ELSEVIER

Artificial Intelligence in Medicine 29 (2003) 39–60

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection

Peter Antal[a,*], Geert Fannes[a], Dirk Timmerman[b], Yves Moreau[a], Bart De Moor[a]

[a]*Electrical Engineering Department ESAT/SCD, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Heverlee, Leuven, Belgium*

[b]*Department of Obstetrics and Gynecology University Hospitals, Katholieke, Universiteit Leuven, Herestraat 49, B-3000 Leuven, Belgium*

Received 24 January 2002; received in revised form 4 December 2002; accepted 17 March 2003

Abstract

Incorporating prior knowledge into black-box classifiers is still much of an open problem. We propose a hybrid Bayesian methodology that consists in encoding prior knowledge in the form of a (Bayesian) belief network and then using this knowledge to estimate an informative prior for a black-box model (e.g. a multilayer perceptron). Two technical approaches are proposed for the transformation of the belief network into an informative prior. The first one consists in generating samples according to the most probable parameterization of the Bayesian belief network and using them as virtual data together with the real data in the Bayesian learning of a multilayer perceptron. The second approach consists in transforming probability distributions over belief network parameters into distributions over multilayer perceptron parameters. The essential attribute of the hybrid methodology is that it combines prior knowledge and statistical data efficiently when prior knowledge is available and the sample is of small or medium size. Additionally, we describe how the Bayesian approach can provide uncertainty information about the predictions (e.g. for classification with rejection). We demonstrate these techniques on the medical task of predicting the malignancy of ovarian masses and summarize the practical advantages of the Bayesian approach. We compare the learning curves for the hybrid methodology with those of several belief networks and multilayer perceptrons. Furthermore, we report the performance of Bayesian belief networks when they are allowed to exclude hard cases based on various measures of prediction uncertainty.

Keywords: Knowledge acquisition; Belief network; Bayesian multilayer perceptron; Informative prior distribution; Classification with rejection

[*] Corresponding author. Tel.: +32-16-32-18-01.
E-mail address: peter.antal@esat.kuleuven.ac.be (P. Antal).

0933-3657/$ – see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0933-3657(03)00053-8

1. Introduction

In this paper, we describe a two-step methodology to construct classifiers that consists of (1) the formalization of domain knowledge and (2) the combination of this formalized knowledge with data. This methodology stands in the Bayesian statistical framework and relies on a dual representation of belief networks and black-box regression models. For the formalization step, we use belief networks to elicit and represent the domain knowledge in the Bayesian framework. For the combination step, we introduce a method for the transformation of belief networks into an *informative prior distribution* for black-box models and describe the use of virtual prior samples for the same purpose. We evaluate our methodology on the medical classification task of discriminating preoperatively between benign and malignant ovarian tumors. The results show that the methodology based on the dual representation is superior to both the pure belief network and the multilayer perceptron. Additionally, we describe how the Bayesian approach can provide uncertainty information about the predictions and how to use this information for classification with rejection (when the hard cases can be excluded based on various measures of prediction uncertainty).

The paper is organized as follows: Section 2 reviews the application of domain knowledge when constructing a classifier, and outlines our motivation for a methodology based on a dual representation. Section 3 summarizes the Bayesian approach in general and specifically in classification. Section 4 describes the preoperative prediction of malignancy in ovarian tumors, which serves as an evaluation domain, and introduces the data set. Section 5 introduces the prior belief network, summarizes its construction and its application for tumor classification. Section 6 introduces the concept of *informative prior distribution* for parametric black-box models together with the two methods we propose for the construction of such an informative prior. Section 7 first presents the performance of the prior belief network against medical experts. We then report the comparison of the different prior transformation methods by presenting the performance of learning methods for belief networks, multilayer perceptrons and hybrid models. Additionally, we use various Bayesian uncertainty measures to identify "hard" cases. Finally, we report the performance of belief networks when rejection is allowed based on these measures. Sections 8 and 9 contain the discussion and conclusion.

2. Motivation for the hybrid application of belief networks and multilayer perceptrons

In the paper we make the following assumptions about the classification task we consider:

(1) The classification is binary with continuous and nominal input variables and the prediction of the class probability and of an uncertainty measure about this probability is advantageous.
(2) The size of the sample is small or medium with respect to the learnability of the problem and missing data are infrequent.
(3) A large amount of prior knowledge is available about the domain, the variables, the dependencies between variables, and the quantification of these dependencies.

These assumptions are inspired partly by the ovarian tumor problem described in Section 4 and our mathematical derivations will be formulated in this context. It is however important to stress that the methods proposed can straightforwardly be extended to multiclass classification and to regression, simply by rewriting the mathematical details accordingly.

If standard statistical tools, such as logistic regression, do not give satisfactory results, we need to use more complex models like data-driven black-box methods (such as multilayer perceptrons, decision trees and kernel-based methods) or more knowledge-oriented white-box methods (such as belief networks).

In the case of black-box methods, the possibilities to incorporate prior knowledge in the model or in the learning process are limited, even though this incorporation is frequently essential. It is generally confined to the selection of the input variables, of the model structure, of the learning algorithm, and to the management of missing values. In the Bayesian context, an inherent problem for black-box parametric models is that it is not possible to directly construct an informative prior distribution.

In the case of white-box methods, particularly for belief networks, the possibilities to incorporate domain knowledge in the model are greatly enhanced. A prior distribution for the parametrization of a given model structure or for the model structures can be constructed by established methods [36,52]. However, the sample complexity of parameter learning [20] is in practice frequently higher than the black-box models. The full scale Bayesian inference or general structure learning is hindered by the superexponential cardinality of the structure space and the high sample complexity [18,25,26]. Additionally, the general structure learning methods—the data-dependent terms and regularization terms—are optimal for learning the joint distribution (i.e. for complex model discovery). It means that they are more appropriate to solve a much harder task than is necessary in a standard classification. As a solution, special belief network classifiers with restricted structures were suggested (tree augmented networks, TAN) [15,24,39], which decrease the sample complexity of learning and are biased towards classification. Partly related to this problem, general structure-learning algorithms have failed in our preliminary experiments to achieve a good quantitative performance, while multilayer perceptrons reached nearly the performance of expert diagnosticians [54], but only in the large sample region (i.e. for all the cases occurred in 2 years in a large, referal medical center). Because small or medium size samples are frequent in medical problems and a good prior belief network was available [9], we decided to investigate the incorporation of prior knowledge into a multilayer perceptron as an evaluation of a general methodology for such problems.

Our two-step methodology combines certain complementary advantages of belief networks and multilayer perceptrons by (1) formalizing the prior knowledge with the belief network and (2) incorporating the formalized knowledge into the Bayesian learning and inference of the multilayer perceptron (Fig. 1). The use of a black-box method in the inductive phase provides a computationally and statistically efficient solution to refine jointly the a priori structure, the prior over the parameters, and the a priori discretization corresponding to the a priori belief network.

