01_59

Ovarian cancer classification with rejection by Bayesian Belief Networks

Peter Antal Geert Fannes Frank De Smet Bart De Moor

Electrical Eng. Dept. ESAT/SISTA
Katholieke Universiteit Leuven
Kasteelpark Arenberg 10, B-3001 Heverlee (Leuven), Belgium

Abstract

Belief Networks in the Bayesian approach provide a well- established methodology to fuse prior knowledge and statistical observations for an enriched decision support. In this paper we investigate one of the advantages of the Bayesian approach - the provided additional uncertainty information for predictions - in a medical classification problem. We perform a Bayesian analysis using Belief Network models to discriminate between benign and malignant ovarian masses. We report the performance of such Bayesian Belief Network models if the exclusion of some data points is allowed based on various uncertainty measures of the prediction.

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 and it provides more detailed information for decision support. These properties are particularly attractive in medical applications, since detailed uncertainty information can be vital in a medical decision and frequently abundant prior domain knowledge is available beside the statistical data. Under certain conditions Belief Networks are especially suitable for Bayesian modeling, that is to formalize the prior domain knowledge, to update it by observations and to perform inference in a Bayesian way [4]. In the paper we investigate a Belief Network model from the Bayesian perspective to discriminate between benign and malignant ovarian masses.

The paper is organized as follows: Section 2 reviews the Bayesian approach in classification problems. Section 3 recapitulates the medical problem which will serve as a test case, introduces the data and defines relevant performance measures. In Section 4 we discuss the applied Belief Network model and the algorithms used to approximate the Bayesian performance measures. Section 5 presents the performance of the model using thresholds based on various uncertainty measures of the prediction to exclude some data points. In Section 6 we summarize our findings about having a detailed Bayesianist prediction in this medical problem.

2 Bayesian Classification

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 rationale 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* $g(\boldsymbol{x}, \omega) \in \{0, 1\}$ where $\omega \in \Theta$ 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. For a *probabilistic regression* model $P(T = 1|\boldsymbol{x}, \omega) = f(\boldsymbol{x}, \omega) \in [0, 1]$ it means there is a prior distribution $p_{\Omega}(.)$ over the model parameters $\omega \in \Theta$. $F_{\Omega|\underline{\boldsymbol{d}}}$ denotes the random variable for the predicted posterior class probability (as a scalar in the $[0, 1]$ interval).

We assume the existence of a labeled training set $\underline{\boldsymbol{d}} = \{\boldsymbol{x}_k, t_k\}_{k=1}^n$, $(\boldsymbol{x}_k, t_k) \in \mathbb{R}^l \times \{0, 1\}$, where $\boldsymbol{x}$ is a real valued l-dimensional input vector and t is the corresponding class label. In the paper we use capitals for random variables, bold indicates a vector and a bold underline indicates a matrix.

01-59

Antal P., Fannes G., De Smet F., De Moor B., ``Ovarian cancer classification with rejection by Bayesian Belief Networks", in *Proc. of the Bayesian Models in Medicine workshop, the European Conference on Artificial Intelligence in Medicine (AIME'01)*, Cascais, Portugal, Jul. 2001, pp. 23-27., Lirias number: 182418.

Using the observed data $\underline{d}$ and applying Bayes' rule, the prior distribution can be transformed to the posterior distribution $p_{\Omega}(\omega|\underline{d})$ given by

$$\frac{p_T(t_1,\ldots,t_n|\omega,x_1,\ldots,x_n)p_{\Omega}(\omega|x_1,\ldots,x_n)}{p_{\underline{D}}(\underline{d})}$$

that is, by

$$L(\omega|\underline{d})p_{\Omega}(\omega)$$

where $L(\omega|\underline{d})$ denotes the probability of the data given the parameters.

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 x, in the random variable $f(x,\Omega)$ where Ω is a random parameter vector. In this way we have uncertainty information about the predicted class probability.

We can simplify this result to scalar values for the class probabilities $P(T=1|x,\underline{d})$. 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 expectation of the class probability in the posterior parameter probability space $f(x) := E_{\Omega|\underline{d}}[f(x,\omega)]$. A further simplification is to discretize this scalar value using some user specified threshold λ, deriving a binary decision function

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

These three distinct levels, the distribution of the class probabilities ($f(x,\Omega)$), the class probabilities ($f(x)$) and the class labels ($g_{\lambda}(x)$) provide diminishing possibilities for decision support, though the burdening statistical and computational complexity should be considered too.

