Selecting Decision Rules from Tree Ensembles

Damir Spisic[1], Jing Xu[2]
[1]IBM Watson Analytics, 71 S. Wacker Dr. 6th Fl., Chicago, IL 60606, USA
[2]IBM SPSS Predictive Analytics, 2&3 Floor, Building C, Outsourcing Park Phase I, No. 11 Jinye 1st Rd., High Tech Zone, Xi'an, China

Abstract

Tree ensembles are among the most popular and successful machine learning models due to their high prediction accuracy. Their shortcomings lie in difficulty of interpretation and drawing insights. While contributing tree models are easy to interpret, this transparency is lost when the tree models are combined into an ensemble. In this article, we describe a method to detect useful decision rules from a given tree ensemble. We exploit the fact that a tree ensemble offers a very large pool of interpretable decision rules. These decision rules can be used as basis for discovery of direct insights into important relationships supported by the ensemble. Novel metrics are proposed to select the top decision rules that are both the most interesting and consistent with the ensemble predictions. Interestingness that we consider is high prediction accuracy for categorical targets and high difference from the overall average for continuous targets. Consistency is defined in terms of predictions generated by a decision rule and the ensemble. We demonstrate the effectiveness of our approach through an example.

Key Words: decision rules, tree ensemble, model interpretation

1. Introduction

Tree ensembles combine predicted values from each tree within the ensemble by voting for categorical targets and by averaging for continuous targets. Different trees in an ensemble are usually generated using bagging, random forests or boosting methods. Bagging is based on re-sampling data records from training data set, random forests are based on re-sampling both records and attributes, while boosting is based on dynamically changing the record weights for generating each tree model. These and other similar tree ensemble methods improve model accuracy by reducing the prediction variance inherent to single tree models.

Tree models generate a number of decision rules that are easy to understand and apply. They often provide direct insights into important relationship. Unfortunately, the ease of interpretation for a single tree is lost when they are combined into an ensemble. Tree ensembles are usually more accurate than a single tree, but are very non-transparent from the user perspective. They offer no interpretable insights into important relationships supported by the data.

In order to bridge this gap, it is necessary to provide tools for interpreting interesting tree ensemble results. While there is much work available considering various decision rule generating systems, none of it provides a method to generate the most interesting rules

that are consistent with the tree ensemble predictions and where ensemble predictions are the most accurate. Interesting rules for categorical targets are the ones with high accuracy. For continuous target, we consider the rules that predict either high or low values to be the most interesting.

Shyr et al. [1] discuss insight discovery in a single tree model. Rules are grouped in high and low groups and unusual rules are detected. Our work discovers the most accurate or unusual rules obtained from all tree models in a given ensemble such that they are consistent with the ensemble predictions. On the other hand, paper by Liu et al [2] offers an example of an approach where random forest is replaced by a reduced set of decision rules approximating the entire ensemble model. Our goal is finding a subset of the most interesting rules consistent with the ensemble model.

In Section 2 we describe how candidate rules are based on the leaf nodes from generated trees in the ensemble. Sub-sections provide details of the proposed novel metrics that combine interestingness and conformance with ensemble predictions. The top decision rules ranked by the computed metrics are the most interesting rules that are consistent with the ensemble predictions. Also provided is a concrete example of obtaining the top decision rule from a tree ensemble. In Section 3 we state the conclusions.

2. Interesting Decision Rules

As described above, the ensemble technique can provide more accurate and stable predictions than a single classifier. On the other hand, an ensemble works as a black box, and its predictions can be difficult to interpret. This is also true when the base models are decision trees.

Our work aims to interpret predictions from a tree ensemble using concrete decision rules. Rather than interpreting some typical results of the ensemble model, we focus on extracting the most interesting rules that are valuable for users to learn. Wide variety of decision rules provided by the tree ensemble forms a foundation for interesting decision rules search.

Figure 1 illustrates a tree ensemble model where interesting decision rules corresponding to the leaf nodes are marked according to their interestingness indexes.

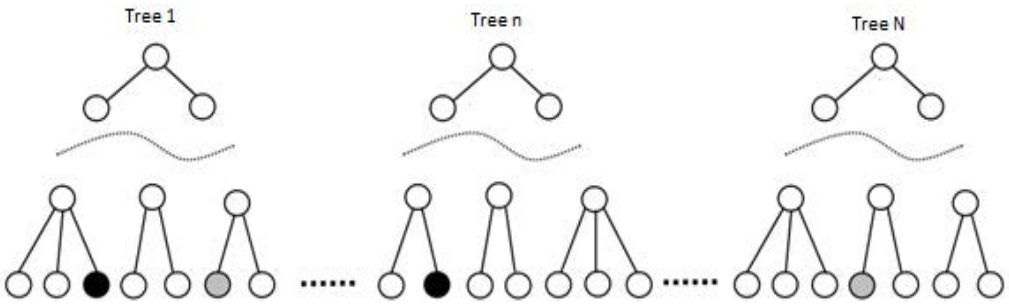

Figure 1: An example of tree ensemble with interesting decision rules corresponding to the leaf nodes

Our approach is data driven. In addition to the tree ensemble, we use an evaluation data. The general workflow is illustrated in Figure 2.

