"Dimitrie Cantemir" Christian University
Knowledge Horizons - Economics
Volume 10, No. 3, pp. 39 - 45
P-ISSN: 2069-0932, E-ISSN: 2066-1061
© 2018 Pro Universitaria
www.orizonturi.ucdc.ro

DECISION TREES

Liviu Constantin STOICA

PhD, Academy of Economic Studies, Bucharest, email: stoica.liviu.constantin@gmail.com

Abstract *In this article we have called "Decision trees" I studied in the first part their general characteristics, its structure and classification. In the second part of the article we model the process of launching a new product on the market using the decision tree. To make the tree, we first made a table with the likelihood of launching the product on the market, after which we made the decision tree and came to the answer I wanted to find out, that is, I learned that the company decides whether it wants to test the new product on the market.*

Key words:
Algorithm, decision tree, entropy, modeling, structure
JEL Codes:
M1

1. INTRODUCTION

Decision trees (or classification) have a structure used to divide large collections of items into smaller sets by applying simple rule-making sequences. This technique builds the tree to model the classification process. Once it is built, it is applied to each tuple (article) of the database and the result of the classification.

The decision tree is a classifier in the form of an arborescence structure in which there is a leaf node and a decision node. The leaf node (or the event node) tells me the value of the target attribute (and is represented by a square), and the decision node specifies some tests to perform on a single attribute (and is represented by a circle).

Decision trees are among the most important methods of classification in Data Mining [A] and assist in decision-making [B].

Decision trees are used to select the best direction in situations where uncertainty arises. The vast majority of a company's decisions are part of this category. I can give as a prime example a manufacturer who has to decide the size of the stock before knowing the demand. Another example may be a stock market player who has to decide whether what he buys before he is sold can make a profit. In these examples, the decision maker encounters an unknown person who can choose the correct one with absolute certainty.

Decision trees can be made from right to left and from top to bottom by performing many tests and trying to get to the best sequence to predict the purpose. Each test creates branches leading to different states until this test ends in a leaf node. Their structure is made up of internal nodes that show me an attribute-by-branch test that is the result of the test and the nodes of the leaves representing the class labels.

A major advantage of decision trees is that they provide us with a graphical representation of successive decision-making processes.

The decision trees (see figure 1) contain the following elements: alternative points, decision points, points of interest, natural states and gains.

2. THE STRUCTURE OF AN DECISION TREES

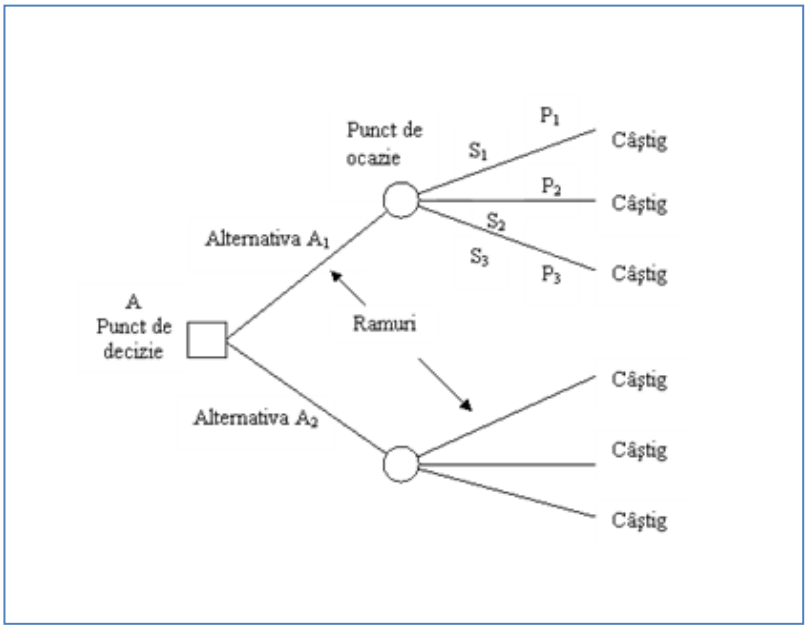

Figure 1 - The general structure of the decision tree

- Decisive decision points and decision nodes are marked with a square, and the decision maker chooses an alternative to running the action that is in a finite number of existing alternatives. These points are presented in the form of branches that depart from the doctor's section of the decision point.
- Occasion points are represented by a circle and describe the occasion of an event that is expected at the moment of preoccupation. I mean, out of an infinite number of natural states one of these is about to emerge. The natural states are represented by branches at the right-hand points. While trees indicate that a decision is made at risk, estimated probabilities of natural states are written above the branches. Every natural state can be followed by a win, a point of opportunity or a decision.

