Incorporation of Prior Knowledge in Black-Box Models: Comparison of Transformation Methods from Bayesian Network to Multilayer Perceptrons

P. Antal * M.Sc. | G. Fannes * M.Math. | H. Verrelst * M.Eng. | B. De Moor * M.Eng. Ph.D. | J. Vandewalle * M.Eng. Ph.D.

Abstract

Fusing domain knowledge and data aims to exploit two kinds of information and combine the advantages of knowledge engineering and inductive techniques. In this context we tested and analyzed hybrid methods that use Bayesian networks (to incorporate efficiently the prior background knowledge), multi-layer perceptrons (to efficiently exploit the data from the domain) and a connection between these two representations. These techniques can be used to define an "informative" prior or an "informative" cost for black-box models. We compare various hybrid combination methods and suggest a novel solution for the application of the informative prior for black-box models (multilayer perceptrons) that avoids the symmetry problems in the weight space.

1 Introduction

The optimal integration of prior background knowledge and data is a challenging practical and theoretical task in classification problems. The possible approaches to this task can be divided in two related categories: (1) adaptive knowledge-based methods and (2) knowledge-based inductive methods. In the first group the probabilistic domain models, particularly the Bayesian networks, were suggested for the integration: these originally knowledge-based techniques are applied in a more and more adaptive way. Inductive techniques similarly try to utilize the prior knowledge, for example by selecting an optimal function class for classification or defining special prior knowledge-based cost functions [6, 1, 11].

Because of its general nature, the Bayesian network knowledge representation became prevalent to describe the background knowledge and for the incorporation of examples. However, the technique has disadvantages in the function learning context w.r.t. black-box methods such as the sample and computational complexity of learning and the computational complexity of inference.

To combine optimally the strengths of a white-box technique (e.g. Bayesian network (BN)) and a black-box technique (e.g. multi-layer perceptrons (MLP), support vector machines) we analyze the following general *hybrid* strategy for building classifier systems:

1. Apply a fixed structure Bayesian network model with a prior distribution over its parameters to describe the background knowledge with confidence.
2. Use various transformation methods to incorporate this knowledge into a black-box model.
3. Do standard optimization in the classical statistical framework (CSF) using the real data D_n^r or Bayesian simulation in the black-box models in the Bayesian statistical framework (BSF).

We focus on methods mainly in the Bayesian statistical framework, since the inherent subjectivity of prior knowledge makes it more suitable for the integration.

The paper is organized as follows: Section 1 summarizes the approaches to integrate prior domain knowledge and data. Section 2 recapitulates a Bayesian network representation to describe the prior background knowledge in a Bayesian manner. In Section 3 and 4 5 we discuss hybrid methods to use this prior knowledge and their theoretical properties. In Section 6 a Bayesian method is analyzed in which the prior background knowledge represented by a Bayesian network

*Electrical Eng. Dept., Katholieke Universiteit Leuven, Kardinaal Mercierlaan 94, B-3001 Heverlee (Leuven), Belgium

We thank J. Suykens and Y. Moreau for their helpful comments.

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 *Proc. of the Workshop on Fusion of Domain Knowledge with Data for Decision Support, The Sixteenth Conference on Uncertainty in Artificial Intelligence*, Stanford University, USA, Jun. 2000, pp. 7-12., Lirias number: 178783.

is transformed to a prior distribution over the parameters of a multi-layer perceptron model. We analyze problems related to symmetries in the weight space of multi-layer perceptrons that have significant effects on the applicability of multi-layer perceptron models in the Bayesian statistical framework. Section 7 presents results comparing the performance of these methods.

