Ensembles and Random Forest

Gary Seamans April 18, 2018

Short discussion on Ensembles. Citatations for additional reading are provided.

Machine learning ensembles use multiple techniques in order to improve prediction. Originally, ensemble classifiers used the results of multiple classifiers and Bayesian averaging to come up with a final prediction (Dietterich, 2000). Many other techniques are used for creating machine learning ensembles. Banfield, Hall, Bowyer, and Kegelmeyer (2007) lists five different techniques used for creating machine learning ensembles:

Ensemble Techniques

- Bagging
- Boosting
- Random Subspaces
- Random Forests™
- Randomized C4.5

Bagging involves taking the output from multiple techniques and creating a model that can then be used for prediction. ("Improve Predictive Performance in R with Bagging | R-bloggers," n.d.) is an example of using R to implement bagging.

Boosting is a technique, similar to *bagging*, for taking multiple weak learners, in machine learning, and combining them to create a better learner/precdictor. There are different boosting techiques, AdaBoost (Adaptive Boosting) (Freund and Schapire, 1995) is one of the earliest and best known techniques. There are a variety of boosting methodologies that are supported in R including *mboost* for which there is a very nice tutorial (Hofner, Mayr, Robinzonov, and Schmid, 2014). ("Improving Adaboosting with decision stumps in R | R-bloggers," n.d.) has a good description of AdaBoost and an example of applying it to *decision stumps*.

Random Subspaces is very much like *bagging* with the difference being that the features used for training are randomly selected for each of the multiple techniques that are being combined. This is a technique that might be very useful with the *high-dimensional* data that is commonly encountered with Big Data EDA (Li and Zhao, 2009).

Random Forests is a technique that uses the most common class, or the mean, of multiple decision trees to determine the output. ("Random Forest Using R," n.d.) is an excellent tutorial on random forest and implementing the techinque in R. Random Forests™is an extension of Random Forest that includes *bagging* and random feature selection. In a *random forest* each of the decision trees has a *vote* in determining the final class of an object.

C4.5 is a decision tree generation algorithm. Quinlan (1996) describes the C4.5 algorithm and modifications to C4.5 to support both *bagging* and *boosting*. Interestingly the author *Quinlan* no longer supports C.45 and has already published an R package for *C5.0*.

The experimental results in (Banfield et al., 2007) showed that in 65% of the tests none of the other ensemble techniques showed a statistically significant improvement over *bagging*. However, the two best techniques appeared to be *boosting* and *Random Forests*™, each scored multiple wins against *bagging* while never losing.

References

Banfield, R. E., Hall, L. O., Bowyer, K. W., and Kegelmeyer, W. P. (2007). A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(1), 173–180. http://doi.org/10.1109/TPAMI.2007.250609

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15). Springer. http://doi.org/10.1007/3-540-45014-9_1

Freund, Y., and Schapire, R. E. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23–37). Springer. http://doi.org/10.1007/3-540-59119-2_166

Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2014). Model-based boosting in R: A hands-on tutorial using

the R package mboost. *Computational Statistics*, *29*(1-2), 3–35.

Improve Predictive Performance in R with Bagging | R-bloggers. (n.d.). Retrieved from https://www.r-bloggers.com/improve-predictive-performance-in-r-with-bagging/

Improving Adaboosting with decision stumps in R | R-bloggers. (n.d.). Retrieved from https://www.r-bloggers.com/improving-adaboosting-with-decision-stumps-in-r/

Li, X., and Zhao, H. (2009). Weighted random subspace method for high dimensional data classification. *Statistics and Its Interface*, *2*(2), 153.

Quinlan, J. R. (1996). Bagging, boosting, and C4. 5. In *AAAI/IAAI, Vol. 1* (pp. 725–730). Retrieved from http://www.cs.ecu.edu/~dingq/CSCI6905/readings/BaggingBoosting.pdf

Random Forest Using R: Step by Step Tutorial. (n.d.). Retrieved from http://dni-institute.in/blogs/random-forest-using-r-step-by-step-tutorial/