Studien zur Klassifikation, Bd. 19 (SK 19)

Klassifikation und Ordnung

Tagungsband
12. Jahrestagung der Gesellschaft für Klassifikation e.V.
Darmstadt, 17. – 19. März 1988

Herausgeber: Rudolf WILLE

Frankfurt/Main INDEKS VERLAG 1989

Studien zur Klassifikation, Bd. 19 (SK 19)

Herausgeber

Gesellschaft für Klassifikation e. V.

Zdzisław Pawlak

APPROXIMATE CLASSIFICATION AND ROUGH SETS

Contents:

0. Introduction

The problem I am going to deal with is the following one. Suppose we are given a finite set of objects U and assume that each object is characterized by some preassumed features. Moreover let us assume that the objects are classified (partitioned) into some classes $X_1, X_2, \ldots, X_n$. The problem we are going to discuss consists in finding the characterization of each class X_i of the partition in terms of objects features belonging to the class.

In other words we ask whether the knowledge of features of object x enable us to classify x to proper class X_i. Thus our problem reduces to the membership question, i.e. whether the object x belong to a given set X or not.

For example, suppose that the objects are patients suffering from a certain disease and each patient is characterized in terms of some symphtoms, like body temperature, blood preasure etc. Suppose moreover that patients are classified according to age into three classes, young, middle and old. The question arises whether there are specific symphtoms of the disease for each class of age.

Very many problems, in particular in artificial intelligence, (like pattern recognition, machine learning, inductive inference, and others) - can be reduced to the above scheme.

We shall discuss briefly the problem in the framework of the rough sets theory (see (5)), which seems to be very well suited to discuss this kind of questions - and has found many real-life applications (see (1,3,5,7,8)).

Studien zur Klassifikation, Bd.19(1989)p.186-190

1. Information Systems

In this section we are going to show that our original problem can be reduced to the analysis of a certain kind of data table, called here an information system (see(4)).

The information system can be also considered as a context matrix introduced and investigated by R. Wille in concept analysis (see(8)).

Example of such a table is shown below.

Patient	Gasometry	Dyspnea	Pulmonary stasis	Heart rate	Hepato-megaly
p_1	37	1	1	82	0
p_2	43	2	4	76	8
p_3	42	1	1	71	1
p_4	49	0	2	80	5
p_5	48	1	3	92	6

Table 1.

In the table 1 an excerpt from a medical data file is given. Objects are patients $p_1, p_2, \ldots$ which are described in terms of some symphtoms.

Generaly speaking an information system is a finite table columns of which are labelled by "attributes", and rows - by "objects". For each objects and attribute an attribute-value is uniquelly associated, in the table.

For example, heart rate for patiant p_4 is 80.

Formally any *information system* can be presented as $S=(U,A,V,f)$, where U is a finite set of *objects*, A - finite set of *attributes*, $V = \bigcup_{a \in A} V_a$, V_a - *set of values* of attribute a (*domain of* a) and $f: U \times A \to V$ is an *information function* such that $f(x,a) \in V_a$ for any $a \in A$.

It is easy to see that some objects may have the same values of attributes, i.e. they are indiscernible by those attributes. In other words every subset $B \subseteq A$ of attributes generates an indiscernibility relation $\tilde{B}$ - which is an equivalent relation. That is to say that each subset of attributes B generates a partition B^* of all objecs and blocks of the partition B^* are

equivalence classes of the indiscernibility relation $\tilde{B}$.
Now our problem can be formulated as follows: given an information system $S = (U, A, V, f)$ and a partition D^* of U generated by a certain subset of attributes $D \subseteq A$. Give the description of each class of the partition D^* in terms of a subset of attributes $B \subseteq A$.
In turns out that in general case this problem can not be solved positively. That means that in some cases classes X_i of the partition D^* can be described with some approximation only, employing the set of attributes $B \subseteq A$. We shall consider this problem with some detail in the next section.