Beside the research on belief network classifiers, the sequential application of the white-box and black-blox techniques arises from the standpoint of black-box learning. The appearance of learning theories [14,21,60] made it possible to formalize how the

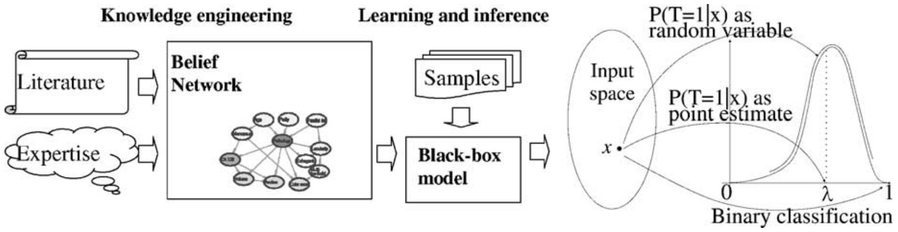


Fig. 1. The methodology based on (1) belief networks for knowledge acquisition, (2) a prior transformation method, and (3) black-box models for learning and classification (left). Classification can target three levels: the class label (discrimination), the class probability (regression), and the distribution of the class probability (Bayesian prediction) (right).

incorporation of domain knowledge in inductive techniques reduces the statistical complexity of learning (in the classical statistical context [1,29,34] and in the Bayesian context [35]). On the practical side, Abu-Mostafa [1] and Niyogi et al. [45] reported methods for exploiting a priori known regularities and symmetries in the input space. Another approach, the knowledge-based artificial neural network, used the prior knowledge for selecting an appropriate multilayer perceptron architecture [57]. This method formalizes the domain knowledge in propositional logic to construct the structure of a multilayer perceptron. Further works reported results about the inductive refinement of the initial network structure, the extension of the translation and transformation of symbolic rules into a feedforward artificial neural network (for surveys about knowledge-based neurocomputing, see [17,49]). Sowmya [50] extended the symbolic paradigm for the transformation of domain theories into a feedforward neural network by proposing belief networks with certain local models for knowledge modeling. Another proposal [42] similarly emphasized the appropriateness of belief networks for prior knowledge formalization and described a mapping of belief networks onto stochastic neural networks to support parallel computations. The potential of *Bayesian* belief networks for supporting the construction of a classifier was surveyed in [4].

3. Introduction to Bayesianism

The Bayesian approach is attractive for machine learning because it combines valuable subjective prior information in a principled way. Furthermore, the Bayesian prediction provides more detailed information for decision support than a class label or class probability. For overviews, see [12,28,43].

3.1. Bayesian classification

Starting from a prior distribution expressing the initial beliefs in the parameters of the model, we can use the observations to transform this prior into the posterior distribution for the model parameters (which expresses the beliefs after observing the data). Using this posterior distribution over the model parameters, random variables can be defined for the predicted class probability for a case and performance measures for a model class on a data set.

To treat the binary classification task, we introduce the following notations: we use capitals for random variables, bold for vectors, and bold underlined for matrices, $\boldsymbol{\omega} \in \mathbb{R}^n$ denotes the model parameters, and when distinction is necessary, $\boldsymbol{\omega}$ denotes the parameters of the multilayer perceptron and $\boldsymbol{\theta}$ the parameters of the belief network. We assume the existence of a labeled training set $\underline{\boldsymbol{d}} = \{(\boldsymbol{x}_k, t_k)\}_{k=1}^n, (\boldsymbol{x}_k, t_k) \in \mathbb{R}^d \times \{0, 1\}$, where $\boldsymbol{x}$ is a real-valued l-dimensional input vector and t is the corresponding class label.

For a *probabilistic regression* model $P(T = 1|\boldsymbol{x}, \boldsymbol{\omega}) = f(\boldsymbol{x}, \boldsymbol{\omega}) \in [0, 1]$, we assume a prior distribution over the model parameters $\boldsymbol{\omega}$. The corresponding random vector is denoted with $\boldsymbol{\Omega}$ and $p_{\boldsymbol{\Omega}}(\cdot)$ denotes the distribution. Using the observed data $\underline{\boldsymbol{d}}$ and applying Bayes' rule, the prior distribution can be transformed into the posterior distribution $p_{\boldsymbol{\Omega}}(\boldsymbol{\omega}|\underline{\boldsymbol{d}})$ by

$$p_{\boldsymbol{\Omega}}(\boldsymbol{\omega}|\underline{\boldsymbol{d}}) = \frac{p_{\boldsymbol{T}}(t_1, \ldots, t_n|\boldsymbol{\omega}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)p_{\boldsymbol{\Omega}}(\boldsymbol{\omega})}{p_{\underline{\boldsymbol{D}}}(\underline{\boldsymbol{d}})} \propto L(\boldsymbol{\omega}|\underline{\boldsymbol{d}})p_{\boldsymbol{\Omega}}(\boldsymbol{\omega})$$

$L(\boldsymbol{\omega}|\underline{\boldsymbol{d}})$ denotes the probability of the data given the parameters and $F_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}(\boldsymbol{x}) \equiv f(\boldsymbol{x}, \boldsymbol{\Omega})$ denotes the induced random variable on $[0, 1]$ for the predicted posterior class probability. Note that the distribution over the parameters of an input–output model is independent of the input values (i.e. $p_{\boldsymbol{\Omega}|\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}$ is equal to $p_{\boldsymbol{\Omega}}$).

For the assessment of the performance of a probabilistic classifier, we have chosen the area under the receiver operating characteristic (ROC) curve (AUC), which is a utility-independent performance measure particularly widespread in medical applications (for the definition and interpretations of the ROC curve, see [31,32]). There are two possibilities to apply and report this performance measure in the Bayesian framework, depending on the formalization of the reporting problem. A first possibility is to integrate out the model parameters, which results in scalar class probabilities from which the area under the ROC curve can be computed directly. A second possibility is to interpret this area as a function of the model parameters for a fixed data set—that is, as the random variable $\mathrm{AUC}(\Omega, \underline{\boldsymbol{d}})$. In this paper we follow the second approach.

3.2. Classification with rejection based on Bayesian prediction

In binary classification problems, we are interested, for a given $\boldsymbol{x}$, in the random variable $f(\boldsymbol{x}, \boldsymbol{\Omega})$ corresponding to the probability $P(T = 1|\boldsymbol{x}, \boldsymbol{\omega})$ for Class 1, where $\boldsymbol{\Omega}$ is a random parameter vector. In this way the distribution of $f(\boldsymbol{x}, \boldsymbol{\Omega})$ gives us uncertainty information about the predicted class probability. One possible use of the Bayesian prediction of the class probability is to compute the expected loss corresponding to an imaginary reporting situation. If it exceeds a certain threshold τ, then no subsequent real decision is taken, for example, further examinations can follow. A rejection scheme in this case can be defined by the loss function L of the imaginary reporting situation, by the imaginary reporting $\tilde{f}(\boldsymbol{x})$ and by the posterior belief represented by the random vector $\boldsymbol{\Omega}|\underline{\boldsymbol{d}}$:

Definition 1.

$$g_{\lambda,\tau}(\boldsymbol{x}) = \begin{cases} \text{``rejected'' if } \tau < \int_{\mathbb{R}^n} L(f(\boldsymbol{x}, \boldsymbol{\omega}), \tilde{f}(\boldsymbol{x}))\, \mathrm{d}P_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}} \\ \text{else 1 if } \lambda < \tilde{f}(\boldsymbol{x}) \\ \text{else 0} \end{cases}$$

Using this scheme, we define the following quantities to characterize the "hard" cases. We use the induced measure by the random variable $f(\boldsymbol{x}, \boldsymbol{\Omega})$ on $[0, 1]$ in the definitions, which is denoted by $P_{f(\boldsymbol{x},\boldsymbol{\Omega})}$.

Definition 2.

$$\delta_{L_1}(\boldsymbol{x}) = \int_0^1 |y - \mathrm{Median}(f(\boldsymbol{x}, \boldsymbol{\Omega}))| \, \mathrm{d}P_{f(\boldsymbol{x},\boldsymbol{\Omega})} \tag{1}$$

$$\delta_{\mathrm{Var}}(\boldsymbol{x}) = \int_0^1 (y - \mathrm{Mean}(f(\boldsymbol{x}, \boldsymbol{\Omega})))^2 \, \mathrm{d}P_{f(\boldsymbol{x},\boldsymbol{\Omega})} \tag{2}$$

$$\delta_{\mathrm{H}}(\boldsymbol{x}) = \int_0^1 \log(y) \, \mathrm{d}P_{f(\boldsymbol{x},\boldsymbol{\Omega})} \tag{3}$$

δ_{L_1} defines the expected loss in the case of the L_1 loss function and the corresponding optimal reporting of the median. δ_{Var} is the variance of the Bayesian prediction of the class probability, it defines the expected loss in the case of the L_2 loss function and the corresponding optimal reporting of the mean. δ_{H} defines the uncertainty from the information theoretic point of view as an entropy.