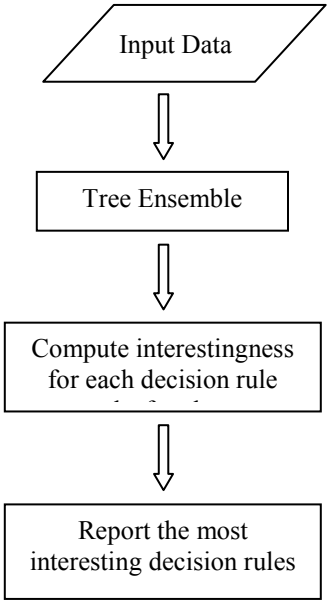

Figure 2: Workflow of detecting interesting decision rules from a tree ensemble

The interesting decision rules based on leaf nodes are defined as those which have high prediction accuracy and also high agreement with the predictions of the tree ensemble. We support both categorical and continuous targets. Depending on the measurement level of the target, a novel index measure is defined to quantify interestingness of the corresponding decision rules.

2.1 Classification

Classification models predict categorical targets. Their prediction is accurate when the predicted category is the same as the observed category for the given data record. We use the following notation in order to describe details of our proposal.

Table 1: Notation for Interesting Rules

t	A leaf node from a decision tree in the given decision tree ensemble
R_t	Decision rule based on the node t (e.g. the most frequent class is predicted)
$E(t)$	Event that the ensemble prediction is accurate on a data record determined by node t
$\bar{E}(t)$	Event that the ensemble prediction is inaccurate on a data record determined by node t
$R(t)$	Event that the prediction rule R_t is accurate on a data record determined by node t
$\bar{R}(t)$	Event that the prediction rule R_t is inaccurate on a data record determined by node t
$P(\cdot)$	Probability of an event

The following procedure is used for detection of interesting rules R_t.

1. Let each record in the evaluation dataset traverse all the base trees, and determine the data records for each leaf node t.
2. For each leaf node t, repeat
   a) Estimate $P(R(t))$ using the proportion of data records determined by the node t where rule R_t is accurate.
   b) Estimate $P(E(t))$ using the proportion of data records determined by the node t where the ensemble model is accurate. Clearly, we have $P(\bar{E}(t)) = 1 - P(E(t))$.
   c) Estimate $P(E(t)R(t))$ using the proportion of data records determined by the node t that are predicted accurately by both the ensemble model and the rule R_t.
   d) Estimate $P(\bar{E}(t)\bar{R}(t))$ using the proportion of data records determined by the node t that are predicted inaccurately by both the ensemble model and the rule R_t.
   e) Compute the first sub-index of interestingness on prediction agreement between the ensemble model and the rule R_t as $I_1^t = P(E(t)R(t)) + P(\bar{E}(t)\bar{R}(t))$.
   f) Compute the second sub-index of interestingness on the rule R_t accuracy $I_2^t = P(R(t))$.
   g) Compute the third sub-index of interestingness on the ensemble model accuracy $I_3^t = P(E(t))$.
   h) Compute the overall interestingness index as the product of sub-indices $I_t = I_1^t * I_2^t * I_3^t$.
3. Report a specified number of top decision rules R_t with the highest interestingness index I_t. Optionally, one can report all rules whose interestingness index is larger than a specified threshold.

Note that in order to make the decision rules more representative, it requires each leaf node to have a reasonable number of records. Several methods can be applied. One is to simply exclude small leaf nodes from selection. Another method is to collapse small leaf nodes into their parent nodes so that larger collapsed leaf nodes can be obtained.

2.2 Regression

Probability has been used to derive the interestingness index for classification models, but it is not suitable for regression models where the target is continuous. We propose different measures to evaluate interestingness of decision rules for continuous targets.

Per conformance requirement of the ensemble model and the decision rule R_t, we expect that the predictions from the ensemble model and the rule R_t are as similar as possible. This can be quantified by several measures based on prediction differences on data records determined by the node t. Typical measures include mean square error and mean absolute error. Smaller measure values signify improved conformance between the ensemble and the decision rule.

Main shortcoming of these measures is that they depend on the scale of the target variable. It is necessary to standardize the used measure in order to obtain useful interestingness sub-index which ranges in values between 0 and 1. A simple standardizing strategy is to divide the measure value for each leaf node by the

corresponding mean or median of measure values across all leaf nodes. For the nodes with large measure values, the standardized measure value could be larger than 1. For such leaf nodes, we simply truncate the measure value at 1. Given the standardized measure value $D(t)$, we define the first sub-index of interestingness based on the conformance requirement as $I_1^t = 1 - D(t)$.

