Roughfication of Numeric Decision Tables: The Case Study of Gene Expression Data

Dominik Ślęzak[1], Jakub Wróblewski[2,1]

[1] Infobright Inc.
218 Adelaide St. W, Toronto, ON, M5H 1W8 Canada
[2] Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland

Abstract. We extend the standard rough set-based approach to be able to deal with huge amounts of numeric attributes versus small amount of available objects. We transform the training data using a novel way of non-parametric discretization, called *roughfication* (in contrast to *fuzzification* known from fuzzy logic). Given *roughfied* data, we apply standard rough set attribute reduction and then classify the testing data by voting among the obtained decision rules. Roughfication enables to search for reducts and rules in the tables with the original number of attributes and far larger number of objects. It does not require expert knowledge or any kind of parameter tuning or learning. We illustrate it by the analysis of the gene expression data, where the number of genes (attributes) is enormously large with respect to the number of experiments (objects).

Keywords: Rough Sets, Discretization, Reducts, Gene Expression Data.

1 Introduction

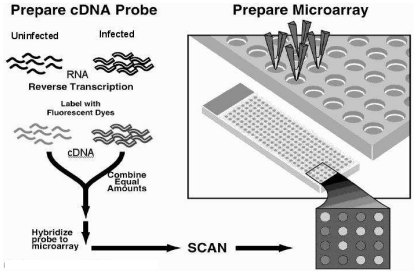
DNA microarrays provide a huge quantity of information about genetically conditioned susceptibility to diseases [1, 2]. However, a typical gene expression data set, represented as an information system $\mathbb{A} = (U, A)$ [9, 14], has just a few objects-experiments $u \in U$, while the number of attributes-genes $a \in A$ is counted in thousands. Moreover, preciseness of measuring gene expressions, i.e. the values $a(x) \in \mathbb{R}$, is still to be improved. Both these issues yield a problem for methods assuming data to be representative enough.

We solve the above problems using *roughfication*,[1] already applied to gene clustering [6] and classification [13]. Foe a given $\mathbb{A} = (U, A)$, we produce a new system $\mathbb{A}^* = (U^*, A^*)$, where U^* corresponds to $U \times U$ and A^* – to the original A. Every $a^* \in A^*$ labels a given $(x, y) \in U^*$ with symbolic value "$\geq a(x)$" iff $a(y) \geq a(x)$, and "$< a(x)$" otherwise. This simple trick provides us with a larger universe, where the number of attributes remains unchanged. This way, it is very different from other *discretization* techniques, which keep U unchanged while exploding the amount of possible attributes in A (cf. [5, 8]).

[1] In [6, 13] we also used the terms *rough discretization* and *rank-based approach*.

Given *roughfied*, symbolic data, we apply the rough set methods for finding optimal (approximate) decision reducts and rules [11, 18]. We extend our research reported in [13] by thorough experimental analysis, involving different reduction and voting options. We show that the proposed simple mechanism provides results comparable to far more complex methods (cf. [15]).

The paper is organized as follows: Section 2 contains basics of gene expression data. Section 3 – basics of rough sets and attribute reduction. Section 4 – *roughfication* and its usage in classification. Section 5 – results of 8-fold cross-validation analysis of the breast cancer-related data with 24 biopsies-objects, 12,625 genes-attributes, and binary decision. Section 6 concludes the paper.



$\Rightarrow$

gene	exp1	exp2	.....	exp8	exp9
No.1	2.52	0.59	.....	-0.46	-0.83
No.2	0.61	0.06	.....	-2.03	-0.59
No.3	-0.19	0.65	.....	-0.37	-0.07
No.4	0.62	-0.47	.....	0.00	-0.10
.	.	.	.....	.	.
.	.	.	.....	.	.
.	.	.	.....	.	.
No.n	.	.	.....	.	.

Fig. 1. Microarrays provide the gene expression data. A sample of 9 experiments from *Synovial Sarcoma* data is illustrated. We have $n = 5520$ genes in this data set (cf. [6]).

2 Gene Expression Data

The DNA microarray technology [1] enables simultaneous analysis of characteristics of thousands of genes in the biological samples of interest. It is automated, much quicker, and less complicated than the previous methods of molecular biology, allowing scientists to study no more than a few genes at a time.

