Decision tree

Decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision trees as a form of representation of a set of rules are very popular among SI programmers due to readability and the requirements for memory resources for storing the database.

This fact is illustrated in the table below with an example of a rule base, and a picture showing the corresponding decision tree. This database is a set of the following rules:

$$IF \ldots THEN \ldots$$

The rule database consists of 4 attributes and assigned to them values.

$$attribut1 = \{a, b, c\},$$

$$attribut2 = \{k, e, f\},$$

$$attribut3 = \{d, m\},$$

$$attribut4 = \{n, w\}.$$

Algorithm for constructing the optimal decision tree

First, you need to answer the question of how to create an optimal tree. The algorithm for constructing an optimal decision tree is implemented by a recursive function, which takes as an argument the base of rules.

The first step is to choose an attribute that will be the root of the tree. We create a new node and choose the best attribute, i.e. the node that will give the tree simplicity. This simplicity is understood as the smallest number of nodes and leaves and the shortest possible paths connecting the root with the leaves.

The individual evaluation functions for the attributes, the entropies of the rule set P with regard to the t attribute, are defined as follows:

$$E_t(P) = \sum_{r \in R_t} \left[\frac{\mid P_{tr} \mid}{\mid P \mid} \sum_{d \in C} - \frac{\mid P_{tr}^d \mid}{\mid P_{tr} \mid} \log_2 \frac{\mid P_{tr}^d \mid}{\mid P_{tr} \mid} \right],$$

where:

- R_t - a set of values for the t attribute;
- P_{tr} - a subset of the set P containing those rules for which the attribute t takes value r;
- C - set of all categories;
- P_{tr}^d - a subset of the P_{tr} set containing rules whose decision part is category d.

Attention. In the case of the expression $0 \log_2 0$, we take its value as 0.

No.	attribute 1	attribute 2	attribute 3	attribute 4	value 1
1.	a	k	d	n	u
2.	a	k	d	w	u
3.	b	k	d	n	o
4.	c	e	d	n	x
5.	a	e	d	n	u
6.	b	f	m	w	o
7.	c	f	m	n	x
8.	c	f	d	w	o
9.	b	e	m	w	o
10.	a	f	m	n	o
11.	b	k	m	n	o
12.	b	f	d	n	o
13.	a	e	m	w	o
14.	c	e	d	n	x
15.	c	e	m	w	o

The rule for the first row:

$$IF$$

$$(attribute\ 1\ IS\ a) AND (attribute\ 2\ IS\ k) AND (attribute\ 3\ IS\ d) AND (attribute\ 4\ IS\ n)$$

$$THEN$$

$$(value\ 1\ IS\ u).$$

As a valid test for a given node, we choose the attribute with the smallest value of entropy. So, we calculate $E_{attribute1}(P)$, $E_{attribute2}(P)$, $E_{attribute3}(P)$, $E_{attribute4}(P)$.

$$E_{attribute1}(P) = E_{att1}(P) =$$

$$\frac{|P_{\text{att1},\ a}|}{|P|}\left[-\frac{|P^u_{\text{att1},\ a}|}{|P_{\text{att1},\ a}|}\log_2\frac{|P^u_{\text{att1},\ a}|}{|P_{\text{att1},\ a}|} - \frac{|P^o_{\text{att1},\ a}|}{|P_{\text{att1},\ a}|}\log_2\frac{|P^o_{\text{att1},\ a}|}{|P_{\text{att1},\ a}|} - \frac{|P^x_{\text{att1},\ a}|}{|P_{\text{att1},\ a}|}\log_2\frac{|P^x_{\text{att1},\ a}|}{|P_{\text{att1},\ a}|}\right] +$$

