

Trees with leaves and without leaves

Roberta Siciliano, Valerio A. Tutore, Massimo Aria, Antonio D'Ambrosio

Abstract Tree is a connected and oriented graph to describe a hierarchical data structure with a set of linked nodes and the end-nodes called "leaves". Tree data structures are considered for supervised classification as well as for nonparametric regression in many fields of applications. Trees with leaves are exploratory trees to learn about the dependence relationship between the target variable and the set of predictors. Ensemble methods and random forest combine more tree structures to define an accurate decision rule for that there is no one tree structure and prediction is made by trees without the possibility to interpret their leaves. This approach can be fruitfully considered for incremental missing data imputation. This paper provides some methodological proposals to improve stability in tree growing and to reduce the computational cost while assuring accuracy in decision rule production.

Key words: Data Mining and Machine Learning, Recursive Partitioning, Ensemble, Incremental missing data imputation

1 Introduction

Data can be hierarchically organized in a connected and oriented graph, the so-called tree, characterized by a set of linked nodes, in which any two nodes are connected by exactly one simple path, the starting-node is the "root" and the end-nodes are the "leaves". Two properties are satisfied: the shape property, where each node has a fixed number r of child nodes (for $r = 2$ it is assumed a binary tree); the heap property, where each node is greater than or equal to each of its children according to some comparison predicate which is fixed for the entire data structure. Trees are widely used in supervised classification and non parametric regression for data mining and prediction [Hastie, Friedman and Tibshirani, 2001] as well as for intelligent

Department of Mathematics and Statistics, University of Naples Federico II, e-mail: (roberta, v.tutore, aria, antdambr)@unina.it

data analysis [Berthold and Hand, 2003]. As in CART methodology [Breiman *et al.*, 1984], each node of the tree includes objects which are recursively partitioned such to reduce the impurity of a target or response variable as explained by a set of available predictors. Partitioning of the objects is determined by the best partition of any predictor's modalities into r groups, where the number of candidate partitions of each predictor depends on the number of distinct modalities assumed by the predictor, which is very high if the domain is real. Impurity can be measured by an heterogeneity index (Gini index, entropy measure, misclassification error) in case of categorical response (usually dummy variable) for classification trees, whereas by variance or deviance in case of numerical response for regression trees. To each leave of the tree is assigned a response value/class, the set of leaves describes a partition of the given sample of objects, each path of the tree gives the sequential conditions of the predictors measurement which is necessary to belong to each final leave. Trees with leaves can be used for exploratory analysis in order to understand the dependence relationships between the target variable and the predictors. A new object for that only the predictors measurements are known can be slide down until falling into one of the leaves so to be assigned to a response value/class. The quality of the prediction can be evaluated in terms of misclassification rate or mean square error estimates based on learning sample (too optimistic), test sample (which requires large sample size), cross-validation (for small sample size). In the trade-off between bias and variance it is necessary to identify the proper tree model complexity. Large tree structures are inaccurate because of a large variance (too much sensitivity sample) whereas tree-models with too few leaves are inaccurate because of a large bias (not enough flexibility). CART suggests to grow the maximal expanded tree, then find a sequence of nested pruned trees (cutting off at each step the weakest link in terms of a cost-complexity parameter), finally select the best pruned tree on the basis of the cross-validation or test sample estimate of the misclassification rate or mean square error. Main focus of recent literature is to outperform the decision rule in terms of accuracy such to answer the bias-variance dilemma with alternative solutions. Enhancements are provided by ensemble methods [Breiman, 1996; Freund and Shapire, 1997; Dietterich, 2000] and random forest [Breiman, 2001]. All these approaches do not provide one tree structure for prediction denying the interpretability advantage of the tree graph to describe the hierarchical dependence relationships. The final assignment of one object is induced by a suitable combination of tree structures.

This paper is structured in the following way. In section 2, enhancements of partitioning criteria to grow exploratory trees are discussed. In section 3. a bootstrapped splitting tree is introduced in order to yield more stable trees; results on real and simulated data sets analyses are shown. Section 4 recalls ensemble methods. This approach is computationally intensive so that a methodological proposal is described in section 5 to speed up the overall procedure; empirical evidence is shown in real and simulated data set analyses. All these results turn to be particularly convenient in incremental approach for missing data imputation and data fusion as discussed in section 6.

