Chapter 8
Trained Decision Trees for a More Meaningful Accuracy (150 Patients)

General Purpose

Traditionally, decision trees are used for finding the best predictors of health risks and improvements (Chap. 53). However, this method is not entirely appropriate, because a decision tree is built from a data file, and, subsequently, the same data file is applied once more for computing the health risk probabilities from the built tree. Obviously, the accuracy must be close to 100 %, because the test sample is 100 % identical to the sample used for building the tree, and, therefore, this accuracy does not mean too much. With neural networks this problem of duplicate usage of the same data is solved by randomly splitting the data into two samples, a training sample and a test sample (Chap. 12 in Machine learning in medicine part one, pp 145–156, Artificial intelligence, multilayer perceptron modeling, Springer Heidelberg Germany, 2013, from the same authors). The current chapter is to assess whether the splitting methodology, otherwise called partitioning, is also feasible for decision trees, and to assess its level of accuracy. Decision trees are both appropriate for data with categorical and continuous outcome (Chap. 53).

Primary Scientific Question

Can inflammatory markers adequately predict pneumonia severities wit the help of a decision tree. Can partitioning of the data improve the methodology and is sufficient accuracy of the methodology maintained.

Example

Four inflammatory markers (CRP (C-reactive protein), ESR (erythrocyte sedimentation rate), leucocyte count (leucos), and fibrinogen) were measured in 150 patients. Based on x-ray chest clinical severity was classified as A (mild infection), B (medium severity), C (severe infection). A major scientific question was to assess what markers were the best predictors of the severity of infection.

CRP	leucos	fibrinogen	ESR	x-ray severity
120,00	5,00	11,00	60,00	A
100,00	5,00	11,00	56,00	A
94,00	4,00	11,00	60,00	A
92,00	5,00	11,00	58,00	A
100,00	5,00	11,00	52,00	A
108,00	6,00	17,00	48,00	A
92,00	5,00	14,00	48,00	A
100,00	5,00	11,00	54,00	A
88,00	5,00	11,00	54,00	A
98,00	5,00	8,00	60,00	A
108,00	5,00	11,00	68,00	A
96,00	5,00	11,00	62,00	A
96,00	5,00	8,00	46,00	A
86,00	4,00	8,00	60,00	A
116,00	4,00	11,00	50,00	A
114,00	5,00	17,00	52,00	A

CRP = C-reactive protein (mg/l)
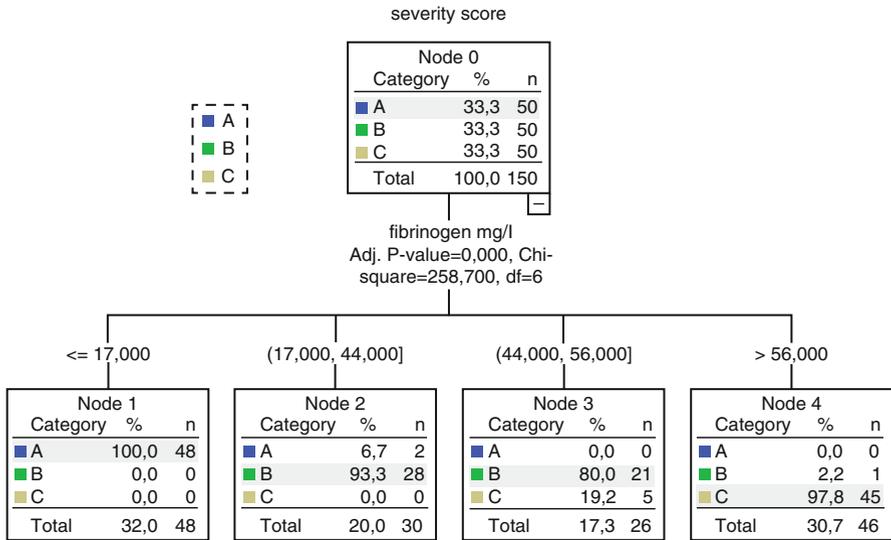leucos = leucyte count (*10^9/l)
fibrinogen = fibrinogen level (mg/l)
ESR = erythrocyte sedimentation rate (mm)
x-ray severity = x-chest severity pneumonia score (A–C = mild to severe)

The first 16 patients are in the above table, the entire data file is in "decisiontree" and can be obtained from "extras.springer.com" on the internet. We will start by opening the data file in SPSS.

