References

Archer, K.J. & Kimes, R.V. (2008). Empirical characterization of random forest variable importance measures. *Comput. Stat. Data Anal.*, **52**(4), 2249–2260.
Azen, R. & Budescu, D.V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychol. Methods*, **8**(2), 129–48.
Bureau, A., Dupuis, J., Falls, K., Lunetta, K.L., Hayward, B., Keith, T.P. & Eerdewegh, P.V. (2005). Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol.*, **28**(2), 171–182.
Diaz-Uriarte, R. & de Andrés, S.A. (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**(3).
Hapfelmeier, A., Hothorn, T., Ulm, K. & Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Stat. Comput.*, **24**(1), 21–34.
Liaw, A. & Wiener, M. (2002). Classification and regression by `randomForest`. *R News*, **2**(3), 18–22.
Little, R. & Rubin, D. (1986). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.
Lunetta, K.L., Hayward, L.B., Segal, J. & Eerdewegh, P.V. (2004). Screening large-scale association study data: Exploiting interactions using random forests. *BMC Genetics*, **5**(32).
Nicodemus, K. & Shugart, Y.Y. (2007). Impact of linkage disequilibrium and effect size on the ability of machine learning methods to detect epistasis in case–control studies. In *Proceedings of the Sixteenth Annual Meeting of the International Genetic Epidemiology Society*, Vol. 31, North Yorkshire, UK, pp. 611.
Rodenburg, W., Heidema, A.G., Boer, J.M., Bovee-Oudenhoven, I.M., Feskens, E.J., Mariman, E.C. & Keijer, J. (2008). A framework to identify physiological responses in microarray based gene expression studies: selection and interpretation of biologically relevant genes. *Genet. Epidemiol.*, **33**(1), 78–90.
Strobl, C. (2013). Data mining. In *The Oxford Handbook on Quantitative Methods*, Ed. T. Little, pp. 678–700. USA, Chapter 29: Oxford University Press.

[*Received March 2014, accepted March 2014*]

International Statistical Review (2014), 82, 3, 352–357 doi:10.1111/insr.12065

Antonio Ciampi

Department of Epidemiology and Biostatistics, McGill University,1020 Pine Ave. West, Montreal H3A 1A2, Quebec, Canada
E-mail: antonio.ciampi@mcgill.ca

Fifty years ago, Morgan & Sonquist (1963) introduced the now famous AID algorithm. AID uses a sample to construct a tree-structured predictor for a specified continuous variable y given a specified vector of covariates x. The result is therefore, in contemporary language, a *regression* tree. The first paper to introduce *classification* trees in the modern sense appeared 9 years later, in 1972: it presented the algorithm THAID, as mentioned in the review (REF). We have to wait 12 more years to see both kinds of trees reunited in the very influential book by Breiman *et al*. (1984), 'Classification and Regression Trees'. This title is often abbreviated as CART, which is somewhat confusing, because the acronym CART™ also denotes the proprietary software associated to the book. Notwithstanding the great merits of the CART book, it is more than fair to consider the 1963 paper as the seminal work for the *research area* known as 'Classification and Regression Trees', the object of Prof. Loh's excellent review. One could argue that the 1963 paper, together with the CART book, is also at the origin of flexible statistical modelling (beyond variable selection algorithms in regression) such as MARS (Friedman, 1991) and PIMPLE (Breiman, 1991), and, indeed, of *statistical* machine learning (Hastie *et al.*, 2009) . Yet one cannot disagree with the author's decision of restricting the review to classification and regression trees, and to draw the line at 'forests' and similar ensemble learning. As the old saying goes, 'Grab all, loose all'.

But what is exactly 'Classification and Regression Trees' as a research area? The answer is open to debate. One might say that it includes research on algorithms that construct a tree from data through some kind of recursive partitioning coupled with a rule or set of rules to determine tree size. It should be noted that partitioning also includes some choices as regards the handling of missing data. Choosing how to partition and how to determine tree size is not trivial, and there is a bewildering array of perfectly respectable approaches to these tasks, often leading to minimally divergent results. One might add that the terminal nodes (leaves) of the tree represent a simple prediction or classification rule, depending on whether we are concerned, respectively, with regression or classification trees. How simple? In a strict sense, the rule should be constant on a leaf, i.e. the same prediction/classification is attached to all observational units belonging to the same leaf.

