Learning Terminological Naïve Bayesian Classifiers Under Different Assumptions on Missing Knowledge

Pasquale Minervini Claudia d'Amato Nicola Fanizzi

Department of Computer Science
University of Bari

URSW 2011 ◇ Bonn, October 23, 2011

Contents

Introduction & Motivations

- Uncertainty is inherently present in real-world knowledge
- In the SW context difficulties arise modeling real-world domains using only purely logical formalisms
- Several approaches for coping with unceratin knowledge have been proposed (probabilistic, fuzzy,...)
  - usually probabilistic information is assumed to be available
  - the CWA is adopted

- Exploiting an already populated ontology, a method capturing probabilistic information could be of help
  - the OWA has to be taken into account

Paper Contributions

Proposal of a **Terminological naïve Bayesian classifier** for *predicting class-membership probabilistically*

- it is a naïve Bayesian network modeling the conditional dependencies between a learned set of Description Logic (complex) concepts and a target concept
- **it deals with the incomplete knowledge due the OWA** by considering different ignorance models:
  - *Missing Completely at Random*
  - *Missing at Random*
  - *Informatively Missing*

Knowledge Base Representation

Assumption: resources, concepts and relationships are defined in terms of a *representation that can be mapped to some DL* language (with the standard model-theoretic semantics)

$\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$

- *T-box* $\mathcal{T}$ is a set of definitions
- *A-box* $\mathcal{A}$ contains extensional assertions on concepts and roles e.g. $C(a)$ and $R(a, b)$
- The set of the individuals (resources) occurring in $\mathcal{A}$ will be denoted $\mathsf{Ind}(\mathcal{A})$

Basics of Bayesian Networks...

- A Bayesian network (BN) is a *DAG* $\mathcal{G}$ representing the *conditional dependencies* in a set of random variables
- Each vertex in $\mathcal{G}$ corresponds to a random variable X_i
- Each edge in $\mathcal{G}$ indicates a *direct influence* relation between the two connected random variables
- A set of *conditional probability distributions* $\theta_{\mathcal{G}}$ is associated with each vertex
- Each vertex X_i in $\mathcal{G}$ is *conditionally independent* of any subset $S \subseteq Nd(X_i)$ of vertices that are not descendants of X_i

...Basics of Bayesian Networks

- The *joint probability distribution* $\Pr(X_1, \ldots, X_n)$ over a set of random variables $\{X_1, \ldots, X_n\}$ is computed as

$$\Pr(X_1, \ldots, X_n) = \prod_{i=1}^{n} \Pr(X_i \mid \mathit{parents}(X_i));$$

- Given a BN, it is possible to evaluate inference queries by marginalization
- To decrease the inference complexity the **naïve Bayes network** is often considered
  - it is assumed that the presence (or absence) of a particular feature (random variable) of a class is unrelated to the presence (or absence) of any other feature, given the class variable (random variable)

Terminological Naïve Bayesian Network: Definition

A Terminological Bayesian Network (TBN) $\mathcal{N}_{\mathcal{K}}$, w.r.t. a DL KB $\mathcal{K}$, is defined as a pair $\langle \mathcal{G}, \Theta_{\mathcal{G}} \rangle$, where:

- $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a directed acyclic graph, in which:
  - $\mathcal{V} = \{F_1, \ldots, F_n, C\}$ is a set of vertices, each F_i representing a DL (eventually complex) concepts defined over $\mathcal{K}$ and C representing a target concept
  - $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is a set of edges, modeling the dependence relations between the elements of $\mathcal{V}$;
- $\Theta_{\mathcal{G}}$ is a set of **conditional probability distributions** (CPD), one for each $V \in \mathcal{V}$, *representing the conditional probability of the feature concept given its parents in the graph*