4. Classification of ovarian masses

We shall illustrate our theoretical developments on a real-world medical problem relating to cancer. Ovarian malignancies represent the greatest challenge among gynecologic cancers. Early detection is of primary importance for the survival of the patient, since currently more than two thirds of the patients are diagnosed only at an advanced stage and therefore have poor prognosis. A reliable test to discriminate between benign and malignant tumors before surgery would be of considerable help to clinicians. It would help them to recognize patients for whom treatment with minimally invasive surgery or conservative management suffices versus those for whom referral to a gynecologic oncologist is needed for more aggressive treatment. There are two different types of information for the development of such predictive models: the biological and medical information about the disease and the growing amount of patient data.

4.1. Domain

The abundant background knowledge is diverse: for example, the MEDLINE collection of abstracts from biomedical journal papers contain tens and thousands of items about ovarian cancer. Factors known to affect the risk of malignancy are parity (number of pregnancies), infertility treatment, duration of lactation, oral contraceptives, foreign bodies (carcinogens), family history of breast and ovarian cancer, genetic deficiencies, age, age at menopause and hysterectomy. Beside clinical data, additional measurements and observations are the following: bilaterality of the tumor, pelvic pain, morphological descriptors of the mass (such as smoothness and solidness), descriptors of its echogenicity and vascularization, level of several serum tumor markers, such as CA125, amount of fluid in the abdominal cavity and the day of the cycle. While the effect of some of these variables can be quantified reliably (such as the effect of the family history and genetic deficiencies), other effects are only qualitatively known and highly subjective (such as the use of the vascularization indices).

4.2. Data

In addition to the prior background information, data have been collected prospectively from 300 consecutive patients who were referred to a single institution (University Hospitals Leuven, Belgium) from August 1994 till June 1997. The data collection protocol excludes other causes with similar symptoms, such as infection or ectopic pregnancy and ensures that the patients with persistent extrauterine pelvic mass undergo surgery. This eliminates the possibility of false negatives and the quality of this single center study provides reliable pathology values as gold standard (for a detailed description, see [54–56]). Univariate statistics of the data set are presented in Table 1.

4.3. Previous studies

The first predictive models were based on single variables (such as CA125, resistance index) or risk indices (Lerner's scoring system, risk of malignancy index (RMI)). Standard statistical studies indicate that a multimodal approach—the combination of several variables—is necessary for the discrimination between benign and malignant tumors. Therefore, several studies [9,55] have applied logistic regression, multilayer perceptrons

Table 1
Univariate statistics for the benign (72.3%) and malignant (27.7%) subpopulations in the ovarian cancer data set [54]

Variable	Benign	Malignant
Age	47.77 (15.60)	58.81 (15.18)
CA 125	110.3 (976.6)	1235.0 (3757.4)
Color score	1.98 (0.84)	3.20 (0.95)
Parity	1.495 (1.397)	1.578 (1.719)
Resistance index	0.634 (0.163)	0.543 (0.168)
Ascites (present) (%)	6.45	48.2
Bilateral (yes) (%)	14.4	41.0
Postmenopause (yes) (%)	35.0	69.9
Papillation (present) (%)	10.6	66.3
Locularity		
Unilocular (%)	43.7	4.8
Unilocular—solid (%)	4.3	14.5
Multilocular (%)	29.9	4.8
Multilocular—solid (%)	13.4	38.5
Solid (%)	8.7	37.4

We report the mean and standard deviation in the benign and the malignant subpopulation for the continuous variables, and the occurrences (%) of the values for the nominal variables.

Table 2
Performance of previous models (area under the ROC curve): univariate discrimination based on the level of serum CA125, discrimination based on the risk of malignancy index, logistic regression, multilayer perceptron, and two belief networks

Model	CA125	RMI	LR	MLP	Naïve BN	Parametric BN
Area under ROC	0.874	0.891	0.904	0.951	0.938	0.952
Standard error	0.034	0.032	0.060	0.039	0.037	0.034

and belief networks. These models predicted the scalar class probabilities and they were developed and tested in the classical statistical framework on a smaller data set (Table 2).

.

5. Bayesian belief networks

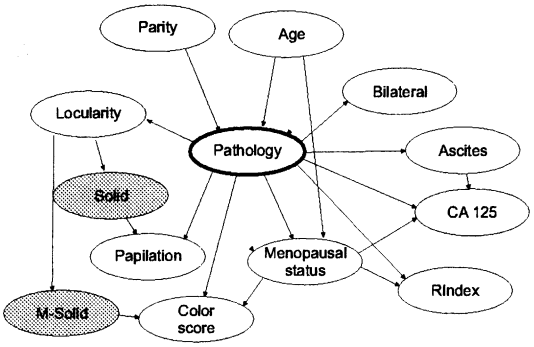
A belief network represents a joint probability distribution over a set of variables [46]. We assume that these are discrete variables, partitioned into three sets $\boldsymbol{X}, Y, \boldsymbol{Z}$: set of inputs, output, and intermediate variables, respectively. The model consists of a qualitative part (a directed graph) and quantitative parts (dependency models). The vertices V_i of the graph represent the random variables X_i and the edges define the independency relations (each variable is independent of its nondescendants given its parents [46]). There is a probabilistic dependency model for each variable that describes its dependency on its parents.

In this paper, we use standard multinomial dependency models which directly encode the conditional probabilities of the child conditioned on parental value configurations. Following Spiegelhalter et al. [52], we assume parameter independence and use Dirichlet distributions to define a distribution over these parameters of the dependency models. In this case, the prior background knowledge is formalized as a single belief network structure and a prior density over the parameters is given by

$$p(\boldsymbol{\theta}) = \prod_{i=1}^{m} \prod_{j=1}^{\mathrm{pa}_i} \mathrm{Dirichlet}(\theta_{ij1}, \ldots, \theta_{ijr_i} | N_{ij1}, \ldots, N_{ijr_i}) \tag{4}$$

where the N_{ijk} hyperparameters can be interpreted as the number of previously seen examples for $X_i = k$ and j is an one-based index for all possible parental configurations pa_i (equal to the product of the cardinality of parental values). Furthermore, the posterior update of a Dirichlet prior over the parameters of a multinomial distribution is also a Dirichlet distribution for multinomial sampling: the hyperparameters are simply updated according to the ‘‘counting’’ interpretation. Because of the assumption of parameter independence, the local Dirichlet models can be updated independently if they are fully observable according to the decomposition of the directed graph.

In a simplified approach the prior hyperparameters are derived using a single sample size N, which can be interpreted as the number of complete cases seen a priori [27]. In this case, the posterior inference $P(T = 1|\boldsymbol{x}, \boldsymbol{\Theta})$ follows a Dirichlet (Beta) distribution. However, this approach is frequently too strict in practice, for example, in our case different parts of the prior belief network are quantified by different experts or studies and consequently the



Variable	Source
Bilaterality	[58]
CA125	[23]
Color score	[59]
Locularility	[30]
Papillation	[30]
Pathology	[11, 16, 22, 33, 38, 58, 61]
Resistance index	[53]

Fig. 2. The prior belief network model for the ovarian cancer problem. The predicted variable "Pathology" is indicated by the thick node. The input variables that are not used in the inductive methods—except in the parametric learning with the fixed a priori BN structure —are grayed (left). Relevant publications provide information on logical dependency and parametric dependency for certain variables beside the expert's opinion (right) [11,16,22,23,30,33,38,53,58,59,61].

network cannot be characterized by a single *prior sample size N*. Consequently, we perform the Bayesian inference in two steps. At first, a parametrization $\boldsymbol{\theta}$ is drawn by direct sampling from the Dirichlet distributions [47], subsequently an exact inference computation is performed using a probability propagation algorithm in tree of cliques. For the theoretical foundations and implementational issues, see [2,37,51].

