Katholieke Universiteit Leuven

Departement Elektrotechniek

ESAT-SISTA/2000-03

Bayesian predictive models for Ovarian Cancer Classification: evaluation of Logistic Regression, Multi Layer Perceptron and Belief Network models in the Bayesian Context [1]

Peter Antal and Geert Fannes and Bart Demoor and Joos Vandewalle and Sabine Vanhuffel [2] and Dirk Timmerman

December 13

Accepted for the Tenth Belgian-Dutch Conference on Machine Learning, Tilburg University

[1]This report is available by anonymous ftp from *ftp.esat.kuleuven.ac.be* in the directory *pub/sista/fannes/reports/Bayesian-OC.ps.gz*

[2]K.U.Leuven, Dept. of Electrical Engineering (ESAT), Research group SISTA, Kasteelpark Arenberg 10, 3001 Leuven, Belgium, Tel. 32/16/32 17 09, Fax 32/16/32 19 70, WWW: *http://www.esat.kuleuven.ac.be/sista*. E-mail: *joos.vandewalle@esat.kuleuven.ac.be*.

Abstract

Bayesian statistics provides a natural framework to incorporate prior knowledge, compare model classes and to provide predictions with uncertainty information. We perform a Bayesian analysis in a classification problem to discriminate between benign and malignant ovarian masses. Belief Network, Multi Layer Perceptron and Logistic Regression models are used with non-informative and with newly introduced informative priors. Additionally, we investigate a method to handle missing values based on a related prior domain model. We assume that the prior domain model is formalized as a Belief Network, since this representation allows an optimal and feasible solution to handle missing input values and incorporate prior knowledge into black-box regression models through informative priors.
In this paper we evaluate these two applications of Belief Network models by comparing the performance of various models in the Bayesian framework and summarize the practical advantages of the Bayesian approach in this medical classification problem.

Bayesian predictive models for Ovarian Cancer Classification: evaluation of Logistic Regression, Multi Layer Perceptron and Belief Network models in the Bayesian Context

P. Antal	G. Fannes	B. De Moor	J. Vandewalle	S. Van Huffel	D. Timmerman
M. Sc.	M. App. Math.	Prof. Dr. M. Eng.	Prof. Dr. M. Eng.	Prof. Dr. M. Eng.	Prof. Dr. Med.

[ ]

Abstract

Bayesian statistics provides a natural framework to incorporate prior knowledge, compare model classes and to provide predictions with uncertainty information. We perform a Bayesian analysis in a classification problem to discriminate between benign and malignant ovarian masses. Belief Network, Multi Layer Perceptron and Logistic Regression models are used with non-informative and with newly introduced informative priors. Additionally, we investigate a method to handle missing values based on a related prior domain model. We assume that the prior domain model is formalized as a Belief Network, since this representation allows an optimal and feasible solution to handle missing input values and incorporate prior knowledge into black-box regression models through informative priors.
In this paper we evaluate these two applications of Belief Network models by comparing the performance of various models in the Bayesian framework and summarize the practical advantages of the Bayesian approach in this medical classification problem.

1 Introduction

The Bayesian approach is becoming more attractive for the machine learning community because it can cope with the valuable subjective prior information in a principled way, it has an implicit model complexity regularization mechanism and it provides more detailed information for decision support. Its principled nature originates from the fact that it expresses all forms of uncertainty in terms of probabilities. The statistical inferences including learning and prediction are achieved by applying the rules of probability theory.

Starting with a prior distribution expressing the initial beliefs concerning the parameter values of the model, we can use the observations to transform this into the posterior distribution for the model parameters expressing the beliefs after observing the data. Using this posterior distribution over the model parameters, useful random variables can be defined for functions depending on the model parameters, like predictions and error measures.

