Insight Discovery for Decision Tree Models

Jing Shyr [1], Jane Chu[1], Weicai Zhong[2]
[1]IBM SPSS Predictive Analytics, 200 W. Madison St. 23th 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

The decision tree model is a popular data mining tool in predictive analytics. The goal of building it in most applications is for prediction only. The question of identifying which leaf nodes have different target distributions from the root node remains unanswered. Such a missing part should provide further insights and understanding into the predictive structure of the data while it is often overlooked by a tree model user. It might be possible to find some of such leaf nodes in checking the tree diagram when the number of leaf nodes is small or the target distributions between leaf nodes and the root node are very different. However, it becomes more challenging or even impossible when there exist hundreds of leaf nodes or the target distributions between leaf nodes and the root node are not that different. In this paper we propose a systematic and efficient system to identify these leaf nodes based on several tests and present the results in an intuitive way with graphs and texts so it is easy for a tree model user to discover insights. In addition, all tests are based on already computed statistics in the leaf nodes, therefore there is little extra computational cost.

Key words: data mining, decision tree, tree diagram, data insight.

1. Introduction

The decision tree building process is to create a model that can predict the value of a target based on a set of predictor variables. It is a popular data mining tool for classification and prediction in business analytics because it is easy to interpret and simple to understand as it doesn't require many model assumptions like traditional parametric models have. These tree models can be regression trees (for continuous targets) and classification trees (for categorical targets). In these tree structures, each leaf (terminal) node represents a segment of the data so it provides a distribution of the target while the segments are defined by the values of the predictors by the path from the root to the leaf. Thus a path from the root to the leaf can be viewed as a rule or a customer profile and a whole tree represents a set of rules or a set of customer profiles.

After a tree is built, the output usually includes a tree diagram and some summary statistics. A tree diagram displays how the data is classified into different profiles. It is usually used to discover the differences among leaf nodes. The statistics for each leaf node includes the target mean and average squared error for the continuous target and the frequency percentage for each target category for the categorical target. Since building a decision tree in most applications focuses mainly on predicting the target value for each node and the overall accuracy of the tree, see Rokach and Maimon (2005), these statistics are used for the prediction purpose only. The question of identifying which leaf nodes

have different distributions from the root node remains unanswered. Such a missing part should provide further insights and understanding into the predictive structure of the data while it is often overlooked by a tree model user.

Let's use an example to illustrate. Figure 1 is a tree diagram which is generated by IBM SPSS Statistics software, see IBM (2012), after building a tree model on the Titanic data. It shows survival state of passengers (dead/alive) based on some passengers' characteristics, such as sex, travel class and age group. There are 11 nodes: Node 0 is the root node and 7 of them are the leaf nodes (Nodes 3, 5, 6, 7, 8, 9, 10) which represent 7 different passenger profiles. A user usually checks these leaf nodes to discover which passenger profiles are more likely to be alive or dead and understand the data structure. For example, Female passengers with the first and second travel classes have much higher percentages to be alive (96.6% for Node 6 and 87.7% for Node 7), but female passengers with the third travel class have less than 50% changes to be alive (45.9% for Node 8).

