# Approximate Bayesian Networks: Applications and Extraction from Data

Dominik Ślęzak[1], Jakub Wróblewski[1]

**Abstract**
Bayesian network (BN) is a directed acyclic graph encoding probabilistic independence statements between variables. BN with decision attribute as a root can be applied to classification of new cases, by synthesis of conditional probabilities propagated along the edges. We consider approximate BNs, which almost keep entropy of a decision table. They have usually less edges than classical BNs. They enable to model and extend the well-known Naive Bayes approach. Experiments show that classifiers based on approximate BNs can be very efficient. Genetic-based algorithmic framework for extracting approximate BNs from data is proposed.

## 1. Introduction

Bayesian network (BN) is a directed acyclic graph (DAG) designed to encode knowledge about probabilistic conditional independence (PCI) statements between considered variables, within a given probabilistic space [8]. Its expressive power increases while removing the edges, unless it causes a loss of control of exactness of derivable PCI-statements. When mining real-life data, one needs less accurate, approximate criteria of independence. We base such an approximation on the information measure of entropy [5], by letting a reasonably small increase of its quantity during the edge reduction. It leads to approximate BNs corresponding to approximate PCI-statements, introduced in [9,10].
BN can model the flow of information in decision tables, while reasoning about new cases. Necessary probabilities can be calculated directly from training data, by substituting the foregoing decision values in a loop. One can maximize the product of such probabilities and choose the most probable decision value. This is, actually, an example of the bayesian reasoning approach (cf. [2]).
We analyze how the strategies of choosing the approximation threshold and searching for corresponding approximate BNs can influence the new case classification results. We extract optimal DAGs from data in a very basic way, just to provide a material for simulations. In parallel, we consider application of order-based genetic algorithms (o-GAs), which are known to be able to deal with complex NP-hard optimization problems efficiently [4,6].
Exemplary applications of o-GAs to searching for optimal rough-set-based decision reducts were developed in [13,14]. We show that the same framework can be used in case of approximate decision reducts and Markov boundaries. Furthermore, we discuss possibility of its extension in order to cope with directed acyclic graphs inducing approximate BNs. Various other approaches to extraction of classical BNs (cf. [3]) are worth generalizing onto the approximate case as well.
Although BN-related framework can be regarded as purely probabilistic, let us stress its relationship to the rough set theory [7], by means of correspondence between fundamental notions, like e.g. these of decision reduct and Markov boundary (cf. [9,10]), as well as between optimization problems concerning extraction of approximate BNs and rough-set-based models from data (cf. [11,12]).

## 2. Probabilities in information systems

Following [7], we represent data as information systems – tuples $\boldsymbol{A}=(U,A)$. Each attribute $a\in A$ is identified with function $a:U\rightarrow V_a$, for $V_a$ denoting the set of all possible values on $a$. Let us assume ordering $A=<a_1,\ldots,a_n>$. For any $B\subseteq A$, consider $B$-information function, which labels objects $u\in U$ with vectors $<a_{i1}(u),\ldots,a_{im}(u)>$, where values of $a_{ij}\in B$, $j=1,\ldots,m$, occur due to the ordering on $A$. We denote this function by $B:U\rightarrow V_B$, where $V_B=\{B(u): u\in U\}$ is the set of all vectors on $B$ occurring in $\boldsymbol{A}$.
Classification problems concern distinguished decisions to be predicted under information provided over conditional attributes. For this purpose, one represents data as a decision table $\boldsymbol{A}=(U,A\cup\{d\})$, $d\notin A$. One can use various classification methodologies, provided, e.g., by statistical calculus [2]. Occurrence of $v_d\in V_d$ conditioned by $w_B\in V_B$, can be expressed as probability

[1] Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, Poland

$$P_A ( v_d / w_B ) = |\{ u \in U: B(u) = w_B \wedge d(u) = v_d \}| \,/\, |\{ u \in U: B(u) = w_B \}|$$

For a given $\alpha \in [0,1]$, we say that $\alpha$-inexact decision rule $(B=w_B) \Rightarrow_\alpha (d=v_d)$ is satisfied iff $P_A(v_d/w_B) \geq \alpha$, 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 prior probability $P_A(w_B)=|\{u \in U: B(u)=w_B\}| \,/\, |U|$. It corresponds to the chance that an object $u \in U$ will satisfy the rule's left side. One can consider such probabilities not only for the case of a distinguished decision attribute at the right side of a rule. In case of bayesian approaches to the new case classification one uses probabilistic rules with decisions involved in their left sides.