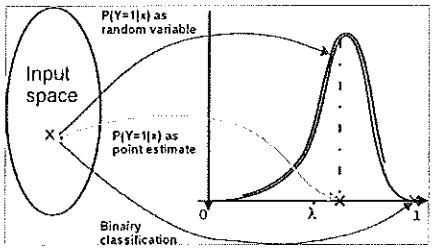

Figure 1: Three levels of predictions.

A more refined scheme allows rejection based on the uncertainty of the prediction of the class probability $F_{\Omega|\underline{d}}$

$$g_{\lambda,\sigma}(x) := \begin{cases} \text{"}rejected\text{" if } \delta[F_{\Omega|\underline{d}}] \geq \sigma \\ 1 \text{ if } E_{\Omega|\underline{d}}[f(x,\omega)] \geq \lambda \\ 0 \text{ else.} \end{cases}$$

We are using the following uncertainty measures derived by the transformations of the random variable $F_{\Omega|\underline{d}}$ into a scalar δ:

$$\begin{aligned} \delta_{L_1}[F_{\Omega|\underline{d}}] &= min(E_{F_{\Omega|\underline{d}}}[f], 1 - E_{F_{\Omega|\underline{d}}}[f]) \\ \delta_{Var}[F_{\Omega|\underline{d}}] &= Var_{F_{\Omega|\underline{d}}}[f] \\ \delta_{L_1,Var}[F_{\Omega|\underline{d}}] &= \delta_{Var}[F_{\Omega|\underline{d}}] - \alpha\delta_{L_1}[F_{\Omega|\underline{d}}] \\ \delta_{H}[F_{\Omega|\underline{d}}] &= H(F_{\Omega|\underline{d}}) \\ \delta_{Bayes}[F_{\Omega|\underline{d}}] &= min(\int_0^{1/2} dF_{\Omega|\underline{d}}, \int_{1/2}^{1} dF_{\Omega|\underline{d}}) \end{aligned}$$

3 Classification of Ovarian Masses

Ovarian malignancies represent the greatest challenge among 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: prior knowledge and data.

The available relevant medical literature and expert knowledge is abundant and very diverse (for an overview, see [5]). 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 [5]). Since the data set is mostly complete with respect to the used model in the paper we used only this subset. Univariate statistics of data set are presented in Table 1.

	Age	Parity	CA 125	CS	RI
$\widehat{E}[.\|0]$	47.77	1.50	110.34	1.98	0.12
$\widehat{E}[.\|1]$	58.62	1.57	1222.299	3.20	0.41
$\widehat{Std}[.\|0]$	15.60	1.40	976.56	0.84	0.77
$\widehat{Std}[.\|1]$	15.18	1.73	3779.64	0.95	0.46

Table 1: Univariate statistics for the benign(0) and malignant(1) subpopulation in the ovarian cancer data set.

Standard statistical studies indicate that a multimodal approach – the combination of various types of variables – is necessary for the discrimination between benign and malignant tumors. Therefore Logistic Regression models, Multilayer Perceptrons and Belief Networks were previously applied [5; 1]. 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, such as misclassification rate 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.

4 Bayesian Belief Networks

A Belief Network represents a joint probability distribution over a set of variables (see e.g. [2]). 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 (see e.g. [4; 3]). In this case the prior background knowledge is formalized as a fixed Belief Network structure and the prior distribution $p_{\Omega}(.)$ over the model parameters $\omega \in \Theta$ 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).

Using such Dirichlet distributions, an expert can express his belief in parametrizations and for complete samples the posterior distribution $p_{\Omega}(\omega|\underline{d})$ has the same analytic formula with updated hyperparameters [4; 3].

We built the Belief Network from the available prior knowledge from expert and literature in a "heterogeneous" way incorporating biological models of the underlying mechanism quantifiable by the literature, parts quantified by a medical expert and parts quantified by previously published studies [1]. The structure of the Belief Network model is shown on Fig. 2.

The target random variables to be estimated are hierarchical: the inference $P(T = 1|\Omega, x^{obs}, z, \underline{d})$ and the performance related $MR(\Omega, \underline{d})$. We sample the posterior distribution $p_{\Omega}(\omega|\underline{d})$ by direct sampling from the updated Dirichlet distributions and compute the conditional probabilities of malignancy for the drawn parametrizations by an exact inference algorithm using a join tree (see [2]). Based on these predictions the corresponding MR values can be computed.

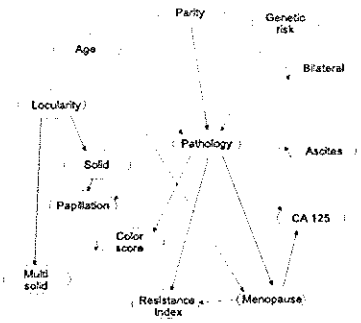

Figure 2: The BN model structures.