Prof. Loh's review goes a little beyond the strict definition. Firstly, it does include a number of regression tree algorithms, including his own, that fit linear models at each node, thus relaxing the requirement of a constant predictor at each leaf. A minor criticism is that it would have been useful to explicitly note that the idea of fitting linear models in terminal nodes is not exclusive to *regression* trees but can also be implemented for *classification* trees; however, doing this would somewhat disrupt the scheme of the review, which is based on keeping regression and classification separate.

Secondly, the author also considers tree construction algorithms for predicting count data, censored data, multivariate binary data and longitudinal continuous data (functional data). He may have included, but this is by no means an important flaw, trees for predicting multivariate continuous data (beyond longitudinal data) (Gillo & Shelly, 1974): indeed, in one of the RECPAM articles (Ciampi *et al.*, 1991) that he reviews, there are examples of such trees. Minor details aside, the author commendably transcends the old identification of regression trees with trees to predict a continuous, scalar variable.

Thirdly, this review also mentions, although not in great depth, that 'hard' partitioning may be replaced by 'soft' partitioning, i.e. at a given node, an observational unit may be assigned to the issuing branches probabilistically rather than sharply.

Finally, the author discusses some global tree-construction algorithms, which look for the optimal tree within (a large subset of) all possible trees: an application of the genetic algorithm (REF), as well as two Bayesian 'model averaging' approaches are mentioned (REF).

In developing the review, again the author shows wisdom in concentrating on algorithms that have been extensively applied and validated. The variations in partitioning rules (including the treatment of missing data) and size determining rules are clearly if succinctly outlined. Comparative work is cited, and a simple and enlightening original comparison of the performance of several algorithms on a 'classic' data set is presented. The author cites in detail his own work, and I find this totally acceptable: indeed, one of Prof. Loh's major accomplishments is that of having blazed a trail within all the possible variants of tree-growing algorithms obtaining superior accuracy of prediction, high computational efficiency and major reduction of the inherent biases in the original CART (1984) approach.

One feature of this review I have particularly appreciated is the stress on tree algorithms that were developed within the machine-learning tradition at a time when the compartmentalisation separating statisticians from computer scientist was watertight. Indeed, Quinlan's work was just as influential among computer scientists as the CART book was among statisticians: the review clearly re-establishes the balance. In my opinion, if tree research will continue to advance in the near future, it will be because disciplinary boundaries are falling. Nowadays it is not rare to find, especially in the new generations, accomplished researchers who are excellent in both statistics and computing, regardless of their disciplinary background.

And now the hard question would be, is there a future for tree research? Prof. Loh points out that there are hard problems left to solve within the strict definition of regression and classification trees. He cites, in particular, the handling of missing data, the inclusion in tree-growing algorithms of longitudinal *predictor variables* and the introduction of splits based not only on unique variables (monothetic splits or nodes) but also on linear combinations of the original variables. Moreover, he hints at new developments that include incremental tree construction algorithms for streaming data: again, such algorithms go a little beyond the strict definition of classification and regression trees, which were originally conceived in a static setting, with a well-defined 'learning' data set. However, the main message of his concluding remarks is that tree-growing researchers are faced with a real dilemma: *either* one sticks to the classic definition of trees, and in so doing, accepts intrinsic limitations in predictive accuracy; *or* one pursues predictive accuracy by extending the definition of tree-growing algorithms towards ensemble learning, and sacrifices, in exchange, the advantage of highly interpretable predictions. It is difficult not to agree with this view, which I would define as 'realistically pessimist'.

However, when faced with a dilemma of this importance, it may be useful to step back and take a fresh look at the premises that have lead us to the dilemma. Here is a short list of questions that occurred or re-occurred to me while reading Prof. Loh's paper. I use these questions as headings for grouping some considerations related to them.

