Int. J. Man–Machine Studies (1984) **20**, 469–483

Rough classification

ZDZISŁAW PAWLAK

Institute of Computer Science, Polish Academy of Sciences, P.O. Box 22, 00-901 Warsaw, PKiN, Poland

(*Received 6 May 1983*)

This article contains a new concept of approximate analysis of data, based on the idea of a "rough" set. The notion of approximate (rough) description of a set is introduced and investigated. The application to medical data analysis is shown as an example.

1. Introduction

This article is concerned with "approximate" classification of objects, based on the concept of a "rough" set introduced in Pawlak (1982). The idea of approximate classification was introduced in Pawlak (1983), where an algorithm for approximate classification was outlined.

The article discusses in more detail the concept of "rough" classification. A program for approximate classification—based on the rough set concept—has been developed (see Fila & Wilk, 1983) and aplied for computer-assisted medical diagnosis. Results of computation are briefly discussed.

We have used standard mathematical notation throughout this paper and we assume that the reader is familiar with basic notions of set theory and topology.

2. Basic notions

2.1. SETS AND THEIR APPROXIMATIONS

In this section we recall after Pawlak (1982) the notion of an upper and a lower approximation of a set, which are basic concepts in our approach to approximate (rough) classification.

By an *approximation space* A we mean ordered pair $A = (U, R)$, where U is a set called the *universe* and R is a binary relation over U, called an indiscernibility relation. We assume that R is an equivalence relation. If $(x, y) \in R$ we say that x and y are indiscernible in A. Equivalence classes of the relation R are called *elementary sets*, or *atoms*, in A. We assume that the empty set is also elementary for every approximation space A.

Any finite union of elementary sets in A will be called a *definable* set in A. The family of all definable sets in A will be denoted by Def (A).

Let $X \subset U$. By an *upper approximation* of X in A, in symbols $\bar{A}X$, we mean the least definable set in A containing set X; by a *lower approximation* of set X in A, in symbols $\underline{A}X$, we mean the greatest definable set in A, contained in X; set $Bn_A(X) = \bar{A}X - \underline{A}X$ will be called a *boundary* of X in A.

0020-7373/84/050469 + 15S03.00/0

2.2. PROPERTIES OF APPROXIMATIONS

Approximation space $A=(U,R)$ defines uniquely the topological space $T_A=(U,\mathrm{Def}(A))$, where $\mathrm{Def}(A)$ is topology for U, and it is the family of open and closed sets in T_A. The family of all elementary sets in A is a base T_a.

The lower and upper approximation of X in A are interior and closure operations respectively in the topological space T_A.

Thus $\underline{A}X$ and $\bar{A}X$ have the following properties:

(A1) $\underline{A}X \subset X \subset \bar{A}X$,
(A2) $\underline{A}U = \bar{A} = U$,
(A3) $\underline{A}\varnothing = \bar{A}\varnothing = \varnothing$,
(A4) $\bar{A}(X \cup Y) = \bar{A}X \cup \bar{A}Y$,
(A5) $\underline{A}(X \cup Y) \supset \underline{A}X \cup \underline{A}Y$,
(A6) $\bar{A}(X \cap Y) = \underline{A}X \cap \underline{A}Y$,
(A7) $\underline{A}(X \cap Y) \subset \underline{A}X \cap \underline{A}Y$,
(A8) $\bar{A}(-X) = -\underline{A}X$ and
(A9) $\underline{A}(-X) = -\bar{A}X$.

Moreover in topological space T_A we have the properties:

(A10) $\underline{A}\underline{A}X = \bar{A}\underline{A}X = \underline{A}X$ and
(A11) $\bar{A}\bar{A}X = \underline{A}\bar{A}X = \bar{A}X$.

2.3. UNDEFINABLE SETS

Let us notice that set X is definable in A iff $\underline{A}X = \bar{A}X$: otherwise set X is *undefinable* in A.

We introduce four classes of undefinable sets in A.

Let X be an undefinable set in A.

(B1) If $\underline{A}X \neq \varnothing$ and $\bar{A}X \neq U$, X will be called *roughly definable* in A.
(B2) If $\underline{A}X \neq \varnothing$ and $\bar{A}X = U$, X will be called *externally undefinable* in A.
(B3) If $\underline{A}X = \varnothing$ and $\bar{A}X \neq U$, X will be called *internally undefinable* in A.
(B4) If $\underline{A}X = \varnothing$ and $\bar{A}X = U$, X will be called *totally undefinable* in A.