5 Results

We investigated the advantages of having a more detailed probabilistic prediction in the Bayesian framework. At first we manually evaluated the Bayesian predictions of the Belief Network model from a medical point of view. We noticed that the predictions for misclassified cases are more uncertain, e.g. they have higher variances $Var_{\Omega|\underline{d}}[f(x,\omega)]$ which is one measure for the 'uncertainty'. Generally spoken, the cases with a high value for $Var_{\Omega|\underline{d}}[f(x,\omega)]$ were also hard to classify by a medical professional, in contrast with cases with a low value for $Var_{\Omega|\underline{d}}[f(x,\omega)]$, that were almost always straightforward to predict.

To identify automatically these medically hard cases, we tried to quantify the uncertainty of the prediction by the δ-measures introduced in section 2. Fig. 3 and 4 show the correlation between the δ_{Var}, δ_E and δ_H measures. Correct classified samples are denoted with "*" and incorrect ones with a "o".

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 referring such a patient to an expert or further examinations. To investigate the efficiency of the identification of hard cases, we computed the misclassification rate when various proportions of the most uncertain cases are rejected. Fig. 5 shows the misclassification rates after excluding various proportions of the data set based on δ-measures (δ_{L_1}, δ_{Var} and δ_H) as defined in Section 2, Fig. 6 shows the same for the rejected partition. In these experiments, we partitioned the data set described in Section 3 randomly to a test (50%) and training (50%) set, this was repeated

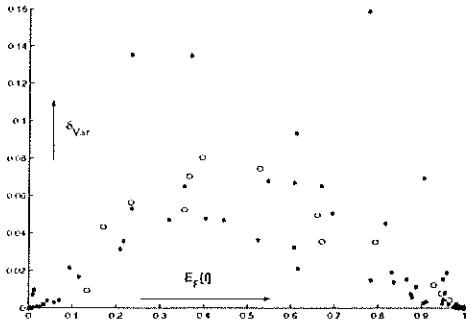

Figure 3: Correlation between $E_{F_{\Omega|\underline{d}}}[f]$ and δ_{Var}. Correct classified samples are denoted with "*" and incorrect ones with "o".

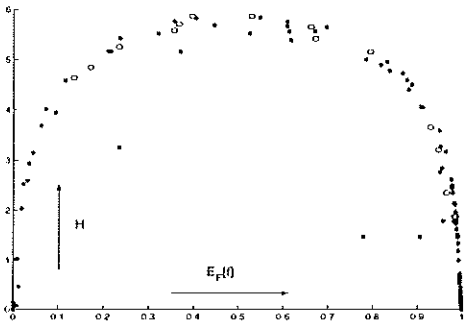

Figure 4: Correlation between $E_{F_{\Omega|\underline{d}}}[f]$ and δ_H. Correct classified samples are denoted with "*" and incorrect ones with "o".

30 times to eliminate dependency on separation. The reported results are based on the test set.

Tables 2 and 3 show the misclassification rates that are achieved for 'non-rejected' respectively 'rejected' samples for varying uncertainty measures defined by Eq. 1

6 Discussion

Since the Bayesian approach is becoming more and more popular as an efficient inductive method for integrating prior knowledge and statistical data, the question arises how we can use other potential advantages of this framework. One attractive candidate is the detailed Bayesian prediction of class probabilities, since it may allow automatic identification of the uncertain cases for special treatment. To test this idea, we experimented with representing the uncertainty of a prediction by a scalar and investigated the classification performance when cases with uncertainty above a given threshold are 'rejected' as defined in 2.

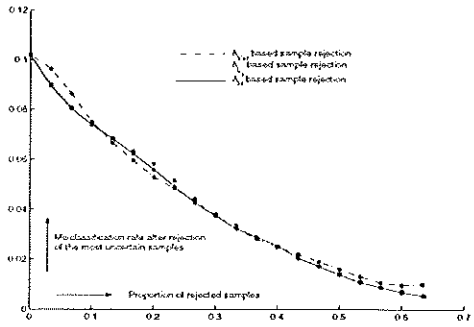

Figure 5: The misclassification rate on the test set after rejecting varying proportions.

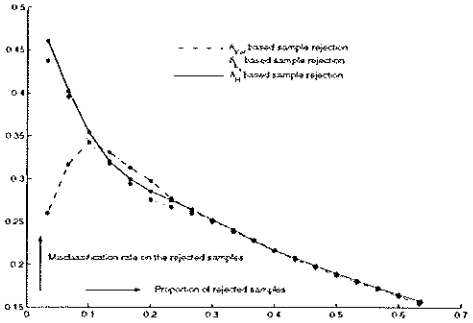

Figure 6: The misclassification rate on the rejected data points for varying proportions.