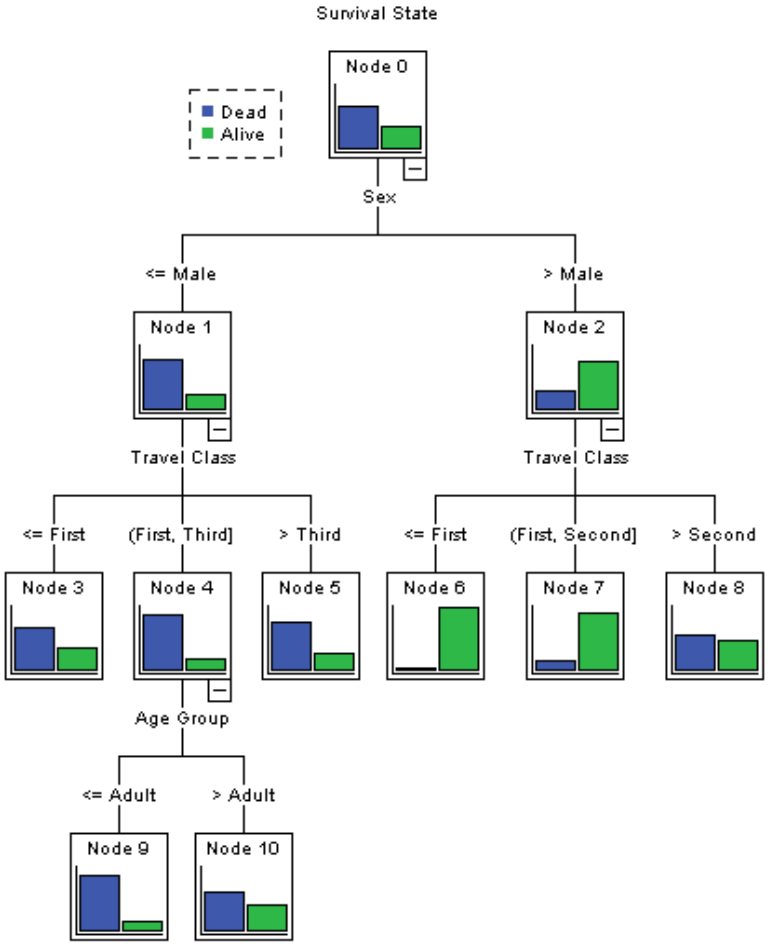

FIG. 1 The tree diagram showing passengers' survival state on the Titanic data

Now let's use the tree diagram to compare the target distributions in the leaf nodes and the root node. It is easy to spot Nodes 6 and 7 have completely different target

distributions from Node 0 because the proportions of survival state (dead/alive) of the nodes are opposite, but it is not easy to decide whether other leaf nodes have different distributions from the root node. It becomes more challenging or even impossible when the target has more than 2 categories and/or there exist hundreds of leaf nodes by checking the leaf nodes in the tree diagram alone. Hence a systematic and efficient method is needed to determine which leaf nodes have different target distributions from the root node and present the results to the user.

For the continuous target in a regression tree, testing target distributions between leaf nodes and the root node is usually reduced to testing target means. So we can extend the analyses proposed in Shyr, et al (2011) by considering the leaf nodes in a regression tree as categories of a predictor in a linear regression model. For the categorical target in a classification tree, testing target distribution between leaf nodes and the root node can be based on Pearson's one-sample chi-squared test, see Agresti (2002).

The rest of this paper is organized as follows: Section 2 will describe the decision tree insight discovery system in details while a few concluding remarks are in Section 3.

2. Decision Tree Insight Discovery System

Given a decision tree built on a data set with a target of interest and a set of predictors and it doesn't matter which tree building method is used, the tree insight discovery system proposes some node-by-node analyses based on some basic statistics in order to provide some insights in addition to the usual prediction purpose. The system includes two parts:

(1) Summary Computation: it extracts and summarizes useful information from the statistics in the leaf nodes based on some statistical tests.
(2) Insight Presentation: it displays the results from the Summary Computation part in both graphical and text insights to the user.

Depending on the target is continuous or categorical, the analyses and insights would be different. We will describe these analyses and insights in the next two subsections in details.

2.1 For Regression Trees (Continuous Targets)

Suppose a regression tree is built for the target *Y* with *M* leaf nodes and the following statistics for all leaf nodes and also the root node are computed:

N_i	The number of records for the i^{th} leaf node, $i = 1, \cdots, M$
$\bar{y}_i$	The observed target mean for the i^{th} leaf node, $i = 1, \cdots, M$ $\bar{y}_i = \frac{1}{N_i} \sum_{k=1}^{N_i} y_{ik}$ where y_{ik} is the target value of the k^{th} record in the i^{th} leaf node.
s_i	The standard error of $\bar{y}_i$ for the i^{th} leaf node, $i = 1, \cdots, M$ $s_i = \sqrt{\frac{1}{N_i(N-M)} \sum_{i=1}^{M} \sum_{k=1}^{N_i} (y_{ik} - \bar{y}_i)^2}$ where $N = \sum_{i=1}^{M} N_i$ is the total sample size.

$\bar{y}$	The overall target mean or the observed target mean for the root node, $\bar{y} = \frac{1}{N}\sum_{i=1}^{M}\sum_{k=1}^{N_i} y_{ik}$

First Summary Computation part would receive the above statistics and two analyses would be conducted:

(a) Grouping analysis:
Under the assumption that the overall target mean is known and fixed, it will partition all leaf nodes into a high group and a low group (with a possible medium group) by the following steps:

- Sort the leaf nodes by their observed target means, $\bar{y}_1, \cdots, \bar{y}_M$, in descending order. Without loss of generality, assume $\bar{y}_1 \geq \cdots \geq \bar{y}_M$.