Let us give some intuitive meaning of the definitions introduced above.

If set X is roughly definable in A it means that we can define set X with some "approximation", i.e. define its lower and upper approximations in A.

If set X is externally undefinable in A it means that we are unable to exclude any element $x \in U$ being possibly a member of X.

If set X is internally undefinable in A it means that we are unable to say for sure that any $x \in U$ is a member of X.

If set X is totally undefinable it means that we are unable to define even its approximations (both approximations in this case are trivial, i.e. $\underline{A}X = \varnothing$, and $\bar{A}X = U$).

2.4. ACCURACY OF APPROXIMATION

In this section we introduce a measure of accuracy of an approximation of a set in the approximation space A. The measure is defined for finite sets only.

An accuracy measure of set X in the approximation space $A=(U,R)$ is defined as

$$\mu_A(X) = \frac{\underline{\mu}_A(X)}{\bar{\mu}_A(X)} = \frac{\mathrm{card}(\underline{A}X)}{\mathrm{card}(\bar{A}X)}.$$

Instead of $\mu_A(X)$ we shall also write $\mu_R(X)$.

Notice that $0 \le \mu_A(X) \le 1$, and $\mu_A(X) = 1$ if X is definable in A; if X is undefinable in A, then $\mu_A(X) < 1$.

2.5. APPROXIMATION OF FAMILIES OF SETS

Let $A = (U, R)$ be an approximation space and let $F = \{X_1, X_2, \dots, X_n\}$, $X_i \subset U$, be a family of subsets of the universe U.

By lower (upper) approximation of F in A, in symbols $\underline{A}F(\bar{A}F)$, we understand the family

$$\underline{A}F = \{\underline{A}X_1, \underline{A}X_2, \dots, \underline{A}X_n\}$$

and

$$\bar{A}F = \{\bar{A}X_1, \bar{A}X_2, \dots, \bar{A}X_n\}$$

respectively.

If F is a partition of U, i.e.

$$X_i \cap X_j = \varnothing \quad \text{for every } i, j,\ 1 \le i, j \le n,$$

$$\bigcup_{i=1}^{n} X_i = U,$$

we call then F a *classification* of U and X_i are called *classes* or *blocks* of F.

If F is a classification of U we shall write C(U) instead of F, and the corresponding approximations of C(U) in A are denoted by $\bar{A}(C(U))$ and $\underline{A}(C(U))$ or in short $\bar{C}(U)$ and $\underline{C}(U)$ when A is understood.

The number

$$\eta_A C(U) = \operatorname{card}\left(\bigcup_{i=1}^{n} \underline{A}X_i\right) \Big/ \operatorname{card} U$$

will be called the *quality* of the classification $C(U) = \{X_1, \dots, X_n\}$ in A and the number

$$\beta_A C(U) = \operatorname{card}\left(\bigcup_{i=1}^{n} \underline{A}X_i\right) \Big/ \operatorname{card}\left(\bigcup_{i=1}^{n} \bar{A}X_i\right)$$

will be called the *accuracy* of the classification C(U) in A. Instead of $\eta_A C(U)$ and $\beta_A C(U)$ we shall also write $\eta_R C(U)$ and $\beta_R C(U)$, respectively.

3. Information systems and classification

3.1. INFORMATION SYSTEMS

In this section we shall consider a special kind of approximation spaces needed when classifying objects on basis of their properties, and we identify properties with some attributes characteristic of those objects. With each attribute a set of values is associated. Description of an object is given when one value for each attribute is chosen.

The above idea can be expressed more precisely by means of the notion of an information system introduced in Pawlak (1981).

By an information system S we mean an ordered quadruple

$$S = (U, Q, V, \rho),$$

where U is a set called the *universe* of S—elements of U are called objects; Q is a set of *attributes*, $V(=\bigcup_{q\in Q} V_q)$ is a set of *values* of attributes—V_q will be called the *domain* of q and $\rho : U \times Q \rightarrow V$ is a *description function*, such that $\rho(x, q) \in V_q$ for every $q \in Q$ and $x \in U$.

