Comparing of classifier algorithms on two data sets

Jiani Li

1. Introduction

1.1. C4.5 decision tree

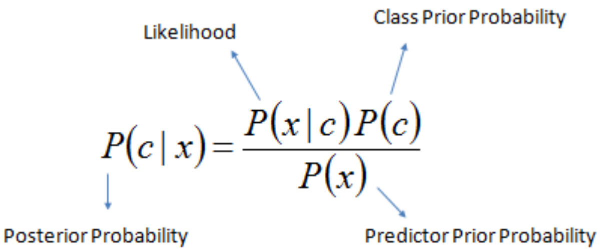
A decision tree is a graph using a branching method to illustrate all the possible outcomes of a decision. It takes as input an object or situation, and outputs a class. Each node of tree is a test of the value of one of the properties. Branches from a node correspond to possible values for a test.

1.2. K-NN classifer

k-NN classifer finds the classes of the k-nearest neighbors (based on some distance metric) and finds the class in majority and assigns that class to the test pattern. It is a kind of lazy algorithm.

1.3. Naïve bayes classifer

Naïve bayes classifier is based on Bayes rule of conditional probability with independence assumptions between predictors.


Likelihood
Class Prior Probability
$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability

$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

- $P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
- $P(c)$ is the prior probability of *class*.
- $P(x|c)$ is the likelihood which is the probability of *predictor* given *class*.
- $P(x)$ is the prior probability of *predictor*.

2. Experiments and Results

2.1. Experimental procedure

1) Preprocess – choose the data set (iris.arff and car.arff)
2) classify – choose classfier: trees -> J48 (C4.5 decision tree), lazy -> Ibk (k-NN), bayes -> NaiveBayes (Naïve bayes)
3) classify – test options: Using training set / Cross-validation Folds 10, start.

2.2. Results

2.2.1. Fisher's classic Iris study

For k-NN, the correctly and incorrectly classfied instances for k=3,5,7 is given in table 1.

Table 1. Performance of k-NN classifier with different k for iris study

Test Options	Training Set			Cross-Validation		
k	k = 3	k = 5	k = 7	k = 3	k = 5	k = 7
Correctly Classified Instances	145 / 96.6667%	144 / 96%	145 / 96.6667%	143 / 95.3333%	143 / 95.3333%	145 / 96.6667%
Incorrectly Classified Instances	5 / 3.3333%	6 / 4%	5 / 3.3333%	7 / 4.6667%	7 / 4.6667%	5 / 3.3333%

k = 7 generates the best results.

Explanation: When k is too small, there may be the problem of overfitting. The boundary data may be classified wrongly to other classifications. And outliers can also disturb classification. In the table, both k = 3 and k = 5 have the problem of overfitting (has a better performance in training set but worse performance in cross-validation test). When k is large, the influence by outliers is reduced. But when k is too large, there may be the problem of underfitting. Set k to be odd numbers greater than 7, the results are not greater than when k = 7. So that k = 7 generates the best results.

Table 2. Performance of different classfiers for iris study

Test Options	Training Set			Cross-Validation		
Classifier Performance	C4.5 Decision Tree	Naïve Bayes Classifier	k-NN Classifier	C4.5 Decision Tree	Naïve Bayes Classifier	k-NN Classifier
Correctly Classified Instances	147 / 98%	144 / 96%	145 / 96.6667 %	144 / 96%	144 / 96%	145 / 96.6667 %
Incorrectly Classified Instances	3 / 2%	6 / 4%	5 / 3.3333 %	6 / 4%	6 / 4%	5 / 3.3333 %
Kappa statistic	0.97	0.94	0.95	0.94	0.94	0.95
Mean absolute error	0.0233	0.0324	0.0337	0.035	0.0342	0.0387
Root mean squared error	0.108	0.1495	0.1165	0.1586	0.155	0.1282
Relative absolute error	5.2482 %	7.2883 %	7.585 %	7.8705 %	7.6997 %	8.7166 %
Root relative squared error	22.9089 %	31.7089 %	24.7235 %	33.6353 %	32.8794 %	27.1942 %