- Conduct a one-sample t-test for the leaf node with the largest target mean with the hypothesis $H_0 : \bar{y}_1 = \bar{y}$ vs. $H_A : \bar{y}_1 > \bar{y}$. We use the one-tailed test because it will provide more power. The *t* statistic and its corresponding *p*-value are

$$t = \frac{\bar{y}_1 - \bar{y}}{s_1}$$

$$p = 1 - \text{Prob}(t_{(N-M)} \leq t)$$

where $t_{(N-M)}$ is a random variable which following a *t* distribution with $(N - M)$ degrees of freedom. Here we use a model-based method to compute the standard errors for the more robust test results hence the degrees of freedom would be the same for all t-tests.

If $p < \alpha$, where α is a significant level (the default is 0.05), then the high group is formed by including the leaf node with the largest target mean.

- Repeat the same test for the node with the second highest target mean ($\bar{y}_2$), then for the node with the third highest target mean ($\bar{y}_3$), etc. until there is no leaf node can be added into the high group. I.e., the test is not significant when $p \geq \alpha$.

- Similarly, conduct a one-sample t-test for the leaf node with the smallest target mean with the hypothesis $H_0 : \bar{y}_M = \bar{y}$ vs. $H_A : \bar{y}_M < \bar{y}$. The *t* statistic and its corresponding *p*-value are

$$t = \frac{\bar{y}_M - \bar{y}}{s_M}$$

$$p = 1 - \text{Prob}(t_{N-M} \leq |t|)$$

If $p < \alpha$, then the low group is formed by including the leaf node with the smallest target mean.

- Repeat the same tests for the nodes with $\bar{y}_{M-1}$, $\bar{y}_{M-2}$, etc. until there is no leaf node can be added into the low group.

- If there still exist some leaf nodes after forming the high and low groups, they are grouped into the medium group.

- The results are a list of leaf nodes for the high, low and medium groups with relevant test statistics.

(b) Unusual node detection analysis:
It detects possible unusual nodes in the high and low groups based on an outlier detection method with the following steps:

- Calculate the median of M target means, $\bar{y}_1, \cdots, \bar{y}_M$, incorporating the number of records in each leaf node. Denote the median as *MED*,

$$MED = \text{median}\left(\bar{y}_{1,N_1}, \cdots, \bar{y}_{M,N_M}\right),$$

where $\bar{y}_{i,N_i}$ is a set which contains $\bar{y}_i$ value with N_i of them, for $i = 1, \cdots, M$.

- Calculate the median absolute deviation (*MAD*) of M target means, again incorporating with the number of records in each cell

$$MAD = \text{median}\left(|\bar{y}_1 - MED|_{N_1}, \cdots, |\bar{y}_M - MED|_{N_M}\right),$$

where $|\bar{y}_i - MED|_{N_i}$ is a set which contains $|\bar{y}_i - MED|$ value with N_i of them, for $i = 1, \cdots, M$.

- Compute the modified z-score for the i^{th} leaf node, $i = 1, \cdots, M$

$$z_i = \begin{cases} \dfrac{\bar{y}_i - MED}{1.4826 \times MAD} & \text{if } MAD \neq 0 \\ \dfrac{\bar{y}_i - MED}{1.2533 \times MeanAD} & \text{if } MAD = 0 \end{cases}$$

where $MeanAD = \frac{1}{N}\sum_{i=1}^{M} N_i |\bar{y}_i - MED|$.

- Detect unusual nodes:

If $z_i > 3$, the leaf node *i* has an unusually high target mean in the high group.
If $z_i < -3$, the leaf node *i* has an unusually low target mean in the low group.

- The results are a list of unusual nodes in the high and low groups with relevant test statistics.

Second the Insight Presentation part would receive the output from the Summary Computation part and display some graphical and text insights. For instance, the test results for grouping analysis and text insights can be displayed in the following table:

Test Results for Grouping Analysis

Group	Leaf node #	t[a]	Sig
High	1	#.##	.##
	5	#.##	.##
Low	6	#.##	.##
	3	#.##	.##
	8	#.##	.##

[a] Each t has df= $N - M$.

Any leaf nodes in the medium group are unlisted.

For this data set, there's an interesting grouping of leaf nodes with higher target means versus those with lower target means.

And the test results for unusual node detection analysis and text insights can be displayed in the following table:

Test Results for Unusual node detection analysis

Type	Leaf node #	Modified z
Unusually High	1	#.##
	5	#.##
Unusually Low	8	#.##
	3	#.##

These leaf nodes have very different target means from the overall mean.

For graphical insights, we can highlight the leaf nodes in the high and low groups in different colors in the tree diagram. Similar can be done for unusually high and low leaf nodes.