We introduce function $\rho_x : Q \rightarrow V$ such that $\rho_x(q) = \rho(x, q)$ for every $q \in Q$ and $x \in U$; ρ_x will be called the *description* of x in S.

For the sake of simplicity, function ρ_x will be written as a sequence of attribute values $v_{i_1}, v_{i_2}, \ldots, v_{i_n}$ assuming that $v_{i_j} \in V_{q_j}$. Of course, the order of values in this sequnce is immaterial.

We say that objects $x, y \in U$ are *indiscernible* with respect to $q \in Q$ in A, iff $\rho_x(q) = \rho_y(q)$, and we shall write $x_{\tilde q} y$; certainly $_{\tilde q}$ is an equivalence relation. Objects $x, y \in U$ are indiscernible with respect to $P \subset Q$ in S, in symbols $x_{\tilde P} y$, iff $\tilde P = \bigcap_{p\in P} \tilde p$.

In particular, if $P = Q$ we say that x and y are indiscernible in S and write $x_{\tilde S} y$ instead of $x_{\tilde Q} y$.

Obviously, P is an equivalence relation, thus each information system $S = (U, Q, V, \rho)$ defines uniquely an approximation space $A_S = (U, \tilde S)$, where $\tilde S$ is the indiscernibility relation generated by the information system S.

If $x \in U$ and ρ_x is the description of x in S, then we assume that ρ_x is also the description of the equivalence class of the relation $\tilde S$ containing x.

We say that subset $X \subset U$ is describable in S iff X is definable in A_S; if X is undefinable in A_S, X will be called *nondescribable* in S. Description of a describable set in S consists of all descriptions of its elementary sets. Description of an empty set is denoted by ψ.

Example 1

Suppose we are given information system $S = (U, Q, V, \rho)$ where

$$U = \{x_1, x_2, \ldots, x_{10}\}, \qquad Q = \{p, q, r\}, \qquad V_p = \{0, 1, 2\},$$

$$V_q = \{0, 1\}, \qquad V_r = \{0, 1, 2, 3\}$$

and information function ρ is given by the table:

U	p	q	r
x_1	1	0	3
x_2	0	1	1
x_3	0	1	1
x_4	1	1	0
x_5	1	1	0
x_6	2	0	1
x_7	0	1	1
x_8	2	0	1
x_9	2	0	2
x_{10}	1	0	3

There are the following elementary sets in the system:

$$E_1=\{x_1, x_{10}\}, \qquad E_2=\{x_2, x_3, x_7\},$$

$$E_3=\{x_4, x_5\}, \qquad E_4=\{x_6, x_8\}, \qquad E_5=\{x_9\}.$$

For example, sets

$$X_1=\{x_1, x_2, x_3, x_9, x_{10}\}=E_1\cup E_2$$

and

$$X_2=\{x_2, x_3, x_4, x_5, x_6, x_7, x_8\}=E_2\cup E_3\cup E_4$$

are describable in S, and sets

$$X_3=\{x_1, x_2, x_3, x_7, x_8\} \quad \text{and} \quad X_4=\{x_1, x_3, x_9\}$$

are nondescribable in S.

We can introduce the following four classes of nondescribable sets in a information system S.

Let $X\subset U$ be a nondescribable set in S. Then:

(C1) if X is roughly definable in A_S then X is called *roughly describable* in S;

(C2) if X is externally undefinable in A_S, then X is called *externally nondescribable* in S;

(C3) if X is internally undefinable in A_S, then X is called *internally nondescribable* in S and

(C4) If X is totally undefinable in A_S, then X is called *totally nondescribable* in S.

The meaning of these definitions is obvious. They simply say that there are several grades of nondescribability, from approximate describability to total nondescribability. In other words, if we are given some properties (attributes) of an object, and we want to characterize a subset of objects by means of these properties, the task can end in failure, because only describable sets can be uniquely characterized by a given set of attributes.

Example 2

Let us consider the information system as in example 1. Then set

$$Y_1=\{x_1, x_2, x_4, x_5\}$$

is roughly describable in S; set

$$Y_2=\{x_1, x_2, x_3, x_4, x_6, x_9\}$$

is externally nondescribable in S and set

$$Y_3=\{x_1, x_2, x_5, x_8\}$$

is internally nondescribable in S.

There are no totally nondescribable sets in this system.