2 Formalization of prior background information

A Bayesian network represents a joint probability distribution over a set of variables (see e.g. [5]). We assume that these are discrete variables, partitioned into three sets $\mathbf{X}$, Y *in* $\{c_0, c_1\}$, $\mathbf{Z}$: set of input, output, and intermediate variables respectively. The model consists of a qualitative part (a directed graph) and quantitative parts (dependency models). The vertices of the graph represent the variables and the edges define the qualitative dependency-independency relations among the variables. There is a dependency model for every vertex (i.e., for the corresponding variable) to describe its probabilistic dependency on the parents (i.e., on the corresponding variables). Assuming parameter independence we use Dirichlet distributions as dependency models [5]. In this case the prior background knowledge is formalized as a single Bayesian network structure and a prior density over the parameters is given by:

$$\begin{aligned} p(\theta) &= \prod_{i=1}^{m} p(\theta_i) = \prod_{i=1}^{m}\prod_{j=1}^{\mathrm{pa}_i} p(\theta_{ij}) \\ &= \prod_{i=1}^{m}\prod_{j=1}^{\mathrm{pa}_i} \mathrm{Dirichlet}(\theta_{ij1},\ldots,\theta_{ijr_i}|N_{ij1},\ldots,N_{ijr_i}) \\ &\propto \prod_{i=1}^{m}\prod_{j=1}^{\mathrm{pa}_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}-1} \end{aligned} \quad (1)$$

where N_{ijk} can be interpreted as the number of previously seen examples in which the value of the ith variable is k with parental configuration $\mathrm{pa}_i = j$ (a one based index for all possible parental configurations).

Under certain conditions [3] the prior background knowledge formalized in the Bayesian network can be interpreted as N prior seen complete cases to be used to quantify the prior Bayesian network's hyperparameters. In practice the network cannot be characterized by a single *prior sample size* N (e.g. because various parts of the model are quantified by different experts or studies).[1]

[1]For simplicity we assume that the prior $\mathcal{P}(\Theta)$ is specified completely in a Bayesian way (e.g., by using Dirichlets priors), but in general point value specifications (or even hybrid specifications) can be allowed and most of our results and methods can be modified to cover this case.

3 Pure or hybrid representations

After constructing a prior domain model, it seems natural to apply the observations on this model (i.e., to fit a Bayesian network model to the observations (in CSF) or to simulate the Bayesian inference with the Bayesian network (in BSF). This method is hindered in general by high sample and computational complexity for learning, the practically discrete nature of the related prior knowledge and its focusedness on a single restrictive structure, as well as the high computational complexity of the inference in Bayesian networks.

Beside the knowledge-based approaches, another methodology to the classification and prediction task is the inductive one. Assume that the black-box model's inputs and output are the same as for the Bayesian network: $\mathbf{X}$ and Y. The black-box techniques in CSF tries to select a function $f(\mathbf{x})$ from a function class C using the data D_n^r that approximates the posterior probability that observation x belongs to class c_1 ($P(c_1|\mathbf{x})$). In the BSF the aim is to faithfully incorporate the prior domain knowledge in a model that is powerful enough to model the data and do an efficient Bayesian inference.

In our context the main deficiency of black-box models is the inability to incorporate prior knowledge (e.g. the lack of an *informative* prior distribution or cost). Consequently the performance characteristics of the model types sharply differ as shown in Figure 4 in Section 7 (for details about the speed of convergence, see e.g. [8] for MLP, [2] for the learning of Bayesian network parameters, [9] for Bayesian network structure). This means that the efficiency of the incorporation of prior domain knowledge and the learning characteristics should be evaluated together for each model type. The following two-step hybrid algorithm aims to combine the pro's of the two methods: use the understandable white-box models to derive and describe efficiently the prior domain knowledge and use black-box techniques to exploit the samples efficiently.

4 Incorporation of prior knowledge

For some problems, the prior knowledge formalized as a Bayesian network can be characterized by a prior virtual sample size N, the number of previously seen examples. The following method generalizes this and provides an universal tool to use the prior knowledge formalized as a Bayesian network in black-box methods.

