Eriksson, L., Trygg, J. & Wold, S. (2009). PLS-trees®, a top-down clustering approach. *J. Chemom.*, **23**, 569–580.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *The Ann. of Stat.*, **19**, 1–67.

Gillo, M.W. & Shelly, M.W. (1974). Predictive modeling of multi-variable and multivariate data. *J. Amer. Stat. Assoc.*, **69**, 646–653.

Hastie, T., Tibshirani, R. & Friedman, J.H. (2009). *The Elements of Statistical Learning*, 2nd ed. New York: Springer.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J. & Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Comput.*, **3**, 79–87.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Boca Raton: Chapman and Hall/CRC.

Tenenhaus, M., Esposito Vinzi, V., Chatelinc, Y.M. & Lauro, C. (2005). PLS Path Modeling. *Comput. Stat. & Data Anal.*, **48**, 159–205.

Vermunt, J.K. (1997). *Log-linear Models for Event Histories*, Advanced Quantitative Techniques in the Social Sciences Series, vol. 8. Thousand Oakes: Sage Publications.

[*Received April 2014, accepted April 2014*]

International Statistical Review (2014), 82, 3, 357–359 doi:10.1111/insr.12061

Hongshik Ahn

State University of New York, Stony Brook, NY, USA
E-mail: hongshik.ahn@stonybrook.edu

The author presented a nice review of classification and regression trees by providing a discussion of major developments of the methods in the last 50 years. The author has made a great contribution to this field through developing fast and unbiased algorithms and applying the methods to various application areas. There has been a remarkable improvement in tree-structured methods. Due to the rapid advancement of computing capacity, even more computer intensive methods such as ensemble approach have been introduced.

Here, we will focus on discussing the properties of ensemble methods. There is a trade-off between a single tree and an ensemble method. Ensemble methods give higher prediction accuracy than a single tree in general. However, the ensemble method cannot compete with a single tree in interpretability as the author pointed out.

Three ensemble voting approaches, bagging, boosting and random subspace (Ho, 1998), have received attention. Because bagging and boosting were discussed in the paper, I will briefly discuss random subspace. Random subspace method combines multiple classification trees constructed in randomly selected subspaces of the variables. The final classification is obtained by an equal weight voting of the base trees. Ahn *et al.* (2007) proposed classification by ensembles from random partitions (CERP). CERP is similar to random subspace, but the difference is that base classifiers in an ensemble are obtained from mutually exclusive sets of predictors in CERP to increase diversity, whilst they are obtained by a random selection with overlap in random subspace.

The improvement in prediction accuracy in an ensemble from a single tree can be illustrated using a binomial model. If we assume independence amongst the n classifiers and equal prediction accuracy p of each classifier, where n is odd, the prediction accuracy of an ensemble classifier with majority voting is strictly increasing when $p > 0.5$ and strictly decreasing when $p < 0.5$ (Lam & Suen, 1997). The improvement of the prediction accuracy can be calculated using the beta-binomial model (Williams, 1975) when the the accuracies of the classifiers are positively correlated and using the extended beta-binomial model (Prentice, 1986) when they are negatively correlated.

The improvement of the ensemble accuracy illustrated earlier is valid under the assumption of equal accuracy of the base classifiers and equal correlation amongst the classifiers. Without these constraints, Breiman (2001) obtained the upper bound for the generalisation error. Convergence of the generalisation error rate depends on the average correlation, and it converges to zero when the classifiers are independent.

Logistic Regression Ensembles (LORENS: Lim *et al.*, 2010) is a logistic regression ensemble. LORENS uses the CERP algorithm to classify binary responses using the logistic regression model as a base classifier. This method enables class prediction by an ensemble of logistic regression models for a high-dimensional data set, which is impossible by a single logistic regression model due to the restriction that the sample size needs to be larger than the number of predictors. It is not as computer intensive as tree-based ensemble methods, whilst it does not lose the ensemble accuracy for high-dimensional data.

Recently, Kim *et al.* (2011) proposed weight-adjusted voting for ensemble (WAVE of classifiers). This method assigns unique voting weights to each classifier in the ensemble. Using an iterative process, a weight vector for the classifiers and another weight vector for the instances are obtained in the learning phase of model formation. They then proved the convergence of these vectors. After the final iteration, hard-to-classify instances get higher weights and subsequently, better performing classifiers on the hard-to-classify instances are assigned larger weights. Because a closed-form solution of the weight vectors can be obtained, WAVE does not need the iteration process.

In the evaluation of the performance of the classification methods, the sensitivity and specificity, positive predictive value, negative predictive value and receiver operating characteristic (ROC) curve also need to be considered. Most of the widely used classification methods have difficulties with unbalanced class sizes and almost always favour the majority class in order to increase the prediction accuracy.