3.2. ATTRIBUTE DEPENDENCES AND REDUCED INFORMATION SYSTEMS

By means of the indescernibility relation, we can easily define some important features of information systems; first of all the most important one—dependency at attributes.

Let $S=(U, Q, V, \rho)$ be an information system and let $p, q \in Q$.

(a) Attribute p is said to be *dependent* on attribute q in S, $(q \rightarrow p)$ iff $\tilde{q} \subset \tilde{b}$.
(b) Attributes p, q are called *independent* in S iff neither $p \rightarrow q$ nor $q \rightarrow p$ hold.

The meaning of these two definitions is obvious. For more details see Pawlak (1981).

Example 3
Consider the information system $S=(U, Q, V, \rho)$ such that $U=\{x_1, x_2, x_3, x_4, x_5\}$, $Q=\{q_1, q_2, q_3, q_4\}$, $V_{q_1}=\{0, 1\}$, $V_{q_2}=\{0, 1\}$, $V_{q_3}=\{0, 1\}$, $V_{q_4}=\{0, 1, 2\}$, and function ρ given by the table:

U	q_1	q_2	q_3	q_4
x_1	0	0	0	0
x_2	0	1	0	2
x_3	1	1	0	1
x_4	1	1	0	1
x_5	0	1	1	2

It is easy to see that $q_4 \rightarrow q_2$ and $q_4 \rightarrow q_1$ because $\tilde{q}_4 \subset \tilde{q}_2$ and $\tilde{q}_4 \subset \tilde{q}_1$.
For a later purpose we introduce some new definitions.

(c) A subset $P \subset Q$ is said to be *independent* in S iff for every $P' \subset P$, $\tilde{P}' \supset \tilde{P}$.
(d) A subset $P \subset Q$ is said to be *dependent* in S iff there exists a $P' \subset P$ such that $\tilde{P}' = \tilde{P}$.
(e) A subset $P' \subset P$ is said to be *superfluous* in P iff $\widetilde{P-P'} = \tilde{P}$.
(f) A subset $P \subset Q$ is called *reduct* of Q in S iff $Q-P$ is superfluous in Q and P is independent in S; the corresponding system $S'=(U, P, V, \rho')$ is called *reduced system* (ρ' is the restriction of ρ to set $U \times P$).

Example 4
In the information system considered in example 3 the set of attributes Q is dependent in S and sets $\{q_1, q_2, q_3\}$, $\{q_3, q_4\}$ are reducts of Q.

Note that a system can have more than one reduct!

Now we give some properties of attributes, which enable us to simplify the decision procedure whether the set of attributes is dependent or not, and the procedure for finding reducts of the set of attributes. The proofs are by simple computation.

Fact 1. If a set of attributes Q is independent in S then all its different attributes are pairwise independent in S.

Fact 2. Subset $P \subset Q$ is dependent in S iff there exists $P' \subset P$ such that P' is superfluous in P.

Fact 3. If $P \subset Q$ is independent in S then every $P' \subset P$ is also independent in S.

Fact 4. If $P \subset Q$ is dependent in S, then for every $P' \supset P$ and $P' \subset Q$, P' is dependent in S.

Let $P=\{p_1, p_2, \ldots, p_n\}$, $P \subset Q$ and let $P_i = P - \{p_i\}$. $1 \leq i \leq n$.

Fact 5. Set $\mathrm{P} \subset \mathrm{Q}$ is independent in S iff for every i $(1 \le i \le n)\tilde{\mathrm{P}}_i \supset \tilde{\mathrm{P}}$.

Fact 6. Set $\mathrm{P} \subset \mathrm{Q}$ is independent in S iff for every i $(1 \le i \le n)$ $\mathrm{card}(\mathrm{U}/\tilde{\mathrm{P}}_i) < \mathrm{card}(\mathrm{U}/\tilde{\mathrm{P}})$.

By Facts 5 and 6 in order to check whether set $\mathrm{P} \subset \mathrm{Q}$ is independent or not in S it is enough to check for every attribute whether removing of this attribute increases the number of elementary sets or not in the system. This leads to very simple algorithm.

If the set of attributes is dependent we can be interested in finding all reduced systems.

The reduction algorithm can be based on the following property.

Fact 7. If $\mathrm{P} \subset \mathrm{Q}$ is superfluous in Q and $\{p\}$ is superfluous in $\mathrm{Q}-\mathrm{P}$, then $\mathrm{P} \cup \{p\}$ is superfluous in Q.