Microarrays rely on DNA sequences fabricated on glass slides, silicon chips, or nylon membranes. Each slide (DNA chip) contains samples of many genes in fixed spots. It may represent cDNA (most popular, used also in this paper), DNA or oligonucleotide. Microarray production starts with preparing two samples of mRNA. The sample of interest is paired with a healthy control sample. Fluorescent labels are applied to the control (green) and the actual (red) samples. Then the slide is washed and the color intensities of gene-spots are scanned, indicating to what extent particular genes are expressed. Figure 1 illustrates the process.

The analysis of such prepared data can lead to discoveries of important dependencies in gene sequences, structures, and expressions. The cDNA microarray data sets are often analyzed to track down the changes of the gene activations for different types of tumors. This information could be then applied to identifying tumor-specific and tumor-associated genes. However, a large number of gathered numerical data makes this analysis particularly hard.

3 Rough Sets, Reducts, Classification

In the rough set theory [9], we analyze information system $\mathbb{A} = (U, A)$, where attributes $a \in A$ correspond to functions $a : U \to V_a$ from universe U into value sets V_a. In this paper, A corresponds to the set of genes, U – to the set of experiments, and functions $a : U \to \mathbb{R}$, reflect gene expressions of humans or other organisms, measured in certain conditions. For classification purposes, we distinguish decision $d \notin A$ to be determined using A. Then, we talk about decision systems $\mathbb{A} = (U, A \cup \{d\})$. In the case of gene expression data d is usually symbolic, reflecting some classes of diseases or behaviors. For instance, in Section 5, d labels two types of behaviors related to breast cancer.

Rough set-based classifiers are a good illustration of tradeoff between accuracy and complexity. In this paper, we focus on decision reducts – minimal subsets $B \subseteq A$ that (almost) determine d. Smaller reducts induce shorter and more general rules. Often, it is even better to remove attributes to get shorter rules at a cost of slight loss of decision determination [11], which can be expressed in many ways. We refer to the original rough set positive region [9] and to its *probabilistic* counterpart [12]. Both measures base on the indiscernibility classes $[u]_B \subseteq U$ defined, for every $u \in U$ and $B \subseteq A$, as $[u]_B = \{x \in U : \forall_{a \in B}\, a(x) = a(u)\}$. The positive region is defined as $POS(B) = \{u \in U : [u]_B \subseteq [u]_{\{d\}}\}$. Each $u \in POS(B)$ induces an "if-then" rule saying that if a new object is equal to u on all attributes $a \in B$, then it can be classified to the u's decision class. One can use also *inexact* rules, especially if the exact ones require too many attributes. Even given no $[u]_B \subseteq [u]_{\{d\}}$ we can still refer to *rough memberships* $\mu_B^v(u) = |\{x \in [u]_B : d(x) = v\}|/|[u]_B|$,[2] where $v \in V_d$ [10]. Whenever a new object is equal to u on all $a \in B$, our belief that its value on d equals to v relates to $\mu_B^v(u)$. In [12] it is shown that such a classification strategy relates to *probabilistic* positive region $POS_\mu(B) = \{(u, \mu_B^{d(u)}(u)) : u \in U\}$ and, more precisely, to its *cardinality* represented as in the theory of fuzzy sets [19]:[3]

$$|POS_\mu(B)| = \sum_{u \in U} \mu_B^{d(u)}(u) = \sum_{u \in U} \frac{|[u]_{B \cup \{d\}}|}{|[u]_B|} \tag{1}$$

By a decision reduct we mean $B \subseteq A$ such that $POS(B) = POS(A)$ and there is no $C \subsetneq B$ such that $POS(C) = POS(A)$ [9]. In this paper,[4] by an *ε-approximate decision* and *ε-approximate μ-decision reducts* we mean subsets $B \subseteq A$ such that, respectively, $|POS(B)| \geq (1-\varepsilon)|POS(A)|$ and $|POS_\mu(B)| \geq (1-\varepsilon)|POS_\mu(A)|$, and there are no $C \subsetneq B$ holding analogous inequalities. In both cases, threshold $\varepsilon \in [0, 1)$ expresses willingness to reduce more attributes (and simplify rules) on the cost of losing the (probabilistic) positive region's strength.[5] In the special case of $\varepsilon = 0$, we are interested in complete preserving of the regions, i.e. in equalities $POS(B) = POS(A)$ and $POS_\mu(B) = POS_\mu(A)$, respectively.

[2] By $|X|$ we denote cardinality of the set X.
[3] Formula (1), after additional division by $|U|$, was denoted in [12] by $E(B)$.
[4] There are many formulations of approximate reducts in the literature [12, 18].
[5] $|POS|, |POS_\mu| : 2^A \to [0, |U|]$ are monotonic with respect to inclusion [9, 12].

