Tests of Hypothesis About the Accuracy of Decision Trees

Bill Qualls
MS • MBA • MSEd • PMP

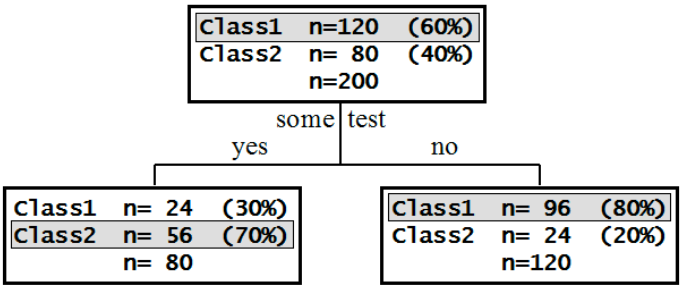
This paper will demonstrate the process with which we can perform tests of hypothesis about decision trees; specifically, test H_0: decision tree does not improve classification accuracy vs. H_1: decision tree does improve classification accuracy.

Assume we have the following classifications in our data:

```
Class1  n=120  (60%)
Class2  n= 80  (40%)
```

If we know nothing else about the data, then our best guess for classifying the data is to label everything as Class1. Indeed, we would expect to be correct about 60% of the time.

We create decision trees because we believe that we can do better than that. Assume we have the following decision tree:



The classification matrix is as follows:

		Predicted		
		Class1	Class2	Total
Observed	Class1	96	24	120
	Class2	24	56	80
	Total	120	80	200

The accuracy of the decision tree is (96+56)/200 = .76, or 76%. It would appear that we can make better classifications with the decision tree (76%) than we can without (60%). But is the difference statistically significant? We can perform a formal test of hypothesis.

Step 1: State the hypothesis.

Let p = the proportion of cases classified correctly. Then...
H_0: p = .60 (without tree) vs. H_1: p > .60 (with tree)

Step 2: Determine the confidence level to be used.

We will use the standard $\alpha = .05$.

Step 3: Specify the test statistic to be used.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

Step 4: Determine the rejection region.

We will reject H_0 if $z > 1.645$.

Step 5: Calculate the observed value of the test statistic.

$$z = \frac{.76 - .60}{\sqrt{\frac{(.60)(.40)}{200}}} = \frac{.16}{.0346} = 4.62$$

Step 6: Conclusion.

Since the observed value of the test statistic (4.62) is greater than the critical value (1.645), we reject the null hypothesis and conclude that the decision table provides greater accuracy in classification than can be obtained without it.

Caveat.

The reader is reminded that statistical significance does not guarantee clinical significance!

Exercise.

Does the following decision tree provide greater accuracy in classification than can be obtained without it? Perform a formal test of hypothesis.