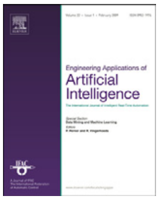
Engineering Applications of Artificial Intelligence 25 (2012) 284–294

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Inductive discovery of laws using monotonic rules

Jerzy Błaszczyński [a], Salvatore Greco [b], Roman Słowiński [a,c,*]

[a] Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland
[b] Faculty of Economics, University of Catania, 95129 Catania, Italy
[c] Systems Research Institute, Polish Academy of Sciences, 01-447 Warsaw, Poland

ARTICLE INFO

Article history:
Received 18 January 2011
Received in revised form
2 August 2011
Accepted 1 September 2011
Available online 29 September 2011

Keywords:
Knowledge discovery
Inductive learning
Classification
Rough Sets
Dominance-based Rough Set Approach
Decision rules
Monotonicity

ABSTRACT

We are considering knowledge discovery from data describing a piece of real or abstract world. The patterns being induced put in evidence some laws hidden in the data. The most natural representation of patterns-laws is by "*if…, then…*" decision rules relating some conditions with some decisions. The same representation of patterns is used in multi-attribute classification, thus the data searched for discovery of these patterns can be seen as classification data. We adopt the classification perspective to present an original methodology of inducing general laws from data and representing them by so-called monotonic decision rules. Monotonicity concerns relationships between values of condition and decision attributes, e.g. the greater the mass (condition attribute), the greater the gravity (decision attribute), which is a specific feature of decision rules discovered from data using the Dominance-based Rough Set Approach (DRSA). While in DRSA one has to suppose a priori the presence or absence of positive or negative monotonicity relationships which hold in the whole evaluation space, in this paper, we show that DRSA can be adapted to discover rules from any kind of input classification data, exhibiting monotonicity relationships which are unknown a priori and hold in some parts of the evaluation space only. This requires a proper non-invasive transformation of the classification data, permitting representation of both positive and negative monotonicity relationships that are to be discovered by the proposed methodology. Reported results of a computational experiment confirm that the proposed methodology leads to decision rules whose predictive ability is similar to the best classification predictors. It has, however, a unique advantage over all competitors because the monotonic decision rules can be read as laws characterizing the analyzed phenomena in terms of easily understandable "*if…, then…*" decision rules, while other predictor models have no such straightforward interpretation.

1. Introduction

Knowledge discovery from data describing a piece of real or abstract world is a field of computer science that concerns the process of automatically searching the data for patterns that can be considered knowledge about this piece of the world. The patterns are to evidence by induction some *laws* hidden in the data. The most natural representation of patterns-laws is by "*if…, then…*" decision rules relating some conditions with some decisions. The same representation of patterns is used in multi-attribute classification, thus the data searched for discovery of these patterns can be seen as classification data. In this paper, we adopt the classification perspective to present an original methodology of inducing general laws from data and representing them by so-called *monotonic decision rules*.

Classification concerns a set of objects described by a set of attributes. Commonly, in the set of attributes there is at least one called decision attribute (also called dependent variable, output variable or predictor), and others called condition attributes (also called independent variables, input variables or features). The decision attribute makes a partition of the set of objects into classes, thus the value set of the decision attribute is composed of class labels. Analysis of classification data aims at discovering relationships between decision attribute and condition attributes, which can be seen as patterns or laws characterizing the world described by the data. The type of discovered relationships depends on the character of decision and condition attributes. The character of the attributes which is pertinent for our study concerns the presence or absence of an order in their value sets. Specifically, we distinguish *ordinal* and *non-ordinal* attributes, both decision and condition.

Moreover, the nature of the classification problem may require that discovered relationships show a *monotonic dependency* between values of some ordinal condition attributes and values of the ordinal decision attribute, e.g. the greater the mass (condition attribute), the greater the gravity (decision attribute). This was precisely a specific feature of the approach to knowledge discovery from data presented in a series of publications on Dominance-based Rough Set Approach (DRSA) (see, e.g. Greco et al., 2001, 2007; Słowiński et al., 2009). Let us remember that

---

* Corresponding author at: Institute of Computing Science, Poznań University of Technology, 60-965 Poznań, Poland.
E-mail addresses: jblaszczynski@cs.put.poznan.pl (J. Błaszczyński), salgreco@unict.it (S. Greco), rslowinski@cs.put.poznan.pl (R. Słowiński).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.09.003

DRSA has been developed as a generalization of rough set theory (Pawlak, 1991) based on the concept of monotonicity for reasoning about ordinal data.

In this paper, we shall extend the consideration of monotonicity on non-ordinal classification. In fact, until now, monotonicity has been considered only in case there is a clear and predefined correspondence between decision and condition attributes, such that we know a priori that the value of the ordinal decision attribute is consistently monotonically increasing or decreasing with respect to the value of a given ordinal condition attribute. Very often, however, the user does not know a priori if such a monotonic dependency is present, and if so, what type of monotonicity it is. For instance, investigating how a given substance is influencing a new disease, the medical doctor could not know how the amount of the substance in the blood is related to the gravity of this disorder. What is more, (s)he could not know a priori if the relationship between the amount of substance and the gravity of disorder is monotonically positive or negative. In such a case, (s)he would like to discover from the data if such a monotonic relationship exists and if it is positive or negative. Moreover, one should not limit the interest to global monotonicity, which holds in all the space of variation of decision and condition attributes. We believe that monotonicity is something more than a relationship between quantities to be investigated. We claim that it is a universal category permitting to analyze any data, giving easily understandable laws explaining important aspects of the studied phenomena. In this sense, we consider the concept of monotonicity also when the monotonicity is not global, i.e. there are local relationships of monotonicity that can be positive in some part of the investigated space, and negative in other parts of the same space. For instance, investigating the effects of some medicines in treating a certain disease, we can have that until a certain point, the greater the dose the better the result, but after that point the further increase of the dose may have a negative effect. Such types of phenomena cannot be studied using the methodology supposing that a positive or negative monotonicity holds in all the space, and trying to discover relationships respecting this assumption. Therefore, to deal with classification data describing such phenomena, we propose a new methodology that attempts to discover local monotonicity relationships, without assuming a priori a specific and constant direction of this monotonicity.

Taking into account the distinction of ordinal and non-ordinal attributes, as well as presence or absence of monotonicity relationships, we can consider the following types of classification problems:

(α) non-ordinal classification, where all condition attributes and the decision attribute are non-ordinal,
(β) non-ordinal classification, where all condition attributes are ordinal, but the decision attribute is non-ordinal,
(γ) ordinal classification, where all condition attributes are non-ordinal, but the decision attribute is ordinal,
(δ) ordinal classification, where at least one condition attribute is ordinal, and the decision attribute is also ordinal, but there is no monotonic relationship between values of ordinal condition and decision attributes,
(ε) ordinal classification, where at least one condition attribute is ordinal, the decision attribute is also ordinal, and there is a monotonic relationship between values of ordinal condition and decision attributes.

The following examples illustrate, respectively, the above types of classification problems:

(α) taxonomy of plants, where condition attributes are morphological features of the plants and the class labels correspond to types of plants,
(β) taxonomy of plants, where condition attributes are measures of morphological features and the class labels correspond to types of plants,
(γ) rating of films, where condition attributes are characteristics of films and spectators, and the class labels express the attractiveness of the films (in terms of one, two, three, etc. stars),
(δ) classification of comfort states, where the room temperature is the ordinal condition attribute, and the ordinal decision attribute is the comfort level, but it is not true that the higher the temperature, the higher the comfort, nor vice versa,
(ε) classification of students, where the ordinal condition attributes are course grades, and the ordinal decision attribute is the overall student evaluation, while there exists a monotonic relationship between each course grade and the overall student evaluation.

It is clear that the classification problem (ε) is the most specific, because other classification problems use less background information than (ε), however, we shall show that any classification problem can be formulated in its terms, exploiting the concept of monotonicity also when it is neither predefined nor global.

The relationships between decision and condition attributes are particular for each classification problem type. We are interested in relationships (patterns-laws) in the form of "*if..., then...*" decision rules. In the condition (*if*) part of the rules there is a conjunction of elementary conditions concerning the values taken by particular condition attributes, while in the decision (*then*) part of the rules there is a decision about the class label given to an object satisfying the condition part.

The rules are induced from classification data presented in a *classification table* whose rows correspond to objects and columns to condition and decision attributes. To induce decision rules for the (ε) classification problem, DRSA presented in Greco et al. (1998, 1999, 2001, 2005) and Słowiński et al. (2005, 2009). These rules are called *monotonic* for their syntax of the form:

> *if* evaluation of object a is greater (or smaller) than given values of some condition attributes, *then a* belongs to at least (at most) given class.

