






3.1 Autonomous Tree Growth Options (Self-Study)



Autonomous Decision Tree Defaults

Default Settings


 Maximum Branches	2
 Splitting Rule Criterion	Logworth
 Subtree Method	Average Profit
 Tree Size Options	Bonferroni Adjust Split Adjust Maximum Depth Leaf Size

108 ...


The behavior of the tree algorithm in SAS Enterprise Miner is governed by many parameters that can be divided into four groups:

- the number of splits to create at each partitioning opportunity
- the metric used to compare different splits
- the method used to prune the tree model
- the rules used to stop the autonomous tree growing process


The defaults for these parameters generally yield good results for an initial prediction.




Tree Variations: Maximum Branches



Complicates split search



Trades height for width




Uses heuristic shortcuts

109
...

SAS Enterprise Miner accommodates a multitude of variations in the default tree algorithm. The first involves the use of multiway splits instead of binary splits. Theoretically, there may not be a clear advantage in doing multiway splits. However, for a given data set, an analyst can try and explore if multiway splits produce better results than 2-way splits.

The inclusion of multiway splits complicates the split-search algorithm. A simple linear search becomes a search whose complexity increases geometrically in the number of splits allowed from a leaf. To combat this complexity explosion, the Tree tool in SAS Enterprise Miner resorts to heuristic search strategies.



Tree Variations: Maximum Branches

Property	Value
Train	
Variables	...
Interactive	...
Import Tree Model	No
Tree Model Data Set	...
Use Frozen Tree	No
Use Multiple Targets	No
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	...
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000


Maximum branches in split

Exhaustive search size limit


110
...

Two fields in the Properties panel affect the number of splits in a tree. The Maximum Branch property sets an upper limit on the number of branches emanating from a node. When this number is greater than the default of 2, the number of possible splits rapidly increases. To save computation time, a limit is set in the Exhaustive property as to how many possible splits are explicitly examined. When this number is exceeded, a heuristic algorithm is used in place of the exhaustive search described above.


The heuristic algorithm alternately merges branches and reassigns consolidated groups of observations to different branches. The process stops when a binary split is reached. Among all considered candidate splits, the one with the best worth is chosen. The heuristic algorithm initially assigns each consolidated group of observations to a different branch, even if the number of such branches is more than the limit allowed in the final split. At each merge step, the two branches that degrade the worth of the partition the least are merged. After the two branches are merged, the algorithm considers reassigning consolidated groups of observations to different branches. Each consolidated group is considered in turn, and the process stops when no group is reassigned.

The SAS logo, consisting of the letters "sas" in a stylized font, with the tagline "THE POWER TO KNOW." to its right.


Tree Variations: Splitting Rule Criterion

A small icon showing a grey rectangular area with a red line graph plotted on it, representing a splitting rule criterion.

Yields similar splits

A small icon showing a blue tree with a blue arrow pointing towards it, representing a tree structure.

Grows enormous trees

A small icon showing a square with a blue and yellow gradient, representing a multi-level input.

Favors many-level inputs

111

In addition to changing the number of splits, you can also change how the splits are evaluated in the split-search phase of the tree algorithm. For categorical targets, SAS Enterprise Miner offers three separate split-worth criteria. Changing from the chi-squared default criterion typically yields similar splits if the number of distinct levels in each input is similar. If not, the other split methods tend to favor inputs with more levels due to the multiple comparison problem discussed above. You can also cause the chi-squared method to favor inputs with more levels by turning off the Bonferroni adjustments.

