Fundamenta Informaticae 53 (2002) 365–390
IOS Press

Approximate Entropy Reducts

Dominik Ślęzak*
Department of Computer Science, University of Regina
Regina, SK, S4S 0A2, Canada
and
Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland

Abstract. We use information entropy measure to extend the rough set based notion of a reduct. We introduce the Approximate Entropy Reduction Principle (AERP). It states that any simplification (reduction of attributes) in the decision model, which approximately preserves its conditional entropy (the measure of inconsistency of defining decision by conditional attributes) should be performed to decrease its prior entropy (the measure of the model's complexity). We show NP-hardness of optimization tasks concerning application of various modifications of AERP to data analysis.

1. Introduction

The data based reasoning relates to the analysis of information within the acquired samples of objects. The theory of *rough sets* [8, 12, 13] assumes that a universe of known objects is the only source of knowledge usable to construct the reasoning models, stated by means of functional and logical, possibly uncertain, dependencies between attributes. A special case is concerned with classification problems, where the goal is to approximate values of a distinguished *decision attribute* under information provided by *conditional attributes*. For this purpose, one stores data within *decision systems*, where each object belongs to one of the predefined *decision classes*.

Classification of new objects can be easily performed by analogy, by application of *"if..then.."* *rules* previously calculated over the universe of a given system. According to the *Minimum Description Length Principle (MDLP)*, introduced in [20] and adapted to the theory of rough sets in [4, 23], the optimization problems concerned with the search of the simplest *patterns* and *decision rules*, as well as *information* and *decision reducts* are crucial. Although proved to be NP-hard [24], solutions of those problems can be

*Address for correspondence: Banacha 14/33, 02-097 Warsaw, Poland

approximated sufficiently enough by using various heuristics [3, 17, 34]. Their computational efficiency, as well as degrees of correct classification for the obtained decision models, are important advantages, which should be taken into account while developing various extensions of the "classical" rough set based approach.

We consider one of such extensions, especially dedicated to the analysis of strongly inconsistent data, which occur very often in, e.g., medical domain [17, 18]. We are interested in tools enabling reconsideration of rules and reducts for systems, which support no "if..then.." dependencies in an exact form. We focus on *approximate reducts* – irreducible subsets of attributes inducing collections of rules, which approximate decision classes accurately enough, or, in other words, approximate them almost at the same level of precision as provided by the initial decision model, before starting the reduction process.

We propose to label each subset $B \subseteq A$ of available attributes A with its *information entropy* $H(B)$ calculated directly from data, basing on data partition generated by the values of elements of B (cf. [5, 7, 21]). We show how to interpret $H(B)$ as a predictive complexity of the model induced by B. In case of classification problem, we label B also with *conditional entropy* $H(d/B)$, which is the degree of the model's inaccuracy in predicting the values of decision attribute d (cf. [4, 25]).

Entropy provides decision models induced by subsets of attributes with mutually opposing measures leading to one of the most known versions of MDLP, where the entropy based complexity and inaccuracy measures are combined within the one global optimization measure [4, 10, 20]. In this paper, we concentrate on the *Approximate Entropy Reduction Principle (AERP)*, where we minimize complexity $H(B)$ under the constraint of keeping inaccuracy $H(d/B)$ at the reasonably low level with respect to $H(d/A)$. This is a special case of the *Approximate Reduction Principle* developed in [30, 32], which is a generalization of the rough set reduction laws [8, 13, 23, 24].

We show NP-hardness of optimization problems concerned with the application of AERP to extraction of data models from decision systems and information systems. Although strongly related to the previous studies on time complexity of approximate reduction [28, 29, 30], this is the first paper devoted in an exhaustive way to interpretation and extension of the fundamental rough set reduction principles in terms of the measures of information entropy. It contains novel mathematical results related, e.g., to the modified version of AERP, for $H(B/d)$ instead of $H(B)$. It reflects the tradeoff between *precision* and *sensitivity* of the rule based decision models, widely used in machine learning [9, 11], providing the basis for the *Relative Operating Characteristic (ROC)* approach [19].

The paper is organized as follows: In Section 2 we recall the basic concepts concerned with prior and conditional probabilities derived from data. We pay a special attention to probabilities corresponding to the object related "if..then.." rules, where objects are understood as the known cases, gathered within available information and decision systems. In Subsection 2.3 we interpret and compare conditional probabilities with the tools of the theory of rough sets, developed in purpose of modeling inexact dependencies between attributes. We analyze various – probabilistic and rough set based – ways of expressing the notion of *defining decision* by conditional attributes.

In Section 3 we recall basic notions and results concerned with the measures of information and conditional entropy. We discuss how to use entropy to evaluate subsets of attributes, as well as inexact decision rules extracted from data. We show how to interpret entropy in terms of the rough set based notions of *rough membership functions* and *rough membership distributions*. In particular, we show that the entropy measures can be obtained as geometric averages of the basic machine learning factors of quality of "if..then.." rules.

In Section 4 we recall the concepts of information and decision reducts. We show the correspondence between *μ-decision reducts*, which preserve the probabilistic information about conditions→decision dependencies during the reduction process, and basic notions of *probabilistic conditional independence*. We express the conditions for information and decision reducts in terms of entropy. We also discuss other measures enabling to model the principle of keeping the level of *(in)consistency* of defining decision during the reduction process. In Subsection 4.4 we use entropy to introduce the concept of *approximate preserving of information*, which leads to specification of the notions of *(H, ε)-approximate information reducts* and *(H, ε)-approximate μ-decision reducts*.

In Section 5 we show the NP-hardness of various approximate reduction tasks. In Subsection 5.1 we formulate basic optimization problems and results, which lead to specification of two versions of AERP, in Definitions 5.3 and 5.4. In Subsection 5.2 we prove NP-hardness of the task of finding minimal (H, ε)-approximate information and decision reducts. Finally, in Subsection 5.3, we do the same with reducts optimized with respect to the entropy based complexity and sensitivity measures.

2. Probabilities in data

2.1. Prior probabilities in information systems

In the theory of *rough sets* [8, 12, 13], one represents data in terms of an *information system* $\mathbb{A} = (U, A)$, where U denotes the *universe* of *objects* and each *attribute* $a \in A$ is identified with function $a : U \to V_a$, for V_a denoting the set of values on a. Let us enumerate elements of A as $a_1, \ldots, a_{|A|}$. For any $B \subseteq A$ and $u \in U$, we define *B-information vector*

$$B(u) = \langle a_{i_1}(u), \ldots, a_{i_{|B|}}(u) \rangle \tag{1}$$

where $i_1 < i_2 < \cdots < i_{|B|}$ and for any $j = 1, \ldots, |B|$ we denote by $a_{i_j}(u)$ the value of attribute $a_{i_j} \in B$. The set of all B-information vectors, which occur in $\mathbb{A}$, takes the following form:

$$V_B^U = \{B(u) : u \in U\} \tag{2}$$

Each subset $B \subseteq A$ induces a partition over the universe U. Partition classes correspond to particular B-information vectors. We obtain partition space

$$U/B = \{E_w : w \in V_B^U\} \tag{3}$$

where each class $E_w \subseteq U$ is defined by

$$E_w = \{u \in U : B(u) = w\} \tag{4}$$

The described partition is referred to as *indiscernibility relation* $IND_{\mathbb{A}}(B)$, where subsets $E_w \subseteq U$ are called *indiscernibility classes* of objects, which cannot be distinguished from each other by the values on B [8, 12, 13].

One can easily label partition classes with their cardinalities, measured relatively with respect to the universe. It enables to define the prior probability distribution over B-information vectors. The data derived probability of occurrence of $w \in V_B^U$ on B is simply

$$P_{\mathbb{A}}(w) = |\{u \in U : B(u) = w\}| \,/\, |U| \tag{5}$$

One can define probabilities for all possible combinations of values on elements of B, i.e.

$$V_B = \prod_{a \in B} V_a \tag{6}$$

Obviously, we have

$$V_B^U = \{w \in V_B : P_{\mathbb{A}}(w) > 0\} \tag{7}$$

Pairs (B, w), for $B \subseteq A$ and $w \in V_B^U$, are often referred to as *patterns* or *templates* (cf. [3, 8]), interpreted as conjunctions of *descriptors* (a, v), for $a \in B$ and $v \in V_a$. In this context, quantity $P_{\mathbb{A}}(w)$ reflects the *strength* of (B, w), i.e. the chance that a randomly chosen object $u \in U$ will support the considered pattern.

Usually, one implicitly assumes that $P_{\mathbb{A}}(w)$ estimates also the probability of satisfying (B, w) by objects outside the currently known universe. Such an assumption occurs also in case of the data driven conditional probabilities considered in the next subsection, unless one has a reason to suspect that the training data set is not representative enough, in a statistical sense.

2.2. Conditional probabilities in information and decision systems

Conditional probabilities are usually derived in purpose of expressing a chance of occurrence of a given pattern under information about occurrence of another one. In the basic form, it leads to the analysis of *association rules* [1], where both the left and right sides consist of conditions involving disjoint subsets of attributes $B, C \subseteq A$ and their values $w \in V_B^U$, $w' \in V_C^U$. The *precision* of the rule $B = w \Rightarrow C = w'$ is defined as

$$P_{\mathbb{A}}(w'/w) = |\{u \in U : B(u) = w \wedge C(u) = w'\}| \,/\, |\{u \in U : B(u) = w\}| \tag{8}$$

It turns out that the rules with positive precision correspond to *object related association rules* $B = B(u) \Rightarrow C = C(u)$ induced by elements of U. In this case, (8) takes the form

$$P_{\mathbb{A}}(C(u)/B(u)) = |\{u' \in U : (B \cup C)(u) = (B \cup C)(u')\}| \,/\, |\{u' \in U : B(u) = B(u')\}| \tag{9}$$

Prior probability $P_{\mathbb{A}}(B(u)) = |\{u' \in U : B(u) = B(u')\}|/|U|$ reflects the chance that an object $u' \in U$ will be *recognized*, i.e., it will satisfy the left side of the considered rule.

In the object related case, we do not need to bother with intersections between attribute sets occurring at both sides of the rule. For instance, we can express how $B \subseteq A$ determines the rest of attributes $A \setminus B$ by considering the bunch of probabilities

$$P_{\mathbb{A}}(A(u)/B(u)) = |\{u' \in U : A(u) = A(u')\}| \,/\, |\{u' \in U : B(u) = B(u')\}| \tag{10}$$

for particular objects $u \in U$. We will use such probabilities in purpose of defining the global measures of dependencies between attributes.

