Attribute Discovery and Rough Sets

Jarosław Stepaniuk
Institute of Computer Science
Bialystok University of Technology
Wiejska 45A, 15-351 Bialystok, Poland
email: jstepan@ii.pb.bialystok.pl

Abstract

Most knowledge discovery methods assume that the original representation space is adequate, that is, the initial attributes are sufficiently relevant to the problem at hand. In real-world applications discovery of new attributes and selection of relevant attributes are applied frequently in data pre-processing. In the paper we discuss rough set based approach to attribute discovery. We consider discovery of adequate attributes for structural objects. We present two algorithms for extracting new attributes.

1 INTRODUCTION

Knowledge discovery, data mining and machine learning are some of the Artificial Intelligence tools that help human to process data, and make use of data. Researchers and practitioners realize that in order to use these tools effectively, an important part is pre-processing in which data is processed before it is presented to any learning, discovering, or visualizing algorithm. Attribute transformation and attribute selection are applied frequently in data pre-processing for real-world applications.

Attribute transformation is a process through which a new set of attributes is created. Assuming the original set A of attributes consists of $a_1, a_2, \ldots, a_n$, some variants of attribute transformation can be defined below.

Attribute transformation process can augment the space of attributes by inferring or creating additional attributes. After attribute construction, we may have additional m attributes $a_{n+1}, a_{n+2}, \ldots, a_{n+m}$. For example, a new attribute a_k ($n < k \leq n + m$) could be constructed by performing a logical operation of a_i and a_j from the original set of attributes.

Attribute transformation process can also extract a set of new attributes from the original attributes through some functional mapping. After attribute extraction, we have $b_1, b_2, \ldots, b_m$ ($m < n$), $b_i = f_i(a_1, a_2, \ldots, a_n)$, and f_i is a mapping function. For instance for real valued attributes a_1 and a_2, for every object x we can define $b_1(x) = c_1 * a_1(x) + c_2 * a_2(x)$ where c_1 and c_2 are constants.

Attribute selection is different from attribute transformation in that no new attributes will be generated, but only a subset of original attributes is selected and the attribute space is reduced.

In machine learning, the idea of constructive induction has been proposed (Michalski 1983), (Michalski and Wnek 1993), (Matheus and Randell 1989), (Muggleton 1987), (Zhang and Lu 1994). A constructive induction system performs

a double, mutually intertwined search, one for the most suitable representation space, and second for the best concept description in this space.
A structured object consists of a finite set of elementary objects and a finite set of relations between them. The elementary objects as well as the relations can be of different types. In this paper we assume that structured objects are comprised of elementary objects all being of the same type, which are treated as the elements of a universe of relational structure. In order to express discernibility between structured objects (relational structures), we use formulas of first order logic. Constructed formulas are treated as new binary attributes.
In the paper we discuss rough set based approach (see for example (Pawlak 1991), (Slowinski 1992), (Ziarko 1994)) to attribute discovery problem. The relevance function (quality of a set of attributes) is based on cardinality of positive region, which is defined as union of lower approximations of all decision classes (Pawlak 1991). We also use some randomization test (Düntsch and Gediga 1996) for evaluation of the final set of attributes.
We present two examples of our approach.
In the first example we discuss problem of employing rough sets and formulas of first order logic (Monk 1976) in automatic attribute discovery (Bazan et al. 1995), (Skowron and Stepaniuk 1991), (Skowron and Stepaniuk 1991a). We prove that discovery of attributes based on first-order logic is in some sense equivalent to graph isomorphism problem. We also present extraction of binary attribute, which is defined by formula of first order logic. We discuss an example application related to handwritten digits recognition.
Second example is based on joining of some set of attributes, with finite (small) number of values into new attribute.