5.1. Belief network for ovarian cancer

In the prior belief network, we used 13 variables from which the continuous and integer variables were discretized according to the medical literature and expert knowledge. We built a "heterogeneous" belief network containing biological models of the underlying mechanism quantifiable by the literature: parts quantified by a medical expert and parts quantified from previous studies (for a more detailed description of the model construction process, see [9]). The prior belief network is shown in Fig. 2.

Because of the extensive and complex usage of the prior knowledge, we used a strict documentation method to track the route of the prior information from studies into the model, which has lead to the introduction of the *Annotated Belief Networks*. For its applicability on information retrieval, knowledge engineering, and structure learning, see [5,7,8].

6. Bayesian multilayer perceptrons

A multilayer perceptron defines a nonlinear input–output mapping defined by the layers of summation and elementary nonlinear mappings [13]. The model used in this paper is defined by the following formula:

$$f(\boldsymbol{x}, \boldsymbol{\omega}) = \sigma\left[\sum_{i=1}^{5}\left(\omega_i \tanh\left[\sum_{j=1}^{10}(\omega_{ij}x_j + \omega_{i0})\right]\right)\right]$$

in which the tranfer function is a hyperbolic tangent and the logistic function $\sigma(x) = 1/(1+\mathrm{e}^{-x})$; $\boldsymbol{\omega}$ contains all the parameters including the bias parameters ω_{i0}. In the case of the ovarian tumor classification, there are four nominal inputs (locularity, menopausal status, papillation, and bilaterality) and six continuous (color score, resistance index, age, level of CA125, parity, and amount of ascites). The number of neurons in the hidden layer is five.

We interpret the output as a posterior probability, that is, $P(T=1|\boldsymbol{x}) = f(\boldsymbol{x},\boldsymbol{\omega})$. As summarized in Section 3.1 in the case of a complete data set and the existence of a prior distribution, this interpretation gives a posterior distribution for the model parameters. We used multiple imputation for the missing input data based on the prior probabilistic model from the expert (since only the measurement CA125 is missing at random in 33 cases) [28]. For the Bayesian inference we sample from the posterior distribution $p_{\Omega}(\boldsymbol{\omega}|\boldsymbol{d})$ by using the hybrid Monte Carlo method [41,43].

6.1. Noninformative prior

The parameters of the multilayer perceptron are the weights and biases of the composing neurons, which are hard to interpret in a multilayer model. This prohibits the incorporation of prior knowledge into this model by directly specifying a *prior* distribution over the parameters (versus the intuitive interpretation of the Dirichlet used in the belief network). In practice, the prior is used only for penalizing the complexity of the model, for example, based on the L_2 norm of the parameter vector $\boldsymbol{\omega}$ by using a Gaussian prior distribution $\mathscr{N}(\mathbf{0},\mathscr{I})$[13,43].

6.2. Informative prior from belief network

In the simplest situation explained in Section 5, the encoded prior knowledge comes from N previously seen complete cases, which could be used as data in the Bayesian inference with the multilayer perceptron. Therefore, we examine the possibility of generating a certain number of *prior samples* from the domain model.

However, if the belief network is hyperparametrized from heterogeneous sources (e.g. various parts of the model are quantified by different experts or studies), then such a prior complete data set is not available. Even if a complete prior data set is available and reconstructed, then its direct usage as real data loses certain aspects of the prior uncertainty modeling (e.g. network structure, Dirichlet assumption). So we introduce a second method that transforms the domain knowledge encoded in the prior distribution of the belief network into an *informative prior distribution* for the multilayer perceptron.

6.2.1. Informative prior by prior samples

Let us first assume that the prior domain knowledge consists of N complete cases, called the prior sample D_N^{prior}, which will be used together with the real data D_N^{real}. Assuming a noninformative prior distribution $P(\omega)$ for the multilayer perceptron, the Bayesian update is defined as follows:

Definition 3.

$$P(\omega|D_N^{\text{prior}}, D_N^{\text{real}}) \propto P(D_N^{\text{real}}|\omega)P(\omega|D_N^{\text{prior}}) = P(D_N^{\text{real}}|\omega)P'(\omega) \tag{5}$$

This grouping of the terms illustrate the effect of the prior sample D_N^{prior} to transform the noninformative prior distribution $P(\omega)$ into an informative prior $P'(\omega) = P(\omega|D_N^{\mathrm{prior}})$. If the prior data set follows the real conditional distribution, then the effect of this Bayesian update by the prior sample is the same as by real data. A problematic issue is the selection of the prior sample size—particularly, if the prior belief network is heterogeneously hyperparametrized. It is difficult in general, because it means selecting an optimal complexity regularization, consequently it can be influenced by the real sample size, the problem, the general correctness of the prior domain model and in practice even by the inference scheme. In our domain the results of this method proved robust for changing the virtual sample size in the 15–50 range.

In our experiments, we use a stochastic scheme to generate a prior sample equivalent to N samples. We fix the most probable a priori parametrization in the belief network. Instead of generating a sample of size N, we generate a larger number of prior samples that we rescale to an *effective sample size* in the update and inference process. This approach reduces the impact of stochastic effects.

6.2.2. Informative prior by direct transformation

The second method for transforming the domain knowledge encoded in the prior distribution of the belief network into an *informative prior distribution* over the MLP parameters ($p_{\text{MLP-I}}$) is based on the direct transformation of the a priori distribution between model spaces. To formalize this transformation we follow the definition of Neal [44] instead of our earlier approach [6] which corresponds to the asymptotic case ($p_{\text{MLP-NI}}$ denotes a noninformative distribution over the MLP parameters):

Definition 4.

$$p_{\text{MLP-I}}(\omega) = \sum_{d_1,\ldots,d_k} p_{\text{MLP-NI}}(\omega|d_1,\ldots d_k) \int \prod_{i=1}^{k} p(d_i|\theta) p_{\text{BN-I}}(\theta)\,\mathrm{d}\theta \qquad (6)$$

This defines a transformation of the prior probability measure $p_{\boldsymbol{\Theta}}$ over the parameter space of the *donor* model (the belief network in our case is denoted by BN-I) to a prior probability measure $p_{\boldsymbol{\Omega}}$ over the parameter space of the *recipient* model (the MLP model). The main steps for the practical implementation of the mapping in the case of multilayer perceptron are listed below.

Algorithm 1 (Construction of an informative MLP prior).

(1) Generate prior sample.
  (a) Generate belief network parametrizations $\{\theta_1,\ldots,\theta_l\}$.
  (b) Generate a block of prior samples from each parametrization $\{D_1^p,\ldots,D_l^p\}$.
(2) Generate a multilayer perceptron parametrization from the posterior based on each block of samples, resulting in a block of MLP parametrizations $\{\omega_1,\ldots,\omega_l\}$.
(3) Estimate the transformed distribution $p_{\text{MLP-I}}(\omega)$ from the MLP parametrizations with a mixture of Gaussians.

The belief network parametrizations are generated from the Dirichlet distribution by standard methods. The sample blocks are generated according to the belief network parametrizations. Next, a simple preprocessing of the input variables is necessary: (1) perform one-out-of-c coding for nominal variables [13] and (2) resample in the discretization intervals for discretized continuous variables. To generate the multilayer perceptron parametrizations from the posterior, we used the hybrid Monte Carlo Markov Chain [43].

Finally, to estimate the transformed prior distribution $p_{\text{MLP-I}}(\omega)$ over the black-box model parameter space $\boldsymbol{\Omega}$ from the trained perceptrons, we used a mixture of Gaussians (for the approximation with one kernel, see [13]):

$$p(\omega) \approx \sum_{i=1}^{L} \alpha_i \mathcal{N}(\omega|\mu_i, \boldsymbol{\Sigma_i}), \quad \text{where } 0 \leq \alpha_i \leq 1, \sum_{i=1}^{L} \alpha_i = 1$$

Because the total number of symmetries (due to possible permutations and sign symmetries) in a multilayer perceptron with k hidden layers and L_i neurons in layer i is given by $\prod_{i=1}^{k} 2^{L_i} L_i!$, this prior estimation requires the proper management of symmetries in the parameter space [48]. We applied a heuristic clustering algorithm exploiting the symmetries to map the parametrizations into clusters with minimal within-cluster variance (for a more detailed description, see [6]).