Classification by ensembles from random partitions uses a different threshold from 0.5 in classification by logistic regression tree ensemble for unbalanced data. In a two-way classification, when r is the proportion of the positive responses in a data set, a threshold of r tends to give a better balance and a threshold of $1-r$ results in the highest accuracy (Chen *et al.*, 2006). Whilst a threshold of $1-r$ tends to yield the highest prediction accuracy, it worsens the balance by predicting more samples to the majority class. Pazzani *et al.* (1994) and Domingos (1999) assign a high cost to the misclassification of the minority class in order to improve a balance between sensitivity and specificity.

References

Ahn, H., Moon, H., Fazzari, M.J., Lim, N., Chen, J.J. & Kodell, R.L. (2007). Classification by ensembles from random partitions of high-dimensional data. *Comput. Stat. Data Anal.*, **51**, 6166–6179.

Chen, J.J., Tsai, C.A., Moon, H., Ahn, H. & Chen, C.H. (2006). The use of decision threshold adjustment in class prediction. *SAR & QSAR Environ. Res.*, **17**, 337–351.

Domingos, P. (1999). Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164. San Diego, California: ACM Press.

Ho, T.K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**, 832–844.

Kim, H., Kim, H., Moon, H. & Ahn, H. (2011). A weight-adjusted voting algorithm for ensembles of classifiers. *J. Korean Stat. Soc.*, **40**, 437–449.

Lam, L. & Suen, C.Y. (1997). Application of majority voting to pattern recognition: an analysis of its behavior and performance. *IEEE Trans. Syst. Man Cybern.*, **27**, 553–568.

Lim, N., Ahn, H., Moon, H. & Chen, J.J. (2010). Classification of high-dimensional data with ensemble of logistic regression models. *J. Biopharm. Stat.*, **20**, 160–171.

Pazzani, M., Merz, C., Murphy, P., Ali, K., Hume, T. & Brunk, C. (1994). Reducing misclassification costs: Knowledge-intensive approaches to learning from noisy data. In *Proceedings of the 11th International Conference on Machine Learning,* ML-94, pp. 217–225. New Brunswick:New Jersey.

Prentice, R.L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Assoc.*, **81**, 321–327.

Williams, D.A. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, **31**, 949–952.

[*Received March 2014, accepted March 2014*]

International Statistical Review (2014), 82, 3, 359–361 doi:10.1111/insr.12060

Chi Song and Heping Zhang

Department of Biostatistics, Yale University School of Public Health
E-mail: heping.zhang@yale.edu

We wish to congratulate the author for a nice overview of tree-based methods, and the author clearly highlighted the recursive partitioning technique (Friedman, 1977; Breiman *et al.*, 1984; Zhang & Singer, 2010) behind the tree-based methods. As the author summarized, there are two major types of tree methods: classification trees and regression trees, as precisely reflected in the title of the classical book by Breiman *et al.* (1984). In our own experience, for regression problems, other nonparametric methods, including adaptive splines (Friedman, 1991) that are based on a similar partitioning technique, appear more desirable than regression trees, with the exception of survival analysis (Zhang, 1997; 2004; Zhang & Singer, 2010).

With the advent of high-throughput genomic technologies, classification trees have become one of the most common and convenient bioinformatic tools. In what follows, we would like to share some of the recent developments in this area.

Genome-wide association studies (GWASs) collect data for hundreds of thousands or millions of single-nucleotide polymorphisms (SNPs) to study diseases of complex inheritance patterns, which can be recorded qualitatively (e.g. breast cancer) or in a quantitative scale (e.g. blood pressure). GWASs typically employ the case–control design, and the logistic regression model is generally applied to assess the association between each of the SNPs and the disease response, although more advanced techniques, especially nonparametric regression, have been proposed to incorporate multiple SNPs and interactions.

A clear advantage of classification trees is that they make no model assumption and that they can select important variables (or features) and detect interactions among the variables. Zhang & Bonney (2000) was among the early applications of tree-based methods to genetic association analysis. Since then, interests in tree-based genetic analyses have grown substantially. For example, Chen *et al.* (2007) developed a forest-based method on haplotypes instead of SNPs to detect gene–gene interactions, and importantly, they detected both a known variant and an unreported haplotype that were associated with age-related macular degeneration. Wang *et al.* (2009) further demonstrated the utility of this forest-based approach. Yao *et al.* (2009) applied GUIDE to the Framingham Heart Study and detected combinations of SNPs that affect the disease risk. García-Magariños *et al.* (2009) demonstrated that the tree-based methods were effective in detecting interactions with pre-selected variables that were marginally associated with the disease outcome but were susceptible to the local maximum problem when many noise variables were present. Chen *et al.* (2011) combined the classification tree and Bayesian search