Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Classifying ovarian tumors using Bayesian Multi-Layer Perceptrons and Automatic Relevance Determination: A multi-center study

Ben Van Calster, Dirk Timmerman, Ian T. Nabney, Lil Valentin, Caroline Van Holsbeke, and Sabine Van Huffel, *Senior Member, IEEE*

***Abstract*—Ovarian masses are common and a good pre-surgical assessment of their nature is important for adequate treatment. Bayesian Multi-Layer Perceptrons (MLPs) using the evidence procedure were used to predict whether tumors are malignant or not. Automatic Relevance Determination (ARD) is used to select the most relevant of the 40+ available variables. Cross-validation is used to select an optimal combination of input set and number of hidden neurons. The data set consists of 1066 tumors collected at nine centers across Europe. Results indicate good performance of the models with AUC values of 0.93-0.94 on independent data. A comparison with a Bayesian perceptron model shows that the present problem is to a large extent linearly separable. The analyses further show that the number of hidden neurons specified in the ARD analyses for input selection may influence model performance.**

I. Introduction

OVARIAN tumors are commonly observed by gynecologists and are highly lethal compared to other tumors of the female reproductive system. The outcome of surgery depends on the pre-surgical assessment of the tumor [1]. Therefore, an accurate prediction is desirable and mathematical models can assist experts in their judgments. Several models for the prediction of malignancy of ovarian tumors have already been developed such as scoring systems, logistic regression models, multi-layer perceptrons, and support vector machines. Disadvantages of these studies are that i) data are drawn from a single centre, ii) samples are relatively small, and iii) data collection is not standardized. These issues influence model performance on new data.

Therefore, the International Ovarian Tumor Analysis (IOTA) Group conducted a multi-center study by collecting a large sample of patients with a persistent adnexal mass using a standardized protocol [2]. More than 40 variables [3] are available to construct predictive models for assessing the malignancy of an adnexal mass. The multi-centre nature of the study is thought to enhance the generalizing ability of the models across several centers and populations such that it performs well in various situations.

The present paper uses the IOTA database to build a predictive model using Bayesian multi-layer perceptrons (MLP) using the evidence framework [4]-[6]. By using a hidden layer connecting the input and output layers, MLPs are 'universal approximators' [7]: they can approximate any continuous function to an arbitrary degree. Such models can capture any non-linearity in the decision boundary.

II. Bayesian Multi-Layer Perceptrons

A. Multi-Layer Perceptrons

In a two-layer feed-forward MLP [4] addressing a binary classification problem with target t, an input vector $\mathbf{x}$ of size p is linked to the output y (which values are kept between 0 and 1 by using the logistic transfer function) through paths via the k neurons in the hidden layer. The strength of these paths is expressed by the weight vector $\mathbf{w}$. The output activation y can be seen as the posterior probability $P(t=1|\mathbf{x})$. Values for the weights can be obtained by using the cross-entropy error function when comparing the network output activations y with t. This means that maximum likelihood (ML) methodology is used to obtain the weights.

B. Bayesian Multi-Layer Perceptrons

(This part is largely based on [6].) ML solutions for parameter estimation problems such as the MLP weights do not take into account uncertainty in the estimates. Instead of looking for point estimates, Bayesian analysis incorporates uncertainty by looking for posterior probability distributions for the model parameters. The posterior distribution is obtained by confronting a prior distribution with the data D. The prior distribution reflects our prior knowledge about the model parameters. This knowledge is often vague such that usually non-informative priors are used. The posterior

Manuscript received April 3, 2006. Research supported by *Research Council KUL*: GOA-AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants, *Flemish Government*: FWO (PhD/postdoc grants, projects, G.0407.02 (support vector machines), G.0360.05 (EEG, Epileptic), G.0519.06 (Noninvasive brain oxygenation), FWO-G.0321.06 (Tensors/Spectral Analysis), research communities (ICCoS, ANMMM)); IWT (PhD Grants), *Belgian Federal Science Policy Office* IUAP P5/22 (`Dynamical Systems and Control: Computation, Identification and Modelling'), *EU*: BIOPATTERN (FP6-2002-IST 508803), ETUMOUR (FP6-2002-LIFESCIHEALTH 503094), Healthagents (IST–2004–27214).

B. Van Calster* and S. Van Huffel are with the Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Leuven, Belgium (* phone/fax: +321632 1857/1970; e-mail: bvancals@esat.kuleuven.be).

D. Timmerman and C. Van Holsbeke are with the Department of Obstetrics and Gynecology, University Hospitals Leuven, Leuven, Belgium.

