Int. J. Man-Machine Studies (1986) **24**, 413-433

Rough classification of patients after highly selective vagotomy for duodenal ulcer

ZDZISŁAW PAWLAK
Institute of Computer Science, Polish Academy of Sciences, P.O. Box 22, 00-901 Warszawa, Poland

KRZYSZTOF SŁOWIŃSKI
Department of Surgery, F. Raszeja Mem. Hospital, 60-833 Poznań, Poland

AND

ROMAN SŁOWIŃSKI
Institute of Control Engineering, Technical University of Poznań, 60-965 Poznań, Poland

(*Received 7 April 1985*)

The concept of "rough" sets is used to approximate the analysis of an information system describing 77 patients with duodenal ulcer treated by highly selective vagotomy (HSV). The patients are described by 11 attributes. The attributes concern sex, age, duration of disease, complication of duodenal ulcer and various factors of gastric secretion. Two values of sex and age are distinguished, five values of complications and three or four values of secretion attributes, according to norms proposed in this paper. For each patient, the result of treatment by HSV is expressed in the Visick grading which corresponds to four classes. Using the method of rough classification it is shown that the given norms ensure a good classification of patients. Afterwards, some minimum sets of attributes significant for high-quality classification are obtained. Upon analysis of values taken by attributes belonging to these sets a "model" of patients in each class is constructed. This model gives indications for treatment by HSV.

1. Introduction

In this paper we are concerned with "rough" classification of patients with duodenal ulcer treated by highly selective vagotomy (HSV). The HSV consists in vagal denervation of stomach area secreting hydrochloric acid (HCl) (cf. Holle & Anderson, 1974). To be more precise, small branches of nerves of Latarjet (stomach branches of the Vagus) passing on the lesser curve of the stomach (above the incisura) have to be ligated and divided. This denervation results, among others, in lower acid secretion which helps to heal the duodenal ulcer. Contrary to traditional methods of surgical treatment of duodenal ulcer, the HSV does not need gastrotomy and does not remove any part of the stomach. Due to this operation the pyloric area of the stomach is spared with the result that gastric emptying proceeds normally and a drainage operation (such as pyloroplasty, gastroenterostomy and antrectomy) is not required. This is a new and safe treatment method of duodenal ulcer which gives both a low frequency of complications and very low mortality (cf. Dorricot *et al.*, 1978; Dunn *et al.*, 1980; Hautefeuille & Picaud, 1983). For these reasons HSV is presently under intensive study (cf. Goligher *et al.*, 1978).

0020-7373/86/050413+21$03.00/0

In the Department of Surgery of the F. Raszeja Mem. Hospital in Poznań, directed by Professor J. Fibak, this therapy was introduced in 1976. The data concerning 77 cases was collected and analysed statistically by Słowiński (1983). The statistical analysis concerned an investigation of dependencies among the particular attributes describing patients and results of the HSV in the Visick grading, using correlation factors. The observation of these dependencies does not allow, however, the determination of a minimum set of attributes significant for a high-quality classification of patients. For this reason, a tool permitting measurement of the quality of classification depending on the attributes applied has been searched for. The theory of rough classification presented by Pawlak (1984) provides such a tool. It is based on the concept of the rough set introduced by Pawlak (1982).

In section 2 we outline the theory of rough classification. Then, in section 3, we present the information system concerning the set of patients assuming some norms which define domains of particular attributes. The minimum sets of attributes ensuring high-quality classification of patients are calculated in section 4. In section 5, using the minimum sets of attributes we set up a "model" of patients in each class which can be used to foresee the results of HSV for new patients. This also allows for a verification of indications for HSV. The final section gives conclusions.

2. Rough classification

2.1. INFORMATION SYSTEMS

We are given a set of objects and each object under consideration is characterized in terms of some features. Moreover, the set of objects is classified into a disjoint family of classes, and we want to characterize each class in terms of features of objects belonging to this class. In order to do so we exploit the concept of an information system and rough classification following Pawlak (1985).

By an information system we understand the 4-tuple $S=(U, Q, V, \rho)$, where U is a finite set of *objects*, Q is a finite set of *attributes*,

$$V=\bigcup_{q\in Q} V_q$$

and V_q is a *domain* of attribute q, and $\rho: U\times Q\rightarrow V$ is a total function such that $\rho(x, q)\in V_q$ for every $q\in Q$, $x\in U$, called *information function.*

Any pair (q, v), $q\in Q$, $v\in V_q$ is called a *descriptor* in S.

Thus, an information system may be considered as a finite table in which columns are labelled by attributes, rows are labelled by objects and the entry in column q row x has the value $\rho(x, q)$. Each row in the table represents *information* about some object in S. An example of information system is shown in Table 1.

2.2. INDISCERNIBILITY RELATION

Let $S=(U, Q, V, \rho)$ be an information system and let $P\subseteq Q$, $x, y\in U$. We say that x and y are *indiscernible* by set of attributes P in S (denotation $x\tilde{P}y$) iff $\rho(x, q)=\rho(y, q)$ for every $q\in P$. One can easily check that $\tilde{P}$ is an *equivalence* relation on U for every $P\subseteq Q$. Equivalence classes of relation $\tilde{P}$ are called P-*elementary sets* in S. Q-elementary sets are called *atoms* of S.

TABLE 1
An example of information system

U	p	q	r
x_1	1	0	2
x_2	0	1	1
x_3	2	0	0
x_4	1	1	0
x_5	1	0	2
x_6	2	0	0
x_7	0	1	1
x_8	1	1	0
x_9	1	0	2
x_{10}	0	1	1

For example, in the information system from Table 1 $\{p\}$-elementary sets and atoms are as follows:

(a) $\{p\}$-elementary sets:

$X_1 = \{x_1, x_4, x_5, x_8, x_9\}$
$X_2 = \{x_2, x_7, x_{10}\}$
$X_3 = \{x_3, x_6\}$

(b) atoms:

$Z_1 = \{x_1, x_5, x_9\}$
$Z_2 = \{x_2, x_7, x_{10}\}$
$Z_3 = \{x_3, x_6\}$
$Z_4 = \{x_4, x_8\}$

Any finite union of P-elementary sets will be called P-*definable set* in S. Information system S is *selective* iff all atoms in S are one element sets, i.e. $\tilde{Q}$ is an identity relation.