The above syntax takes into account that condition and decision attributes are ordinal and monotonically related. In this paper, we want to show that it is advantageous to use DRSA and monotonic decision rules also in case of all remaining classification problems (α)−(δ), after a non-invasive transformation of classification data for these problems. In fact, DRSA was developed and applied in case of supposed monotonicity relationship between values of condition and decision attributes. This is exactly the case of the (ε) classification problem. In case of the other classification problems, monotonic relationships are not known a priori and may change from positive to negative in many points of the range of variation of condition attributes, thus one must be able to discover *local monotonicity relationships*. A local monotonicity relationship becomes *global* if it is positive or negative in the whole evaluation space. For example, considering room temperature as condition attribute, and comfort as decision attribute, instead of assuming that the higher (or the lower) the room temperature, the higher the comfort, it is reasonable to allow splitting this monotonicity relationship into two local relationships: until some value of the temperature the monotonic relationship is positive, and it is negative over this value. Even in case of binary attributes, corresponding to presence/absence of a property indicated by condition and decision attributes, the concept of monotonicity makes sense, because the presence of one property may be more credible when another property holds, or vice versa. For example, considering weather condition (sunny/rainy) as condition attribute, and playing golf (yes/no) as decision attribute, it is reasonable to expect that "yes" decision is more credible under sunny than rainy weather, which corresponds to monotonicity relationship among 0–1 coded condition and decision attributes.

Adaptation of DRSA to discovery of local monotonicity relationships in classification problems $(\alpha)-(\delta)$ is the main contribution of this paper. Due to this adaptation, one can apply DRSA without declaring a priori where are the turning points of the monotonicity relationship in the condition attribute space, because the proposed method is able to discover them by itself. DRSA induction of decision rules with local monotonic relationships has the following advantages over the use of specific induction methods for problems $(\alpha)-(\delta)$:

- it can handle monotonic relationships known a priori, as well as monotonic relationships that are not known a priori and have to be discovered,
- it can induce general laws involving local monotonicity relationships, i.e. it is able to discover decision rules with "interval" elementary conditions: "attribute $a_i \in [r_i^1, r_i^2]$",
- it can handle ordinal and non-ordinal condition and decision attributes,
- the non-ordinal attributes can be nominal, numerical or binary,
- it does not need transformation of the value sets (scales) of condition attributes,
- it does not need discretization of numerical condition attributes,
- it is able to discover rules with elementary conditions on nominal attributes of the type: "attribute $a_i \in \{v_i^1, \ldots, v_i^k\}$",
- it can induce rules providing arguments pro and cons assignment of an object to a given class,
- the monotonic rules, together with a specially proposed classification scheme, have at least as good predictive ability as other well known predictors, while they are much more comprehensible than any other forms of relationships between condition and decision attributes.

Adaptation of DRSA to classification problems $(\alpha)-(\delta)$ needs a proper transformation of classification table. This transformation is non-invasive, i.e. it does not bias the matter of discovered relationships.

The intuition which stands behind this transformation is the following. In case of ordinal condition attributes, for which the presence and the sign of the monotonicity relationship between values of condition and decision attributes is known a priori, no transformation is required and DRSA can be applied directly. Each non-ordinal condition attribute, for which the presence or absence and the possible sign of the monotonicity relationship is not known a priori, is doubled and for the first attribute in the pair it is supposed that the monotonicity relationship is potentially positive, while for the second attribute, that it is potentially negative. Due to this transformation, using DRSA one will be able to find out if the actual monotonicity is global or local, and if it is positive or negative. The decision attributes are transformed such that:

- in case of a non-ordinal decision attribute, each value of this attribute representing a given feature is replaced by a new decision attribute with two values corresponding to the presence and absence of this feature, respectively,
- in case of an ordinal decision attribute, each value of interest t, is replaced by a new decision attribute with two values corresponding to original values smaller than t and greater or qual to t, respectively.

More precisely, given a finite set of objects (universe) U described by condition and decision attributes, we assume that the decision attribute makes a partition of U into a finite set of classes $X_1, X_2, \ldots, X_n$. To discover rules relating values of condition attributes with class assignment, in case of non-ordinal classification problems, we have to consider n ordinal binary classification problems with two sets of objects: class X_t and its complement $\neg X_t$, $t = 1, \ldots, n$, which are number-coded by 1 and 0, respectively. We also assume, without loss of generality, that the value sets of all non-ordinal condition attributes are number-coded. While this is natural for numerical attributes, nominal attributes must be binarized and get 0–1 codes for the absence or presence of a given nominal value. In this way, the value sets of all non-ordinal attributes get ordered (as all sets of numbers are ordered). Now, to apply DRSA on a classification problem different from (ε), we transform the data table such that each number-coded attribute is cloned (doubled). It is assumed that the value set of each original number-coded attribute is positively monotonically dependent on the decision, i.e. the greater the value of the condition attribute, the higher the number code (rather 1 than 0) of the class assignment, and the value set of its clone is negatively monotonically dependent on the decision, i.e. the greater the value of the condition attribute, the lower the number code (rather 0 than 1) of the class assignment. Then, using DRSA, we get rough approximations of class X_t and its complement $\neg X_t$, $t = 1, \ldots, n$. These approximations serve to induce "*if..., then...*" decision rules recommending assignment to class X_t (argument pros) or to its complement $\neg X_t$ (argument cons). In this way, any classification problem $(\alpha)-(\delta)$ can be transformed to an ordinal classification problem with monotonicity constraints. Due to cloning of attributes with opposite monotonicity relationships, we can have rules that cover a subspace in the condition attribute space, which is bounded from the top and from the bottom. This leads (without discretization) to more synthetic rules than those resulting from induction techniques specific to classification problems $(\alpha)-(\delta)$.

The plan of this paper is the following. In the next section, we recall DRSA designed for classification problem (ε). Section 3 is devoted to transformation of classification table in case of the remaining classification problems $(\alpha)-(\delta)$. Then, in Section 4, we present the classification scheme which says how to use decision rules for ordinal and non-ordinal classification. In Section 5, we give an illustrative example which permits to follow all steps of the transformation of classification data, and shows the monotonic rules obtained from DRSA applied on the transformed data. Section 6 reports a computational experiment aiming at the comparison of the proposed approach with other predictors of classification. The last section includes conclusions and remarks on future research.

2. Dominance-based rough set approach

This section reminds the main concepts of the Dominance-based Rough Set Approach (DRSA) (for a more complete presentation see, for example, Greco et al., 1999, 2001, 2005, 2007; Słowiński et al., 2005, 2009).

2.1. Data representation—classification table

Information about objects (classification examples) is represented in the form of an information table. The rows of the table are labeled by objects, whereas columns are labeled by attributes and entries of the table are attribute-values. Formally, an *information table* (system) is the 4-tuple $\mathbf{S} = \langle U, Q, V, \phi \rangle$, 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 the value set of the attribute q, and $\phi : U \times Q \to V_q$ is a total function such that $\phi(x,q) \in V_q$ for every $q \in Q$, $x \in U$, called an information function. The set Q is, in general, divided into set C of condition attributes and set D of decision attributes. When it is the case, **S** is called a *classification table*. Furthermore, it is supposed that the set of decision attributes D is a singleton $\{d\}$.

Condition attributes whose value sets are ordered are called *ordinal attributes*. Without loss of generality, for ordinal attribute $q \in C$, $\phi : U \to \mathbb{R}$, for all objects $x,y \in U, \phi(x) \geq \phi(y)$ means "x is evaluated at least as high as y on ordinal attribute q", which is denoted by $x \succcurlyeq_q y$. Therefore, it is supposed that $\succcurlyeq_q$ is a complete preorder, i.e. a strongly complete and transitive binary relation, defined on U on the basis of evaluations $\phi(\cdot)$. Ordinal attribute q may have positive or negative monotonic relationship with the decision attribute d (which is also ordinal). Positive relationship means that the greater the value of the condition attribute the higher the class label (i.e. the value of decision attribute), and negative relationship means that the greater the value of condition attribute the lower the class label. For the sake of simplicity, we assume that all condition attributes in set C are ordinal.

Furthermore, values of decision attribute d make a partition of U into a finite number of decision classes, $\mathbf{X} = \{X_t, t = 1, \ldots, n\}$, such that each $x \in U$ belongs to one and only one class $X_t \in \mathbf{X}$. It is supposed that the classes are ordered, i.e. for all r, $s \in \{1, \ldots, n\}$, such that $r > s$, the objects from X_r are in higher class than the objects from X_s. More formally, if $\succcurlyeq$ is a *comprehensive weak order relation* on U, i.e. if for all x, $y \in U$, $x \succcurlyeq y$ means "x is ranked at least as high as y", it is supposed: $[x \in X_r, y \in X_s, r > s] \Rightarrow$ $[x \succcurlyeq y$ and *not* $y \succcurlyeq x]$. The above assumptions are typical for consideration of *ordinal classification problems with monotonicity constraints*, also called *multiple criteria sorting problems*.