In a binary classification task this means the following. We are primarily interested in the correct classification of an observation $\boldsymbol{x} \in \mathbb{R}^l$. This can be achieved by constructing a *binary decision function* $g(\boldsymbol{x}, \boldsymbol{\omega}) \in \{0, 1\}$ where $\boldsymbol{\omega} \in \Theta$ are the model parameters. A more informative predictive model provides not only a class label, but also the class probabilities, though it is a more complex task both from a statistical and computational point of view. As a further step in improving the decision support, uncertainty information can be provided for the class probabilities, for example the posterior distribution of class probabilities in the Bayesian framework. In this paper we follow the Bayesian approach to solve the classification problem for two main reasons: to incorporate prior background information in a general and principled way and to provide detailed information with clear semantics for decision support. Therefore we use a *probabilistic regression* model $P(T = 1|\boldsymbol{x}, \boldsymbol{\omega}) = f(\boldsymbol{x}, \boldsymbol{\omega})$ with model parameters $\boldsymbol{\omega} \in \Theta$ and a prior distribution $p_\Omega(.)$ over Θ. We assume a supervised learning scheme, that is the existence of a labeled training set $\underline{\boldsymbol{d}} = \{\boldsymbol{x}_k, t_k\}, \;\; (\boldsymbol{x}_k, t_k) \in \mathbb{R}^l \times \{0, 1\}$

Using the observed data $\underline{\boldsymbol{d}}$ and applying Bayes' rule, the prior distribution can be transformed to the posterior distribution

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

Once we have this posterior distribution for the model parameters, we can define random variables related to predictions, performance, etc. In classification problems for example, we are interested, for a given $\boldsymbol{x}$, in the random variable $f(\boldsymbol{x},\Omega)$ where Ω is a random vector in a probability space $(\Theta, B, P_\Omega(.))$ or $(\Theta, B, P_{\Omega|\underline{\boldsymbol{d}}}(.))$. In this way we have uncertainty information about the predicted class probability.

We can simplify this result to scalar values for the class probabilities $P(T=1|\boldsymbol{x},\underline{\boldsymbol{d}})$. The optimal step back depends on the cost function attached to the reported scalar value. Assuming the L_2 loss function, the optimal strategy is to report the mean $E_{\Omega|\underline{\boldsymbol{d}}}[f(\boldsymbol{x},\boldsymbol{\omega})] = f(\boldsymbol{x})$. A further simplification is to discretize this scalar value using some user specified threshold λ, deriving a binary decision function $g_\lambda(.)$.

$$g_\lambda(\boldsymbol{x}) = \begin{cases} 1 \text{ if } E_{\Omega|\underline{\boldsymbol{d}}}[f(\boldsymbol{x},\boldsymbol{\omega})] \geq \lambda \\ 0 \text{ else} \end{cases}$$

These three distinct levels ($P(T=1|\boldsymbol{x},\Omega)$,$P(T=1|\boldsymbol{x},\omega)$ and $g_\lambda(\boldsymbol{x})$) provide different possibilities for decision support. This is illustrated by the increasing information provided by the class labels, class probabilities and distributions of class probabilities, though the burdening statistical and computational complexity should be considered too.

The paper is organized as follows: Section 1 reviews the Bayesian approach in classification problems. Section 2 recapitulates the medical problem which will serve as a test case, introduces the data and defines relevant performance measures. In Section 3 we discuss the applied Bayesian models, particularly the applied prior distributions and the missing value management. In Section 4 we describe shortly the algorithms used to approximate the Bayesian performance measures. Section 5 presents the performance of the models and the comparison of the models in the Bayesian framework. In Section 6 we summarize the advantages of Bayesianism in this classification problem and we evaluate the proposed informative priors and informative missing value management.

2 Classification of Ovarian Masses

Ovarian malignancies represent the greatest challenge among the gynaecologic cancers. A reliable preoperative prediction in terms of benign and malignant ovarian tumors would be of considerable help to clinicians selecting an appropriate treatment. There are two sources of information to construct such predictive models: the prior knowledge and the data.

