# Approximate Markov Boundaries and Bayesian Networks: Rough Set Approach

Dominik Ślęzak

Polish-Japanese Institute of Information Technology
Koszykowa 86, 02-008 Warsaw, Poland
email: slezak@pjwstk.edu.pl

**Abstract.** We consider approximate bayesian networks, which almost keep the information entropy of data and encode knowledge about approximate dependencies between features. We develop the rough set based framework for extraction of such networks from empirical data, by relating the notion of an approximate rough membership decision reduct and the notion of an approximate Markov boundary.

## 1 Introduction

Bayesian network (BN) is a directed acyclic graph (DAG) designed to encode knowledge about probabilistic conditional independence (PCI) statements between random variables, within the product probabilistic space ([1,2,8]). Its expressive power increases while removing the edges, unless it causes a loss of control of exactness of derivable statements. Bayesian networks can be applied, e.g., to model the flow of information while reasoning about new cases by analogy to records gathered in decision tables ([14,16]).

Exact PCI is too accurate while mining real life data, because of the risk of possible noises or fluctuations. Thus, one needs to generalize fundamental concepts and results concerning BNs, to let them deal with appropriately defined approximate PCI-statements. The idea of basing this generalization on the rough set framework ([6]) originates from the fact that it provides a wide range of tools for expressing data inconsistency, in particular, those related to rough membership functions ([7]). The notion of a rough membership decision reduct (cf. [12]) corresponds to the probabilistic notion of a Markov boundary (MB), crucial for the BN related models ([8]).

Various methods for the reduction of noises and redundant information by the approximate preserving of rough membership information (cf. [15]) can be applied to approximate the concepts related to MBs and BNs. We focus on approximations dedicated to the information measure of entropy ([4,5]). The rough set based Reduction Principle ([6,9]) suggests – in this particular case – constructing decision models by reducing conditional attributes, unless it causes too large increase of entropy reflecting inaccuracy of conditions→decision dependencies. In the same way, one can extract approximate BNs with possibly low number of edges, but still *almost* preserving

entropy of data ([13,14,16]). This idea can be also referred to the well known Minimum Description Length Principle ([4,10]), by means of the balance between the entropy based complexity and accuracy of modeling data.

The paper continues research developed in [13,14,16], as well as the Ph.D. Thesis [15] supervised by Professor Andrzej Skowron, at Warsaw University, Poland. Let us also refer the reader to [17], where algorithms for extraction of approximate BNs from data are discussed, and [18], where experiments concerning their application to the new case classification are presented.

## 2 Data based probabilistic models

### 2.1 Rough membership functions

We represent data as an information system $\mathbf{A} = (U, A)$, where 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 write $A = \langle a_1, \ldots, a_n \rangle$ according to some ordering over the set of attributes. For any $B \subseteq A$ and $u \in U$, let us consider vector

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

where $a_{i_j}(u)$ denote the values of successive attributes $a_{i_j} \in B$, $j = 1, \ldots, m$, $m = |B|$. The set of all vectors of values on $B$, which occur in $\mathbf{A}$, takes the following form:

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

The classification task is concerned with a distinguished decision to be predicted under information provided over the rest of attributes. For this purpose, let us represent data as a decision table $\mathbf{A} = (U, A \cup \{d\})$, $d \notin A$. Let $V_d = \{v_1, \ldots, v_r\}$, $r = |V_d|$. For each $k = 1, \ldots, r$, let us define the $k$-th decision class $X_k \subseteq U$ by

$$X_k = \{u \in U : d(u) = v_k\} \tag{3}$$

One can operate with probability of occurrence of $v_k \in V_d$ conditioned by $w_B \in V_B^U$:

$$P_{\mathbf{A}}(v_k / w_B) = \frac{|\{u \in X_k : B(u) = w_B\}|}{|\{u \in U : B(u) = w_B\}|} \tag{4}$$

It corresponds to the $\alpha$-inexact decision rule $B = w_B \Rightarrow_\alpha d = v_k$, with the accuracy level $\alpha = P_{\mathbf{A}}(v_k / w_B)$. The strength of the rule is provided by the prior probability $P_{\mathbf{A}}(w_B) = |\{u \in U : B(u) = w_B\}| \,/\, |U|$. It expresses the chance that an object $u \in U$ will be recognized, i.e., it will satisfy the left side of the rule.

Probabilities were introduced to the theory of rough sets in [7], as rough membership functions:

**Definition 1.** Let $\mathbf{A} = (U, A)$, $B \subseteq A$ and $X \subseteq U$ be given. The $B$-rough membership function $\mu_X^B : U \to [0, 1]$ is defined by

$$\mu_X^B(u) = \frac{|[u]_B \cap X|}{|[u]_B|} \tag{5}$$

where $[u]_B = \{u' \in U : B(u) = B(u')\}$ is the $B$-indiscernibility class of $u$.