The task of analysis can be also concerned with a distinguished *decision* to be predicted under information provided over the rest of attributes. In this case, we represent data as a *decision system* $\mathbb{A} = (U, A \cup \{d\})$, $d \notin A$. Let $V_d = \{1, \ldots, |V_d|\}$. For each $k = 1, \ldots, |V_d|$, we define the k-th decision class $X_k = \{u \in U : d(u) = k\}$. Probability

$$P_{\mathbb{A}}(k/w) = |\{u \in X_k : B(u) = w\}| \,/\, |\{u \in U : B(u) = w\}| \tag{11}$$

of $k \in V_d$ conditioned by $w \in V_B^U$ corresponds to the precision of *decision rule* $B = w \Rightarrow d = k$. The strength of the rule is provided by $P_{\mathbb{A}}(w)$. One can also consider quantity

$$P_{\mathbb{A}}(w/k) = |\{u \in X_k : B(u) = w\}| \,/\, |X_k| \tag{12}$$

which is referred to as the *sensitivity* of the rule. Keeping the balance between (11) and (12) is the idea of the *Relative Operating Characteristic (ROC)* considered in, e.g., [19, 35].

Just like before, we can restrict ourselves to *object related decision rules*. Given $B \subseteq A$ and $u \in U$, one defines the precision and sensitivity of $B = B(u) \Rightarrow d = d(u)$ by, respectively:

$$P_{\mathbb{A}}(d(u)/B(u)) = |\{u' \in X_{d(u)} : B(u) = B(u')\}| \,/\, |\{u' \in U : B(u) = B(u')\}| \tag{13}$$

and

$$P_{\mathbb{A}}(B(u)/d(u)) = |\{u' \in X_{d(u)} : B(u) = B(u')\}| \,/\, |X_{d(u)}| \tag{14}$$

These quantities can be used while optimizing the choice of the subset of conditional attributes for construction of the decision model. One can also replace sensitivity $P_{\mathbb{A}}(B(u)/d(u))$ with the rule's strength $P_{\mathbb{A}}(B(u))$ or complexity of its left side's description (e.g. the number of descriptors). This is one of possible interpretations of the *Minimum Description Length Principle* [9, 20] within the theory of rough sets, where the model induced by $B \subseteq A$ is evaluated by the combination of two factors: its quality and the length of its description (cf. [23, 24]).

2.3. Rough membership functions and distributions

Probabilities provide the most popular, but not the only one way of handling inexact dependencies in data. For example, one can derive from a given $\mathbb{A} = (U, A)$ the *rough set approximations*, which express the degree of defining each $X \subseteq U$ by means of the partition space U/B, $B \subseteq A$:

$$\underline{B}_{\mathbb{A}}^X = \{E \in U/B : E \subseteq X\} \qquad \overline{B}_{\mathbb{A}}^X = \{E \in U/B : E \cap X \neq \emptyset\} \tag{15}$$

The above kind of information is less detailed than the probabilistic one. Probabilities can be expressed in the theory of rough sets by means of *rough membership function* $\mu_X^B : U \rightarrow [0, 1]$ introduced in [14], where

$$\mu_X^B(u) = |[u]_B \cap X| \,/\, |[u]_B| \tag{16}$$

for $[u]_B = \{u' \in U : B(u) = B(u')\}$ denoting the B-indiscernibility class of u. Rough membership function enables to rewrite (15) equivalently as

$$\underline{B}_{\mathbb{A}}^X = \left\{u \in U : \mu_X^B(u) = 1\right\} \qquad \overline{B}_{\mathbb{A}}^X = \left\{u \in U : \mu_X^B(u) > 0\right\} \tag{17}$$

We say that B *defines* X, iff $\underline{B}_{\mathbb{A}}^X = \overline{B}_{\mathbb{A}}^X = X$, or, equivalently, there is $\mu_X^B(u) = 1 \Leftrightarrow u \in X$ and $\mu_X^B(u) = 0 \Leftrightarrow u \notin X$. Still, handling values of μ_X^B enables to consider degrees of membership between 0 and 1. For instance, it is done in the *Variable Precision Rough Set Model* [36], where conditions concerned with $\mu_X^B(u)$ in (17) are appropriately weakened.

In case of a decision system $\mathbb{A} = (U, A \cup \{d\})$, we are interested in approximating decision classes. We say that $B \subseteq A$ *defines* d, iff it defines each decision class $X_k \subseteq U$, $k = 1, \ldots, |V_d|$. Equivalently, B defines d, iff, e.g., there is:

- $IND_{\mathbb{A}}(B) \subseteq IND_{\mathbb{A}}(\{d\})$, i.e. for any $u, u' \in U$, if $d(u) \neq d(u')$, then $B(u) \neq B(u')$.
- *The B-positive region* $POS_{\mathbb{A}}(d/B) = \bigcup_k \underline{B}_{\mathbb{A}}^{X_k}$ ([13]) equals to the whole universe U.
- There is equality $P_{\mathbb{A}}(d(u)/B(u)) = 1$, for any element $u \in U$.

Obviously, for any $B \subseteq A$, $u \in U$ and $k = 1, \ldots, |V_d|$, the following equality holds:

$$\mu_{X_k}^B(u) = P_{\mathbb{A}}(k/B(u)) \tag{18}$$

Thus, in case of considering decision classes, we have an equivalence between conditional probabilities and rough membership functions. The difference is that probability, as a function, operates on vectors of values. On the other hand, rough membership function operates on the set of instances understood as objects in a decision system.

We can continue the above comparison at the level of probabilistic distributions. By the *rough membership distribution* we mean the function labeling each $u \in U$ with vector

$$\overrightarrow{\mu}_d^B(u) = \left\langle \mu_{X_1}^B(u), \ldots, \mu_{X_{|V_d|}}^B(u) \right\rangle \tag{19}$$

which is an element of the ($|V_d|$-1)-dimensional simplex of probabilistic distributions. Obviously, B defines d, iff vector $\overrightarrow{\mu}_d^B(u)$ takes the form of one of the simplex vertices, for any $u \in U$.

One can see that there are many possibilities of expressing the notion of defining decision. All of them can be used as the reference points while stating the criteria for *approximate defining*. We can claim that $B \subseteq A$ *approximately defines* d, iff $IND_{\mathbb{A}}(B)$ is *almost* contained in $IND_{\mathbb{A}}(\{d\})$, or iff $POS_{\mathbb{A}}(d/B)$ is *almost* equal to U, or iff for *almost* all $u \in U$ probability $P_{\mathbb{A}}(d(u)/B(u))$ is *sufficiently close* to 1, or vector $\overrightarrow{\mu}_d^B(u)$ is *sufficiently close* to a vertex of the simplex. In this paper we focus on the approaches based mainly on probabilistic tools. We refer the reader to [16, 17, 27, 28, 30] for other ways of introducing the notion of approximate defining.

3. Information entropy

3.1. Basic properties of prior entropy

As a measure of information, *entropy* was first considered in [21]. It was used in purpose of evaluating the degree of information we get from the fact that a given random event occurred. Entropy is related directly to the event's probability $p > 0$. It is defined by $h(p) = -\log_2 p$. For instance, given $B \subseteq A$ and $w \in V_B^U$ for some $\mathbb{A} = (U, A)$, one can state the entropy of pattern (B, w) as equal to $-\log_2 P_{\mathbb{A}}(w)$. In its generalized form [5, 7], entropy is used to evaluate random distributions $\overrightarrow{p} = \langle p_1, \ldots, p_r \rangle$ with the expected degree of information

$$H(\overrightarrow{p}) = -\sum_{k:\, p_k > 0} p_k \log_2 p_k \tag{20}$$

For instance, by operating with probabilistic distributions considered in Subsection 2.1, we can label each subset $B \subseteq A$ with its entropy

$$H_{\mathbb{A}}(B) = -\sum_{w \in V_B^U} P_{\mathbb{A}}(w) \log_2 P_{\mathbb{A}}(w) \tag{21}$$

interpreted as the average degree of information about elements of the universe, which can be obtained from knowledge concerning their values on B. As another example, we can consider the entropy of rough membership distribution $\overrightarrow{\mu}_d^B(u)$ given by (19), for some $B \subseteq A$ and $u \in U$, in a decision system $\mathbb{A} = (U, A \cup \{d\})$. It equals to

$$H\left(\overrightarrow{\mu}_d^B(u)\right) = -\sum_{k \in \partial_d^B(u)} \mu_{X_k}^B(u) \log_2 \mu_{X_k}^B(u) \tag{22}$$

where $\partial_d^B(u)$ is the value of the *generalized decision function*, which is given by formula (47) in Subsection 4.2. Quantity of $H\left(\overrightarrow{\mu}_d^B(u)\right)$ can be used to express the degree of the lack of precise information about decision value of $u \in U$, given knowledge about its values on $B \subseteq A$. Due to the basic properties of entropy [5, 7], we have, for any $\overrightarrow{p} = \langle p_1, \ldots, p_r \rangle$, inequalities

$$0 \leq H\left(\overrightarrow{p}\right) \leq \log_2 |\{k : p_k > 0\}| \tag{23}$$

where $H\left(\overrightarrow{p}\right) = 0$ iff $|\{k : p_k > 0\}| = 1$, and $H\left(\overrightarrow{p}\right) = \log_2 |\{k : p_k > 0\}|$ iff any positive coordinate of $\overrightarrow{p}$ equals to $1/|\{k : p_k > 0\}|$. In case of (22), it means that

$$0 \leq H\left(\overrightarrow{\mu}_d^B(u)\right) \leq \log_2 \left|\partial_d^B(u)\right| \tag{24}$$

where $H\left(\overrightarrow{\mu}_d^B(u)\right) = 0$ iff $P_{\mathbb{A}}(d(u)/B(u)) = 1$, and $H_d^B(u) = \log_2 \left|\partial_d^B(u)\right|$ iff for all elements $k \in \partial_d^B(u)$ we have the same values of rough membership function, i.e. $\mu_{X_k}^B(u) = 1/\left|\partial_d^B(u)\right|$. Hence, we see that $H\left(\overrightarrow{\mu}_d^B(u)\right)$ is minimal, iff B determines the decision value $d(u)$. $H\left(\overrightarrow{\mu}_d^B(u)\right)$ is maximal, iff B provides no knowledge about $d(u)$, i.e. $\partial_d^B(u) = V_d$ and all decision values $k = 1, \ldots, |V_d|$ are equally probable within the B-indiscernibility class $[u]_B \subseteq U$. In the same way, for any $\mathbb{A} = (U, A)$ and $B \subseteq A$, we get inequalities

$$0 \leq H_{\mathbb{A}}(B) \leq \log_2 \left|V_B^U\right| \tag{25}$$

where $H_{\mathbb{A}}(B) = 0$, iff we have $\left|V_B^U\right| = 1$, and $H_{\mathbb{A}}(B) = \log_2 \left|V_B^U\right|$, iff for each $w \in V_B^U$ we have $P_{\mathbb{A}}(w) = 1/\left|V_B^U\right|$.