The available relevant medical literature and expert knowledge is abundant and very diverse (for an overview, see [?]). In addition to the prior background information, data were collected prospectively from 300 consecutive patients who were referred to a single institution (University Hospitals Leuven, Belgium) from August 1994 until June 1997. The data collection protocol ensure that the patients had an apparent persistent extrauterine pelvic mass and excludes other causes that may have similar symptoms such as infection or pregnancy, so the primary aim is differentiation between benign and malignant masses (for a detailed description, see [?]). Univariate statistics of data set are presented in Table 1

	Age	Parity	CA 125	Col score	Resistance index	Ascites	Locularity	Bilat
$\widehat{E}[.\|Benign]$	47.77	1.50	110.34	1.98	0.12	0.20	1.39	0.12
$\widehat{E}[.\|Malignant]$	58.62	1.57	1222.299	3.20	0.41	1.07	2.86	0.40
$\widehat{Std}[.\|Benign]$	15.60	1.40	976.56	0.84	0.77	0.54	1.38	0.33
$\widehat{Std}[.\|Malignant]$	15.18	1.73	3779.64	0.95	0.46	0.94	1.20	0.49

Table 1: Univariate statistics for the ovarian cancer data set.

Standard statistical studies indicate that a multi-modal approach – the combination of various types of variables – is necessary for the discrimination between benign and malignant tumors. Therefore Logistic Regression models, Multi Layer Perceptron models and Belief Network models were previously applied [**?**, 2]. These models predicted the scalar class probabilities and they were developed and tested in the classical statistical framework.

A natural step to provide more detailed information for medical decision support is to apply the Bayesian approach to provide the distribution of class probabilities. We can use the classical statistical performance measures for the evaluation of the models in the Bayesian framework, since any performance measure is a function of the model parameters (for fixed observations/test data). These performance measures then become random variables which provide more information than a point estimate.

Because in the medical literature the Receiver Operator Characteristics (ROC) curve was advocated to assess and compare the performance of probabilistic classifiers, we use three performance measures: miss-classification rate $MR(\Omega, \underline{\boldsymbol{d}})$, ROC curve $ROC(.,\Omega, \underline{\boldsymbol{d}})$ and area under the ROC curve $AUC(\Omega, \underline{\boldsymbol{d}})$ (for the definition and interpretations of the ROC curve, see e.g.[**?**]).

For the comparison of two models $m_1, m_2 \in M$, we use the Bayes factors defined as

$$Ratio = \frac{P(M = m_1|\underline{\boldsymbol{d}})}{P(M = m_2|\underline{\boldsymbol{d}})} = \frac{P_{\underline{\boldsymbol{D}}}(\underline{\boldsymbol{d}}|M = m_1)}{P_{\underline{\boldsymbol{D}}}(\underline{\boldsymbol{d}}|M = m_2)} \times \frac{P(M = m_1)}{P(M = m_2)}$$

In brief, the application of Bayesianism in this medical problem offers the possibility of the incorporation of prior background knowledge, more detailed predictions for decision support and a principled way to report and evaluate model performances.

3 Applied Bayesian Models

A Logistic Regression (LR), a Multi Layer Perceptron (MLP) and a Belief Network (BN) model are discussed in the paper. Since the Logistic Regression models can be interpreted as special Multi Layer Perceptron models without hidden layers and sigmoidal activation function, we refer to these probabilistic regression models as MLPs. For each model we define and investigate two types of prior distribution: *informative* and *non-informative* priors, depending on the amount of incorporated prior background knowledge. Additionally, an auxiliary probabilistic domain model is introduced for the MLP models, which are not capable of coping with incomplete input samples. This prior probabilistic domain model is not updated by the observed samples and is only used for coping with missing input values in the original data set[1]. Table 2 shows the components of the investigated Bayesian models.

Input-output models	BN	MLP	LR
Missing value model	implicit	prior BN	
Prior distribution	Informative or Non-informative		

Table 2: Parts of Bayesian models

The following combinations are used in the paper.

3.1 Belief network models

A Belief Network model defines a joint probability distribution over the domain variables. It is a probabilistic *white-box* model, since it consists of a graph model about the conditional independencies of the domain variables and a quantitative part specifying the conditional probabilities for the domain variables. Fig. 1 shows the graphical structure of the used Belief Network model. For details, please see [2]. It is worth to note from a computational point of view that all the variables are discrete.