Given $\mathbf{A} = (U, A \cup \{d\})$, we are especially interested in subsets $X \subseteq U$, which are decision classes. One can see that for any $B \subseteq A$, $u \in U$ and $k = 1, \ldots, r$, the following equality holds:

$$\mu_{X_k}^B(u) = P_{\mathbf{A}}(v_k / B(u)) \tag{6}$$

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 table.

### 2.2 Decision reducts and Markov boundaries

The fundamental rough set principle is concerned with the reduction of possibly large amount of redundant information ([6]). It corresponds to the notion of a decision reduct – an irreducible subset of conditional attributes, which enables to define (approximate) decision classes with the same accuracy as the whole set. Specification of such an accuracy requires a formal way of expressing inexact conditions→decision dependencies. One can operate with the rough set based set approximations, generalized decision functions or, for instance, rough membership functions (cf. [12,15]). In that last case, the notion of a reduct takes the following form:

**Definition 2.** Let $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We say that $B$ preserves rough membership information about $d$ ($\mu$-preserves $d$, in short), iff for any $u \in U$ and $k = 1, \ldots, r$ we have equality

$$\mu_{X_k}^B(u) = \mu_{X_k}^A(u) \tag{7}$$

We say that $B$ is a rough membership decision reduct ($\mu$-decision reduct, in short), iff it satisfies (7) and none of its proper subsets does it.

The following is a simple generalization of the result shown in [11]:

**Theorem 1.** *([15]) The problem of finding minimal $\mu$-decision reduct for a given decision table $\mathbf{A} = (U, A \cup \{d\})$ is NP-hard.*

One can also adapt from [11] the idea of basing the process of searching for $\mu$-decision reducts on boolean reasoning:

**Proposition 1.** *([15]) Let $\mathbf{A} = (U, A \cup \{d\})$ be given. Let us consider boolean formula $\alpha_\mu \equiv \bigwedge_{u,u':\exists_k \mu^A_{X_k}(u) \neq \mu^A_{X_k}(u')} \bigvee_{a:a(u)\neq a(u')} a$. Subset $B \subseteq A$ is a $\mu$-decision reduct for $\mathbf{A}$, iff $\alpha_B \equiv \bigwedge_{a\in B} \overline{a}$ is a prime implicant of $\alpha_\mu$.*[1]

Criterion (7) can be equivalently rewritten in various ways (cf. [15]).

**Proposition 2.** *([15]) Let $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. $B$ $\mu$-preserves $d$, iff for any $u \in U$ we have equality*

$$\mu^B_{X_{d(u)}}(u) = \mu^A_{X_{d(u)}}(u) \tag{8}$$

*where $X_{d(u)}$ denotes the decision class, which $u$ belongs to.*

One can draw a correspondence between the notions of a $\mu$-decision reduct and a Markov boundary [8] – irreducible subset of random variables, which makes a distinguished variable probabilistically independent from the rest of variables. Let us formulate probabilistic conditional independence in terms of an arbitrary information system $\mathbf{A} = (U, A)$, where $P_\mathbf{A}$ is treated as the probability distribution over the product of discrete random variables $A$.

**Definition 3.** Let $\mathbf{A} = (U, A)$ and mutually disjoint subsets $X, Y, Z \subseteq A$ be given. We say that $Y$ makes $X$ conditionally independent from $Z$, iff for all possible configurations of $w_X$, $w_Y$ and $w_Z$ – being vectors of values over $X$, $Y$ and $Z$, respectively – we have implication

$$P_\mathbf{A}(w_Y, w_Z) > 0 \Rightarrow P_\mathbf{A}(w_X/w_Y, w_Z) = P_\mathbf{A}(w_X/w_Y) \tag{9}$$

We will call the fact of such an independence as a conditional independence statement (PCI-statement, in short), and denote it by $I_\mathbf{A}(X/Y/Z)$.

**Proposition 3.** *([15]) Let information system $\mathbf{A} = (U, A)$ and mutually disjoint subsets $X, Y, Z \subseteq A$ be given. PCI-statement $I_\mathbf{A}(X/Y/Z)$ holds, iff for all objects $u \in U$ we have*

$$P_\mathbf{A}(X(u)/Y(u)) = P_\mathbf{A}(X(u)/Y(u), Z(u)) \tag{10}$$

Let us consider arbitrary $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$. One can see that for $X = \{d\}$, $Y = B$ and $Z = A \setminus B$ conditions (8) and (10) are equivalent. Hence, $B \subseteq A$ preserves rough membership information about $d$, iff PCI-statement $I_\mathbf{A}(\{d\}/B/A \setminus B)$ holds. Moreover, $B$ is a $\mu$-decision reduct, iff none of its proper subsets satisfies an analogous PCI-statement, i.e. iff it is a Markov boundary of $d$ with respect to the product distribution $P_\mathbf{A}$ over $A \cup \{d\}$.

The above observation provides us with a strong correspondence between theory of rough sets and probabilistic calculus. As a corollary, we obtain the well known fact that the problem of finding minimal Markov boundary for a random variable, within a product probabilistic space, is NP-hard (cf. [1]). Moreover, one can adapt the rough set based algorithmic tools (cf. [9]) to search for Markov boundaries.

