Kevin Maher

Dr. Siripun Sanguansintukul

MSDS680

March 2, 2017

Analyzing the UCI Abalone Data Set with Two Classifiers

The assignment is to analyze and classify abalone in the UCI Abalone Data Set using two different classifiers (Nash, et. al. 1994). We are to apply additional measurements of performance in addition to accuracy.

The first challenge was how to group the abalone by age for classification. I chose to use a Random Forest model plus the accuracy metric and kappa statistic for this purpose. Experimentation produced a reasonable model with 400 trees. My goal here was to pick dividing points between young, adult and old abalone that the model could detect best. In essence, I choose to use a Random Forest model to distinguish between young, adult and old abalone. The best unweighted kappa statistic was with a cutoff of less than 7 years for young abalone and greater than 13 years old for old abalone (see figure 1). This model does not produce the highest accuracy because the model tends to predict adult abalone well but not young or old abalone. Because kappa is a measure of how often the model is correct by more than mere chance, I chose it as the best measure for this purpose. Otherwise one could simply build a model that optimized predicting adult abalone but did poorly at finding young or old abalone.

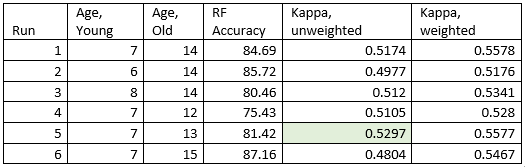


Figure 1. Finding Reasonable Cutoff Points for Age Groupings.

The Random Forest model was tuned using the “train” method in the R caret package (Langer, 2016). This provided training data as well as a cross validation score for the model (see figure 2). The cross validation gave somewhat lower accuracies than had been observed above for the same age cutoffs. The training algorithm also recommended that the “mtry” parameter be set to 2, but this actually reduced accuracy and kappa in my case. Accuracy was reduced from 81.42% from run 5 of figure 1 to 80.86% and unweighted kappa from 0.5297 to 0.5121. The original model with “mtry” set to 5 was retained as the benchmark Random Forest model.

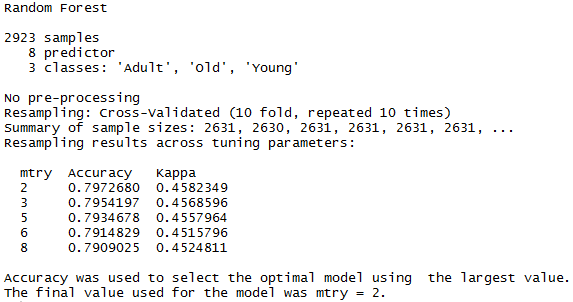


Figure 2. Tuning the Random Forest Model.

Several accuracy statistics describe the Random Forest model. There is the overall accuracy of 81.42% reported and the kappa statistic of 0.5297. The kappa statistic indicates that this is only a moderately good model (Lantz, 2015). A confusion matrix (see figure 3) shows that the model does a good job of identifying adult abalone. However, the model does only a fair job of predicting young abalone and is only a little better than a coin toss at detecting old abalone. This is confirmed by the sensitivity statistics (see figure 3). The sensitivity statistic is the also called the true positive rate, while over 91% for adult abalone, it is only around 53% for old abalone and 75% for young abalone. The specificity statistic gives the expected mirror of these results. This statistic is the true negative rate, and is high for old and young abalone but low for adult abalone (Lantz, 2015). Again, this confirms that the model does not do well except for adult abalone. The sensitivity statistic shows that the model is especially weak at predicting old abalone.

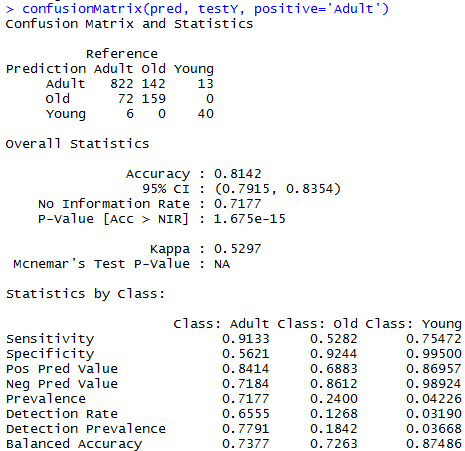


Figure 3. Confusion Matrix and Statistics for Random Forest Model.

The second classifier attempted was a Support Vector Machine (SVM). Because of the high number of tuning options for the SVM the R function “tune.svm” was used in an effort to obtain a reasonably optimal SVM. Both radial and polynomial kernels were tried, over a range of cost and gamma values. The polynomial was also tried for 2nd, 3rd and 4th degree. For the radial kernel the tuning algorithm recommended a cost of 20 and a gamma of 0.125. This produced an accuracy of 81.18% and an unweighted kappa of 0.5004 (see figure 5). The radial kernel SVM did not provide a better model by either metric than the Random Forest attempted earlier. The polynomial kernel performed worse with 80.46% accuracy and an unweighted kappa of 0.4593.