Because of the white box nature of the Belief Network models, the prior knowledge available from experts and the literature can be directly represented in an *informative* prior distribution for this model using the following

[1]Note that the BN missing value mechanism is fundamentally different from this in two important respects: first, the mechanism is immediately provided since a BN model defines a joint probability distribution, second the prior distributions for Belief Network models (and therefore the underlying mechanism to cope with the missing values) is updated by the data.

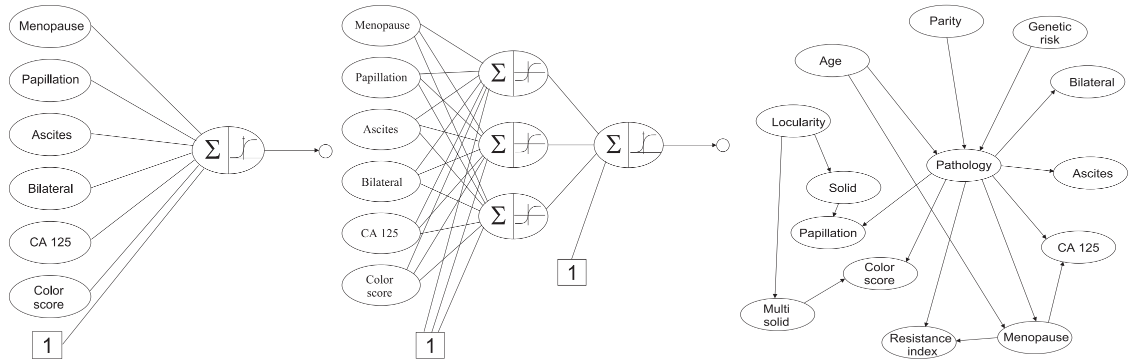

Figure 1: The LR, MLP and BN Network model structures.

technique. Assuming that the parameter independence holds, we use the Dirichlet family to represent the prior distribution [1]. Using a Dirichlet, an expert can express his belief in parametrizations, instead of giving only point estimation for the parameters. The *non-informative* distribution for Belief Network is a uniform distribution corresponding to a Dirichlet distribution with all hyperparameters equal to one.

Since a Bayesian network model is inherently capable of performing inference for incomplete cases, there is no additional probabilistic model for handling missing values. Consequently, we evaluate two Belief Network models: the same structure with *informative* and *non-informative* prior distribution.

3.2 Logistic Regression and Multi-Layer Perceptron models

These probabilistic regression models are defined over a continuous input space specifying an input-output mapping which can be interpreted as a conditional probability

$$f(.,\boldsymbol{\omega}) \; : \; \mathbb{R}^l \;\rightarrow\; [0,1] \quad (1)$$

$$\boldsymbol{x} \;\rightarrow\; f(\boldsymbol{x},\boldsymbol{\omega}) = P(T=1|\boldsymbol{x},\boldsymbol{\omega}) \quad (2)$$

The MLP and LR model structures used in the paper are shown on Fig. 1 (for details about data preprocessing, input features selection and MLP structures please see [**?**, **?**]).

Since the Multi Layer Perceptron (MLP) model is a *black-box* model, it is not possible to specify directly a prior distribution incorporating the available prior knowledge in a general way. To solve this problem we use the Belief Network models as a tool to acquire and represent the prior domain knowledge. The prior distribution for the BN is then transformed into a prior distribution for the MLP parameters. This complex estimation method results in a mixture of Gaussians over the MLP parameter space Θ expressing the prior background knowledge (see [**?**] for details). The *non-informative* prior for the MLP model is a non-restrictive, complexity based prior $\mathcal{N}(\mathbf{0},\sigma^2\boldsymbol{I})$ (cfr. weight decay for neural networks in the classical framework).

