Approximate Bayesian Networks

Dominik Ślęzak[1,2]

[1] Polish-Japanese Institute of Information Technology
Koszykowa 86, 02–008 Warsaw, Poland
[2] Institute of Mathematics, Warsaw University
Banacha 2, 02–097 Warsaw, Poland

Abstract. We introduce the notion of an approximate Bayesian network, which almost keeps the information entropy of data and encodes knowledge about approximate dependencies between features. Presented theoretical results, as well as relationships to fundamental concepts of the rough set theory, provide a novel methodology of applying the Bayesian net models to the real life data analysis.

1 Introduction

Bayesian network (BN) is a directed acyclic graph (DAG) designed to encode knowledge about conditional independence statements between considered variables, within a given probabilistic space ([1,2,10]). Roughly speaking, the power of such an encoding increases while removing DAG's edges, unless it causes a loss of control of exactness of derivable statements. BN-structures can be also used to model data and the flow of information while reasoning about new cases by analogy to records gathered in data tables (cf. [18,19]).

Classical BN corresponds to the notion of exact probabilistic independence, which is too accurate while mining real life data, because of the risk of possible noises or fluctuations. Thus, one needs a methodology of generalization of fundamental concepts and results concerning Bayesian networks, to let them deal with appropriately defined approximate independence statements.

The idea of basing such a generalization onto the rough set framework ([7]) originates in the fact that it provides a wide range of tools for expressing data inconsistency, in particular, those related to frequency-based rough membership functions ([8,9]). The notion of a rough membership decision reduct (cf. [13,14,16,17]) corresponds to the probabilistic notion of a Markov boundary, crucial for effective construction of BN-models ([10]). Various criteria of the reduction of noises and redundant information by the approximate preserving of rough membership information ([12–14,16,17,20]) can be thus used to approximate the concepts related to Markov boundaries and Bayesian networks.

We focus on approximations dedicated to the information measure of entropy ([3,4,6,15]), by letting a reasonably small increase of its quantity while reducing conditions (cf. [14,16,18,19]). After recalling the basics of data analysis in Section 2, we thus introduce the notion of an entropy-based approximate decision reduct in Section 3. In Section 4 we consider possibilities of using the approximate decision reduct framework for searching for Bayesian data models (cf. [3,8,10,15,19]). In Section 5 we introduce the entropy-based notion of an approximate conditional independence, which generalizes the classical probabilistic model and its properties providing a kind of graphical representation of probabilistic information ([10,11]). In Section 6 we recall the notions related to Bayesian networks and show how to generalize them onto the case of the analysis of approximate dependencies between features. Section 7 contains the proof of the main result concerning approximate Bayesian networks and Section 8 concludes the paper with final remarks.

2 Frequencies in Data

Data can be represented 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 all possible values on a ([7]). Let us write $A = \langle a_1, \ldots, a_n \rangle$ according to some ordering over the set of attributes. For any $B \subseteq A$, one can consider information function $B : U \to V_B^U$, which labels objects $u \in U$ with vectors $B(u) = \langle a_{i_1}(u), \ldots, a_{i_m}(u) \rangle$, where values of successive $a_{i_j} \in B$, $j = 1, \ldots, m$, occur due to the ordering assumed on A. The set $V_B^U = \{B(u) : u \in U\}$ gathers all vectors of values on B supported in $\mathbf{A}$.

Reasoning about data can be stated, e.g., as the classification problem concerning a distinguished decision to be predicted under information provided over the rest of attributes. For this purpose, one represents data as a decision table $\mathbf{A} = (U, A \cup \{d\})$, $d \notin A$. To express conditions$\to$decision dependencies, one can use frequencies of occurrence of $v_d \in V_d$ conditioned by $w_B \in V_B^U$, provided by

$$P_{\mathbf{A}}(v_d / w_B) = \frac{|\{u \in U : B(u) = w_B \wedge d(u) = v_d\}|}{|\{u \in U : B(u) = w_B\}|} \tag{1}$$

Then, for a given $\alpha \in [0, 1]$, α-inexact decision rule $(B = w_B) \Rightarrow_\alpha (d = v_d)$ is satisfied iff $P_{\mathbf{A}}(v_d / w_B) \geq \alpha$, i.e., iff for at least $\alpha \cdot 100\%$ of objects $u \in U$ such that $B(u) = w_B$ we have also $d(u) = v_d$. The strength of the rule is provided by quantity $P_{\mathbf{A}}(w_B) = |\{u \in U : B(u) = w_B\}| \, / \, |U|$. It corresponds to the chance that an object $u \in U$will satisfy the rule's left side.

In the rough set literature, frequencies are best known as rough membership functions, introduced in [9] to measure degrees of inclusion of indiscernibility classes into concepts being approximated.

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

$$\mu_B^X(u) = |[u]_B \cap X| \, / \, |[u]_B| \tag{2}$$

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

The general rough set principle of reduction of possibly large amount of redundant conditional information takes here the following form:

Definition 2. Let $\mathbf{A} = (U, A \cup \{d\})$ be given. Let $V_d = \{v_1, \ldots, v_r\}$, where, for each $k = 1, \ldots, r$, $X_k = \{u \in U : d(u) = v_k\}$ is called the k-th decision class. We say that $B \subseteq A$ preserves rough membership information iff

$$\forall_{u \in U} \forall_{k=1,\ldots,r} \left[\mu_B^{X_k}(u) = \mu_A^{X_k}(u)\right] \tag{3}$$

We say that B is a rough membership decision reduct (μ-decision reduct, in short) iff it satisfies (3) and none of its proper subsets does it.