3.2. Basic properties of conditional entropy

Conditional entropy of random variable C conditioned by B evaluates the degree of information we would still obtain from the knowledge about the value of C, under already provided knowledge about the value of B (cf. [4, 5, 7]). If we interpret B and C as subsets of A, within an information system $\mathbb{A} = (U, A)$, the entropy of C conditioned by B takes the form of

$$H_{\mathbb{A}}(C/B) = H_{\mathbb{A}}(B \cup C) - H_{\mathbb{A}}(B) \tag{26}$$

and reflects the degree of information about a randomly chosen object $u \in U$, which we would additionally gain from the knowledge about $C(u)$, if we already know $B(u)$.

Properties of conditional entropy correspond to the notion of *probabilistic conditional independence* [15]. Let us formulate it in terms of $\mathbb{A} = (U, A)$, where $P_{\mathbb{A}}$ is treated as the probability distribution over the product of discrete random variables A. Given mutually disjoint subsets $B, C, D \subseteq A$, we say that

C makes D independent from B, iff for all possible configurations of vectors $w_B \in V_B$, $w_C \in V_C$ and $w_D \in V_D$ we have implication

$$P_{\mathbb{A}}(w_B, w_C) > 0 \Rightarrow [\, P_{\mathbb{A}}(w_D/w_B, w_C) = P_{\mathbb{A}}(w_D/w_C)\,] \tag{27}$$

or, equivalently ([30, 31]), iff for all objects $u \in U$ we have equality

$$P_{\mathbb{A}}(D(u)/C(u)) = P_{\mathbb{A}}(D(u)/(B \cup C)(u)) \tag{28}$$

In particular, D is independent from B, iff for any $u \in U$ we have $P_{\mathbb{A}}(D(u)) = P_{\mathbb{A}}(D(u)/B(u))$.

According to [5], we have the following inequality:

$$H_{\mathbb{A}}(D/B \cup C) \leq H_{\mathbb{A}}(D/C) \tag{29}$$

where equality holds, iff C makes D independent from B. In particular, for $C = \emptyset$, we have $H_{\mathbb{A}}(D/B) \leq H_{\mathbb{A}}(D)$, where equality holds, iff D is independent from B.

Given a decision system $\mathbb{A} = (U, A \cup \{d\})$, we can use $H_{\mathbb{A}}(d/B)$ [1] to label each $B \subseteq A$ with the amount of uncertainty concerning d under the information about B provided. Several modifications of conditional entropy were proposed within the theory of rough sets (cf. [4, 25]). In this paper, we refer to the classical meaning of $H_{\mathbb{A}}(d/B)$. In such a case, inequality (29) leads to the following constrains:

Proposition 3.1. (cf. [5]) Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We have inequalities

$$0 \leq H_{\mathbb{A}}(d/A) \leq H_{\mathbb{A}}(d/B) \leq H_{\mathbb{A}}(d) \leq \log_2 |V_d| \tag{30}$$

where:

- $H_{\mathbb{A}}(d/A) = 0$ iff A defines d.
- $H_{\mathbb{A}}(d/A) = H_{\mathbb{A}}(d/B)$ iff B makes d independent from $A \setminus B$.
- $H_{\mathbb{A}}(d/B) = H_{\mathbb{A}}(d)$ iff d is independent from B.
- $H_{\mathbb{A}}(d) = \log_2 |V_d|$ iff for any $k \in V_d$ we have $P_{\mathbb{A}}(k) = 1/|V_d|$.

The above result shows that conditional entropy is yet another measure encoding the state of defining decision by conditional attributes. In Section 4, we will apply it to define the notion of approximate defining in the probabilistic framework. Further, we will use it to introduce the notion of *approximate preserving of decision information* while reducing attributes.

3.3. Rough membership interpretation of entropy

Let $\mathbb{A} = (U, A)$ and disjoint $B, C \subseteq A$ be given. Any $w \in V_B^U$ induces the new probabilistic distribution over V_C^U, under the condition $B = w$. Its entropy equals to

$$H_{(B,w)}(C) = -\sum_{w' \in V_C^U :\, P_{\mathbb{A}}(w'/w) > 0} P_{\mathbb{A}}(w'/w) \log_2 P_{\mathbb{A}}(w'/w) \tag{31}$$

[1] To simplify notation, we will write $H_{\mathbb{A}}(d/\cdot)$, $H_{\mathbb{A}}(d)$, $H_{\mathbb{A}}(\cdot/d)$, instead of $H_{\mathbb{A}}(\{d\}/\cdot)$, $H_{\mathbb{A}}(\{d\})$, $H_{\mathbb{A}}(\cdot/\{d\})$.

One can equivalently rewrite $H_{\mathbb{A}}(C/B)$ as the average value of the above quantities, i.e.:

$$H_{\mathbb{A}}(C/B) = \sum_{w \in V_B^U} P_{\mathbb{A}}(w) H_{(B,w)}(C) \tag{32}$$

For $\mathbb{A} = (U, A \cup \{d\})$, where we are interested in $C = \{d\}$, we can express $H_{\mathbb{A}}(d/B)$ in terms of local entropies of rough membership distributions $\overrightarrow{\mu}_d^B(u)$, for $u \in U$:

Proposition 3.2. ([30, 31]) Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We have equality

$$H(d/B) = \frac{1}{|U|} \sum_{u \in U} H\left(\overrightarrow{\mu}_d^B(u)\right) \tag{33}$$

where, for any $u \in U$, $H\left(\overrightarrow{\mu}_d^B(u)\right)$ is the entropy of distribution $\overrightarrow{\mu}_d^B(u)$, given by (22).

□

In Subsection 3.1 we considered $H\left(\overrightarrow{\mu}_d^B(u)\right)$ as expressing the degree of the lack of information about decision value of $u \in U$. Hence, equality (33) justifies the choice of $H_{\mathbb{A}}(d/B)$ as a measure of *inconsistency of defining d by B*.

Proposition 3.3. ([30, 31]) Let $\mathbb{A} = (U, A)$ and $B, C \subseteq A$ be given. We have

$$H_{\mathbb{A}}(B) = -\log_2(G_{\mathbb{A}}(B)) \qquad H_{\mathbb{A}}(C/B) = -\log_2(G_{\mathbb{A}}(C/B)) \tag{34}$$

where

$$G_{\mathbb{A}}(B) = \sqrt[|U|]{\Pi_{u \in U} P_{\mathbb{A}}(B(u))} \qquad G_{\mathbb{A}}(C/B) = \sqrt[|U|]{\Pi_{u \in U} P(C(u)/B(u))} \tag{35}$$

are, respectively, the geometric average of the strength of patterns $B = B(u)$, and the geometric average of the precision of association rules $B = B(u) \Rightarrow C = C(u)$, for $u \in U$.

As a special case, for $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$, we obtain equality

$$H_{\mathbb{A}}(d/B) = -\frac{1}{|U|} \sum_{u \in U} \log_2 P_{\mathbb{A}}(d(u)/B(u)) = -\log_2(G_{\mathbb{A}}(d/B)) \tag{36}$$

where

$$G_{\mathbb{A}}(d/B) = \sqrt[|U|]{\Pi_{u \in U} P(d(u)/B(u))} \tag{37}$$

reflects the degree of defining d by means of decision rules generated by B. It leads to the following way of expressing conditional entropy in terms of rough membership functions:

Proposition 3.4. ([30, 31]) Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We have equality

$$H_{\mathbb{A}}(d/B) = -\frac{1}{|U|} \sum_{u \in U} \log_2 \mu^B_{X_{d(u)}}(u) \tag{38}$$

As a conclusion, we obtain formulas (33) and (38) providing an interpretation of data based conditional entropy in terms of basic rough set notions. Moreover, we draw a correspondence between entropy and average qualities of object related "if..then.." rules, expressed by (35) and (37).

4. Reducts in terms of probabilities

4.1. Information and decision reducts

So far, we considered various measures and criteria enabling to evaluate data based models induced by subsets of attributes. Due to the *Ockham's Razor* principle, formalized in terms of the *Kolmogorov Complexity* [10], we should simplify such models, unless it causes a loss of information encoded in a model. It is a reference to the *Minimum Description Length Principle (MDLP)* introduced in [20] within the statistical framework, as well as a general principle of the theory of rough sets, concerned with reducing attributes and shortening decision rules [13, 23].

Definition 4.1. Let $\mathbb{A} = (U, A)$ and $B \subseteq A$ be given. We say that B *defines* $\mathbb{A}$, iff the following equality holds:

$$IND_{\mathbb{A}}(B) = IND_{\mathbb{A}}(A) \tag{39}$$

i.e. iff B defines each $a \in A$ (equivalently, $a \in A \setminus B$). We say that B is an *information reduct*, iff it defines $\mathbb{A}$ and none of its proper subsets does it.

Proposition 4.1. Let $\mathbb{A} = (U, A)$ and $B \subseteq A$ be given. B is an information reduct, iff we have

$$H_{\mathbb{A}}(B) = H_{\mathbb{A}}(A) \tag{40}$$

and inequalities $H_{\mathbb{A}}(B \setminus \{a\}) < H_{\mathbb{A}}(A)$ hold for any $a \in B$.

Obviously, entropy is not the only measure enabling to express the meaning of information reduct in a numeric way. For instance, one can consider the *discernibility measure*

$$Disc_{\mathbb{A}}(B) = |\{(u, u') \in U \times U : B(u) \neq B(u')\}| \tag{41}$$

which leads to the analogous characteristics:

Proposition 4.2. Let $\mathbb{A} = (U, A)$ and $B \subseteq A$ be given. B is an information reduct, iff we have

$$Disc_{\mathbb{A}}(B) = Disc_{\mathbb{A}}(A) \tag{42}$$

and inequalities $Disc_{\mathbb{A}}(B \setminus \{a\}) < Disc_{\mathbb{A}}(A)$ hold for any $a \in B$.

Just like we did it in case of the criteria for approximate defining of decision attribute in Subsection 2.3, we can generalize conditions (40) and (42) to be able to talk about *approximate defining of information system*. For instance, we can set up an *approximation threshold* $\varepsilon \in [0, 1)$ and say that B defines $\mathbb{A}$ *ε-approximately*, iff

$$Disc_{\mathbb{A}}(B) \geq (1 - \varepsilon) Disc_{\mathbb{A}}(A) \tag{43}$$

i.e. iff B discerns *ε-almost* pairs of objects, which can be discerned by the whole A. We will continue this idea in Subsection 4.4, by referring to the information entropy measures. For now, let us focus on decision systems $\mathbb{A} = (U, A \cup \{d\})$, which are *consistent*, i.e. such that A defines d. In such cases, introducing the notion of a decision reduct is very simple:

Definition 4.2. Let $\mathbb{A} = (U, A)$ and $B \subseteq A$ be given. We say that B is a *decision reduct*, iff it defines d and none of its proper subsets does it.