1 What is the Root of this Dilemma?

The essential feature of tree growing, which is also the reason for its popularity, is the reduction of one *global* optimisation problem–given a learning set, which is considered as a sample from a target population, find the *best* predictor–to a sequence of *local* problems of decreasing sizes according to the general cognitive strategy of 'divide and conquer'. However, no matter how large our learning set is, one is led very quickly to work locally on fairly small data sets: and this is perhaps the root of most problems. The smaller the subsamples, the more variable and unstable are the choices of splits, and the less generalisable are the results to other future samples from the same population. There is no way out of this, *unless* we broaden somewhat the definition of tree growing, to make it *a little less local*. This leads to the next question.

2 Should We Redefine Tree Growing and How?

The author and I agree on the fact that ensemble learning, at least as realised in available algorithms, is only distantly related to tree growing. From the point of view of 'tree growers', the loss in interpretability of such algorithms is too large. To take the example of random forest, the trees of the forest are far too numerous to help developing an interpretation; moreover, there is too much randomness in the generation of each tree.

On the other hand, some promising novel ideas have been put forward leading to algorithms that deserve to be considered as part of the tree-growing family. The Bayesian tree approach, although similar to ensemble learning, does recover some interpretability. Indeed, the analyst has the choice of using model averaging for prediction, while basing interpretation on a (usually) small number of trees: the one(s) with largest posterior probability.

TARGET, a tree-growing approach based on the genetic algorithm, also yields a 'best' tree, although it does *not* proceed by recursive partitioning. However, it is still not known whether in practice the theoretical superiority of the global search does translate into substantially superior predictive accuracy.

Trees with soft nodes were proposed in an attempt to gain predictive accuracy while mitigating the inevitable loss of interpretability. Predictive accuracy is increased by using at each node, *all* data–but with observational units weighted according to the probability of belonging to the node. Loss of interpretability is mitigated by retaining the monothetic feature of classical trees (one split, one variable). However, trees with soft nodes do not seem very useful when there are many categorical predictors. Also, empirical results obtained so far seem to indicate that the gain in predictive accuracy of soft trees with respect to hard trees is real but unimpressive.

All in all, it seems that the most promising ways to extend tree growing beyond its strict definition is to look for an interesting compromise between interpretability and predictive accuracy. But. . .

3 What is Interpretability? or, Better, What Kind of Interpretability do We Really Need?

At first sight, interpretability of a tree seems to hinge on the monothetic nature of the nodes. However, a closer look suggests that this point of view may be misleading: monothetic splits are *not* always what we need. In fact, as already noted, Prof. Loh identifies the introduction of splits based on linear combination of variables as one of the most important open problems in tree-growing research. When do we need to depart from monothetic splits? For instance, consider medical data: typically, predictors are categorical variables based on qualitative observations of symptoms and signs, and/or imprecise measurements of indices that are known to be, at best, proxy of some underlying construct, e.g. 'cardiovascular health'. Is it really useful to choose, say, 'elevated total blood cholesterol' to create a node? Looking for splits based on a linear combination of variables is a natural alternative to looking for a monothetic split; however, how to do this remains problematic, and in fact, we risk to loose interpretability *without* improving predictive accuracy. So, it may be that new ideas are needed to look for polythetic splits: such ideas may arise from a creative interplay of clinical expertise and statistical modelling. Using techniques such as PLS regression at each node (Eriksson *et al.*, 2009), one may extract from the data a split defining statement of the kind clinicans are used to while making diagnosis and prognosis, e.g. '*if* the subject has one or more characteristics of the following list. . . ., *then go left*'.