2.3. REPRESENTATION OF AN INFORMATION SYSTEM

A P-*representation* of S is an information system $S_P = (U/\tilde{P}, P, V_P, \rho_P)$, where $U/\tilde{P}$ is the family of all equivalence classes of relation $\tilde{P}$,

$$V_P = \bigcup_{q \in P} V_q$$

and $\rho_P : U/\tilde{P} \times P \to V_P$ is the information function such that $\rho_P([x]_{\tilde{P}}, q) = \rho(x, q)$ for every $x \in U$, $q \in P$ ($[x]_{\tilde{P}}$ denotes an equivalence class of relation $\tilde{P}$ containing object x).

Thus in P-representation of S, objects are P-elementary sets, and information function ρ_P is an extension of function ρ for P-elementary sets. Of course, every P-representation of any S, $P \subseteq Q$, is selective.

2.4. APPROXIMATION OF SETS IN AN INFORMATION SYSTEM

By P-*lower* (P-*upper*) approximation of set X in S, we understand the sets $\underline{P}X$ ($\overline{P}X$) defined as follows

$$\underline{P}X = \{x \in U: [x]_{\tilde{P}} \subset X\}$$

$$\overline{P}X = \{x \in U: [x]_{\tilde{P}} \cap X \neq 0\}$$

Set $Bn_P(X) = \overline{P}X - \underline{P}X$ is referred to as a P-*boundary* of X in S.

It is easy to check that each information system $S = (U, Q, V, \rho)$ and subset of attributes $P \subseteq Q$ defines topological space $T_S = (U, Def_P(S))$, where $Def_P(S)$ is the family of all P-definable sets in S, and the lower and upper approximations are interior and closure in topological space T_S, hence approximations have the following properties:

(1) $\underline{P}X \subseteq X \subseteq \overline{P}X$.
(2) $\underline{P}\emptyset = \overline{P}\emptyset = \emptyset$; $\underline{P}U = \overline{P}U = U$,
(3) $\underline{P}(X \cup Y) \supseteq \underline{P}X \cup \underline{P}Y$,
(4) $\overline{P}(X \cup Y) = \overline{P}X \cup \overline{P}Y$,
(5) $\underline{P}(X \cap Y) = \underline{P}X \cap \underline{P}Y$,
(6) $\overline{P}(X \cap Y) \subseteq \overline{P}X \cap \overline{P}Y$,
(7) $\underline{P}(-X) = -\overline{P}(X)$,
(8) $\overline{P}(-X) = -\underline{P}(X)$.

Moreover, for topological space we have:

(9) $\underline{P}\underline{P}X = \overline{P}\underline{P}X$,
(10) $\overline{P}\overline{P}X = \underline{P}\overline{P}X$.

Set $\underline{P}X$ is called P-*positive region* of set X in S; set $Bn_P(X)$ is called P-*doubtful region* of set X in S and set $U - \overline{P}X$ is called P-*negative region* of set X in S.

Examples of approximations in the information system presented in Table 1 are given below:

Let $X = \{x_1, x_2, x_3, x_6\}$ and $Q = \{p, q, r\}$,

$$\underline{p}X = X_3 = (x_3, x_6\},$$

$$\overline{p}X = X_1 \cup X_2 \cup X_3 = U,$$

$$\underline{Q}X = Z_3 = \{x_3, x_6\},$$

$$\overline{Q}X = Z_1 \cup Z_2 \cup Z_3 = \{x_1, x_2, x_3, x_5, x_6, x_7, x_9, x_{10}\}.$$

2.5. ACCURACY OF APPROXIMATION

With every subset $X \subseteq U$ we associate number $\mu_P(X)$, called an *accuracy of approximation* of set X by P in S, or in short accuracy of X, defined as

$$\mu_P(X) = \frac{\underline{\mu}_P(X)}{\overline{\mu}_P(X)} = \frac{\text{card}\,(\underline{P}X)}{\text{card}\,(\overline{P}X)}$$

Let us notice that $0 \leq \mu_P(X) \leq 1$ and $\mu_P(X) = 1$ if set X is P-definable in S.

2.6. NON-DEFINABLE SETS

Let us notice that set X is P-definable in S iff $\underline{P}X = \overline{P}X$; otherwise set X is non-definable in S and belongs to one of the following classes:

(a) Set X is *roughly* P-definable in S iff $\underline{P}X \neq 0$ and $\overline{P}X \neq U$.
(b) Set X is *internally* P-non-definable in S iff $\underline{P}X = \emptyset$.
(c) Set X is *externally* P-non-definable in S iff $\overline{P}X = U$.
(d) Set X is *totally* P-non-definable in S iff $\underline{P}X = \emptyset$ and $\overline{P}X = U$.

2.7. APPROXIMATION OF FAMILIES OF SETS

Let S be an information system, $P \subseteq Q$, and let $\mathscr{X} = \{X_1, X_2, \ldots, X_n\}$, $X_i \subseteq U$, be a family of subsets of U.

By P-lower (P-upper) approximation of $\mathscr{X}$ in S denoted $\underline{P}\mathscr{X}(\bar{P}\mathscr{X})$, we mean sets $\underline{P}\mathscr{X} = \{\underline{P}X_1, \underline{P}X_2, \ldots, \underline{P}X_n\}$ and $\bar{P}\mathscr{X} = \{\bar{P}X_1, \bar{P}X_2, \ldots, \bar{P}X_n\}$ respectively.

If $\mathscr{X}$ is a *classification* of U, i.e. $X_i \cap X_j = \emptyset$ for every $i, j \leq n$, $i \neq j$ and $\bigcup_{i=1}^{n} X_i = U$, X_i *are called classes* of $\mathscr{X}$.

If every class of $\mathscr{X}$ is P-definable then classification $\mathscr{X}$ will be called P-definable.

If $\mathscr{X}$ is a classification of U then

$$\beta_P(\mathscr{X}) = \frac{\sum_{i=1}^{n} \text{card}(\underline{P}X_i)}{\sum_{i=1}^{n} \text{card}(\bar{P}X_i)}$$

is called the *accuracy of approximation of* $\mathscr{X}$ by P in S, or simply accuracy of classification $\mathscr{X}$.