2 Enhancements in partitioning criteria

Recursive partitioning procedure can be understood as a greedy searching algorithm which looks forward, at each node, for the best partition of objects into r groups induced by the best partition of any predictor's modalities. The shape property is satisfied so that r is fixed for each node and typically it is equal to 2 yielding to binary trees. The common predicate of the heap property is defined by the optimality partitioning criterion to find the best partition of the objects into r subgroups such to provide the highest decrease of the impurity of the response variable at the top node, which is always greater or equal to the sum of the impurities in the child nodes weighted by the percentages of objects within each subnode.

As an alternative to CART, in TWO-STAGE methodology for univariate trees (one response variable) and multivariate trees (more response variable) [Mola and Siciliano, 1992; Siciliano and Mola, 2000], the best partition of objects is selected from the best partition of the best predictor's modalities (or a subset of best predictors' modalities) according to global and local predictability measures (i.e. Goodman tau index for qualitative responses, Pearson correlation coefficient for numerical responses). On the basis of suitable properties of predictability measures, FAST algorithms [Siciliano and Mola, 1996, 1997], which can be understood as locally iterative two-stage partitioning, have been introduced in order to yield the same CART tree, saving high computational cost. TWO-STAGE approach has offered the opportunity to think over alternative partitioning criteria, where statistical indexes are replaced by statistical modeling [Siciliano and Mola, 2004]. Some goodness of fit measures can be considered in the common predicate where model fit improves when passing from the top node to r subnodes. An hybrid two-stage approach based on model selection, for choosing the most predictive variables, and partitioning upon model parameters allows to improve classification of multiclass responses as in latent budget trees for multiple classification [Siciliano, 1999] and in ternary trees by nonsymmetric correspondence analysis [Siciliano and Mola, 1998], as well as to deal with nonlinear dependence relations in nonlinear canonical correlation analysis for optimal scaling trees [Tutore and Mooijaart, 2007]. Furthermore, partitioning criteria have been introduced to deal with three-way data structures [Tutore *et al.*, 2007], preference rankings [D'Ambrosio and Heiser, 2009], multilevel hierarchy of data [Giordano and Aria, 2010]. The concept of retrospective split for binary trees (easily extendable to r -way trees), introduced in Two-Stage Discriminant Trees for regression trees [Mola and Siciliano, 2002], consists in looking backward to the most coherent partition of any predictor's modalities once that the optimal (theoretical) partition of the objects is defined in terms of the response distribution discarding the predictors. Tree growing based on just the optimal partitions of the objects yields to the optimal tree, such that retrospectively it can be associated an hierarchy of posterior prediction models [Siciliano, Aria, D'Ambrosio, 2008] as well as the best predictors' modalities partitions such as in two-class trees for non parametric regression [Siciliano and Aria, 2010].

3 Stability improvement in tree-growing

A weak point of exploratory trees is stability since the best predictor's modalities which generate the best partition of objects at each node may change upon data perturbation. In regression trees, a robust procedure based on M-estimators has been proposed [Galimberti *et al.*, 2007]. A bootstrap estimation method is proposed in order to select the best partition of the objects upon weighted averaging of the decreases of impurity measures in bootstrapped samples. In our experiment we used five well known dataset for the classification case (German credit, Australian credit, Breast cancer Wisconsin and Pima Indians Diabetes from UCI Machine learning repository, and Credit from the repository of the French software SPAD) as well as two simulated dataset for the regression case. The simulations were performed by using different random distributions for the set of predictors (Normal, Uniform, Discrete uniform, Binomial), whereas the response variable was generated by a non linear link with other variables. It is known that tree-based methods are unstable, i.e. they return different tree-structures after little changes in the data input. For that reason we performed our tests by randomly deleting, for each considered dataset, a proportion of cases equal respectively to 1%, 5% and 10%. Then, for each level of data perturbation, we ran k times ($k = 50$) both the classical CART and our Bootstrapped Splitting Tree. We used the simple matching coefficient as stability measure. For both the methodologies we compared all the possible couples of trees, by creating in this way a matrix in which we reported the matching of any of the internal nodes in terms of both the variable generating the split and the split tail. In this way, for each perturbation level, we obtained a similarity matrix among all the trees, then we computed the averaged similarity that is reported in the columns 6 and 7 of the table 1. It is known that the more the matching coefficient is close to 1, the more the objects (in our case the trees) are similar.

