

Abstract

Decision-tree ensemble creation techniques are powerful concepts in predictive data mining. Using a unique statistical test called 5x2-fold cross validation, approaches to creating classifier ensembles can be assessed in terms of performance and accuracy. Through experiments with more than 50 publicly-available datasets, we provide an analysis of each technique through the 5x2 cross-validation experiments in comparison with the standard 10-fold cross validation statistical tests conducted in the past. Our results show that the 5x2-fold cross validation experiments reduce the elevated Type I error inherently present in the 10-fold cross validation method.

Research Findings

Data Set	Boosting 1000	Boosting 50	Random Subspaces	Random Trees B	Random Forests - Ign	Random Forests - 1	Random Forests - 2
zip	+/+	+/+	+/+	+/+	+/+	+/	+/+
letter	+/+	+/+	+/+	+/+	+/+	+/+	+/+
pendigits	+/+	+/	+/+	+/	+/+		+/
heart-c				+/	+/		+/
krk	+/+	+/	-/-		+/	-/	+/+
waveform				+/	+/		+/
page				+/			+/
sat	+/			+/			
sonar		+/		+/			
autos				+/			
led-24		-/	-/-				+/+
promoters	+/				+/+		
splice	/+	+/	+/			-/-	-/-
threenorm					+/+	/+	
vote				+/			
nursery	+/+	+/+	-/-	-/	-/	-/	-/
dna	+/		+/	+/		-/-	-/-
car	+/+	+/+	-/-	-/-	-/-	-/-	-/-
tic-tac-toe	+/	/+	-/-	-/	-/	-/	-/
breast-y		-/					
horse-colic				-/			
yeast				-/			
shuttle			-/-				
phoneme			-/-	-/			
krkp			-/-	-/	-/	-/	-/
sick			-/	-/	-/	-/-	-/-

Table 1. Shown as: 10 Fold CV/5x2-Fold CV.

Using more than 50 publicly available datasets (not all datasets are included above), this table shows the all the statistically significant results when evaluating decision tree classifier performance. Notice the number of statistically significant results for 10-fold cross validation and 5x2-fold cross validation. This method was originally proposed by Thomas Dieterich and relies on the idea that learning curves have the same relative ranking for algorithms with different training sizes.

Methodology

In this test, we perform five replications of two-fold cross validation. S_1 and S_2

In each replication, the available data are randomly partitioned into two equal sized sets

$$P^{(1)} = P_A^{(1)} - P_B^{(1)} \text{ and } P^{(2)} = P_A^{(2)} - P_B^{(2)}$$

Each learning algorithm (A or B) is trained on each set and tested on the other set. This produces four error estimates:

$$P_A^{(1)} \text{ and } P_B^{(1)} \text{ (trained on } S_1 \text{ and tested on } S_2)$$

$$P_A^{(2)} \text{ and } P_B^{(2)} \text{ (trained on } S_2 \text{ and tested on } S_1)$$

Subtracting the corresponding error estimates gives us two estimated error differences:

$$S^2 = (P^{(1)} - p)^2 + (P^{(2)} - p)^2 \text{ where } p = (P^{(1)} + P^{(2)})/2$$

Let S be the variance computed from the i th replication, and let $P1(1)$ and $P(1)$ from the first of five replications. Then define the following statistic:

$$\bar{t} = \frac{P_1^{(1)}}{\sqrt{\frac{1}{5} \sum_{i=1}^5 S_i^2}}$$

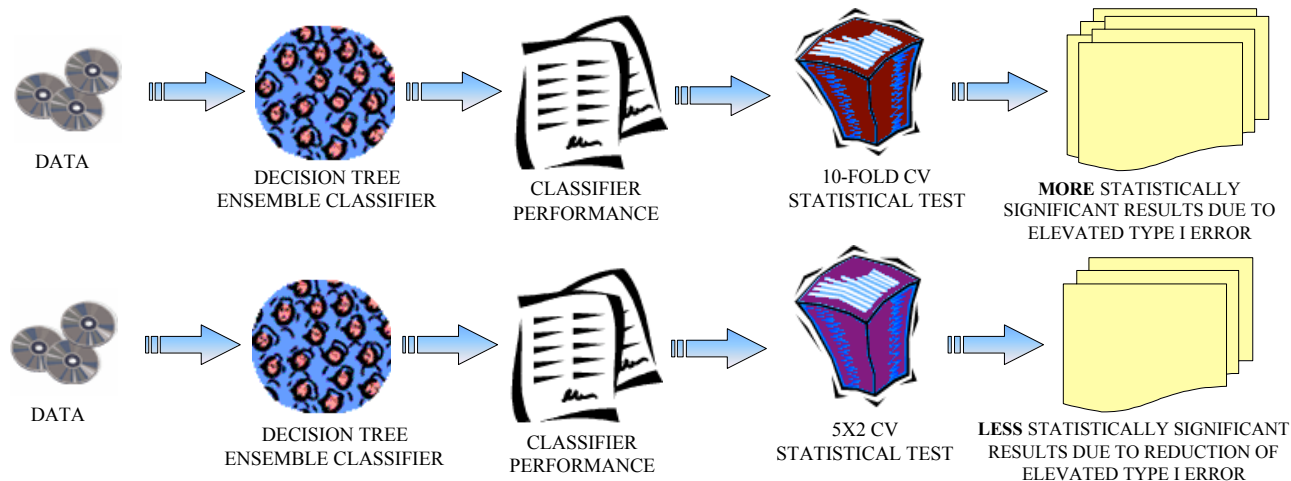
which we will call the 5x2-fold cross validation statistic.

Decision Tree Ensemble Creation Techniques

- ❖ **Bagging.** Bagging creates each ensemble of decision trees through bootstrap aggregation or random sampling with replacement from the set of training data.
- ❖ **Boosting.** Also known as AdaBoost.M1, this ensemble creation technique creates classifiers using a training set with weights assigned to every example.
- ❖ **Random Subspaces.** Proposed by Ho, this ensemble creation technique selects random subsets of the available features to be used in training the individual classifiers in an ensemble.
- ❖ **Random Trees.** For this ensemble creation technique, at each node in the decision tree the twenty best tests are determined and one of them is randomly selected for use at that node.
- ❖ **Random Forests (Three Variations).** Used with bagging, this ensemble creation technique blends elements of random spaces with bagging in a way that is specific to using decision trees as the base classifier.

Conclusion

Our research has shown that the 5x2-fold cross validation methodology results in less statistically-significant wins and losses when evaluating decision tree creation ensemble techniques, a consequence of reducing the elevated Type I error inherent in the 10-fold cross validation methodology.



Acknowledgment

Many thanks to Robert Banfield and Larry Shoemaker