The other essential pillar of interpretability for a tree-based predictor is its simple architecture: a hierarchy of (hard) nodes. It is possible, in my opinion, to make this framework more flexible without completely loosing interpretability. Perhaps we may consider interpretable an architecture consisting of hierarchically structured 'black boxes', each being based on a limited number of variables that in some intuitive sense 'go together'. Moreover, 'soft nodes' rather than 'hard nodes' could link these black boxes. For example, suppose we want to predict cardiovascular mortality: we may aim to construct a predictor by stringing together a black box based on measurements of blood lipid levels, another black box based on family history data and yet another black box based on demographics. If such a predictor works well, it is possible that a knowledgeable user may find it interpretable. Arguably, he or she may *prefer* this alternative view, as it recognises some of the complexities of the specific predictive task. Now, such system of black boxes linked by soft nodes already exists in machine learning, and is known as 'hierarchy of experts' (Jacobs *et al.*, 1991). However, to the best of my knowledge, there is no popular algorithm for *constructing* hierarchy of experts from data, including, as possible, domain specific knowledge. In other words, the concept of hierarchy of experts does provide an excellent framework to imagine algorithms, but is not (yet) a ready-to-use discovery tool. There is here a great opportunity for tree growers: they could use their unique expertise to build problem-specific hierarchy of experts using both data and knowledge bases.

The last question concerns the role of tree-growing research in the context of new challenges arising from the increasing complexity of available data. Volume can be seen as an aspect of complexity; in this respect, I will not add to what Prof. Loh has already mentioned in this review, citing, among others, recent papers on tree-growing algorithm for streaming data. Instead, I wish to briefly discuss another type of complexity, which cannot be dealt with without rethinking prediction and prediction accuracy.

4 What Kind of Prediction Tools do We Really Need to Explore New Types of Data?

Again, I will discuss an example from clinical biostatistics. It becomes increasingly common to collect longitudinal data not only for a particular outcome variable but also for several clinical indices and for several categorical outcome variables. In other words, data become available that summarise the history of a disease as observed on a population of patients over a time window of considerable width. Clearly, a first task for the analyst is to develop the appropriate statistical models for the stochastic process underlying life history data: this task that has been successfully accomplished for a broad variety of situations (REF) (Skrondal & Rabe-Hesketh, 2004; Tenenhaus *et al.*, 2005; Vermunt, 1997). But then, typically, the analyst is also asked to assess the impact of covariates, e.g. patient and treatment characteristics, on the type of disease history that a patient is likely to experience. This is a prediction task, in a very true sense, but is not a standard one: we are very far from the classical problem of predicting a continuous or categorical variable. The statistician's automatic reflex would be to develop some (generalised) linear regression model that should describe the dependence of *some features* of the disease history process on the covariate of interest. However, these features will be represented by a high-dimensional parameter, so that a hypothetic regression model would be extremely hard to interpret. In contrast, a tree-growing approach, if it could be developed, would lead to a fairly straightforward interpretation. *If it could be developed…* Developing this approach is a serious but not impossible task. The tree-growing approach has been formulated and reformulated in abstract terms by several authors, leading to some of the extensions reviewed by Prof. Loh. Further and bolder developments are possible. Conceptually, all we need to do is to define a reasonable measure of 'goodness of split' for the appropriate stochastic process underlying the available data. *If this can be accomplished*, then virtually any tree-growing algorithm can be adapted to the new situation. The adaptation will be, in general, far from trivial and will require new statistical and computational developments: in other words, a great amount of original tree research may be produced, well beyond the present perspective.

To conclude, I wish to thank Prof. Loh for an excellent review of the status of tree research as it celebrates its 50th anniversary. Because I recognise that it would be very hard to do better, I have focussed on potential for future development. I am cautiously optimistic about the *next* 50 years of tree research. The reason for my optimism is the increasing cooperation of researchers from several disciplines that have in the past ignored each other, often 'rediscovering the wheel'. The reason for my caution is that the task of forming the next generation of tree researchers is fraught with many obstacles, but a discussion of this is well beyond the scope of my contribution.

References

Breiman, L. (1991). The Π method for estimating multivariate functions from noisy data. *Technometrics*, **33**, 125–143.

Ciampi, A., du Berger, R., Taylor, G. & Thiffault, J. (1991). RECPAM: a computer program for recursive partition and amalgamation for survival data and other situations frequently occurring in biostatistics. III. Classification according to a multivariate construct. Application to data on Haemophilus influenzae type B meningitis. *Comput. Methods and Prog. in Biomed.*, **36**, 51–61.

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.