By this property we can eliminate superfluous attributes step by step from the system; after exhausting all possible patterns of reduction we get all reducts of Q in S.

In order to explain the above ideas in more detail let us first define the notion of *representation* of an information system.

Let $\mathrm{S} = (\mathrm{U}, \mathrm{Q}, \mathrm{V}, \rho)$ be an information system. The system $\mathrm{S}^* = (\mathrm{U}/\tilde{\mathrm{S}}, \mathrm{Q}, \mathrm{V}, \rho^*)$ will be called the *representation* of S where

$$\rho^* : \mathrm{U}/\tilde{\mathrm{S}} \times \mathrm{Q} \to \mathrm{V}$$

and

$$\rho^*(\mathrm{X}, q) = v, \qquad \mathrm{X} \in \mathrm{U}/\tilde{\mathrm{S}}, \qquad q \in \mathrm{Q}$$

iff

$$\rho(x, q) = v$$

for all $x \in \mathrm{X}$.

In other words, if we omit all duplicate rows in the table of function ρ and replace objects by elementary sets containing these objects so we obtain the representation of the system.

Example 5

Let us consider information system as in example 3, i.e.

U	q_1	q_2	q_3	q_4
x_1	0	0	0	0
x_2	0	1	0	2
x_3	1	1	0	1
x_4	1	1	0	1
x_5	0	1	1	2

For the sake of simplicity throughout the remainder of this article we will identify the notion of the information system with the table of the information function.

The representation of this system has the form

$U/\tilde{S}$	q_1	q_2	q_3	q_4
$\{x_1\}$	0	0	0	0
$\{x_2\}$	0	1	0	2
$\{x_3, x_4\}$	1	1	0	1
$\{x_5\}$	0	1	1	2

Thus each row in the table is the description of an elementary set, and we can treat the whole table as the description of the whole information system.

In order to simplify the notation the above table will be also presented as follows:

1	2	3	4
0	0	0	0
0	1	0	2
1	1	0	1
0	1	1	2

The set of attributes in this system is dependent because by removing attribute 4 we obtain the system

1	2	3
0	0	0
0	1	0
1	1	0
0	1	1

with the same number of elementary sets as the original system.

After removing attribute 3 from the last system we obtain the system

1	2
0	0
0	1
1	1
0	1

in which the second and the fourth rows are the same, which means that the second and the fourth elementary sets are "glued" together and in this way we get a smaller number of elementary sets so attribute 3 is not superfluous. Proceeding in this way we get that 1, 2, 3 and 3, 4 are the only reducts of set of attributes 1, 2, 3, 4.

The corresponding reduced systems are the following

1	2	3
0	0	0
0	1	0
1	1	0
0	1	1

3	4
0	0
0	2
0	1
1	2

One can easily see that each elementary set in these systems has a different description and that removing any attribute (column) from the system changes this property.

If a set of attributes Q is independent in S and we remove subset P from Q, then we obtain independent set of attributes Q−P in S again. If the relations $\tilde{Q}$ and $\widetilde{Q-P}$ differ "a little", we can say that set P is roughly superfluous in S.

More exactly, we say that set $P \subset Q$ is *$\mathscr{C}$-superfluous* in S iff

$$\mu_{\tilde{Q}}(X) - \mu_{\widetilde{Q-P}}(X) \leq \mathscr{C}$$

for every $X \subset U$, and consequently we say that $P \subset Q$ is *$\mathscr{C}$-reduct* of Q in S iff

$$\mu_{\tilde{Q}}(X) - \mu_{\tilde{P}}(X) \leq \mathscr{C}$$

for every $X \subset U$.

To this end let us remark that sometimes we are interested in removing superfluous ($\mathscr{C}$-superfluous) attributes not for the whole set of objects U, but for a certain subset X of U.

In such a case we can simply use the same methods as before assumming only that the universe of the system is not U but X.

4. Example of application

4.1. THE PROGRAM

On the basis of the presented approach a program has been developed (see Fila & Wilk, 1983) which

(i) computes lower and upper approximations of sets,
(ii) checks whether a set of attributes is dependent or independent,
(iii) computes reducts of a set of attributes and
(iv) computes accuracy of approximation.

