Decision tables and decision spaces

Z. Pawlak[1]

Abstract

Abstract. In this paper an Euclidean space, called a *decision space* is associated with every decision table. This can be viewed as a generalization of the indiscernibility matrix, basic tool to find reducts in information systems. Besides, the decision space gives a clear insight in the decision structure imposed by the decision table.

1. Introduction

This paper concerns granular structure of decision tables imposed by the indiscernibility relation on data. With every decision rule in a decision table three coefficients are associated: the *strength*, the *certainty* and the *coverage factors* of the rule. It is shown that these coefficients satisfy Bayes' theorem and the total probability theorem. This enables us to use Bayes' theorem to discover patterns in data in a different way than that offered by standard Bayesian inference technique employed in statistical reasoning. Besides, it is shown that the decision rules define a relation between condition and decision granules, which can be represented by a flow graph. The certainty and coverage factors determine „flow of information" in the graph, which shows clearly the relationship between condition and decision granules determined by the decision table. The decision structure of a decision table can be represented in a „decision space", which is Euclidean space, in which dimensions of the space are determined by values of the decision attribute, points in the space are condition granules and coordinates of the points are strengths of the corresponding rules. Distance in the decision space between condition granules allows to determine how „distant" are decision makers in view of their decisions. This idea can be viewed as a generalization of the indiscernibility matrix, basic tool to find reducts in information systems. Besides, the decision space gives a clear insight in the decision structure imposed by the decision table.

2. Information systems and decision tables

In this section we define basic concept of rough set theory, information system.

An information system is a data table, whose columns are labeled by attributes, rows are labeled by objects of interest and entries of the table are attribute values.

Formally, an *information system* is a pair $S = (U,A)$, where U and A, are non-empty finite sets called the *universe*, and the set of *attributes*, respectively such that $a:U \rightarrow V_a$, where V_a, is the set of all *values* of a called the *domain* of a. Any subset B of A determines a binary relation $I(B)$ on U, which will be 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 the information system two disjoint classes of attributes, called *condition* and *decision attributes*, respectively, then the system will be called a *decision table* and will be denoted by $S = (U, C, D)$, where C and D are disjoint sets of condition and decision attributes, respectively and $C \cup D = A$.

$C(x)$ and $D(x)$ will be referred to as the condition class and the decision class induced by x, respectively.

[1] Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, ul. Bałtycka 5, 44 100 Gliwice, Poland

Thus the decision table describes decisions (actions, results etc.) determined, when some conditions are satisfied. In other words each row of the decision table specifies a decision rule which determines decisions in terms of conditions.

An example of a simple decision table is shown below.

decision rule	*age*	*sex*	*profession*	*disease*
1	*old*	*male*	*yes*	*no*
2	*med.*	*female*	*no*	*yes*
3	*med.*	*male*	*yes*	*no*
4	*old*	*male*	*yes*	*yes*
5	*young*	*male*	*no*	*no*
6	*med.*	*female*	*no*	*no*

Table 1. Decision table

In the table *age*, *sex* and *profession* are condition attributes, whereas *disease* is the decision attribute.

The table contains data concerning relationship between age, sex, profession and certain vocational disease.

3. Decision rules

In what follows we will describe decision rules more exactly.

Let $S = (U, C, D)$ be a decision table. Every $x \in U$ determines a sequence $c_1(x),\ldots, c_n(x), d_1(x),\ldots, d_m(x)$ where $\{c_1,\ldots, c_n\} = C$ and $\{d_1,\ldots, d_m\} = D$.

The sequence will be called a *decision rule induced by x* (in *S*) and denoted by $c_1(x),\ldots, c_n(x) \rightarrow d_1(x),\ldots, d_m(x)$ or in short $C \xrightarrow[x]{} D$.

The number $supp_x(C,D) = |C(x) \cap D(x)|$ will be called a *support* of the decision rule $C \xrightarrow[x]{} D$ and the number

$$\sigma_x(C,D) = \frac{supp_x(C,D)}{|U|},$$

will be referred to as the *strength* of the decision rule $C \xrightarrow[x]{} D$, where $|X|$ denotes the cardinality of X.

In Table 2 below a modified version of Table 1 is shown.

decision rule	*age*	*sex*	*profession*	*disease*	*support*	*strength*
1	*old*	*male*	*yes*	*no*	200	0.18
2	*med.*	*female*	*no*	*yes*	70	0.06
3	*med.*	*male*	*yes*	*no*	250	0.23
4	*old*	*male*	*yes*	*yes*	450	0.41
5	*young*	*male*	*no*	*no*	30	0.03
6	*med.*	*female*	*no*	*no*	100	0.09

Table 2. Support and strength

This decision table can be understood as an abbreviation of a bigger decision table containing 1100 rows. Support of the decision rule means the number of identical decision rules in the original decision table.