For the second sub-index of interestingness on the decision rule predictions, we consider the rules that predict either high or low values of the target variable to be the most interesting. This can be achieved by comparing each decision rule predictions with the root node predictions.

Two independent samples t-test are used for the comparison of the decision rules based on a leaf node and the root node. We consider a decision rule to be more interesting if its prediction distribution is more significantly different from the root node predictions. Moreover, we can compute an effect size measure corresponding to the t-test, and transform the effect size measure into an interestingness sub-index with values between 0 and 1. Appropriate interpolation methods can be used for this purpose. For example, assume that the p-value from the t-test is $p_{value}(t)$, the effect size measure is $E_s(t)$, and $f(x)$ is a monotone cubic interpolation function. Then we define the second sub-index of interestingness as

$$I_2^t = \begin{cases} 0, & p_{value}(t) \geq sig \\ f(E_s(t)), & p_{value}(t) < sig \end{cases},$$

where sig is the specified significance level (default 0.05).

The third sub-index of interestingness is an indicator of the ensemble model accuracy. A fairly straightforward choice is to compute R squared metric for the ensemble model on data records determined by the node t. R squared is computed as 1 minus the relative error, where relative error is given by the sum of squared errors for the ensemble model divided by the total sum of squares. R squared has values between 0 and 1, where 1 indicates a perfect fit. We therefore specify $I_3^t = R^2(t)$.

In general, we define the interestingness index for decision rule at leaf node t as $I_t = I_1^t * I_2^t * I_3^t$. Accordingly, we report the most interesting decision rules using the same criteria as described for classification models.

2.3 Example

We use a data sample to demonstrate effectiveness of our approach. The data sample, tree_credit.sav, defines a classification problem: predicting credit rating level (bad or good) by demographic and behavior characteristics of customers such as age, income, car loans etc. This data is available in IBM SPSS Modeler installation directory.

For this experiment, the data is randomly divided into a training (50%) and a testing (50%) partition. We build a CART tree model and a random forest model respectively using the training partition. Then we extract decision rules from the two models, and evaluate them on the testing partition.

The rules in the table below are obtained from the CART model. They are ranked by their prediction accuracy.

Table 2: Rules Based on CART Model

Rule #	Decision Rule	Rule Accuracy
1	(Income > Low) and (Credit_cards = Less than 5)	92.8%
2	(Income = Low) and (Credit_cards = 5 or more)	89.5%
3	(Income = High) and (Credit_cards = 5 or more)	81.8%
4	(Age <= 30) and (Income = Medium) and (Credit_cards = 5 or more)	74.9%
5	(Income = Low) and (Credit_cards = Less than 5)	70.0%
6	(Age > 30) and (Income = Medium) and (Credit_cards = 5 or more)	56.4%

In contrast, rules in the next table are obtained from a random forest model according to the proposed method.

Table 3: Rules Based on Random Forest Model

Rule #	Decision Rule	Rule Accuracy	Forest Accuracy	Interestingness Index
1	(Age > 33) and (Income > Medium) and (Credit_cards = {Less than 5})	99.0%	99.0%	0.98
2	(Income > Medium) and (Age > 41) and (Car_loans = {More than 2})	93.9%	93.9%	0.88
3	(Income > Medium) and (Age > 41) and (Credit_cards = {5 or more})	92.8%	92.8%	0.86

The rules detected from the random forest achieve much higher prediction accuracy comparing with the ones from the single tree. There are multiple reasons. One is that tree models are unstable and thus less likely to generate high quality rules by themselves. In contrast, random forest includes large number of trees (about 100) that provide much bigger set of rules to search from. Moreover, the search is guided so that rule predictions match ensemble predictions. This helps to avoid overfitting, and improves generalization of detected rules.

3. Conclusions

Proposed method succeeds in finding interesting rules conforming to tree ensembles. Rules based on an ensemble can be considerably more accurate than the rules from a single decision tree. In addition, they provide insights into ensemble model and also benefit from being consistent with the ensemble predictions.

There are additional challenges to consider. Rules detected by this method can frequently overlap. In order to remove redundant results, it is necessary to extract a set of relatively independent representatives of the most interesting rules. Another suggested direction has been to consider creating rules that are consistent with a given model from scratch. This approach would be applicable to any predictive model, not only to ensembles based on decision trees.

References

[1] Shyr, J., Chu, J. and Zhong, W.C. (2013). Insight Discovery for Decision Tree Models. In *JSM Proceedings*, Section on Statistical Learning and Data Mining. Alexandria, VA: American Statistical Association. 2571-2579.

[2] Liu, S., et al. (2014). Learning accurate and interpretable models based on regularized random forests regression. *BMC Systems Biology*, 8(Suppl 3):S5.