The manual evaluation by a medical expert confirmed that the derived uncertainty measures from the Bayesian prediction realistically model the subjective uncertainty of a human decision maker. To evaluate the efficiency of the rejection methods based on these measures we investigated their effect on the classification performance. As Table 2 shows without rejection the misclassification rate is 10.2% while in the rejected sets it can be above 40% for small rejected sets and in the most interesting region it is still between $20-30\%$. For example, if we set our rejection threshold to exclude 20% of the cases, the misclassification rate drops to 5%. In practice, this means that a decision support system can be specified to classify 80% of the cases with a low misclassification rate and identify the remaining 20% as hard cases that need special considerations.

As Table 2 and Fig. 6 illustrates, the effect of various rejection methods based on different δ-measures are similar and it also holds for the δ_{L1}, which is a non-Bayesian uncertainty measure. However, they have slightly different characteristics which can be interesting for various decision support strategies or problems.

% Reject.	δ_{L_1}	δ_{Var}	$\delta_{L_1,Var}$	δ_H	δ_{Bayes}
0	10.1	10.1	10.1	10.1	10.1
6.66	8.08	8.64	8.15	8.03	8.09
13.3	6.85	6.65	6.81	6.81	6.73
20	5.82	5.27	5.67	5.58	5.30
26.6	4.42	4.32	4.38	4.25	*
33.3	3.34	3.21	3.25	3.21	*
40	2.51	2.48	2.46	2.52	*
46.6	1.71	1.89	1.82	1.70	*
53.3	1.11	1.30	1.25	1.06	*
60	0.716	0.966	0.700	0.667	*

Table 2: The misclassification rate on the test set after rejecting varying proportions.

% Reject.	δ_{L_1}	δ_{Var}	$\delta_{L_1,Var}$	δ_H	δ_{Bayes}
0	*	*	*	*	*
6.66	39.5	31.7	38.5	40.2	39.2
13.3	31.8	33.0	32.0	32.0	32.4
20	27.5	29.7	28.2	28.5	29.1
26.6	25.9	26.2	26.1	26.4	*
33.3	23.8	24.1	24.0	24.0	*
40	21.6	21.7	21.7	21.6	*
46.6	19.8	19.6	19.7	19.8	*
53.3	18.1	17.9	17.9	18.1	*
60	16.4	16.3	16.4	16.5	*

Table 3: The misclassification rate on the rejected samples.

7 Conclusions

In this paper we investigated one of the advantages of the Bayesian approach - the provided additional uncertainty information for predictions - in a medical classification problem. We performed a Bayesian analysis using Belief Network models to discriminate between benign and malignant ovarian masses allowing the exclusion of some data points.

We introduced various uncertainty measures for characterizing the confidence in the prediction. We evaluated the medical applicability of these uncertainty measures in the problem. Furthermore, we demonstrated that a classifier with 'rejection' can efficiently identify the hard cases in an automated way, consequently its performance on the remaining cases improves significantly. In practice, this may result in a decisions support method where the normal cases can be more accurately classified by the system while the difficult cases are classified as 'rejected' requiring special investigations. Though the examined uncertainty measures behave similarly for this modeling method and problem, their slightly different characteristics can be utilized in various decision support strategies or problems. In general, their comparison needs further investigation.

Acknowledgements

Bart De Moor is Full Professor at the K.U.Leuven. Peter Antal is a Research Assistant with the K.U.Leuven. Geert Fannes is a Research Assistant with the F.W.O. Vlaanderen. Frank De Smet is a research assistent with the K.U.Leuven. 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 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. 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] P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, and I. Vergote, *Bayesian networks in ovarian cancer diagnosis: Potential and limitations*, Proc. of the 13th IEEE Symp. on Comp.-Based Med.Sys., 2000, Houston, pp. 103–109.

[2] E. Castillo, J. M. Gutiérrez, and A. S. Hadi, *Expert systems and probabilistic network models*, Springer, 1997.

[3] D. Heckerman et al., *Learning bayesian networks: The combination of knowledge and statistical data*, Machine Learning **20** (1995), 197–243.

[4] D. J. Spiegelhalter et al., *Bayesian analysis in expert systems*, Statistical Science **8** (1993), no. 3, 219–283.

[5] D. Timmerman et al., *Artificial neural network models for the pre-operative discrimination between malignant and benign adnexal masses.*, Ultrasound Obstet. Gynecol. **13** (1999), 17–25.