2 BASIC CONCEPTS

Information systems (Pawlak 1991) (sometimes called data tables, attribute-value systems, condition-action tables, knowledge representation systems) are used for representing knowledge. Rough sets have been introduced as a tool to deal with inexact, uncertain or vague knowledge in artificial intelligence applications. In this section we recall some basic notions related to information systems and rough sets.
An *information system* is a pair $A = (U, A)$, where U is a non-empty, finite set called the *universe* and A - a non-empty, finite set of attributes, i.e. $a: U \to V_a$ for $a \in A$, where V_a is called the *value set* of a. Elements of U are called objects and interpreted as, for example, cases, states, processes, patients, observations. Attributes are interpreted as features, variables, characteristic conditions, etc.
Every information system $A = (U, A)$ and non-empty set $B \subseteq A$ determine a *B-information function* $Inf_B : U \to P(B \times \bigcup_{a \in B} V_a)$ defined by $Inf_B(x) = \{(a, a(x)): a \in B\}$.

We define B - indiscernibility relation as follows: $xIND(B)y$ iff $Inf_B(x)=Inf_B(y)$.
For every subset $X \subseteq U$ we define the lower approximation $L_B(X)$ and the upper approximation $U_B(X)$ as follows:
$L_B(X) = \{x \in U: [x]_B \subseteq X\}$,
$U_B(X) = \{x \in U: [x]_B \cap X \neq \emptyset\}$.
Some illustration of the approximations is presented on Figure 1 (a set U of all objects is represented as the global rectangle and B - indiscernibility classes are

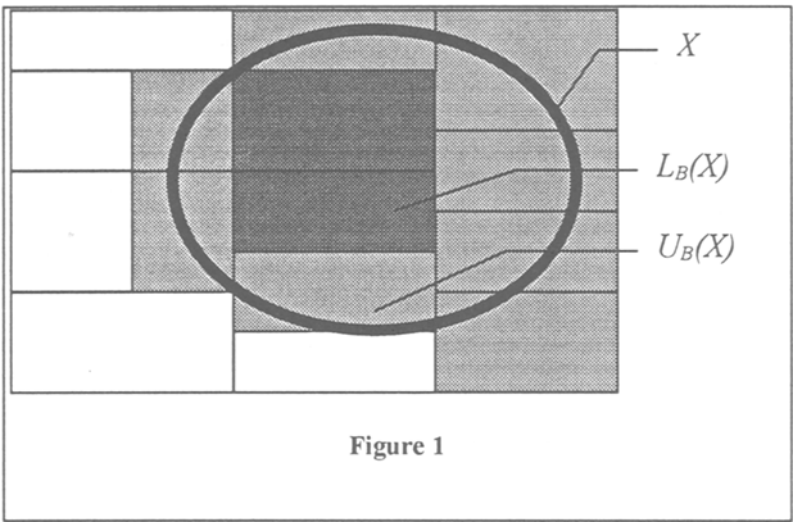

Figure 1

represented as small rectangles).
We consider a special case of information systems called decision tables. A decision table (Pawlak 1991) is any information system of the form $A = (U, A \cup \{d\})$, where $d \notin A$ is a distinguished attribute called *decision*. The elements of A are called *conditions*. One can interpret a decision attribute as a kind of classification of the universe of objects given by an expert, decision-maker, operator, physician, etc. The cardinality of the image $d(U) = \{k: d(x)=k \text{ for some } x \in U\}$ is called the rank of d and is denoted by $r(d)$. We assume that the set V_d of values of the decision d is equal to $\{1,...,r(d)\}$. Let us observe that the decision d determines the partition $CLASS_A(d) = \{X_1,...,X_{r(d)}\}$ of the universe U, where $X_k = \{x \in U: d(x)=k\}$ for $1 \leq k \leq r(d)$. $CLASS_A(d)$ will be called *the classification of objects in A determined by the decision d*. The set X_k is called the k-*th decision class of A*. The set $POS(B,\{d\})$ is called the positive region of classification $CLASS_A(d)$ and is equal to the union of all lower approximations of decision classes. Some example of positive region is presented on Figure 2 (a set U of all objects is represented as the global rectangle, indiscernibility classes are represented as small rectangles, and there are three decision classes).