2.2. Rough approximations

The sets to be approximated are called *upward union* and *downward union* of classes, respectively:

$$X_t^{\geq} = \bigcup_{s \geq t} X_s, \quad X_t^{\leq} = \bigcup_{s \leq t} X_s, \quad t = 1, \ldots, n. \tag{1}$$

The statement $x \in X_t^{\geq}$ means "x belongs to at least class X_t", while $x \in X_t^{\leq}$ means "x belongs to at most class X_t". Let us remark that $X_1^{\geq} = X_n^{\leq} = U$, $X_n^{\geq} = X_n$ and $X_1^{\leq} = X_1$. Furthermore, for t=2,…,n,

$$X_t^{\leq} = U - X_{t-1}^{\geq} \quad \text{and} \quad X_{t-1}^{\geq} = U - X_t^{\leq}. \tag{2}$$

The key idea of the rough set approach is representation (approximation) of knowledge generated by decision attributes, using "*granules of knowledge*" generated by condition attributes. In DRSA, where condition attributes are ordinal and decision classes are ordered, the represented knowledge is a collection of upward and downward unions of classes and the "granules of knowledge" are sets of objects defined using a *dominance relation*. Dominance relation is defined with respect to $P \subseteq C$. *x dominates y*, denoted by *xDy*, if for every ordinal attribute $q \in P$, $\phi(x,q) \geq \phi(y,q)$. The relation of dominance is reflexive and transitive, that is it is a partial preorder.

Given a set of ordinal attributes $P \subseteq C$ and $x \in U$, the "granules of knowledge" used for approximation in DRSA are:

- a set of objects dominating x, called *dominating set*, $D^+(x) = \{y \in U : yDx\}$,
- a set of objects dominated by x, called *dominated set*, $D^-(x) = \{y \in U : xDy\}$.

Remark that the "granules of knowledge" defined above have the form of upward (positive) and downward (negative) *dominance cones* in the evaluation space.

Let us recall that the *dominance principle* (or Pareto principle) requires that an object x dominating object y on all considered ordinal attributes (i.e. x having evaluations at least as high (good) as y on all considered attributes) should also dominate y on the decision (i.e. x should be assigned to at least as high (good) decision class as y).

Violation of the dominance principle leads to *inconsistency w.r.t. dominance*. Given $P \subseteq C$, the inclusion of an object $x \in U$ to the upward union of classes $X_t^{\geq}$, t=2,…,n, is inconsistent w.r.t. dominance if one of the following conditions holds:

- x belongs to class X_t or higher but it is dominated by an object y belonging to a class lower than X_t, i.e. $x \in X_t^{\geq}$ but $D^+(x) \cap X_{t-1}^{\leq} \neq \emptyset$,
- x belongs to a lower class than X_t but it dominates an object y belonging to class X_t or higher, i.e. $x \notin X_t^{\geq}$ but $D^-(x) \cap X_t^{\geq} \neq \emptyset$.

If, given a set of ordinal attributes $P \subseteq C$, the inclusion of $x \in U$ to $X_t^{\geq}$, where $t = 2, \ldots, n$, is inconsistent w.r.t. dominance, then x belongs to $X_t^{\geq}$ *with some ambiguity*. Thus, x belongs to $X_t^{\geq}$ *without any ambiguity* if $x \in X_t^{\geq}$ and there is no inconsistency w.r.t. dominance. This means that all objects dominating x belong to $X_t^{\geq}$, i.e. $D^+(x) \subseteq X_t^{\geq}$.

Furthermore, *x possibly belongs to $X_t^{\geq}$* if one of the following conditions holds:

- according to decision attribute d, x belongs to $X_t^{\geq}$,
- according to decision attribute d, x does not belong to $X_t^{\geq}$, but it is inconsistent w.r.t. dominance with an object y belonging to $X_t^{\geq}$.

In terms of ambiguity, x possibly belongs to $X_t^{\geq}$, if x belongs to $X_t^{\geq}$ with or without any ambiguity. Due to the reflexivity of the dominance relation, the above conditions can be summarized as follows: *x possibly belongs* to class X_t or higher, if among the objects dominated by x there is an object y belonging to class X_t or higher, i.e. $D^-(x) \cap X_t^{\geq} \neq \emptyset$.

The *lower approximation* of $X_t^{\geq}$, denoted by $\underline{X}_t^{\geq}$, and the *upper approximation* of $X_t^{\geq}$, denoted by $\overline{X}_t^{\geq}$, are defined as follows ($t = 1, \ldots, n$):

$$\underline{X}_t^{\geq} = \{x \in U : D^+(x) \subseteq X_t^{\geq}\}, \tag{3}$$

$$\overline{X}_t^{\geq} = \{x \in U : D^-(x) \cap X_t^{\geq} \neq \emptyset\}. \tag{4}$$

Analogously, one can define the lower approximation and the upper approximation of $X_t^{\leq}$ as follows ($t = 1, \ldots, n$):

$$\underline{X}_t^{\leq} = \{x \in U : D^-(x) \subseteq X_t^{\leq}\}, \tag{5}$$

$$\overline{X}_t^{\leq} = \{x \in U : D^+(x) \cap X_t^{\leq} \neq \emptyset\}. \tag{6}$$

The lower and upper approximations so defined satisfy the following properties for all $P \subseteq C$:

$$\underline{X}_t^{\geq} \subseteq X_t^{\geq} \subseteq \overline{X}_t^{\geq}$$

and

$$\underline{X}_t^{\leq} \subseteq X_t^{\leq} \subseteq \overline{X}_t^{\leq}, \quad t = 1, \ldots, n,$$

$$\underline{X}_t^{\geq} = U - \overline{X}_{t-1}^{\leq}$$

and

$$\overline{X}_t^{\geq} = U - \underline{X}_{t-1}^{\leq}, \quad t = 2, \ldots, n,$$

$$\underline{X}_t^{\leq} = U - \overline{X}_{t+1}^{\geq}$$

and

$$\overline{X}_t^{\leq} = U - \underline{X}_{t+1}^{\geq}, \quad t = 1, \ldots, n-1.$$

2.3. *Variable consistency rough approximations*

In DRSA, lower approximation of a union of ordered classes contains only consistent objects. This definition of the lower approximation appears to be too restrictive in practical applications. In the consequence, lower approximations may be even empty, preventing generalization of data in terms of decision rules. This observation has motivated research on generalizations of definition of lower approximation.

One of the possibilities is a generic definition of extended lower approximation, which is defined as Variable Consistency Dominance-based Rough Set Approach (VC-DRSA) (Błaszczyński et al., 2009; Greco et al., 2000b). This definition allows to include the lower approximation objects with sufficient evidence for membership to approximated union of decision classes. The evidence is quantified by *consistency measures*. In Błaszczyński et al. (2009), we distinguished gain-type and cost-type consistency measures, and we specified conditions that must be satisfied by these measures. For $P \subseteq C, y \in U$, given a gain-type (resp. cost-type) object consistency measure $\Theta(y)$ and a gain-threshold (resp. cost-threshold) θ, the lower approximation of $X_t^{\geq}$, and the lower approximation of $X_t^{\leq}$ are defined as

$$\underline{X}_t^{\geq} = \{y \in X_t^{\geq} : \Theta_{X_t^{\geq}}(y) \propto \theta_{X_t^{\geq}}\}, \qquad (7)$$

$$\underline{X}_t^{\leq} = \{y \in X_t^{\leq} : \Theta_{X_t^{\leq}}(y) \propto \theta_{X_t^{\leq}}\},$$

where $\propto$ denotes $\geq$ in case of a gain-type object consistency measure and a gain-threshold, or $\leq$ for a cost-type object consistency measure and a cost-threshold. In the above definition, $\theta_{X_t^{\geq}} \in [0, A_{X_t^{\geq}}]$, and $\theta_{X_t^{\leq}} \in [0, A_{X_t^{\leq}}]$ are technical parameters indicating a limit degree of consistency of objects belonging to the corresponding lower approximation.

2.4. *Decision rules induced from rough approximations*

The lower approximations of upward and downward unions of classes can serve to induce "*if..., then...*" decision rules. In DRSA, such rules are called certain because their credibility is full. In VC-DRSA, decision rules induced from lower approximations are, in general, not fully credible, so they are characterized by a consistency measure. Using DRSA or VC-DRSA, one can induce decision rules with the following syntax:

$$\text{if } q_{i_1}(x) \succcurlyeq t_{i_1} \wedge \cdots \wedge q_{i_p}(x) \succcurlyeq t_{i_p}, \text{ then } x \in X_t^{\geq}, \qquad (8)$$

$$\text{if } q_{i_1}(x) \succcurlyeq t_{i_1} \wedge \cdots \wedge q_{i_p}(x) \succcurlyeq t_{i_p}, \text{ then } x \in X_t^{\leq}, \qquad (9)$$

where $q_{i_1}, \ldots, q_{i_p}$ denote ordinal attributes, and t_{i_j} denotes a value taken from the value set of attribute q_{i_j}, $i_j \in \{i_1, \ldots, i_p\} \subseteq \{1, \ldots, |C|\}$. We use symbols $\succcurlyeq$ and $\succcurlyeq$ to indicate weak order relation and inverse weak order relation w.r.t. the specified ordinal attribute, respectively.