[1] For the definition of prime implicant let us refer to [11].

### 2.3 Bayesian networks

A bayesian network (BN) has the structure of a directed acyclic graph (DAG) $\mathcal{D} = (A, \overrightarrow{E})$, where $\overrightarrow{E} \subseteq A \times A$. The objective of BN is to encode PCI-statements involving groups of probabilistic variables corresponding to elements of $A$, in terms of the following graph-theoretic notion ([8]):

**Definition 4.** Let DAG $\mathcal{D} = (A, \overrightarrow{E})$ and mutually disjoint subsets $X, Y, Z \subseteq A$ be given. We say that $Y$ d-separates $X$ from $Z$, iff any path between any node in $X$ and any node in $Z$ comes through a serial or diverging connection covered by some element of $Y$, or a converging connection not covered by $Y$, having no descendant in $Y$.[2] We will denote such a fact by $\langle X/Y/Z \rangle_{\mathcal{D}}$.

Let us formulate the notion of a bayesian network within the rough set framework, i.e. for probabilistic distributions induced by information systems:

**Definition 5.** Let $\mathbf{A} = (U, A)$ and DAG $\mathcal{D} = (A, \overrightarrow{E})$ be given. We say that $\mathcal{D}$ is a bayesian network for $\mathbf{A}$, iff for any mutually disjoint subsets $X, Y, Z \subseteq A$, if $\langle X/Y/Z \rangle_{\mathcal{D}}$ holds, then $I_{\mathbf{A}}(X/Y/Z)$ holds as well.

The following Theorem 2 was originally stated in [8] in terms of arbitrary probabilistic distributions and Markov boundaries. Here we simply rewrite it in the language of information systems, by basing on the analogy between Markov boundaries and $\mu$-decision reducts.

**Theorem 2.** *([8]) Let $\mathbf{A} = (U, A)$, with arbitrary ordering $A = \langle a_1, \ldots, a_n \rangle$, $n = |A|$, be given. Let us assume that for each decision*

$$\mathbf{A}_i = (U, \{a_1, \ldots, a_{i-1}\} \cup \{a_i\}) \tag{11}$$

*where $i = 1, \ldots, n$, subset $B_i \subseteq \{a_1, \ldots, a_{i-1}\}$, which $\mu$-preserves $a_i$ in $\mathbf{A}_i$, is given.*[3] *DAG $\mathcal{D} = (A, \overrightarrow{E})$ defined by putting*

$$\overrightarrow{E} = \bigcup_{i=1}^{n} B_i \times \{a_i\} \tag{12}$$

*is a BN for $\mathbf{A}$. Moreover, each BN for $\mathbf{A}$ can be constructed in this way.*

The above result provides a methodology of searching for bayesian networks for experimental data. A number of algorithmic approaches to such a search has been proposed (cf. [2,17]). The most widely applied optimization criterion corresponds to minimization of the network edges, although another optimization measures are considered as well ([18]). In [3] it is shown that the task of finding bayesian network with minimal number of edges is NP-hard.

[2] Descriptions '*serial*', '*diverging*' and '*converging*' correspond to directions of arrows meeting within a given path, in a given node; $b$ is said to be a descendant of $a$, iff there is a directed path from $a$ towards $b$ in $\mathcal{D}$.

[3] For $\mathbf{A}_1$ we put $B_1 = \emptyset$. Actually, it may happen that $B_i$ is $\emptyset$ also for some $i > 1$, if the prior distribution for $a_i$ is the same as the one conditioned by $\{a_1, \ldots, a_{i-1}\}$.

In [1] one can find a discussion about the meaning of the number of edges in BNs for various modeling tasks. Here, let us consider the example from [14], concerned with the task of classification of new cases. Let us assume that new cases may have no known values over some conditional attributes. A question is then whether we can reduce irrelevant input information in each particular situation. Since a new case can have unknown values over any subset of $A$, we need a tool for an effective derivation of possible reductions of all $B \subseteq A$. Let us state the requirement for such a tool as follows:

*Problem 1.* Given $\mathbf{A} = (U, A \cup \{d\})$, construct a procedure being able to assign to each input $B \subseteq A$ possibly small output $C \subseteq B$, such that PCI-statement $I_{\mathbf{A}}(\{d\}/C/B \setminus C)$ is satisfied. The assigning procedure should neither store in memory relevant subsets for all $B \subseteq A$ nor run intermediate calculations over $\mathbf{A}$ for each particular input.

