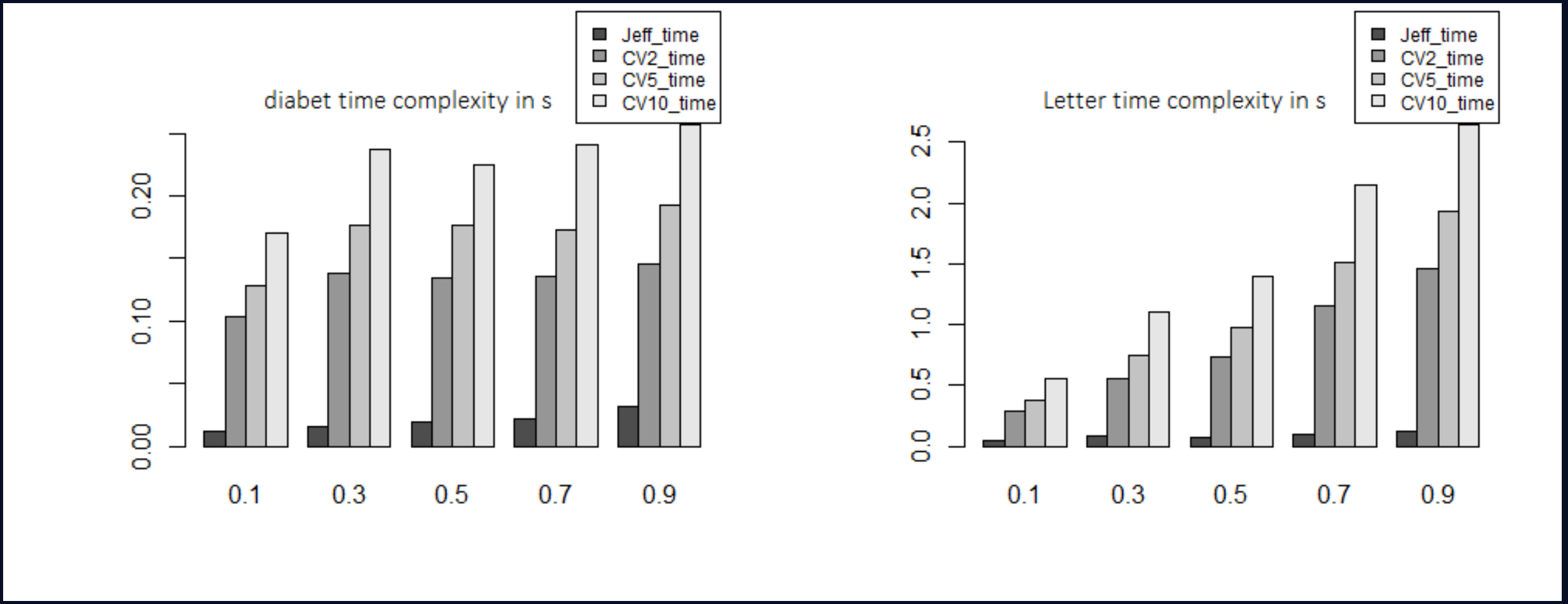
LETTER MSE AND PEARSON



The results of the proposed method on the letter dataset show more clearly its ability to provide more accurate estimation of the generalization error both in terms of MSE and Pearson correlation, in fact it gets the best performances for each proportion unless for the first proportion in the MSE plot.

From the plots we can see that the estimation method with the highest standard deviation is the cross-validation that uses 2 folds while the best one in terms of standard deviation is still the proposed method.

COMPUTATIONAL EFFICIENCY



From the plots of the time complexity of the different estimation methods of the generalization error used in the paper is clear that the proposed method is orders of magnitude more computationally efficient than all others method for each input size.

CONCLUSIONS

1 ACCURACY

From the MSE and Pearson plots we can assert that the implemented method tends to be more accurate on the estimation of the generalization error for both small-sized dataset (see diabet) and larger dataset (see letter).

2 EFFICIENCY

The running times needed for the different methods demonstrate that the implemented method is way more efficient with respect to the cross-validation method.

The 2 esperiments conducted on the Diabet and Letter dataset show that the most useful use of the proposed method is on large datasets since it is able to get more accurate estimation of the error than the cross-validation method with a run time significantly lower.