7. Results

First, we investigated the performance of the prior belief network. The misclassification rate is 12.0% on the data set and the mean of the Bayesian area under the ROC curve is 0.905. To get a more detailed understanding of the performance of the model we compared its predictions with those of medical experts. In a previous study six ultrasonographers (denoted by A–F in Table 3) have evaluated the 300 patients based on the corresponding medical records and ultrasound images. Two of them were highly experienced (A and B), one moderately experienced (C) and three less experienced (D, E, and F) [54]. Since these classifications were based on the original observations (such as images), in a recent experiment an expert (G) has performed the classification using only the discrete values of the variables present in the prior belief network. In this experiment the expert has also rated the cases as 1 = very certain benign, 2 = uncertain benign, 3 = uncertain, 4 = uncertain malignant, and 5 = very certain malignant, which is an aggregate expression of the mixed nature of the adnexal mass (“fuzziness”) and the probabilistic uncertainty.

Table 3 presents the number of previously performed examinations, the misclassification rate and the agreements with the prior model (Cohen’s kappa) for the diagnosticians [10]. The correspondence with the prior model is the highest for expert A (indicated in bold)

Table 3
Expert agreement with the prior domain model in discriminating benign and malignant adnexal masses [54]

Observer	A	B	C	D	E	F	G	Prior-BN
#	$\geq$ 4000	$\geq$ 10000	$\geq$ 1000	200	300	300	300	–
MR	8.3	8.3	11.0	17.7	7.7	13.3	18.0	12.0
κ	**0.713**	0.687	0.650	0.577	0.503	0.590	0.577	–

Number of examinations (#), misclassification rate (MR), Cohen’s kappa (κ).

which is in line with our expectation because expert A participated in the construction of the prior model.

To evaluate the effect of the prior incorporation methods, we compare them in a retrospective setup with standard learning algorithms for belief networks and multilayer perceptrons. For belief networks, we present results for four models with a *noninformative* prior distribution (BN-naïve, BN-fixed-noninformative, BN-TAN and BN-general) and one with an *informative* prior distribution (BN-fixed-informative). For the MLP model class, we describe one model with a *noninformative* prior (MLP-noninformative), one with *prior samples* (MLP-prior sample) and one with an *informative* prior (MLP-informative) as explained in Sections 6.1, 6.2.1 and 6.2.2. The BN-fixed-noninformative and BN-fixed-informative methods use the same fixed structure as the prior domain model shown in Fig. 2, in all the other learning methods two deterministic variables that are functions of the variable Locularity were removed, because these auxiliary variables were introduced to support knowledge elicitation.

The BN-fixed-noninformative, BN-fixed-informative, and the BN-naïve methods perform only parameter learning (the BN-naïve method uses a naïve Bayes structure, a tree where all variables are direct children of the predicted variable pathology). The BN-TAN method searches in the space of generalized tree-augmented networks, which are extended naïve belief network structures [15,24]. Finally, the BN-general method searches the space of directed acyclic graphs (in each crossvalidation session 10^5 random orderings of the variables are generated, for each ordering the parental sets are evaluated exhaustively up to three parents, and if necessary, by the greedy (not exhaustive) K2 algorithm [18]).

In the *prior sample* (MLP-prior sample) method, 1000 samples are generated from the prior belief network rescaled to an effective sample size of 30 in the Bayesian inference scheme as explained in Section 6.2.1. In the *informative prior* (MLP-informative) method, the *informative prior* was estimated on 5000 multilayer perceptron parametrizations using the mixture of three Gaussian kernels. Each multilayer perceptron parametrization is computed from an independently drawn belief network parametrization by training the multilayer perceptron on 1000 random samples produced by the belief network, as explained in Section 6.2.2.

For the Bayesian inference, direct sampling was used for the belief network and hybrid Monte Carlo methods were used for multilayer perceptrons to draw 100 parametrizations from the a posteriori distribution, thus performing 100 inferences for the test set. This process was repeated for 100 cross-validation sessions (different partitions of the data set into test and training set). Fig. 3 shows the detailed effect of the prior incorporation for varying proportions of samples used in the training set.

Two characteristic points from this learning curve (the small and large sample region) are shown in Figs. 4 and 5 corresponding to the 5–95% and 75–25% training–test proportions.

We investigated the advantages of having a more detailed probabilistic prediction in the Bayesian framework. We noticed that the predictions for misclassified cases are more uncertain (e.g. they have higher variances $\mathrm{Var}_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x},\boldsymbol{\omega})]$ which is one measure for the “uncertainty”). Generally spoken, the cases with a high value for $\mathrm{Var}_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x},\boldsymbol{\omega})]$ were also hard to classify for a medical professional, in contrast with cases with a low value for $\mathrm{Var}_{\boldsymbol{\Omega}|\underline{\boldsymbol{d}}}[f(\boldsymbol{x},\boldsymbol{\omega})]$ that were almost always straightforward to predict.

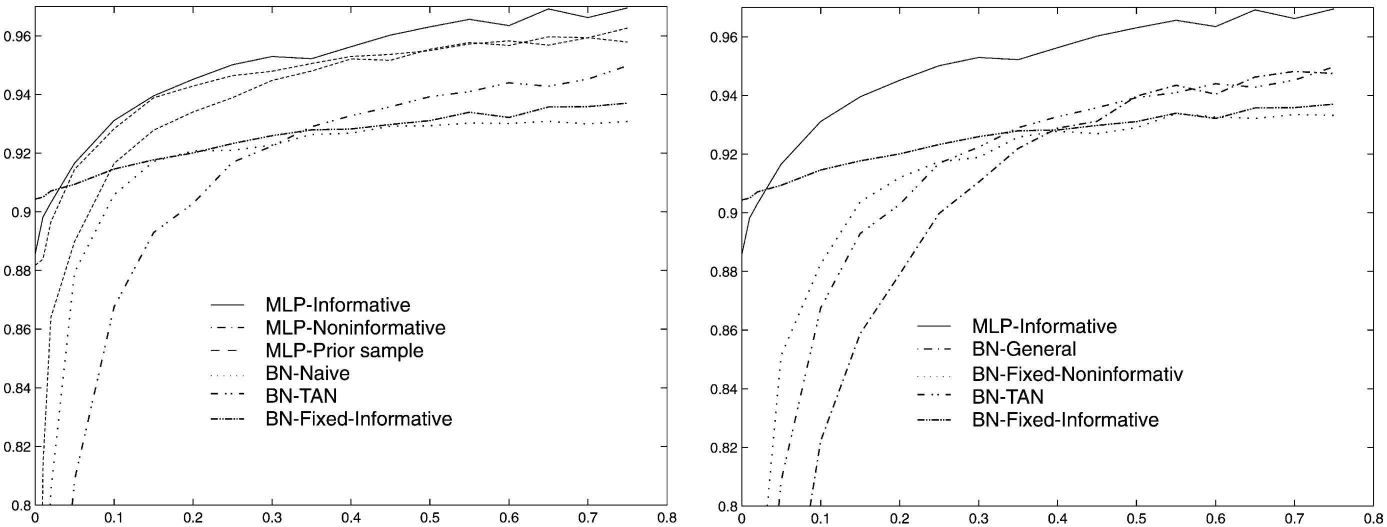
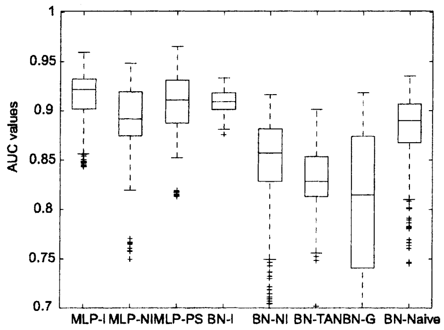


Fig. 3. The learning curves for the multilayer perceptron models using an informative prior (MLP-informative), a noninformative prior (MLP-noninformative) or prior samples (MLP-prior sample). For the belief network models, the learning curves correspond to the naïve Bayes structure (BN-naïve) with noninformative prior, a search in the generalized tree-augmented networks (BN-TAN) with noninformative prior, and to the fixed prior structure in combination with the informative prior (BN-fixed informative) (left). The other figure shows the learning curves for the multilayer perceptron and belief network models using an informative prior (MLP-informative and BN-fixed informative) in comparison with three belief network models using a noninformative prior in combination with a search over the generalized tree-augmented network space (BN-TAN), the fixed prior structure (BN-fixed noninformative) and a general belief network structure learning algorithm (BN-general) (right). The *x* axis indicates the proportion of samples used for training while the *y* axis represents the corresponding area under the ROC curve.