The coefficient

$$\gamma_P(\mathscr{X}) = \frac{\sum_{i=1}^{n} \text{card}(\underline{P}X_i)}{\text{card}(U)}$$

is called *quality of approximation of classification* $\mathscr{X}$ by set P of attributes, or shortly quality of classification $\mathscr{X}$. It expresses the ratio of all P-correctly classified objects to all objects in the system.

2.8. REDUCTION OF ATTRIBUTES

We say that set of attributes $R \subseteq Q$ *depends* on set of attributes $P \subseteq Q$ in S (denotation $P \rightarrow R$) iff $\tilde{P} \subseteq \tilde{R}$. It can be shown that $P \rightarrow R$ iff $\tilde{P}$-representation of $S_{P \cup R}$ is selective. For example, in order to check whether $\{p, q\} \rightarrow r$ in the information system S given in Table 1, we first compute $\{p, q, r\}$-representation of S, which is the following system $S_{\{p,q,r\}}$:

$U/\widetilde{\{p, q, r\}}$	p	q	r
Z_1	1	0	2
Z_2	0	1	1
Z_3	2	0	0
Z_4	1	1	0

Now, removing the attribute r we obtain $\widetilde{\{p, q\}}$-representation of $S_{\{p,q,r\}}$ which is selective (all rows of this representation are different), thus $\widetilde{\{p, q\}} \rightarrow r$ holds, i.e. $\widetilde{\{p, q\}} \subset \tilde{r}$, which is equivalent to $\widetilde{\{p \cup q\} \cup \{r\}} = \widetilde{\{p, q\}}$.

In the sequel we are using the following definitions:

(a) Set $P \subseteq Q$ is *independent* in S iff for every $P' \subset P$, $\tilde{P}' \supset \tilde{P}$.

(b) Set $P \subseteq Q$ is *dependent* in S iff there exist $P' \subset P$, such that $\tilde{P}' = \tilde{P}$.

(c) Set $P \subseteq Q$ is a *reduct* of Q in S iff P is the greatest independent set in Q.

TABLE 2
Domains of attributes of the POD information system

No.	Attribute	Values 0	1	2	3	4	Remarks	
1	Sex	♂	♀	—	—	—		
2	Age (years)	≤35	>35	—	—	—		
3	Duration of the disease (years)	≤0.5	0.5–3	>3	—	—		
4	Complications of ulcer	none	acute haemorrhage	multiple haemorrhages	perforation	pyloric stenosis		
5	HCl concentration (mmol/100 ml)	≤2	2–4	>4	—	—	♂	
		≤1	1–3	>3	—	—	♀	
6	Volume of gastric juice per hour (ml)	≤70	70–150	>150	—	—	♂♀	basic
7	Volume of residual gastric juice (ml)	≤50	50–100	>100	—	—	♂♀	gastric
8	Basic acid output (BAO) (mmol/h)	≤2	2–3	>3	—	—	♂	secretion
		≤1	1–2	>2	—	—	♀	
9	HCl concentration (mmol/100 ml)	≤10	10–15	>15	—	—	♂	gastric
		≤7	7–10	>10	—	—	♀	secretion
10	Volume of gastric juice per h (ml)	≤100	100–250	>250	—	—	♂♀	stimulated
11	Maximal acid output (MAO) (mmol/h)	≤15	15–25	25–40	>40	—	♂	by
		≤15	15–18	18–30	>30	—	♀	histamine

TABLE 3
The POD information system

No.	Sex	Age	Duration of the disease	Complications of ulcer	Basic gastric secretion: HCl concentration	Basic gastric secretion: Vol. of gastric juice	Basic gastric secretion: Vol. of residual gastric juice	Basic gastric secretion: BAO	Gastric secretion stimulated by histamine: HCl concentration	Gastric secretion stimulated by histamine: Vol. of gastric juice	Gastric secretion stimulated by histamine: MAO	Visick grading
1	1	1	2	0	2	1	0	2	2	1	2	1
2	0	0	1	1	2	0	0	2	2	1	2	1
3	0	0	2	0	2	1	0	2	2	0	0	1
4	0	1	1	0	2	0	0	2	2	1	3	1
5	1	0	0	0	2	1	0	2	2	1	1	3
6	0	0	2	1	2	0	1	2	1	1	1	1
7	0	0	1	1	2	0	0	0	1	0	0	1
8	1	0	1	1	2	0	0	2	2	0	0	1
9	0	1	2	0	2	1	0	2	2	1	3	1
10	0	0	2	3	2	1	0	2	2	2	3	1
11	0	1	1	0	2	2	0	2	2	1	3	1
12	0	1	2	2	1	1	0	0	0	1	0	2
13	0	1	2	3	2	1	0	2	1	1	0	2
14	0	0	1	0	2	0	0	0	2	1	1	1
15	0	1	2	2	2	1	0	2	2	2	3	1
16	0	0	1	1	2	2	0	2	1	1	0	1
17	0	0	2	0	2	2	0	2	2	1	3	1
18	1	1	2	0	2	0	0	2	0	0	0	4
19	0	1	2	0	2	1	0	2	2	2	3	1
20	0	0	2	0	1	2	0	2	0	1	0	2
21	0	0	2	0	0	2	1	0	0	1	0	1
22	0	0	2	4	1	2	2	2	0	1	0	1
23	0	0	0	1	2	2	0	2	0	1	1	1

TABLE 3 (cont.).