The rough set classifier is based on a collection of *optimal* (approximate) decision reducts and the resulting rules [14, 18]. Relaying on a larger set of reducts improves both applicability and accuracy of rough set classifiers. A chance that a new object is recognized by at least one out of many rules is getting higher when operating with diversified subsets of attributes. Appropriate synthesis of information based on different attributes also helps in predicting the right decisions (cf. [3]). In this paper, we compare two techniques of voting – by rules' *supports* and by rules' counting, referring to both of the above-considered types of approximate reducts. We obtain four variants described in Figure 2. We apply them in Section 5 to classification of real-life data.

$\|POS(B)\| \geq (1-\varepsilon)\|POS(A)\|$ Only exact decision rules used $\|[u]_B\|$ added for $d(u)$	$\|POS(B)\| \geq (1-\varepsilon)\|POS(A)\|$ Only exact decision rules used 1 added for $d(u)$
$\|POS_\mu(B)\| \geq (1-\varepsilon)\|POS_\mu(A)\|$ Also inexact decision rules used $\|[u]_B\| \cdot \mu_B^v(u)$ added for each v	$\|POS_\mu(B)\| \geq (1-\varepsilon)\|POS_\mu(A)\|$ Also inexact decision rules used $\mu_B^v(u)$ added for each v

Fig. 2. Four techniques of classification, each described by three lines. Line 1 indicates constraints for subsets $B \subseteq A$ that can be contained in the classifier. (E.g., in Section 5 we always choose 5 best found subsets.) Line 2: the types of rules. (Note that they are related to the types of reducts.) Line 3: the weights being added to decision classes, whenever a rule generated by B is applicable to u. The decision for u is finally chosen as $v \in V_d$ with the highest overall weight, summed up over all applicable rules.

4 Roughfication

Standard rule-based methods are hardly applicable to real-valued data systems unless we use discretization [7, 8] or switch to more advanced techniques, requiring more parameters and/or expert knowledge (cf. [16, 19]). Machine learning methods have serious problems while dealing with disproportions between attributes and objects. Gene expression technology is still quite imprecise, which causes additional problems with data representativeness. As a result, many approaches, including those in Section 3, cannot be applied straightforwardly.

We suggest a new way of data preparation, called *roughfication*. Given $\mathbb{A} = (U, A \cup \{d\})$, we create a new system $\mathbb{A}^* = (U^*, A^* \cup \{d^*\})$, where $U^* \equiv U \times U$ and $A^* \cup \{d^*\} \equiv A \cup \{d\}$. It is illustrated by Figure 3. For every $a \in A$, $V_a = \mathbb{R}$, a new attribute $a^* \in A^*$ has the value set $V_a^* = \bigcup_{u \in U} \{\text{“}\geq a(x)\text{”}, \text{“}< a(x)\text{”}\}$. For every $x, y \in U$, we put $a^*(x,y) =$“$\geq a(x)$” iff $a(y) \geq a(x)$, and $a^*(x,y) =$“$< a(x)$” otherwise. For symbolic attributes, in particular for $d \notin A$, we put $d^*(x,y) = d(y)$, i.e. $V_d^* = V_d$. It is important that V_a^* can be treated as a symbolic-valued domain during further calculations, e.g., those described in Section 3. When the classifier is ready, the values in V_a^* begin to be interpreted in a non-symbolic way again. Continuing with analogy to fuzzy sets [19], we may say that the data is first *roughfied*, then the learning process is performed, and finally the resulting classifier is *deroughfied* to deal with new objects.

	a	b	d
$u1$	3	7	0
$u2$	2	1	1

	a	b	d
$u3$	4	0	1
$u4$	0	5	2

====>

	a^*	b^*	d^*
$(u1, u1)$	≥ 3	≥ 7	0
$(u1, u2)$	< 3	< 7	1
$(u1, u3)$	≥ 3	< 7	1
$(u1, u4)$	< 3	< 7	2
$(u2, u1)$	≥ 2	≥ 1	0
$(u2, u2)$	≥ 2	≥ 1	1
$(u2, u3)$	≥ 2	< 1	1
$(u2, u4)$	< 2	≥ 1	2

	a^*	b^*	d^*
$(u3, u1)$	< 4	≥ 0	0
$(u3, u2)$	< 4	≥ 0	1
$(u3, u3)$	≥ 4	≥ 0	1
$(u3, u4)$	< 4	≥ 0	2
$(u4, u1)$	≥ 0	≥ 5	0
$(u4, u2)$	≥ 0	< 5	1
$(u4, u3)$	≥ 0	< 5	1
$(u4, u4)$	≥ 0	≥ 5	2