	$\widehat{E}[.]$	$\widehat{Std}[.]$
MLP-Informative	0.915	0.026
MLP-Noninformative	0.889	0.039
MLP-Prior sample	0.907	0.032
BN-Fixed-Informative	0.909	0.011
BN-Fixed-Noninformative	0.851	0.040
BN-TAN	0.809	0.037
BN-General	0.738	0.113
Naïve-BN	0.879	0.039

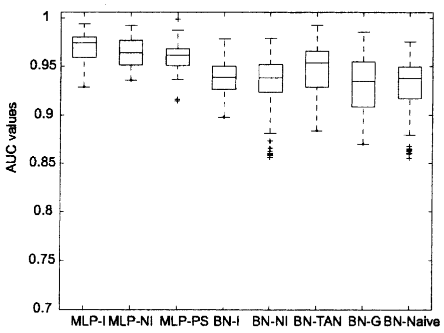
Fig. 4. The a posteriori distribution of the area under the ROC curve at the 5–95% training–test proportion for the MLP-informative (MLP-I), MLP-noninformative (MLP-NI), MLP-priorsample (MLP-PS), BN-fixed-informative (BN-I), BN-fixed-noninformative (BN-NI), BN-TAN, BN-general (BN-G), and BN-naïve models.

To evaluate this observation quantitatively, we introduced the following definitions for "uncertainty" based on the experts prediction from the experiments.

Definition 5.

- Expert is uncertain: expert G labeled a case as uncertain (uncertainty score of 2, 3, or 4).
- Expert misclassifies: either expert A or B misclassifies.
- Less experienced misclassifies: at least two out of the experts C, D, E, and F misclassify.
- Majority misclassifies: at least two out of the six experts misclassify.
- BN-I misclassifies: the model BN-fixed-informative itself misclassifies.

To quantify the efficiency of the uncertainty measures defined in Definition 2 to differentiate these groups, we computed the corresponding area under the ROC curve using the BN-fixed-informative model. Table 4 shows the AUC values for various definitions of "hard cases" versus uncertainty measures.



	$\widehat{E}[.]$	$\widehat{Std}[.]$
MLP-Informative	0.969	0.015
MLP-Noninformative	0.964	0.015
MLP-Prior sample	0.959	0.016
BN-Fixed-Informative	0.937	0.021
BN-Fixed-Noninformative	0.933	0.024
BN-TAN	0.950	0.026
BN-General	0.947	0.027
Naïve-BN	0.930	0.025

Fig. 5. The a posteriori distribution of the area under the ROC curve at the 75–25% training–test proportion for MLP-informative (MLP-I), MLP-noninformative (MLP-NI), MLP-priorsample (MLP-PS), BN-fixed-informative (BN-I), BN-fixed-noninformative (BN-NI), BN-TAN, BN-general (BN-G) and BN-naïve models.

Table 4
Identification of "hard cases" (based on expert errors and assessments) using the predicted uncertainty measures from the BN-fixed-informative (BN-I) model (based on the Bayesian predictions)

	Expert is uncertain	Expert misclassifies	Less experienced misclassifies	Majority misclassifies	BN-I misclassifies
δ_{L_1}	0.642	0.776	0.793	0.796	0.775
δ_{Var}	0.610	0.753	0.776	0.766	0.777
δ_{H}	0.678	0.800	0.805	0.817	0.775

Average area under the ROC values are reported on the test set, 70–30% training–test proportion, 30 times cross-validation.

One promising possibility of having a quantification for the uncertainty of the prediction is to allow the rejection of the most uncertain cases, which in practice can mean further examinations or referring such a patient to an expert. To investigate the effect of rejection, we computed the area under the ROC curve when various proportions of the most uncertain cases are rejected. Fig. 6 shows the AUC values after excluding various proportions of the data set based on the uncertainty measures and the same for the rejected partition. In these experiments, we partitioned the data set randomly into a training (70%) set and test (30%), this was repeated 100 times to eliminate dependency on separation. The reported results are based on the test set.

8. Discussion

As more and more domain knowledge becomes available beside statistical data, machine learning increasingly needs methods that integrate domain knowledge [3,40]. The first step in this prior incorporation is the acquisition and formalization of the heterogeneous a priori information. Because of space limitations, this research direction cannot be followed and covered in this paper, however the experiences of building the reported prior domain model and particularly recent attempts to broaden the scope of the model has lead us to the *Annotated Belief Network*, the semantic annotation of the belief network model with expert knowledge and documents (for results on its potential for knowledge engineering and machine learning, see [8] and [5,7]). The second step is the incorporation of the formalized prior knowledge in a task-specific model. The prior incorporation methods described in Section 6.2 make it possible to integrate probabilistic domain knowledge into black-box models. This hybrid use of knowledge-oriented and data-driven methods can be particularly advantageous in constructing a classifier when the size of the sample is small or medium and a large amount of domain knowledge is available.

The efficiency and simplicity of the *prior sample* method makes it attractive. This method has the advantage that the generation of the prior data set from a belief network is straightforward, computationally simple and that it can be applied to nonparametric models, such as support vector machines. Prior incorporation significantly enhances the performance for small sample sizes. More surprisingly, we also observed slight improvements when the size of the real data exceeds the effective prior sample size by a factor of two to four (the 0.2–0.4 region in Fig. 3). The rescaled large *prior sample* block—which may contain

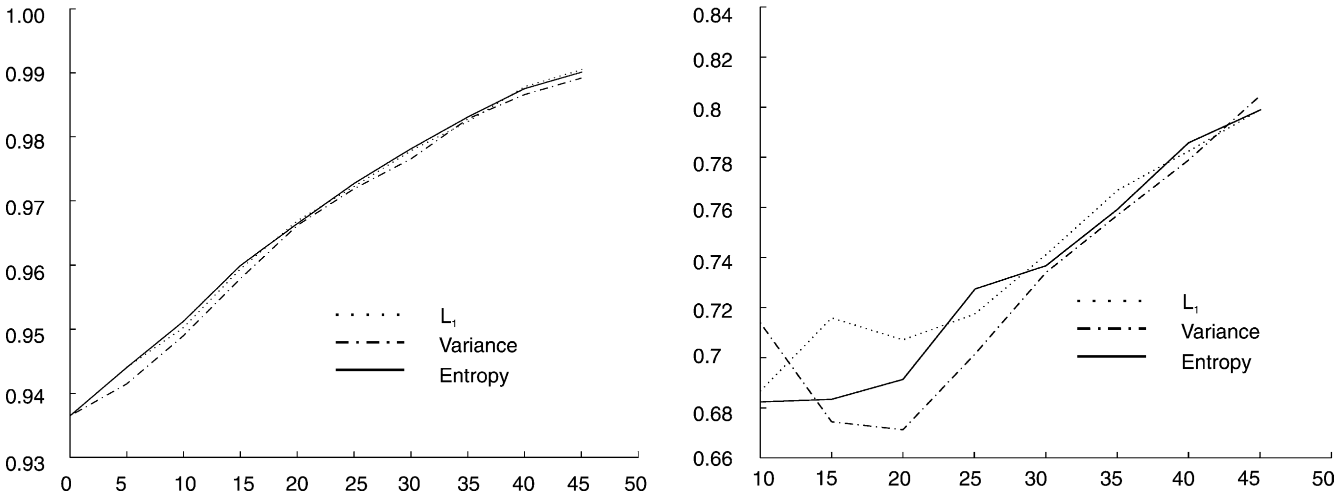


Fig. 6. Classification by the BN-fixed-informative model with rejecting varying number of samples: AUC values on the test set (left) and on the rejected set (right).

infrequent samples as vital hints from the prior belief network—is probably the source of this improvement.