One can easily show the following equivalent forms of (3):

Proposition 1. *Let $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. The following conditions are equivalent:*

- *B preserves rough membership information.*
- *B makes d conditionally independent on $A \setminus B$, in terms of $P_{\mathbf{A}}$.*
- *B satisfies, for each $u \in U$, equality $\mu_{d/B}(u) = \mu_{d/A}(u)$, for*

$$\mu_{d/B}(u) = P_{\mathbf{A}}(d(u)/B(u)) = \mu_B^{X_{d(u)}}(u) \tag{4}$$

Several alternative definitions of a frequency-based decision reduct were proposed within the rough set framework (see e.g. [13,14,17]). We focus on the above one because it clearly emphasizes the analogies between frequency based and rough set based methodologies of data analysis. The following correspondence is of the greatest importance:

*If we treat $P_{\mathbf{A}}$ as the empirical probability distribution spanned over $A \cup \{d\}$, then each μ-decision reduct is actually a **Markov boundary** of d within A, i.e., irreducible subset $B \subseteq A$, which makes d independent on the rest of A.*

3 Approximate Decision Reducts

The notion of a μ-decision reduct enables us to handle inconsistencies but its conditions turn out to be too rigorous with respect to possible noises or fluctuations in real life data. A solution is to set up a numeric measure labeling subsets of conditional attributes with their capability of defining decision in terms of rough membership information. Then we would be likely to focus on subsets, which approximately preserve it under the attribute reduction.

Each $B \subseteq A$ induces in $\mathbf{A} = (U, A \cup \{d\})$ the bunch of $\mu_{d/B}(u)$-inexact decision rules $(B = B(u)) \Rightarrow_{\mu_{d/B}(u)} (d = d(u))$, for successive objects $u \in U$. One can thus measure the quality of B by arithmetic or geometric average of accuracy of those rules, i.e., respectively, by

$$E_{\mathbf{A}}(d/B) = \frac{1}{|U|} \sum_{u \in U} \mu_{d/B}(u) \qquad G_{\mathbf{A}}(d/B) = \sqrt[|U|]{\Pi_{u \in U} \mu_{d/B}(u)} \tag{5}$$

For instance, in [17,20] we consider various generalizations of the following:

Definition 3. Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We say that B preserves (E,ε)-approximately rough membership information iff

$$E_{\mathbf{A}}(d/B) \geq (1-\varepsilon)E_{\mathbf{A}}(d/A) \tag{6}$$

We say that B is an (E,ε)-approximate μ-decision reduct iff it satisfies (6) and none of its proper subsets does it.

Analogous notion can be formulated by replacing (6) with condition

$$G_{\mathbf{A}}(d/B) \geq (1-\varepsilon)G_{\mathbf{A}}(d/A) \tag{7}$$

One can see that $E_{\mathbf{A}}(d/A) = G_{\mathbf{A}}(d/A) = 1$ iff $\mathbf{A}$ is consistent, i.e., iff each $[u]_A$ belongs to some decision class. Then, we can consider subsets $B \subseteq A$, which almost define d, by means of $E_{\mathbf{A}}(d/B) \geq 1-\varepsilon$ or $G_{\mathbf{A}}(d/B) \geq 1-\varepsilon$.

$G_{\mathbf{A}}$ corresponds to the measure of conditional entropy, known from the information theory ([4,6]), occurring in the rough set, statistical and machine learning applications in various forms (cf. [3,5,15,16]). Let $\mathbf{A} = (U, A \cup \{d\})$, where $V_d = \{v_1, \ldots, v_r\}$, and $B \subseteq A$ be given. Let us put

$$H_{\mathbf{A}}(d/B) = \sum_{w_B \in V_B^U} P_{\mathbf{A}}(w_B) h(\langle P_{\mathbf{A}}(v_1/w_B), \ldots, P_{\mathbf{A}}(v_r/w_B)\rangle) \tag{8}$$

where $h : \triangle_{r-1} \to [0,1]$ is defined over the (r-1)-dimensional simplex

$$\triangle_{r-1} = \{s = \langle s[1], \ldots, s[r]\rangle \in [0,+\infty)^r : s[1] + \ldots + s[r] = 1\} \tag{9}$$

by attaching to each probabilistic distribution $s \in \triangle_{r-1}$ its entropy

$$h(s) = -\sum_{k:\, s[k]>0} s[k]\log_2(s[k]) \tag{10}$$

Proposition 2. *For any* $\mathbf{A} = (U, A \cup \{d\})$ *and* $B \subseteq A$, *we have*

$$H_{\mathbf{A}}(d/B) = -\log_2(G_{\mathbf{A}}(d/B)) \tag{11}$$

Proof. $-|U| \cdot H_{\mathbf{A}}(d/B) =$
$= -\sum_{w_B \in V_B^U} |\{u \in U : B(u) = w_B\}| h(\langle P_{\mathbf{A}}(v_1/w_B), \ldots, P_{\mathbf{A}}(v_r/w_B)\rangle)$
$= \sum_{w_B, v_d : P_{\mathbf{A}}(v_d/w_B)>0} |\{u \in U : (B,d)(u) = (w_B, v_d)\}| \log_2(P_{\mathbf{A}}(v_d/w_B))$
$= \sum_{w_B, v_d : P_{\mathbf{A}}(v_d/w_B)>0} |\{u \in U : (B,d)(u) = (w_B, v_d)\}| \log_2(P_{\mathbf{A}}(d(u)/B(u)))$
$= \sum_{u \in U} \log_2(P_{\mathbf{A}}(d(u)/B(u))) = \log_2(\Pi_{u \in U} P_{\mathbf{A}}(d(u)/B(u)))$

By taking the logarithm of both sides of inequality (7), we get the following condition (12), expressing the information approximation in terms of entropy:

Definition 4. Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U, A \cup \{d\})$ and $B \subseteq A$ be given. We say that B preserves (H,ε)-approximately rough membership information iff

$$H_{\mathbf{A}}(d/B) + \log_2(1-\varepsilon) \leq H_{\mathbf{A}}(d/A) \tag{12}$$

We say that B is an (H,ε)-approximate μ-decision reduct iff it satisfies (12) and none of its proper subsets does it.