Because Gini reduction and entropy reduction criteria lack the significance threshold feature of the chi-squared criterion, they tend to grow enormous trees. Pruning and selecting a tree complexity based on validation profit limit this problem to some extent.

sas THE POWER TO KNOW

Tree Variations: Splitting Rule Criterion

Property	Value
Splitting Criterion	ProbF
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChsq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	1
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Scoring	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Imports	
Observation Based Imports	No
Number Single Var Imports	5
Logworth Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	

Split Criterion

Categorical Criteria

- Chi-square logworth
- Entropy
- Gini

Interval Criteria

- Variance
- Prob-F logworth

Logworth adjustments

112

A total of five choices in SAS Enterprise Miner can evaluate split worth. Three (chi-squared logworth, entropy, and Gini) are used for categorical targets, and the remaining two (variance and ProbF logworth) are reserved for interval targets.

Both chi-squared and ProbF logworths are adjusted (by default) for multiple comparisons. It is possible to deactivate this adjustment.

The split worth for the entropy, Gini, and variance options are calculated as shown below. Let a set of cases S be partitioned into p subsets S_1, \dots, S_p so that

$$S = \bigcup_{i=1}^p S_i$$

Let the number of cases in S equal N and the number of cases in each subset S_i equal n_i . Then the worth of a particular partition of S is given by the following:

$$worth = I(S) - \sum_{i=1}^p w_i I(S_i)$$


where $w_i = n_i/N$ (the proportion of cases in subset S_i), and for the specified split worth measure, $I(\cdot)$ has the following value:

$$I(\cdot) = \sum_{\text{classes}} p_{\text{class}} \cdot \log_2 p_{\text{class}} \quad \text{Entropy}$$


$$I(\cdot) = 1 - \sum_{\text{classes}} p_{\text{class}}^2 \quad \text{Gini}$$

$$I(\cdot) = \frac{1}{N} \sum_{\text{node cases}} (y_{\text{case}} - \bar{Y})^2 \quad \text{Variance}$$

Each worth statistic measures the change in $I(S)$ from node to branches. In the Variance calculation, \bar{Y} is the average of the target value in the node with Y_{case} as a member.



Tree Variations: Subtree Method




Controls Complexity

113
...

SAS Enterprise Miner features two settings for regulating the complexity of subtree: modify the pruning process or the Subtree method.

By changing the model assessment measure to average square error, you can construct what is known as a *class probability tree*. It can be shown that this action minimizes the imprecision of the tree. Analysts sometimes use this model assessment measure to select inputs for a flexible predictive model such as neural networks.

You can deactivate pruning entirely by changing the subtree to **Largest**. You can also specify that the tree be built with a fixed number of leaves.



Tree Variations: Subtree Method

Property	Value
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Import	
Observation Based Import	No
Number Single Var Import	5
P-Value Adjustment	
Bonferroni Adjustment	Yes

Assessment
Largest
N


Pruning options

Pruning metrics

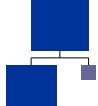
Decision
Average Square Error
Misclassification
Lift

114
...


The pruning options are controlled by two properties: Method and Assessment Measure.




Tree Variations: Tree Size Options



Avoids orphan nodes




Controls sensitivity



Grows large trees

115
...

The family of adjustments that you modify most often when building trees are the rules that limit the growth of the tree. Changing the minimum number of observations required for a split search and the minimum number of observations in a leaf prevents the creation of leaves with only one or a handful of cases. Changing the significance level and the maximum depth allows for larger trees that can be more sensitive to complex input and target associations. The growth of the tree is still limited by the depth adjustment made to the threshold.



Tree Variations: Tree Size Options

Property	Value
Splitting rules	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChsq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Leaf	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Imports	
Observation Based Imports	No
Number Single Var Imports	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	Before
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes
Output Variables	

Logworth threshold

Maximum tree depth

Minimum leaf size

Threshold depth adjustment

116

Changing the logworth threshold changes the minimum logworth required for a split to be considered by the tree algorithm (when using the chi-squared or ProbF measures). Increasing the minimum leaf size avoids orphan nodes. For large data sets, you might want to increase the maximum leaf setting to obtain

additional modeling resolution. If you want big trees and insist on using the chi-squared split-worth criterion, deactivate the Split Adjustment option.