As mentioned previously, an auxiliary probabilistic domain model for the MLP models is introduced, which is represented as a Belief Network[2]. We use the prior Belief Network described earlier with its most probable a priori parametrization. We assume that the missing at random condition [**?**] holds

$$P(\boldsymbol{\mathcal{I}}|\boldsymbol{X}^{obs},T,\boldsymbol{X}^{miss},\boldsymbol{Z}) = P(\boldsymbol{\mathcal{I}}|\boldsymbol{X}^{obs},T,\boldsymbol{Z})$$

Here $\boldsymbol{\mathcal{I}}$ is a binary random vector, where $\mathcal{I}_i$ denotes the observed or missing status of the i^{th} variable. $\boldsymbol{X}$, T and $\boldsymbol{Z}$ are respectively the input variables, the output variable (i.e. class label) and those variables that occur in the data set and in the attached prior domain model, but not in the MLP model. Note, that the extended MLP models can use the additional variables $\boldsymbol{Z}$ to cope with incomplete input samples. We denote variables with observed and missing values with *obs* and *miss* superscript (e.g. $\boldsymbol{X}^{miss}$ denotes the input variables with unknown value).

[2]Note that the applied Belief Network is discrete valued, so we use additional probabilistic models for transformation between discrete and continuous values.

Now we can write

$$
\begin{aligned}
P(\boldsymbol{X}^{miss}|\boldsymbol{X}^{obs},\ T,\ \boldsymbol{Z},\ \boldsymbol{\mathcal{I}}) &= \frac{P(\boldsymbol{X}^{miss},\ \boldsymbol{\mathcal{I}}|\boldsymbol{X}^{obs},\ T,\ \boldsymbol{Z})}{P(\boldsymbol{\mathcal{I}}|\boldsymbol{X}^{obs},\ T,\ \boldsymbol{Z})} && (3)\\
&= \frac{P(\boldsymbol{\mathcal{I}}|\boldsymbol{X}^{miss},\ \boldsymbol{X}^{obs},\ T,\ \boldsymbol{Z})\ P(\boldsymbol{X}^{miss}|\boldsymbol{X}^{obs},\ T,\ \boldsymbol{Z})}{P(\boldsymbol{\mathcal{I}}|\boldsymbol{X}^{obs},\ T,\ \boldsymbol{Z})} && (4)\\
&= P(\boldsymbol{X}^{miss}|\boldsymbol{X}^{obs},\ T,\ \boldsymbol{Z}) && (5)
\end{aligned}
$$

Since $P(\boldsymbol{X}^{miss}|\boldsymbol{X}^{obs},\ T,\ \boldsymbol{Z},\ \mathrm{BN})$ can be computed from the auxiliary Belief Network, we can use both the complete and incomplete samples for Bayesian learning. The posterior is given by

$$
\begin{aligned}
& p_\Omega(\boldsymbol{\omega}|\boldsymbol{t}, \underline{\boldsymbol{x}}^{obs}, \underline{\boldsymbol{z}}, \mathrm{BN}) && (6)\\
= & \int p_\Omega(\boldsymbol{\omega}|\boldsymbol{t}, \underline{\boldsymbol{x}}^{obs}, \underline{\boldsymbol{x}}^{miss}) \prod_{i=1}^{n} dP_{\boldsymbol{X}^{miss}}(\boldsymbol{x}_i^{miss}|t_i, \boldsymbol{x}_i^{obs}, \boldsymbol{z}_i, \mathrm{BN}) && (7)
\end{aligned}
$$

where we expanded with $\underline{\boldsymbol{x}}^{miss} = \boldsymbol{x}_1^{miss}, \ldots, \boldsymbol{x}_n^{miss}$ and assumed that the samples are independent.

The inference for these probabilistic regression models with auxiliary probabilistic models can be written as follows

$$
\begin{aligned}
P(T=1|\boldsymbol{x}^{obs},\ \boldsymbol{z},\ \boldsymbol{\mathcal{I}},\ \boldsymbol{\omega}) &= \int P(T=1|\boldsymbol{x}^{obs},\ \boldsymbol{z},\ \boldsymbol{\mathcal{I}},\ \boldsymbol{\omega},\ \boldsymbol{x}^{miss}) dP_{\boldsymbol{X}^{miss}}(\boldsymbol{x}^{miss}|\boldsymbol{x}^{obs},\ \boldsymbol{z},\ \boldsymbol{\mathcal{I}}) && (8)\\
&= \int f(\boldsymbol{x}, \boldsymbol{\omega}) dP_{\boldsymbol{X}^{miss}}(\boldsymbol{x}^{miss}|\boldsymbol{x}^{obs},\ \boldsymbol{z}) && (9)
\end{aligned}
$$

where we used the missing at random assumption.