Command:

click Classify….Tree….Dependent Variable: enter severity score….Independent Variables: enter CRP, Leucos, fibrinogen, ESR….Growing Method: select CHAID….click Output: mark Tree in table format….Criteria: Parent Node type 50, Child Node type 15….click Continue…. ….click OK.

severity score

Node 0		
Category	%	n
■ A	33,3	50
■ B	33,3	50
■ C	33,3	50
Total	100,0	150

■ A
■ B
■ C

fibrinogen mg/l
Adj. P-value=0,000, Chi-square=258,700, df=6

| <= 17,000 | (17,000, 44,000] | (44,000, 56,000] | > 56,000 |

Node 1		
Category	%	n
■ A	100,0	48
■ B	0,0	0
■ C	0,0	0
Total	32,0	48

Node 2		
Category	%	n
■ A	6,7	2
■ B	93,3	28
■ C	0,0	0
Total	20,0	30

Node 3		
Category	%	n
■ A	0,0	0
■ B	80,0	21
■ C	19,2	5
Total	17,3	26

Node 4		
Category	%	n
■ A	0,0	0
■ B	2,2	1
■ C	97,8	45
Total	30,7	46

The above decision tree is displayed. A fibrinogen level <17 is 100 % predictor of severity score A (mild disease). Fibrinogen 17–44 gives 93 % chance of severity B, fibrinogen 44–56 gives 81 % chance of severity B, and fibrinogen >56 gives 98 % chance of severity score C. The output also shows that the overall accuracy of the model is 94.7 %, but we have to account that this model is somewhat flawed, because all of the data are used twice, one, for building the tree, and, second, for using the tree for making predictions.

Downloading the Knime Data Miner

In Google enter the term "knime". Click Download and follow instructions. After completing the pretty easy download procedure, open the knime workbench by clicking the knime welcome screen. The center of the screen displays the workflow editor. Like the canvas in SPSS Modeler, it is empty., and can be used to build a stream of nodes, called workflow in knime. The node repository is in the left lower angle of the screen, and the nodes can be dragged to the workflow editor simply by left-clicking. The nodes are computer tools for data analysis like visualization and statistical processes. Node description is in the right upper angle of the screen. Before the nodes can be used, they have to be connected with the "file reader" node, and with one another by arrows, drawn, again, simply by left clicking the small triangles attached to the nodes. Right clicking on the file reader enables to configure from your computer a requested data file....click Browse....and download from the appropriate folder a csv type Excel file. You are set for analysis now.