Looking at the last two columns of the table, we can say that if all the predictors are categorical then the average matching coefficients of the Bootstrapped Splitting Tree are higher than the ones of the classical CART for each perturbation level. The same occurrence happens if most of the predictors are categorical (Australian credit, German credit, Simulation 2). If all the predictors are numerical, or if most of the predictors are numerical, it seems that classical CART returns more stable trees than the Bootstrapped Splitting Tree (except for Breast Cancer Wisconsin at the highest level of data perturbation).

4 Ensemble methods

Ensemble methods are learning algorithms that develop a population of simple models (like trees), called weak learner, from the perturbed training set combining them to form a composite prediction rule, which is generally more accurate than the single trees whence it is formed by. Ensemble of classifiers works by constructing a set of weak learners and then classifying new data points by taking a vote of their predic-

Dataset	Sample size	Numerical predictors	Categorical predictors	Proportion of data perturbation	CART (Avg. matching coefficient)	Our criterion (Avg. matching coefficient)
Breast Cancer Wisconsin	699	9	0	1,0%	0,8134	0,8076
				5,0%	0,6089	0,4392
				10,0%	0,3091	0,3894
Pima Indians Diabetes	768	8	0	1,0%	0,8024	0,7643
				5,0%	0,6128	0,5285
				10,0%	0,3840	0,3612
Credit (Cisia)	468	0	11	1,0%	0,8215	0,9201
				5,0%	0,5163	0,6320
				10,0%	0,4061	0,4706
Australian credit	690	6	8	1,0%	0,6169	0,6878
				5,0%	0,4865	0,5418
				10,0%	0,3474	0,4373
German credit	1000	8	12	1,0%	0,6532	0,7488
				5,0%	0,5003	0,5905
				10,0%	0,3259	0,4248
Simulation 1 (regression)	5000	10	3	1,0%	0,7235	0,6454
				5,0%	0,5190	0,4781
				10,0%	0,3267	0,2908
Simulation 2 (regression)	5000	4	12	1,0%	0,7561	0,8523
				5,0%	0,6011	0,7494
				10,0%	0,4245	0,5612

Table 1 Tree instability: Bootstrapped Splitting Tree versus CART

tions. Even though there exists several ways to build ensemble [Dietterich, 2000], the most popular ensemble methods, such as Bagging [Breiman, 1996], Boosting [Freund and Schapire, 1997] and Random Forest [Breiman, 2001], work by manipulating the training examples through re-sampling methods. In perturbing the data input Bagging uses V -fold bootstrap replication, whereas Boosting uses V -fold weighted-bootstrap replications in the sense that the probability of each object to be included in the subsequent sample increases if the same object is misclassified by the weak learner. Random Forest uses at the same time V -fold bootstrap replication and a random selection of predictors. The latter ensemble is as accurate as Boosting, and it allows to obtain a ranking of variable importance. All of these algorithms aggregate the object decisions by voting, but none of these ensemble methods allows to preserve the final tree-structure: if we are interested in the accuracy of the prediction then we can use an ensemble because it is absolutely more accurate than a single decision tree, but the interpretation of the the tree-structure is irreparably lost because the aggregation process compromises the construction of a unique prediction tree structure.

5 Computational cost saving in ensemble algorithms

We used the same datasets to compare the accuracy of the AdaBoost algorithm (Freund and Schapire, 1997) if Two-Stage algorithm were used instead of the classical CART as weak learner for the ensemble procedure. In general, an ensemble of classifiers is more accurate than any of its individual members if they are accurate and diverse (Dietterich, 2000), i.e. if they have error rates lower than random choices and they provide different errors on new data points. Two-Stage is an accurate classifier and, as it belongs to the category of tree-based methods, is "diverse" because it is unstable. With respect to CART, it has a computational cost really lower. We performed 200 boosting iterations for each dataset by using the CART and the Two-Stage as weak learner. Table 2 shows a comparison between the two algorithms in terms of both computational cost and accuracy. The column named average computing cost shows, on average, the number of computations made by the CART and the Two-Stage to build a single tree. The column computational cost reduction shows the computational saving in using Two-Stage instead of CART. In example, in building the final boosted classifier for the Breast Cancer Wisconsin dataset with the Two-Stage, about 73% of computations were saved with respect to CART.