With every decision rule $C \xrightarrow[x]{} D$ we associate a *certainty factor* of the decision rule, denoted $cer_x(C, D)$ and defined as follows:

$$cer_x(C,D) = \frac{|C(x) \cap D(x)|}{|C(x)|} = \frac{\sigma_x(C,D)}{\pi(C(x))},$$

where $C(x) \neq \emptyset$ and $\pi(C(x)) = \frac{|C(x)|}{|U|}$.

The certainty factor may be interpreted as a conditional probability that y belongs to $D(x)$ given y belongs to $C(x)$, symbolically $\pi_x(D|C)$, i.e., $cer_x(C, D) = \pi_x(D|C)$.

If $cer_x(C, D) = 1$, then $C \xrightarrow[x]{} D$ will be called a *certain decision* rule; if $0 < cer_x(C, D) < 1$ the decision rule will be referred to as an *uncertain decision rule*.

Besides, we will also use a *coverage factor* of the decision rule, denoted $cov_x(C, D)$ defined as

$$cov_x(C,D) = \frac{|C(x) \cap D(x)|}{|D(x)|} = \frac{\sigma_x(C,D)}{\pi(D(x))},$$

where $D(x) \neq \emptyset$ and $\pi(D(x)) = \frac{|D(x)|}{|U|}$.

Similarly

$$cov_x(C,D) = \pi_x(C \mid D).$$

If $C \xrightarrow[x]{} D$ is a decision rule then $D \xrightarrow[x]{} C$ will be called an *inverse decision rule*. The inverse decision rules can be used to give *explanations* (*reasons*) for decisions.

4. Probabilistic properties of decision tables

Decision tables have important probabilistic properties which are discussed next.

Let $C \xrightarrow[x]{} D$ be a decision rule, then the following properties are valid:

$$\sum_{y \in C(x)} cer_y(C,D) = 1 \tag{1}$$

$$\sum_{y \in D(x)} cov_y(C,D) = 1 \tag{2}$$

$$\pi(D(x)) = \sum_{y \in D(x)} cer_y(C,D) \cdot \pi(C(y)) = \sum_{y \in D(x)} \sigma_y(C,D) \tag{3}$$

$$\pi(C(x)) = \sum_{y \in C(x)} cov_y(C,D) \cdot \pi(D(y)) = \sum_{y \in C(x)} \sigma_y(C,D) \tag{4}$$

$$cer_x(C,D) = \frac{cov_x(C,D) \cdot \pi(D(x))}{\sum_{y \in C(x)} cov_y(C,D) \cdot \pi(D(y))} = \frac{\sigma_x(C,D)}{\pi(C(x))} \tag{5}$$

$$cov_x(C,D) = \frac{cer_x(C,D) \cdot \pi(C(x))}{\sum_{y \in D(x)} cer_y(C,D) \cdot \pi(C(y))} = \frac{\sigma_x(C,D)}{\pi(D(x))} \tag{6}$$

That is, any decision table, satisfies (1) - (6). Observe that (3) and (4) refer to the well known *total probability theorem*, whereas (5) and (6) refer to *Bayes' theorem*.

Thus in order to compute the certainty and coverage factors of decision rules according to formula (5) and (6) it is enough to know the strength (support) of all decision rules only. The strength of decision rules can be computed from data or can be a subjective assessment.

Certainty and coverage factors for the decision table presented in Table 2 are shown in Table 3.

decision rule	*strength*	*certainty*	*coverage*
1	0.18	0.31	0.34
2	0.06	0.40	0.13
3	0.23	1.00	0.43
4	0.41	0.69	0.87
5	0.03	1.00	0.06
6	0.09	0.60	0.17

Table 3. Certainty and coverage factors

Let us observe that according to formulas (5) and (6) the certainty and coverage factors can be computed employing only the strength of decision rules.

In Table 2 decision rules 3 and 5 are certain, whereas the remaining decision rules are uncertain.

This means that medium age male having the profession and young male not having the profession are certainly healthy. Old males having the profession are most probably ill (probability = 0.69) and medium age females not having the profession are most probably healthy (probability = 0.60).

The inverse decision rules say that healthy persons are most probably medium age males having the profession (probability = 0.43) and ill persons are most probably old males having the profession (probability = 0.87).

