Decision tables and case based reasoning

Zdzisław Pawlak

Institute for Theoretical and Applied Informatics
Polish Academy of Sciences
ul. Bałtycka 5, 44-100 Gliwice, Poland
and
Warsaw School of Information Technology
ul. Newelska 6, 01-447 Warsaw, Poland
e-mail: zpw@ii.pw.edu.pl

Abstract.

1 Introduction

2 Decision tables

Formally, by an *information system* we will understand a pair $S = (U, A)$, where U and A, are finite, nonempty sets called the *universe*, and the set of *attributes,* respectively. With every attribute $a \in A$ we associate a set V_a of its *values*, called the *domain* of a. Any subset B of A determines a binary relation I(B) on U, called an *indiscernibility relation*, and defined as follows: $(x, y) \in I(B)$ if and only if $a(x) = a(y)$ for every $a \in A$, where a(x) denotes the value of attribute a for element x.

Obviously I(B) is an equivalence relation. The family of all equivalence classes of I(B), i.e., a partition determined by B, will be denoted by U/I(B), or simply by U/B. An equivalence class of I(B), i.e., block of the partition U/B, containing x will be denoted by B(x).

If (x, y) belongs to I(B), we will say that x and y are B-*indiscernible* (*indiscernible with respect to* B). Equivalence classes of the relation I(B) (or blocks of the partition U/B) are referred to as B-*elementary sets* or B-*granules.*

If we distinguish in an information system two disjoint classes of attributes, called *condition* and *decision attributes*, respectively, then the system will be called a *decision system*, denoted by $S = (U, C, D)$, where C and D are disjoint sets of condition and decision attributes, respectively.

An example of a decision table is shown in Table 1.

Table 1. Decision table

case	disease	age	test	support
1	yes	old	+	320
2	yes	old	−	130
3	yes	middle	+	70
4	yes	middle	−	50
5	yes	young	−	30
6	no	old	+	80
7	no	old	−	40
8	no	young	−	280

In the table 8 cases of patients stuffering from a certain disease, together with test result is presented. In the table the number of cases of every "type" of patients is given and is called *support* of the type. *Disease* and *age* are condition attributes, whereas *test* is condition attribute.

Every row (case) in the decision table determine a *decision rule*. For example, row 3 determines a decision rule

$$\textit{if } (\textit{disease, yes}) \textit{ and } (\textit{age, middle}) \textit{ then } (\textit{test}, +).$$

Normalized support for a decision rule will be called *strength* of the decision rule. For example, strength of the above decision rule is $70/1000 = 0.07$.

With every decision table we associate a decision graph $G = (N, B, \sigma), B \subseteq N \times N$, where N is a set of *nodes* B set of *branches* (pair of notes) and σ is the strength of the branch. Nodes represent conditions and decision of the decision table, whereas branches represents decision rules.

With every branch (decision rules) (x, y) of a flow graph G we associate the *certainty factor*

$$\mathrm{cer(x,y)} = \frac{\sigma\mathrm{(x,y)}}{\sigma\mathrm{(x)}}$$

and the *coverage factor*

$$\mathrm{cov(x,y)} = \frac{\sigma\mathrm{(x,y)}}{\sigma\mathrm{(y)}}$$

where $\sigma(x) \neq 0$ and $\sigma(y) \neq 0$.

These coefficients are widely used in data mining (see, e.g., [5,11,12]) but they can be traced back to Łukasiewicz [4], who used them first in connection with his research on logic and probability.

3 Properties of decision rules

The following properties are immediate consequences of definitions given above:

$$\sum_{\mathrm{y \in O(x)}} \mathrm{cer(x,y)} = 1 \tag{1}$$

$$\sum_{x \in I(y)} \mathrm{cov}(x, y) = 1 \tag{2}$$

$$\sigma(x) = \sum_{y \in O(x)} \mathrm{cer}(x, y)\sigma(x) = \sum_{y \in O(x)} \sigma(x, y) \tag{3}$$

$$\sigma(y) = \sum_{x \in I(y)} \mathrm{cov}(x, y)\sigma(y) = \sum_{x \in I(y)} \sigma(x, y) \tag{4}$$

$$\mathrm{cer}(x, y) = \frac{\mathrm{cov}(x, y)\sigma(y)}{\sigma(x)} \tag{5}$$

$$\mathrm{cov}(x, y) = \frac{\mathrm{cer}(x, y)\sigma(x)}{\sigma(y)} \tag{6}$$

The above properties have a probabilistic flavor, e.g., equations (3) and (4) have a form of total probability theorem, whereas formulas (5) and (6) are Bayes' rules [10]. However, in our approach, these properties are interpreted in a deterministic way and they describe flow distribution among branches in the network.

Decision graph associated with Table 1 is given in Fig. 1.

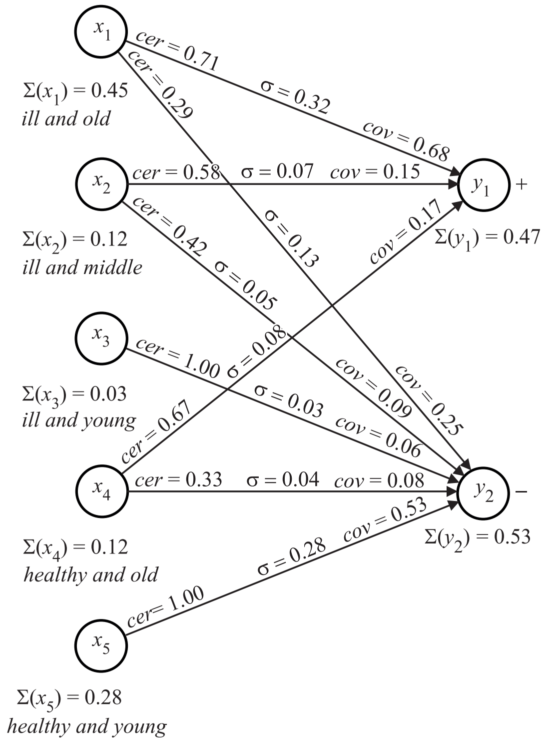

Fig. 1. tytul

4 Dependencies in decision tables

Let x and y be nodes in a decision graph $G = (N, \mathcal{B}, \sigma)$, such that $(x, y) \in \mathcal{B}$. Nodes x and y are *independent* in G if

$$\sigma(x, y) = \sigma(x)\sigma(y). \tag{7}$$

From (7) we get

$$\frac{\sigma(x, y)}{\sigma(x)} = cer(x, y) = \sigma(y), \tag{8}$$

and

$$\frac{\sigma(x, y)}{\sigma(y)} = cov(x, y) = \sigma(x). \tag{9}$$

If

$$cer(x, y) > \sigma(y), \tag{10}$$

or

$$cov(x, y) > \sigma(x), \tag{11}$$

x and y are *positively depends* on x in G.
similarly, if

$$cer(x, y) < \sigma(y), \tag{12}$$

or

$$cov(x, y) < \sigma(x), \tag{13}$$

then x and y are *negatively dependent* in G.

Relations of independency and dependences are symmetric ones, and are analogous to those used in statistics.