Let us conclude this part with two important properties of the above notion:

Proposition 3. *The notions of a μ-decision reduct and an $(H,0)$-approximate μ-decision reduct are equivalent.*

Proof. For $\varepsilon = 0$, (12) takes the form of $H_{\mathbf{A}}(d/B) \leq H_{\mathbf{A}}(d/A)$. According to [4], we know that $H_{\mathbf{A}}(d/B) \geq H_{\mathbf{A}}(d/A)$, where equality holds iff B makes d conditionally independent on $A \setminus B$. Thus, one has (12) iff $H_{\mathbf{A}}(d/B) = H_{\mathbf{A}}(d/A)$ iff B preserves rough membership information.

Theorem 1. *The problem of finding minimal (H,ε)-approximate μ-decision reduct is NP-hard, for any constant $\varepsilon \in [0,1)$.*

Proof. In [17] we prove the NP-hardness of the Minimal Graph α-Covering Problem for any $\alpha \in (0,1]$ – a generalization of the classical complexity result for $\alpha = 1$. Then, we show how to construct, for any $\varepsilon \in [0,1)$, the polynomial reduction of the Minimal Graph $\alpha(\varepsilon)$-Covering Problem (for appropriately chosen $\alpha(\varepsilon) \in (0,1]$) to the problem of finding a minimal (E,ε)-approximate μ-decision reduct. Actually, analogous reduction can be constructed also for H instead of E, i.e., the polynomial reduction of the Minimal Graph $\alpha(\varepsilon)$-Covering Problem (for $\alpha(\varepsilon)$ chosen slightly differently than in case of E) to the problem being considered here.

4 Approximate Bayesian Models

Rough set applications to the data analysis are usually based on the search for minimal (inexact) decision rules of various kinds ([12–14,16,17,20]). Recently, some interest on possibilities of combining the rough set framework with the Bayesian techniques arises (see e.g. [8,19]). Bayesian models contain the rules with decision situated at their left side. In general, they are related to the analysis of distribution $P_{\mathbf{A}}(A/d)$, letting an object $u \in U$ be classified as, e.g., having decision value

$$v = \arg\max_{v_d \in V_d} [prior(v_d) P_{\mathbf{A}}(A(u)/v_d)] \tag{13}$$

for $prior : V_d \to [0,1]$ expressing prior probabilities of decision values, their frequencies in observed data, subjective preferences of an expert, etc..

Let us set up an arbitrary ordering $A = \langle a_1, \ldots, a_n \rangle$ and denote by V_i the set of all values of a_i. We can decompose $P_{\mathbf{A}}(A/d)$ by noting that for any supported combination of values $v_d \in V_d$, $v_i \in V_i$, $i = 1, \ldots, n$, one has

$$P_{\mathbf{A}}(v_1, \ldots, v_n/v_d) = \prod_{i=1}^{n} P_{\mathbf{A}}(v_i/v_d, v_1, \ldots, v_{i-1}) \tag{14}$$

The attribute reduction can be now related to sub-component distributions:

Proposition 4. *Let* $\mathbf{A} = (U, A \cup \{d\})$, $A = \langle a_1, \ldots, a_n \rangle$, *be given. Let us assume that for each table* $\mathbf{A}_i = (U, \{d, a_1, \ldots, a_{i-1}\} \cup \{a_i\})$, $i = 1, \ldots, n$, *a* μ*-decision reduct* B_i *has been found. Then, for any given* $u \in U$, *the decision value calculated by (13) is equal to*

$$v = \arg\max_{v_d \in V_d} \left[prior(v_d) \prod_{i:\, d \in B_i} P_{\mathbf{A}}(a_i(u)/v_d, (B_i \setminus \{d\})(u)) \right] \tag{15}$$

Proof. First, let us consider the case of $v_d \in V_d$ such that $P_{\mathbf{A}}(A(u)/v_d) = 0$. Then there is such $i = 1, \ldots, n$ that $P_{\mathbf{A}}(a_i(u)/v_d, a_1(u), \ldots, a_{i-1}(u)) = 0$ and $P_{\mathbf{A}}(v_d, a_1(u), \ldots, a_{i-1}(u)) > 0$. Subset B_i must contain d because otherwise $P_{\mathbf{A}}(a_i(u)/B_i(u)) > 0$. Thus, $P_{\mathbf{A}}(a_i(u)/v_d, (B_i \setminus \{d\}(u))$ occurs in the product in (15) and makes it equal to 0 for the considered v_d.

Now, let us consider $v_d \in V_d$ such that $P_{\mathbf{A}}(A(u)/v_d) > 0$. In this case, the value of the product in (15) is positive. We obtain

$$\frac{\prod_{i=1}^{n} P_{\mathbf{A}}(a_i(u)/v_d, a_1(u), \ldots, a_{i-1}(u))}{\prod_{i:\, d \in B_i} P_{\mathbf{A}}(a_i(u)/v_d, (B_i \setminus \{d\})(u))} = \prod_{i:\, d \notin B_i} P_{\mathbf{A}}(a_i(u)/B_i(u)) \tag{16}$$

what is a positive quantity independent on the choice of v_d. Thus, one can see that for any $v_k, v_l \in V_d$ inequality $P_{\mathbf{A}}(A(u)/v_k) \leq P_{\mathbf{A}}(A(u)/v_l)$ holds iff substitution of v_k to the product in (15) gives not more than in case of v_l.