Induction of decision rules is a complex problem and many algorithms have been introduced to solve it. Algorithms proposed specifically for DRSA and VC-DRSA have been described in Greco et al. (2000a), Blaszczyński and Slowiński (2003), and Błaszczyński et al. (2011).

Once the rules are induced, they can be used to classify objects. The *standard classification method* used with DRSA and VC-DRSA has been presented in Greco et al. (2002). In this procedure, an object covered by a set of rules is assigned to a class (or a set of contiguous classes) resulting from intersection of unions of decision classes suggested by the rules. In Blaszczyński et al. (2007), we presented a *new classification method* for DRSA and VC-DRSA. It is based on a notion of a class score coefficient associated with a set of rules covering the classified object. The object is assigned to a class getting the highest score.

3. Transformation of classification table

The transformation method which is described below allows application of DRSA to classification problems $(\alpha)-(\delta)$. It should also be applied to all non-ordinal condition attributes present in classification problem (ε).

We assume, without loss of generality, that the value sets of both decision attribute (class labels) and condition attributes are number-coded. As in classification problems (α),(β) the complete ordering of classes $X_1, X_2, \ldots, X_n$ induced by number-coded class labels is not entering, in general, into some monotonic relationships with value sets of condition attributes, we have to consider n binary ordinal classification problems with two sets of objects: class X_t and its complement $\neg X_t$, $t = 1, \ldots, n$, which are number-coded by 1 and 0, respectively. This means that in the t-th ordinal binary classification problem, set X_t is interpreted by DRSA as union $X_1^{\geq}$ and set $\neg X_t$ as union $X_0^{\leq}$, $t = 1, \ldots, n$. Classification problems (γ), (δ) and (ε) can be handled by DRSA without altering the original number codes of the class labels. In case of any classification problem (ordinal or not) with two classes only (X_1, X_2, $n = 2$), we consider it as an ordinal binary classification problem, where union $X_1^{\geq}$ is composed of all objects belonging to X_1, and union $X_0^{\leq}$ is composed of all objects belonging to X_2.

As to non-ordinal condition attributes, we have to distinguish two kinds of them: numerical and nominal. While numerical attributes are obviously number coded, nominal condition attributes must be binarized and get 0–1 codes that represent absence or presence of a given nominal value.

To allow discovery of some local monotonic relationships between the values of condition attributes and the assignment of objects to union $X_1^{\geq}$ or to union $X_0^{\leq}$, we clone each of the number-coded condition attribute. Every cloned attribute is supposed to enter into one of the two possible monotonicity relations with the class assignment: positive or negative. Positive relationship means that the greater the value of the condition attribute, the higher the number code (rather 1 than 0) of the class assignment, and negative relationship means that the greater the value of the condition attribute, the lower the number code (rather 0 than 1) of the class assignment. As a result, we get pairs of number-coded attributes with supposed inverse monotonic relationships to the class assignment. The redundancy in description of objects by attributes is necessary because it is the monotonic rules that are to discover the correct direction of the monotonicity relationship.

Formally, the original classification table $\mathbf{S}$ including set U of objects described by set A of attributes is transformed into classification table $\mathbf{S}'$ including number-coded and cloned (possibly binarized) non-ordinal condition attributes. In case of classification problems (γ), (δ), and (ε), $\mathbf{S}'$ is handled by DRSA without altering the original number codes of the class labels. In case of classification problems (α) and (β), classification table $\mathbf{S}'$ undergoes one more transformation: it is replaced by n classification tables $\mathbf{S}^{\mathbf{t}\prime}, t = 1, \ldots, n$, that represent each of the binary ordinal classification problems resulting from transformation of the original decision attribute.

In classification problems of type $(\alpha)-(\delta)$ each one of the condition attributes has to be transformed. In case of classification problem (ε), only non-ordinal condition attributes are to be transformed.

The transformation of each non-ordinal condition attribute from A is made individually, depending on its type:

(1) numerical (number-coded),
(2) nominal.

Each numerical (number-coded) attribute a_i is represented in $\mathbf{S}'$ (or in $\mathbf{S}^{\mathbf{t}'}, t = 1, \ldots, n$), as a pair of ordinal attributes q_i', and q_i''.

The first one in the pair, q'_i, is set to have positive monotonic relationship with (possibly transformed) decision attribute, while the second one, q''_i, is set to have negative monotonic relationship with the decision attribute, i.e. the second one gets the opposite ordering. In other words, evaluation of each object $x \in U$ by numerical attribute a_i is repeated twice in $\mathbf{S}'$, and the first evaluation $a_i(x)$ is renamed to $q'_i(x)$, while the second evaluation $a_i(x)$ is renamed to $q''_i(x)$.

Illustration of the transformation of a numerical attribute is presented in Fig. 1. Note that the transformation does not introduce inconsistency w.r.t. dominance. For each object x, and set of transformed ordinal attributes $P' \subseteq C'$, the dominance cones $D^+(x)$, and $D^-(x)$ are composed of x and all objects that have exactly the same description by P' as x. Still, object x may be inconsistent w.r.t. dominance if there are some other objects that have the same description by P' as x and at least one of them has a different class label than x. Nevertheless, such inconsistency is also present in the original classification table $\mathbf{S}$.

Each nominal attribute a_j that has value set composed of k distinct values is binarized, such that the presence or absence of the l-th value of this attribute is coded by a new ordinal attribute q_{jl} taking value 1 or 0, respectively, $l=1,\dots,k$. Then, the binary attribute q_{jl} is represented in $\mathbf{S}'$ (or in $\mathbf{S}^{t'}, t=1,\dots,n$), by a pair of binary ordinal attributes, q'_{jl} and q''_{jl}, $l=1,\dots,k$. The first one in each of the pairs, q'_{jl}, is set to have positive monotonic relationship with the decision attribute, while the second one, q''_{jl}, is set to have negative monotonic relationship with the decision attribute. In other words, evaluation of each object $x \in U$ by nominal attribute a_j, denoted by $a_j(x)$, is first described by k binary attributes q_{jl}, such that $q_{jl}(x)=1$ if $l=a_j(x)$, and $q_{jl}(x)=0$ otherwise, $l=1,\dots,k$. Then, each binary evaluation $q_{jl}(x)$ is repeated twice in $\mathbf{S}'$, and the first one is renamed to $q'_{jl}(x)$, while the second one is renamed to $q''_{jl}(x)$.

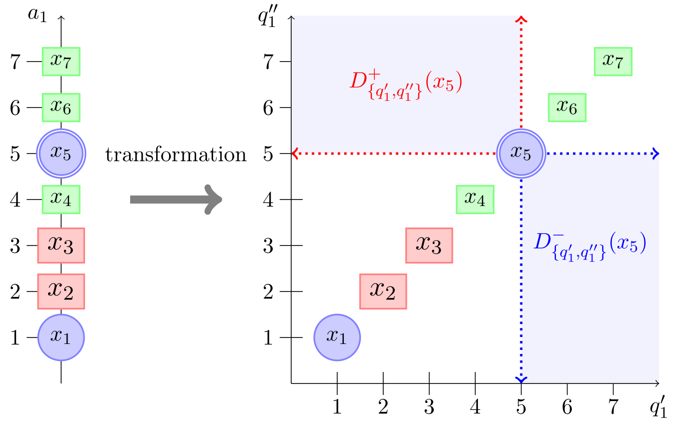


Fig. 1. Illustration of transformation of numerical attribute a_1.

Illustration of the transformation of a nominal attribute is presented in Fig. 2. The transformation of the nominal attribute does not introduce new inconsistency w.r.t. dominance.

4. Classification

In case of binary classification problems (α) and (β), and classification problems (γ), (δ), and (ε), the assignment of a (new) object x by a set of monotonic decision rules induced by DRSA or VC-DRSA from the transformed classification table $\mathbf{S}'$, is performed according to the scheme presented in Blaszczyński et al. (2007). The classification scheme needs to be updated in case of transformed classification problems of type (α) and (β), if the number of decision classes $n>2$. In this section, we present the updated scheme.

We consider a (new) object x to be assigned to X_t (considered as union $X_1^{\geq}$) or to $\neg X_t$ (considered as union $X_0^{\leq}$) by decision rules induced from $\mathbf{S}^{t'}$ ($t \in \{1,\dots,n\}$). The classification scheme is based on a notion of class score coefficient associated with a set of rules covering the object to be classified. Let us remind three situations that may occur in case of classification by a given set of rules:

1. None of the rules cover object x.
2. Exactly one decision rule covers object x.
3. Several rules cover object x.

Situation 1 results in object x being assigned to all considered decision classes.