No.	Sex	Age	Duration of the disease	Complications of ulcer	Basic gastric secretion: HCl concentration	Basic gastric secretion: Vol. of gastric juice	Basic gastric secretion: Vol. of residual gastric juice	Basic gastric secretion: BAO	Gastric secretion stimulated by histamine: HCl concentration	Gastric secretion stimulated by histamine: Vol. of gastric juice	Gastric secretion stimulated by histamine: MAO	Visick grading
24	0	0	1	3	0	0	0	0	1	1	0	3
25	0	1	1	0	2	1	0	2	2	1	2	1
26	0	0	1	0	2	1	0	2	2	1	2	1
27	0	1	2	0	0	2	1	2	2	0	0	1
28	1	1	1	0	2	0	0	1	1	0	0	1
29	0	0	1	0	1	2	0	2	1	1	0	1
30	0	0	2	0	2	1	0	2	0	1	0	1
31	0	1	1	0	2	1	0	2	1	1	2	1
32	0	1	2	0	1	0	0	0	0	1	0	1
33	1	1	2	1	1	0	2	1	0	0	0	1
34	0	0	1	0	2	2	0	2	2	1	3	1
35	0	0	1	0	2	1	0	2	0	1	2	1
36	0	0	2	0	2	1	0	2	0	2	0	1
37	0	1	1	0	2	0	0	2	0	1	0	1
38	0	0	1	1	0	2	1	1	0	1	0	1
39	0	0	2	0	2	2	0	2	2	2	3	4
40	0	1	2	0	2	2	2	2	0	2	2	1
41	0	0	2	2	0	0	0	0	0	1	0	1
42	0	0	1	0	2	2	0	2	0	1	0	4
43	0	1	1	3	2	2	2	2	0	1	0	3
44	0	1	0	1	2	1	0	2	0	1	0	2
45	0	0	1	2	2	1	1	2	2	1	3	2

46	0	1	2	1	2	1	1	2	2	1	2	1
47	1	0	2	1	2	1	0	2	2	1	1	1
48	0	1	1	0	2	1	0	2	1	1	2	1
49	0	0	2	3	2	0	0	2	1	1	0	1
50	0	0	2	3	1	0	0	2	0	2	1	1
51	1	0	1	3	2	1	1	2	0	1	0	2
52	0	1	2	1	2	0	0	2	1	0	0	1
53	0	0	1	0	2	1	0	2	1	1	0	1
54	0	0	2	0	1	2	0	2	2	1	1	2
55	0	1	2	3	2	1	0	2	1	1	0	1
56	1	0	2	1	2	2	2	2	2	2	3	1
57	0	1	1	1	1	1	0	2	2	1	2	1
58	0	0	0	3	2	2	2	2	1	2	3	1
59	0	1	2	4	2	2	0	2	2	1	3	1
60	1	0	2	0	1	0	1	0	1	1	0	1
61	1	1	1	0	2	0	0	2	2	2	3	1
62	0	0	2	2	2	1	0	2	1	1	1	2
63	0	1	1	3	1	2	1	2	1	2	2	4
64	0	0	1	2	1	1	0	2	1	0	0	2
65	1	0	0	0	2	0	1	2	2	0	0	3
66	1	0	1	0	1	0	0	0	1	0	0	1
67	0	1	2	0	1	1	0	1	0	1	0	1
68	0	0	1	0	2	1	1	2	1	1	0	1
69	0	0	2	0	1	2	1	2	0	1	0	1
70	1	0	2	0	2	0	1	2	2	2	3	1
71	1	1	2	2	1	1	0	2	2	0	0	1
72	0	0	2	0	2	1	0	2	0	1	0	1
73	0	0	2	4	2	0	2	0	0	1	1	2
74	0	1	2	0	0	0	0	0	0	1	0	1
75	1	0	2	0	2	2	2	2	2	2	3	4
76	0	0	2	0	2	1	0	2	1	2	3	1
77	0	0	1	0	1	2	0	2	1	2	2	1

If P is independent in S then for every $p, q \in P$ neither $p \to q$ nor $q \to p$, i.e. all attributes from P are pairwise independent. Also, if P is independent in S, then for every $P' \subset P$, card $(U/\tilde{P}') <$ card $(U/\tilde{P})$. Thus, in order to check whether set $P \subseteq Q$ is independent or not in S it is sufficient to check for every attribute whether or not its removal increases the number of elementary sets in the system. This leads to a very simple algorithm. If P is dependent in S, then there exists $P' \subset P$, independent in S, such that $P' \to P - P'$; the greatest P' is of course a reduct of P in S.

Let us notice that an information system may have more than one reduct. For example, in the information system shown in Table 1 there are three following reducts $\{p, q\}$, $\{p, r\}$ and $\{q, r\}$.

3. The POD information system

In this section we present the information system called POD (*Pre**O*perating *D*ata) concerning 77 patients (objects) with duodenal ulcer treated by HSV. The following set Q of attributes has been used: (1) sex; (2) age; (3) duration of the disease; (4) complications of ulcer; (5) basic HCl concentration; (6) basic volume of gastric juice per hour; (7) volume of residual gastric juice; (8) basic acid output (BAO); (9) HCl concentration under histamine; (10) volume of gastric juice per hour under histamine; (11) maximal acid output (MAO).

Attributes 1–4 concern anamnesis, and the remaining attributes preoperative gastric secretion examined using the histaminic test of Kay (1967). For all attributes, except 1 and 4, we have established some norms resulting from clinical experience; these norms correspond to intervals of "low", "medium" or "high" values. Different norms are assumed for men (♂) and women (♀). Domains of all attributes are shown in Table 2.

The POD information system is presented in Table 3. The last column of the table gives the overall grading of operation results in the modified Visick classification (Goligher *et al.*, 1978). The grading was derived using the following definitions:

(1) **Excellent:** absolutely no symptoms, perfect result.

(2) **Very good:** patient considers result perfect, but interrogation elicits mild occasional symptoms easily controlled by minor adjustment to diet.

(3) **Satisfactory:** mild or moderate symptoms not controlled by care, causing some discomfort, but patient and surgeon satisfied with result, which does not interfere seriously with life or work.

(4) **Unsatisfactory:** moderate or severe symptoms or complications which interfere considerably with work or enjoyment of life; patient or surgeon dissatisfied with result; includes all cases with recurrent ulcer and those submitted to further operation, even though the latter may have been followed by considerable symptomatic improvement.

The Visick grading defines classification $\mathscr{X}$ of set U composed of 77 patients, i.e. $\mathscr{X} = \{X_1, X_2, X_3, X_4\}$, $X_i \cap X_j = \varnothing$ for every $i, j \leq 4$; $i \neq j$ and $\bigcup_{i=1}^{4} X_i = U$. X_i are classes of $\mathscr{X}$.

4. Reduction of attributes and quality of classification

A simple analysis of correlation factors between the attributes and the Visick grading, as well as clinical observation, gives evidence of non-equal importance of the 11 attributes in determining indications for the treatment of duodenal ulcer by HSV (see Słowiński, 1983). The statistical analysis does not allow, however, for a complex evaluation of the importance of a subset of attributes. Using the method of rough classification for the analysis of the POD information system we hoped to find some minimum sets of attributes ensuring a high quality of classification, i.e. the most significant attributes which should be taken into account in order to set up a "model" of patients in every class, which could be used for anticipation of a result of HSV for a new patient.