Similarly, the *informative prior* method immediately achieves the same performance as the prior belief network and gives better performance than the noninformative multilayer perceptron for any amount of real data available in the experiment. It means that the estimated prior is efficient in the small sample region and not restrictive in the large sample region; it has a balanced, lasting positive impact. Beside this statistical (machine learning) aspect, the related computational aspect of the inference similarly provides certain advantages. The high computational complexity of deriving the informative prior (when compared to simply generating blocks of samples in the *prior sample* method) is compensated by a lower complexity in the inference. Note that the additional blocks of prior samples slows down the likelihood and gradient computations in the Bayesian inference in MLP models. This precomputation property of the *informative prior* method is similarly relevant in the belief network context. Remember that, for belief networks (even using only a single structure with single parametrization), the computational complexity of the exact inference or its approximation is NP hard [19]. Consequently, the transformation itself—the precomputation of the collapse of a complex general Bayesian model into a task specific, simpler classification model—can be advantageous, if the computational efficiency of the Bayesian inference is important—for example, if a regular classification task in medical decision support involves only a relatively small, but fixed subpart of a complex a priori belief network covering the overall domain.

Multilayer perceptrons perform generally better than belief networks, but part of this difference comes from the refinement of the a priori discretization. Indeed, the performance of a corresponding noninformative multilayer perceptron with the original nominal inputs is similarly worse (for $70-30\%$ train–test ratio, the mean AUC is 0.945 and the misclassification rate is 9%).

Figs. 4 and 5 give a more detailed characterization of the different models, showing that in the small sample region (5%) the prior based methods have the best performance (BN-fixed-informative, MLP-informative, and MLP-prior sample). From the *tabula rasa* methods (MLP-noninformative, BN-general, BN-TAN, and BN-naïve) the BN-naïve has the best performance. Another effect of the incorporation of the prior domain knowledge is that the performance of these models has smaller variance. In the large sample region (75%) the MLP methods have the best performance, specifically the MLP-informative is still slightly better than the MLP-noninformative.

Table 4 presents results on identifying ''hard'' cases with various Bayesian uncertainty measures. The high AUC values show that uncertainty measures based on the Bayesian prediction (Definition 2) characterize well the uncertainty concepts based on expert errors and assesments (Definition 5).

Fig. 6 shows that without rejection the AUC is 0.937 while in the rejected sets it can be below 0.7. For example, if we set the rejection threshold to exclude 25% of the cases, the AUC rises to 0.97. In practice, this means that a decision support system can be specified to classify 75% of the cases with a lower misclassification rate (from 9 to 6%) and identify the remaining 25% as hard cases that need special considerations. The effect of various rejection methods based on different uncertainty measures is similar and their slightly different characteristics need further investigations.

9. Conclusion

We investigated two approaches to incorporate domain knowledge into Bayesian multilayer perceptrons based on the formalization of the prior knowledge as a Bayesian belief network. The *prior sample* method generates an artificial data set equivalent to a certain ''effective sample size'', which is used together with real data in the Bayesian inference. The *informative prior* method approximates the informative prior of the belief network with a prior for the multilayer perceptron.

These techniques were elaborated for a prototype task—classification in presence of rich prior domain knowledge and moderate sample size—that we presented in Section 2; however many elements of our methods are relevant to a wider range of problems. Our prototype task stemmed from the real-world task of classifying ovarian tumors, which we used to evaluate our approach and demonstrate its effectiveness.

We investigated another advantage of the Bayesian approach—the additional uncertainty information for predictions—in a medical classification problem. We introduced various uncertainty measures for characterizing the Bayesian prediction. The analysis of their correspondence with various human uncertainty assessments based on expert errors and assessments shows their potential for characterizing the ''hard cases''. Subsequently, we demonstrated that a classifier with rejection based on these can exclude a certain subset to improve significantly its performance on the remaining cases.

In future research we plan to investigate the presented prior incorporation methods for belief networks also (as recipient models), where the expert-structure- and literature-based priors for structures can be incorporated in the Bayesian model update beside parameter transformation [5]. Note that, whereas the structure of the current recipient MLP model is fixed, it has a built-in capability for joint refinement of the discretization and the structure with reference to classification.

We conclude with the view that our hybrid methodology offers new insights and challenges on how classification can benefit from the integration of knowledge engineering and machine learning by formalizing, transforming, and incorporating general domain knowledge into the learning of a specialized classifier.

Acknowledgements

Bart De Moor is a full Professor at the K.U. Leuven, Belgium. Dirk Timmerman is an associate professor at the University Hospitals K.U. Leuven. Yves Moreau is an Assistant Professor at the K.U. Leuven and a postdoctoral researcher with the Belgian Fund for Scientific Research (F.W.O.—Vlaanderen). Peter Antal is a Research Assistant with the K.U. Leuven. Geert Fannes is a Research Assistant with the F.W.O.—Vlaanderen. Research supported by GOA-Mefisto 666, IDO, G.0115.01, G.0240.99, G.0407.02, G.0413.03, G.0388.03, G.0229.03, ICCoS, ANMMM, Bil. Int. Collaboration Hungary/Poland, STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA, DWTC, IUAP IV-02 (1996-2001) and IUAP V-22 (2002–2006), CAGE; ERNSI; Data4s, Electrabel, Elia, LMS, IPCOS, VIB.

References

[1] Abu-Mostafa YS. Hints and the VC dimension. Neural Comput 1993;5(2):278–88.

[2] Van Allen T, Greiner R, Hooper P. Bayesian error-bars for belief net inference. In: Breese J, Koller D, editors. Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI-2001). Morgan Kaufmann; 2001. p. 522–9.

[3] Altman RB. Challenges for intelligent systems in biology. IEEE Intell Syst 2002;16(6):14–8.

[4] Antal P. Applicability of prior domain knowledge formalised as Bayesian network in the process of construction of a classifier. In: Proceedings of the 24th Annual Conference of the IEEE Industrial Electronic Society (IECON '98); 1998. p. 2527–31.

[5] Antal P, Fannes G, Moreau Y, Timmerman D, De Moor B. Using literature and data to learn Bayesian networks as clinical models of ovarian tumors. Artif Intell Med, 2003, in press. Special issue on Bayesian Models Med.

[6] Antal P, Fannes G, Verrelst H, De Moor B, Vandewalle J. Incorporation of prior knowledge in black-box models: comparison of transformation methods from Bayesian network to multilayer perceptrons. In: Workshop on Fusion of Domain Knowledge with Data for Decision Support, 16th Uncertainty in Artificial Intelligence Conference; 2000. p. 42–8.

[7] Antal P, Glenisson P, Fannes G, Mathijs J, Moreau Y, De Moor B. On the potential of domain literature for clustering and Bayesian network learning. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (ACM-KDD-2002); 2002. p. 405–14.

[8] Antal P, Meszaros T, De Moor B, Dobrowiecki T. Domain knowledge based information retrieval language: an application of annotated Bayesian networks in ovarian cancer domain. In: Proceedings of the 15th IEEE Symposium on Computer-Based Medical Systems (CBMS-2002); 2002. p. 213–8.

[9] Antal P, Verrelst H, Timmerman D, Moreau Y, Van Huffel S, De Moor B, Vergote I. Bayesian networks in ovarian cancer diagnosis: potential and limitations. In: Proceedings of the 13th IEEE Symposium on Computer-Based Medical System (CBMS-2000); 2000. p. 103-9.