It can be said that decision trees are analysis schemes that can help decision-makers by designing possible outcomes.

a) Classification by induction of decision trees

The most commonly used algorithms that implement decision trees are: CART, CHAID, ID3 / C4,5 and C5,0 and SPRINT. The CART

(Classification and Regression Trees) algorithm was introduced in 1984 by Breiman and is based on classifications and regressions, and its construction is based on binary attribute division, and the CHIID (Chisquared Automatic Interaction Detection) algorithm uses chi-square "For segmentation, and the number of branches differs from two to the number of predicate categories. Another algorithm is the algorithm ID3 / C4.5 and C5.0.

The ID3 algorithm was introduced in 1986 by Quinlan Ross and is used to generate a decision tree from a set of data based on Hunt's algorithm. These algorithms produce trees that have multiple branches for a single node, combining decision trees into a single classifier using information gained for division, and the SPRINT algorithm is used in large data sets, and division is based on the value of a single attribute.

The most commonly used algorithm among the above is the ID3 / C4.5 algorithm, which algorithm adopts as its top-down algorithm that searches only in a part of the search space. The C4.5 algorithm is an extension of the ID3 algorithm, extending the classification range from enumeration attributes to some of the numeric type.

Let S be a lot of data. I suppose I want to classify this set after the X attribute, with *card(X)=m*; it follows that I will have distinct classes C_i, with $i=\overline{1,m}$. Let s_i the number of S elements belonging to the class S belonging to the class C_i. Then the information needed to classify a lot of training is:

$$I(s_1, s_2, \ldots s_3) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

Formula ... Shannon's formula for entropy

Where p_i= the probability that an element belongs to the class C_i , $p_i=\frac{s_i}{s}$

Let the attribute *A* with *card(A)=v*, $A=\{a_1, a_2, \ldots a_v\}$. Attribute *A* partition *S* in *v* subsets $\{S_1, S_2, \ldots S_v\}$, where S_j contains the items $y \in S$, $A(y)=a_i$. If the attribute *A* would be selected as the test attribute (ie best for the classification that most efficiently divides the data), then these subparts correspond to the arcs that start from the node containing the set *S*. Let $s_{ij}=card(C_i \cap S_j)$.

Entropy, ie the expected information can be obtained by partitioning the set S by attribute A is given by:

$$E(A)=\sum_{j=1}^{v} \frac{s_{1j}+\ldots+s_{mj}}{s} I(s_{1j},\ldots s_{mj}).$$

Formula ... - Entropy calculation

The lower the entropy, the more partitioned sets are more homogeneous. The information obtained by dividing the tree after node A is $G(A)=I(s_1, s_2, \ldots s_m)-E(A)$.

The algorithm calculates the information obtained by each attribute. The attribute with the most information gathered by going through the tree is chosen as a test attribute for the S trained set. A node is created and labeled with the chosen attribute, and for each value of that attribute arcs are created, the set of trains is partitioned after that attribute.

b) Advantages of decision trees

In 1997, M. Kambr, L. Winston, W. Gong, S. Cheng, and J. Han stated that decision trees require restrictions on the data being studied. According to Bounsaythip C. and Rinta-Runsala in 2001, the main advantages are that they are easy to understand and produce efficient models. They also claim to be applicable to real issues, for example, in trade issues. [C]

3. MODELING THE LANGUAGE PROCESS OF A PRODUCT BASED ON THE DECISION ARBOR [D]

Decision trees are methods that analyze multiple variables by applying complex situations. Decision trees are capable of completing and replacing some forms of statistical analysis (example: multiple linear regression), some techniques and tools used in data mining (eg neural networks) and some forms developed by business intelligence. Decision trees are based on algorithms that identify different segmentation modes that are part of a series of data that gives us the branches of the tree.

I still want to model the process of launching a new product that has the following advantages, ie: students and students have a 15% discount, people over 65 have a 10% discount, and if the product is bought in installments, then the rate its annual rate is 6.75% per year.