2.2.2. Car Evaluation

Table 3. Performance of different classfiers for car evaluation

Test Options	Training Set		Cross-Validation	
Classifier Performance	C4.5 Decision Tree	Naïve Bayes Classifier	C4.5 Decision Tree	Naïve Bayes Classifier
Correctly Classified Instances	1664 / 96.2963 %	1505 / 87.0949%	1596 / 92.3611 %	1478 / 85.5324 %
Incorrectly Classified Instances	64 / 3.7037 %	223 / 12.9051%	132 / 7.6389 %	250 / 14.4676 %
Kappa statistic	0.9198	0.7065	0.8343	0.6665
Mean absolute error	0.0248	0.1112	0.0421	0.1137
Root mean squared error	0.1114	0.2218	0.1718	0.2262
Relative absolute error	10.8411 %	48.5842%	18.3833 %	49.6626 %
Root relative squared error	32.9501 %	65.5935%	50.8176 %	66.9048 %

3. Discussion and Conclusion

3.1. Fisher's classic Iris study

Table 2 illustrates the performance of three classifiers for iris study. **The accuracy on training set is ranked as: C4.5 decision tree > k-NN > Naïve Bayes, while the accuracy on cross-validation is ranked as: k-NN > Naïve Bayes ≈ C4.5 decision tree.**

C4.5 decision tree has the greatest discrepancy between training set and cross validation test set accuracy (2% accuracy decreased). That may be because of overfitting in decision tree. When setting k to be small in k-NN classfier, there will also be the problem of overfitting. But here since we use k = 7, the overfitting problem for k-NN is greatly reduced.

As for Naïve bayes, it assumes independence between attributes, but for the iris study, the attributes are not rigorous independent. That might be the reason why the performance of Naïve bayes classifier is not better than the other two classifiers in this problem.

3.2. Car Evaluation

Table 3 illustrates the performance of three classifiers for car evaluation. **The accuracy on training set is ranked as: C4.5 decision tree > Naïve Bayes, the accuracy on cross-validation is still ranked as: C4.5 decision tree > Naïve Bayes.** For this much larger problem, there is also the problem of overfitting for both of the two classifiers.

C4.5 decision tree still has the greatest discrepancy between training set and cross validation test set accuracy (3.94% accuracy decreased, whereas only 1.57% accuracy decreased for Naïve bayes). It is still may be because of the problem of overfitting, and the overfitting problem is much severer for decision tree than for naïve bayes. We see in the decision tree, the tree spilts into too many branches, but splitting a lot leads to a complex tree and raises the probability of overfitting.

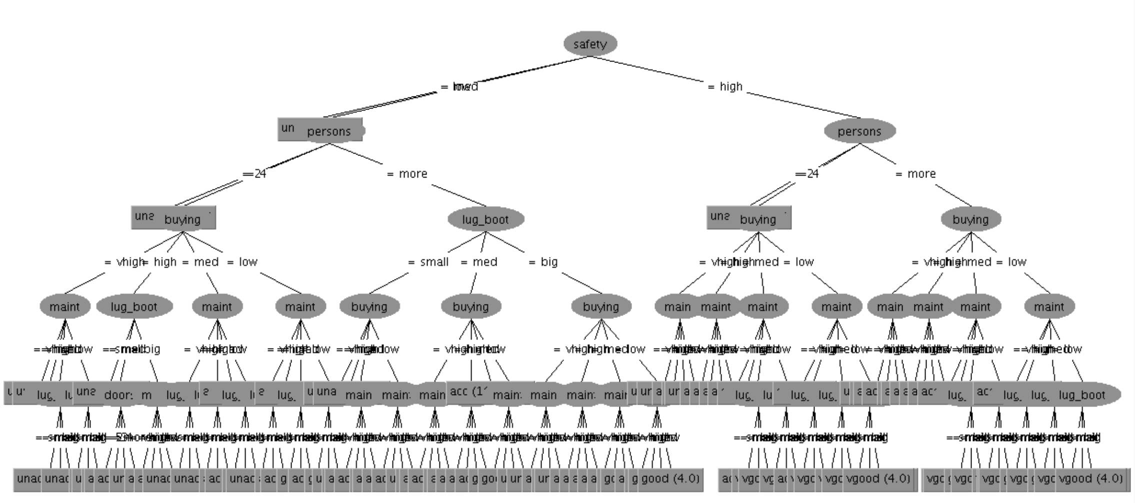


Figure 1. Decision tree for car evaluation

However, the overall accuracy of C4.5 decision tree is much greater than that of Naïve bayes. This may be still due to Naïve bayes' s strong assumption of independence between attributes, whereas the attributes of cars are dependent between each other, i.e., more doors imply more persons.