In the case of a **Terminological Naïve Bayesian Network**,

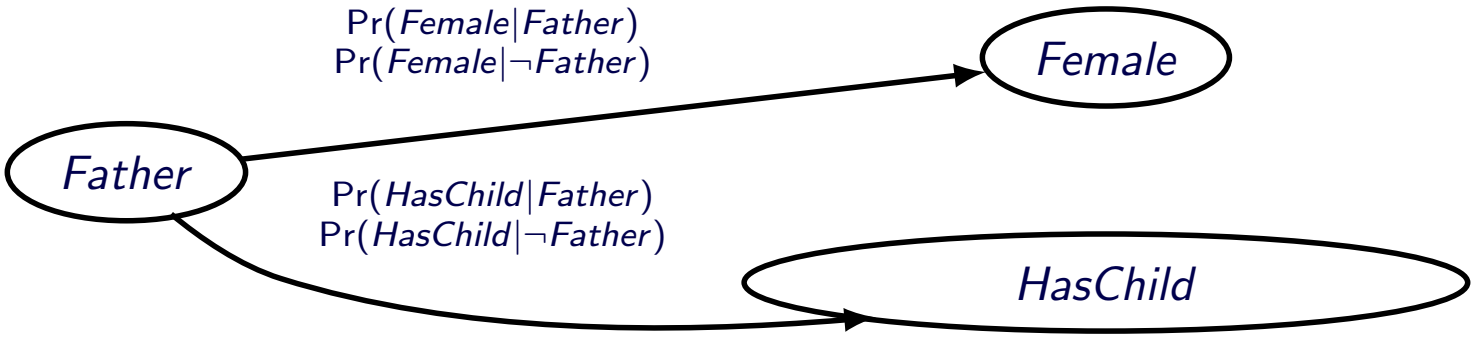
$\mathcal{E} = \{\langle C, F_i \rangle \mid i \in \{1, \ldots, n\}\}$, namely $\forall i, j \in \{1, \ldots, n\}$ and $i \neq j$ F_i is independent of F_j given the value of the target concept

Terminological Naïve Bayesian Network: Example

Given:

- *a set of DL feature concepts*
  $\mathcal{F} = \{Female, HasChild := \exists hasChild.Person\}$ (variable names are used instead of complex feature concepts)
- *a target concept* *Father*

the **Terminological Naïve Bayesian Network** is:



Learning a TBN: Problem Definition

Given:

- a *target concept* C
- a DL KB $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$
- the *sets of of positive, negative and neutral examples* for C, denoted with $Ind_C^+(\mathcal{A})$, $Ind_C^-(\mathcal{A})$ and $Ind_C^0(\mathcal{A})$, so that:
  - $\forall a \in Ind_C^+(\mathcal{A}) : \mathcal{K} \models C(a)$,
  - $\forall a \in Ind_C^-(\mathcal{A}) : \mathcal{K} \models \neg C(a)$,
  - $\forall a \in Ind_C^0(\mathcal{A}) : \mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a)$;
- an *ignorance model*
- a *scoring function* **score** *for a TBN* $\mathcal{N}_\mathcal{K}$ *w.r.t.* $Ind_C(\mathcal{A})$

Find:

a network $\mathcal{N}_\mathcal{K}^*$ maximizing the scoring function

$$\mathcal{N}_\mathcal{K}^* \leftarrow \arg\max_{\mathcal{N}_\mathcal{K}} \; score(\mathcal{N}_\mathcal{K}, Ind_C(\mathcal{A})))$$

TBN: the Learning Algorithm...

function $learn(\mathcal{K}, Ind_C(\mathcal{A}))$

{The TBN is initialized as containing only the target concept node}
$\mathcal{N}^*_{\mathcal{K}} = \langle \mathcal{G}, \Theta_{\mathcal{G}} \rangle$; $\quad \mathcal{G} = \langle \mathcal{V} \leftarrow \{C\}, \mathcal{E} \leftarrow \emptyset \rangle$;
$\mathcal{N}_{\mathcal{K}} \leftarrow \emptyset$;
repeat
$\mathcal{N}_{\mathcal{K}} \leftarrow \mathcal{N}^*_{\mathcal{K}}$;
{A new network is created, having one more node and different parameters than the previous one}
$Network = \langle c', \mathcal{N}'_{\mathcal{K}}, s' \rangle \leftarrow extend(\mathcal{N}_{\mathcal{K}}, Ind_C(\mathcal{A}))$;
$\mathcal{N}^*_{\mathcal{K}} \leftarrow \mathcal{N}'_{\mathcal{K}}$;
{Possible stopping conditions: *a*) *improvements in score below a threshold*; *b*) *reaching a maximum number of nodes*}
until stopping criterion on *Network*;
return $\mathcal{N}_{\mathcal{K}}$;

