Machine Learning: Homework 4

Laurent Hayez

October 24, 2016

Exercise 1. *By default, WEKA takes the class from the last column in the arff file and uses all other columns as attributes. Take the zoo dataset and only use 'hair', 'airborne', and 'type' as attributes to predict the class 'eggs' with a C4.5 tree (J48). Visualize the tree.*

Solution. Using a J48 pruned tree, WEKA outputs the following tree:

```
J48 pruned tree
------------------

hair = false: true (58.0/4.0)
hair = true
|   type = mammal: false (39.0/1.0)
|   type = bird: false (0.0)
|   type = reptile: false (0.0)
|   type = fish: false (0.0)
|   type = amphibian: false (0.0)
|   type = insect: true (4.0)
|   type = invertebrate: false (0.0)

Number of Leaves  :  8

Size of the tree :  10
```

□

Exercise 2. *The C4.5 algorithm implemented in WEKA allows us to choose between a pruned and an unpruned tree. Build both trees for the Bank-account dataset (bank.arff),* *visualize the trees, and compare the results. What is the purpose of using pruning? Which are the two pruning strategies that exist? Explain them briefly with an example.*

Solution. The unpruned tree is

```
J48 unpruned tree
------------------

children = YES
|   income <= 30099.3
|   |   car = YES
|   |   |   sex = MALE
|   |   |   |   region = INNER_CITY: NO (12.0/1.0)
|   |   |   |   region = RURAL
|   |   |   |   |   married = YES
|   |   |   |   |   |   age <= 27: NO (2.0)
|   |   |   |   |   |   age > 27: YES (3.0)
|   |   |   |   |   married = NO: NO (2.0)
|   |   |   |   region = TOWN: NO (8.0/2.0)
|   |   |   |   region = SUBURBAN: YES (1.0)
|   |   |   sex = FEMALE
|   |   |   |   region = INNER_CITY
|   |   |   |   |   married = YES: NO (6.0/2.0)
|   |   |   |   |   married = NO
|   |   |   |   |   |   income <= 23475.6: YES (2.0)
|   |   |   |   |   |   income > 23475.6: NO (2.0)
|   |   |   |   region = RURAL: NO (2.0)
|   |   |   |   region = TOWN
|   |   |   |   |   married = YES: NO (3.0/1.0)
|   |   |   |   |   married = NO: YES (3.0/1.0)
|   |   |   |   region = SUBURBAN: NO (4.0/1.0)
|   |   car = NO
|   |   |   married = YES
|   |   |   |   region = INNER_CITY
|   |   |   |   |   income <= 13106.6: NO (3.0)
|   |   |   |   |   income > 13106.6
|   |   |   |   |   |   mortgage = YES: YES (7.0/1.0)
|   |   |   |   |   |   mortgage = NO
|   |   |   |   |   |   |   income <= 18923: YES (3.0)
|   |   |   |   |   |   |   income > 18923: NO (4.0/1.0)
|   |   |   |   region = RURAL: YES (6.0/2.0)
|   |   |   |   region = TOWN
|   |   |   |   |   mortgage = YES: NO (5.0/2.0)
|   |   |   |   |   mortgage = NO
|   |   |   |   |   |   sex = MALE: YES (2.0/1.0)
|   |   |   |   |   |   sex = FEMALE
|   |   |   |   |   |   |   age <= 33: NO (3.0)
|   |   |   |   |   |   |   age > 33
|   |   |   |   |   |   |   |   age <= 50: YES (2.0)
|   |   |   |   |   |   |   |   age > 50: NO (2.0)
|   |   |   |   region = SUBURBAN: NO (3.0/1.0)
|   |   |   married = NO
|   |   |   |   mortgage = YES: NO (8.0/1.0)
|   |   |   |   mortgage = NO
|   |   |   |   |   region = INNER_CITY: NO (5.0/1.0)
|   |   |   |   |   region = RURAL: YES (1.0)
|   |   |   |   |   region = TOWN: NO (4.0/1.0)
|   |   |   |   |   region = SUBURBAN: YES (4.0/2.0)
|   income > 30099.3
|   |   sex = MALE
|   |   |   married = YES
|   |   |   |   region = INNER_CITY
|   |   |   |   |   car = YES
|   |   |   |   |   |   income <= 42603.9
|   |   |   |   |   |   |   age <= 37: YES (2.0)
|   |   |   |   |   |   |   age > 37: NO (2.0)
|   |   |   |   |   |   income > 42603.9: YES (3.0)
|   |   |   |   |   car = NO: YES (5.0/1.0)
|   |   |   |   region = RURAL: YES (3.0/1.0)
|   |   |   |   region = TOWN: YES (4.0/1.0)
|   |   |   |   region = SUBURBAN: YES (2.0)
|   |   |   married = NO: YES (12.0/1.0)
|   |   sex = FEMALE: YES (26.0/1.0)
```

```
children = NO
|   married = YES
|   |   mortgage = YES
|   |   |   region = INNER_CITY
|   |   |   |   income <= 39547.8: YES (12.0/3.0)
|   |   |   |   income > 39547.8: NO (4.0)
|   |   |   region = RURAL: NO (3.0/1.0)
|   |   |   region = TOWN: NO (9.0/2.0)
|   |   |   region = SUBURBAN: NO (4.0/1.0)
|   |   mortgage = NO
|   |   |   car = YES: NO (27.0/2.0)
|   |   |   car = NO
|   |   |   |   income <= 28421.7
|   |   |   |   |   region = INNER_CITY: NO (15.0/4.0)
|   |   |   |   |   region = RURAL: YES (1.0)
|   |   |   |   |   region = TOWN: YES (2.0/1.0)
|   |   |   |   |   region = SUBURBAN: YES (1.0)
|   |   |   |   income > 28421.7: NO (11.0)
|   married = NO
|   |   mortgage = YES
|   |   |   age <= 39
|   |   |   |   age <= 28: NO (4.0)
|   |   |   |   age > 28: YES (5.0/1.0)
|   |   |   age > 39: NO (11.0)
|   |   mortgage = NO: YES (20.0/1.0)

Number of Leaves  :     53

Size of the tree :      91
```