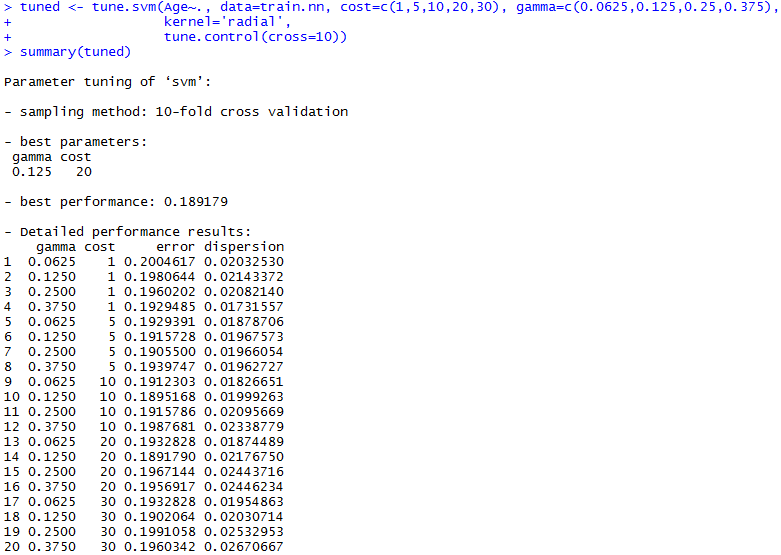


Figure 4. Tuning an SVM with a Radial Kernel.

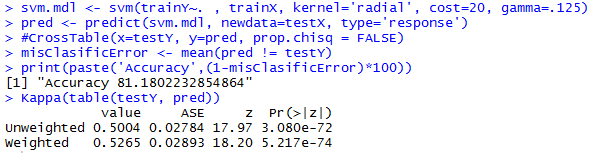


Figure 5. SVM Accuracy – Radial Kernel

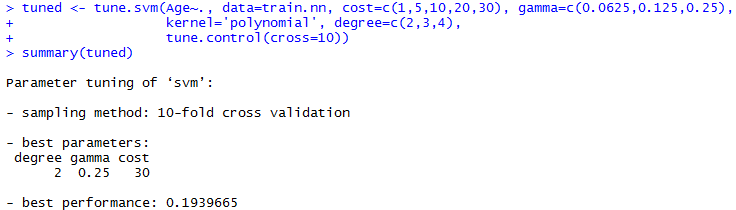


Figure 6. Tuning the SVM with a Polynomial Kernel.

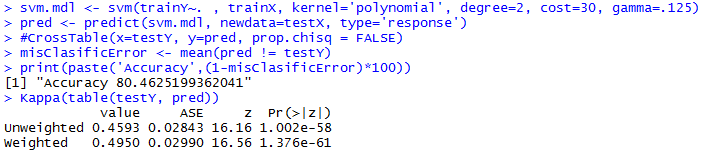


Figure 7. SVM Accuracy – Polynomial Kernel

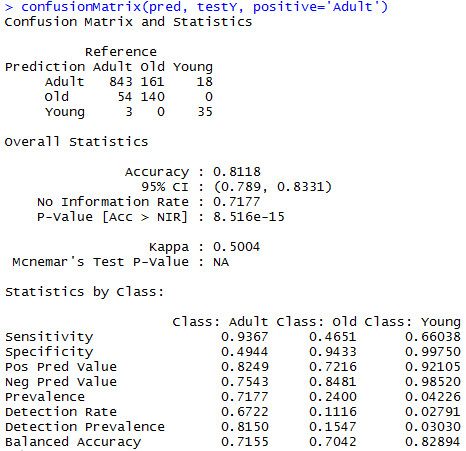


Figure 8. SVM Confusion Matrix – Radial Kernel

Especially a problem for the SVM model is that it scores a sensitivity of only 0.4651 for the “Old” class, correctly finding only 140 of 301 abalone in this class. The “Positive Predicted Value” statistic shows an accuracy of 72.16% (proportion of .7216, see figure 8). This means that a high proportion of abalone predicted as old actually should have been in the adult class, in this case 54 / 194 or 27.8%. Like the random Forest model, the SVM does not do well with old abalone. This tends to indicate that there is not much physical difference in the data between old and adult abalone.

One of the goals was to try different methods of measuring accuracy. In addition to raw accuracy; kappa, sensitivity and specificity have been presented here as other methods of determining a model’s fitness. Also, cross validation accuracy has been explored as a technique for building an optimal model. .

References:

Nash, Warwick & Sellers, Tracy & Talbot, Simon & Cawthorn, Andrew (1994). Abalone Data Set. University of California, Irvine, Machine Learning Repository. Retrieved from: <https://archive.ics.uci.edu/ml/datasets/Abalone>.

Lantz, Brett (2015). *Machine Learning with R, 2nd Edition*. Packt Publishing. Birmingham, UK.

Langer, David (2016). Introduction to Data Science with R - Cross Validation. YouTube. Retrieved from: <https://www.youtube.com/watch?v=84JSk36og34>.

Attachments:

Abalone-age.R: R code file for the exercise