Omitting the technical details, it is possible to define a mapping $\mathcal{T} : \Theta \to \Omega$ that transforms a prior dis-

tribution $\mathcal{P}(\Theta)$ over the Bayesian network parameter space to a prior probability distribution $\mathcal{Q}(\Omega)$ over the black-box model parameter space . The outline of this mapping is the following: the black-box model $f_\omega(\mathbf{x})$ is used for approximating the conditional distribution of the output class $P(c_1|\mathbf{x})$ conditioned on the input $\mathbf{x}$, which is defined by the Bayesian network. Thus we can define a mapping from every Bayesian network parametrization $\theta \in \Theta$ to the "best" approximating black-box function parametrization $\omega \in \Omega$ (see 6 for practical issues). This link between the two model types makes the suggested hybrid two-step method possible.

5 Methods and their relations

The previous method can be regarded as an optimal solution for the hybrid approach in BSF. Other methods that can be used to implement the hybrid approach are the following:

- Probabilistic combination of the prediction of the prior Bayesian network and the adapted black-box model.
- Generation of prior sample.
- Defining an *informative* cost function (in CSF) or defining an *informative* prior distribution (in BSF).

We focus only on the last two methods.

5.1 Using prior sample

After selecting a well-balanced N that characterizes the confidence w.r.t. the input-output relation, the prior probabilistic domain model can be used to generate random examples. In the CSF the optimal solution for prior sample generation is to use a Bayesian network with point value specification given by the mean of the Dirichlet distributions. Since the generalization error between a target f^* and the selected function $\hat{f}$ w.r.t. the data can be written as $\|f^* - \hat{f}\| = \mathcal{O}(1/n)$ [8] then, if the mean is the "true" conditional distribution $P(c_1|\mathbf{x})$ and input distribution ($\mathcal{D}$), these prior virtual samples can be used as real samples, resulting $\|f^* - \hat{f}\| = \mathcal{O}(1/(n+N))$. Related results about the general case when the mean of the prior only approximates the "true" distribution can be found in [4]. According to the real and prior sample we can decompose the cost function in two terms:

$$\begin{aligned} \mathrm{Cost}(\omega, D) &= \frac{1}{(n+N)} \sum_{i=1}^{n+N} -\ln Br(y_i|f_\omega(\mathbf{x}_i)) \\ &= \frac{q}{n}\mathrm{Cost}^{real}(\omega, n) + (1-q)\lambda'_N(\omega) \end{aligned}$$

Where $Br(y|\zeta) = \zeta^y(1-\zeta)^{1-y}$ denote the Bernoulli distribution. We analyze the "prior" penalty term $\lambda'_N(\omega)$ based on the prior sample in Section 5.2.

To analyze the effect of prior samples in the BSF where the mean is used for sample generation, we can write:

$$\begin{aligned} P(\omega|D_N^p, D_n^r) &\propto P(D_n^r|\omega)P(\omega|D_N^p) \\ &= P(D_n^r|\omega)P'(\omega). \end{aligned} \quad (2)$$

The noninformative prior distribution $P(\omega)$ is transformed into an informative prior $P'(\omega) = P(\omega|D_N^p)$ using the prior sample D_N^p. Similarly, if the prior mean defines the same conditional and input distribution, then the effect of this Bayesian update by the generated prior sample is the same as by real data. In Section 6 we suggest a method that approximates the ideal prior transformation as described in Section 4. It uses a *Bayesian* data generation method in which randomly choosen Bayesian network parametrizations are used to generated blocks of prior samples.

In short, the effect of prior samples in CSF is the same as an *informative prior* penalty term. In the BSF the prior sample transforms the noninformative prior to an informative prior.