Situation 2 is relatively simple. The classification involves calculation of a score coefficient that reflects relevance between rules and the suggested class assignment. For rule r_{X_t} covering object x and having decision part "*then* $x \in X_1^{\geq}$", a value of $score_{r_{X_t}}(X_t,x)$ is calculated as

$$score_{r_{X_t}}(X_t,x) = \frac{|\|\Phi_{r_{X_t}}\| \cap X_t|^2}{|\|\Phi_{r_{X_t}}\||\,|X_t|}, \tag{10}$$

where $\|\Phi_{r_{X_t}}\|$ denotes the set of objects verifying the condition part of rule r_{X_t}, and $|\|\Phi_{r_{X_t}}\||$, $|X_t|$ and $|\|\Phi_{r_{X_t}}\| \cap X_t|$ denote cardinalities of the corresponding sets: the set of objects verifying $\Phi_{r_{X_t}}$, the set of objects belonging to class X_t, and the set of objects verifying $\Phi_{r_{X_t}}$ and belonging to class X_t. Note that $score_{r_{X_t}}(X_t,x)$ is

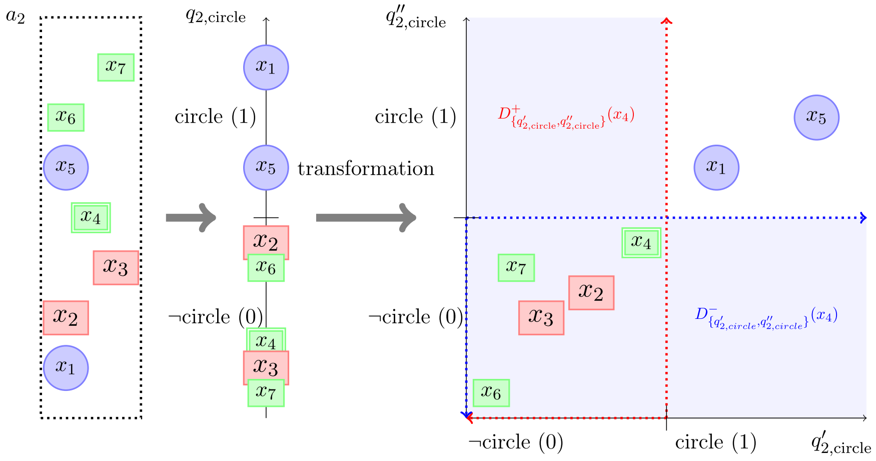


Fig. 2. Illustration of transformation of nominal attribute a_2.

a product of confidence and coverage of all rules matching the description of object x and suggesting its assignment to class X_t. From probabilistic point of view, coefficient $score_{r_{X_t}}(X_t,x)$ can be presented as a product of two conditional probabilities. The first conditional probability (coverage), $\Pr(\|\Phi_{r_{X_t}}\| \,|\, X_t)$, says what is the probability of covering object x by a rule suggesting assignment to X_t, given that object x belongs to X_t. The second one (confidence), $\Pr(X_t \,|\, \|\Phi_{r_{X_t}}\|)$, says what is the probability that object x belongs to X_t, given x is covered by a rule suggesting assignment to X_t. Thus, coefficient $score_{r_{X_t}}(X_t,x)$ can be rewritten as $\Pr(\|\Phi_{r_{X_t}}\| \cap X_t)^2/\Pr(\|\Phi_{r_{X_t}}\|)\Pr(X_t)$. Note that $\Pr(\|\Phi_{r_{X_t}}\| \cap X_t) = \Pr(\|\Phi_{r_{X_t}}\|)\Pr(X_t)$ when the two events: covering of object x by a rule suggesting assignment to X_t and object x belongs to X_t are independent. Coefficient $score_{r_{X_t}}(X_t,x)$ is thus measuring the relevance between these two events.

Analogously, for rule $r_{\neg X_t}$ matching x and having decision part "*then* $x \in X_0^{\leq}$", a value of $score_{r_{\neg X_t}}(\neg X_t,x)$ is calculated as

$$score_{r_{\neg X_t}}(\neg X_t,x) = \frac{\left| \|\Phi_{r_{\neg X_t}}\| \cap \neg X_t \right|^2}{\left| \|\Phi_{r_{\neg X_t}}\| \right| \left| \neg X_t \right|}. \tag{11}$$

If object x is covered by a rule suggesting the classification decision "*then* $x \in X_1^{\geq}$", the final score for class X_t and object x is

$$score(X_t,x) = score_{r_{X_t}}(X_t,x), \tag{12}$$

If, however, object x is covered by a rule suggesting the classification decision "*then* $x \in X_0^{\leq}$", the final score for class X_t and object x is

$$score(X_t,x) = -score_{r_{\neg X_t}}(\neg X_t,x). \tag{13}$$

Situation 3 requires that we divide the set of rules covering object x into two subsets: those that suggest assignment of x to X_t and those that suggest assignment of x to $\neg X_t$. Then, we calculate the value of score coefficient $score^{+}_{r_{X_t}}(X_t,x)$ for rules covering object x and having decision part "*then* $x \in X_1^{\geq}$":

$$score^{+}_{r_{X_t}}(X_t,x) = \frac{\left|(\|\Phi_1\| \cap X_t) \cup \cdots \cup (\|\Phi_k\| \cap X_t)\right|^2}{\left|\|\Phi_1\| \cup \cdots \cup \|\Phi_k\|\right| \left|X_t\right|}. \tag{14}$$

We also calculate the value of score coefficient $score^{-}_{r_{\neg X_t}}(X_t,x)$ for rules covering object x and having decision part "*then* $x \in X_0^{\leq}$":

$$score^{-}_{r_{\neg X_t}}(X_t,x) = \frac{\left|(\|\Phi_1\| \cap \neg X_t) \cup \cdots \cup (\|\Phi_l\| \cap \neg X_t)\right|^2}{\left|\|\Phi_1\| \cup \cdots \cup \|\Phi_l\|\right| \left|\neg X_t\right|}. \tag{15}$$

The value of the final score for class X_t and object x is

$$score(X_t,x) = score^{+}_{r_{X_t}}(X_t,x) - score^{-}_{r_{\neg X_t}}(X_t,x). \tag{16}$$

Let us observe, that analogously to (14) and (15), we can calculate the score coefficients for the complement of class X_t:

$$score^{+}_{r_{\neg X_t}}(\neg X_t,x) = \frac{\left|(\|\Phi_1\| \cap \neg X_t) \cup \cdots \cup (\|\Phi_l\| \cap \neg X_t)\right|^2}{\left|\|\Phi_1\| \cup \cdots \cup \|\Phi_l\|\right| \left|\neg X_t\right|}, \tag{17}$$

$$score^{-}_{r_{X_t}}(\neg X_t,x) = \frac{\left|(\|\Phi_1\| \cap X_t) \cup \cdots \cup (\|\Phi_k\| \cap X_t)\right|^2}{\left|\|\Phi_1\| \cup \cdots \cup \|\Phi_k\|\right| \left|X_t\right|}. \tag{18}$$

The value of the final score for class $\neg X_t$ is calculated analogously to (16):

$$score(\neg X_t,x) = score^{+}_{r_{\neg X_t}}(\neg X_t,x) - score^{-}_{r_{X_t}}(\neg X_t,x). \tag{19}$$

Let us observe that, taking into account (14), (15), (17), (18), the following properties hold:

$$score^{+}_{r_{X_t}}(X_t,x) = score^{-}_{r_{X_t}}(\neg X_t,x), \tag{20}$$

$$score^{+}_{r_{\neg X_t}}(\neg X_t,x) = score^{-}_{r_{\neg X_t}}(X_t,x). \tag{21}$$

Moreover, according to definition (16), properties (20) and (21), we get

$$\begin{aligned} score(X_t,x) &= score^{+}_{r_{X_t}}(X_t,x) - score^{-}_{r_{\neg X_t}}(X_t,x) \\ &= score^{-}_{r_{\neg X_t}}(\neg X_t,x) - score^{+}_{r_{\neg X_t}}(\neg X_t,x) \\ &= -score(\neg X_t,x). \end{aligned}$$

When classifying object x, the final score $score(X_t,x)$ is calculated for each class X_t, $t = 1,\ldots,n$. If for at least one of the classes, classification situation is different from *situation* 1, and at least one of the final scores is positive, the class with the highest value of the score is selected for the final assignment of x. Otherwise, the classification result is unknown for x.

5. Illustrative example

Let us consider the following illustrative classification problem, which is described by non-ordinal attributes only (classification problem of type (α)). The objects are patients after radical prostatectomy, and the decision attribute specifies if there is recurrence of the disease or not. Moreover, in case of recurrence it may be local (return of the cancer in the same place) or not (occurrence of the cancer in other places). Thus, the values of decision attribute Recurrence are: "no", "local", and "other". Condition attributes that describe patients are the following. The first two are integer valued attributes: Age and Gleason score. Tumor Volume is a nominal attribute with value set composed of three values: "small", "medium", and "large". The values of PSA are continuous. Classification table with the exemplary set of patients after radical prostatectomy is presented in Table 1. Observe that there are two patients in the table, whose description is inconsistent. Patient 5 and patient 6 have the same description by condition attributes, yet the first one had local recurrence, while the second one did not have.