Obviously, the above equivalence can be considered only over the input vectors occurring in data. In case of combinations not included in V_A^U, it remains to trust into the generalization ability of the classification model related to formula (15). According to the general rough set principle of reduction, one can regard that that ability as dependent on solving the following:

Optimization problem: *For any* $\mathbf{A} = (U, A \cup \{d\})$, *find such ordering* $A = \langle a_1, \ldots, a_n \rangle$ *that* μ*-decision reducts* B_i *for* $\mathbf{A}_i$ *provide minimal* $\sum_{i=1}^{n} |B_i|$.

In [19] we consider also the approximate Bayesian models:

Definition 5. Let $\varepsilon \in [0, 1)$ and $\mathbf{A} = (U, A \cup \{d\})$, $A = \langle a_1, \ldots, a_n \rangle$, be given. We say that collection $\mathcal{B} = \langle B_1, \ldots, B_n \rangle$ of subsets $B_i \subseteq \{d, a_1, \ldots, a_{i-1}\}$, $i = 1, \ldots, n$, is (H, ε)-approximately consistent with $\mathbf{A}$ iff

$$\sum_{i=1}^{n} H_{\mathbf{A}}(a_i/B_i) + \log_2(1 - \varepsilon) \leq H_{\mathbf{A}}(A/d) \tag{17}$$

Condition (17) means that the aggregate entropy of $\mathcal{B}$ approximates the information entropy of distribution $P_{\mathbf{A}}(A/d)$. One can obtain such $\mathcal{B}$ by decomposing $\varepsilon \in [0, 1)$ onto $\varepsilon_1, \ldots, \varepsilon_n \in [0, 1)$, satisfying $(1 - \varepsilon_1) \cdot \ldots \cdot (1 - \varepsilon_n) \geq 1 - \varepsilon$, and setting up components $B_i \in \mathcal{B}$ as (H, ε_i)-approximate μ-decision reducts for particular decision tables $\mathbf{A}_i$. Indeed, then we have $H_{\mathbf{A}}(A/d) =$

$$\begin{array}{l} = \sum_{i=1}^{n} H_{\mathbf{A}}(a_i/d, a_1, \ldots, a_{i-1}) \\ \geq \sum_{i=1}^{n} [H_{\mathbf{A}}(a_i/B_i) + \log_2(1 - \varepsilon_i)] \\ = \sum_{i=1}^{n} H_{\mathbf{A}}(a_i/B_i) + \log_2 [\prod_{i=1}^{n} (1 - \varepsilon_i)] \\ \geq \sum_{i=1}^{n} H_{\mathbf{A}}(a_i/B_i) + \log_2(1 - \varepsilon) \end{array} \tag{18}$$

The choice of $\varepsilon_i \in [0,1)$, $i = 1,\ldots,n$, influences the degree of keeping classification calculations, like e.g. those in (15), close to observed data. On the other hand, by appropriate tuning of approximation coefficients, one can get simplified and more efficient model, which is still reliable enough.

Optimization problem (ε): *For any $\mathbf{A} = (U, A \cup \{d\})$, find such ε-decomposition $\varepsilon_1,\ldots,\varepsilon_n \in [0,1)$ and ordering $A = \langle a_1,\ldots,a_n\rangle$ that (H,ε_i)-approximate μ-decision reducts B_i for $\mathbf{A}_i$ provide minimal $\sum_{i=1}^n |B_i|$.*

Obviously the above problem is substantially more complex than, e.g., that of finding minimal (H,ε)-approximate μ-decision reducts – the effective search for reducts is worth almost nothing unless an "appropriate" strategy of choosing the attribute ordering is provided.

5 Approximate Independence

The property of preserving rough membership information by a given subset of conditional attributes is a special case of the following:

Definition 6. Let $\mathbf{A} = (U, A \cup \{d\})$ and mutually disjoint $X, Y, Z \subseteq A \cup \{d\}$ be given. We say that Y makes X conditionally independent on 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

$$(P_{\mathbf{A}}(w_Y, w_Z) > 0) \Rightarrow (P_{\mathbf{A}}(w_X/w_Y, w_Z) = P_{\mathbf{A}}(w_X/w_Y)) \tag{19}$$

Proposition 5. *Let $\mathbf{A} = (U, A \cup \{d\})$ and mutually disjoint $X, Y, Z \subseteq A \cup \{d\}$ be given. Then Y makes X conditionally independent on Z iff*

$$\forall_{u \in U} \left[P_{\mathbf{A}}(X(u)/Y(u)) = P_{\mathbf{A}}(X(u)/Y(u), Z(u))\right] \tag{20}$$

Just like in the special case of decision tables, a kind of approximate version of independence seems to be necessary. We restrict ourselves to the one based on entropy, although there are also other possibilities (see e.g. [16,18]).

Definition 7. Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U, A \cup \{d\})$ and mutually disjoint $X, Y, Z \subseteq A \cup \{d\}$ be given. We say that Y makes X conditionally (H,ε)-approximately independent on Z iff

$$H_{\mathbf{A}}(X/Y) + \log_2(1-\varepsilon) \le H_{\mathbf{A}}(X/Y \cup Z) \tag{21}$$

Proposition 6. *Let $\varepsilon \in [0,1)$ and $\mathbf{A} = (U, A \cup \{d\})$ be given. For mutually disjoint $X, Y, Z \subseteq A \cup \{d\}$, let us denote by $I_{\mathbf{A}}^{H,\varepsilon}(X/Y/Z)$ the statement that Y makes X conditionally (H,ε)-approximately independent on Z. Let mutually disjoint $X, Y, Z, W \subseteq A \cup \{d\}$ be given. The following rules of reasoning about conditional independence statements are satisfied:*
Symmetry:

$$I_{\mathbf{A}}^{H,\varepsilon}(X/Y/Z) \Rightarrow I_{\mathbf{A}}^{H,\varepsilon}(Z/Y/X) \tag{22}$$