Note that the usage of the auxiliary Belief Network model is different in the definition of the posterior distribution and in the definition of $f(\boldsymbol{x}^{obs}, \boldsymbol{\omega})$: $P_{\boldsymbol{X}^{miss}}(\boldsymbol{x}^{miss}|t, \boldsymbol{x}^{obs}, \boldsymbol{z})$ versus $P_{\boldsymbol{X}^{miss}}(\boldsymbol{x}^{miss}|\boldsymbol{x}^{obs},\ \boldsymbol{z})$.

Recapitulate it briefly, we investigate four probabilistic regression models: LR and MLP model with informative or non-informative priors, each with the prior auxiliary BN model.

4 Inference Algorithms

Various stochastic methods are used to estimate the distribution of the prediction and performance related random variables. An additional problem is that the computation of the functions defining these random variables are complex in itself.

In the case of the Belief Network prediction, the computation of the class probability is an NP-hard problem both in exact or in stochastic formulation. In the case of MLP prediction, we have to average out the unknown inputs for an incomplete input sample, it means that we have to sample the corresponding values from the auxiliary Belief Network. Then we compute the miss-classification rate, the ROC curve and the area under the ROC curve based on these predictions.

Omitting the details of the computations, Table 3 summarizes the used algorithms.

4.1 Computation of Prediction and Performance Measures for BN

The target random variables are $P_{\mathrm{BN}}(T=1|\Omega, \boldsymbol{x}^{obs}, \boldsymbol{z}, \underline{\boldsymbol{d}}), AUC_{\mathrm{BN}}(\Omega, \underline{\boldsymbol{d}})$ and $MR_{\mathrm{BN}}(\Omega, \underline{\boldsymbol{d}})$. We sample the posterior distribution $p_\Omega(\boldsymbol{\omega}|\underline{\boldsymbol{d}})$ by direct sampling from the analytic posterior and compute the conditional probabilities of malignancy for the drawn parametrizations by an exact inference algorithm. Based on these predictions the corresponding AUC and MR values can be computed.

4.2 Computation of Prediction and Performance Measures for MLP

Now the target random variable are $P_{\mathrm{MLP}}(T=1|\Omega, \boldsymbol{x}^{obs}, \boldsymbol{z}, \underline{\boldsymbol{d}}), AUC_{\mathrm{MLP}}(\Omega, \underline{\boldsymbol{d}})$ and $MR_{\mathrm{MLP}}(\Omega, \underline{\boldsymbol{d}})$.

Random variable	**Function computation**	**Bayesian simulation over Θ**
$P_{\mathrm{BN}}(T=1\|\Omega, \boldsymbol{x}^{obs}, \boldsymbol{z}, \underline{\boldsymbol{d}})$	PPTC [1]	Direct sampling from the analytic posterior
$P_{\mathrm{MLP}}(T=1\|\Omega, \boldsymbol{x}^{obs}, \boldsymbol{z}, \underline{\boldsymbol{d}})$	MLP+direct sampling for averaging out missing values	Hybrid Monte Carlo method[3]
$AUC_{\mathrm{BN}}(\Omega, \underline{\boldsymbol{d}})$	PPTC+AUC computation	Direct sampling from the analytic posterior
$AUC_{\mathrm{MLP}}(\Omega, \underline{\boldsymbol{d}})$	MLP+direct sampling for averaging out missing values+AUC computation	Hybrid Monte Carlo method
$MR_{\mathrm{BN}}(\Omega, \underline{\boldsymbol{d}})$	PPTC+MR computation	Direct sampling from the analytic posterior
$MR_{\mathrm{MLP}}(\Omega, \underline{\boldsymbol{d}})$	MLP+direct sampling for averaging out missing values+MR computation	Hybrid Monte Carlo method

