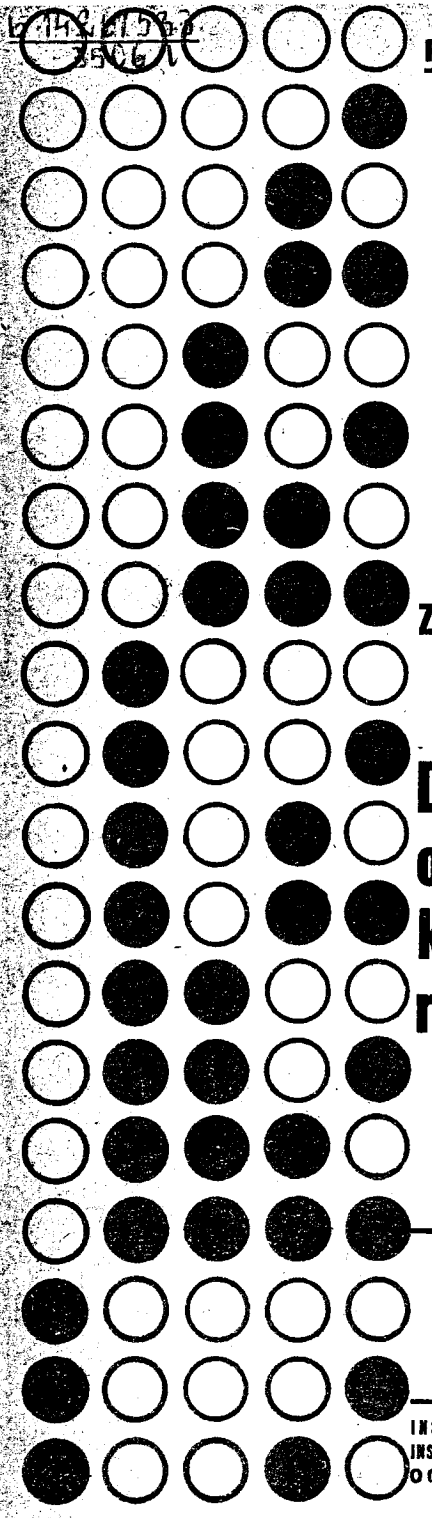
PRACE IPI PAN • ICS PAS REPORTS



Zdzisław Pawlak

Discrimination power of attributes in knowledge representation system

533

January 1984

WARSZAWA

INSTYTUT PODSTAW INFORMATYKI POLSKIEJ AKADEMII NAUK
INSTITUTE OF COMPUTER SCIENCE POLISH ACADEMY OF SCIENCES
00-901 WARSAW, P.O. Box 22, POLAND

Zdzisław Pawlak

DISCRIMINATION POWER OF ATTRIBUTES IN KNOWLEDGE REPRESENTATION SYSTEMS

533

Warsaw, January 1984

2.a
3.a

Pracę zgłosił Andrzej Blikle

Mailing address: Zdzisław Pawlak
Institute of Computer Science
Polish Academy of Sciences
P.O. Box 22
00-901 Warszawa, PKiN

ISSN 0138-0643

P r i n t e d a s a m a n u s c r i p t
N a p r a w a c h r ę k o p i s u

Nakład 350 egz. Ark. wyd. 0,35; ark. druk. 1,00.
Papier offset. kl. V, 70 g, 70 x 100. Oddano do
druku w styczniu 19[illegible]. Z. N. Zam. nr 10/84

Abstract • **Содержание** • Streszczenie

In this paper we consider the problem how selection of attributes affects the accuracy of description of a given set of objects. To investigate this problem we employ the rough sets approach. An example of medical data analysis by this method is given.

О мощности дискриминационных атрибутов в системах представления знаний

В работе иисследуется проблема, каким образом подбор атрибутов влияет на точность описания заданного множества объектов. Для исследования этой проблемы используется подход вытекающий из концепции приближенных множеств. Приводится пример применения этого метода к анализу медицинских данных.