Dataset	Average computing cost		Computational cost reduction (%)	Badness of fit	
	CART	TWO-STAGE		CART	TWO STAGE
Breast Cancer Wisconsin	777	207	0,7336	0,1115	0,1312
Pima Indians Diabetes	15244	2158	0,8584	0,1633	0,1799
Credit (Cisia)	765	132	0,8275	0,1760	0,1690
Australian Credit	10746	456	0,9576	0,1427	0,1496
German Credit	17313	1271	0,9266	0,2433	0,2303
Simulation 1 (regression)	9862	839	0,9149	3,9274	3,7642
Simulation 2 (regression)	19402	2381	0,8733	11,3268	10,3526

Table 2 Computational cost using CART versus TWO-STAGE in boosting algorithm

The more interesting result is shown in the last two columns of the table, in which are reported the badness of fit measures for the final boosted classifiers. For the classification case we refer to the misclassification ratio as well as for the regression case we report the root mean squared error (the Boosting algorithm used for the regression problem was the one proposed by Grey and Poggi, 2006). The empirical evidence shows as these measures are absolutely similar, with the difference that if we use the Two-Stage algorithm, the same ensemble classifier is built by saving lots of computations. Another remark is that even if the Two-Stage is used in its worst working-situation (i.e. when all predictors are numerical as in the first two cases in the table), the ensemble classifier is pragmatically as accurate as the CART was been used.

6 Speeding up the BINPI algorithm

Incremental Non Parametric missing data Imputation (INPI) approach was introduced in a pioneer work by Siciliano and Conversano (2002). The basic assumption is that monotone missing data pattern are derived from the Missing At Random mechanism. The methodology considers a lexicographic ordering of the missing data in order to perform iteratively tree-based data imputation, yielding to a conditional mean estimation by non-parametric model. The resulting missing data imputation is performed incrementally because, in each iteration, the imputed values are added to the completely observed ones and concur to the imputation of the remaining missing data [Conversano and Siciliano, 2009]. Boosted Incremental Non Parametric Imputation (BINPI) [Siciliano *et al.*, 2006] provides more robust estimates with respect to the departure from the underlying model, and the incremental imputation of each variable at time (instead of each single data at time) allows a more efficient algorithm, thus reducing the computational cost of the overall procedure. It can be proved that this approach reduces the problem of error back-propagation due to the incremental procedure. Incremental algorithms have been also considered in data fusion [D'Ambrosio *et al.*, 2007], with the possibility to check the validity of pre-fusion conditions by a non parametric test [Aria and D'Ambrosio, 2008]. Robust Tree-based Incremental Imputation Methods for missing data imputation and data fusion have been widely discussed with empirical evidence on real and simulated data sets analyses [D'Ambrosio *et al.*, 2010]. To speed-up such robust methods based on ensembles it can be very useful to consider a tree-growing based on two-stage partitioning criterion as weak learner.

Acknowledgements The work for this paper is funded by Department of Mathematics and Statistics.

References

1. Aria M., D'Ambrosio A. A non parametric pre-grafting procedure for data fusion, *Proceedings of the MTISD 2008 (Metodi, Modelli e Tecnologie dell'Informazione a Supporto delle Decisioni)*, Coordinamento SIBA, Universit del Salento. (2008)
2. Berthold, M., Hand, D.J. *Intelligent Data Analysis*, Springer-Verlag (2003)
3. Breiman, L. Random Forests, *Machine Learning*, 45(1), 5-32. (2001)
4. Breiman, L. Bagging Predictors, *Machine Learning*, 26, 46-59. (1996)
5. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. *Classification and Regression Trees*. Wadsworth International Group, Belmont, California. (1984)
6. Conversano, C., Siciliano, R. Incremental Tree-Based Missing Data Imputation with Lexicographic Ordering, *Journal of Classification*, 26(3), 361-379. (2009)
7. D'Ambrosio A., Aria M., Siciliano R. Robust Tree-based Incremental Imputation Method for Data Fusion. *Lecture Notes in Computer Science 4723, Advances in Intelligent Data Analysis*, Springer-Verlag, pp 174-183. (2007)
8. D'Ambrosio A., Aria M., Siciliano R. Robust Incremental Tree-Based Methodology for Missing Data Imputation and Data Fusion, *submitted to Journal of Classification, under review* (2010)