5. Decision tables and flow graphs

With every decision table we associate a *flow graph*, i.e., a directed acyclic graph defined as follows: to every decision rule $C \xrightarrow[x]{} D$ we assign a *directed branch x* connecting the *input node C*(*x*) and the *output node D*(*x*). Strength of the decision rule represents a *throughflow* of the corresponding branch. The throughflow of the graph is governed by formulas (1),...,(6).

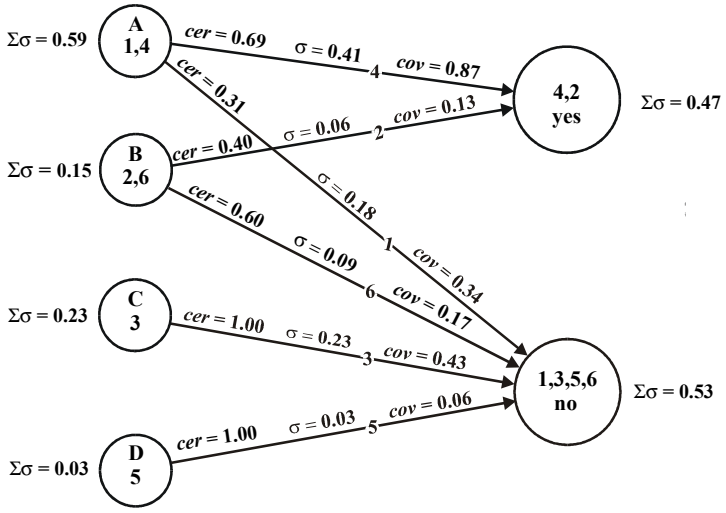

Fig. 1. Flow graph

Flow graph associated with decision table presented in Table 2 is shown in Fig. 1.

The application of flow graphs to represent decision tables gives a very clear insight into the decision process. Classification of objects in this representation boils down to finding the maximal output flow in the flow graph, whereas explanation of decisions is connected with the maximal input flow associated with the given decision (see also [2] and [3]).

6. Decision Space

With every decision table having one *n*-valued decision attribute we can associate *n*-dimensional Euclidean space, where values of the decision attribute determine *n* axis of the space and condition attribute values (equivalence classes) determine point of the space. Strengths of decision rules are to be understood as coordinates of corresponding points.

Distance $\delta(x, y)$ between point *x* and *y* in an *n*-dimensional decision space is defined as

$$\delta(x,y)=\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where $x = (x_1,...,x_n)$ and $y = (y_1,...,y_n)$ are vectors of strengths of corresponding decision rules.

Decision space for Table 1 is shown in Figure 2

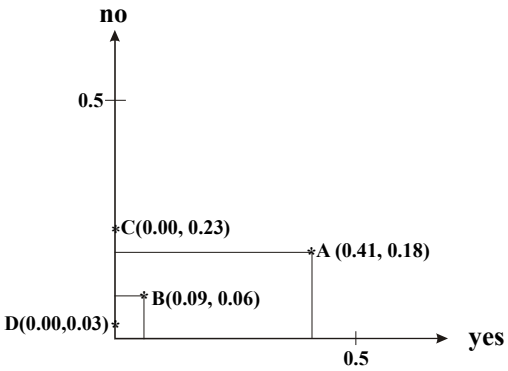

Fig. 2. Decision space

Distances between granules A, B, C and D are shown in Table 4.

	A	B	C	D
A				
B	0.4511			
C	0.4130	0.1523		
D	0.4365	0.0849	0.2000	

Table 4. Distance matrix

It follows from the above example that group of patients B, C and D are „close" and form a cluster which is „distant" from group A.

7. Conclusions

Decision spaces associated with decision tables can be viewed as a generalization of discernibility matrices. Besides, they enable us to get deeper insight in decision processes determined by decision tables.

References

1. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems, in: Słowiński (ed.), Intelligent Decision Support, - Handbook of Applications and Advances in Rough Set Theory, Kluwer Academic Publishers, Dordrech, 1992, 311-362.
2. Ford, L.R., Fulkerson, D. R.: Flows in Networks. Princeton University Press, Princeton, New Jersey. 1962.
3. Pawlak, Z.: New look on Bayes' theorem - the rough set outlook. Proceeding of International Workshop on Rough Set Theory and Granular Computing (RSTGC-2001), Matsue, Shimane,

Japan, May 20-22, S. Hirano, M. Inuiguchi and S. Tsumoto (eds.), Bull. of Int. Rough Set Society vol. 5 no. 1/2, 2001, 1-8.