O mocy dyskryminacyjnych atrybutów w systemach reprezentacji wiedzy

W pracy badany jest problem, jak dobór atrybutów wpływa na dokładność opisu zadanego zbioru obiektów. Do badania tego problemu użyto podejścia oferowanego w koncepcji zbiorów przybliżonych. Podano przykład zastosowania tej metody do analizy danych medycznych.

1. INTRODUCTION

In this paper we analyze how selection of attributes affects the accuracy of description of a given set of objects. We employ in our considerations the rough set approach (see Pawlak (1982), Pawlak (1984), Orłowska and Pawlak (1984)).

With each set of attributes we associate the uncertainty coefficient, which shows how the set of attributes under consideration contributes to the accuracy of description of a given set of objects.

An example of medical data analysis by means of the introduced concepts is shown at the end of the paper.

2. KNOWLEDGE REPRESENTATION SYSTEM. INDISCERNIBILITY

We recall after Pawlak (1981), the notion of a Knowledge Representation System, which is the departure point of our considerations.

By a Knowledge Representation System we mean a system

$$S = (U, A, V, \varrho)$$

where

U - is a set of objects

A - is a set of attributes

$V = \bigcup_{a \in A} V_a$ - is a set of values of attributes

$\varrho : U \times A \to V$ - is an information function

Set V_a, $a \in A$ will be referred to as domain of the attribute a.

Function $\varrho_x : A \to V$ such that $\varrho_x(a) = \varrho(x,a)$ for every $a \in A$, $x \in U$ will be called information about x in S.

Let B be a non-empty subset of attributes A. We say that objects $x,y \in U$ are B-indiscernible in S, $x \underset{B}{\sim} y$, iff

$$\varrho_x(a) = \varrho_y(a) \quad \text{for every} \quad a \in B$$

Obviously $\underset{B}{\sim}$ is an equivalence relation for any $B \subseteq A$.

Equivalence classes of relation $\underset{B}{\sim}$ are called B-elementary sets in A. A-elementary sets are called simply elementary sets in S.

When $B = \{a\}$, $a \in A$ is a single attribute we shall write $\underset{a}{\sim}$ instead of $\underset{\{a\}}{\sim}$. B-elementary set containing object $x \in U$, will be denoted by $[x]^S_{\underset{B}{\sim}}$, or $[x]_{\underset{B}{\sim}}$ when S is understood.

Let us notice that

$$\underset{B}{\sim} = \bigcap_{a \in B} \underset{a}{\sim}$$

for every $B \subseteq A$.

Subset $X \subseteq U$ will be called a B-definable set in S if X is union of same B-elementary sets in S; an empty set is B-definable for every $B \subseteq A$.

3. APPROXIMATION OF SETS IN KNOWLEDGE REPRESENTATION SYSTEM

Let $S = (U,A,V,\varrho)$ be a Knowledge Representation System, let $X \subseteq U$ and let $B \subseteq A$ $(B \neq \emptyset)$.

A lower B-approximation of X in S ($\underline{B}_S(X)$ or $\underline{B}(X)$ when S is understood) we define as follows:

$$\underline{B}(X) = \{x \in U : [x]_{\underset{B}{\sim}} \subset X\}$$

An upper B-approximation of X in ($\overline{B}_S(X)$ or $\overline{B}(X)$ when S is understood) we mean set

$$\overline{B}(X) = \{x \in U : [x]_{\underset{B}{\sim}} \cap X \neq \emptyset\}$$

In particular when $B = \{a\}$ we write $\underline{a}(X)$, $\overline{a}(X)$ instead of $\underline{\{a\}}\,X$, $\overline{\{a\}}\,X$ respectively.

Some properties of approximations are given in Pawlak(1982).

The number

$$\alpha_B(X) = \frac{|\underline{B}(X)|}{|\overline{B}(X)|}$$

will be called accuracy (uncertainty) coefficient of X with respect to B in S, where $|Y|$ denotes cardinality of Y.