$$\frac{|P_{\text{att1},\ b}|}{|P|}\left[-\frac{|P^u_{\text{att1},\ b}|}{|P_{\text{att1},\ b}|}\log_2\frac{|P^u_{\text{att1},\ b}|}{|P_{\text{att1},\ b}|} - \frac{|P^o_{\text{att1},\ b}|}{|P_{\text{att1},\ b}|}\log_2\frac{|P^o_{\text{att1},\ b}|}{|P_{\text{att1},\ b}|} - \frac{|P^x_{\text{att1},\ b}|}{|P_{\text{att1},\ b}|}\log_2\frac{|P^x_{\text{att1},\ b}|}{|P_{\text{att1},\ b}|}\right] +$$

$$\frac{|P_{\text{att1},\ c}|}{|P|}\left[-\frac{|P^u_{\text{att1},\ c}|}{|P_{\text{att1},\ c}|}\log_2\frac{|P^u_{\text{att1},\ c}|}{|P_{\text{att1},\ c}|} - \frac{|P^o_{\text{att1},\ c}|}{|P_{\text{att1},\ c}|}\log_2\frac{|P^o_{\text{att1},\ c}|}{|P_{\text{att1},\ c}|} - \frac{|P^x_{\text{att1},\ c}|}{|P_{\text{att1},\ c}|}\log_2\frac{|P^x_{\text{att1},\ c}|}{|P_{\text{att1},\ c}|}\right] =$$

$$\frac{5}{15}\cdot\left[-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} - \frac{0}{5}\log_2\frac{0}{5}\right] +$$

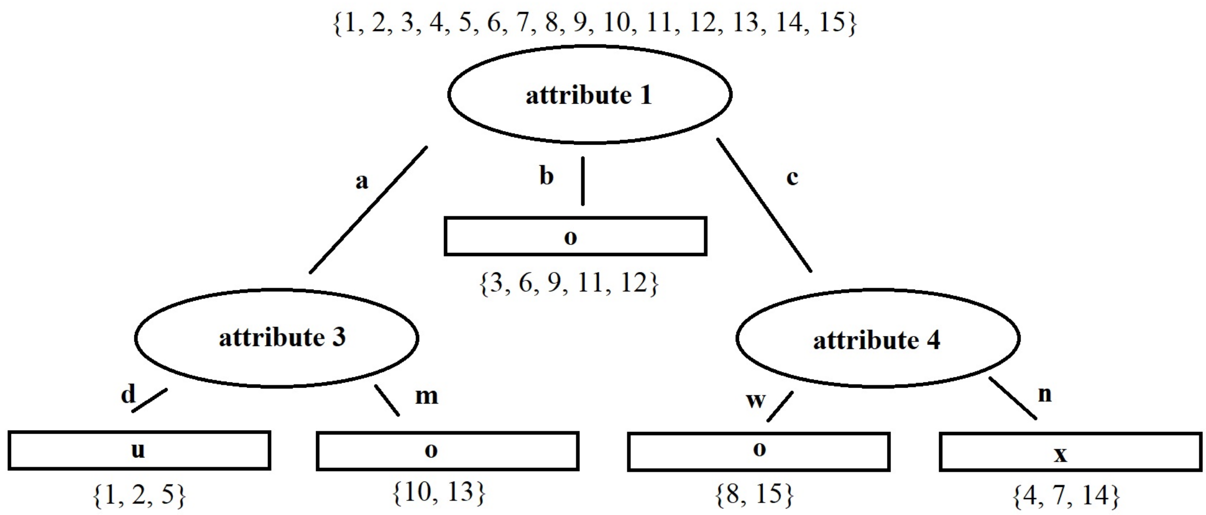
$$\frac{5}{15}\cdot\left[-\frac{0}{5}\log_2\frac{0}{5} - \frac{5}{5}\log_2\frac{5}{5} - \frac{0}{5}\log_2\frac{0}{5}\right] +$$

$$\frac{5}{15} \cdot \left[-\frac{0}{5}\log_2\frac{0}{5} - \frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} \right] = 0,647.$$

Analogously we calculate $E_{attribute2}(P)$, $E_{attribute3}(P)$, $E_{attribute4}(P)$.

Since the attribute with the lowest entropy value is attribute 1, the root of our tree will be this attribute.

The simplest decision tree.



Rysunek 1: The simplest decision tree.