## 3. Probabilistic decision reducts

Each pair $(B,u) \in P(A) \times U$ generates approximate decision rule pointing at the $d(u)$-th decision class. It is described by means of the following parameters:

**Definition:** Let $\boldsymbol{A}=(U,A \cup \{d\})$, $B \subseteq A$ and $u \in U$ be given. By the accuracy and support coefficients for $(B,u)$ we mean, respectively, quantities

$$\mu_{d/B}(u) = P_A(d(u)/B(u)) \qquad \mu_B(u) = P_A(B(u))$$

In the context of the above coefficients, the rough-set-based principle of reduction of redundant information [7] corresponds to the following notion:

**Definition:** Let $\boldsymbol{A}=(U,A \cup \{d\})$ be given. $B \subseteq A$ $\mu$-preserves $d$ iff

$$\forall_{u \in U}\, \mu_{d/B}(u) = \mu_{d/A}(u)$$

$B$ is a $\mu$-decision reduct iff it satisfies the above condition and none of its proper subsets does it.

The above property is an example of a probabilistic conditional independence (PCI) statement. PCI is defined over subsets of variables considered within a discrete product probabilistic space, over all possible configurations of vectors of values. Since we deal with probabilistic distributions derived from information systems, let us focus on the following, equivalent [10] definition:

**Definition:** Let $\boldsymbol{A}=(U,A)$ and $X,Y,Z \subseteq A$ be given. We say that $Y$ makes $X$ independent from $Z$ iff

$$\forall_{u \in U}\, P_A(X(u)/Y(u)) = P_A(X(u)/(Y \cup Z)(u))$$

**Corollary:** Let $\boldsymbol{A}=(U,A \cup \{d\})$ and $B \subseteq A$ be given. $B$ is a $\mu$-decision reduct iff it is a Markov boundary of $d$ within $A$, i.e., it is an irreducible subset, which makes $d$ independent from the rest of $A$.

## 4. Entropy-based approximations

Each $B \subseteq A$ induces in $\boldsymbol{A}=(U,A \cup \{d\})$ the bunch of inexact decision rules $(B=B(u)) \Rightarrow \mu_{d/B}(u)\ (d=d(u))$ for particular objects $u \in U$. One can measure the quality of $B$ in terms of such rules.

**Definition:** Let $\boldsymbol{A}=(U,A \cup \{d\})$ and $B \subseteq A$ be given. We put

$$H_A(B) = -1/|U| \sum_{u \in U} \log_2 \mu_B(u) \quad \text{and} \quad H_A(d/B) = -1/|U| \sum_{u \in U} \log_2 \mu_{d/B}(u)$$

$H_A$ corresponds to the measure of information entropy [5,10]. Let us focus on the following way of approximate preserving of accuracy under the conditional attribute reduction.

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

$$H_A( d/B ) + \log_2( 1 - \varepsilon) \leq H_A( d/A)$$

We say that $B$ is an $\varepsilon$-approximate $\mu$-decision reduct ($\varepsilon$-approximate Markov boundary) iff it satisfies the above condition and none of its proper subsets does it.

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

$$H_A( X / Y) + \log_2 (1- \varepsilon) \leq H_A( X / Y \cup Z )$$

Such a criterion of *approximate* independence is more robust to possible fluctuations in real life data. Moreover, we have equivalence of the notions of independence and 0-approximate independence.

## 5. Bayesian networks

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

**Definition:** Let DAG $D=(A,E)$ and $X,Y,Z \subseteq A$ be given. We say that $Y$ $d$-separates $X$ from $Z$ iff any path between any $x \in X \setminus Y$ and any $z \in Z \setminus Y$ comes through: (1) a serial or diverging connection covered by some $y \in Y$, or (2) a converging connection not covered by $Y$, having no descendant in $Y$.