4. SETS SEPARABLE BY SETS OF ATTRIBUTES

If we remove some attributes from set of attributes A in system $S = (U,A,V,\varrho)$ some elementary sets can glue together forming $(A - B)$-elementary sets in S, where B is a subset of removed attributes from A. Union of $(A - B)$-elemen-

tary sets obtained by "glue together" some A-elementary sets when removing subset of attributes B from A, can be defined as follows:

$$S_p(B) = \{x \in U : [x]_{\widetilde{A-B}} \in U / \widetilde{A-B} - U / \widetilde{A}\}$$

Set $S_p(B)$ characterize discrimination power of set of attributes B in S, for set B splits (A - B)-elementary sets from $S_p(B)$ into A-elementary sets.

The number

$$\delta_B = \frac{|S_p(B)|}{|U|}$$

can be used as a measure of the <u>discrimination power</u> of attributes B in S, and will be called <u>discrimination coefficient</u> of B in S accordingly.

Of course $0 \leq \delta_B \leq 1$ for every $B \subseteq A$.

In other words δ_B says what part of set U can be split by attribute set B.

5. SETS DECIDABLE BY SETS OF ATTRIBUTES

Given subset $X \subseteq U$ of objects in system $S = (U, A, V, \varrho)$. We might be interested how subset $B \subseteq A$ of attributes affects the accuracy of approximation of set X in S.

To answer this question we introduce set

$$X_B = Fr_{A-B}(X) - Fr_A(X)$$

called set <u>decidable</u> by set of attributes B in S, where

$Fr_B(X) = \overline{B}(X) - \underline{B}(X)$ is called <u>boundary</u> of X with respect to B in S.

Set X_B is simply set of objects in S membership of which to set X is dependent upon set of attributes B; in other words set X_B says how the boundary region of set X changes when removing set of attributes B from system S.

We can split set X_B into two sets X_B^+ and X_B^- ($X_B = X_B^+ \cup X_B^-$) such that

$$X_B^+ = \underline{A}(X) - \underline{A - B}(X)$$

$$X_B^- = \overline{A - B}(X) - \overline{A}(X)$$

called <u>positively</u> and <u>negatively decidable</u> by B, respectively.

Elements of set X_B^+ are those objects which are positively decided being members of set X, when examining their properties expressed by attributes B; elements of set X_B^- are those objects which are positively excluded being members of set X, on the basis of their properties expressed by attributes B.

We say that set $B \subset A$ is <u>superfluous</u> for X in S, iff $X_B = \emptyset$.

In order to express how the accuracy of approximation of set X is affected by set of attributes $B \subset A$, we may examine how the accuracy of approximation changes when removing set of attributes B from A, i.e. examine the dif-

ference $\eta_{A-B}(X) - \eta_A(X)$.

Another possibility is to examine how the boundary of set X is affected by set of attributes $B \subset A$. In order to do this we introduce the following coefficient

$$\beta_B(X) = \frac{|X_B|}{|Fr_{A-B}(X)|} = 1 - \frac{|Fr_A(X)|}{|Fr_{A-B}(X)|}$$

which simply says what part of the boundary of set X is decidable by set of attributes B in S.

However, the most intuitive way of expressing discernibility power of set of attributes B is to give simply the numbers $|X_B|$, $|X_B^-|$ and $|X_B^+|$. These numbers say how many objects are classified by set of attributes B (positively or negatively) and this information seems to be of primary importance.

6. EXAMPLE

A file of 150 patients suffering from heart disease was investigated. The set of patients was divided by experts into six classes according to their health status; class one contains all patients with the least disease advance and class six contains all patients with the greatest disease advance.

With every patient seven items of information (symptoms) were associated. The problem was whether these symptoms

can be used to define the stage of disease advance with accordance to expert classification.

The results of computation are shown in tables below.