We applied a computer program developed on the basis of the approach presented in section 2, which (1) computes lower and upper approximations of sets; (2) checks whether a set of attributes is dependent or independent; (3) computes reducts of a set of attributes; and (4) computes the accuracy of approximation and the quality of classification.

4.1. FOUR-CLASS CLASSIFICATION

Let us start with the remark that the POD information system is almost selective. Indeed, the number of atoms equals 74 and only three of them are two-element sets: $\{13, 55\}$, $\{30, 72\}$, $\{31, 48\}$. Moreover, only one two-element atom is composed of patients belonging to different classes: 13 belongs to class 1 and 55 to class 2. This proves that the norms for descriptors proposed by us are well-defined. The norms would have been inappropriately defined if many atoms had been composed of patients belonging to different classes.

TABLE 4
Accuracy of approximation of each class by Q

Class	Number of patients	Lower approx.	Upper approx.	Accuracy
1	58	57	59	0·97
2	10	9	11	0·82
3	4	4	4	1·00
4	5	5	5	1·00

Table 4 shows the accuracy of description of each class by Q. It can be seen that classes 3 and 4 are Q-definable and classes 1 and 2 are roughly Q-definable in POD, although the accuracy of their approximation is very high. The quality of classification by set Q equals 0·95.

According to section 2.8, in order to check whether a set of attributes is dependent or not, we have to remove one attribute at a time and compute the number of elementary sets for each case. If the set, say P, is independent then reduction of one attribute, say r, results in the equality of at least two rows of the reduced information system. These equal rows (P-elementary sets) are clustered together creating a new $(P-\{r\})$-elementary set and thus we get a smaller number of elementary sets. Let us notice, however, that if the clustered rows belong to the same P-lower approximation of set X_i $(i=1,\ldots,n)$ then $\mu_P(X_i)=\mu_{P-\{r\}}(X_i)$ for $i=1,\ldots,n$, and $\beta_P(\mathscr{X})=\beta_{P-\{r\}}(\mathscr{X})$, $\gamma_P(\mathscr{X})=\gamma_{P-\{r\}}(\mathscr{X})$; otherwise the signs of equality are replaced by $\geqslant$.

Proceeding in this way we have found out that set Q of all 11 attributes is dependent and has two following reducts of nine attributes:

2	3	4	5	6	7	9	10	11
1	2	3	4	5	6	7	9	10

Then, we have removed the particular attributes one at a time from these reducts and observed the decrease of the accuracy of classes X_i $(i = 1, \ldots, 4)$ and the quality of classification $\mathscr{X}$. We have done this with the aim of finding the smallest set of attributes which would give a relatively high quality of classification.

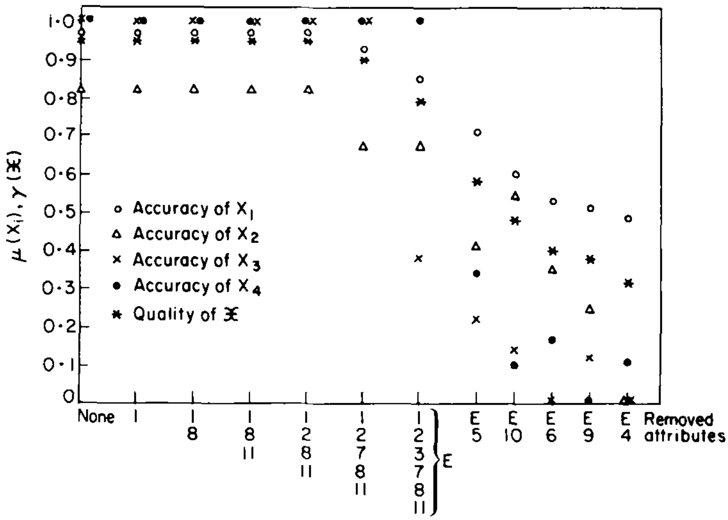

FIG. 1. Accuracy of classes and quality of four-class classification vs removed attributes.

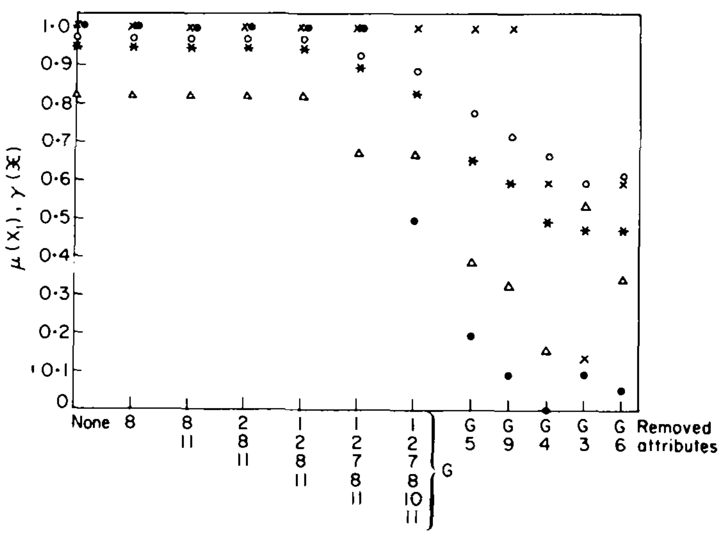

FIG. 2. Accuracy of classes and quality of four-class classification vs removed attributes.

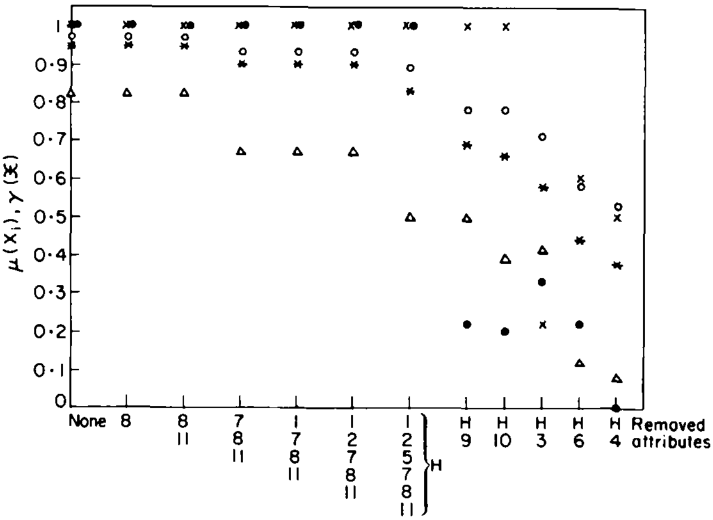

FIG. 3. Accuracy of classes and quality of four-class classification vs removed attributes.