Decomposition:

$$I_{\mathbf{A}}^{H,\varepsilon}(X/Y/Z \cup W) \Rightarrow I_{\mathbf{A}}^{H,\varepsilon}(X/Y/Z) \tag{23}$$

Weak union:

$$I_{\mathbf{A}}^{H,\varepsilon}(X/Y/Z \cup W) \Rightarrow I_{\mathbf{A}}^{H,\varepsilon}(X/Y \cup Z/W) \tag{24}$$

Dynamic contraction:

$$I_{\mathbf{A}}^{H,\varepsilon}(X/Y \cup Z/W) \wedge I_{\mathbf{A}}^{H,\varepsilon}(X/Y/Z) \Rightarrow I_{\mathbf{A}}^{H,\varepsilon(2-\varepsilon)}(X/Y/Z \cup W) \tag{25}$$

Proof. We have $H_{\mathbf{A}}(X/Y) = H_{\mathbf{A}}(X \cup Y) - H_{\mathbf{A}}(Y)$, where

$$H_{\mathbf{A}}(B) = -\frac{1}{|U|}\sum_{u \in U} \log_2 P_{\mathbf{A}}(B(u)) \tag{26}$$

for any $B \subseteq A \cup \{d\}$. Thus, (21) can be rewritten as inequality

$$H_{\mathbf{A}}(X \cup Y \cup Z) \geq H_{\mathbf{A}}(X \cup Y) + H_{\mathbf{A}}(Y \cup Z) - H_{\mathbf{A}}(Y) + \log_2(1-\varepsilon) \tag{27}$$

It is equivalent to both $I_{\mathbf{A}}^{H,\varepsilon}(X/Y/Z)$ and $I_{\mathbf{A}}^{H,\varepsilon}(Z/Y/X)$ – it implies (22). To show (23), let us assume its left side. Then we have

$$H_{\mathbf{A}}(X/Y) + \log_2(1-\varepsilon) \leq H_{\mathbf{A}}(X/Y \cup Z \cup W) \leq H_{\mathbf{A}}(X/Y \cup Z) \tag{28}$$

where the second inequality is provided by [4]. The left side of (24) provides

$$H_{\mathbf{A}}(X/Y \cup Z \cup W) - \log_2(1-\varepsilon) \geq H_{\mathbf{A}}(X/Y) \geq H_{\mathbf{A}}(X/Y \cup Z) \tag{29}$$

Finally, the left side of (25) implies that $H_{\mathbf{A}}(X/Y \cup Z \cup W) \geq$

$$\geq H_{\mathbf{A}}(X/Y \cup Z) + \log_2(1-\varepsilon) \geq H_{\mathbf{A}}(X/Y) + \log_2(1-\varepsilon) + \log_2(1-\varepsilon) \tag{30}$$

what gives the right side, for $2\log_2(1-\varepsilon) = \log_2(2(1-\varepsilon))$.

It can be shown, by generalizing Proposition 3, that the notions of conditional independence and conditional (H,ε)-approximate independence are equivalent for $\varepsilon = 0$. As a consequence, 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 (cf. [10,11]). Moreover, stability of the degrees of approximation in (22–24), as well as the polynomial bound for their aggregation in (25), enable to regard Definition 7 as providing a dynamically stable model of the semi-graphoid-based inference.

6 Approximate Bayesian Networks

Bayesian networks (BN) have the structure of a directed acyclic graph (DAG) $\mathcal{D} = (V, \overrightarrow{E})$, where $\overrightarrow{E} \subseteq V \times V$. The objective of the BN-based methodology is to encode conditional independence statements involving groups of probabilistic variables corresponding to elements of V, in terms of the following graph-theoretic notion ([10]):

Definition 8. Let DAG $\mathcal{D} = (V, \overrightarrow{E})$ and mutually disjoint $X, Y, Z \subseteq V$ 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;

where:

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

Let us formulate the notion of a Bayesian network in terms of data analysis:

Definition 9. Let $\mathbf{A} = (U, A \cup \{d\})$ and DAG $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ be given. We say that $\mathcal{D}$ is a Bayesian net for $\mathbf{A}$ iff for any mutually disjoint $X, Y, Z \subseteq A \cup \{d\}$ such that Y d-separates X from Z this is also true that Y makes X conditionally independent on Z, in terms of distribution induced by $\mathbf{A}$.

Theorem 2. *([10]) Let $\mathbf{A} = (U, A \cup \{d\})$, $A = \langle a_1, \ldots, a_n \rangle$, be given. Let us assume that for each table $\mathbf{A}_i = (U, \{d, a_1, \ldots, a_{i-1}\} \cup \{a_i\})$ a μ-decision reduct B_i is provided. Then DAG $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$, where*

$$\overrightarrow{E} = \bigcup_{i=1}^{n} \{\langle b, a_i \rangle : b \in B_i\} \tag{31}$$

is a Bayesian network for $\mathbf{A}$.

The above construction corresponds to the model of the Bayesian classification considered in Proposition 4. Actually, one can treat DAG defined by (31) as the visualization of the flow and synthesis of information while searching for the frequency-based weights of decision classes. The Optimization Problem considered in Section 4 can be now regarded as corresponding to the extraction of optimal Bayesian networks from data. The quantity of $\sum_{i=1}^{n} |B_i|$ relates to the number of edges in (31). It should be minimized to achieve possible large number of conditional independence statements derivable from the graphical structure of $\mathcal{D}$.