**Definition:** Let $\boldsymbol{A}=(U,A)$ and DAG $D=(A,E)$ be given. We say that $D$ is a bayesian network (BN) for $\boldsymbol{A}$ iff for any $X,Y,Z \subseteq A$, if $Y$ $d$-separates $X$ from $Z$, then $Y$ makes $X$ conditionally independent from $Z$.

In [10] the following approach to approximation of the notion of BN was proposed:

**Definition:** Let $\varepsilon \in [0,1)$, $\boldsymbol{A}=(U,A)$ and DAG $D=(A,E)$ be given. We say that $D$ is $\varepsilon$-approximately consistent with $\boldsymbol{A}$ iff

$$H_A(D) + \log ( 1 - \varepsilon ) \leq H_A(A)$$

where $H_A(D)=\sum_{a \in A} H_A( a / \{b \in A: < b,a > \in E\} )$.

**Definition:** Let $\varepsilon \in [0,1)$, $\boldsymbol{A}=(U,A)$, $D=(A,E)$ be given. We say that $D$ is an $\varepsilon$-approximate bayesian network ($\varepsilon$-BN) iff for any $X,Y,Z \subseteq A$, if $Y$ $d$-separates $X$ from $Z$, then $Y$ makes $X$ $\varepsilon$-approximately independent from $Z$.

**Theorem:** [10,11] Let $\varepsilon \in [0,1)$ and $\boldsymbol{A}=(U,A)$ be given. Each DAG, which is $\varepsilon$-approximately consistent with $\boldsymbol{A}$ is an $\varepsilon$-approximate BN for $\boldsymbol{A}$.

## 6. BN-based classification

Bayesian decision models are related to the analysis of approximations of distribution $P_A(A(u)/v_d)$. One can let $u \in U$ be classified as having decision value $v_u$ defined by the following law:

$$v_u = \arg\max_{vd} [ \text{prior}(v_d)\, P_A (A(u)/v_d) ]$$

for prior:$V_d \rightarrow [0,1]$. Let us set up $A=<a_1,\ldots,a_n>$ and denote by $V_i$ the set of all values of $a_i$. We decompose $P_A(A/d)$ by noting that for any combination of values $v_d \in V_d$, $v_i \in V_i$, $i=1,\ldots,n$, one has

$$P_A (v_1,\ldots,v_n/v_d) = \prod_{i=1,\ldots,n} P_A (v_i/v_d,v_1,\ldots,v_{i-1})$$

**Proposition:** [10,11] Let $\boldsymbol{A}=(U,A \cup \{d\})$, $A = < a_1, \ldots, a_n >$, be given. Assume that for each table $\boldsymbol{A}_i = (U,\{d,a_1,\ldots,a_{i-1}\} \cup \{a_i\})$, $i=1,\ldots,n$, a $\mu$-decision reduct $B_i$ has been found. For any $u \in U$, we have

$$v_u = \arg\max_{vd} [ \text{prior}(v_d) \prod_{i:\, d \in Bi} P_A (a_i(u)/v_d, (B_i \setminus \{d\})(u))$$

The above way of classifying objects corresponds to BN for $\boldsymbol{A}=(U,A \cup \{d\})$. We obtain a scheme of the bayesian classification, where conditional probabilities are propagated along the DAG structure, beginning with decision as the root.

## 7. Order-based genetic algorithms

In the *order-based genetic algorithm* (o-GA [4]) a chromosome is a permutation. There are various methods of recombination (crossing-over) considered in literature (cf. [13]). In case of a *hybrid algorithm*, its heuristic part is launched in order to calculate fitness value of a chromosome. The heuristic procedure is used to obtain actual results, while genetic part of algorithm optimizes the parameters of heuristic part. As an example, let us present the hybrid algorithm for searching for approximate $\varepsilon$-approximate $\mu$-decision reducts, calculated for specified settings of parameter $\varepsilon \in [0,1)$. It is a generalization of the algorithm developed in [13] for searching for classical rough-set-based decision reducts: (1) Let $B=A$ be the set of all attributes and let $A=<a_1,\ldots,a_n>$ be the ordering consistent with input permutation. Let $\varepsilon \in [0,1)$ be given; (2) For $i=1$ to $n=|A|$ repeat: Let $B \leftarrow B \setminus \{a_i\}$; If $B$ does not $\mu$-preserve $d$ $\varepsilon$-approximately, then $B \leftarrow B \cup \{a_i\}$.