The store can choose one of the following options, ie: can give up the launch, sell the product immediately, and test the sale of the product to a limited number of people.

If tested, they can have the following results:

a) the new product may be chosen by less than 25% of the person;

b) the new product may be chosen by less than 25% of the person, but not less than 50% of them will buy it again;

c) the new product may be chosen by more than 25% of the person, but at least 50% of them will be purchased for the second time.

Next, exemplify in table AA the probabilities of launching the product on the market

	< 25 %	**> 25 %**		**Probability**
		Return < 50%	**Return > 50%**	
Success (S)	0.18	0.22	0.20	0.60
Failure (E)	0.26	0.10	0.04	0.40
Probability	**0.44**	**0.32**	**0.24**	**1**

Table AA - Product launch probabilities on the market

As a result of the test, the test shows whether the product has been put to the attention of individuals, then the company has two possibilities: to market the product or to give up the new product. These two situations will be presented by a diagram in the form of a decision tree (figure ...).

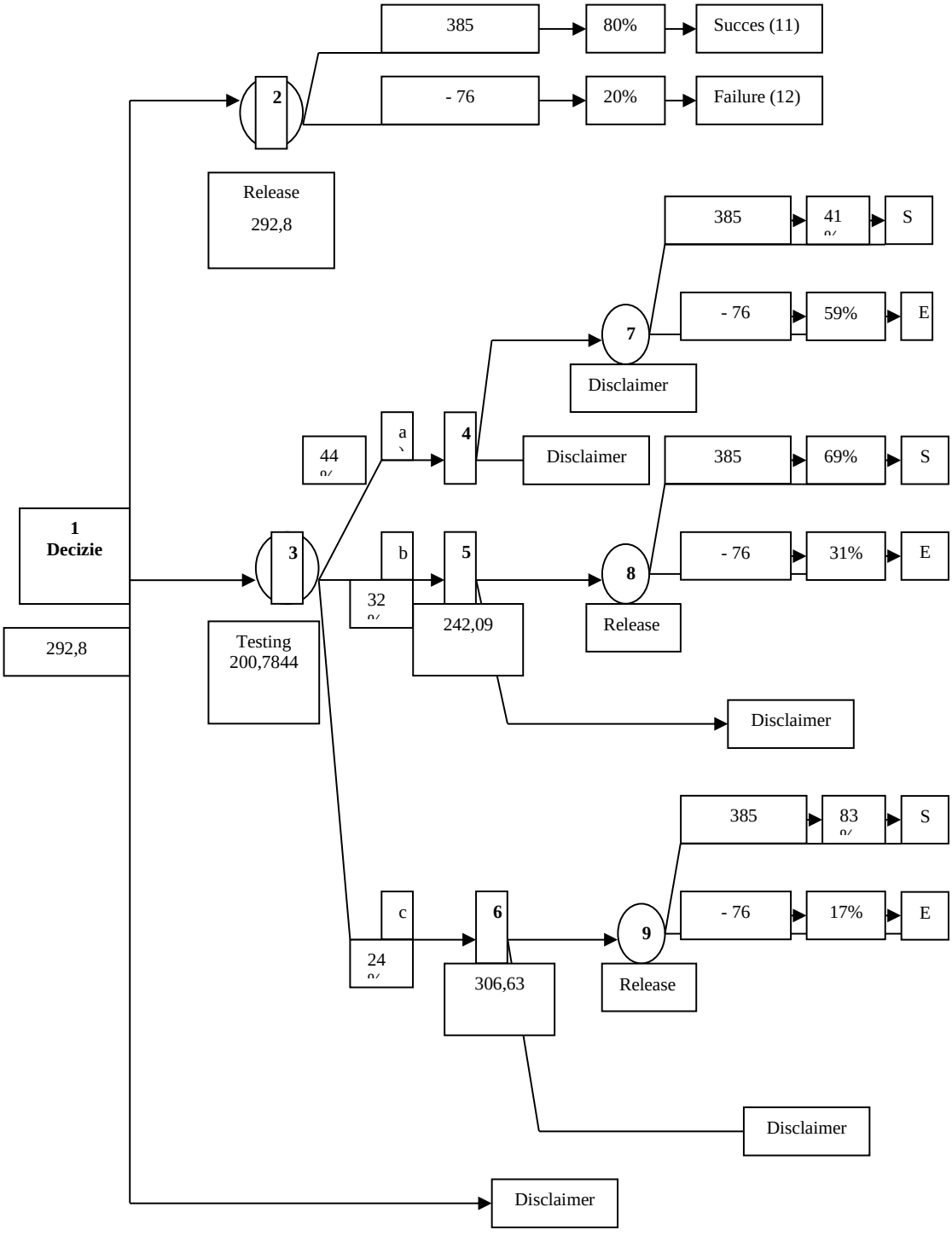

Figure ... - Decision tree

To calculate the values in nodes 1, 2 and 3 in the figure above, I first have to move to the right by performing the inverse calculations starting from nodes 7, 8, 9 to nodes 1, 2 and 3.

The probability of success or failure in the case of:		
a).	success (S):	0,18 / 0,44 = 0.41 (41 %)
	failure (E):	0.26 / 0.44 =0.59 (59 %)
b).	success (S):	0.22 / 0.32 = 0.69 (69 %)
	failure (E):	0.10 / 0.32 = 0.31 (31 %)
c).	success (S):	0.20 / 0.24 = 0.83 (83 %)
	failure (E):	0.04 / 0.24 = 0.17 (17 %)

Table BB - Probabilities in case of success or failure

And data is taken from table AA, ie:

Case : a) S=0,16; E=0,28; P=0,44;

b) S=0,24; E=0,08; P=0,32;

c) S=0,2; E=0,04; P=0,24.

On the basis of the percentages, I calculate the values of the nodes 2, 7, 8 and 9:

- node 2: 385 * 80% - 76 * 20% = 292,8;
- node 7: 385 * 36% - 76 * 64% = 89,96
- node 8: 385 * 75% - 76* 25% = 269,75
- node 9: 385 * 83% - 76 * 17% = 306,63

The examples given above can be described in Table … as follows:

Node	Success / Failure	Value	Percent
Node 2	success (S)	385	80%
	failure (E)	-76	20%
Node 7	success (S)	385	41%
	failure (E)	-76	59%