Obviously, one can redefine the above notion in terms of all previously considered criteria for defining decision. As an example, let us do it in terms of conditional entropy.

Proposition 4.3. Let a consistent $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. B is a decision reduct, iff we have $H_{\mathbb{A}}(d/B) = 0$ and inequalities $H_{\mathbb{A}}(d/B \setminus \{a\}) > 0$ hold for any $a \in B$.

4.2. Decision reducts in terms of conditional independence

In many practical situations, we have to consider models based on *inconsistent* decision systems $\mathbb{A} = (U, A \cup \{d\})$, where A does not define d, i.e. it does not provide exact decision rules covering the whole universe (cf. [17, 18]). Then, the question arises, what kind of a reduction criterion should be used instead of keeping (approximate) defining of decision. In [30, 31] we discussed the correspondence between decision reducts and the notions of the theory of probabilistic independence [15]. It is related to the idea of keeping decision information in terms of rough membership distributions while reducing conditional attributes.

Definition 4.3. Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We say that B *μ-preserves* d, iff for any $u \in U$ we have equality $\overrightarrow{\mu}_d^B(u) = \overrightarrow{\mu}_d^A(u)$ or, equivalently [30, 31]:

$$P_{\mathbb{A}}(d(u)/B(u)) = P_{\mathbb{A}}(d(u)/A(u)) \tag{44}$$

B is a *μ-decision reduct*, iff it μ-preserves d and none of its proper subsets does it.

Below we generalize Proposition 4.3. In consistent decision systems $\mathbb{A} = (U, A \cup \{d\})$ we have always $H_{\mathbb{A}}(d/A) = 0$. Then, μ-decision reducts simply coincide with decision reducts.

Proposition 4.4. Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. B is a μ-decision reduct, iff

$$H_{\mathbb{A}}(d/B) = H_{\mathbb{A}}(d/A) \tag{45}$$

and inequalities $H_{\mathbb{A}}(d/B) > H_{\mathbb{A}}(d/B \setminus \{a\})$ hold for any $a \in B$.

Due to (28), which states the sufficient and satisfactory condition for conditional independence between any mutually disjoint subsets of attributes, one can see that (44) is equivalent to stating that B makes d independent from $A \setminus B$. Hence, we can say that B is a μ-decision reduct, iff it is a *Markov boundary* of d with respect to the product distribution $P_{\mathbb{A}}$ over $A \cup \{d\}$, i.e. it is an irreducible subset of variables, which provides exactly the same probabilistic information about d as the whole A.

Obviously, probabilities do not provide the only possible framework for introducing the notion of conditional independence between attributes (cf. [22]). Given any such framework, we could redefine the notion of a decision reduct as an appropriate analogy of probabilistic Markov boundary. For instance, given $\mathbb{A} = (U, A)$ and mutually disjoint $B, C, D \subseteq A$, let us say that C makes D *O-independent* from B [26, 30], iff for all possible combinations of vectors $w_B \in V_B$, $w_C \in V_C$, $w_D \in V_D$ we have implication

$$P_{\mathbb{A}}(w_B, w_C) > 0 \Rightarrow [\, P_{\mathbb{A}}(w_D/w_C) > 0 \Leftrightarrow P_{\mathbb{A}}(w_D/w_B, w_C) > 0 \,] \tag{46}$$

It is weaker than its probabilistic analogy (27), because, under condition $P_{\mathbb{A}}(w_B, w_C) > 0$, we check only whether w_D is *plausible* for w_C iff it is *plausible* for (w_B, w_C), instead of demanding a perfect equality $P_{\mathbb{A}}(w_D/w_C) = P_{\mathbb{A}}(w_D/w_B, w_C)$. It corresponds to one of the most popular ways of dealing with the inconsistencies in the theory of rough sets, namely to the notion of the *generalized decision function* $\partial_d^B : U \to \mathcal{P}(V_d)$ defined by

$$\partial_d^B(u) = \{k : X_k \cap [u]_B \neq \emptyset\} = \left\{k : \mu_{X_k}^B(u) > 0\right\} \tag{47}$$

which labels each $u \in U$ with the set of *plausible* decision values (cf. [13, 16, 23, 27]).

Definition 4.4. Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We say that B *∂-preserves* d, iff for any $u \in U$ we have equality $\partial_d^B(u) = \partial_d^A(u)$. We say that B is a *∂-decision reduct*, iff it ∂-preserves d and none of its proper subsets does it.

Proposition 4.5. ([26]) Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. B is a ∂-decision reduct, iff it is an irreducible subset of A, which makes d O-independent from $A \setminus B$.

4.3. Decision reducts in terms of degrees of (in)consistency

Generalized decision functions model *the degree of inconsistency of defining decision* by means of cardinalities of their value sets. For any $\mathbb{A} = (U, A \cup \{d\})$, $B \subseteq A$ and $u \in U$, we have always inclusion $\partial_d^A(u) \subseteq \partial_d^B(u)$, i.e. the reduction of attributes causes potentially the loss of information corresponding to the sets of plausible decision values. In particular, for consistent $\mathbb{A} = (U, A \cup \{d\})$, we have always $\left|\partial_d^A(u)\right| = 1$ and any reduction to $B \subseteq A$ should be allowed, iff there is still $\left|\partial_d^B(u)\right| = 1$, for any $u \in U$. Actually, the notions of ∂-decision reduct and decision reduct are equivalent for consistent decision systems.

Exactly the same situation occurs in case of probabilities, encoded in terms of conditional entropy. Inequalities (30) show that the reduction of attributes causes potential growth of conditional entropy, i.e., due to Subsection 3.3, potential average decrease of precision of object related decision rules. Hence, according to the characteristics provided by Proposition 4.4, we should try to keep conditional entropy, interpreted as the degree of inconsistency, possibly at the same level while reducing attributes.

Yet another possibility refers to the following *decision discernibility measure*, considered in [28, 30] (as well as, e.g., during the *discretization process* in [3]):

$$Disc_{\mathbb{A}}(d/B) = \left|\left\{(u, u') \in U \times U : B(u) \neq B(u') \wedge d(u) \neq d(u')\right\}\right| \tag{48}$$

It counts all pairs $u, u' \in U$, which *are discerned* by $B \subseteq A$ (i.e. $B(u) \neq B(u')$) and *should be discerned* because of d (i.e. $d(u) \neq d(u')$). We say that B is a *discerning decision reduct*, iff

$$Disc_{\mathbb{A}}(d/B) = Disc_{\mathbb{A}}(d/A) \tag{49}$$

which can equivalently [28, 30] rewritten as

$$IND_{\mathbb{A}}(B) \setminus IND_{\mathbb{A}}(\{d\}) = IND_{\mathbb{A}}(A) \setminus IND_{\mathbb{A}}(\{d\}) \tag{50}$$

and none of its proper subsets does it. Due to (49), we are interested in subsets $B \subseteq A$, which discern all pairs $u, u' \in U$, which *should* and *could* (i.e. $A(u) \neq A(u')$) be discerned.

Finally, we can express the consistency level in terms of positive regions, already discussed in Subsection 2.3. For $B \subseteq A$, one has always inclusion $POS_{\mathbb{A}}(d/B) \subseteq POS_{\mathbb{A}}(d/A)$, where equality holds, iff B induces exactly the same lower approximations of particular decision classes as the whole A. In such a case, we would search for subsets B keeping consistency understood as the size of those approximations, i.e. satisfying equality

$$POS_{\mathbb{A}}(d/B) = POS_{\mathbb{A}}(d/A) \tag{51}$$

It turns out that all the mentioned criteria for keeping (in)consistency level while reducing the attributes can be compared with condition (44) for μ-preserving decision.

Proposition 4.6. (cf. [28, 30]) Let $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. If B satisfies (49), then it also μ-preserves d. If B μ-preserves d, then it also ∂-preserves d. Finally, if B ∂-preserves d, then it also satisfies (51).

In the following subsection, we are going to focus on the approximate versions of decision reducts corresponding to the criteria considered in the above proposition. Although, intuitively, it should be possible to keep similar implications for appropriately tuned approximations, there are no theoretical results in this area. We refer the reader to [28] for more examples concerning the comparison of such approximations.

4.4. Approximate entropy reducts

Conditions of preserving the degree of the model (in)consistency while reducing attributes turn out to be too rigorous with respect to possible noises and fluctuations in data. A solution would be to weaken them just like we did in case of discerning information reducts, as proposed in Subsection 4.1. For instance, continuing the topic of discernibility measures, we could be interested in subsets of attributes discerning *almost all* pairs of objects, which should and could be discerned. It leads to the following notion:

Definition 4.5. Let $\varepsilon \in [0, 1)$, $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We say that B is an *ε-approximate discerning decision reduct*, iff it satisfies inequality

$$Disc_{\mathbb{A}}(d/B) \geq (1 - \varepsilon) Disc_{\mathbb{A}}(d/A) \tag{52}$$

and of its proper subsets does it.

We can also try to express the concept of approximate preserving of rough membership information. One can discuss various approaches to weakening of μ-preserving conditions, based either on analyzing distances between probabilistic distributions or on aggregate quality measures [28, 29, 30, 31]. In Section 2, we proposed to evaluate $B \subseteq A$ according to the object related decision rules $B = B(u) \Rightarrow d = d(u)$. Hence, let us consider constraint

$$G_{\mathbb{A}}(d/B) \geq (1 - \varepsilon) G_{\mathbb{A}}(d/A) \tag{53}$$

which focuses our attention on subsets $B \subseteq A$ inducing rules $B = B(u) \Rightarrow d = d(u)$ being on average *almost* as precise as those induced by the whole set of conditions, i.e. $A = A(u) \Rightarrow d = d(u)$, for $u \in U$. Obviously, besides geometric average (37), one can consider also other measures, e.g. the

arithmetic average measures discussed in [29]. Here, by taking the logarithm of both sides of inequality (53), we get the following equivalent condition (54), expressing the precision approximation in terms of entropy:

Definition 4.6. Let $\varepsilon \in [0,1)$, $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We say that B (H,ε)-*approximately* μ*-preserves* d, iff

$$H_{\mathbb{A}}(d/B) + \log_2(1-\varepsilon) \leq H_{\mathbb{A}}(d/A) \tag{54}$$

We say that B is an (H,ε)-*approximate* μ*-decision reduct*, iff it satisfies (54) and none of its proper subsets does it.