Table 1 contains numbers of objects in lower and upper approximations of each class for different sets of attributes. (We identify symptoms with attributes). Column marked by "non" contains results of computation for the whole set of seven attributes; column marked by number i, contains results of computation for set of attributes without attribute number i. (C - denotes class, and NP - number of patients)

Table 1

C	N.P.	non	1	2	3	4	5	6	7
X_1	10	4 15	1 16	3 16	2 17	4 15	3 16	3 17	4 16
X_2	46	35 54	35 59	85 59	35 60	37 57	33 58	13 57	35 59
X_3	42	39 45	34 52	34 51	33 51	34 49	33 56	32 58	29 53
X_4	33	30 36	24 40	14 40	28 39	29 37	14 39	16 38	22 48
X_5	15	15 15	13 19	13 17	15 15	14 16	14 16	11 20	12 19
X_6	4	4 4	2 5	4 4	4 4	4 4	4 4	0 7	1 6

Table 2 contains accuracy of approximation for each class for the whole set of seven attributes and sets of attributes without one attribute.

Table 2

C	non	1	2	3	4	5	6	7
X_1	0,27	0,06	0,19	0,12	0,27	0,19	0,18	0,25
X_2	0,72	0,59	0,59	0,58	0,65	0,57	0,54	0,59
X_3	0,87	0,65	0,67	0,65	0,69	0,59	0,55	0,55
X_4	0,83	0,60	0,60	0,72	0,78	0,62	0,58	0,46
X_5	1,00	0,68	0,76	1,00	0,88	0,82	0,55	0,63
X_6	1,00	0,40	1,00	1,00	1,00	1,00	0,00	0,17

One can see that attributes 2, 3, 4 and 5 are superfluous for class 6; attribute 3 is superfluous for class 5 and other classes have no superfluous attributes.

In table 3 the cardinalities of boundary regions for corresponding sets of attributes are given.

Table 3

C	non	1	2	3	4	5	6	7
X_1	11	15	13	15	11	13	14	12
X_2	15	24	24	25	20	25	26	24
X_3	6	18	17	18	15	23	26	24
X_4	6	16	16	11	8	15	12	26
X_5	0	6	4	0	2	2	9	7
X_6	0	3	0	0	0	0	7	5

One can easily read from this table which attributes are "most important" for each class. Attributes which give the bigest boundary, when removed from the set of attributes, are most characteristic for the class; thus for example attributes 1, 3 and 6 are most characteristic for class 1; attributes 3, 5 and 6 - for class 2; 5, 6 and 7 - for class 4, and 6, 7 for classes 5 and 6. The same results one can obtain reading table 2.

In table 4 cardinalities of decidable sets for corresponding sets of attributes are shown.

Table 4

C	1	2	3	4	5	6	7
X_1	4	2	4	0	2	3	1
X_2	9	9	10	5	10	11	9
X_3	12	11	12	9	17	20	18
X_4	10	10	5	2	9	6	20
X_5	6	4	0	2	2	9	7
X_6	3	0	0	0	0	7	5

In the next two tables cardinalities of positively and negatively decidable sets of corresponding attributes are given respectively.

Table 5

C	1	2	3	4	5	6	7
X_1	3	1	2	0	1	1	0
X_2	4	4	4	2	6	8	4
X_3	5	5	6	5	6	7	10
X_4	6	6	2	1	6	4	8
X_5	2	2	0	1	1	4	3
X_6	2	0	0	0	0	4	3

Table 6

C	1	2	3	4	5	6	7
X_1	1	1	2	0	1	2	1
X_2	5	5	6	3	4	3	5
X_3	7	6	6	4	11	13	8
X_4	4	4	3	1	3	2	12
X_5	4	2	0	1	1	5	4
X_6	1	0	0	0	0	3	3

Meaning of data in these tables is obvious.

REFERENCES

Pawlak, Z., (1981) Information Systems - Theoretical Foundations, Information Systems, vol. 6 no.3 pp. 205 - 218.

Pawlak, Z., (1982) Rough Sets, International Journal of Information and Computer Sciences, vol. 11 no.5, pp. 341 - 356.

Pawlak, Z., (1984) Rough Classification, International Journal of Man-Machine Studies (to appear).

Orłowska, E. and Pawlak, Z., (1984) Foundations of Knowledge Representation, Springer-Verlag (submitted).