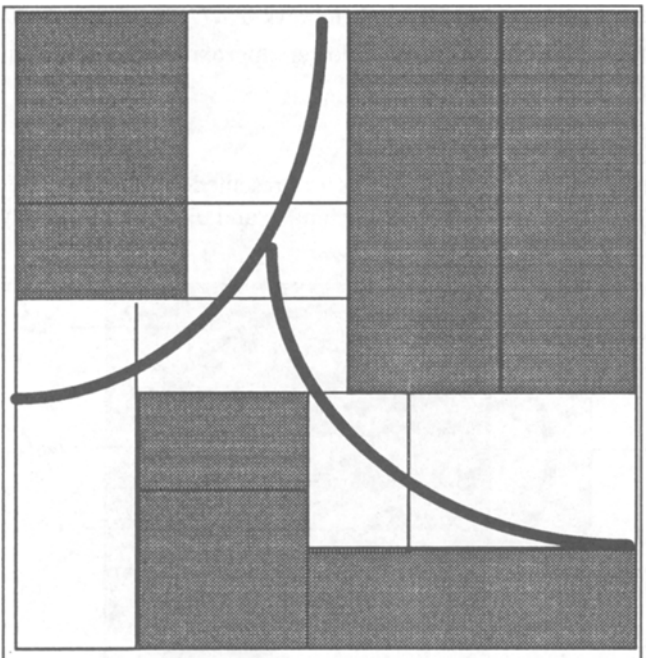

Figure 2

For evaluation of the new set of attributes we use randomization test (Düntsch and Gediga 1996). Here we only briefly describe some concepts of this approach.
Let $\boldsymbol{A} = (U, A \cup \{d\})$ be a decision table and let σ be a permutation of the set of objects U. We define new decision table $\boldsymbol{A}^{\sigma} = (U, A^{\sigma} \cup \{d\})$ such that $A^{\sigma} = \{a^{\sigma}: a \in A\}$ and $a^{\sigma}(x) = a(\sigma(x))$ for every object $x \in U$.
Let us consider coefficient

$$p(POS_R \geq POS) = \frac{\{\sigma: card(POS(A^{\sigma}, d)) \geq card(POS(A, d))\}}{card(U)!}.$$

For example, if the cardinality of every indiscernibility class is equal to one ($IND(A) = \{(x,x): x \in U\}$), then for every permutation we obtain positive region equal to $POS(A,\{d\}) = U$, hence $p(POS_R \geq POS) = 1$. In general, if the coefficient $p(POS_R \geq POS)$ is low (conventionally less than 0.05) the assumption of randomness (random dependency between condition attributes A and decision attribute d) can be rejected.

3 STRUCTURAL OBJECTS

In this section we formulate discernibility formula problem and prove that this problem is equivalent in polynomial time to graph isomorphism problem. We discuss some incremental algorithm for construction of suitable formulas. Next we discuss application of presented algorithm in searching for suitable attributes for handwritten digits recognition problem.

3.1 DISCERNIBILITY FORMULA PROBLEM

Relational structures provide the possibility to describe structural data in an appropriate way. In order to express discernibility between structured objects represented by relational structures we use formulas of first order logic. We prove that the discernibility formula problem is equivalent in polynomial time to graph isomorphism problem.

Let us consider signature $\sigma_l = (=,P_1,...,P_l)$, where $P_1,...,P_l$ are predicate symbols and $l>0$ is a given natural number. Let $FOL(\sigma_l)$ be a set of all formulas of first order logic (Monk 1976) constructed over signature σ_l.

Discernibility Formula Problem

Instance: Two finite relational structures $G = (V,R_1,...,R_l)$ and $H = (V',S_1,...,S_l)$, where V and V' are non-empty, finite sets and $R_1,...,R_l$ and $S_1,...,S_l$ are relations on V and V', respectively.