Proposition 4.7. The notion of an $(H,0)$-approximate μ-decision reduct is equivalent to the notion of a μ-decision reduct.

For a consistent decision system $\mathbb{A} = (U, A \cup \{d\})$, where $G_{\mathbb{A}}(d/A) = 1$, the notion of an $(H,0)$-approximate μ-decision reduct is equivalent to the notion of a decision reduct. In such a case, by using positive approximation threshold $\varepsilon > 0$, we obtain condition

$$G_{\mathbb{A}}(d/B) \geq 1-\varepsilon \tag{55}$$

leading to smaller subsets $B \subseteq A$, which still ε*-approximately define* d. The same methodology can be applied to approximation of information reducts. We can refer to probabilities corresponding to the association rules $B = B(u) \Rightarrow A = A(u)$ and consider inequality

$$G_{\mathbb{A}}(A/B) \geq 1-\varepsilon \tag{56}$$

Just like in case of decision reducts, by taking the logarithm of both sides of (56), we get the following equivalent condition (57):

Definition 4.7. Let $\varepsilon \in [0,1)$, $\mathbb{A} = (U, A)$ and $B \subseteq A$ be given. We say that B (H,ε)-*approximately defines* $\mathbb{A}$, iff

$$H_{\mathbb{A}}(B) \geq H_{\mathbb{A}}(A) + \log_2(1-\varepsilon) \tag{57}$$

We say that B is an (H,ε)-*approximate information reduct*, iff it satisfies (57) and none of its proper subsets does it.

Proposition 4.8. The notion of an $(H,0)$-approximate information reduct is equivalent to the notion of an information reduct.

As a conclusion, we obtain conditions (54) and (57), which enable to intuitively express the notions of approximate preserving information in probabilistic framework, for the appropriately tuned threshold $\varepsilon \in [0,1)$.

Condition (54) can be generalized onto a very powerful notion of *approximate conditional probabilistic independence*, which has similar properties as the classical probabilistic model considered in [15]. It enables, e.g., to derive from data *approximate Bayesian networks*, based on *approximate Markov boundaries*. We refer the reader to [30, 31] for more details concerning this approach.

5. Complexity of the approximate reduction

5.1. Definitions and basic results

Let us consider the model optimization tasks concerned with the *Approximate Reduction Principle (ARP)*, which states that any simplification of the model, which *approximately* preserves its *precision*, should be performed to decrease its *complexity* (cf. [30, 32]). Let us begin with the problems related to Definitions 4.6, 4.7.

Definition 5.1. Let $\varepsilon \in [0, 1)$ be given. By *the Minimal* (H, ε)*-Approximate Decision Reduct Problem (MH*ε*DRP)* we mean the task of finding for each given decision system $\mathbb{A} = (U, A \cup \{d\})$ a minimal (in sense of the number of elements)[2] subset $B \subseteq A$ satisfying (54). By *the Minimal* (H, ε)*-Approximate Information Reduct Problem (MH*ε*IRP)* we mean the task of finding for each given information system $\mathbb{A} = (U, A)$ a minimal subset $B \subseteq A$ satisfying (57).

According to Propositions 4.7 and 4.8, the above problems generalize the original rough set based principles of the attribute reduction, for $\varepsilon = 0$.

Proposition 5.1. MH0DRP is equivalent to MμDRP – the task of finding for each given decision system a minimal μ-decision reduct. If we additionally restrict ourselves to the consistent decision systems, then MH0DRP becomes to be equivalent to MDRP – the task of finding for each given consistent decision system a minimal decision reduct. In the same way, MH0IRP is equivalent to MIRP – the task of finding for each given information system a minimal information reduct.

MDRP and MIRP are well known to be NP-hard [24]. As an example, let us sketch the proof of NP-hardness of MDRP. It is based on polynomial reduction of *the Minimal Dominating Set Problem (MDSP)* – the task of finding for each given undirected graph $\mathcal{G} = \left(A, \overline{E}\right)$ a minimal (in sense of the number of elements) subset $B \subseteq A$, which *covers* the whole set of vertices A, i.e. satisfies equality $Cov_{\mathcal{G}}(B) = A$, where

$$Cov_{\mathcal{G}}(B) = B \cup \left\{a \in A : \exists_{b \in B}(a, b) \in \overline{E}\right\} \tag{58}$$

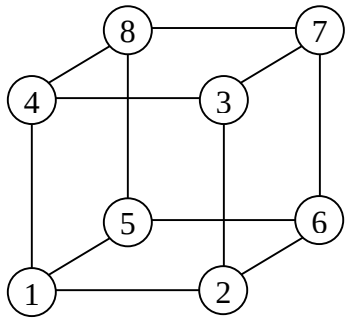
MDSP is reported as NP-hard in [6]. In purpose of reducing it to MDRP, one constructs for each $\mathcal{G} = \left(A, \overline{E}\right)$ a decision system $\mathbb{A}_{\mathcal{G}} = (U_{\mathcal{G}}, A \cup \{d_{\mathcal{G}}\})$ such that any minimal decision reduct $B \subseteq A$ for $\mathbb{A}_{\mathcal{G}}$ corresponds to a minimal subset, which satisfies (58). The idea of such a construction is very simple – we illustrate it in Figure 1. In the same way one proves the NP-hardness of MμDRP mentioned in Proposition 5.1, comparable to the problem of finding minimal Markov boundaries [28].

In [35] it was proposed to evaluate decision reducts with the number of induced decision rules. This is a step towards optimization of reducts by means of complexity of the corresponding decision models. Given decision reduct $B \subseteq A$ for $\mathbb{A} = (U, A \cup \{d\})$, the number of unique rules equals to the number of B-indiscernibility classes, i.e. $\left|V_B^U\right|$.

Definition 5.2. By *the Minimal Rule Decision Reduct Problem (MRDRP)* we mean the task of finding for each given consistent decision system $\mathbb{A} = (U, A \cup \{d\})$ a minimal decision reduct $B \subseteq A$, which satisfies condition

$$\left|V_B^U\right| = min_{C \subseteq A:\, C\ defines\ d} \left|V_C^U\right| \tag{59}$$

[2] In all foregoing specifications of the problems of finding various types of reducts, by a minimal subset of attributes we will mean a subset, which is minimal in sense of the number of elements.



$U_{\mathcal{G}}$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	$d_{\mathcal{G}}$
u_1	1	1	0	1	1	0	0	0	0
u_2	1	1	1	0	0	1	0	0	0
u_3	0	1	1	1	0	0	1	0	0
u_4	1	0	1	1	0	0	0	1	0
u_5	1	0	0	0	1	1	0	1	0
u_6	0	1	0	0	1	1	1	0	0
u_7	0	0	1	0	0	1	1	1	0
u_8	0	0	0	1	1	0	1	1	0
u_9	0	0	0	0	0	0	0	0	1

Figure 1. Example of $\mathcal{G} = \left(A, \overline{E}\right)$, where $A = \{1, \ldots, 8\}$, and appropriately constructed decision system $\mathbb{A}_{\mathcal{G}} = (U_{\mathcal{G}}, A \cup \{d_{\mathcal{G}}\})$, where objects $u_1, \ldots, u_8$ and attributes $a_1, \ldots, a_8$ correspond to vertices of $\mathcal{G}$. Values $a_i(u_j)$, $i, j = 1, \ldots, 8$, are equal to 1, iff $i = j$ or $(i, j) \in \overline{E}$. The rest of the system is specified in such a way that there is an equivalence between subsets defining $d_{\mathcal{G}}$ in $\mathbb{A}_{\mathcal{G}}$ and subsets dominating $\mathcal{G}$.

MRDRP was reported as NP-hard in [30, 35], in various forms. Here, let us show it by using a very simple reduction of NP-hard MIRP, specified in Proposition 5.1.

Theorem 5.1. MRDRP is NP-hard.

Proof:
For any information system $\mathbb{A} = (U, A)$, we construct decision system $\mathbb{A}^* = (U, A \cup \{d_{\mathbb{A}}\})$ by adding column $d_{\mathbb{A}}$, which takes different values for different A-indiscernibility classes in $\mathbb{A}$. Then, any subset $B \subseteq A$ defines $d_{\mathbb{A}}$ in $\mathbb{A}^*$, iff it defines $\mathbb{A}$. According to specification of $d_{\mathbb{A}}$, each such B must satisfy equality $\left|V_B^U\right| = \left|V_A^U\right|$. Hence, if B is the solution of MRDRP for $\mathbb{A}^*$, then it is the solution of MIRP for $\mathbb{A}$. □

The same methodology can be applied to other types of reducts. We can also define other complexity measures of decision models induced by considered subsets of attributes (cf. [30]). It leads to the ARP Principle specified at the beginning of this subsection. ARP was originally formulated in [32], as stating that, given a rule based decision model, any simplification, which approximately preserves *the expected chance of correct classification* should be performed to increase *the expected chance of the new case recognition*. In this paper, we concentrate on the *Approximate Entropy Reduction Principle (AERP)*, where we use entropy to express both precision (corresponding to the chance of correct classification) and complexity (intuitively opposite to the chance of the new case recognition, i.e. to the average strength of decision rules).

Definition 5.3. Let $\varepsilon \in [0, 1)$ be given. By *the H-Strength Optimal (H, ε)-Approximate Decision Reduct Problem (StHεDRP)* we mean the task of finding for each given $\mathbb{A} = (U, A \cup \{d\})$ a minimal (H, ε)-approximate μ-decision reduct $B \subseteq A$, which satisfies condition

$$H_{\mathbb{A}}(B) = min_{C \subseteq A:\, H_{\mathbb{A}}(d/C) + \log_2(1-\varepsilon) \leq H_{\mathbb{A}}(d/A)} H_{\mathbb{A}}(C) \tag{60}$$

Constraint in (60) is the same as in case of MHεDRP – it corresponds to condition (54). Optimization goals are, however, different. Let us note that in case of MHεDRP we are searching for $B \subseteq A$, which, for a given decision system $\mathbb{A} = (U, A \cup \{d\})$, satisfies condition

$$|B| = min_{C \subseteq A:\, H_{\mathbb{A}}(d/C) + \log_2(1-\varepsilon) \leq H_{\mathbb{A}}(d/A)} |C| \tag{61}$$

It means that the only important optimization criterion in MHεDRP is the number of attributes: first, we restrict ourselves to the family of subsets $B \subseteq A$, which satisfy (54); then, we choose one of subsets with minimal number of elements from the obtained family. In case of StHεDRP we are searching for a minimal (H, ε)-approximate μ-decision reduct $B \subseteq A$, which has minimal value of $H_{\mathbb{A}}(B)$. Such a reduct is not necessarily the one, which has minimal possible value of $|B|$. Just like before, we first consider the family of subsets, which satisfy (54); then, however, we additionally restrict to *subfamily* of subsets, which minimize $H_{\mathbb{A}}(B)$; finally, we choose a minimal element from such obtained subfamily.