I. T. Nabney is with the Neural Computing Research Group (NCRG), Aston University, Birmingham, UK.

L. Valentin is with the Department of Obstetrics and Gynecology, University Hospital Malmö, Malmö, Sweden.

distribution is computed using Bayes' rule:

$$p(\mathbf{w} \mid D, \mathcal{H}_i) = \frac{p(D \mid \mathbf{w}, \mathcal{H}_i) p(\mathbf{w} \mid \mathcal{H}_i)}{p(D \mid \mathcal{H}_i)}$$

where $\mathcal{H}_i$ represents the model type (including its assumptions), $p(\mathbf{w}|\mathcal{H}_i)$ is the prior distribution of $\mathbf{w}$, $p(D|\mathbf{w},\mathcal{H}_i)$ is the likelihood of the data as a function of $\mathbf{w}$, and $p(D|\mathcal{H}_i)$ is the normalization term representing the evidence of the particular model $\mathcal{H}_i$. Model predictions are made by integrating over the posterior distribution:

$$p(\mathbf{y} \mid \mathbf{x}^*, D) = \int p(\mathbf{y} \mid \mathbf{x}^*, \mathbf{w}) p(\mathbf{w} \mid D) \mathbf{dw} ,$$

with point predictions being the expected value of this distribution. This approach requires the computation of complex integrals that are very often not easily soluble in practice. One Bayesian approach, the *evidence* procedure [4]-[6], is to find approximations to the posterior. This method ends up with fixed parameter estimates and is therefore a Bayesian method close to ML.

In Bayesian methods, also hyperparameters are involved. In a Bayesian MLP, the prior distribution for $\mathbf{w}$ is Gaussian with zero mean and a hyperparameter α representing the inverse variance of the prior. This hyperparameter is related to the complexity of the network because a prior with larger variance can give rise to larger weights which tends to make the decision boundary more curved.

To find the most probable model parameter values $\mathbf{w}_{MP}$, $C(\mathbf{w})$ is minimized. For classification networks, this is the cross-entropy cost function augmented with

$$\alpha E_W = \frac{\alpha}{2} \sum_{i=1}^{W} w_i^2 \cdot$$

This additional error term regularizes the weight vector by penalizing larger weights to keep the model from overfitting.

The evidence approach to Bayesian modeling tries to find optimal values for the hyperparameters (α_{MP}) rather than integrating over them. Using α_{MP} the optimal weights $\mathbf{w}_{MP}$ are found by approximating the posterior by a Gaussian density around $\mathbf{w}_{MP}$. The optimal weights are found by minimizing $C(\mathbf{w})$. The evidence procedure proceeds as follows: a) initialize the weights/hyperparameters, b) train the network by minimizing $C(\mathbf{w})$ using standard optimization algorithms, c) update the hyperparameters, d) iterate steps b and c until convergence. To make a final prediction in the sense of a class probability, we need to take the network output $y = y(\mathbf{x},\mathbf{w})$ and integrate over the posterior weight distribution to take into account the uncertainty in $\mathbf{w}_{MP}$.

Advantages of a Bayesian framework for MLPs are that weight decay is done automatically, no separate validation set is needed (like in cross-validation), and it is able to apply input selection using Automatic Relevance Determination.

C. Automatic Relevance Determination (ARD)

In many applications it is not clear which variables are important for making accurate classifications. To apply the ARD framework, a separate hyperparameter α_i is assigned to each group of weights fanning out from the i^{th} input variable. During the evidence procedure, these hyperparameters are iteratively re-estimated. At the end of the training process, those hyperparameters that have reached large values are constraining weights from the corresponding input to be small. Hence such inputs have little impact on the model and can be dropped.

III. Method

A. Data and preprocessing

Data of 1066 tumors (266 malignant, 25%) were collected at nine centers from five countries: Sweden, Belgium, the UK, France, and Italy. When a patient had two masses, data from the worst mass were used. The data was randomly divided into a training set of 754 cases and a test set of the remaining 312 cases (71%-29%). This partition was stratified for outcome and center.

Because normalizing or a similar rescaling technique is preferable for neural network modeling, we rescaled the continuous variables into the [-1; +1] interval. Binary variables were coded as -1 versus +1.

B. Input selection

Since the results can vary depending on the initial values of the weights and hyperparameters, the ARD model was run 10 times. Based on the resulting α_i values, the inputs were ranked with respect to their relevance. The median of the 10 ranks was used as the main criterion to select which input(s) to drop. The ARD analysis was redone each time one or a few inputs were dropped. One disadvantage of ARD is that the endpoint of the pruning is subjective. We decided not to use a simple endpoint but to continue with the suggested input sets of size 6 to 12.