Node 8	success (S)	385	69%
	failure (E)	-76	31%
Node 9	success (S)	385	83%
	failure (E)	-76	17%

According to the chart, I noticed that in node 4 I have two possibilities, ie 157.85 people can choose the new product from the company or even give up. It can be noticed that in node 5 there are 265.65 people who want to buy the new product, and if node 6 is studied it can be seen that the new product is only wanted for 24% of the questioned persons, ie:

113,01*44% + 242,09*32% + 306,63*24% = 200,7844.

If I study node 1, I have three possibilities, namely: the company to give up the launch of the new product on the market and then no person will benefit from it, the second possibility is to launch the new product on the market and it will be purchased by 200,7844 people, and the last option would be to test the new product.

So the company decides whether to test the product in time to launch it on the market because its cost is 92.01 (number of people launched – number of people tested= 292,8 – 200,7844 = 92,01).

4. CONCLUSIONS

The present article entitled "Decision trees" was structured in three parts. In the first part we studied decision trees in general, in the second part we made a structure of it emphasizing their classification by stating Shannon's formula for entropy and its calculation formula. In the third part we made an application for launching a new product, and after completing the decision tree, we concluded that the company decides whether to test the product before launching at a cost of 92.01.

5. BIBLIOGRAPHY:

[A] M.J. Berry, G.S. Linoff, "Data mining techniques: For marketing, sales, and customer support", John Wiley & Sons, Inc., 1997

[B] Shucheng Gong, Hongyan Liu, "Constructing Decision Trees for Unstructured Data", International Conference on Advanced Data Mining and Applications, pp 475-487, 2014

[C] Kamber M., Winstone L., Gong W., Cheng S., and Han J., "Generalization and Decision Tree Induction: Efficient Classification inData Mining", Proceedings of 1997 Int'l Workshop on Research Issues on Data Engineering (RIDE'97), Birmingham, England, April 1997, pg. 111-120

[D]Lungu Ion, Stancu Ana-Maria Ramona, "Modelarea procesului de lansare a unui produs pe baza arborelui de decizie", Conferința Internațională Universitatea Ovidius-Constanța, pp 203-206, ISSN 1582-9383, 2014