Note: the above data file cannot be read by the file reader, and must first be saved as csv type Excel file. For that purpose command in SPSS: click File....click Save as....in "Save as" type: enter Comma Delimited (*.csv)....click Save. For your convenience it has been made available in extras.springer.com, and entitled "decisiontree".
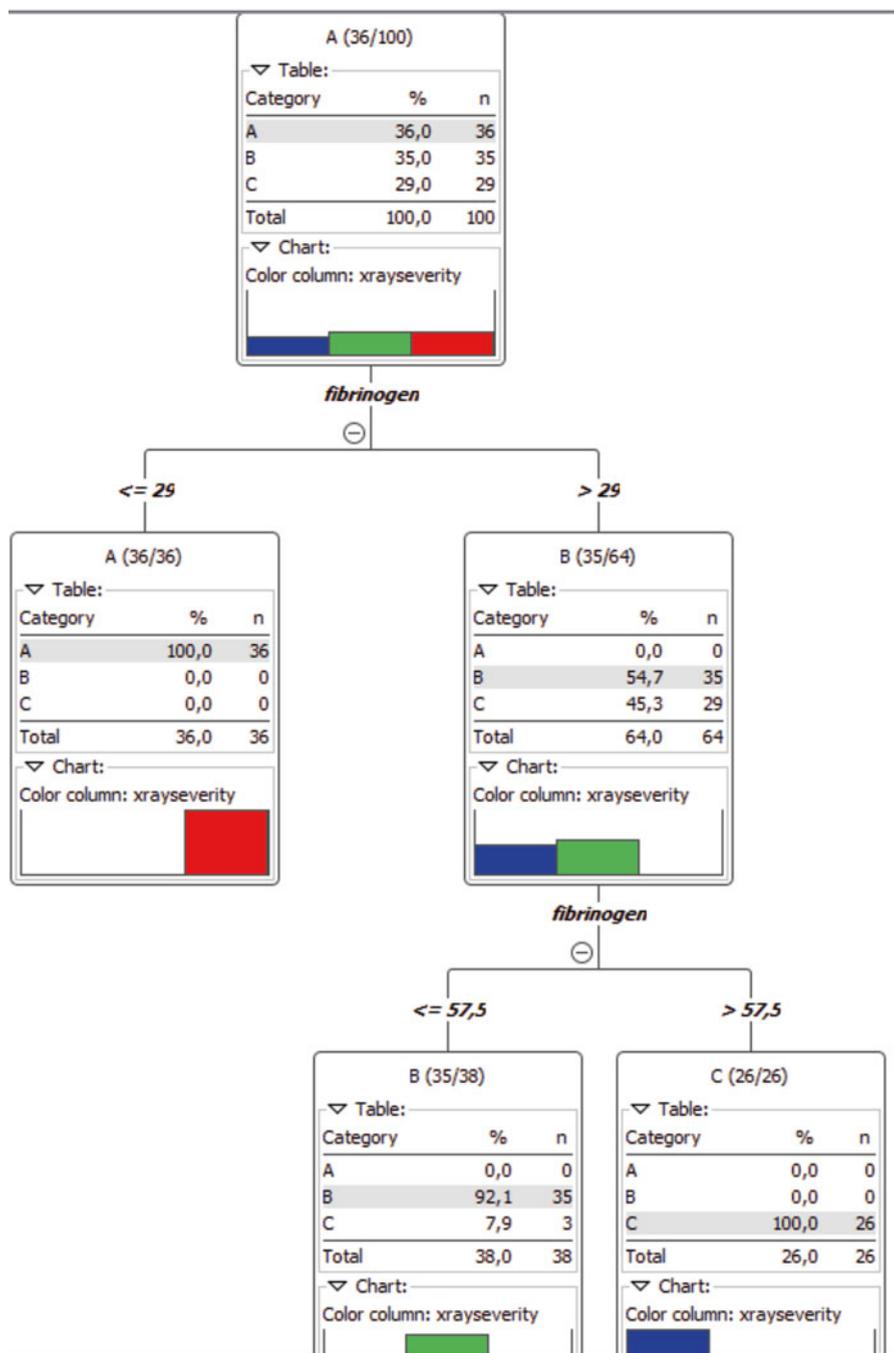
Knime Workflow