Table 3: Algorithms

At first we sample the unknown input variables conditioned on all the knowns using the auxiliary probabilistic BN model described by 3. Then we sample from the posterior distribution based on the completed data set $p_\Omega(\boldsymbol{\omega}|\boldsymbol{t}, \underline{\boldsymbol{x}}^{obs}, \underline{\boldsymbol{x}}^{miss})$ by the Hybrid Monte Carlo method. Finally we compute the predictions for the test set and the corresponding AUC and MR values for the drawn parametrization by averaging over the missing input variables using the auxiliary Belief Network as defined in Equation 8.

5 Results

We report results for the six models described in Section 3.1 using the algorithms summarized in Section 4. We partitioned the data set described in Section 2 randomly to a test (30%) and training (70%) set, the reported results are based on the test set. This was repeated 30 times to eliminate dependency on separation. The simulations were carried out by the Simulation Environment for Bayesian and Neural Networks (SEBANN). The model classes are denoted as follows: Logistic Regression model with *informative* prior (LR-I) and with *non-informative* (LR-N), for Multi Layer Perceptron model with *informative* prior (MLP-I) and with *non-informative* (MLP-N) and for Belief Network model with *informative* prior (BN-I) and with *non-informative* prior (BN-N).

The Bayesian ROC curves are shown on Fig. 2 with the mean ROC curve $E_{\Omega|\underline{\boldsymbol{d}}}[ROC(x, \boldsymbol{\omega})]$ and the confidence curves $E_{\Omega|\underline{\boldsymbol{d}}}[ROC(x, \boldsymbol{\omega})] \pm \sqrt{Var_{\Omega|\underline{\boldsymbol{d}}}[ROC(x, \boldsymbol{\omega})]}$.

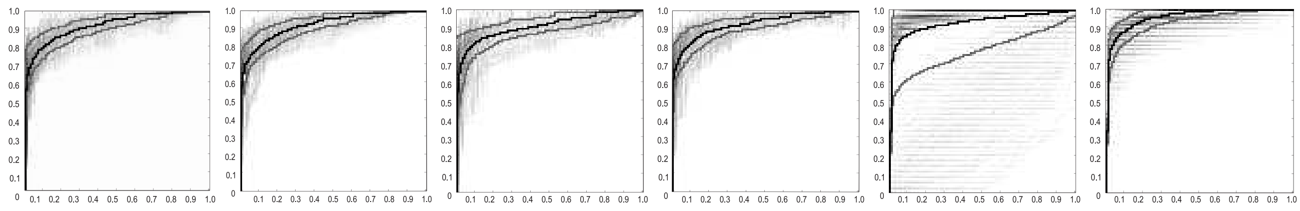

Figure 2: Bayesian ROC curves together with mean and confidence curves.

For comparing the performance of the models Fig. 3 shows for each of the six models the distribution of the area under the ROC Curve and the variance of the prediction for the correct and incorrect the predictions. Note that most of the outliers in the Logistic Regression with ınoninformative prior (number 5) are omitted to concentrate on the most interesting region. A more detailed plot about prediction error (L_1 distance) versus the certainty is given for the MLP with *informative* prior.

Note that the boxplots show the medians and quartiles of the ROC areas and misclassification rates. Table 4