Still, only by basing the process of DAG's creation on an approximate μ-decision reducts, we can expect a structure with the number of edges substantially lower than $n(n+1)/2$. Let us thus recall the notion of an (H, ε)-approximately consistent Bayesian model introduced in Section 4 and adapt it to the current framework as follows:

Definition 10. Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U, A \cup \{d\})$ and DAG $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ be given. We say that $\mathcal{D}$ is (H,ε)-approximately consistent with $\mathbf{A}$ iff

$$\sum_{a \in A \cup \{d\}} H_{\mathbf{A}}(a/\pi_{\mathcal{D}}(a)) + \log(1-\varepsilon) \leq H_{\mathbf{A}}(A \cup \{d\}) \tag{32}$$

where $\pi_{\mathcal{D}}(a) = \{b \in A \cup \{d\} : \langle b, a\rangle \in \overrightarrow{E}\}$ is the set of parents of a in $\mathcal{D}$.

Condition (32) seems to keep the aggregate information induced by $\mathcal{D}$-based local conditional distributions somehow close to that encoded within the whole of $P_{\mathbf{A}}(A \cup \{d\})$. This idea can be compared to that of studying the Bayesian likelihood of a DAG under data observed, proposed in [2]. There, likelihood means the probability $\mathbf{P}(\mathbf{A}/\mathcal{D})$ of obtaining a table with frequency distribution equal to $P_{\mathbf{A}}$ by random sampling based on probabilities $P_{\mathbf{A}}(a/\pi_{\mathcal{D}}(a))$, $a \in A \cup \{d\}$. An application of the Stirling's approximation to the formula for $\mathbf{P}(\mathbf{A}/\mathcal{D})$ derived in [2] leads to the following (cf. [1,6]):

$$-\frac{\log_2(\mathbf{P}(\mathbf{A}/\mathcal{D}))}{|U|} \approx \sum_{a \in A} H_{\mathbf{A}}(a/\pi_{\mathcal{D}}(a)) - H_{\mathbf{A}}(A \cup \{d\}) \tag{33}$$

It encourages to tune $\varepsilon \in [0,1)$ in purpose of searching for the balance between likelihood and generalization abilities of DAG-models determined by condition (32). Still, the fundamental question is whether the closeness (likelihood) understood in terms of (H,ε)-approximate consistency implies somehow the closeness understood in terms the quality of information about dependencies between features.

Definition 11. Let $\varepsilon \in [0,1)$, $\mathbf{A} = (U, A \cup \{d\})$ and DAG $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ be given. We say that $\mathcal{D}$ is an (H,ε)-approximate Bayesian network for $\mathbf{A}$ iff for any mutually disjoint $X, Y, Z \subseteq A \cup \{d\}$ such that Y d-separates X from Z this is also true that Y makes X conditionally (H,ε)-approximately independent on Z.

The following result seems to answer to the question stated above positively. In particular, it generalizes Theorem 2, since any DAG $\mathcal{D}$ built on the basis of μ-decision reducts is $(H,0)$-approximately consistent with a given $\mathbf{A}$, as well as any $(H,0)$-approximate Bayesian network is a Bayesian network.

Theorem 3. *Let $\varepsilon \in [0,1)$ and $\mathbf{A} = (U, A \cup \{d\})$ be given. Each DAG which is (H,ε)-approximately consistent with $\mathbf{A}$ is an (H,ε)-approximate Bayesian network for $\mathbf{A}$.*

7 Proof of Theorem 3

Proof. Let DAG $\mathcal{D} = (A \cup \{d\}, \overrightarrow{E})$ be given. Assume that for some mutually disjoint $X, Y, Z \subseteq A \cup \{d\}$ subset Y d-separates X from Z in $\mathcal{D}$. Let us denote such a d-separation statement by $\langle X/Y/Z\rangle_{\mathcal{D}}$. Consider subsets

$$\begin{aligned} X' &= \{a \in A \cup \{d\} \setminus Y : \langle \{a\}/Y/Z\rangle_{\mathcal{D}}\} \\ Z' &= \{a \in A \cup \{d\} \setminus Y : \langle \{a\}/Y/X\rangle_{\mathcal{D}} \wedge \neg\langle \{a\}/Y/Z\rangle_{\mathcal{D}}\} \end{aligned} \tag{34}$$

One can see that $X \subseteq X'$ and $Z \subseteq Z'$, subsets X', Y, Z' are mutually disjoint and d-separation statement $\langle X'/Y/Z' \rangle_{\mathcal{D}}$ holds. In particular:

1. For any $a \in X'$, we have $\pi_{\mathcal{D}}(a) \subseteq X' \cup Y$, because of implication

$$\langle \{a\}/Y/Z \rangle_{\mathcal{D}} \Rightarrow \forall_{b \in \pi_{\mathcal{D}}(a) \setminus Y} \langle \{b\}/Y/Z \rangle_{\mathcal{D}} \tag{35}$$

2. For any $a \in Z'$, we have $\pi_{\mathcal{D}}(a) \subseteq Z' \cup Y$, because of implications

$$\begin{array}{l} \langle \{a\}/Y/X \rangle_{\mathcal{D}} \Rightarrow \forall_{b \in \pi_{\mathcal{D}}(a) \setminus Y} \langle \{b\}/Y/X \rangle_{\mathcal{D}} \\ \neg \langle \{a\}/Y/Z \rangle_{\mathcal{D}} \Rightarrow \forall_{b \in \pi_{\mathcal{D}}(a) \setminus Y} \neg \langle \{b\}/Y/Z \rangle_{\mathcal{D}} \end{array} \tag{36}$$