and the pruned tree is

```
J48 pruned tree
------------------

children = YES
|   income <= 30099.3
|   |   car = YES: NO (50.0/15.0)
|   |   car = NO
|   |   |   married = YES
|   |   |   |   income <= 13106.6: NO (9.0/2.0)
|   |   |   |   income > 13106.6
|   |   |   |   |   mortgage = YES: YES (12.0/3.0)
|   |   |   |   |   mortgage = NO
|   |   |   |   |   |   income <= 18923: YES (9.0/3.0)
|   |   |   |   |   |   income > 18923: NO (10.0/3.0)
|   |   |   married = NO: NO (22.0/6.0)
|   income > 30099.3: YES (59.0/7.0)
children = NO
|   married = YES
|   |   mortgage = YES
|   |   |   region = INNER_CITY
|   |   |   |   income <= 39547.8: YES (12.0/3.0)
|   |   |   |   income > 39547.8: NO (4.0)
|   |   |   region = RURAL: NO (3.0/1.0)
|   |   |   region = TOWN: NO (9.0/2.0)
|   |   |   region = SUBURBAN: NO (4.0/1.0)
|   |   mortgage = NO: NO (57.0/9.0)
|   married = NO
|   |   mortgage = YES
|   |   |   age <= 39
|   |   |   |   age <= 28: NO (4.0)
|   |   |   |   age > 28: YES (5.0/1.0)
|   |   |   age > 39: NO (11.0)
|   |   mortgage = NO: YES (20.0/1.0)

Number of Leaves  :     17

Size of the tree :      31
```

We see that the unpruned tree is quite huge while the pruned tree is more human readable. There are more errors in the pruned tree, but as some data can be noisy (ex. someone entered income=1000000 instead of income=100000), the pruning can sometimes remove these kind of errors. Thus it is classified as an error by the algorithm, but in fact it can prevent further errors.

The two pruning strategies are

Post-pruning: Suppose we have a decision tree that returns true or false. Suppose we built the whole decision tree and one branch leads to 10 instances that return true, and one instance that returns false. Then we can prune this branch by saying it always lead to true (with one error). For a "real" example, take the first part of the unprunned tree for the file `bank.arff`. For the category sex=MALE, we almost exclusively have NO (we have 24 NO with 3 errors and 4 YES). That means the next attributes are not very relevant for the classification. Hence we can prune this branch and say if sex=MALE, return NO leading to 28 NO with 7 errors. The same happens with sex=FEMALE, hence we can also prune this branch, and simply put if car=YES, return NO. Note that this is the result of the pruned tree with the J48 algorithm.

Pre-pruning: The pre-pruning is based on a statistical signifiance test (e.g. chi-squared test). Instead of cutting the branches of a tree at the end, they are cut during the building of the tree. One branch stops growing when there is no statistically significant association between an attribute and the class at a particular node, i.e., if an attribute is not relevant for the choice of the output.

For example, suppose we are creating the tree for the `bank.arff` file. We wonder if we can prune the branch car=YES. Suppose we computed that the next attribute for the decision is "sex" (with information gain for example). We wish to see it this attribute is significant. Then we can use the chi-squared test for example. We start by computing the following table:

	class=YES	class=NO	
sex=MALE	7	21	28
sex=FEMALE	8	14	22
	15	35	50

Then we compute $\mu_{j,CLASS} = \frac{n_{CLASS} \cdot n_j}{n}$ for $j = 1, 2$ and $CLASS = YES,\ NO$:

$$\mu_{1,YES} = 8.4 \quad \mu_{2,YES} = 6.6 \quad \mu_{1,NO} = 19.6 \quad \mu_{2,NO} = 15.4.$$

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_{i,YES} - \mu_{i,YES})^2}{\mu_{i,YES}} + \frac{(n_{i,NO} - \mu_{i,NO})^2}{\mu_{i,NO}} = 0.7575.$$

Table 1: Results of different classifiers on different datasets

	Simple rule	Naïve Bayes	Decision tree (J48 pruned)
Bank	68%	63%	81%
Iris	96%	96%	98%
Soybean	40.85%	93.7%	96.34%
Vote	95.6%	90.34%	97.24%
Zoo	100%	100%	99.01%

For this example, we have a degree of freedom of 1, and setting $\alpha = 0.95$ we have a limit value of 3.84. As $0.7575 < 3.84$, we conclude that the attribute “sex” is not significant and that we can cut this branch. □

Exercise 3. *Take the five datasets and use the three classifiers Naïve Bayes, Simple Rule, and C4.5. For each dataset-classifier tuple, retrieve the percentage of instances correctly classified. What are your conclusions (without doing any statistical tests)?*

Solution. The results of the classifiers are in Table 1. The first observation is that on these examples, the J48 pruned tree algorithm gives the best results, except for the zoo dataset. This is due to the fact that the animals are uniquely identified by their names, which can be seen as a subset of the category they are classified in (e.g. wolfs $\subset$ mammals) and the 1rule uses this information for classification. The decision tree is less good on this example because it is pruned.

Otherwise there is only the soybean dataset on which the simple rule is not good. Otherwise it is at least as good as Naïve Bayes, promoting the Ockham’s razor principle.

In conclusion, if one class is clearly identifiable with one attribute, the simple rule is a good choice, but if the classification is based on multiple attributes, a decision tree will be efficient, and a bit more precise than Naïve Bayes. □