A knime workflow for the analysis of the above data example is built, and the final result is shown in the underneath figure.
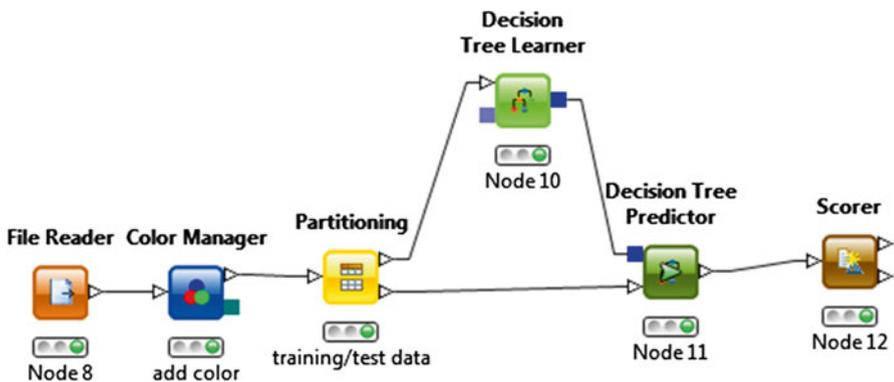
In the node repository click and type color....click the color manager node and drag to workflow editor....in node repository click again color....click the Esc button of your computer....in the node repository click again and type partitioning....the partitioning node is displayed....drag it to the workflow editor....perform the same actions and type respectively Decision Tree Learner, Decision Tree Predictor, and Scorer....Connect, by left clicking, all of the nodes with arrows as indicated above.... Configurate and execute all of the nodes by right clicking the nodes and then the texts "Configurate" and "Execute"....the red lights will successively turn orange and then green....right click the Decision Tree Predictor again....right click the text "View: Decision Tree View".

The underneath decision tree comes up. It is pretty much similar to the above SPSS tree, although it does not use 150 cases but only 45 cases (the test sample from which 100 were resampled). Fibrinogen is again the best predictor. A level <29 mg/l gives you 100 % chance of severity score A. A level 29–57.5 gives 92.1 % chance of Severity B, and a level over 57.5 gives 100 % chance of severity C.

Right clicking the scorer node gives you the accuracy statistics, and shows that the sensitivity of A, B, an C are respectively 100, 93.3, and 90.5 %, and that the overall accuracy is 94 %, slightly less than that of the SPSS tree (94.7 %), but still pretty good. In addition, the current analysis is appropriate, and does not use identical data twice.

A (36/100)

▽ Table:

Category	%	n
A	36,0	36
B	35,0	35
C	29,0	29
Total	100,0	100

▽ Chart:

Color column: xrayseverity

fibrinogen
⊖

<= 29   *> 29*

A (36/36)

▽ Table:

Category	%	n
A	100,0	36
B	0,0	0
C	0,0	0
Total	36,0	36

▽ Chart:

Color column: xrayseverity

B (35/64)

▽ Table:

Category	%	n
A	0,0	0
B	54,7	35
C	45,3	29
Total	64,0	64

▽ Chart:

Color column: xrayseverity

fibrinogen
⊖

<= 57,5   *> 57,5*

B (35/38)

▽ Table:

Category	%	n
A	0,0	0
B	92,1	35
C	7,9	3
Total	38,0	38

▽ Chart:

Color column: xrayseverity

C (26/26)

▽ Table:

Category	%	n
A	0,0	0
B	0,0	0
C	100,0	26
Total	26,0	26

▽ Chart:

Color column: xrayseverity

Conclusion

Traditionally, decision trees are used for finding the best predictors of health risks and improvements. However, this method is not entirely appropriate, because a decision tree is built from a data file, and, subsequently, the same data file is applied once more for computing the health risk probabilities from the built tree. Obviously, the accuracy must be close to 100 %, because the test sample is 100 % identical to the sample used for building the tree, and, therefore, this accuracy does not mean too much. A decision tree with partitioning of a training and a test sample provides similar results, but is scientifically less flawed, because each datum is used only once. In spite of this, little accuracy is lost.

Note

More background, theoretical and mathematical information of decision trees and neural networks are in Chap. 53, and in Machine learning in medicine part one, Chap. 12, pp 145–156, Artificial intelligence, multilayer perceptron modeling, Springer Heidelberg Germany, 2013, both from the same authors.