Let us show how to reach the above goal by basing on a previously found BN for $\mathbf{A}$. Let us introduce some helpful DAG-related notions:

**Definition 6.** Let $\mathcal{D} = (A, \overrightarrow{E})$ and $a \in A$ be given. By the set of parents of $a$ in $\mathcal{D}$ we mean

$$\pi_{\mathcal{D}}(a) = \left\{ a' \in A : \langle a', a \rangle \in \overrightarrow{E} \right\} \tag{13}$$

Let $B \subseteq A$, $a \notin B$, be given. By the set of $B$-ascendants of $a$ in $\mathcal{D}$ we mean

$$\pi^B_{\mathcal{D}}(a) = \{ b \in B : b \rightarrow_{A \setminus B} a \} \tag{14}$$

where $b \rightarrow_{A \setminus B} a$ means that either $\langle b, a \rangle \in \overrightarrow{E}$ or there exists in $\overrightarrow{E}$ a directed path from $b$ towards $a$, with all intermediate nodes in $A \setminus B$.[4]

**Definition 7.** Let DAG $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ be given. Let $d$ be the root of $\mathcal{D}$, i.e. $\pi_{\mathcal{D}}(d) = \emptyset$. For any $B \subseteq A$, the $\mathcal{D}$-boundary of $d$ with respect to $B$ is defined by

$$bound_{\mathcal{D}}(B) = \bigcup_{a \in B:\, d \rightarrow_{A \setminus B} a} \pi^B_{\mathcal{D}}(a) \cup \{a\} \tag{15}$$

**Proposition 4.** *Let $\mathbf{A} = (U, A \cup \{d\})$ and BN $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ with the root in $d$ be given. For any $B \subseteq A$, we have $bound_{\mathcal{D}}(B) \subseteq B$. Moreover, for subset $C = bound_{\mathcal{D}}(B)$, PCI-statement $I_{\mathbf{A}}(\{d\}/C/B \setminus C)$ is satisfied.*

Due to the above fact, the wanted procedure can be based on any BN $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ with the root in $d$. It allows to extract reductions of arbitrary subsets just on the basis of the DAG structure, without necessity of calculations over a data table. Hence, it is worth searching for BNs with less edges because they provide smaller $\mathcal{D}$-boundaries for particular subsets.

[4] Let us note that $\pi_{\mathcal{D}}(a) = \pi^A_{\mathcal{D}}(a)$.

## 3 Approximate probabilistic models

### 3.1 Information entropy

Condition of preserving probability (rough membership) distribution while reducing attributes in Definition 2 turns out to be too rigorous with respect to possible noises and fluctuations. It is usually impossible to derive relatively small Markov boundaries ($\mu$-decision reducts) as well as bayesian networks from real life data. A solution would be to weaken classical probabilistic criteria. One can discuss various approaches to such a weakening, based either on keeping distances between probabilistic distributions or quality measures which encode them ([15,16]). For this purpose, we propose to consider the well known measure of information entropy ([4,5]).

**Definition 8.** Let $\mathbf{A} = (U, A)$ and $X \subseteq A$ be given. By entropy of $X$ with respect to probability distribution $P_{\mathbf{A}}$ we mean quantity $H_{\mathbf{A}}(X) =$

$$= -\sum_{w_X : P_{\mathbf{A}}(w_X) > 0} P_{\mathbf{A}}(w_X) \log_2 P_{\mathbf{A}}(w_X) = -\sum_{w_X \in V_X^U} P_{\mathbf{A}}(w_X) \log_2 P_{\mathbf{A}}(w_X) \tag{16}$$

Let $Y \subseteq A$, $X \cap Y = \emptyset$, be given. By entropy of $X$ under $Y$ we mean:

$$H_{\mathbf{A}}(X/Y) = H_{\mathbf{A}}(X \cup Y) - H_{\mathbf{A}}(Y) \tag{17}$$

Although we focus on information systems $\mathbf{A} = (U, A)$, the above notion can be stated for an arbitrary discrete product distribution. In the particular case of $P_{\mathbf{A}}$ it is possible to provide the following interpretation:

**Proposition 5.** *Let $\mathbf{A} = (U, A)$ and disjoint subsets $X, Y \subseteq A$ be given. We have equality*

$$H_{\mathbf{A}}(X/Y) = -\frac{1}{|U|} \sum_{u \in U} \log_2 P_{\mathbf{A}}(X(u)/Y(u)) \tag{18}$$

According to the above equalities, $H_{\mathbf{A}}(X/Y)$ is a kind of average measure of accuracy of association rules $\bigwedge_{a \in Y}(a = a(u)) \Rightarrow \bigwedge_{a \in X}(a = a(u))$, $u \in U$. In the same way, $H_{\mathbf{A}}(Y)$ could be regarded as the average of supports of the left sides of such rules. By referring to the well known facts, one can see that the entropy based measures are monotonic in the following sense:

**Proposition 6.** *(cf. [5]) Let information system $\mathbf{A} = (U, A)$ and mutually disjoint subsets $X, Y, Z \subseteq A$ be given. We have the following inequalities:*

$$H_{\mathbf{A}}(Y \cup Z) \geq H_{\mathbf{A}}(Y) \qquad H_{\mathbf{A}}(X/Y \cup Z) \geq H_{\mathbf{A}}(X/Y) \tag{19}$$

*Equality in the former above case holds, iff $Y$ determines $Z$.*[5] *Equality in the latter above case holds, iff $Y$ makes $X$ independent from $Z$.*

[5] Within the rough set theory ([6]), one says that $Y$ determines $Z$, iff for each pair of objects $u, u' \in U$, if $Z(u) \neq Z(u')$, then $Y(u) \neq Y(u')$.