IF $a \geq 3$ AND $b \geq 7$ THEN $d = 0$
IF $a \geq 3$ AND $b < 7$ THEN $d = 1$
IF $a \geq 2$ AND $b < 1$ THEN $d = 1$
IF $a < 2$ AND $b \geq 1$ THEN $d = 2$
IF $a \geq 4$ AND $b \geq 0$ THEN $d = 1$
IF $a \geq 0$ AND $b < 5$ THEN $d = 1$

Fig. 3. Top left: Original $\mathbb{A} = (U, A \cup \{d\})$, $U = \{u1, u2, u3, u4\}$, $A = \{a, b\}$, $V_d = \{0, 1, 2\}$. Right: *Roughfied* $\mathbb{A}^* = (U^*, A^* \cup \{d^*\})$. Bottom left: Exact rules based on A^*.

Figure 3 illustrates a decision system obtained using roughfication. Region $POS(a^*, b^*) = \{(u1, u1), (u1, u3), (u2, u3), (u2, u4), (u3, u3), (u4, u2), (u4, u3)\}$ is covered by six exact rules. Although formally we should write, e.g., IF $a^* =$"≥ 3" AND $b^* =$"≥ 7" THEN $d^* = 0$, we use a *deroughfied* notation, as used while classifying new objects. Note that one set of attributes can induce more than one rule. For a new object $u \notin U$, the rules may point at different decisions. For instance, if $a(u) = 4$ and $b(u) = 7$, then the first rule in Figure 3 will yield $d = 0$ while the fifth one will give us $d = 1$. Also, some rules may be more applicable than the others. For example, whenever conditions of the third rule are satisfied, the sixth rule is applicable too. We may say that such rules are able to represent more and less typical patterns for decision classes. It indicates that roughfication should be followed by well-designed voting mechanisms.

Continuing with Figure 3, let us consider $POS(\{a^*\}) = \{(u2, u4), (u3, u3)\}$ and $POS(\{b^*\}) = \{(u1, u1), (u2, u3), (u4, u2), (u4, u3)\}$. Given that $|POS(A^*)|$ equals 7, the set $\{a^*\}$ begins to be an ε-approximate reduct for $\varepsilon \geq 3/7$ and $\{b^*\}$ – for $\varepsilon \geq 5/7$. Note that $POS(\{a^*\})$ and $POS(\{b^*\})$ are disjoint, covering together 6 our of 7 elements of $POS(A^*)$. Hence, $\{a^*\}$ and $\{b^*\}$ might work well together, as a reduct collection. The only missing element is $(u1, u3)$ and, accordingly, the second rule in Figure 3 cannot be shortened.

Finally, we briefly illustrate $POS_\mu(A^*)$ in Figure 4. The rules not occurring in Figure 3 are, e.g.: IF $a < 3$ AND $b < 7$ THEN $d = 0$ with weight 0.5 AND $d = 2$ with weight 0.5; or, e.g.: IF $a \geq 2$ AND $b \geq 1$ THEN $d = 0$ with weight 0.5 AND $d = 1$ with weight 0.5. As illustrated in Figure 2, such decision weights participate in voting in the case of inexact rule-based reduction/classification.

$u1$	a^*	b^*	d^*	μ
$u1$	≥ 3	≥ 7	0	1
$u2$	< 3	< 7	1	1/2
$u3$	≥ 3	< 7	1	1
$u4$	< 3	< 7	2	1/2

$u2$	a^*	b^*	d^*	μ
$u1$	≥ 2	≥ 1	0	1/2
$u2$	≥ 2	≥ 1	1	1/2
$u3$	≥ 2	< 1	1	1
$u4$	< 2	≥ 1	2	1

$u3$	a^*	b^*	d^*	μ
$u1$	< 4	≥ 0	0	1/3
$u2$	< 4	≥ 0	1	1/3
$u3$	≥ 4	≥ 0	1	1
$u4$	< 4	≥ 0	2	1/3

$u4$	a^*	b^*	d^*	μ
$u1$	≥ 0	≥ 5	0	1/2
$u2$	≥ 0	< 5	1	1
$u3$	≥ 0	< 5	1	1
$u4$	≥ 0	≥ 5	2	1/2

Fig. 4. The $POS_\mu(A^*)$ coefficients displayed in the column "μ". The objects in U^* are now split onto subtables with respect to the first elements in pairs (ui, uj), $i, j = 1, ..., 4$.