According to (34), minimization of $H_{\mathbb{A}}(B)$ is equivalent to maximization of $G_{\mathbb{A}}(B)$, which is the geometric average of the strength of object related decision rules induced by B. Hence, searching for minimal subsets satisfying (60) leads to the decision rule based models with relatively (up to the choice of $\varepsilon \in [0, 1)$) high precision and possibly high average strength of the component rules. Similarly, instead of the average strength, we could optimize the average sensitivity. Given $\mathbb{A} = (U, A \cup \{d\})$ and $B \subseteq A$, we have

$$H_{\mathbb{A}}(B/d) = -\log_2(G_{\mathbb{A}}(B/d)) \tag{62}$$

where

$$G_{\mathbb{A}}(B/d) = \sqrt[|U|]{\Pi_{u \in U} P(B(u)/d(u))} \tag{63}$$

reflects the average sensitivity of decision rules $B = B(u) \Rightarrow d = d(u)$, for $u \in U$. It leads to the following modification of AERP:

Definition 5.4. Let $\varepsilon \in [0, 1)$ be given. By *the H-Sensitivity Optimal (H, ε)-Approximate Decision Reduct Problem (SeHεDRP)* we mean the task of finding for each given $\mathbb{A} = (U, A \cup \{d\})$ a minimal (H, ε)-approximate μ-decision reduct $B \subseteq A$, which satisfies condition

$$H_{\mathbb{A}}(B/d) = min_{C \subseteq A:\, H_{\mathbb{A}}(d/C) + \log_2(1-\varepsilon) \leq H_{\mathbb{A}}(d/A)} H_{\mathbb{A}}(C/d) \tag{64}$$

5.2. Minimal approximate entropy reducts

The main goal of this subsection is to show that the optimization problems formulated in Definition 5.1 are NP-hard. We base on the result, which has been already presented in [29], as helpful while considering another class of the approximate reduction problems. We recall it here in a slightly improved form.

Definition 5.5. Let $\alpha \in [0, 1)$, $\mathcal{G} = \left(A, \overline{E}\right)$ and $B \subseteq A$ be given. We say that B *α-approximately dominates $\mathcal{G}$*, iff

$$|Cov_{\mathcal{G}}(B)| \geq (1 - \alpha)|A| \tag{65}$$

Definition 5.6. Let $\alpha \in [0, 1)$ be given. By *the Minimal α-Approximate Dominating Set Problem (MαDSP)* we mean the task of finding for each given graph $\mathcal{G} = \left(A, \overline{E}\right)$ a minimal (in sense of the number of elements) subset $B \subseteq A$ satisfying (65).

Obviously, MDSP discussed in the previous subsection is a special case of MαDSP.

Theorem 5.2. ([29]) For any $\alpha \in [0,1)$, MαDSP is NP-hard.

Proof:
We reduce MDSP to MαDSP, for any $\alpha \in (0,1)$. We provide the method of constructing for any $\mathcal{G} = \left(A, \overline{E}\right)$ such graph $\mathcal{G}_\alpha = \left(A_\alpha, \overline{E}_\alpha\right)$, that the solution of M$\alpha$DSP for $\mathcal{G}_\alpha$ provides the solution of MDSP for $\mathcal{G}$. Let us put $\overline{E}_\alpha = \overline{E}$ and $A_\alpha = A \cup \left\{a^*_1, \ldots, a^*_{n(\alpha)}\right\}$, for

$$n(\alpha) = \left\lfloor \frac{|A|\alpha}{1-\alpha} \right\rfloor \tag{66}$$

As an example, for $\alpha = 0.3$ we obtain $n(0.3) = 3$ and $\mathcal{G}_{0.3}$ presented in Figure 2.

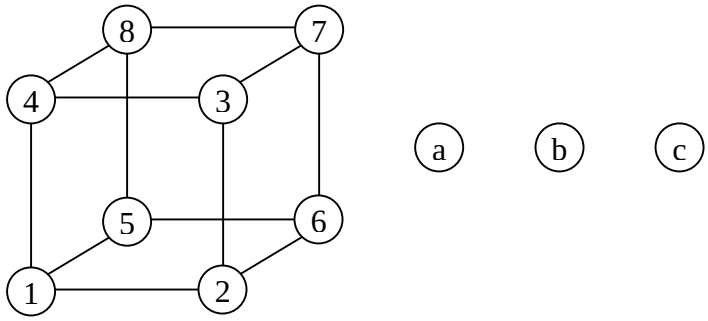


Figure 2. $\mathcal{G}_{0.3} = \left(\{1,\ldots,8,a,b,c\}, \overline{E}\right)$ resulting from extension of $\mathcal{G}$ from Figure 1. One can see that the minimal set, which 0.3-approximately dominates $\mathcal{G}_{0.3}$ can take the form of $B_{0.3} = \{1,7\}$, $\{2,8\}$, $\{3,5\}$ or $\{4,6\}$. In all these cases, $B_{0.3}$ is already the solution of MDSP for $\mathcal{G}$, i.e. $B^0 = B$ and $B^* = \emptyset$.

Quantity $n(\alpha)$ satisfies inequality

$$|A| - 1 < (1-\alpha)(|A| + n(\alpha)) \leq |A| \tag{67}$$

It implies that the number of vertices in $\mathcal{G}_\alpha$ can be bounded as follows:

$$|A_\alpha| = |A| + n(\alpha) \leq |A| \left(1 + \left\lceil \frac{\alpha}{1-\alpha} \right\rceil\right) \tag{68}$$

Let us assume that a minimal subset $B_\alpha \subseteq A_\alpha$, which α-approximately dominates $\mathcal{G}_\alpha$, is given. We show how to get from $B_\alpha \subseteq A_\alpha$ a minimal subset $B \subseteq A$ dominating $\mathcal{G}$. Let us express B_α as disjoint set theoretic sum $B_\alpha = B^0 \cup B^*$, where $B^0 \subseteq A$ and $B^* \subseteq \{a^*_1, \ldots, a^*_{n(\alpha)}\}$. Consider the subset dominating $\mathcal{G}$, defined as

$$B = B^0 \cup \left(A \setminus Cov_{\mathcal{G}}(B^0)\right) \tag{69}$$

We show that if B_α is a minimal set, which α-approximately dominates $\mathcal{G}_\alpha$, then B is a minimal set, which dominates $\mathcal{G}$. Let us notice that each subset dominating $\mathcal{G}$ α-approximately dominates $\mathcal{G}_\alpha$. This is because, according to (67), we have

$$|A| \geq (1-\alpha)(|A| + n(\alpha)) = (1-\alpha)|A_\alpha| \tag{70}$$

Hence, if B defined by (69) is a minimal set, which α-approximately dominates $\mathcal{G}_\alpha$, then B must be a minimal set, which dominates $\mathcal{G}$ as well. It remains to show that $|B| \leq |B_\alpha|$. Let us notice that

$$|B^*| \geq |A| - \left|Cov_{\mathcal{G}}\left(B^0\right)\right| \tag{71}$$

because otherwise we would have

$$|Cov_{\mathcal{G}_\alpha}(B_\alpha)| = \left|Cov_{\mathcal{G}}\left(B^0\right)\right| + |B^*| < |A| \tag{72}$$

which, according to (67), would imply

$$|Cov_{\mathcal{G}_\alpha}(B_\alpha)| \leq |A| - 1 < (1-\alpha)(|A| + n(\alpha)) = (1-\alpha)|A_\alpha| \tag{73}$$

and B_α would not α-approximately dominate $\mathcal{G}_\alpha$. Hence:

$$|B| = \left|B^0 \cup \left(A \setminus Cov_{\mathcal{G}}\left(B^0\right)\right)\right| = \left|B^0\right| + \left(|A| - |Cov_{\mathcal{G}}\left(B^0\right)|\right) \leq \left|B^0\right| + |B^*| = |B_\alpha| \tag{74}$$

□

As an example of the application of Theorem 5.2, let us consider following problem:

Definition 5.7. Let $\varepsilon \in [0,1)$ be given. By *the Minimal ε-Approximate Discerning Decision Reduct Problem (MεDDRP)* we mean the task of finding for each given decision system $\mathbb{A} = (U, A \cup \{d\})$ a minimal subset $B \subseteq A$ satisfying (52).

Theorem 5.3. (cf. [30]) For any $\varepsilon \in [0,1)$, MεDDRP is NP-hard.

Proof:
Let $\mathcal{G} = \left(A, \overline{E}\right)$ be given. We show that by considering the decision system $\mathbb{A}_{\mathcal{G}}$, constructed due to specification illustrated in Figure 1, MαDSP can be reduced to MεDDRP, for $\alpha = \varepsilon$. It is enough to notice that in such a case $Disc_{\mathbb{A}_{\mathcal{G}}}(d/B)$ equals to $|Cov_{\mathcal{G}}(B)|$. Hence, for $\alpha = \varepsilon$, conditions (52) and (65) are equivalent. □

Now, let us go back to the main topic, concerned with Definition 5.1.

Theorem 5.4. (cf. [30]) For any $\varepsilon \in [0,1)$, MHεDRP is NP-hard.

Proof:
We show how to construct, for each given graph $\mathcal{G} = \left(A, \overline{E}\right)$, decision system $\mathbb{A}^*_{\mathcal{G},\varepsilon} = (U^*_{\mathcal{G},\varepsilon}, A \cup \{d^*_{\mathcal{G},\varepsilon}\})$, for which $|U_{\mathcal{G},\varepsilon}|$ is linearly bounded by $n = |A|$ and there exists such $\alpha(\varepsilon) \in [0,1)$ that each subset $B \subseteq A$ $\alpha(\varepsilon)$-approximately dominates $\mathcal{G}$, iff it (H,ε)-approximately μ-preserves $d^*_{\mathcal{G},\varepsilon}$ in $\mathbb{A}^*_{\mathcal{G},\varepsilon}$. In this way, we will reduce in polynomial time the NP-hard M$\alpha(\varepsilon)$DSP Problem to MHεDRP.