To apply DRSA or VC-DRSA to this classification table, we need to transform Table 1 using the methodology presented in Section 3. The classification tables resulting from this transformation are presented in Tables 2–4. Table 2 concerns classification of objects into two ordered classes: "no" and "¬ no", i.e. to "no", or to "local" or "other". The two classes are coded by 1 and 0, respectively, which corresponds to their order. Similarly, Tables 3 and 4, concern binary classification with ordered classes "local" and "¬ local", and "other" and "¬ other", which are also coded by 1 and 0, respectively.

The inconsistencies found in Table 1 are also present in Tables 2 and 3. This is not surprising because the inconsistencies in Table 1 have been stated for objects belonging to class "no" and to class "local". It is worth stressing that no new inconsistent objects occur after the transformation. As no inconsistencies existed for objects belonging to class "other", Table 4 continues to be composed of consistent objects only. This means that the transformation has *been* non-invasive.

The set of transformed attributes $P' = \{\text{Age}', \text{Age}'', \text{Gleason}', \text{Gleason}'', \text{PSA}', \text{PSA}', \text{V-s}', \text{V-s}'', \text{V-m}', \text{V-m}'', \text{V-l}', \text{V-l}''\}$ includes binary

Table 1
Set of patients after radical prostatectomy.

Id	Age	Gleason	PSA	Volume	Recurrence
1	60	10	2.0	large	other
2	20	7	1.2	large	local
3	40	4	0.1	medium	local
4	45	2	0.8	medium	no
5	50	3	0.3	small	local
6	50	3	0.3	small	no
7	40	7	0.5	small	no
8	25	5	0.4	small	no
9	25	2	0.5	small	no
10	40	4	0.5	small	no

Table 2
Transformed set of patients after radical prostatectomy—binary classification into “no” and “¬ no”.

Id	Age′	Age″	Gleason′	Gleason″	PSA′	PSA″	V-s′	V-s″	V-m′	V-m″	V-l′	V-l″	R-no
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑
1	60	60	10	10	2.0	2.0	0	0	0	0	1	1	0
2	20	20	7	7	1.2	1.2	0	0	0	0	1	1	0
3	40	40	4	4	0.2	0.2	0	0	1	1	0	0	0
4	45	45	2	2	0.8	0.8	0	0	1	1	0	0	1
5	50	50	3	3	0.3	0.3	1	1	0	0	0	0	0
6	50	50	3	3	0.3	0.3	1	1	0	0	0	0	1
7	40	40	7	7	0.6	0.6	1	1	0	0	0	0	1
8	25	25	5	5	0.4	0.4	1	1	0	0	0	0	1
9	25	25	2	2	0.5	0.5	1	1	0	0	0	0	1
10	40	40	4	4	0.5	0.5	1	1	0	0	0	0	1

Table 3
Transformed set of patients after radical prostatectomy—binary classification into “local” and “¬ local”.

Id	Age′	Age″	Gleason′	Gleason″	PSA′	PSA″	V-s′	V-s″	V-m′	V-m″	V-l′	V-l″	R-local
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑
1	60	60	10	10	2.0	2.0	0	0	0	0	1	1	0
2	20	20	7	7	1.2	1.2	0	0	0	0	1	1	1
3	40	40	4	4	0.2	0.2	0	0	1	1	0	0	1
4	45	45	2	2	0.8	0.8	0	0	1	1	0	0	0
5	50	50	3	3	0.3	0.3	1	1	0	0	0	0	1
6	50	50	3	3	0.3	0.3	1	1	0	0	0	0	0
7	40	40	7	7	0.6	0.6	1	1	0	0	0	0	0
8	25	25	5	5	0.4	0.4	1	1	0	0	0	0	0
9	25	25	2	2	0.5	0.5	1	1	0	0	0	0	0
10	40	40	4	4	0.5	0.5	1	1	0	0	0	0	0

Table 4
Transformed set of patients after radical prostatectomy—binary classification into “other” and “¬ other”.

Id	Age′	Age″	Gleason′	Gleason″	PSA′	PSA″	V-s′	V-s″	V-m′	V-m″	V-l′	V-l″	R-other
	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑	↓	↑
1	60	60	10	10	2.0	2.0	0	0	0	0	1	1	1
2	20	20	7	7	1.2	1.2	0	0	0	0	1	1	0
3	40	40	4	4	0.2	0.2	0	0	1	1	0	0	0
4	45	45	2	2	0.8	0.8	0	0	1	1	0	0	0
5	50	50	3	3	0.3	0.3	1	1	0	0	0	0	0
6	50	50	3	3	0.3	0.3	1	1	0	0	0	0	0
7	40	40	7	7	0.6	0.6	1	1	0	0	0	0	0
8	25	25	5	5	0.4	0.4	1	1	0	0	0	0	0
9	25	25	2	2	0.5	0.5	1	1	0	0	0	0	0
10	40	40	4	4	0.5	0.5	1	1	0	0	0	0	0

ordinal attributes V-s′,V-s″,V-m′,V-m″,V-l′,V-l″ resulting from transformation of nominal attribute Volume. For Table 2, $\underline{X}_0^{\leq} = \{1,2,3\}$, while $\underline{X}_1^{\geq} = \{4,7,8,9,10\}$.

Two decision rules induced from P'-lower approximations of $\underline{X}_1^{\geq}$ and $\underline{X}_0^{\leq}$ using DRSA are sufficient to cover all consistent objects from Table 2. These rules are:

1 : *if* Gleason″ ≥ 4 *and* V-s′ ≤ 0, *then* R-no ≤ 0,

2 : *if* PSA′ ≥ 0.4 *and* PSA″ ≤ 0.8, *then* R-no ≥ 1.

The first rule covers all objects from $\underline{X}_0^{\leq}$, while the second rule covers all objects from $\underline{X}_1^{\geq}$. Remark that elementary condition V-s′ ≤ 0 from the first rule can be read as “Volume is not small”. Thus, this elementary condition can be expressed in terms of the original attribute as: Volume $\in$ {medium,large}. Moreover, in the second rule, the elementary conditions based on the cloned numerical attribute, PSA′ ≥ 0.4 and PSA″ ≤ 0.8, can be synthesized into an interval condition expressed in terms of the original attribute as: PSA $\in$ [0.4,0.8]. In consequence of this synthesis, the rules become more readable:

1 : *if* Gleason ≥ 4 *and* Volume $\in$ {medium, large},
then Recurrence is ¬*no*,

2 : *if* PSA $\in$ [0.4,0.8], *then* Recurrence is no.

For Table 3, $\underline{X}_0^{\leq} = \{1,4,7,8,9,10\}$, while $\underline{X}_1^{\geq} = \{2,3\}$. Two decision rules are also sufficient to cover all consistent objects from Table 3. These rules are

3 : *if* Age″ ≥ 25 *and* PSA″ ≥ 0.4, *then* R-local ≤ 0,

4 : *if* Age″ ≤ 40 *and* V-s‴ ≤ 0, *then* R-local ≥ 1.

The first rule covers all objects from $\underline{X}_0^{\leq}$, while the second rule covers all objects from $\underline{X}_1^{\geq}$. The rules can be expressed in terms of the original attributes as

3 : *if* Age ≥ 25 *and* PSA ≥ 0.4, *then* Recurrence is ¬local,

4 : *if* Age ≤ 40 *and* Volume $\in$ {medium,large}, *then*
Recurrence is local.

For Table 4, $\underline{X}_0^{\leq} = \{2,3,4,5,6,7,8,9,10\}$, while $\underline{X}_1^{\geq} = \{1\}$.

Table 5
Patient to be classified.

Id	Age′ ↑	Age″ ↓	Gleason′ ↑	Gleason″ ↓	PSA′ ↑	PSA″ ↓	V-s′ ↑	V-s″ ↓	V-m′ ↑	V-m″ ↓	V-l′ ↑	V-l″ ↓
11	30	30	2	2	0.6	0.6	1	1	0	0	0	0

Two decision rules are, as well, sufficient to cover all consistent objects from Table 4. These rules are

$$5:\ \textit{if}\ \text{PSA}' \leq 1.2,\ \textit{then}\ \text{R-other} \leq 0,$$

$$6:\ \textit{if}\ \text{PSA}'\ \geq 2,\ \textit{then}\ \text{R-other}\ \geq 1.$$

The first rule covers all objects from $\underline{X}_0^{\leq}$, while the second rule covers all objects from $\underline{X}_1^{\geq}$. The rules can be expressed in terms of the original attributes as

$$5:\ \textit{if}\ \text{PSA}\ \leq 1.2,\ \textit{then}\ \text{Recurrence is}\neg\text{other},$$

$$6:\ \textit{if}\ \text{PSA}\ \geq 2,\ \textit{then}\ \text{Recurrence is other}.$$

Note that the rules include elementary conditions of the type "attribute $a_i \in [r_i^1, r_i^2]$" and "attribute $a_i \in \{v_i^1, \ldots, v_i^k\}$". It was possible to discover such rules by DRSA due to the presented transformation of the classification table.

Let us suppose that a new patient (Id=11) is classified by the discovered rules. Description of patient 11 in terms of the transformed attributes from P' is presented in Table 5.