**Proposition:** The result of the above algorithm is an $\varepsilon$-approximate $\mu$-decision reduct. Moreover, for any $\varepsilon$-approximate $\mu$-decision reduct $B$ there exists a permutation such that $B$ is the result.

## 8. Optimal approximate networks

Let us assume that ordering $A=<a_1,...,a_n>$ corresponds to a specified permutation. Let us initiate the structure of $D=(A,E)$ by putting $E = \cup_{i=1,...,n-1} \cup_{j=i+1,...,n} \{<a_i,a_j>\}$. Calculation of fitness should correspond to a strategy of choosing the order of the edge reduction trials. Let $m=n(n-1)/2$, $n=|A|$, for a given $\boldsymbol{A}=(U,A)$. Let permutation $\tau$ over $A$ be given. The scheme of the reduction is the following:

- Let $D$ be complete DAG and let $E=<e_1,...,e_m>$ be any ordering of edges of $D$, consistent with $\tau$.
- For $i$=1 to $m$ repeat: Let $E \leftarrow E \setminus \{e_i\}$; If $D$ is not $\varepsilon$-approximately consistent with $\boldsymbol{A}$, then $E \leftarrow E \cup \{e_i\}$.

Such an approach requires, obviously, an additional procedure for choosing orderings $E =< e_1 ,..., e_m>$ for particular permutations encoded by chromosomes. One can, e.g., apply *greedy* heuristics checking, at each step of reduction, deletion of which remaining edge causes the smallest increase of entropy $H_A(D)$. In general, one has then no guarantee that for given information system $\boldsymbol{A}=(U,A)$ and approximation threshold $\varepsilon \in [0,1)$ there exists permutation over $A$ providing at outcome of the considered algorithm the solution of Problem 2 defined in Section 5. Still, the following holds:

**Proposition:** The result of the above algorithm is always a DAG $D=(A,E)$, which is $\varepsilon$-approximately consistent with $\boldsymbol{A}$. Moreover, it is irreducible, i.e. deletion of any edge (or subset of edges) from $E$ makes criterion of $\varepsilon$-approximate consistency not satisfied any more.

## Acknowledgements

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

## References

1. Bay, S.D.: *The UCI Machine Learning Repository*, http://www.ics.uci.edu/ml
2. Box, G.E.P., Tiao, G.C.: *Bayesian Inference in Statistical Analysis*. Wiley, 1992.
3. Buntine, W.: A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 1996.
4. Davis, L. (ed.): *Handbook of Genetic Algorithms*. Van Nostrand Reinhold, 1991.
5. Kapur, J.N., Kesavan, H.K.: *Entropy Optimization Principles with Applications*, 1992.
6. Michalewicz, Z.: *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, 1994.
7. Pawlak, Z.: *Rough sets – Theoretical aspects of reasoning about data*. Kluwer, 1991.
8. Pearl, J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, 1988.
9. Ślęzak, D.: *Approximate decision reducts* (in Polish). Ph.D. thesis, Institute of Mathematics, Warsaw University, 2001.
10. 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, 2002*.
11. Ślęzak, D., Wróblewski, J.: Order-based genetic algorithms for extraction of approximate bayesian networks from data. In: *Proc. of IPMU'2002*.
12. Ślęzak, D., Wróblewski, J.: Approximate bayesian network classifiers. In: *Proc. of RSCTC'2002*.
13. Wróblewski, J.: Theoretical Foundations of Order-Based Genetic Algorithms. *Fundamenta Informaticae* 28/3-4, IOS Press, 1996, pp. 423-430.
14. Wróblewski, J.: *Adaptive methods of the object classification* (in Polish). Ph.D. thesis. Institute of Mathematics, Warsaw University, Poland, 2001.