...TBN: the Learning Algorithm

function $extend(\mathcal{N}_{\mathcal{K}}, Ind_C(\mathcal{A}))$

$Concept \leftarrow Start$; $Best \leftarrow \emptyset$;

repeat

$Concepts \leftarrow \emptyset$;

for $c' \in \{c' \in \rho^{cl}_{\downarrow}(Concept) \mid |c'| \leq min(|c| + d, maxLen)\}$ **do**

$\mathcal{V}' \leftarrow \mathcal{V} \cup \{c'\}$;

$\mathcal{N}'_{\mathcal{K}} \leftarrow optimalNetwork(\mathcal{V}', Ind_C(\mathcal{A}))$;

$s' \leftarrow score(\mathcal{N}'_{\mathcal{K}}, Ind_C(\mathcal{A}))$;

$Concepts \leftarrow Concepts \cup \{\langle c', \mathcal{N}'_{\mathcal{K}}, s' \rangle\}$;

end for

$Best \leftarrow \arg\max_{\langle c', \mathcal{N}'_{\mathcal{K}}, s' \rangle \in Concepts \cup \{Best\}} s'$;

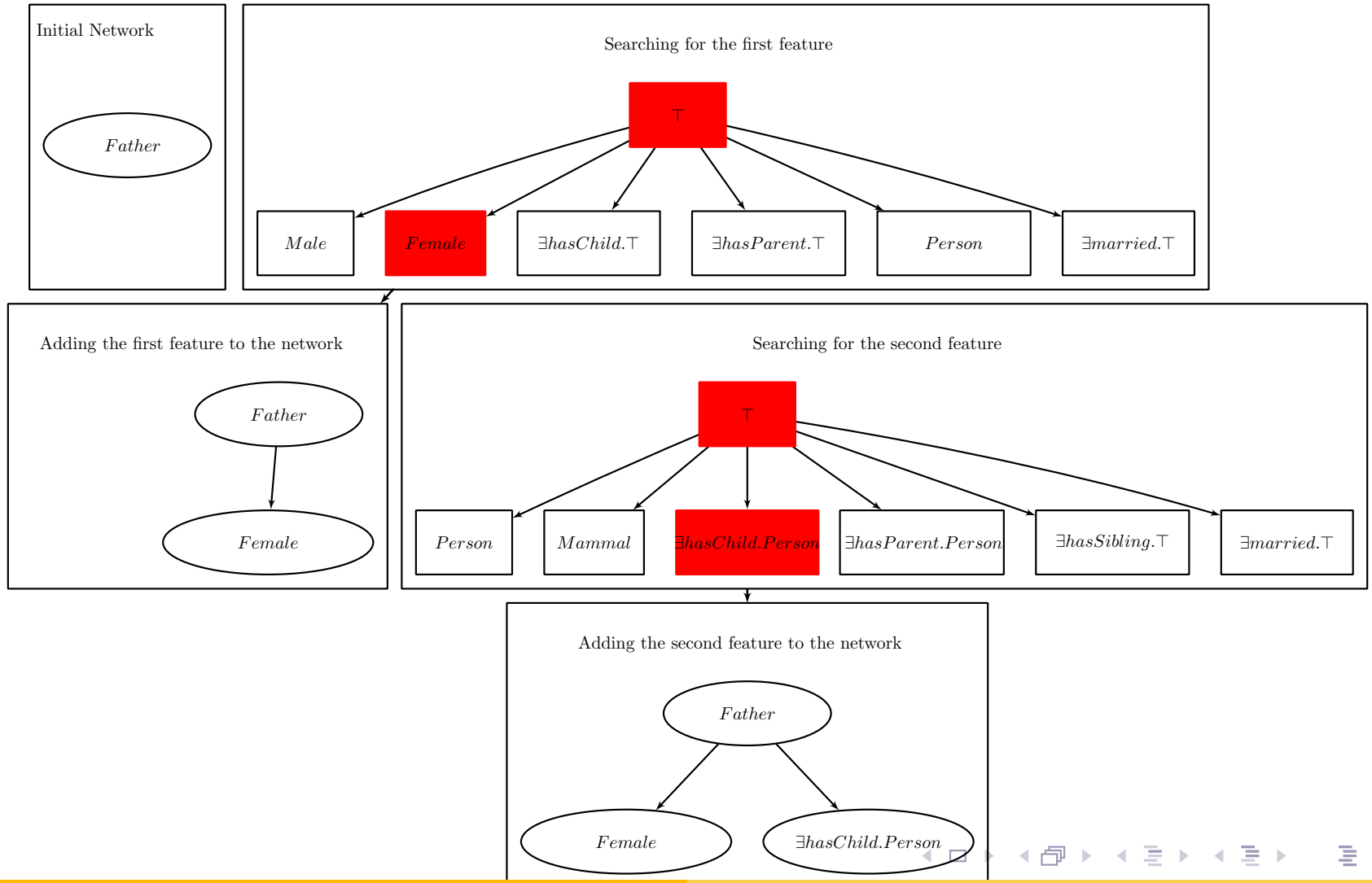
$Concept \leftarrow c : \langle c, \mathcal{N}_{\mathcal{K}}, s \rangle = Best$;