Let us consider a decision table $\mathbf{A} = (U, A \cup \{d\})$ and denote by $H_{\mathbf{A}}(d/B)$ conditional entropy of decision $d$ under the subset of conditional attributes $B \subseteq A$. Proposition 5 enables to express the above quantity as follows:

**Proposition 7.** *([15]) Let $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We have:*

$$H_{\mathbf{A}}(d/B) = -\frac{1}{|U|} \sum_{u \in U} \log_2 \mu^B_{X_{d(u)}}(u) \tag{20}$$

**Proposition 8.** *([15]) Let $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. Then $B$ $\mu$-preserves $d$, iff*

$$H_{\mathbf{A}}(d/B) = H_{\mathbf{A}}(d/A) \tag{21}$$

*Moreover, $B$ is a $\mu$-decision reduct, iff it satisfies (21) and for each $a \in B$ there is $H_{\mathbf{A}}(d/B) > H_{\mathbf{A}}(d/B \setminus \{a\})$.*

### 3.2 Approximate independence

Entropy encodes the notions corresponding to probabilistic conditional independence. In particular, it enables to rewrite the definition of a $\mu$-decision reduct in terms of real quantities derived from data. Below we recall the notion of entropy based approximate conditional independence introduced in [13]. It enables to weaken criterion of the PCI by tuning the approximation parameter $\varepsilon \in [0, 1)$.

**Definition 9.** Let $\varepsilon \in [0, 1)$, $\mathbf{A} = (U, A)$ and mutually disjoint subsets $X, Y, Z \subseteq A$ be given. We say that $Y$ makes $X$ $\varepsilon$-approximately independent from $Z$, iff

$$H_{\mathbf{A}}(X/Y) + \log_2(1 - \varepsilon) \leq H_{\mathbf{A}}(X/Y \cup Z) \tag{22}$$

We will call the fact of such an independence as an $\varepsilon$-approximate conditional independence statement ($\varepsilon$-PCI-statement, in short), and denote it by $I^{\varepsilon}_{\mathbf{A}}(X/Y/Z)$.

The way of involving $\varepsilon \in [0, 1)$ into inequality (22) provides us with the following properties:

**Proposition 9.** *Let $\mathbf{A} = (U, A)$ and mutually disjoint subsets $X, Y, Z \subseteq A$ be given. Then:*

$$\begin{aligned} &I^0_{\mathbf{A}}(X/Y/Z) \Leftrightarrow I_{\mathbf{A}}(X/Y/Z) \qquad \exists_{\varepsilon \in [0,1)} I^{\varepsilon}_{\mathbf{A}}(X/Y/Z) \\ &\forall_{\varepsilon, \varepsilon' \in [0,1)} [(\varepsilon \leq \varepsilon') \Rightarrow (I^{\varepsilon}_{\mathbf{A}}(X/Y/Z) \Rightarrow I^{\varepsilon'}_{\mathbf{A}}(X/Y/Z))] \end{aligned} \tag{23}$$

*Proof.* Because of the lack of space, let us focus on the first part. If $\varepsilon = 0$, then (22) takes the form of $H_{\mathbf{A}}(X/Y) \leq H_{\mathbf{A}}(X/Y \cup Z)$. However, due to Proposition 6, we have $H_{\mathbf{A}}(X/Y) \geq H_{\mathbf{A}}(X/Y \cup Z)$, with equality iff $I_{\mathbf{A}}(X/Y/Z)$. Hence: $I^0_{\mathbf{A}}(X/Y/Z)$ iff $H_{\mathbf{A}}(X/Y) = H_{\mathbf{A}}(X/Y \cup Z)$ iff $I_{\mathbf{A}}(X/Y/Z)$.

One can see that $\varepsilon$-approximate conditional independence generalizes the notions corresponding to PCI. It also satisfies similar properties. The analogy of the following result can be found e.g. in [8], for the exact case of $\varepsilon = 0$.

**Proposition 10.** *([16]) Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U, A)$ and mutually disjoint subsets $X, Y, Z, W \subseteq A$ be given. The following laws are satisfied:*

$$\begin{gathered} I^{\varepsilon}_{\mathbf{A}}(X/Y/Z) \Rightarrow I^{\varepsilon}_{\mathbf{A}}(Z/Y/X) \\ I^{\varepsilon}_{\mathbf{A}}(X/Y \cup Z/W) \wedge I^{\varepsilon}_{\mathbf{A}}(X/Y/Z) \Rightarrow I^{\varepsilon(2-\varepsilon)}_{\mathbf{A}}(X/Y/Z \cup W) \\ I^{\varepsilon}_{\mathbf{A}}(X/Y/Z \cup W) \Rightarrow I^{\varepsilon}_{\mathbf{A}}(X/Y/Z) \wedge I^{\varepsilon}_{\mathbf{A}}(X/Y \cup Z/W) \end{gathered} \tag{24}$$

The above result implies that probabilistic independence satisfies the axioms of so called *semi-graphoids* – the theory being developed in purpose of the graph based reasoning about dependencies among variables ([8]). Behaviour of the degrees of approximation in (24) enables to regard Definition 9 as providing a dynamically stable model of the semi-graphoid based inference.