We illustrate the following procedure in Figure 3. Each $a \in A$ has the set of values $V_a \subseteq \{1, \ldots, nr(\varepsilon)\}$, and decision $d^*_{\mathcal{G},\varepsilon}$ has the set of values $V_{d^*_{\mathcal{G},\varepsilon}} = \{1, \ldots, nr(\varepsilon)\}$, where

$$r(\varepsilon) = \left\lfloor \frac{1}{1-\varepsilon} + 1 \right\rfloor \tag{75}$$

$U^*_{\mathcal{G},\varepsilon}$	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	$d^*_{\mathcal{G},\varepsilon}$
u_1	1	1	1	1	1	1	1	1	1
u_2	2	2	1	2	2	1	1	1	2
u_3	3	3	1	3	3	1	1	1	3
u_4	4	4	4	4	4	4	4	4	4
u_5	5	5	5	4	4	5	4	4	5
u_6	6	6	6	4	4	6	4	4	6
u_7	7	7	7	7	7	7	7	7	7
u_8	7	8	8	8	7	7	8	7	8
u_9	7	9	9	9	7	7	9	7	9
u_{10}	10	10	10	10	10	10	10	10	10
u_{11}	11	10	11	11	10	10	10	11	11
u_{12}	12	10	12	12	10	10	10	12	12
u_{13}	13	13	13	13	13	13	13	13	13
u_{14}	14	13	13	13	14	14	13	14	14
u_{15}	15	13	13	13	15	15	13	15	15
u_{16}	16	16	16	16	16	16	16	16	16
u_{17}	16	17	16	16	17	17	17	16	17
u_{18}	16	18	16	16	18	18	18	16	18
u_{19}	19	19	19	19	19	19	19	19	19
u_{20}	19	19	20	19	19	20	20	20	20
u_{21}	19	19	21	19	19	21	21	21	21
u_{22}	22	22	22	22	22	22	22	22	22
u_{23}	22	22	22	23	23	22	23	23	23
u_{24}	22	22	22	24	24	22	24	24	24

Figure 3. For $\varepsilon = 0.6$ and $\mathcal{G} = \left(A, \overline{E}\right)$ from Figure 1, we construct decision system $\mathbb{A}^*_{\mathcal{G},0.6}$. Each vertex of $\mathcal{G}$ corresponds to one conditional attribute and $r(0.6) = 3$ objects.

For any $k = 1, \ldots, nr(\varepsilon)$, let us put $d^*_{\mathcal{G},\varepsilon}(u_k) = k$ and define $i(k) = \left\lceil \frac{k}{r(\varepsilon)} \right\rceil$. For any $a_j \in A$, $j = 1, \ldots, n$, we put

$$a_j(u_k) = \begin{cases} k & \text{if} \quad i(k) = j \vee \{a_{i(k)}, a_j\} \in \overline{E} \\ (i(k)-1)r(\varepsilon) + 1 & \text{otherwise} \end{cases} \tag{76}$$

For any $k = 1, \ldots, nr(\varepsilon)$ and $B \subseteq A$, we have two possibilities:

1. If there is an element of B connected with $a_{i(k)}$ by an edge in $\mathcal{G}$, then $[u_k]_B = \{u_k\}$. Distribution $\overrightarrow{\mu}_{d^*_{\mathcal{G},\varepsilon}/B}(u_k)$ is then a vertex of simplex $\triangle_{r(\varepsilon)-1}$ and its entropy is 0.

2. If there is no element of B connected with $a_{i(k)}$ by an edge in $\mathcal{G}$, then

$$[u_k]_B = \{u_{(i(k)-1)r(\varepsilon)}, \ldots, u_{i(k)r(\varepsilon)}\} \tag{77}$$

Distribution $\overrightarrow{\mu}_{d^*_{\mathcal{G},\varepsilon}/B}(u_k)$ is then uniform and its entropy is $\log_2 r(\varepsilon)$.

Summarizing, for any given $B \subseteq A$, we have

$$H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(d/B) = \frac{1}{nr(\varepsilon)} \sum_{k=1}^{nr(\varepsilon)} h_{d/B}(u_k) = \frac{1}{n} \left(0 \cdot |Cov_{\mathcal{G}}(B)| + \log_2 r(\varepsilon) \cdot |A \setminus Cov_{\mathcal{G}}(B)|\right) \tag{78}$$

what can be equivalently written as

$$H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(d/B) = \log_2 r(\varepsilon) \left(1 - \frac{|Cov_{\mathcal{G}}(B)|}{n}\right) \tag{79}$$

Let us notice that $H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(d/A) = 0$ and for any $B \subseteq A$ inequality

$$H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(d/B) + \log_2(1-\varepsilon) \leq H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(d/A) \tag{80}$$

is equivalent to inequality

$$\log_2 r(\varepsilon) \left(1 - \frac{|Cov_{\mathcal{G}}(B)|}{n}\right) + \log_2(1-\varepsilon) \leq 0 \tag{81}$$

and further to

$$\frac{|Cov_{\mathcal{G}}(B)|}{n} \geq 1 + \frac{\log_2(1-\varepsilon)}{\log_2 r(\varepsilon)} \tag{82}$$

Let us put

$$\alpha(\varepsilon) = \frac{\log_2(\frac{1}{1-\varepsilon})}{\log_2 r(\varepsilon)} \tag{83}$$

The above implies that for such defined $\alpha(\varepsilon) \in [0, 1)$ the solution of MHεDRP for decision system $\mathbb{A}^*_{\mathcal{G},\varepsilon}$ provides the solution of M$\alpha(\varepsilon)$DSP for undirected graph $\mathcal{G}$. □

We finish with the result concerning the second of problems specified in Definition 5.1.

Theorem 5.5. For any $\varepsilon \in [0, 1)$, MHεIRP is NP-hard.

Proof:
One can use similar construction as in the proof of Theorem 5.5. The only difference is that we skip the decision column – instead of decision system $\mathbb{A}^*_{\mathcal{G},\varepsilon} = (U^*_{\mathcal{G},\varepsilon}, A \cup \{d^*_{\mathcal{G},\varepsilon}\})$, we consider, for any given graph $\mathcal{G} = \left(A, \overline{E}\right)$, information system $\mathbb{A}^*_{\mathcal{G},\varepsilon} = (U^*_{\mathcal{G},\varepsilon}, A)$. One can see that for any $B \subseteq A$ inequality

$$H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(B) \geq H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(A) + \log_2(1-\varepsilon) \tag{84}$$

is equivalent to (82). Hence, by choosing $\alpha(\varepsilon) \in (0, 1)$ as equal to (83), we reduce NP-hard M$\alpha(\varepsilon)$DSP to the considered problem. □

5.3. Optimal approximate entropy reducts

In this subsection we consider two versions of the Approximate Entropy Reduction Principle, formulated in Definitions 5.3 and 5.4. In case of the task of the entropy based sensitivity optimization, we can proceed with the same construction as before.

Theorem 5.6. For any $\varepsilon \in [0,1)$, SeHεDRP is NP-hard.

Proof:
Let us notice that in decision system $\mathbb{A}^*_{\mathcal{G},\varepsilon} = (U^*_{\mathcal{G},\varepsilon}, A \cup \{d^*_{\mathcal{G},\varepsilon}\})$, specified for a given $\mathcal{G} = \left(A, \overline{E}\right)$ in the proof of Theorem 5.4, we have $H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(B/d) = 0$, for any $B \subseteq A$. Hence, the solutions of the problems MHεDRP and SeHεDRP for $\mathbb{A}^*_{\mathcal{G},\varepsilon}$ are identical. It means that the previously presented reduction of M$\alpha(\varepsilon)$DSP to MHεDRP works also for SeHεDRP. □

In case of the entropy based strength optimization, we need to use a slightly modified version of the problem concerning approximately dominating sets.

Definition 5.8. Let $\alpha \in [0,1)$ be given. By *the α-Approximate Minimally Dominating Set Problem (αMDSP)* we mean the task of finding for each given graph $\mathcal{G} = \left(A, \overline{E}\right)$ a minimal (in sense of the number of elements) subset $B \subseteq A$ such that

$$|Cov_{\mathcal{G}}(B)| = min_{C \subseteq A:\, |Cov_{\mathcal{G}}(C)| \geq (1-\alpha)|A|} |Cov_{\mathcal{G}}(C)| \tag{85}$$

Figure 4 illustrates the difference between Definitions 5.8 and 5.6. For $\alpha = 0.2$, subset $B = \{1,7\}$ is a minimal α-approximately dominating set for the above graph $\mathcal{G}$. Actually, it is also a minimal dominating set. It is, however, not a minimally α-approximate dominating set – it satisfies constraint $|Cov_{\mathcal{G}}(B)| \geq (1-\alpha)|B|$ but there is another $B' = \{1,3,8\}$, which satisfies it as well, and such that $|Cov_{\mathcal{G}}(B')| < |Cov_{\mathcal{G}}(B)|$.

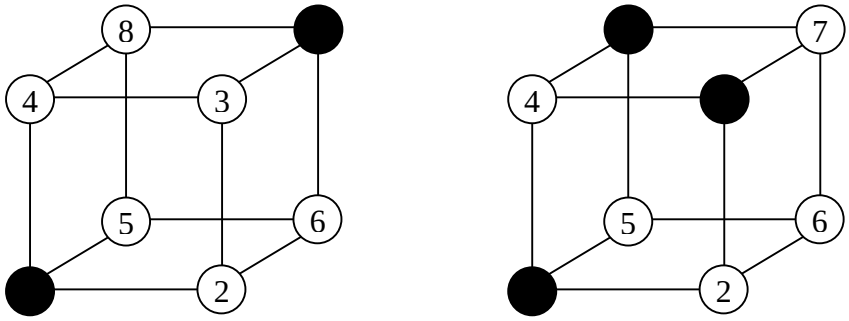


Figure 4. Solutions of the MαDSP and αMDSP Problems, for $\alpha = 0.2$.

Theorem 5.7. ([30]) For any $\alpha \in [0,1)$, αMDSP is NP-hard.

Proof:
Let $\alpha \in [0,1)$ be given. Consider, for a given $\mathcal{G} = \left(A, \overline{E}\right)$, graph $\mathcal{G}_\alpha$ constructed like in the proof of Theorem 5.2. Let us note that for $\mathcal{G}_\alpha$ the family of α-approximate dominating sets coincides with the family of subsets $B \subseteq A$ satisfying condition (85) if considered for $\mathcal{G}_\alpha$. Indeed, all minimal sets α-approximately dominating $\mathcal{G}_\alpha$ must dominate exactly $n = |A|$ vertices. Otherwise, if for a given $B_\alpha \subseteq A_\alpha$ there is $Cov_{\mathcal{G}_\alpha}(B_\alpha) > n$, then B_α includes at least one isolated vertex $a^* \notin A$. Then B_α is not minimal because after removing a^* we would still obtain the subset, which α-approximately dominates $\mathcal{G}_\alpha$. Hence, the way of constructing $\mathcal{G}_\alpha$ for each particular $\mathcal{G}$ implies that MDSP is reducible in polynomial time to αMDSP. □

Theorem 5.8. (cf. [30]) For any $\varepsilon \in [0,1)$, StHεDRP is NP-hard.