5.2 Informative prior and cost

Using the classical data generation method (i.e. using the means of the Dirichlet distributions for prior sample generation) in the BSF, the noninformative prior distribution converges to a Gaussian distribution $N_t(\omega|\mu_N, \Sigma_N)$ ([7], p.291, for conditions see Section 6.1). Continuing Equation 2 we can write

$$\begin{aligned} P(\omega|D_N^p, D_n^r) &\propto P(D_n^r|\omega)P'(\omega) \\ &= P(D_n^r|\omega)N_t(\omega|\mu_N, \Sigma_N). \end{aligned} \quad (3)$$

Equation 3 can be used to interpret the penalty term in the CSF, setting the cost function $C(\omega)$ to $-\ln(P(\omega)|D_N^p)$. Continuing with the Gaussian approximation we can write

$$\begin{aligned} -\ln(P(D_n^r|\omega)N_t(\omega|\mu_N, \Sigma_N)) &= -\ln(P(D_n^r|\omega)) \\ -\frac{1}{2}(\omega-\mu_N)^T\Sigma_N^{-1}(\omega-\mu_N) &- \text{constant} \\ &= -\ln \mathcal{L}(\omega) + \lambda''_N(\omega). \end{aligned}$$

So the convergence to a Gaussian in the BSF shows that with increasing N it is reasonable to assume that the informative penalty term $\lambda''_N(\omega)$ in the CSF is more and more quadratic (N is fixed by the confidence in the prior domain model, it cannot be set arbitrarely large).

Another property of $\lambda'_N(\omega)$ makes it possible to define an exact informative cost function

$$\lambda'(\omega) = \lim_{N\to\infty} \lambda'_N(\omega) = E_{\mathcal{P}}[-\ln Br(y_i|f_\omega(\mathbf{x}_i))]$$

and the cost function $\lambda'_N(\omega)$ based on N prior sample is an estimator of it. Finally, it is possible to derive an ideal prior distribution over the black-box parameter space in BSF (see Section 4). Taking the negative logarithm, we get an exact cost function $\lambda'(\omega)$ in CSF. Figure 1 summarizes the main relations between methods.

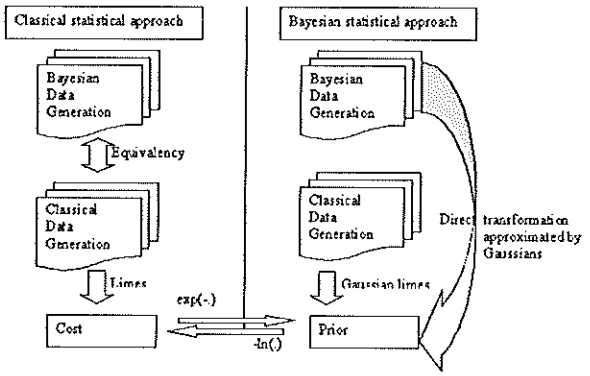

Figure 1: Main relations between methods.

6 Practical issues on the prior transformation to multi-layer perceptrons

In Section 3 we suggested a prior transformation from Bayesian networks to black-box methods.
The main steps for the application of this technique in the case of multi-layer perceptron are shown in Figure 2 and the list below.

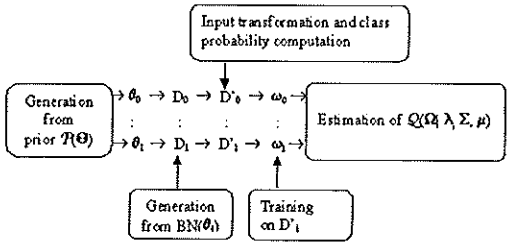

Figure 2: Main steps in transforming the prior

1a. Generate Bayesian network parametrizations $\{\theta_1, \ldots, \theta_l\}$

1b. Generate block of prior samples from each parametrization $\{D_1^p, \ldots, D_l^p\}$

2. Train a multi-layer preceptron for each block of samples resulting in a block of perceptron parametrizations $\{\omega_1, \ldots, \omega_l\}$

3. Estimate the transformed distribution $\mathcal{Q}(\omega)$ from the generated perceptron parametrization with a mixture of Gaussians.