In Figs 1–3, we show three different and most interesting processes of removing attributes from set Q. They are distinguished from other processes by the fact that the decrease of the quality of classification in course of the removal is relatively low. These processes are described in the system of coordinates where the abscissa corresponds to the removed attributes and the ordinate to the accuracy of particular classes and the quality of classification. The list of keys given in Fig. 1 applies to all the three figures. Let us remark that the accuracy of classes as well as the quality of classification holds relatively high until six attributes are removed. In the three processes observed, this corresponds to the cases called E, G and H, respectively. The results for cases E, G and H are presented numerically in Table 5. We can see that in cases G and H quality of classification is the highest, but in case H the accuracy of classes is slightly better than in case G because in H classes 3 and 4 are (Q − H)-definable whereas in G only class 3 is (Q − H)-definable. Moreover, it follows from Figs 2 & 3 that removing one attribute from set Q − H gives (in the worst case) a greater decrease of accuracy and quality than from Q − G. Another important conclusion is that in both cases attributes 3, 4 and 6 have the greatest influence on the accuracy of approximation whereas the two remaining attributes (5, 9 in case G and 9, 10 in case H) have

TABLE 5

Accuracy of classes and quality of classification approximated by a set of five attributes

Case	Removed attributes						Accuracy of classes				Quality of classification
							1	2	3	4	
E	1	2	3	7	8	11	0·85	0·67	0·38	1·00	0·79
G	1	2	7	8	10	11	0·89	0·67	1·00	0·50	0·83
H	1	2	5	7	8	11	0·89	0·5	1·00	1·00	0·83

significantly smaller influence (when they are removed the quality does not decrease below 0·6).

4.2. TWO-CLASS CLASSIFICATION

We have performed a similar computational experiment for aggregated classes 1+2 and 3+4 which correspond to good and bad results of HSV. The aggregated classes are denoted by X'_1 and X'_2, respectively, and the corresponding classification by $\mathscr{X}'$.

Let us remark at the beginning that the aggregation of classes 1 and 2 makes classification $\mathscr{X}'$ Q-definable (see Table 6) since now all two-element atoms are composed of patients belonging to one class only.

TABLE 6
Accuracy of approximation of aggregated classes by Q

Class	Number of patients	Lower approx.	Upper approx.	Accuracy
1+2	68	68	68	1·0
3+4	9	9	9	1·0

The set Q of all 11 attributes is dependent and has the same reducts as set Q for non-aggregated case. However, while removing the attributes one at a time from these reducts, in the same way as in section 4.1, the accuracy of classes X'_i $(i = 1, 2)$ and the quality of classification $\mathscr{X}'$ continues to equal one until five or six attributes are removed. This follows from the fact that a decrease of the number of elementary sets after removing an attribute from a reduct results in clustering together elementary sets including patients of one class only. Three processes of removing attributes from set

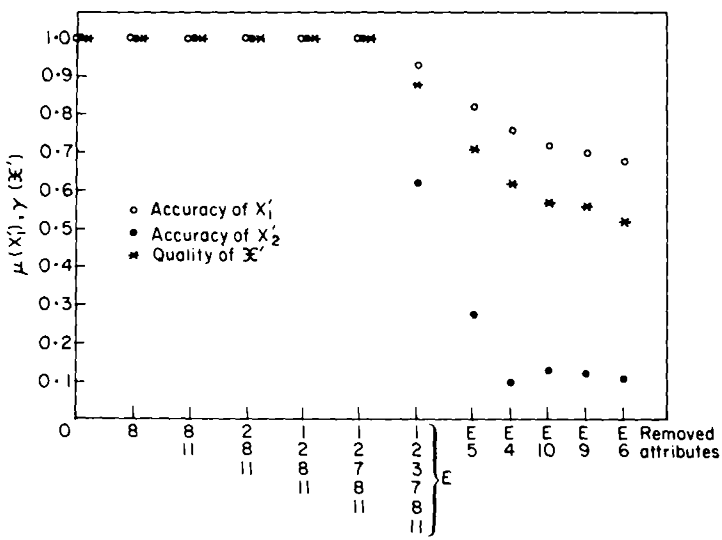

FIG. 4. Accuracy of classes and quality of two-class classification vs removed attributes.

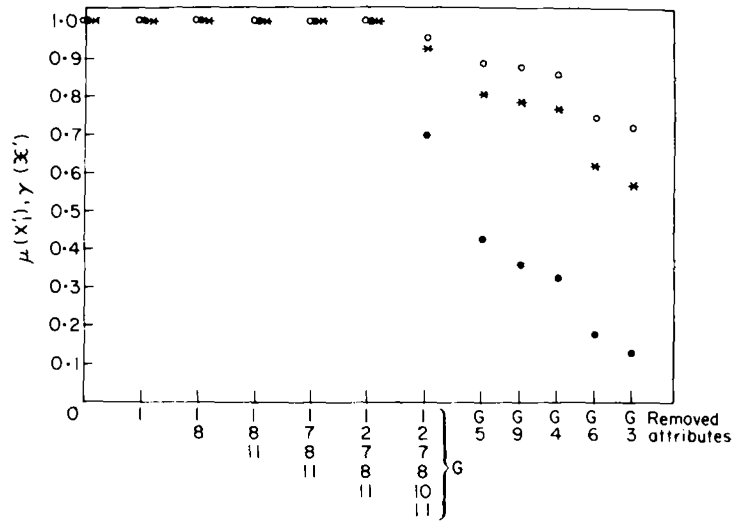

FIG. 5. Accuracy of classes and quality of two-class classification vs removed attributes.