9. D'Ambrosio, A., Heiser, W.J. Decision Trees for Preference Rankings, in *Classification and Data Analysis 2009, Book of Short Papers*, invited paper, CLEUP - Padova, 133-136. (2009)
10. Dietterich, T.G. Ensemble methods in machine learning. In J.Kittler and F.Roli, editors, multiple classifier system. *First International Workshop, MCS 2000, Cagliari, volume 1857 of lecture notes in computer science*. Springer-Verlag. (2000)
11. Freund, Y., and Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1). (1997)
12. Galimberti, G., Pillati, M., Soffritti, G. Robust regression trees based on M-estimators, *Statistica*, 2007(2), 150 - 165 (2007)
13. Gey, S., Poggi, J.M. Boosting and instability for regression trees. *Computational Statistics and Data Analysis*, 50, 533-550. (2006)
14. Giordano, G., Aria, M. Regression trees with moderating effects. *Springer series in New Perspectives in Statistical Modeling and Analysis*, forthcoming (2010)
15. Hastie, T.J., Tibshirani, R.J., Friedman, J.H. *The Elements of Statistical Learning*. Springer Verlag. (2001)
16. Krempel G., Hofer V. Partitioner Trees for Classification: A New Ensemble Method, in *Studies in Computational Intelligence*, vol. 254, 93-112, Springer Berlin/Heidelberg. (2009)
17. Mola, F., Siciliano, R. A two-stage predictive splitting algorithm in binary segmentation. In: Y. Dodge, J. Whittaker (Eds.): *Computational Statistics: COMPSTAT '92*, 1, Physica Verlag, Heidelberg (D), 179-184. (1992)
18. Mola, F., Siciliano, R. A Fast Splitting Procedure for Classification and Regression Trees, *Statistics and Computing*, 7, Chapman Hall, 208-216. (1997)
19. Mola, F., Siciliano, R. Discriminant Analysis and Factorial Multiple Splits in Recursive Partitioning for Data Mining, in Roli, F., Kittler, J. (eds.): *Proceedings of International Conference on Multiple Classifier Systems* (Chia, June 24-26, 2002), 118-126, Lecture Notes in Computer Science, Springer, Heidelberg. (2002)
20. Siciliano, R., Aria, M. Two-Class trees for non-parametric regression analysis. In Fichet, B., Piccolo, D., Verde, R. and Vichi, M. editors *Studies in Classification, Data Analysis and Knowledge Organization*, forthcoming
21. Siciliano, R., Conversano, C. Tree-based Classifiers for Conditional Missing Data Incremental Imputation, *Proceedings of the International Conference on Data Clean* (Jyvaskyla, May 29-31, 2002), University of Jyvaskyla. (2002)
22. Siciliano, R., Mola, F. A Fast Regression Tree Procedure, in A. Forcina, R. Hatzinger, G. Galmacci (eds.): *Statistical Modeling, Proceedings of the 11th International Workshop on Statistical Modeling*, Graphos Ed., 332-340. (1996)
23. Siciliano, R., Mola, F. Ternary Classification Trees: a Factorial Approach, in M. Greenacre, J. Blasius (Eds.): *Visualization of Categorical Data*, cap. 22, , Academic Press, San Diego (CA), 311-323. (1998)
24. Siciliano R., Mola F. Multivariate data analysis and modelling through classification and regression trees. *Computational Statistics & Data Analysis*. (2000)
25. Siciliano R., Aria., D'Ambrosio A. Boosted incremental tree-based imputation of missing data. *Data Analysis, Classification and the Forward Search. Springer series in Studies in Classification, Data Analysis, and Knowledge Organization*. Springer-Verlag, pp. 271-278. (2006)
26. Siciliano, R., Aria, M., D'Ambrosio, A. Posterior Prediction Modelling of Optimal Trees, in *Proceedings in Computational Statistics (COMPSTAT 2008)*, 18th Symposium Held in Porto, Portugal, Brito, Paula (Ed.), Springer-Verlag, pp. 323-334. (2008)
27. Tutore, V.A., Mooijaart, A. Optimal scaling trees. *Classification and data analysis 2007, book of short papers*, 359-362
28. Tutore, V.A., Siciliano, R., Aria, M. Conditional Classification Trees using Instrumental Variables, in *Lecture Notes in Computer Science 4723, Advances in Intelligent Data Analysis*, Springer-Verlag. (2007)