5 Classification Results

We study the breast cancer data downloaded from Gene Expression Omnibus GEO, http://www.ncbi.nlm.nih.gov/projects/geo/gds/gds_browse.cgi?gds=360, analyzed in [15]. It contains 24 core biopsies taken from patients, who are resistant (14 objects) or sensitive (10 objects) to the docetaxel treatment. There are 12,625 genes-attributes. Figure 5 shows the results obtained using the 8-fold cross validation (CV-8). For each split onto 21 training and 3 testing objects, we repeated the following steps: 1) *Roughfy* the training data; 2) Calculate 5 best reducts of a given type; 3) Classify the testing data using voting of a given type. The reported numbers are averaged over the 40 independent CV-8 data splits, with standard deviations low enough to provide the results' credibility.

Given the number of genes-attributes, we applied very simple heuristics to optimize ε-approximate (μ-)decision reducts: 1) Start with $B^* = \emptyset$ and keep randomly adding attributes until we get $|POS(B^*)| \geq (1-\varepsilon)|POS(A^*)|$ (or $|POS_\mu(B^*)| \geq (1-\varepsilon)|POS_\mu(A^*)|$); 2) Also randomly, keep removing attributes from B – if a given attribute cannot be removed without losing the above inequality, keep it in B and try with the next one; 3) Out of 15 such randomly generated reducts,[6] select 5 with minimal cardinality. Obviously, more advanced techniques can be applied to get better collections for each $\varepsilon \in [0, 1)$ [17, 18].

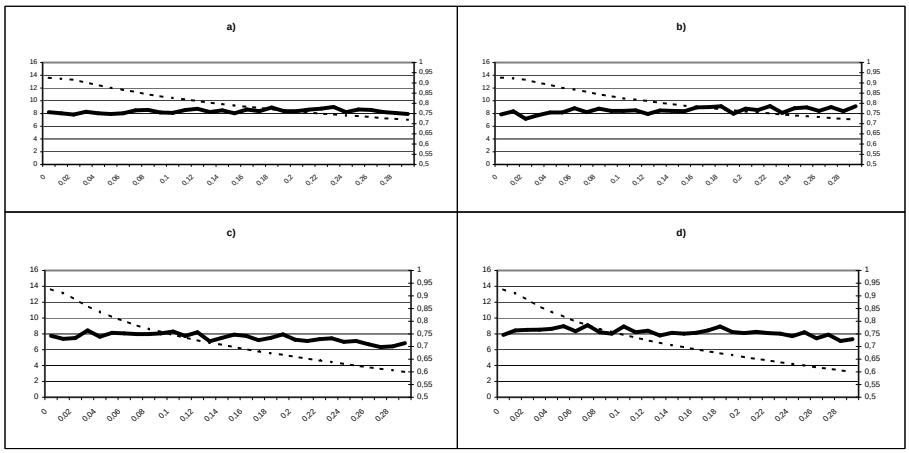


Fig. 5. Accuracy of classification within the CV-8 scheme (solid line) and average the reduct length (dotted line). The following attribute reduction and rule voting settings correspond to those illustrated in Figure 2: a) reduction measure $|POS|$ and voting by the rules' supports; b) reduction measure $|POS|$ and voting by numbers of rules; c) reduction measure $|POS_\mu|$ and voting by the rules' supports times rough memberships to decision classes; d) reduction measure $|POS_\mu|$ and voting by rough memberships. In each case, the horizontal axis shows the values of ε, for which the results were obtained.

[6] Previously mentioned monotonicity of functions $|POS|$ and $|POS_\mu|$, now defined over 2^{A^*}, guarantees that the obtained sets are reducts of appropriate type in $\mathbb{A}^*$.