Answer: „Structures are indiscernible". Otherwise, formula $\alpha \in FOL(\sigma_l)$ such that α is valid in G and is not valid in H.

The graph isomorphism problem is formulated as follows (Leeuven 1990):

Graph Isomorphism Problem

Instance: Two graphs G' and H'.

Answer: „Yes" if there is an isomorphism. Otherwise „no".

Theorem 3.1 The problem of searching for first order formulas distinguishing two finite relational structures is in polynomial time equivalent to the graph isomorphism problem.

Sketch of the proof. The proof is divided into two steps.

STEP 1. For two finite structures the following equivalence is valid: structures are isomorphic if and only if structures are elementary equivalent (Monk 1976). Elementary equivalence means that every first order formula is true in first structure iff that formula is true in second one. Thus if two structures are isomorphic, then there is not first order formula which can distinguish one structure from another. On the other hand if two structures are not isomorphic, then one can construct in polynomial time first order formula which describes one structure and is not true in the second one. For example for structure presented on Figure 3 one can construct the following describing formula: $\exists x_1 \exists x_2 \exists x_3 \exists x_4 (\forall x (x=x_1 \vee x=x_2 \vee x=x_3 \vee x=x_4) \wedge (\neg x_1=x_2 \wedge \neg x_1=x_3 \wedge \neg x_1=x_4 \wedge \neg x_2=x_3 \wedge \neg x_2=x_4 \wedge \neg x_3=x_4) \wedge (x_3 N x_1 \wedge x_4 N x_2 \wedge x_1 S x_3 \wedge x_2 S x_4 \wedge x_2 W x_1 \wedge x_4 W x_3 \wedge x_1 E x_2 \wedge x_3 E x_4 \wedge P_{>25\%}(x_4))$.

STEP 2. Problem of relational structures isomorphism is polynomial time reducible to graph isomorphism problem. For suitable construction of a graph from relational structure see (Miller 1979).

3.2 ALGORITHM FOR CONSTRUCTION OF NEW ATTRIBUTES BASED ON FORMULAS OF FIRST-ORDER LOGIC

In this section we sketch an algorithm of construction of formulas.

ALGORITHM

Input: a set U of relational structures, decision d, parameter *theta* from the interval $[0,1]$

Output: a decision table $(U, A_{new} \cup \{d\})$, such that $card(POS(A_{new},\{d\})) \geq theta*card(U)$.

$A_{new} := \varnothing$ /* Initialize the set A_{new} of attributes as the empty set of attributes */
while $card(POS(A_{new},\{d\})) < theta*card(U)$ **do**
begin
select two objects $x,y \in U$ such that $d(x) \neq d(y)$ and for all $a \in A_{new}$ $a(x)=a(y)$
construct attribute a_{new} such that $a_{new}(x) \neq a_{new}(y)$ using formula of first order logic
$A_{new} := A_{new} \cup \{a_{new}\}$
end.

For construction of formula we propose an algorithm based on an incremental approach.

Suppose the elements of relational structure G are numbered from 1 to n (based on any reasonable scheme). At the kth level of the algorithm we always consider the substructure G(k) of G induced by the elements 1 through k and a substructure H' of H isomorphic to G(k). Essentially the algorithm now proceeds as follows, exploiting a call of the recursive procedure **TEST** which we describe informally.

```
Procedure TEST(k,H');
begin
 if k=n then return true
 else
 begin
  b:=false;
  {consider all possible extensions of H' ...}
  while ¬b and there is an unexplored element v∉H' left do
  begin
   extend H' by v to obtain an induced structure H";
   if the isomorphism between G(k) and H' can be extended to an isomorphism
     between    G(k+1) and H"
   then assign to b the value returned by TEST(k+1,H")
  end
```

```
    return b
  end
end.
```

The suitable formula is constructed if two structures are not isomorphic.