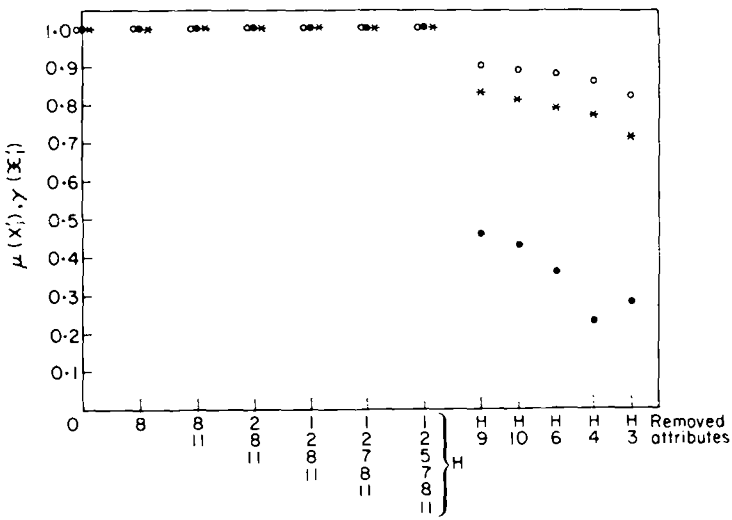

FIG. 6. Accuracy of classes and quality of two-class classification vs removed attributes.

Q are shown in Figs 4–6. The list of keys given in Fig. 4 applies to all the three figures. The numerical results for cases E, G and H are presented in Table 7. We can see that in case H the accuracies of classes and the quality of classification still equal one. Both aggregated classes are (Q − H)-definable and the total number of (Q − H)-elementary sets equals 60; 10 of them are composed of two patients, two of them of three patients and one of four patients, but they all belong to one class in the framework of one composed elementary set. Moreover, from Fig. 6 it follows that by removing one attribute from set Q − H one significantly decreases the accuracy of classes and the quality of classification; attributes 3, 4 and 6 have again the greatest influence.

TABLE 7
Accuracy of aggregated classes and quality of classification approximated by a set of five attributes

Case	Removed attributes						Accuracy of classes 1+2	Accuracy of classes 3+4	Quality of classification
E	1	2	3	7	8	11	0·93	0·62	0·88
G	1	2	7	8	10	11	0·96	0·70	0·93
H	1	2	5	7	8	11	1·00	1·00	1·00

5. Indications for HSV in terms of the most important subset of attributes

The results of the previous section indicate that the most important subset of attributes, denoted by R (N.B. R = Q − H), is composed of:

complications of ulcer (4),
basic volume of gastric juice per hour (6),
duration of the disease (3),
volume of gastric juice per hour under histamine (10), and
HCl concentration under histamine (9).

The above attributes are listed in descending order of the influence on the quality of classification.

In this section we analyse the distribution of values taken by each of the above attributes in the particular classes. We took into account the patients belonging to R-lower approximations of classes only. This distribution is described in Tables 8–12, respectively. The internal part of each table is composed of sectors including three numbers arranged as below:

$$a_{ij}$$
$$b_{ij}[\%] \qquad c_{ij}[\%]$$

where i = number of class, j = value of attribute, a_{ij} = number of patients with descriptor (q, j) belonging to $\underline{R}X_i$ ($q \in R$ is the attribute analysed in a given table), $b_{ij} = a_{ij}/\sum_{i=1}^{4}$ card $(\underline{R}X_i)$, and $c_{ij} = a_{ij}/\sum_{i=1}^{4} a_{ij}$.

The last column in the tables shows characteristic values of the particular attributes following from the distribution. We avoid giving any characteristic value for class i when the maximum values of b_{ij} and c_{ij} are below 5% and 10%, respectively. It can be seen, moreover, that for attributes 9 and 10 we have less "readable" distribution. For this reason, and considering the fact that attributes 9 and 10 have significantly smaller influence on the quality of classification than attributes 4, 6 and 3, we kept only the three latter ones to construct the following "models" of patients belonging to a given class.

Class 1:

without complications of ulcer or acute haemorrhage,
medium or low basic volume of gastric juice per hour,
long or medium duration of the disease.

TABLE 8

Distribution of values takes by attribute 4

Number of class	Values of attribute 4: 0	1	2	3	4	card ($\underline{R}X_i$)	Characteristic value of attribute 4
1	32 46% 84%	13 19% 93%	3 4·3% 43%	4 5·7% 50%	2 2·9% 67%	54	without complications, acute haemorrhage
2	0 0% 0%	1 1·4% 7%	4 5·7% 57%	1 1·4% 13%	1 1·4% 33%	7	multiple haemorrhages
3	2 2·9% 5%	0 0% 0%	0 0% 0%	2 2·9% 25%	0 0% 0%	4	perforation of ulcer
4	4 5·7% 11%	0 0% 0%	0 0% 0%	1 1·4% 13%	0 0% 0%	5	perforation of ulcer, without complications
Total	38	14	7	8	3	70	

TABLE 9
Distribution of values taken by attribute 6

Number of class	Values of attributes 6: 0	1	2	card ($\underline{R}X_i$)	Characteristic value of attribute 6
1	19 27% 83%	22 31% 76%	13 19% 72%	54	medium low
2	1 1·4% 4%	6 8·6% 21%	0 0% 0%	7	medium
3	2 2·9% 9%	1 1·4% 3%	1 1.4% 6%	4	
4	1 1·4% 4%	0 0% 0%	4 5·7% 22%	5	high
Total	23	29	18	70	

TABLE 10
Distribution of values taken by attribute 3

Number of class	Values of attribute 3: 0	1	2	card ($\underline{R}X_i$)	Characteristic value of attribute 3
1	2 2·9% 40%	23 33% 77%	29 41% 83%	54	long medium
2	1 1·4% 20%	3 4·3% 10%	3 4·3% 9%	7	long medium
3	2 2·9% 40%	2 2·9% 7%	0 0% 0%	4	short
4	0 0% 0%	2 2·9% 7%	3 4·3% 9%	5	
Total	5	30	35	70	

TABLE 11
Distribution of values taken by attribute 10

Number of class	Values of attribute 10: 0	1	2	card ($\underline{R}X_i$)	Characteristic value of attribute 10
1	9 13% 75%	33 47% 77%	12 17% 80%	54	medium
2	1 1·4% 8%	6 8·6% 14%	0 0% 0%	7	medium
3	1 1·4% 8%	3 4·3% 7%	0 0% 0%	4	
4	1 1·4% 8%	1 1·4% 2%	3 4·3% 20%	5	high
Total	12	43	15	70	