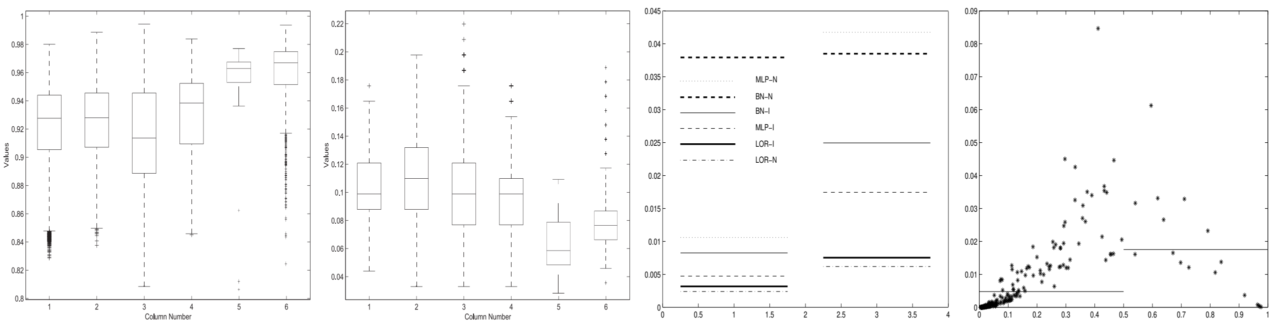

Figure 3: The distributions of the area under the ROC curve, the correctness versus certainty for all the models and prediction error (L_1 distance) versus certainty for MLP-I.

presents the means and variances.

E[.]/Var[.]	LR-N	LR-I	MLP-N	MLP-I	BN-N	BN-I
MR	0.104/0.0283	0.107/0.0301	0.0995/0.0354	0.0958/0.0266	0.101/0.139	0.078/0.026
ROC	0.923/0.0296	0.927/0.0268	0.913/0.0392	0.930/0.0295	0.932/0.151	0.959/0.027

Table 4: The means and variances for the distributions of the miss classification rate and area under the roc curve.

The Bayes-factors for the model classes are presented in Table 5. We assume uniform prior probabilities for the model classes and present the ratios of the likelihoods only against the smallest model likelihood. To eliminate the effect of the sample size we take the n^{th} root of the likelihoods where $n = 300$ is the size of the sample.

	LR-N	LR-I	MLP-N	MLP-I	BN-N	BN-I
	1.29	1.39	1.29	1.45	1.00	1.19

Table 5: Bayes-factors for model comparison as $\sqrt[n]{\frac{P(\underline{\boldsymbol{d}}|.)}{P(\underline{\boldsymbol{d}}|\mathrm{BN-N})}}$.

The results in Table 5 indicate that the posterior probability of MLP and LR models with informative priors are the highest, but the Bayesian model comparison needs further investigation, such as the selection of appropriate priors for the models.

Finally, we performed a manual evaluation of the predictions. The application of the Bayesian models shows that when the model miss-classifies a sample, the distribution of the predicted class probability is widespread. Therefore the introduction of an "uncertain" label can have important consequences on performances and thus on medical decision support. A further interesting medical result that the MLP models are regularly "uncertain" for a well-determined subset of patients, which suggests that the currently used input features are not enough to achieve reliable classification for these patients.

6 Conclusions

In the paper we performed a retrospective analysis in which various Bayesian models were investigated: Logistic Regression, Multi Layer Perceptron and Belief Network models. We reported standard performance measures in the Bayesian framework, investigated the usability of the more detailed predictions of Bayesian models and compared the models in the Bayesian framework. We tested the usability of a technique that derives informative prior distributions for black-box models. Finally, we introduced and demonstrated a technique based on attached Belief Networks as auxiliary probabilistic models to handle incomplete cases for probabilistic regression models.

Currently we are investigating a hybrid Belief Network model composed of a prior Belief Network and a probabilistic regression model, where we treat the prior Belief Network (including the auxiliary model) and a probabilistic regression model together in the Bayesian framework.

Acknowledgements

Joos Vandewalle is Full Professor at the K.U.Leuven. Bart De Moor is Full Professor at the K.U.Leuven. Sabine Van Huffel is a senior research associate with 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. 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, BIL2000/19; from National Fund for Scientific Research (OTKA) under contract number T030586; from National Fund for Scientific Research (OTKA) under contract number F-030763; 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; from the European Community TMR Programme, Networks, project CHRX-CT97-0160; from the Brite Euram Programme, Thematic Network BRRT-CT97-5040 'Niconet'; from the IDO/99/03 project (K.U.Leuven) "Predictive computer models for medical classification problems using patient data and expert knowledge", of the FWO grants G.0326.98 and G.0360.98, the FWO project G.0200.00.

References

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

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

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