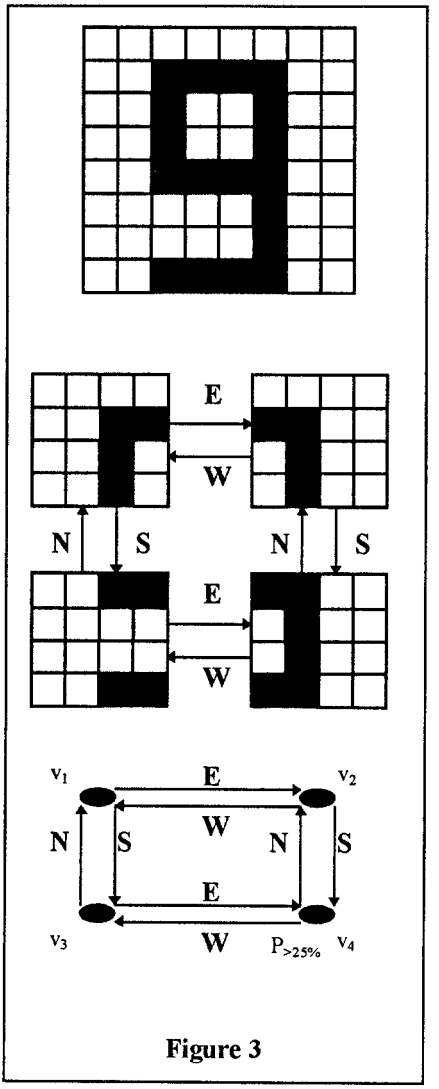

Figure 3

Experimental results based on randomization test show that obtained representation space is acceptable.

Real world application of presented approach is discussed in the next section.

3.3 RELATIONAL STRUCTURES AND HANDWRITTEN DIGITS

We present example of construction of relational structure from image of digit.

Let us consider construction of relational structure presented on Figure 3. Let $V = \{v_1,v_2,v_3,v_4\}$ be the set of elementary objects. Relations are defined as follows: $N = \{(v_3,v_1),(v_4,v_2)\}$, $S = \{(v_1,v_3),(v_2,v_4)\}$, $W = \{(v_2,v_1),(v_4,v_3)\}$, $E = \{(v_1,v_2),(v_3,v_4)\}$, $P_{>25\%} = \{v_4\}$ where N is shortcut for north, S is shortcut for south, W is shortcut for west, etc. and $P_{>25\%}$ means that more than 25% of pixels is black (for example, in the part of the image corresponding to v_4 there are six black pixels and $6/16 > 0.25$).

The constructed relational structure is some representation of digit's image. The problem of automatic construction of suitable relational structure from given image is very interesting.

In paper (Bazan et al. 1995) was presented method of construction of attributes based on multi modal logic which corresponds to some fragment of first order logic.

In our example each recognition system is based on a decision table that has the digit image database as its object space. The decision attribute contains information about the recognition results (expert decision) of the digits. At the beginning, all the digits in the table are totally indiscernible because the attribute set is empty. During the learning process the user will create new attributes for the object set and add them to the decision table. Each new attribute brings into the table new information that may divide existing boundary sets (Pawlak 1991) into smaller ones, which means the positive region is improved. Therefore, with the attributes being added to the decision table, the recognition of objects can become more accurate. The user will continue the process until the positive region is large enough.

One of the most important stages in searching for the appropriate set of attributes contains the strategy for selecting the best attributes from those constructed in the previous steps. Rough set methods provide several approaches to this problem. One of them is based on the idea of reducts (Pawlak 1991) calculated from set of extracted attributes.

4 DISCOVERY OF NEW ATTRIBUTES BASED ON POSITIVE REGION

In this section we present method of attribute construction based on positive region concept.

We present construction of one new attribute from some subset of old attributes. New attribute can have the same number of values as number of decision values.

We use a parameter *theta* $\in [0,1]$, which is · maximal acceptable reduction of positive region.