{Possible stopping conditions: *a*) exceeding a *maximum number of iterations*; *b*) exceeding a *maximum number of refinement steps*}

until Stopping criterion on *Best*;

return *Best*;

Learning a Naïve TBN: Example



The ignorance models

To learn the TBN, different assumptions (ignorance models) *on the nature of the missing information are considered*, given an ideal KB $\mathcal{K}^*$ having additional knowledge:

- **MCAR** (Missing Completely At Random) – the probability that $a \in C^{\mathcal{I}}$ is missing is independent of any kind of (additional) knowledge:
  $\Pr(\mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a) \mid \mathcal{K}^*) = \Pr(\mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a))$;
- **MAR** (Missing At Random) – the probability that $a \in C^{\mathcal{I}}$ is missing depends only from $\mathcal{K}$ and does not depend on additional knowledge:
  $\Pr(\mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a) \mid \mathcal{K}^*) = \Pr(\mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a) \mid \mathcal{K})$;
- **NMAR** (Not Missing At Random or **IM**, Informatively Missing) – the probability that $a \in C^{\mathcal{I}}$ is missing could be not the same if additional knowledge is available
  $\Pr(\mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a) \mid \mathcal{K}^*) \neq \Pr(\mathcal{K} \not\models C(a) \wedge \mathcal{K} \not\models \neg C(a) \mid \mathcal{K})$.

TBN under MCAR assumption

- *Only positive and negative examples* is *considered*
- *Parameters estimated* by the use of the *frequency distribution*
- **score** computed as the log-likelihood on training data:

$$\mathcal{L}(\mathcal{N}_{\mathcal{K}} \mid Ind_C(\mathcal{A})) =$$

$$= \sum_{a \in Ind_C^+(\mathcal{A})} log \Pr(C(a) \mid \mathcal{N}_{\mathcal{K}}) + \sum_{a \in Ind_C^-(\mathcal{A})} log \Pr(\neg C(a) \mid \mathcal{N}_{\mathcal{K}});$$

TBN under MAR assumption

- *Positive, negative and neutral examples* are *considered*
- The *EM* algorithm is adopted for parameters estimation
- **score** is computed as the log-likelihood on training data considering also the neutral examples

$$\mathcal{L}(\mathcal{N}_{\mathcal{K}} \mid Ind_C(\mathcal{A})) = \sum_{a \in Ind^0_C(\mathcal{A})} \sum_{C' \in \{C, \neg C\}} \log \Pr(C'(a) \mid \mathcal{N}_{\mathcal{K}}) \Pr(C' \mid \mathcal{N}_{\mathcal{K}})$$
$$+ \sum_{a \in Ind^+_C(\mathcal{A})} \log \Pr(C(a) \mid \mathcal{N}_{\mathcal{K}}) + \sum_{a \in Ind^-_C(\mathcal{A})} \log \Pr(\neg C(a) \mid \mathcal{N}_{\mathcal{K}});$$