Patient 11 is covered by the following rules:

- rule 2, suggesting assignment to class "no",
- rule 3, dissuading assignment to class "local" (i.e. suggesting assignment to "¬ local"),
- rule 5, dissuading assignment to class "other" (i.e. suggesting assignment to "¬ other").

Thus, according to the procedure described in Section 4, patient 11 is assigned to class "no". This is because the three matching rules produce the following scores:

$$score_{r_{no}}(no, x_{11}) = \frac{5^2}{5 \times 5} = 1,$$

$$score_{r_{\neg local}}(\neg local, x_{11}) = \frac{6^2}{6 \times 6} = 1,$$

$$score_{r_{\neg other}}(\neg other, x_{11}) = \frac{9^2}{9 \times 9} = 1,$$

which leads to the following final score:

$$score(no, x_{11}) = 1,$$

$$score(local, x_{11}) = -1,$$

$$score(other, x_{11}) = -1.$$

6. Results of a computational experiment

The main goal of the computational experiment was to assess the predictive accuracy of the rule classifier presented in Section 4 when applied to non-ordinal classification problems transformed according to the method described in Section 3. All experiments were carried out on 20 data sets from the UCI repository.[1] Characteristics of all these data sets are given in Table 6. The sets of rules used in classification were induced using VC-DomLEM algorithm (Błaszczyński et al., 2011). VC-DomLEM is a sequential covering rule induction algorithm proposed for DRSA and VC-DRSA. Rough membership measure (Wong and Ziarko, 1987, and Pawlak and Skowron, 1994) was used with VC-DomLEM. The following non-ordinal classifiers were included in the comparison: support vector machine (SVM) with linear kernel (Platt, 1998), decision rule classifier RIPPER (Cohen, 1995), and decision tree classifier C 4.5 (Quinlan, 1992). Moreover, MODLEM (Stefanowski, 1998) was included. MODLEM is a rule classifier induced by sequential covering algorithm developed within the classical rough set approach.

Table 6
Characteristics of data sets.

Data set	Objects	Attributes	Classes
Arythmia	452	280	13
Autos	205	26	7
Breast-cancer	286	10	2
Bupa	345	6	2
Credit-g	1000	20	2
crx	690	16	2
Dermatology	366	35	6
Diabetes	768	8	2
Ecoli	336	7	8
Glass	214	9	7
Heart-c	303	14	2
Hypothyroid	3772	30	4
Page-blocks	5473	11	5
Pima	768	8	2
Sonar	208	60	2
Soybean	683	36	19
Spambase	4601	58	2
Vehicle	846	18	4
Vowel	990	14	11
Wine	178	14	3

The classification accuracy was estimated by the stratified 10-fold cross-validation, which was repeated five times to get reproducible results. Table 7 presents the average classification accuracy and its standard deviation for each data set and each classifier. Moreover, for each data set, we calculated the rank of the result obtained by a classifier in comparison with other classifiers. The rank is presented in brackets (the smaller the rank, the better). We show these ranks because they are used in statistical tests described further. The last row of each table shows the average rank obtained by a given classifier. Moreover, for each data set, the best value of the predictive accuracy measure, and those values which are within standard deviation of the best value, are marked as bold.

We used statistical tests to compare differences in predictive accuracy between considered classifiers. First, we applied Friedman test to globally compare performance of six different classifiers on multiple data sets (Demsar, 2006; Kononenko and Kukar, 2007). The null-hypothesis in this test was that all compared classifiers perform equally well in terms of average classification accuracy. Unfortunately, we were not able to reject this hypothesis. This is mostly due to the fact that we applied the weak and conservative nonparametric test. The difference in ranks must be very high in order to conclude by this test that one classifier is better than another.

We continued our experimental comparison with examination of importance of the difference in average classification accuracy

---

[1] See http://www.ics.uci.edu/~mlearn/MLRepository.html.

Table 7
Percentage of correct classifications in repeated 10-fold cross validation.

Data set	VC-DomLEM	Naive Bayes	SVM	RIPPER	C4.5	MODLEM
Arrhythmia	**71.73** (1)	62.17 (6)	**71.15** (2)	69.82 (3)	65.8 (5)	66.02 (4)
	±0.9428	±0.7664	±0.8916	±1.150	±1.361	±1.424
Autos	78.34 (3)	56 (6)	71.71 (5)	72.68 (4)	**79.9** (2)	**81.27** (1)
	±1.004	±2.783	±1.195	±1.272	±1.961	±1.561
Breast-cancer	70.42 (5)	72.45 (2)	70.77 (4)	71.61 (3)	**75.03** (1)	68.32 (6)
	±2.283	±0.5139	±1.119	±1.457	±0.5233	±1.923
Bupa	**69.28** (1)	55.48 (6)	57.97 (5)	66.84 (3)	66.26(4)	**68.7** (2)
	±1.335	±0.8715	±0.1833	±0.6507	±1.830	±1.729
Credit-g	71.8 (4)	74.98 (2)	**75.42** (1)	72.52 (3)	71.78 (5)	71.68 (6)
	±0.938	±0.4167	±0.3487	±0.4578	±0.7414	±1.107
crx	82.84 (5)	78.14 (6)	84.75 (3)	**85.45** (1)	**85.3** (2)	84.09 (4)
	±0.7252	±0.05797	±0.1085	±0.9374	±0.4546	±1.01
Dermatology	95.96 (3)	**97.54** (1)	96.28 (2)	88.42 (6)	93.66 (4)	92.68 (5)
	±0.4372	±0.1728	±0.5354	±0.4764	±0.5573	±0.8536
Diabetes	74.01 (4)	75.5 (2)	**76.82** (1)	74.84 (3)	73.88 (5)	72.73 (6)
	±0.5741	±0.29	±0.1426	±0.5487	±1.493	±0.966
Ecoli	83.93 (2)	**85.9** (1)	83.63 (3)	81.13 (5)	82.02 (4)	80.18 (6)
	±1.048	±0.7192	±0.6787	±0.7669	±0.7669	±0.9336
Glass	**69.07** (3)	48.04 (6)	57.94 (5)	65.89 (4)	**69.16** (2)	**70.37** (1)
	±1.855	±1.733	±0.9802	±0.6608	±2.999	±2.890
Heart-c	79.54 (4)	82.64 (2)	**83.43** (1)	80.26 (3)	76.04 (6)	78.35 (5)
	±1.287	±0.3960	±0.5678	±1.343	±2.225	±0.7976
Hypothyroid	98.94 (4)	95.36 (5)	93.57 (6)	99.38 (2)	**99.55** (1)	99.33 (3)
	±0.05141	±0.04436	±0.03968	±0.06138	±0.05661	±0.04242
Page-blocks	96.63 (4)	90.12 (6)	92.86 (5)	**97.04** (1)	96.85 (2)	96.66 (3)
	±0.09216	±0.1146	±0.06638	±0.07525	±0.1582	±0.1860
Pima	74.17 (5)	75.73 (2)	**76.88** (1)	74.97 (4)	75 (3)	72.42 (6)
	±0.6619	±0.55	±0.2116	±0.7699	±1.633	±1.357
Sonar	75.29 (4)	68.17 (6)	**78.37** (1)	75.87 (3)	72.6 (5)	76.83 (2)
	±0.6521	±0.1923	±1.138	±1.193	±2.803	±1.757
Soybean	92.06 (4)	92.12 (3)	**92.91** (1)	91.86 (5)	**92.83** (2)	91.54 (6)
	±0.4079	±0.1707	±0.3773	±0.4304	±0.5857	±0.5106
Spambase	93.44 (2)	79.63 (6)	90.4 (5)	92.63 (3.5)	92.63 (3.5)	**93.82** (1)
	±0.2583	±0.08908	±0.05252	±0.1457	±0.3583	±0.3538
Vehicle	**75.2** (1)	45.15 (6)	74.09 (2)	68.65 (5)	73.03 (3)	71.4 (4)
	±0.4693	±0.4547	±0.6092	±1.578	±1.682	±1.028
Vowel	**82.5** (1)	63.11 (6)	71.35 (4)	70.36 (5)	79.8 (2)	76.77 (3)
	±0.7182	±0.4444	±0.5481	±1.130	±1.026	±0.667
Wine	97.2 (3)	97.42 (2)	**98.76** (1)	93.48 (4.5)	92.92 (6)	93.48 (4.5)
	±0.7106	±0.762	±0.2247	±0.5729	±0.8408	±1.491
Avg. rank	3.15	4.10	2.9	3.55	3.38	3.92

for each pair of classifiers. We applied Wilcoxon test (Kononenko and Kukar, 2007) with null-hypothesis that the medians of results on all data sets of two compared classifiers are equal. Let us remark that in the paired test, ranks are assigned to values of differences in average classification accuracy between two compared classifiers. We observed significant difference (*p*-values smaller than 0.05) between Naive Bayes and any other classifier, and between VC-DomLEM and MODLEM.