ALGORITHM

Input: a decision table $(U, A \cup \{d\})$, *a parameter theta* $\in [0,1]$,

Output: a new decision table $(U, A_{new} \cup \{d\})$, such that $card(A_{new}) \leq card(A)$ and $card(POS(A_{new},\{d\})) \geq theta * card(POS(A,\{d\}))$.

A_{new}:=A /* Initialize the set A_{new} of attributes as a set of primitive attributes */
while *card(POS(A_{new},{d})) ≥ theta*card(POS(A,{d}))* **and** (there is „good" set C ⊆ A_{new}) **do**
begin
for all equivalence classes *E* of indiscernibility relation *IND(C)* **do**
begin
for *i:=1* **to** *r(d)* **do** μ_i:=*card(E∩X_i)/card(X_i)*
for all *y∈E* **do**
a_{new}(y) is equal to j with the highest μ_j, if a tie occurs, then select j such that card(X_j) is maximal; if another tie occurs, select the minimal j
end
A_{new}:= (A_{new} -C)∪{a_{new}}
end.

Important part of this algorithm is selection of „good" subset *C* of attributes. Taking computational complexity in mind we decided to consider two element subsets of attributes.

After some experiments, we decided to consider the following criteria based on positive region for choosing two attributes:

1. The pair $\{a_1, a_2\} \subseteq A_{new}$ of attributes such that *card(POS(A_{new}-{a_1,a_2},d))* is maximal.
2. The attribute $a_1 \in A_{new}$ such that *card(POS(A_{new}-{a_1},d))* is maximal and attribute $a_2 \in A_{new}$ such that *card(POS((A_{new}-{a_1})-{a_2},d))* is maximal
3. The pair $\{a_1, a_2\} \subseteq A_{new}$ of attributes such that *card(POS(A_{new}-{a_1,a_2},d))* is minimal.
4. The attribute $a_1 \in A_{new}$ such that *card(POS(A_{new}-{a_1},d))* is minimal and attribute $a_2 \in A_{new}$ such that *card(POS((A_{new}-{a_1})-{a_2},d))* is minimal

We use the following two criteria in evaluation of new decision table:

- number of discovered attributes
- result of randomization test.

We observe that the best results are obtained when choosing at every set of the algorithm the pair $\{a_1, a_2\}$ of attributes such that *card(POS(A_{new}-{a_1,a_2},d))* is minimal. In this case final set of attributes is relatively small and coefficient $p(POS_R \geq POS)$ is low.

For new objects, values of attributes from A_{new} are computed as follows: let x_{new} be a new object, values of new attributes are the same as values for object $x \in U$ most similar to x_{new}.
Similarity between objects can be defined in many ways (Stepaniuk 1996). We consider the following operators:

- $sim(x,y) = \prod_{a \in A} s_a(a(x), a(y))$
- $sim(x,y) = \min_{a \in A} s_a(a(x), a(y))$
- $sim(x,y) = \sum_{a \in A} w_a s_a(a(x), a(y))$, where $0 \leq w_a \leq 1$ is a weight assigned to attribute a, for example we use as w_a significance of attribute a (Pawlak 1991) i.e.

$$w_a = \frac{card(POS(A_{new}, \{d\})) - card(POS(A_{new} - \{a\}, \{d\}))}{card(U)}$$

where s_a is a similarity measure for values of attribute $a \in A$. In the simplest case we consider $s_a(v_i, v_j) = \begin{cases} 1 & if \quad v_i = v_j \\ 0 & otherwise \end{cases}$. For more advanced similarity measures see (Stepaniuk 1996).

Thus for all $a_{new} \in A_{new}$ $a_{new}(x_{new}) = a(x)$, where $sim(x_{new}, x) = max\{sim(x_{new}, y) : y \in U\}$.

Conclusions

This paper addressed issues of discovery of new attributes. Discovery is necessary when the original representation space (set of primitive attributes) is not adequate for a problem at hand. We use cardinality of positive region as measure of quality in attribute discovery process. We also consider some randomization test for evaluation of the final set of attributes.

