Published in final edited form as:

Int Stat Rev. 2014 December 1; 82(3): 359–361. doi:10.1111/insr.12060.

Comments on Fifty Years of Classification and Regression Trees

Chi Song and Heping Zhang

Department of Biostatistics, Yale University School of Public Health

We wish to congratulate the author for a nice overview of the tree-based methods, and the author clearly highlighted the recursive partitioning technique (Friedman, 1977; Breiman et al., 1984; Zhang and 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 reected 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 and 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 et al. (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 examples, 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 (FHS) 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 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 and 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 and Zhang (2009) attempted to address the basic questions: 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 analyzing 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.

Acknowledgments

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

References

- Chen M, Cho J, Zhao H. Detecting epistatic snps associated with complex diseases via a bayesian classification tree search method. *Annals of Human Genetics*. 2011; 75:112–121. [PubMed: 21121902]
- Chen X, Liu C, Zhang M, Zhang H. A forest-based approach to identifying gene and gene-gene interactions. *Proceedings of the National Academy of Sciences*. 2007; 104:19199–19203.
- Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*. 2006; 7:3. [PubMed: 16398926]
- Friedman JH. A recursive partitioning decision rule for nonparametric classification. *IEEE Transaction on Computers*. 1977; C-26:404–407.

- Friedman JH. Multivariate adaptive regression splines. *IEEE Transaction on Computers*. 1991; 19:1–141.
- García-Magariños M, López-de Ullibarri I, Cao R, Salas A. Evaluating the ability of tree-based methods and logistic regression for the detection of snp-snp interaction. *Annals of Human Genetics*. 2009; 73:360–369. [PubMed: 19291098]
- Huang E, Cheng S, et al. Gene expression predictors of breast cancer outcomes. *The Lancet*. 2003; 361:1590–1596.
- Pang H, Lin A, et al. Pathway analysis using random forests classification and regression. *Bioinformatics*. 2006; 22:2028–2036. [PubMed: 16809386]
- Wang M, Zhang H. Search for the smallest random forest. *Statistics and its Interface*. 2009; 2:381–388. [PubMed: 20165560]
- Wang M, Zhang M, Chen X, Zhang H. Detecting genes and gene-gene interactions for age-related macular degeneration with a forest-based approach. *Statistics in Biopharmaceutical Research*. 2009; 1:424–430. [PubMed: 20161521]
- Yao L, Zhong W, Zhang Z, Maenner M, Engelman C. Classification tree for detection of single-nucleotide polymorphism (snp)-by-snp interactions related to heart disease: Framingham heart study. *BMC Proceedings*. 2009; 3:S83. [PubMed: 20018079]
- Zhang H. Multivariate adaptive splines for analysis of longitudinal data. *Journal of Computational and Graphical Statistics*. 1997; 6:74–91.
- Zhang H. Mixed effects multivariate adaptive splines model for the analysis of longitudinal and growth curve data. *Statistical Methods in Medical Research*. 2004; 13:63–82. [PubMed: 14746441]
- Zhang H, Bonney G, et al. Use of classification trees for association studies. *Genetic Epidemiology*. 2000; 19:323–332. [PubMed: 11108642]
- Zhang H, Yu C, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. *Proceedings of the National Academy of Sciences*. 2001; 98:6730–6735.
- Zhang H, Yu C, Singer B. Cell and tumor classification using gene expression data: construction of forests. *Proceedings of the National Academy of Sciences*. 2003; 100:4168–4172.
- Zhang H, Legro RS, Zhang J, et al. 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*. 2010; 25:2612–2621. [PubMed: 20716558]