



- 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

strategy, which improved the power to detect high-order gene–gene interactions at the cost of high computation demand.

Tree-based methods are extensively used in gene expression analysis to classify tissue types. Here, the setting is very different from the GWAS applications. In GWAS applications, we deal with a very large number of discrete risk factors (e.g. the number of copies of a particular allele). In expression analysis, the number of variables is large but not so large, usually in the order of tens of thousands, and the variables tend to be continuous. For example, Zhang *et al.* (2001) demonstrated that classification trees can discriminate distinct colon cancers more accurately than other methods. Huang *et al.* (2003) found that aggregated gene expression patterns can predict the breast cancer outcomes with about 90% accuracy using tree models. Zhang *et al.* (2003) introduced deterministic forests for gene expression data in cancer diagnosis, which have a similar power to random forests but are easier in scientific interpretation. Pang *et al.* (2006) developed a random forest method incorporating pathway information and demonstrated that it has low prediction error in gene expression analysis. Furthermore, Díaz-Uriarte & De Andres (2006) demonstrated that random forest can be useful in variable selection by using a smaller set of genes and maintaining a comparable prediction accuracy. Of a related note, Wang & Zhang (2009) attempted to address the following basic question: how many trees are really needed in a random forest? They provided empirical evidence that a random forest can be reduced in size so much to allow scientific interpretation.

As more and more data are generated from new technologies such as the next-generation sequencing, tree-based methods will be very useful for analysing such large and complex data after necessary extensions. Closely related to genomic data analysis is the personalized medicine. Zhang *et al.* (2010) presented a proof of concept that tree-based methods have some unique advantages over parametric methods to identify patient characteristics that may affect their treatment responses. In summary, tree-based methods have thrived in the past several decades, and they will become more useful, and the methodological developments will be more challenging than ever, as more information increases in both size and complexity.

Acknowledgements

This research is supported in part by grant R01 DA016750 from the National Institute on Drug Abuse.

References

- Chen, M., Cho, J. & Zhao, H. (2011). Detecting epistatic SNPs associated with complex diseases via a Bayesian classification tree search method. *Ann. of Human Gen.*, **75**, 112–121.
- Chen, X., Liu, C., Zhang, M. & Zhang, H. (2007). A forest-based approach to identifying gene and gene–gene interactions. *Proceedings of the National Academy of Sciences*, **104**, 19199–19203.
- Díaz-Uriarte, R. & De Andres, S.A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.
- Friedman, J.H. (1977). A recursive partitioning decision rule for nonparametric classification. *IEEE Trans. Comput.*, **C-26**, 404–407.
- Friedman, J.H. (1991). Multivariate adaptive regression splines. *IEEE Trans. Comput.*, **19**, 1–141.
- García-Magariños, M., López-de Ullibarri, I., Cao, R. & Salas, A. (2009). Evaluating the ability of tree-based methods and logistic regression for the detection of SNP–SNP interaction. *Ann. of Human Gen.*, **73**, 360–369.
- Huang, E., Cheng, S., Dressman, H., Pittman, J., Tsou, M., Horng, C.F., Bild, A., Iversen, E.S., Liao, M., Chen C.M., West, M. & Nevins, J.R. (2003). Gene expression predictors of breast cancer outcomes. *The Lancet*, **361**, 1590–1596.
- Pang, H., Lin, A., Holford, M., Enerson, B.E., Lu, B., Lawton, M.P., Floyd, E. & Zhao, H. (2006). Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028–2036.
- Wang, M. & Zhang, H. (2009). Search for the smallest random forest. *Stat. Interface*, **2**, 381–388.

- Wang, M., Zhang, M., Chen, X. & Zhang, H. (2009). Detecting genes and gene–gene interactions for age-related macular degeneration with a forest-based approach. *Stat. Biopharm. Res.*, **1**, 424–430.
- Yao, L., Zhong, W., Zhang, Z., Maenner, M. & Engelman, C. (2009). Classification tree for detection of single-nucleotide polymorphism (SNP)-by-SNP interactions related to heart disease: Framingham Heart Study. *BMC Proceedings*, **3**, S83.
- Zhang, H. (1997). Multivariate adaptive splines for analysis of longitudinal data. *J. Comput. Graph. Statist.*, **6**, 74–91.
- Zhang, H. (2004). Mixed effects multivariate adaptive splines model for the analysis of longitudinal and growth curve data. *Stat. Methods Med. Res.*, **13**, 63–82.
- Zhang, H. & Bonney, G. (2000). Use of classification trees for association studies. *Genet. Epidemiol.*, **19**, 323–332.
- Zhang, H., Legro, R.S., Zhang, J., Zhang, L., Chen, X., Huang, H., Casson, P.R., Schlaff, W.D., Diamond, M.P., Krawetz, S.A., Coutifaris, C., Brzyski, R.G., Christman, G.M., Santoro, N. & Eisenberg, E. (2010). Decision trees for identifying predictors of treatment effectiveness in clinical trials and its application to ovulation in a study of women with polycystic ovary syndrome. *Human Reproduction*, **25**, 2612–2621.
- Zhang, H., Yu, C. & Singer, B. (2003). Cell and tumor classification using gene expression data: Construction of forests. *Proceedings of the National Academy of Sciences*, **100**, 4168–4172.
- Zhang, H., Yu, C., Singer, B. & Xiong, M. (2001). Recursive partitioning for tumor classification with gene expression microarray data. *Proceedings of the National Academy of Sciences*, **98**, 6730–6735.

[Received March 2014, accepted March 2014]

International Statistical Review (2014), **82**, 3, 361–367 doi:10.1111/insr.12062

Thomas Rusch¹ and Achim Zeileis²

¹ IWU Vienna, Vienna, Austria

E-mail: thomas.rusch@wu.ac.at

² Universität Innsbruck, Innsbruck, Austria

E-mail: Achim.Zeileis@R-Project.org

1 Introduction

We thank Wei-Yin Loh for this review paper. He provides a much-needed guide to tree methods currently available as well as the main ideas behind them, indicative of his experience with and knowledge of this topic. His contribution proves to be very valuable in bringing structure into the vast interdisciplinary field of tree algorithms: We found 83 different tree induction algorithms for different response types listed in his paper, and, along the lines of Loh's disclaimer, this is not even an exhaustive list.

The availability of so many different algorithms for fitting tree-structured models directly relates to the main point of our discussion: The tree literature is highly fragmented. Loh hints at that issue already on the first page, and we gladly take it up for discussion: There are so many recursive partitioning algorithms in the literature that it is nowadays very hard to see the wood for the trees.

In the remainder of our discussion paper, we identify causes for and consequences of this fragmentation, discuss what we perceive to be advantages and disadvantages of the current state of the tree algorithm literature and offer suggestions that might improve the situation in the years ahead by retaining advantages and overcoming disadvantages.

2 The Fragmentation of Tree Algorithms

Currently, there is an abundance of different tree algorithms coming from different communities including statistics, machine learning and other fields. We believe that this fragmentation