2. Approximation of Sets

Because we are interested in characterization of blocks of some partition let us first consider characterization of single subset $X \subseteq U$ in terms of some attributes $B \subseteq A$. To this end we introduce the concepts of a lower and upper approximation of $X \subseteq U$ by $B \subseteq A$, denoted $\underline{B}X$ and $\overline{B}X$ respectively, and defined as below:

$$\underline{B}X = \bigcup \{Y \in B^* : Y \subseteq X\}$$

$$\overline{B}X = \bigcup \{Y \in B^* : Y \cap X \neq \emptyset\}.$$

Thus $\underline{B}X$ is the set of all objects which can be shurely classified to X, employing the set of attributes B, and $\overline{B}X$ is the set of all objects which possibly belong to X, i.e. those elements which can not be excluded being elements of X. Of course if $\underline{B}X = \overline{B}X$, then the set X can be uniquely characterized (described) by B. Set of objects which can not be characterized by B are refered to as rough sets (with respect to B).
Thus with each set $X \subseteq U$ we can associate three important regions: $\underline{B}X$ - *B-positive region of X*, $\overline{B}X - \underline{B}X$ - *B-boundary region of X*, and $U - \overline{B}X$ - *B-negative region of X*.
Objects belonging to either B-positve or B-negative region of X can be shurely classified to either X or $-X$, but elements belonging to the boundary region can not be properly classified to neither X nor $-X$ - employing the set of attributes B.

Thus in the general case we can classify elements of U with some approximation only employing some preassumed set of

attributes B - specifying the lower and upper approximation of the set.

3. Partial Dependency of Attributes

Now we are about to present our main claim. The problem of the classification of objects to classes of the partition D^* by means of the set of attributes $C \subseteq A$, reduces to the question whether the set of attributes D depends functionally on the set of attributes C (usually C and D are refered to as a *condition* and *decision attributes*, respectively).

Functional dependency means that the values of the condition attributes C uniquely define the values of the decision attributes D. If it is the case, it is obvious that blocks D^* can be uniquely characterized in terms of condition attributes C; otherwise the characterization is impossible.

In order to express the above considerations more precisely we introduce formally the concept of a *partial dependency of attributes*.

Suppose we are given an information system $S=(U,A,V,f)$ and the subsets of attributes $C,D \subseteq A$. We say that $C \Rightarrow^k D$ $(0 \leq k \leq 1)$, which reads: "D depends in degree k on C" if

$$k = \frac{card(POS_C(D^*))}{card(U)}$$

where $POS_C(D^*) = \bigcup_{Y \in D^*} \underline{C}Y$, is the C-positive region of the classification D^* i.e. the set of all elements of U which can be properly classified to the blocks of the partition D^* employing the set of attributes C. Thus if $k=1$, all elements of U can be properly classified, if $k=0$ non of the elements of U can be properly classified and if $0<k<1$ some elements can be properly classified only and the number k, called the *accuracy of the classification*, gives the ratio of elements which can be properly classified. Thus in this case we are able to give only the lower and upper approximation of each class.

In this way the problem of approximate classification has been reduced to the investigation of partial dependency of attributes in an information system.

4. Conclusion

The approach to the approximate classification presented above induced a wide class of theoretical problems, which are currently investigated by many authors in Poland and abroad. Several succesfull applications of the discussed ideas have proved the approach both useful and interesting.

References

(1) **Arciszewski, T. and Ziarko, W.:** Adaptive Expert System for Preliminary Engineering Design. *Proc. of the 6th International Workshop on Expert Systems and their Applications*, Avignon, France (1986) pp. 695-712.

(2) **Dubois, D. and Prade, H.:** Twofold Fuzzy Sets and Rough Sets - Some Issues in Knowledge Representation. *Fuzzy Sets and Systems* **23** (1987) pp. 3-18.

(3) **Mrózek, A.:** Rough Sets and Dependency Analysis among Attributes in Computer Implementation of Experts Inference Model. *Proc. the 7th International Workshop on Expert Systems and their Applications*, Avignon, France (1987) pp. 597-611.

(4) **Pawlak, Z.:** Information Systems-Theoretical Foundations. *Information Systems* **6** (1981) pp. 205-218.

(5) **Pawlak, Z.:** Rough Sets. *International Journal of Information and Computer Sciences* **11** (1982) pp. 341-356.

(6) **Pawlak, Z., Slowiński, K. and Slowiński, R.:** Rough Classification of Patients after Highly Selective Vagotomy for Duodenal Ulcer. *International Journal of Man - Machine Studies*, **24** (1986) pp. 413-433.

(7) **Yazdi, R. and Ziarko, W.:** Conceptual Scheme Design: A Machine Learning Approach. *Proc. of the Sec. International Symposium on Methodologies for Intelligent Systems*. Charlotte, USA (1987) pp. 379-398.

(8) **Wille, R.:** Liniendiagramme Hierarhischer Begriffs- systeme. *Technische Hochschule, Preprint No* 812, Darmstad 1984.

(9) **Wong, S.K.M. and Ziarko, W.:** INFER - An Adaptive Decision Support System Based on the Probabilistic Approximate Classification. *Proc.of the 6th International Workshop on Expert Systems and their Applications*, Avignon, France (1986), pp. 713-726.