Although statistical significance could not be confirmed, it follows from the results of the experiment that VC-DomLEM is at least comparable to the other classifiers. When we consider the value of the average rank observed in our experiments, VC-DomLEM is better than any other classifier, except SVM. However, according to the results of Friedman test the observed differences in ranks, calculated between all the classifiers, are not significant. On the other hand, the results of Wilcoxon test show that any classifier is performing better than Naive Bayes and that VC-DomLEM is performing better than MODLEM.

Another issue is the matter of comprehensibility of discovered classification patterns (laws). Remark that monotonic decision rules induced using our approach have the interesting property of showing pros and cons for assignment of an objet to each of the considered classes. This is because for each class, there are two kinds of rules: rules suggesting assignment to this class, and rules making an opposite suggestion (assignment to a complement of this class). These positive and negative arguments for the assignment give a better insight into the classification decision for domain experts. On the other hand, the number of induced rules may be too high to be read together. This may be avoided, however, since only a small subset of rules is usually covering a classified object and thus its analysis does not need a big cognitive effort.

Finally, as it was already mentioned in Section 3, monotonic rules induced using our approach can cover multiple values of an attribute. For example, let us consider the set of the decision rules induced for the breast cancer data set from the University Medical Center, Institute of Oncology, Ljubljana (which is labeled “Breast-cancer” in Tables 6 and 7). For these data, the goal is to classify a patient into one of two classes “no recurrence events” or “recurrence events”. In the set of rules induced by VC-DomLEM from the transformed classification table, one can find the following rule:

if tumor size $\notin [45, 49]$

and malignant degree $= 1$

and breast quadrant $\in$ {central,right-low, right-up},

then no recurrence events.

This rule is showing a set of possible conditions for no recurrence events. It is like a scenario or a law for no occurrence events. It says that if the tumor size attribute takes a value from outside the interval [45, 49], and the malignity degree is equal to 1, and the breast quadrant is central, right-low or right-up, then there is no recurrence event. The condition on the tumor size attribute shows two regions of local monotonicity: below 45 (negative) and above 49 (positive). Such a rule could not be induced by VC-DomLEM from the original breast-cancer data set. The two regions of local monotonicity were discovered by this algorithm due to the transformation proposed in this paper.

7. Conclusions

In this paper, we proposed a new approach to induction of laws from data in which we make use of the concept of monotonic relationships between values of condition and decision attributes, without assuming its direction a priori and allowing local monotonicity relationships in subregions of the evaluation space. Indeed, our method is able to discover local and global monotonicity relationships existing in data. The relationships are represented by monotonic decision rules which can be read as laws describing the analyzed phenomena. To enable the discovery of monotonic rules, we propose a non-invasive transformation of the input data, and a way of structuring them into consistent and inconsistent parts using the dominance-based rough set approach (DRSA) or its extension called VC-DRSA. The rule induction algorithm operates on this structure. The monotonic decision rules thus induced put in evidence easily understandable relationships between values of condition and decision attributes that cannot be discovered by traditional data mining methodologies.

The distinctive features of the proposed approach which count for its advantage over other existing methods include:

- it accepts numerical, nominal and binary condition attributes, and does not need any discretization of numerical attributes, which is always arbitrary to some extend,
- it can be used for ordinal and non-ordinal decision attributes,
- it accepts ordinal and non-ordinal condition attributes, and does not need to transform the ordinal scales into cardinal ones, which would claim to say more than the data,
- it discovers regions of local monotonicity relationship between values of condition and decision attributes, i.e. decision rules involve “interval” elementary conditions: “attribute $a_i \in [r_i^1, r_i^2]$”,
- it is able to discover rules with elementary conditions on nominal attributes of the type: “attribute $a_i \in \{v_i^1, \ldots, v_i^k\}$”,
- due to the capacity of discovering local monotonicity and elementary conditions concerning subsets of nominal attribute values, the monotonic rules are compact and relatively stronger than traditional rules,
- it can induce rules providing arguments pros and cons a given decision.
- the monotonic rules, together with a specially proposed classification scheme, have at least as good predictive ability as other well known predictors, while they are much more comprehensible than any other forms of relationships between condition and decision attributes.

Thus, one can conclude that the presented approach provides a very general framework for inducing laws from heterogeneous data. Our future research will be focused on handling missing values of condition attributes, and on generation of the most attractive monotonic rules from the point of view of some pertinent interestingness measures.

Acknowledgments

The first and the third authors wish to acknowledge financial support from the Polish Ministry of Science and Higher Education, Grant no. N519 314435.

References

Błaszczyński, J., Słowiński, R., 2003. Incremaental induction of satisfactory decision rules from dominance based rough approximations. In: Skowron, A., Szczuka, M. (Eds.), Proceedings of the International Workshop on Rough Sets in Knowledge Discovery and Soft Computing, 2003, pp. 40–51.

Błaszczyński, J., Greco, S., Słowiński, R., 2007. Multi-criteria classification—a new scheme for application of dominance-based decision rules. Eur. J. Oper. Res. 181 (3), 1030–1044.

Błaszczyński, J., Greco, S., Słowiński, R., Szeląg, M., 2009. Monotonic variable consistency rough set approaches. Int. J. Approx. Reason. 50 (7), 979–999.

Błaszczyński, J., Słowiński, R., Szeląg, M., 2011. Sequential covering rule induction algorithm for variable consistency rough set approaches. Inf. Sci. 181 (5), 987–1002.

Cohen, W.W., 1995. Fast effective rule induction. in: Proceedings of the Twelfth International Conference on Machine LearningMorgan Kaufmann, pp. 115–123.

Demsar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.

Greco, S., Matarazzo, B., Słowiński, R., 1998. A new rough set approach to evaluation of bankruptcy risk. in: Zopounidis, C. (Ed.), Operational Tools in the Management of Financial RisksKluwer Academic Publishers, Dordrecht, pp. 121–136.

Greco, S., Matarazzo, B., Słowiński, R., 1999. The use of rough sets and fuzzy sets in MCDM. in: Gal, T., Stewart, T., Hanne, T. (Eds.), Advances in Multiple Criteria Decision MakingKluwer Academic Publishers, Boston, pp. 14.1–14.59 (Chapter 14).

Greco, S., Matarazzo, B., Słowiński, R., 2001. Rough sets theory for multicriteria decision analysis. Eur. J. Oper. Res. 129, 1–47.

Greco, S., Matarazzo, B., Słowiński, R., 2002. Rough sets methodology for sorting problems in presence of multiple attributes and criteria. Eur. J. Oper. Res. 138, 247–259.

Greco, S., Matarazzo, B., Słowiński, R., 2005. Decision rule approach. in: Figueira, J., Greco, S., Ehrgott, M. (Eds.), Multiple Criteria Decision Analysis: State of the Art SurveysSpringer-Verlag, Berlin, pp. 507–563 (Chapter 13).

Greco, S., Matarazzo, B., Słowiński, R., 2007. Dominance-based rough set approach as a proper way of handling graduality in rough set theory. Trans. Rough Sets 7, 36–52.

Greco, S., Matarazzo, B., Słowiński, R., Stefanowski, J., 2000a. An algorithm for induction of decision rules consistent with the dominance principle. In: Ziarko, W., Yao, Y.Y. (Eds.), Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science, vol. 2005. Springer, pp. 304–313.

Greco, S., Matarazzo, B., Słowiński, R., Stefanowski, J., 2000b. Variable consistency model of dominance-based rough sets approach. In: Ziarko, W., Yao, Y.Y. (Eds.), Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence, vol. 2005. Springer-Verlag, Berlin, pp. 170–181.

Kononenko, I., Kukar, M., 2007. Machine Learning and Data Mining. Horwood Publishing, Coll House, Westergate, Chichester, West Sussex.

Pawlak, Z., Skowron, A., 1994. Rough membership functions. in: Yager, R.R., Fedrizzi, M., Kacprzyk, J. (Eds.), Advances in the Dempster–Shafer Theory of EvidenceWiley, New York, pp. 251–271.

Platt, J., 1998. Machines using Sequential Minimal Optimization. In: Schoelkopf, B., Burges, C., Smola, A. (Eds.) Advances in Kernel Methods—Support Vector Learning.

Quinlan, J.R., 1992. C.45: Programs for Machine Learning. Morgan Kaufmann.

Słowiński, R., Greco, S., Matarazzo, B., 2005. Rough set based decision support. in: Burke, E.K., Kendall, G. (Eds.), Search Methodologies: Introductory Tutorials in Optimization and Decision Support TechniquesSpringer-Verlag, New York, pp. 475–527 (Chapter 16).

Słowiński, R., Greco, S., Matarazzo, B., 2009. Rough Sets in Decision Making. in: Meyers, R.A. (Ed.), Encyclopedia of Complexity and Systems ScienceSpringer, New York, pp. 7753–7786.

Stefanowski, J., 1998. The rough set based rule induction technique for classification problems. In: Proceedings of the 6th European Conference on Intelligent Techniques and Soft Computing EUFIT-98, pp. 109–113.

Wong, S.K.M., Ziarko, W., 1987. Comparison of the probabilistic approximate classification and the fuzzy set model. Fuzzy Sets Syst. 21, 357–362.