3. For any $a \in Y$, we have $\pi_{\mathcal{D}}(a) \subseteq X' \cup Y$ or $\pi_{\mathcal{D}}(a) \subseteq Y \cup Z'$, because if not $\pi_{\mathcal{D}}(a) \subseteq X' \cup Y$, then there exists $b \in \pi_{\mathcal{D}}(a) \setminus Y$ which is not in X', what means $\neg \langle \{b\}/Y/Z \rangle_{\mathcal{D}}$. Due to implications

$$\begin{array}{l} \exists_{b \in \pi_{\mathcal{D}}(a) \setminus Y} \neg \langle \{b\}/Y/Z \rangle_{\mathcal{D}} \Rightarrow \forall_{b \in \pi_{\mathcal{D}}(a) \setminus Y} \neg \langle \{b\}/Y/Z \rangle_{\mathcal{D}} \\ \exists_{b \in \pi_{\mathcal{D}}(a) \setminus Y} \neg \langle \{b\}/Y/Z \rangle_{\mathcal{D}} \Rightarrow \forall_{b \in \pi_{\mathcal{D}}(a) \setminus Y} \langle \{b\}/Y/X \rangle_{\mathcal{D}} \end{array} \tag{37}$$

   we thus know that $\pi_{\mathcal{D}}(a) \setminus Y \subseteq Z'$.

Let $\mathcal{D}$ be (H, ε)-approximately consistent with $\mathbf{A} = (U, A \cup \{d\})$, $\varepsilon \in [0, 1)$. We will prove that $I_{\mathbf{A}}^{H,\varepsilon}(X'/Y/Z')$, i.e., that the following inequality holds:

$$H_{\mathbf{A}}(X' \cup Y \cup Z') \geq H_{\mathbf{A}}(X' \cup Y) + H_{\mathbf{A}}(Y \cup Z') - H_{\mathbf{A}}(Y) + \log_2(1-\varepsilon) \tag{38}$$

Let us set up such an ordering $A \cup \{d\} = \langle a_0, \ldots, a_n \rangle$ that if $(a_i, a_j) \in \overrightarrow{E}$, then $i < j$. For $i = 1, \ldots, n$, let us consider $\varepsilon_i \in [0, 1)$ which satisfies equality

$$H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) + \log_2(1 - \varepsilon_i) = H_{\mathbf{A}}(a_i/A_{i-1}) \tag{39}$$

where $A_{i-1} = \{a_0, \ldots, a_{i-1}\}$. Then inequality $\sum_{i=1}^{n} \log_2(1-\varepsilon_i) \geq \log_2(1-\varepsilon)$ holds, because otherwise $\mathcal{D}$ would not be able to satisfy (32). Let us note that for any $B \subseteq A$ the following is true:

$$H_{\mathbf{A}}(B) = \sum_{i: a_i \in B} H_{\mathbf{A}}(a_i/B \cap A_{i-1}) \tag{40}$$

Thus, $H_{\mathbf{A}}(X' \cup Y \cup Z') =$

$$\begin{array}{l} = \sum_{i: a_i \in X' \cup Y \cup Z'} H_{\mathbf{A}}(a_i/(X' \cup Y \cup Z') \cap A_{i-1}) \\ \geq \sum_{i: a_i \in X' \cup Y \cup Z'} H_{\mathbf{A}}(a_i/A_{i-1}) \\ = \sum_{i: a_i \in X' \cup Y \cup Z'} [H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) + \log_2(1 - \varepsilon_i)] \\ \geq \sum_{i: a_i \in X' \cup Y \cup Z'} H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) + \sum_{i=1}^{n} \log_2(1 - \varepsilon_i) \\ \geq \sum_{i: a_i \in X' \cup Y \cup Z'} H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) + \log_2(1 - \varepsilon) \end{array} \tag{41}$$

By comparing (38) and (41), one can see that it is enough to show that

$$\sum_{i: a_i \in X' \cup Y \cup Z'} H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) \geq H_{\mathbf{A}}(X' \cup Y) + H_{\mathbf{A}}(Y \cup Z') - H_{\mathbf{A}}(Y) \tag{42}$$

For any $i = 0, \ldots, n$, inclusions proved in points 1.,2.,3. can take the form of

$$\begin{array}{l} a_i \in X' \Rightarrow (\pi_{\mathcal{D}}(a) \subseteq (X' \cup Y) \cap A_{i-1}) \\ a_i \in Z' \Rightarrow (\pi_{\mathcal{D}}(a) \subseteq (Y \cup Z') \cap A_{i-1}) \\ a_i \in Y \Rightarrow [(\pi_{\mathcal{D}}(a) \subseteq (X' \cup Y) \cap A_{i-1}) \vee (\pi_{\mathcal{D}}(a) \subseteq (Y \cup Z') \cap A_{i-1})] \end{array} \tag{43}$$

Thus, they imply

$$\begin{array}{l} a_i \in X' \Rightarrow H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) \geq H_{\mathbf{A}}(a_i/(X' \cup Y) \cap A_{i-1}) \\ a_i \in Z' \Rightarrow H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) \geq H_{\mathbf{A}}(a_i/(Y \cup Z') \cap A_{i-1}) \\ a_i \in Y \Rightarrow H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i)) \geq H_{\mathbf{A}}(a_i/(X' \cup Y) \cap A_{i-1}) \\ \qquad +H_{\mathbf{A}}(a_i/(Y \cup Z') \cap A_{i-1}) - H_{\mathbf{A}}(a_i/Y \cap A_{i-1}) \end{array} \tag{44}$$