The program is very simple and contains about 200 lines in Fortran.

4.2. MEDICAL DIAGNOSIS

As an example, the program has been used for medical data analysis.

A file of 150 patients suffering from heart disease seen in one of the hospitals in Warsaw was used as a data base. All patients have been divided by experts into six classes corresponding to their health status.

With every patient seven items of information (attributes) were associated. For the sake of simplicity attributes were numbered 1, 2, 3, 4, 5, 6 and 7, and their domains were $V_1 = V_2 = V_3 = V_4 = V_5 = \{0, 1, 2\}$, $V_6 = \{0, 1, 2, 3, 4\}$ and $V_7 = \{0, 1, 2, 3\}$.

The problem was to find the description of each class in terms of data available for each patient of this class, check whether the set of attributes is dependent or independent, find reducts for each class, and compute accuracy of descriptions.

4.3. APPROXIMATIONS AND ACCURACY

There were 125 elementary sets in the system under consideration (104 one-element sets, 19 two-element sets, one three-element set and one five-element set).

The table below shows the accuracy of description of each class:

Class number	Number of patients	Lower approx.	Upper approx.	Accuracy
1	10	4	15	0·27
2	46	33	54	0·72
3	42	39	45	0·87
4	33	30	36	0·83
5	15	15	15	1·00
6	4	4	4	1·00

We see that classes 5 and 6 are describable in the system, and the remaining classes are *roughly* describable with the accuracy given in the last column. That is to say that data (symptoms) available from the patients characterize exactly classes 5 and 6 only, and the remaining classes not are characterized exactly by these data; especially class 1 has very low accuracy.

The quality of the whole classification is 0·87 and the accuracy of the whole classification is 0·95 (see section 2.5).

For the sake of simplicity we show only approximations for class 1.

Lower approximation

1	2	3	4	5	6	7
0	1	0	0	1	1	0
0	1	2	0	1	1	0
1	0	0	0	0	1	0
2	0	0	0	1	1	0

Upper approximation

1	2	3	4	5	6	7
0	0	0	0	0	1	0
0	0	0	0	1	1	0
0	0	0	0	1	0	0
0	0	0	0	1	1	1
0	1	0	0	1	1	0
0	1	2	0	1	1	0
1	0	0	0	0	1	0
2	0	0	0	1	1	0

The boundary

1	2	3	4	5	6	7
0	0	0	0	0	1	0
0	0	0	0	1	1	0
0	0	0	0	1	0	0
0	0	0	0	1	1	1

4.4. INDEPENDENCE OF ATTRIBUTES

According to Fact 6 (section 3.2) in order to check whether the set of attributes is dependent or not we have to remove one attribute step-by-step and compute the number of elementary sets for each case.

The results of computation are:

	Removed attribute							
	none	1	2	3	4	5	6	7
Number of elementary sets	125	106	111	113	118	106	100	101

Because the number of elementary sets is always smaller than 125 that means that set of attributes is independent, and consequently all different attributes are pairwise independent.

In the next table we give the accuracy of approximation for each class when removing one attribute.

Class number	Removed attribute							
	none	1	2	3	4	5	6	7
1	0·27	0·06	0·19	0·12	0·27	0·19	0·18	0·25
2	0·72	0·59	0·59	0·58	0·65	0·57	0·54	0·59
3	0·87	0·65	0·67	0·65	0·69	0·59	0·55	0·55
4	0·83	0·60	0·60	0·72	0·78	0·62	0·68	0·46
5	1·00	0·68	0·76	1·00	0·88	0·82	0·55	0·63
6	1·00	0·40	1·00	1·00	1·00	1·00	0·00	0·17

It is easily seen from this table how the attributes influence accuracy of description. For example, removing attribute 4 gives the smallest changes in accuracy. The accuracy without attribute 4 differs at most about 0·18. So we can say that attribute 4 is $\mathscr{C}$-superfluous for the classification ($\mathscr{C} = 0{\cdot}18$).

4.4. REDUCTION OF ATTRIBUTES

In this section we will show reducts of some classes, i.e. minimal sets of attributes necessary for the description of these classes.