The Bayesian network parametrizations are generated from the Dirichlet distribution by standard methods. The sample blocks are generated according to the drawn Bayesian network parametrizations. Next a simple preprocessing is necessary on the input variables in the generated samples for theoretical (identifibility) and practical reasons (*one-out-of-c* coding scheme for nominal variables and resampling in the discretization intervals for discretized continuous variables). It is advantegous to use the prior probabilistic domain model to compute $P(c_1|\mathbf{x})$ for each sample instead of random generated class labels. These class probabilities eliminate a stochastic element (the class labels), consequently the same performance can be achieved with smaller block size (i.e., with less computational complexity). For training the perceptron model on a block of samples we used the scaled conjugate gradient algorithm [12]. Finally, to estimate the transformed prior distribution $\mathcal{Q}(\boldsymbol{\Omega})$ over the black-box model parameter space $\boldsymbol{\Omega}$ by the trained perceptrons, we used a mixture of Gaussians.

$$\mathcal{Q}(\omega) \approx \sum_{i=1}^{L} \alpha_i N_t(\omega|\mu_i, \Sigma_i)$$

$$0 \le \alpha_i \le 1 \quad \sum_{i=1}^{L} \alpha_i = 1$$

6.1 Symmetries in the parameter space

The total number of symmetries (due to possible permutations and sign symmetries) in a multi-layer perceptron with k hidden layers and L_i neurons in layer i is given by $\prod_{i=1}^{k} 2^{L_i} L_i!$. Based on these symmetries it is possible to define a canonical transformation $\mathcal{C}(.)$ by making all biases positive and ordering the nodes in each layer increasingly w.r.t. the biases. We call the range of this transformation the canonical subspace, and $\mathcal{C}(\omega)$ is the canonical (unique) parametrization for $f_\omega(.)$.

An important consequence is that a *consistent distribution*[2] $\mathcal{Q}^*(\omega)$ has identical regions and an op-

[2] If $f_{\omega_1}(.)$ and $f_{\omega_2}(.)$ are equal w.r.t. the L_2 func-

timal density estimation method should exploit that fact.

6.2 Definition and comparison of methods for estimation

A possible solution for exploiting the regularities in the distribution $\mathcal{Q}^*(\omega)$ is the elimination of the symmetries by transforming the generated perceptron parametrizations to a small number of compact clusters, i.e. transforming $\mathcal{Q}^*(\omega)$ to an better estimatable *base distribution*[3]:

The naïve method uses the original data set to estimate the base distributon $\mathcal{Q}^*$. Because of its symmetries w.r.t. origin, the estimated mean will tend to zero as the number of generated networks increases.

The canonical method estimates the base distribution $\mathcal{C}(\mathcal{Q}^*(\omega))$. It transforms the data set to the canonical subspace by applying $\mathcal{C}(.)$. Since this transformation is not continuous, it causes scattering which deteriorates the estimatibility.

The exhaustive method transforms the parametrizations to clusters with minimal within-cluster variance, though this problem is NP hard in L_i.

Heuristic methods can be found easily which construct sub-optimally contracted clusters in a tractable way.

7 Results

For the multi-layer perceptrons, the existence and the quality of an *informative* prior is crucial for any Bayesian method. Therefore we present the performance of various algorithms for density estimation over the multi-layer perceptron weight space. The effect of the incorporation of prior domain knowledge is demonstrated using the "prior transformation" technique in BSF on a real world classification task [10].

7.1 Handling symmetries in prior transformation

The following artificial example illustrates the properties of the methods described in the previous section. Figure 3 shows the Bayesian network structure with its hyperparametrization and the perceptron model for the approximation of the conditional distribution. The pictures shows the input-output mappings $f_{\omega_{1\ldots5}}(.)$ of the means $\omega_{1\ldots5}$ of the estimated Gaussian distributions $\mathcal{Q}_{1\ldots5}(\omega)$ and the average variances $E_{\mathcal{P}(\mathbf{X})}[Var(f_{\omega_{1\ldots5}}(\mathbf{X}))]$ in parenthesis.