where third inequality holds because, on the one hand, $H_{\mathbf{A}}(a_i/Y \cap A_{i-1})$ is not less than both $H_{\mathbf{A}}(a_i/(X' \cup Y) \cap A_{i-1})$ and $H_{\mathbf{A}}(a_i/(Y \cup Z') \cap A_{i-1})$, and, on the other hand, we know from (43) that $H_{\mathbf{A}}(a_i/\pi_{\mathcal{D}}(a_i))$ is not smaller than at least one of them. We obtain that the left side of (42) is not less than

$$\begin{array}{l} \sum_{i:a_i \in X'} H_{\mathbf{A}}(a_i/(X' \cup Y) \cap A_{i-1}) + \\ \sum_{i:a_i \in Z'} H_{\mathbf{A}}(a_i/(Y \cup Z') \cap A_{i-1}) + \\ \sum_{i:a_i \in Y} [H_{\mathbf{A}}(a_i/(X' \cup Y) \cap A_{i-1}) + \\ \qquad H_{\mathbf{A}}(a_i/(Y \cup Z') \cap A_{i-1}) - H_{\mathbf{A}}(a_i/Y \cap A_{i-1})] \end{array} \tag{45}$$

which – after re-grouping its components – appears to be equal to

$$\begin{array}{l} \sum_{i:a_i \in X' \cup Y} H_{\mathbf{A}}(a_i/(X' \cup Y) \cap A_{i-1}) + \\ \sum_{i:a_i \in Y \cup Z'} H_{\mathbf{A}}(a_i/(Y \cup Z') \cap A_{i-1}) - \\ \sum_{i:a_i \in Y} H_{\mathbf{A}}(a_i/Y \cap A_{i-1}) = \\ = H_{\mathbf{A}}(X' \cup Y) + H_{\mathbf{A}}(Y \cup Z') - H_{\mathbf{A}}(Y) \end{array} \tag{46}$$

Thus, finally we get (42), what implies that Y makes X' conditionally (H, ε)-approximately independent on Z'. Now, it is enough to recall that $X \subseteq X'$, $Z \subseteq Z'$ and apply the inference rules described in Proposition 6 to obtain the wanted statement $I_{\mathbf{A}}^{H,\varepsilon}(X/Y/Z)$.

8 Conclusions

Introduced notion of an entropy-based approximate Bayesian network reflects the need of dealing with approximate independence statements in case of the real life data analysis. Presented results provide the framework for the efficient extraction and application of approximate BN-models to the data classification and description tasks.

Acknowledgements: Supported by the grants of Polish National Committee for Scientific Research, No. 8T11C02319, 8T11C02417, 8T11C02519.

References

1. Bouckaert, R.R.: Properties of Bayesian Belief Network Learning Algorithms. In: Proc. of UAI'94, University of Washington, Seattle, Morgan Kaufmann, San Francisco, CA (1994) pp. 102–109.
2. Cooper, F.G., Herskovits, E.: A Bayesian Method for the Induction of Probabilistic Networks from Data. In: Machine Learning, **9**, Kluwer Academic Publishers, Boston (1992) pp. 309–347.
3. Duentsch, I., Gediga, G.: Uncertainty measures of rough set prediction. Artificial Intelligence **106** (1998) pp. 77–107.
4. Gallager, R.G.: Information Theory and Reliable Communication. John Wiley & Sons, New York (1968).
5. Kapur, J.N., Kesavan, H.K.: Entropy Optimization Principles with Applications. Academic Press (1992).
6. Li, M., Vitanyi, P.: An Introduction to Kolmogorov Complexity and Its Applications. Springer Verlag (1997).
7. Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer Academic Publishers, Dordrecht (1991).
8. Pawlak, Z.: Decision rules, Bayes' rule and rough sets. In: Proc. of RSFD-GrC'99, Yamaguchi, Japan, LNAI **1711** (1999) pp. 1–9.
9. Pawlak, Z., Skowron, A.: Rough membership functions. In: Advances in the Dempster Shafer Theory of Evidence, John Wiley & Sons Inc., New York, (1994) pp. 251–271.
10. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann (1988).
11. Pearl, J., Paz, A.: Graphoids: A graph-based logic for reasoning about relevance relations. In: Advanves in Artificial Intelligence II, B. Du Boulay, D. Hogg and L. Steels (eds.), North-Holland, Amsterdam (1987) pp. 357–363.
12. Polkowski, L., Skowron, A. (eds.): Proc. of RSCTC'98, June 22–26, Warsaw, Poland, Springer Verlag, Berlin (1998).
13. Polkowski, L., Skowron, A. (eds.): Rough Sets in Knowledge Discovery. Physica Verlag, Heidelberg (1998), parts **1**, **2**.
14. Polkowski, L., Tsumoto, S., Lin, T.Y. (eds.): Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems. Physica Verlag / Springer Verlag (2000).
15. Rissanen, J.: Modeling by the shortest data description. Authomatica **14** (1978) pp. 465–471.
16. Ślęzak, D.: Approximate reducts in decision tables. In: Proc. of IPMU'96, July 1–5, Granada, Spain (1996) **3**, pp. 1159–1164.
17. Ślęzak, D.: Normalized decision functions and measures for inconsistent decision tables analysis. Fundamenta Informaticae **44/3**, IOS Press (2000) pp. 291–319.
18. Ślęzak, D.: Foundations of Entropy-Based Bayesian Networks: Theoretical Results & Rough Set Based Extraction from Data. In: Proc. of IPMU'00, July 3–7, Madrid, Spain (2000) **1**, pp. 248–255.
19. Ślęzak, D.: Data Models based on Approximate Bayesian Networks. In: Proc. of JSAI RSTGC'2001, May 20–22, Shimane, Japan (2001).
20. Ślęzak, D., Wróblewski, J.: Application of Normalized Decision Measures to the New Case Classification. In: Proc. of RSCTC'00, October 16–19, Banff, Canada (2000).