Finally, let us consider the task of searching for subsets approximately preserving information about a distinguished decision in a decision table.

**Definition 10.** Let $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We say that $B$ $\varepsilon$-approximately preserves rough membership information about $d$ ($\varepsilon$-approximately $\mu$-preserves $d$, in short), iff $I^{\varepsilon}_{\mathbf{A}}(d/B/A \setminus B)$ holds, i.e.:

$$H_{\mathbf{A}}(d/B) + \log_2(1-\varepsilon) \leq H_{\mathbf{A}}(d/A) \tag{25}$$

We say that $B$ is an $\varepsilon$-approximate rough membership decision reduct ($\varepsilon$-approximate $\mu$-decision reduct, in short), iff it satisfies (25) and none of its proper subsets does it.

**Proposition 11.** *Let $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given.*

1. *$B$ 0-approximately $\mu$-preserves $d$, iff it $\mu$-preserves $d$.*
2. *There exists $\varepsilon \in [0,1)$ such that $B$ $\varepsilon$-approximately $\mu$-preserves $d$.*
3. *For $\varepsilon, \varepsilon' \in [0,1)$ such that $\varepsilon \leq \varepsilon'$, if $B$ $\varepsilon$-approximately $\mu$-preserves $d$, then it also does it $\varepsilon'$-approximately.*

By analogy to the case of exact conditional independence, let us propose to treat $\varepsilon$-approximate $\mu$-decision reducts as *$\varepsilon$-approximate Markov boundaries of $d$*. The following is an extension of classical theorem cited in Section 2:

**Theorem 3.** *([15]) Let $\varepsilon \in [0,1)$ be given. The problem of finding minimal $\varepsilon$-approximate $\mu$-decision reduct for a given decision table is NP-hard.*[6]

The above result is also true for other optimization criteria than simply minimization of elements of a reduct-boundary (cf. [15]). Nevertheless, one can deal with such tasks by adapting techniques developed in [9], devoted to searching for the *exact decision reducts* $B \subseteq A$ within *consistent decision tables* $\mathbf{A} = (U, A \cup \{d\})$, where $A$ determines $d$ ([6]).[7]

[6] As a corollary, the problem of finding minimal $\varepsilon$-approximate Markov boundary for a random variable within a product probabilistic space is NP-hard as well.

[7] $\varepsilon$-approximate $\mu$-decision reducts can take non-trivial form also in case of consistent decision tables. It can be easily shown that $\mathbf{A} = (U, A \cup \{d\})$ is consistent,

### 3.3 Approximate bayesian networks

BNs encode knowledge about PCI-statements. According to Theorem 2, they can be extracted by searching for subsets $\mu$-preserving decisions along a previously fixed ordering over attributes. Analogous methodology can be formulated for approximate PCI-statements.

**Definition 11.** Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U,A)$ and DAG $\mathcal{D} = (A, \overrightarrow{E})$ be given. $\mathcal{D}$ is called an $\varepsilon$-approximate bayesian network for $\mathbf{A}$ ($\varepsilon$-BN, in short), iff for any mutually disjoint subsets $X, Y, Z \subseteq A$, if $\langle X/Y/Z \rangle_{\mathcal{D}}$, then $I^{\varepsilon}_{\mathbf{A}}(X/Y/Z)$.

The most important challenge is to establish characteristics enabling extraction of $\varepsilon$-BNs from data. In [13–16] the following approach was developed:

**Definition 12.** Let $\mathbf{A} = (U,A)$ and DAG $\mathcal{D} = (A, \overrightarrow{E})$ be given. By entropy of $\mathcal{D}$ we mean quantity

$$H_{\mathbf{A}}(\mathcal{D}) = \sum_{a \in A} H_{\mathbf{A}}(a/\pi_{\mathcal{D}}(a)) \tag{26}$$

**Definition 13.** Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U,A)$ and DAG $\mathcal{D} = (A, \overrightarrow{E})$ be given. We say that $\mathcal{D}$ is $\varepsilon$-consistent with $\mathbf{A}$, iff the following inequality holds:

$$H_{\mathbf{A}}(\mathcal{D}) + \log_2(1-\varepsilon) \leq H_{\mathbf{A}}(A) \tag{27}$$

**Theorem 4.** *([15,16]) Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U,A)$ and DAG $\mathcal{D} = (A, \overrightarrow{E})$ be given. $\mathcal{D}$ is an $\varepsilon$-BN for $\mathbf{A}$, iff it is $\varepsilon$-consistent with $\mathbf{A}$.*

The above result provides an efficient procedure for checking, whether a given DAG $\mathcal{D} = (A, \overrightarrow{E})$ is an $\varepsilon$-approximate bayesian network – it is enough to calculate $H_{\mathbf{A}}(\mathcal{D})$ and verify the validity of condition (27).