The obtained results are fully comparable with those obtained using far more complicated methodologies, which usually remain difficult to understand for the domain experts (cf. [4, 15]). A huge advantage of the proposed method is that it simply bases on collections of genes' subsets and "if-then" rules operating on inequalities. Besides a need of tuning the level of $\varepsilon \in [0, 1)$, which is actually related to the *simplicity versus accuracy* tradeoff specific for particular users, there are no parameters requiring additional tuning or expert knowledge, including the ones related to discretization [7, 8]. Even the most basic variant (b) in Figure 5, the one using only exact rules and very simplified voting mechanism, provides good results. Operating with $|POS_\mu|$ instead of $|POS|$ enables to reduce attributes faster along the ε-axis, though we need to be more careful with accuracy. The two considered voting mechanisms give quite similar outcomes, though (b) and (d) seem to slightly better than (a) and (c), respectively. Explanation may lay in a fact that the collections of rules derived from roughfied data are able to express more and less typical areas of particular decision classes by themselves (see Section 4), hence any additional weights assigned to the rules are not so necessary. Last but not least, we do not report percentage of *unrecognized* testing objects because it is practically equal to 0. Also, we do not consider any additional filtration of decision rules [14, 18] because we experimentally measured its insignificance to the results. As a summary, the proposed method is very easy to design and implement, comparing to other approaches.

6 Conclusions and Further Research

We introduced a new method for dealing with real-valued decision tables, especially useful in the cases when the sets of attributes are significantly larger than the sets of objects, with very limited applicability of standard classification methods based on both parameterized and non-parameterized models. We showed how to combine it with a classical framework for construction of rough set-based classifiers. We tested the obtained approach against the real-life gene expression data set with 12,625 genes-attributes and 24 measurements-objects. The results turn out to be comparable with those of other classification techniques, while our method remains far simpler in interpretation for the domain experts. Further research is required to fully prove its practical usefulness over a wider range of benchmark data. We also consider adapting the proposed framework for real-valued decisions, as well as dominance/preference-related attributes.

Acknowledgements: Research of the second co-author was supported by the Research Center at the Polish-Japanese Inst. of Information Technology and by grant 3T11C00226 from Polish Ministry of Sci. Research and Higher Education.

References

1. Baldi P., Hatfield W.G.: DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modelling. Cambridge University Press (2002)

2. Chang J.C. et al.: Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. The Lancet, **362** (2003)
3. Dietterich T.: Machine learning research: four current directions. AI Magazine, **18/4** (1997)
4. Draghici, S.: Data Analysis Tools for DNA Microarray. Chapman & Hall (2003)
5. Ganter B., Wille R.: Formal Concept Analysis: Mathematical Foundations. Springer (1997)
6. Grużdź A., Ihnatowicz A., Ślęzak D.: Interactive gene clustering: A case study of breast cancer microarray data. Information Systems Frontiers, **8** (2006)
7. Fang J., Grzymala-Busse J.W.: Leukemia Prediction from Gene Expression Data – A Rough Set Approach. In: Proc. of ICAISC'06 (2006)
8. Nguyen H.S.: Approximate Boolean Reasoning: Foundations and Applications in Data Mining. Transactions on Rough Sets V, Springer, LNCS, **4100** (2006)
9. Pawlak Z.: Rough sets: Theoretical aspects of reasoning about data. Kluwer (1991)
10. Pawlak Z., Skowron A.: Rough membership functions. In: R.R. Yaeger, M. Fedrizzi, and J. Kacprzyk (eds.), Advances in the Dempster Shafer Theory of Evidence. Wiley (1994)
11. Ślęzak D.: Approximate reducts in decision tables. In: Proc. of IPMU'96, vol. **3** (1996)
12. Ślęzak D.: Various approaches to reasoning with frequency-based decision reducts: a survey. In: L. Polkowski, S. Tsumoto, T.Y. Lin (eds.), Rough Set Methods and Applications. Physica-Verlag (2000)
13. Ślęzak D., Wróblewski J.: Rough Discretization of Gene Expression Data. In: Proc. of ICHIT'06, vol. **2** (2006)
14. Słowiński R., Greco S., Matarazzo B.: Rough Set Based Decision Support. Introductory Tutorials on Optimization, Search and Decision Support Methodologies, Springer (2005)
15. Valdes J.J., Barton A.J.: Relevant Attribute Discovery in High Dimensional Data: Application to Breast Cancer Gene Expressions. In: Proc. of RSKT'06 (2006)
16. Wojna A.: Analogy-Based Reasoning in Classifier Construction. Transactions on Rough Sets IV, Springer, LNCS (2005)
17. Wróblewski J.: Theoretical Foundations of Order-Based Genetic Algorithms. Fundamenta Informaticae, **28/3-4** (1996)
18. Wróblewski J.: Ensembles of classifiers based on approximate reducts. Fundamenta Informaticae, **47/3-4** (2001)
19. Zadeh L.A.: Fuzzy Sets. Information and Control, **8** (1965)