Proof:
For a given $\mathcal{G} = \left(A, \overline{E}\right)$, consider decision system $\mathbb{A}^*_{\mathcal{G},\varepsilon} = (U^*_{\mathcal{G},\varepsilon}, A \cup \{d^*_{\mathcal{G},\varepsilon}\})$, constructed just like in the proof of Theorem 5.4. For each subset $B \subseteq A$, we have $H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(B) =$

$$= \frac{1}{nr(\varepsilon)} \sum_{k=1}^{nr(\varepsilon)} \log_2(\mu_B(u_k)) = \frac{1}{n}\left[\log_2\left(\frac{1}{n}\right)|Cov_{\mathcal{G}}(B)| + \log_2\left(\frac{1}{nr(\varepsilon)}\right)|A \setminus Cov_{\mathcal{G}}(B)|\right] \quad (86)$$

The following holds:

$$\forall_{B,C\subseteq A}\,[\,|Cov_{\mathcal{G}}(B)| \leq |Cov_{\mathcal{G}}(C)| \Rightarrow H_{\mathbb{A}}(B) \leq H_{\mathbb{A}}(C)\,] \quad (87)$$

Hence, any H-optimal (H,ε)-approximate reduct $B \subseteq A$ for $\mathbb{A}^*_{\mathcal{G},\varepsilon}$ is also a subset, which $\alpha(\varepsilon)$-approximately dominates $\mathcal{G}$ and satisfies property (85). It implies that $\alpha(\varepsilon)$MDSP is reducible in polynomial time to StHεDRP, by using the same construction as that in the proof of Theorem 5.4. □

Theorem 5.7 can be used also for obtaining similar results. For instance, in Subsection 5.1, we were talking about complexity expressed by the number of decision rules, i.e. $\left|V^U_B\right|$, for a given decision reduct $B \subseteq A$, in $\mathbb{A} = (U, A \cup \{d\})$. We claimed that such a complexity measure could be used also for, e.g., (H,ε)-approximate μ-decision reducts.

Definition 5.9. Let $\varepsilon \in [0,1)$ be given. By *the Minimal Rule (H,ε)-Approximate Decision Reduct Problem (MRHεDRP)* we mean the task of finding for each given $\mathbb{A} = (U, A \cup \{d\})$ a minimal (in sense of the number of elements) (H,ε)-approximate μ-decision reduct $B \subseteq A$, which satisfies condition

$$\left|V^U_B\right| = min_{C\subseteq A:\, H_{\mathbb{A}}(d/C)+\log_2(1-\varepsilon)\leq H_{\mathbb{A}}(d/A)} \left|V^U_C\right| \quad (88)$$

Theorem 5.9. (cf. [30]) For any $\varepsilon \in [0,1)$, MRHεDRP is NP-hard.

Proof:
We reduce $\alpha(\varepsilon)$MDSP to MRHεDRP, just like in the proof of Theorem 5.8. For a given $\mathcal{G} = \left(A, \overline{E}\right)$, we construct decision system $\mathbb{A}^*_{\mathcal{G},\varepsilon} = (U^*_{\mathcal{G},\varepsilon}, A \cup \{d^*_{\mathcal{G},\varepsilon}\})$, in the same way as before. However, instead of $H_{\mathbb{A}^*_{\mathcal{G},\varepsilon}}(B)$ given by (86), we consider

$$\left|V^{U^*_{\mathcal{G}}}_B\right| = r(\varepsilon) \cdot |Cov_{\mathcal{G}}(B)| + |A \setminus Cov_{\mathcal{G}}(B)| \quad (89)$$

One can see that

$$\forall_{B,C\subseteq A}\left[\,|Cov_{\mathcal{G}}(B)|\leq|Cov_{\mathcal{G}}(C)|\Rightarrow\left|V_B^{U^*_{\mathcal{G}}}\right|\leq\left|V_C^{U^*_{\mathcal{G}}}\right|\,\right] \tag{90}$$

It can be used as (87) in the previous proof, to show the wanted reduction. □

As a conclusion, we obtain a wide range of theoretical results, characterizing the search of optimal (H,ε)-approximate μ-decision and information reducts as NP-hard.

6. Conclusions

We introduced the Approximate Entropy Reduction Principle (AERP), which generalizes the rough set based reduction laws by basing on the measure of information entropy. It states that any simplification (reduction of attributes) in the decision model, which approximately preserves its conditional entropy (the measure of inconsistency of defining decision by conditional attributes) should be performed to decrease its prior entropy (the measure of the model's complexity).

Optimization tasks concerned with various versions of AERP are shown to be NP-hard. Theorems 5.1–5.4 and 5.7–5.9 refer to the results, which are already known in similar form (cf. [29, 30, 35]). We attach their proofs, for they are much improved in comparison with the original sources. Theorems 5.5 and 5.6, which refer to the search of minimal (H,ε)-approximate information reducts and (H,ε)-approximate μ-decision reducts optimizing the entropy based measure of sensitivity, respectively, complete the presented analysis.

In the nearest future we are going to adapt the existing rough set based approaches to approximate solving of the reduction problems ([3, 32, 34, 35]) to extracting of optimal approximate entropy reducts from real life data (cf. [33]).

Acknowledgements Supported by Polish National Committee for Scientific Research (KBN) grant No. $8T11C02519$.

References

[1] Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., Verkamo, A.I.: Fast discovery of assocation rules. In: V.M. Fayad, G. Piatetsky Shapiro, P. Smyth, R. Uthurusamy (eds): Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press (1996) pp. 307–328.

[2] An, A., Cercone, N.: Rule quality measures for rule induction systems: Description and evaluation. Computational Intelligence **17/3** (2001) pp. 409–424.

[3] Bazan, J.G., Nguyen, H.S., Nguyen, S.H, Synak, P., Wróblewski, J.: Rough Set Algorithms in Classification Problem. In: [18] (2000) pp. 49–88.

[4] Duentsch, I., Gediga, G.: Uncertainty measures of rough set prediction. Artificial Intelligence **106** (1998) pp. 77–107.

[5] Gallager, R.G.: Information Theory and Reliable Communication. Wiley (1968).

[6] Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to The Theory of NP-Completeness. Freeman and Company (1979).

[7] Kapur, J.N., Kesavan, H.K.: Entropy Optimization Principles with Applications. Academic Press (1992).

[8] Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial. In: S.K. Pal, A. Skowron (eds): Rough Fuzzy Hybridization – A New Trend in Decision Making. Springer Verlag (1999) pp. 3–98.

[9] Kloesgen, W., Żytkow, J.M. (eds): Handbook of Data Mining and Knowledge Discovery. Oxford University Press (2002).

[10] Li, M., Vitanyi, P.: An Introduction to Kolmogorov Complexity and Its Applications. Springer Verlag (1997).

[11] Mitchell, T.: Machine Learning. Mc Graw Hill (1998).

[12] Pawlak, Z.: Information Systems: Theoretical Foundations (In Polish). WNT (1983).

[13] Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer (1991).

[14] Pawlak, Z., Skowron, A.: Rough membership functions. In: Advances in the Dempster Shafer Theory of Evidence. Wiley (1994) pp. 251–271.

[15] Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988).

[16] Polkowski, L.: Rough Sets: Mathematical Foundations. Phisica Verlag (2002).

[17] Polkowski, L., Skowron, A. (eds): Rough Sets in Knowledge Discovery. Physica Verlag (1998), parts **1**, **2**.

[18] Polkowski, L., Tsumoto, S., Lin, T.Y. (eds): Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems. Physica Verlag (2000).

[19] Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. In: Proc. of IMLC'98. Madison, WI (1998).

[20] Rissanen J.: Minimum-description-length principle. In: S. Kotz, N.L. Johnson (eds), Encyclopedia of Statistical Sciences. Wiley (1985) pp. 523–527.

[21] Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal **27** (1948) pp. 379–423, 623–656.

[22] Shenoy, P.P.: Conditional Independence in Valuation-based Systems. International Journal of Approximate Reasoning **10** (1994) pp. 203–234.

[23] Skowron, A.: Extracting laws from decision tables. Computational Intelligence **11/2** (1995) pp. 371–388.

[24] Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: R. Slowiński (ed.): Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory. Kluwer (1992) pp. 311–362.

[25] Ślęzak, D.: Approximate reducts in decision tables. In: Proc. of IPMU'96. Granada, Spain (1996) 3, pp. 1159–1164.

[26] Ślęzak, D.: Rough Set Reduct Networks. In: Proc. of RSSC'97. Durham, NC (1997) 3 pp. 77–80.

[27] Ślęzak, D.: Decomposition and Synthesis of Decision Tables with Respect to Generalized Decision Functions. In: S.K. Pal, A. Skowron (eds): Rough Fuzzy Hybridization – A New Trend in Decision Making. Springer Verlag (1999) pp. 110–135.

[28] Ślęzak, D.: Various approaches to reasoning with frequency-based decision reducts: a survey. In: [18] (2000) pp. 235–285.

[29] Ślęzak, D.: Normalized decision functions and measures for inconsistent decision tables analysis. Fundamenta Informaticae **44/3** (2000) pp. 291–319.

[30] Ślęzak, D.: Approximate decision reducts (in Polish). Ph.D. thesis, Institute of Mathematics, Warsaw University (2001).

[31] Ślęzak D.: Approximate Bayesian Networks. In: B. Bouchon-Meunier, J. Gutierrez-Rios, L. Magdalena, R.R. Yager (eds), Technologies for Contructing Intelligent Systems 2: Tools. Springer Verlag (2002) pp. 313–326.

[32] Ślęzak, D., Wróblewski, J.: Application of Normalized Decision Measures to the New Case Classification. In: Proc. of RSCTC'2000. Banff, Canada (2000) pp. 515–522.

[33] Ślęzak, D., Wróblewski, J.: Order-based genetic algorithms for the search of approximate entropy reducts. In: Proc. of RSFDGrC'2003. Chongqing, China (2003).

[34] Wróblewski, J.: Theoretical Foundations of Order-Based Genetic Algorithms. Fundamenta Informaticae **28/3-4** (1996) pp. 423–430.

[35] Wróblewski, J.: Adaptive methods of object classification (in Polish). Ph.D. thesis, Institute of Mathematics, Warsaw University (2001).

[36] Ziarko, W.: Variable Precision Rough Set Model. Journal of Computer and System Sciences **40** (1993) pp. 39–59.