TABLE 12
Distribution of values taken by attribute 9

Number of class	Values of attribute 5: 0	1	2	card ($\underline{R}X_i$)	Characteristic value of attribute 9
1	15 21% 68%	16 23% 80%	23 33% 82%	54	high medium
2	4 5·7% 18%	2 2·9% 10%	1 1·4% 4%	7	low
3	1 1·4% 5%	1 1·4% 5%	2 2·9% 7%	4	
4	2 2·9% 9%	1 1·4% 5%	2 2·9% 7%	5	
Total	22	20	28	70	

Class 2:
multiple haemorrhages,
medium basic volume of gastric juice per hour,
long and medium duration of the disease.
Class 3:
perforation of ulcer,
short duration of the disease.
Class 4:
perforation of ulcer or without complications of ulcer,
high basic volume of gastric juice per hour.

Proceeding in the same way as above, we have constructed the models for aggregated classes 1+2 and 3+4, corresponding to good and bad results of HSV. We took into account the same set R of attributes which ensures the system to be R-definable (see section 4.2). Again, the distribution for attributes 9 and 10 is not sharp enough to define their characteristic values; thus we have set up the models for aggregated classes using the three remaining attributes: 3, 4 and 6.

Class 1+2:

long or medium duration of the disease,
without complications of ulcer or acute haemorrhage,
medium or low basic volume of gastric juice per hour.

Class 3+4:

short duration of the disease,
perforation of ulcer,
high basic volume of gastric juice per hour.

A comparison of models for classes 1, 2, 3, 4 and 1+2, 3+4 demonstrates the consistency of results for both cases.

These models, however, need an important comment, which follows. The POD information system applies to patients who *have been* operated using HSV, thus the surgeons expected good results of operations on these patients, taking into account clinical experience. Hence, the distribution of patients in classes is not uniform. For these reasons, the results of our analysis may be useful in establishing indications rather than counter-indications for HSV. In other words, the most respresentative models are those of classes 1, 2 and 1+2.

6. Conclusions

Let us summarize the results following from the application of rough classification to the analysis of the information system describing patients after HSV for duodenal ulcer.

(1) The analysis demonstrates that the norms for descriptors proposed in this paper ensure a good classification of patients.

(2) It was shown that from 11 attributes, five of them, namely {3, 4, 6, 9, 10}, provide a high quality of classification, equal to 0.83.

(3) Attributes 9 and 10 have significantly smaller influence on the quality of classification than attributes 3, 4 and 6. For this reason only the latter attributes are used to construct "models" of patients typical for each class.

(4) Attributes 3 and 4 correspond to anamnesis data; attribute 6 is the only important one for classification from among seven attributes concerning gastric secretion examined with the test of Kay. Moreover, attribute 6 belongs to the basic phase of the test where the secretion is not stimulated by histamine.

(5) The "models" of patients typical for classes 1, 2 and 1+2 provide the following indications for HSV in terms of the most important subset of attributes:

without complications of ulcer or with acute haemorrhage,
long or medium duration of the disease,
medium or low basic volume of gastric juice per hour.

To conclude, let us comment on the above results. All the 11 attributes considered in this paper are up to now commonly used by surgeons performing the HSV. It is also a common opinion that they have a non-equal influence on indications for treatment by HSV. It is generally agreed that an absence of complications of the ulcer is a fundamental indication for HSV. Other anamnesis data have been found important as well, but as to attributes corresponding to gastric secretion, no univocal opinion has been voiced. For example, Dorricot *et al.* (1978) did not found any dependence of gastric secretion on results of HSV, using a simple statistical analysis, while Holle & Anderson (1974) hold that attributes concerning gastric secretion are valuable and state that high secretion in the test of Kay is an indication for HSV. Rough sets permit a global analysis of the set of attributes which statistical analysis does not. Using the theory of rough classification we established the minimum set of attributes significant for high-quality classification of patients and, in consequence, we found indications for HSV in terms of these attributes. This is a new result which, although intuitively anticipated by some clinicians, has not been formally proved until now.

The authors wish to thank the referee for his constructive comments on a previous version of this paper.

References

DORRICOTT, N. J., MCNEISH, A. R., ALEXANDER-WILLIAMS, J., ROYSTON, C. M. S., COOKE, W. M., SPENCER, J., DE VRIES, B. C. & MULLER, H. (1978). Prospective randomized multicentre trial of proximal gastric vagotomy or truncal vagotomy and antrectomy for chronic duodenal ulcer: interim results. *British Journal of Surgery*, **65**(3), 152–154.

DUNN, D. C., THOMAS, W. E. G. & HUNTER, J. O. (1980). An evaluation of highly selective vagotomy in the treatment of chronic duodenal ulcer. *Surgery, Gynecology and Obstetrics*, **150**, 845–849.

GOLIGHER, J. C., HILL, G. L., KENNY, T. E. & NUTTER, E. (1978). Proximal gastric vagotomy without drainage for duodenal ulcer: results after 5–8 years. *British Journal of Surgery*, **65**(3), 145–151.

HAUTEFEUILLE, P. & PICAUD, R. (1983). *Les récidives après vagotomies pour ulcère duodénal.* Masson, Paris.

HOLLE, F. & ANDERSON, S. (1974). *Vagotomy.* Berlin, Heidelberg, New York: Springer Verlag.

KAY, A. (1967). *Memorial lecture.* An evaluation of gastric acid secretion tests. *Gastroenterology*, **53**, 834–853.

PAWLAK, Z. (1982). Rough sets. *International Journal of Information and Computer Sciences*, **11**(5), 341–356.

PAWLAK, Z. (1984). Rough classification. *International Journal of Man-Machine Studies*, **20**, 469–483.

PAWLAK, Z. (1985). *Rough sets and some problems of AI.* ICS PAS Reports, no. 565, Warsaw.

SŁOWINSKI, K. (1983). [*Gastric secretion and clinical results of highly selective vagotomy for duodenal ulcer.*] Ph.D. dissertation, Medical Academy, Poznań. [In Polish.]