Let us first consider class 5 which is describable and has the following description:

1	2	3	4	5	6	7
0	0	2	0	2	3	3
0	0	2	2	0	4	2
0	1	2	1	2	2	2
0	2	1	2	1	0	0
0	2	2	1	0	2	3
0	2	2	2	2	3	3
0	2	2	2	2	4	3
1	2	2	2	1	3	3
1	2	2	2	2	2	1
2	2	1	1	1	3	3
2	2	2	2	1	3	2
2	2	2	2	1	3	3
2	2	2	1	2	2	3
2	2	2	2	2	3	0
2	2	2	2	2	4	2

By Fact 7 (section 3.2) we can compute that attributes 2, 3 and 5 are superfluous for class 5 in the system and we can have the following description of class 5:

1	4	6	7
0	0	3	3
0	1	2	2
0	1	2	3
0	2	0	0
0	2	3	3
0	2	4	2
0	2	4	3
1	2	2	1
1	2	3	3
2	1	3	3
2	2	2	3
2	2	3	0
2	2	3	2
2	2	3	3
2	2	4	2

If we consider nondescribable class, for example class 1, then we get the following descriptions:

Lower approximation

1	2	3	4	5	6	7
0	1	0	0	1	1	0
0	1	2	0	1	1	0
1	0	0	0	0	1	0
2	0	0	0	1	1	0

Upper approximation

1	2	3	4	5	6	7
0	0	0	0	0	1	0
0	0	0	0	1	0	0
0	0	0	0	1	1	0
0	0	0	0	1	1	1
0	1	0	0	1	1	0
0	1	2	0	1	1	0
1	0	0	0	0	1	0
2	0	0	0	1	1	0

The boundary

1	2	3	4	5	6	7
0	0	0	0	0	1	0
0	0	0	0	1	0	0
0	0	0	0	1	1	0
0	0	0	0	1	1	1

Reducts of the lower approximation, upper approximation and the boundary are $\{1, 3\}$, $\{1, 2, 3, 5, 6, 7\}$ and $\{5, 6, 7\}$, respectively.

Consequently we have the following descriptions of these sets:

Lower approximation

1	3
0	2
2	0
0	0
1	0

Upper approximation

1	2	3	5	6	7
0	0	0	0	1	0
0	0	0	1	0	0
0	0	0	1	1	0
0	0	0	1	1	1
0	1	0	1	1	0
0	1	2	1	1	0
1	0	0	0	1	0
2	0	0	1	1	0

The boundary

5	6	7
0	1	0
1	0	0
1	1	0
1	1	1

4.6. COMBINED CASE

Sometimes we can be interested in combining same classes together, for example in our case classes 1 and 2, and 5 and 6.

In this case we obtain the following results:

Class number	Number of patients	Lower approx.	Upper approx.	Accuracy
1′	56	54	58	0·93
2′	42	39	45	0·87
3′	33	30	36	0·83
4′	19	19	19	1·00

We see that now the classification is much better described by the attributes than in the previous example.

The quality and accuracy of this classification are both 0·95.

In the table below we give results of computation showing how the accuracy of class description changes when removing one attribute from the system.

Class number	Removed attribute							
	none	1	2	3	4	5	6	7
1′	0·93	0·85	0·82	0·84	0·85	0·80	0·74	0·81
2′	0·87	0·65	0·67	0·65	0·69	0·59	0·55	0·55
3′	0·83	0·60	0·60	0·72	0·78	0·62	0·68	0·46
4′	1·00	0·63	0·71	1·00	0·90	0·90	0·40	0·52

5. Conclusion

The proposed method can be viewed as a new approach to approximate data analysis, especially in approximate classification, approximate clustering, approximate learning algorithms, etc.

Thanks are due to Professor M. Warmus for providing medical data to experimental computations and valuable discussions. I am also indebted to Dr E. Orłowska and Dr E. Pleszczyńska for reading the manuscript and helpful comments and remarks.

References

FILA, I. & WILK, M. (1983). Implementation of the algorithm for approximate classification (to be published).

KONRAD, E., ORŁOWSKA, E. & PAWLAK, Z. (1981). *An Approximate Concept Learning*, Bericht 81-7. Berlin.

PAWLAK, Z. (1981). Information systems, theoretical foundations. *Information Systems*, **6**(3), 205–218.

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

PAWLAK, Z. (1983). Classification of objects by means of attributes. *International Journal of Information and Computer Sciences* (to appear).

TU-HUE LE (1982). *Approximative Mustererkennung*. Technische Universität, Berlin.