TBN under NMAR assumption

- Positive and negative examples are considered
- For the neutral examples, all the possible fillings are considered
- *Robust Bayesian estimation* (RBE) is adopted **to learn conditional probability distributions**
  - *probability intervals* are determined instead of single probability values
- **score**: as for MCAR considering the mean value of the probability intervals

Classifying individuals with a TBN: Example

Given:

- the feature concepts $\mathcal{F} = \{Female, HasChild\}$
- the target concept *Father*
- the naïve TBN of the previous example
- the DL KB $\mathcal{K}$
- an individual a s.t. $\mathcal{K} \models HasChild(a)$ while the membership of a to *Female* is not known

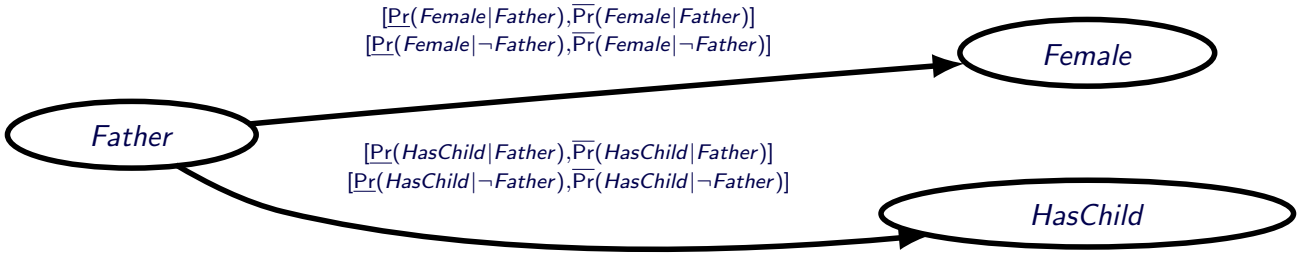
The probability that a is instance of *Father* is given by:

$$\Pr(Father(a)) = \frac{\Pr(Father)\Pr(HasChild \mid Father)}{\sum_{Father' \in \{Father, \neg Father\}} \Pr(Father')\Pr(HasChild \mid Father')};$$

Classifying individuals using RBE: Example

A naïve TBN using Robust Bayesian Estimation for inferring posterior probability intervals in presence of NMAR assumption is s.t.

- *conditional probability contain probability intervals* (defined by upper and lower bound) instead of probability values



Inference on *a* instance of *Father* given that $\mathcal{K} \models HasChild(a)$, is given by the interval $[\underline{\Pr}(Father \mid HasChild), \overline{\Pr}(Father \mid HasChild)]$, where:

$$\underline{\Pr}(Fa(a)) = \underline{\Pr}(Fa \mid HC) = \frac{\underline{\Pr}(HC \mid Fa)\underline{\Pr}(Fa)}{\underline{\Pr}(HC \mid Fa)\underline{\Pr}(Fa) + \overline{\Pr}(HC \mid \neg Fa)\overline{\Pr}(\neg Fa)};$$

$$\overline{\Pr}(Fa(a)) = \overline{\Pr}(Fa \mid HC) = \frac{\overline{\Pr}(HC \mid Fa)\overline{\Pr}(Fa)}{\overline{\Pr}(HC \mid Fa)\overline{\Pr}(Fa) + \underline{\Pr}(HC \mid \neg Fa)\underline{\Pr}(\neg Fa)};$$

Conclusions & Future Work

Conclusions: Proposed a ML method based on the naïve Bayes assumption for estimating the probability that a generic individual belongs to a certain target concept, given

- its membership relation to an induced set of (complex) DL concepts
- an ignorance model for handling incomplete knowledge

Future works:

- experimenting with the method
- finding optimizations of the proposed method

The End

That's all!

Questions ?