2.2 For Classification Trees (Categorical Targets)

Suppose a classification tree is built for the target *Y* (having *J* target categories, denoted as $\{1, \cdots, J\}$) with *M* leaf nodes and the following statistics for all leaf nodes and also the root node are computed:

N_i	The number of records for the i^{th} leaf node, $i = 1, \cdots, M$
N_{ij}	The number of records in the j^{th} target category for the i^{th} leaf node, $i = 1, \cdots, M$ and $j = 1, \cdots, J$.
$p_{\cdot j}$	The overall probability of the j^{th} target category, $j = 1, \cdots, J$ from the whole sample data.

First Summary Computation part would receive the above statistics and two analyses would be conducted:

(a) Grouping analysis:
Under the assumption that the overall target distribution is known and fixed, it will partition all leaf nodes into two groups: a significant group and an insignificant group by the following steps:

- Compute the Person's one sample chi-squared statistic and the corresponding *p*-value for each leaf node i, $i = 1, \cdots, M$:

$$\chi_i^2 = \sum_{j=1}^{J} \frac{(N_{ij} - p_{\cdot j} \times N_i)^2}{p_{\cdot j} \times N_i}$$

$$p = 1 - Pr\,(\chi_{(J-1)}^2 \leq \chi_i^2)$$

where $\chi_{(J-1)}^2$ is a random variable which following a chi-squared distribution with $(J-1)$ degrees of freedom.

If $p < \alpha$, where α is a significant level (the default is 0.05), then the leaf node has a significantly different distribution from the overall distribution and will be added into the significant group.

If $p \geq \alpha$, then the leaf node will be assigned to the insignificant group.

- The results are a list of the leaf nodes in the significant group with relevant test statistics.

(b) Influential target category analysis:
It identifies target categories, which have significantly large frequency differences from that of the root node, based on another chi-squared test in the significant group (suppose it is G) by the following steps:

- Compute the chi-squared statistic and the corresponding *p*-value for the leaf node $i \in G$ and for the j^{th} target category:

$$\chi_{ij}^2 = \frac{(N_{ij} - p_{\cdot j} \times N_i)^2}{N_i \times p_{\cdot j} \times (1 - p_{\cdot j})}$$

$$p = 1 - Pr\,(\chi_1^2 \leq \chi_{ij}^2)$$

where χ_1^2 is a random variable which following a chi-squared distribution with 1 degree of freedom.

If $p < \alpha$, where α is a significant level (the default is $0.05/J$ based on the Bonferroni adjustment method), then the j^{th} target category is an influential target category for the leaf node $i \in G$.

- The results are a list of influential target categories for each significant leaf node with relevant test statistics.

Second the Insight Presentation part would receive the output from the Summary Computation part and display some graphical and text insights. For instance, the test results for Grouping analysis and text insights can be displayed in the following table:

Distribution Test Results

Leaf node #	Chi-square [a]	Sig
1	#.##	.##
5	#.##	.##

[a] Each chi-square has df=J – 1.

Each chi-squared is significant at the .05 level, so reject the null hypothesis that the target distribution at the given node equals the overall distribution.

Some leaf nodes have distributions that are statistically different from the overall distribution of the target.

And the test results for Influential target category analysis and text insights can be displayed in the following table:

Influential Target Category Test

Category	Leaf node #	Chi-square [a]	Sig
A	1	#.##	.##
A	5	#.##	.##
B	1	#.##	.##
C	5	#.##	.##

[a] Each chi-square has df=1.

Each chi-square is significant at the .05 level, so reject the null hypothesis that that the given target category has the same probability in the conditional distribution and the overall distribution.

These target categories have the strongest influence on the comparison with the overall distribution.

For graphical insights, we can highlight the leaf nodes in the significant group and use a glow effect arrow to show the direction of influence (up if the influential target category is higher than expected, down if lower) in the tree diagram.

3. Conclusion

The proposed Decision Tree Insight Discovery system presents insights into a regression tree and a classification tree by comparing the target distribution between leaf nodes and the root node. It has some unique features: it releases the human's burden of extracting the insights from a tree diagram, especially with hundreds of nodes; and there is little extra computational cost.

References

Agresti, A. (2002), *Categorical Data Analysis*, Second Edition, Hoboken, NJ: John Wiley & Sons, Inc.

IBM Inc. (2012), Tree Algorithms, *IBM SPSS Statistics 21 Algorithms*, Chicago, IL, 906–940.

Rokach, L. and Maimon, O. (2005), Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, Part C, 35 (4): 476–487.

Shyr, J., Chu, J. and Han, S. (2011), Category profiling and unusual category detection based on Estimated Marginal Means (EMMEANS). *JSM Proceedings, Social Statistics Section*, Alexandria, VA: American Statistical Association. 4289–4300.