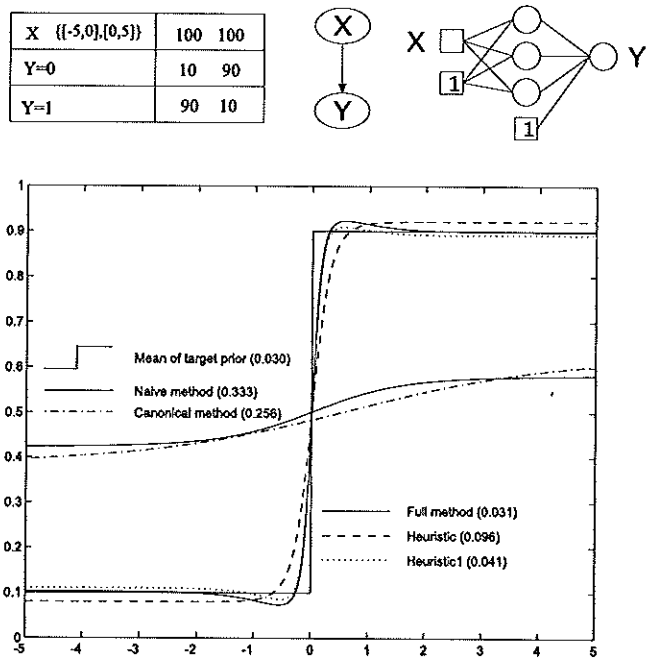

Figure 3: Effect of various symmetry dealing methods on density estimation

As can be seen the *naïve* and *canonical* methods perform poorly (with also large average variances), while the *exhaustive* and its *heuristic* approximations perform much better (additionally with small average variances). Since the same arguments can be extended to multi-mode densities, we can state in general that the symmetries in the perceptron weigth space has enormous impact on the application of such neural models in the Bayesian statistical framework.

7.2 Effect of priors on the performance

We compared the performance of four classifiers. A Bayesian network with a noninformative prior and updated hypers from data. A Bayesian network with informative prior and updated hypers from data A MLP with a noninformative prior and data. A MLP with the transformed informative prior described in 6 and data. For the MLPs, the hybrid MCMC method was used to perform the Bayesian inference [14, 13].

As Figure 4 shows, the effect of the prior depends on the real sample size: it has a large advantageous effect in the small sample region ($[0-0.4]$) and the prior has no restrictive effect in the large sample range.

tion norm, then $\mathcal{Q}(\omega_1)$ should be equal to $\mathcal{Q}(\omega_2)$ almost everywhere.

[3] A base distribution is a distribution so that the superposition of all invariant transformations gives the consistent $\mathcal{Q}^*$.

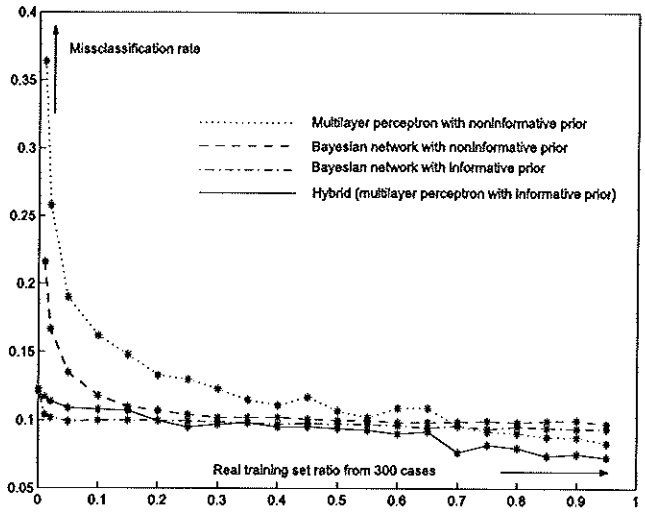

Figure 4: Missclassification rate of various pure and hybrid methods discriminating malignant and benign ovarian masses [10].