Acknowledgments

The author thanks Professor Andrzej Skowron for reviewing an earlier draft of the paper. This research is supported by the Polish State Committee for Scientific Research - grant No. 8T11C01011.

References

[1] J.G. Bazan, H.S. Nguyen, T.T. Nguyen, A. Skowron, J. Stepaniuk (1995) Application of Modal Logics and Rough Sets for Classifying Objects, Proceedings of the Second World Conference on Fundamentals of Artificial Intelligence (eds.) M. de Glas, Z.Pawlak, Paris, July 3-7 1995, pp. 15-26.

[2] I. Düntsch, G. Gediga (1996) Statistical Evaluation of Rough Set Dependency Analysis, International Journal of Human-Computer Studies, to appear.

[3] J. van Leeuwen (ed.) (1990) Handbook of Theoretical Computer Science, Elsevier Science Publishers B.V. 1990.

[4] C.J. Matheus, L.A. Randell (1989) Constructive Induction on Decision Trees, Proceedings of the Eleventh International Joint Conference on Artificial Intelligence, 1989, pp. 645-650.

[5] R.S. Michalski (1983) A Theory and Methodology of Inductive Learning: Developing Foundations for Multistrategy Learning, in Machine Learning: An Artificial Intelligence Approach, vol. 1 (eds.) R.S. Michalski, J.G. Carbonell, T.M. Mitchell, Palo Alto, CA: Morgan Kaufmann, 1983.

[6] R.S. Michalski, J. Wnek (1993) Constructive Induction: An Automated Improvement of Knowledge Representation Spaces for Machine Learning, Proceedings of the Second International Workshop on Intelligent Information Systems, Augustow, Poland, 7-11 June, 1993, pp. 188-236.

[7] G. L. Miller (1979) Graph Isomorphism, General Remarks, Journal of Computer and System Sciences 18, 1979, pp. 128-142.

[8] D. Monk (1976) Mathematical Logic, Springer Verlag, 1976.

[9] S. Muggleton (1987) DUCE, an Oracle-Based Approach to Constructive Induction, Proceedings of the Tenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1987, pp. 287-292.

[10] Z. Pawlak (1991) Rough Sets. Theoretical Aspects of Reasoning about Data, Kluwer, Dordrecht, 1991.

[11] A. Skowron, L. Polkowski, J. Komorowski (1996) Learning Tolerance Relations by Boolean Descriptors: Feature Extraction from Data Tables, Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery, November 6-8, 1996, pp. 11-17.

[12] A. Skowron, J. Stepaniuk (1991) Searching for Classifiers, Proceedings of the First World Conference on Foundations of Artificial Intelligence (eds.) M. de Glas, D. Gabbay, Paris, July 1-5 1991, pp. 447-460.

[13] A. Skowron, J. Stepaniuk (1991a) Towards an Approximation Theory of Discrete Problems, Fundamenta Informaticae 15(2), 1991, pp. 187-208.

[14] R Slowinski (1992) (ed.) Intelligent Decision Support Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, 1992.

[15] J. Stepaniuk (1996) Similarity Based Rough Sets and Learning, Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets, and Machine Discovery, November 6-8, 1996, Tokyo, Japan, pp. 18-22.

[16] J. Zhang, H.H. Lu (1994) A Data-Driven Approach to Feature Construction, Proceedings of the Eighth International Symposium on Methodologies for Intelligent Systems, Lecture Notes in Artificial Intelligence 869, Springer Verlag 1994, pp. 448-457.

[17] W. Ziarko (1994) (ed.) Rough Sets, Fuzzy Sets and Knowledge Discovery, Springer-Verlag 1994.

[18] J.M. Zytkow (1996) Automated Discovery of Empirical Laws, Fundamenta Informaticae, 27(2-3), 1996, pp. 299-318.