[10] Bakerman R, Gottman JM. Observing interaction: an introduction to sequential analysis. Cambridge: Cambridge University Press; 1986.

[11] Berk, JS, Hacker NF (Eds.). Practical gynecologic oncology. 2nd ed. Baltimore, MD: Williams and Wilkins; 1995.

[12] Bernardo JM. Bayesian theory. Wiley: Chichester; 1995.

[13] Bishop CM. Neural networks for pattern recognition. Oxford: Clarendon Press; 1995.

[14] Blumer A, Ehrenfeucht A, Haussler D, Warmuth MK. Learnability and the Vapnik–Chervonenkis dimension. J ACM 1989;36(4):929–65.

[15] Cheng J, Greiner R. Learning Bayesian belief network classifiers: algorithms and system. Lect Notes Comput Sci 2001;2056:141–51.

[16] Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer. Cancer 1994;73(3):643–50.

[17] Cloete I, Zurada JM. Knowledge-based neurocomputing. Cambridge, MA: MIT Press; 2000.

[18] Cooper GF, Herskovits E. A Bayesian method for the induction of probabilistic networks from data. Machine Learn 1992;9:309–47.

[19] Dagum P, Luby M. Approximating probabilistic inference in Bayesian belief networks is NP-hard. Artif Intell 1993;60:141–53.

[20] Dasgupta S. The sample complexity of learning fixed-structure Bayesian networks. Machine Learn 1997;29:165–80.

[21] Devroye L, Györfi L, Lugosi G. A probabilistic theory of pattern recognition. Berlin: Springer; 1996.

[22] Easton DF, Ford D, Bishop DT. Breast and ovarian cancer incidence in BRCA1-mutation. Am J Hum Genet 1995;56:265–71.

[23] Finkler NJ, Benaceraf B, Lavin PT. Comparison of serum CA 125, clinical impression, and ultrasound in the preoperative evaluation of ovarian masses. Obstetrics Gynecol 1998;72(4):659–63.

[24] Friedman N, Geiger D, Goldszmidt M. Bayesian networks classifiers. Machine Learn 1997;29:131–63.

[25] Friedman N, Koller D. Being Bayesian about network structure. In: Boutilier C, Goldszmidt M, editors. Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence (UAI-2000). Morgan Kaufmann; 2000. p. 201–11.

[26] Friedman N, Yakhini Z. On the sample complexity of learning Bayesian networks. In: Horvitz E, Jensen Finn V, editors. Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence (UAI-1996). Morgan Kaufmann; 2000. p. 274–82.

[27] Geiger D, Heckerman D. A characterization of the Dirichlet distribution with application to learning Bayesian networks. In: Besnard P, Hanks S, editors. Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI-1995). Morgan Kaufmann; 2000. p. 196–207.

[28] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. London: Chapman & Hall; 1995.

[29] Geman S, Bienenstock S, Doursat R. Neural networks and the bias/variance dilemma. Neural Comput 1992;4:1–58.

[30] Granberg S, Wikland M, Jansson I. Macroscopic characterization of ovarian tumors and the relation to the histological diagnosis. Gynecol Oncol 1989;35:139–44.

[31] Hand DJ. Construction and assessment of classification rules. Chichester: Wiley; 1997.

[32] Hanley JA, McNeil BJ. The meaning and use of the area under receiver operating characteristic (ROC) curve. Radiology 1982;143:29–36.

[33] Harris R, Whittemore AS, Itnyre J, and the Collaborative Ovarian Cancer Group. Characteristics relating to ovarian cancer risk (i, ii, iii, iv). Am J Epidemiol 1992;136:1175–1220.

[34] Haussler D. Quantifying inductive bias: AI learning algorithms and Valiant's learning framework. Artif Intell 1988;36:177–221.

[35] Haussler D. Bounds on the sample complexity of Bayesian learning using information theory and the Vapnik-Chervonenkis dimension. Machine Learn 1994;14:83–113.

[36] Heckerman D, Geiger D, Chickering D. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learn 1995;20:197–243.

[37] Huang C, Darwiche A. Inference in belief networks: a procedural guide. Amsterdam: Elsevier. Int J Approx Reason 1996;5:225–63.

[38] Langston AA. Hereditary ovarian cancer. Gynecol Oncol Pathol 1997;9:3–7.

[39] Lucas P. Restricted Bayesian network structure learning. In: Blockeel H, Denecker M, editors. Proceedings of 14th Belgian-Dutch Conference on Artificial Intelligence (BNAIC '02); 2002. p. 211–8.

[40] Mitchell TM. Does machine learning really work? AI Mag 1997;18:11–20.

[41] Müller P, Insua RD. Issues in Bayesian analysis of neural network models. Neural Comput 1998;10:571–92.

[42] Myllymaki P. Mapping Bayesian networks to stochastic neural networks: a foundation for hybrid Bayesian-neural systems. Ph.D. dissertation, University of Helsinki, No. A-1995-1; 1995.

[43] Neal RM. Bayesian learning for neural networks. Berlin: Springer; 1996.

[44] Neal RM. Transferring prior information between models using imaginary data. Technical Report No. 0108, Department of Statistics, University of Toronto, July 2001.

[45] Niyogi P, Poggio T, Girosi F. Incorporating prior information in machine learning by creating virtual examples. Proc IEEE 1998;86(11):2196–209.

[46] Pearl J. Probabilistic reasoning in intelligent systems. San Francisco, CA: Morgan Kaufmann; 1988.

[47] Ripley B. Stochastic simulation. Chichester: Wiley; 1987.

[48] Rüger SM, Ossen A. Clustering in weight space of feedforward nets. In: von der Malsburg C, editor. ICANN 96, Lecture Notes in Computer Science. Berlin: Springer; 1996. p. 83–8.

[49] Shavlik JW. An overview of research at Wisconsin on knowledge-based neural networks. In: Proceedings of the International Conference on Neural Networks; 1996. p. 65–9.

[50] Sowmya R. Theory refinement of Bayesian networks with hidden variables. In: Proceedings of 15th International Conference on Machine Learning; 1998.

[51] Spiegelhalter DJ. Local computations with probabilities on graphical structures and their application to expert systems. J R Statist Soc B 1988;50(2):157–224.

[52] Spiegelhalter DJ, Dawid A, Lawritzen S, Cowell R. Bayesian analysis in expert systems. Statist Sci 1993;8(3):219–83.

[53] Tekay A, Jouppila P. Validity of pulsatility and resistance indices in classification of adnexal tumors with transvaginal color Doppler ultrasound. Ultrasound Obstetrics Gynecol 1992;2:338–44.

[54] Timmerman D. Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology. Ph.D. Dissertation, Leuven University Press, D/1997/1869/70; 1997.

[55] Timmerman D. Artificial neural network models for the pre-operative discrimination between malignant and benign adnexal masses. Ultrasound Obstetrics Gynecol 1999;13:17–25.

[56] Timmerman D, Valentin L, Bourne TH, Collins WP, Verrelst H, Vergote I. Terms, definitions and measurements to describe the sonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group. Ultrasound Obstetrics Gynecol 2000;16(5):500–5.

[57] Towell G, Shavlik J. Knowledge-based artificial neural networks. Artif Intell 1994;70:119–65.

[58] National Cancer Institute (US). SEER cancer data; 1998.

[59] Valentin L. Gray scale sonography, subjective evaluation of the color Doppler image and measurement of blood flow velocity for distinguishing benign and malignant tumors of suspected adnexal origin. Eur J Obstetrics Gynecol Reprod Biol 1997;72:63–72.

[60] Vapnik VN. The nature of statistical learning theory. Berlin: Springer; 1995.

[61] Whittemore AS, Gong G, Itnyre J. Prevalence and contribution of BRCA1 mutations in breast cancer and ovarian cancer. Am J Hum Genet 1997;60:496–504.