**Proposition 12.** *Let $\mathbf{A} = (U,A)$ and DAG $\mathcal{D} = (A, \overrightarrow{E})$ be given.*

1. *$\mathcal{D}$ is 0-BN, iff it is BN for $\mathbf{A}$.*
2. *There exists $\varepsilon \in [0,1)$ such that $\mathcal{D}$ is $\varepsilon$-BN for $\mathbf{A}$.*
3. *Given $\varepsilon, \varepsilon' \in [0,1)$ such that $\varepsilon \leq \varepsilon'$, if $\mathcal{D}$ is $\varepsilon$-BN, then $\mathcal{D}$ is $\varepsilon'$-BN for $\mathbf{A}$.*

*Proof.* Just like in case of Proposition 9, let us focus on the first part. Let us notice that $\mathcal{D} = (A, \overrightarrow{E})$ can be identified with a partial ordering over $A$ induced by $\overrightarrow{E}$. Thus, it must be consistent with some linear ordering over $A$. Without the loss of generality, let us assume that this ordering takes the form of $A = \langle a_1, \ldots, a_n \rangle$, $n = |A|$. One can see that

$$H_{\mathbf{A}}(A) = H_{\mathbf{A}}(\{a_1\}) + H_{\mathbf{A}}(\{a_2\}/\{a_1\}) + \ldots + H_{\mathbf{A}}(\{a_n\}/\{a_1, \ldots, a_{n-1}\}) \tag{28}$$

iff $H_{\mathbf{A}}(d/A) = 0$. Hence, if $\mathbf{A}$ is consistent, then condition (25) can be rewritten as $H_{\mathbf{A}}(d/B) \leq -\log_2(1-\varepsilon)$. It can be satisfied by subsets of attributes with potentially less elements than in case of the exact decision reducts. One can regard such subsets as "*almost*" determining decision (cf. [12])

The consistence of $\overrightarrow{E}$ with the above linear ordering means that for each $i = 1, \ldots, n$ we have inclusion $\pi_{\mathcal{D}}(a_i) \subseteq \{a_1, \ldots, a_{i-1}\}$. Hence, by combining (26) and (28) with Proposition 6, we obtain

$$H_{\mathbf{A}}(\mathcal{D}) \geq H_{\mathbf{A}}(A) \tag{29}$$

where equality holds, iff $I_{\mathbf{A}}(\{a_i\}/\pi_{\mathcal{D}}(a_i)/\{a_1, \ldots, a_{i-1}\} \setminus \pi_{\mathcal{D}}(a_i))$ holds for each $i = 1, \ldots, n$, i.e., according to Theorem 2, iff $\mathcal{D}$ is BN for $\mathbf{A}$. For $\varepsilon = 0$ condition (27) takes the form of $H_{\mathbf{A}}(\mathcal{D}) \leq H_{\mathbf{A}}(A)$, which, according to (29), is equivalent to $H_{\mathbf{A}}(\mathcal{D}) = H_{\mathbf{A}}(A)$. Hence, one can see that DAG $\mathcal{D} = (A, \overrightarrow{E})$ is 0-approximately consistent with $\mathbf{A}$ iff $H_{\mathbf{A}}(\mathcal{D}) = H_{\mathbf{A}}(A)$ iff $\mathcal{D}$ is BN for $\mathbf{A}$.

Let us now generalize Theorem 2 in a more straightforward way:

**Theorem 5.** *Let* $\mathbf{A} = (U, A)$, $A = \langle a_1, \ldots, a_n \rangle$, $n = |A|$, *and* $\varepsilon, \varepsilon_1, \ldots, \varepsilon_n \in [0, 1)$ *such that*

$$\varepsilon \geq 1 - (1 - \varepsilon_1) \cdot \ldots \cdot (1 - \varepsilon_n) \tag{30}$$

*be given. Assume that for each* $\mathbf{A}_i$ *a subset* $B_i \subseteq \{a_1, \ldots, a_{i-1}\}$, *which* $\varepsilon_i$*-approximately* $\mu$*-preserves* $a_i$ *in* $\mathbf{A}_i$, *is given. Then DAG* $\mathcal{D} = (A, \overrightarrow{E})$ *defined by (12) is an* $\varepsilon$*-BN for* $\mathbf{A}$. *Moreover, each* $\varepsilon$*-BN can be obtained in this way.*