8 Conclusions

In the paper a general methodology was suggested to transform the prior domain knowledge formalized as a Bayesian network into black-box models, offering various methods to use black-box models with prior background knowledge. A novel approach to derive informative prior distribution for MLPs was presented that avoids the problems caused by the symmetries in the MLP weight space. Results about the applicability of the suggested methods are presented in an artificial and in a real-world example.

Acknowledgements

Joos Vandewalle is Full Professor at the K.U.Leuven. Bart De Moor is a Research Associate at the F.W.O. (Fund for Scientific Research-Flanders). Peter Antal is a Research Assistant with the K.U.Leuven. Geert Fannes is a Research Assistant with the F.W.O. Vlaanderen. Herman Verrelst is a Research Assistant with the I.W.T. (Flemish Institute for Scientific and Technological Research in Industry). This work was carried out at the ESAT laboratory and supported by grants and projects from the Flemish Government: Concerted Research Action GOA-MEFISTO-666 (Mathematical Engineering for Information and Communication Systems Technology) and F.W.O. project G.0262.97: Learning and Optimization: an Interdisciplinary Approach and the F.W.O. Research Communities: ICCoS (Identification and Control of Complex Systems) and Advanced Numerical Methods for Mathematical Modelling and Bilaterale Wetenschappelijke en Technologische Samenwerking Flanders-Hungary, BIL96/19; from the Belgian State, Prime Minister's Office-Federal Office for Sci., Tech. and Cult. Affairs-Interuniversity Poles of Attraction Programme (IUAP P4-02 (1997-2001): Modeling, Identification and Control of Complex Systems. The scientific responsibility is assumed by its authors.

References

[1] Y. S. Abu-Mostafa, *Hints and the vc dimension*, Neural Computation (1993), no. 5, 278–288.

[2] S. Dasgupta, *The sample complexity of learning fixed-structure bayesian networks*, Machine Learning, vol. 29, Kluwer Academic Publishers, 1997, pp. 165–180.

[3] D. Geiger et al., *A characterization of the dirichlet distribution with application to learning bayesian networks*, Proc. of the 11th UAI Conf., Morgan Kauffman Publishers, Houston, 1995, pp. 196–207.

[4] D. Haussler et al., *Bounds on the sample complexity of bayesian learning using information theory and the vapnik-chervonenkis dimension*, Machine Learning 14 (1994), 83–113.

[5] E. Castillo et al., *Expert systems and probabilistic network models*, Springer, 1997.

[6] G. G. Towell et al., *Knowledge-based artificial neural networks*, Artificial Intelligence (1994), no. 70, 119–165.

[7] J. M. Bernardo et al., *Bayesian theory*, Wiley & Sons, 1995.

[8] L. Devroye et al., *A probabilistic theory of pattern recognition*, Springer-Verlag, 1996.

[9] N. Friedman et al., *On the sample complexity of learning bayesian networks*, Proc. of the 12th UAI Conf., 1996.

[10] P. Antal et al., *Bayesian networks in ovarian cancer diagnosis: Potential and limitations*, CBMS2000, 2000, Houston, p. Accepted.

[11] P. Niyogi et al., *Incorporating prior information in machine learning by creating virtual examples*, Proc. of the IEEE, vol. 86-11, 1998.

[12] M. F. Moller, *A scaled conjugate gradient algorithm for fast supervised learning*, Neural Networks 6 (1993), 525–533.

[13] P. Müller and R. D. Insua, *Issues in bayesian analysis of neural network models*, Neural Computation 10 (1998), 571–592.

[14] R. M. Neal, *Bayesian learning for neural networks*, Springer, 1996.