*Proof.* Consider DAG $\mathcal{D} = (A, \overrightarrow{E})$, constructed as described above, consistent with $A = \langle a_1, \ldots, a_n \rangle$. For each $i = 1, \ldots, n$, there is inequality

$$H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) + \log_2(1 - \varepsilon_i) \leq H_{\mathbf{A}}(a_i/\{a_1, \ldots, a_{i-1}\}) \tag{31}$$

because $\pi_{\mathcal{D}}(a_i)$ is assumed to be such a subset of $\{a_1, \ldots, a_{i-1}\}$ that $\varepsilon_i$-approximately $\mu$-preserves $a_i$ in $\mathbf{A}_i$. Hence, if (30) holds, then we have:

$$\begin{array}{l} H_{\mathbf{A}}(\mathcal{D}) + \log_2(1 - \varepsilon) \leq H_{\mathbf{A}}(\mathcal{D}) + \sum_i \log_2(1 - \varepsilon_i) = \\ \sum_i \left(H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) + \log_2(1 - \varepsilon_i)\right) \leq \sum_i H_{\mathbf{A}}(a_i/\{a_1, \ldots, a_{i-1}\}) = H_{\mathbf{A}}(A) \end{array} \tag{32}$$

One can easily see that any $\varepsilon$-BN can be obtained in this way.

Let us finally focus on an example of application of once constructed $\varepsilon$-BN. The following task is analogous to *Problem* 1 considered in Subsection 2.3. Let us assume that for a given $\mathbf{A} = (U, A \cup \{d\})$ it is impossible to construct a BN $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ enabling efficient reduction of irrelevant features, as stated in Proposition 4. A solution is to weaken requirements for irrelevancy and to try to extract less accurate but – according to Proposition 12 – potentially smaller $\varepsilon$-BN:

*Problem 2.* Given $\mathbf{A} = (U, A \cup \{d\})$ and $\varepsilon \in [0, 1)$, construct a procedure being able to assign to each input $B \subseteq A$ possibly small output $C \subseteq B$, such that $\varepsilon$-PCI-statement $I^{\varepsilon}_{\mathbf{A}}(\{d\}/C/B \setminus C)$ is satisfied.

**Proposition 13.** *Let* $\mathbf{A} = (U, A \cup \{d\})$, $B \subseteq A$ *and* $\varepsilon$*-BN* $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ *with the root in* $d$ *be given. Then* $I^{\varepsilon}_{\mathbf{A}}(\{d\}/bound_{\mathcal{D}}(B)/B \setminus bound_{\mathcal{D}}(B))$ *holds.*

Operating with smaller $\varepsilon$-BNs enables to derive smaller $\mathcal{D}$-boundaries and make the classification process more effective. Intuitively, we should follow this strategy unless the approximation threshold becomes *too high*.

## 4 Conclusions

We described a connection between the rough set and probabilistic approaches to data analysis. We discussed the notion of the entropy based approximate conditional independence. We generalized the notion of a bayesian network in purpose of dealing with approximate independence statements.

**Acknowledgements:** This paper is supported by the grant of Polish National Committee for Scientific Research (KBN) No. 8T11C02519.

## References

1. Bouckaert R.R.: Properties of Bayesian Belief Network Learning Algorithms. In: Proc. of UAI'94, Morgan Kaufmann, U.S., 1994, 102-109.
2. Buntine W.: A guide to the literature on learning probabilistic networks from data. IEEE Transactions on Knowledge and Data Engineering, 1996.
3. Chickering D.M., Geiger D., Heckerman D.E.: Learning Bayesian Networks is NP-Hard. Microsoft Research Technical Report MSR-TR-94-17, 1994.
4. Duentsch I., Gediga G.: Uncertainty measures of rough set prediction. Artificial Intelligence **106**, 1998, 77-107.
5. Gallager R.G.: Information Theory and Reliable Communication. Wiley, 1968.
6. Pawlak Z.: Rough sets - Theoretical aspects of reasoning about data. Kluwer, 1991.
7. Pawlak Z., Skowron A.: Rough membership functions. In: Advances in the Dempster Shafer Theory of Evidence. Wiley, 1994, 251-271.
8. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, 1988.
9. Polkowski L., Skowron A. (eds.): Rough Sets in Knowledge Discovery. Physica-Verlag, 1998, parts **1**, **2**.
10. Rissanen J.: Minimum-description-length principle. In: S. Kotz, N.L. Johnson (eds.), Encyclopedia of Statistical Sciences. Wiley, 1985, 523-527.
11. 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, 311-362.
12. Ślęzak D.: Approximate reducts in decision tables. In: Proc. of IPMU'96. Spain, 1996, 1159-1164.
13. Ślęzak D.: Foundations of Entropy-Based Bayesian Networks. In: Proc. of IPMU'00. Spain, 2000, 248-255.
14. Ślęzak D.: Data Models based on Approximate Bayesian Networks. In: Proc. of JSAI RSTGC'2001. Japan, 2001, 89-92.
15. Ślęzak D.: Approximate decision reducts (In Polish). Ph.D. thesis, Institute of Mathematics, Warsaw University, 2001.
16. 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, 313-326.
17. Ślęzak D., Wróblewski J.: Order-based genetic algorithms for extraction of approximate bayesian networks from data. In: Proc. of IPMU'02. France, 2002.
18. Ślęzak D., Wróblewski J.: Approximate bayesian network classifiers. In: Proc. of RSCTC'02. U.S., 2002.