C. Bayesian neural network modeling

All analyses were performed using the NETLAB toolbox for Matlab [6].

Since weight decay parameters are automatically tuned, the only parameters to be tuned are the number of inputs (cf. supra) and the number of hidden neurons. For this, stratified 5-fold cross-validation (5CV) was used. The mean validation cross-entropy and area under the curve (AUC) were recorded as criteria to select a model. The validation AUC was considered because this index will be high when y makes a good separation between the benign and malignant cases.

When a network architecture was chosen, ten such models were fitted on the entire training set using reinitialization of the model parameters. The network with the best training

performance was chosen as final network.

This network was fed to the test data and the AUC was recorded. The results were compared to a linear benchmark model (a perceptron fitted using the evidence framework aided by ARD input selection). The AUC for different models was statistically compared using [8]. Based on a cut-off to make a hard prediction of the outcome, the test set sensitivity, specificity, PPV, and NPV values were also derived. The cut-off value was always developed on the training set where it was checked which value optimized sensitivity and specificity. Priority was given to the optimization of sensitivity. We hoped to achieve a test set sensitivity and specificity of 90% and 75%, respectively.

As a way of evaluating a model type (algorithm) rather than the specific final network (classifier), 100 new random stratified partitions of the data into a training (n = 754) and test (n = 312) set were constructed (*resampling*). The model was fitted on each training set and tested on the respective test set. In this way we obtained 100 test set performance measures of which the median gives a good view of the generalizing ability of a model type.

IV. Results

A. Automatic Relevance Determination

ARD was used to obtain input sets with 6 to 12 variables. Notice that ARD analyses use a hidden layer of user-defined size. We chose 10 hidden neurons in order to allow enough flexibility for eventual nonlinearity to emerge (further referred to as ARD10). This means that the ARD analyses will select variables that are apt for the predefined hidden layer size but not necessarily for other sizes. Therefore, the number of hidden neurons suggested by cross-validation was used to redo the ARD analyses (see next section). We investigated whether this improved model performance.

B. Cross-validation

For the input sets with 6 to 12 variables, 5CV was performed for architectures containing 2 to 15 hidden neurons. The average validation AUC results are shown in Figure 1. In general, two hidden neurons were the best choice. Additional CV analyses suggest a model using 11 inputs and 2 hidden neurons (11-2 architecture; Table 1).

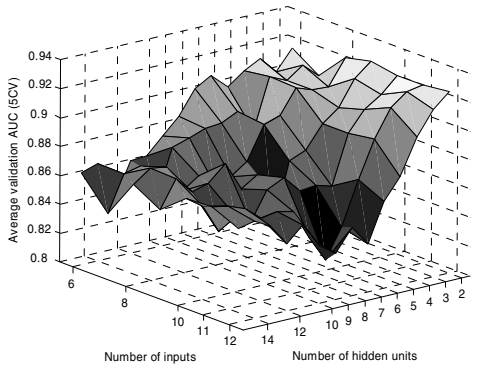
Since CV suggested the use of only two hidden neurons, the ARD analyses were redone with 2 hidden neurons in the hidden layer (ARD2). This time, ARD and subsequent CV again suggested an 11-2 architecture yet only 7 of the inputs were also used in the network based on ARD10 (Table 1).

C. The final networks

Out of the ten fitted 11-2 neural networks based on ARD10 the one with the lowest training set cross-entropy (BMLP11-2a) has training set AUC = 0.943 (SE = .008). For the model based on ARD2, the final network (BMLP11-2b) has training set AUC = 0.952 (SE = .008).

An examination of BMLP11-2a's outputs on the training set suggested to predict malignancy when $y > 0.15$. A similar examination for BMLP11-2b advocated the same cut-off.

Figure 1
Average validation AUC after applying 5CV



D. Linear benchmark model

For the Bayesian perceptron model the same strategy was followed (ARD and CV). Cross-validation suggested that >12 inputs could be selected. We chose the 11 input model, however, since that model has the same number of variables as the Bayesian MLPs that were selected and since 13 inputs was considered too much. The 11 input model (BPER11; Table 1) has training set AUC = 0.940 (SE = .010). Again 0.15 was a good cut-off for predicting malignancy.

Table 1
Overview of selected variables

Input	BMLP 11-2a	BMLP 11-2b	BPER 11
Age (years)	×	×	×
Personal history of ovarian cancer (0-1)		×	×
Hormonal therapy (0-1)	×		
Pelvic pain during examination (0-1)		×	
Ascites (0-1)	×	×	×
Max. diameter of the lesion (mm)	×	×	×
Irregular internal cyst walls (0-1)	×	×	×
Color score of intratumoral blood flow (1-4)	×		×
Blood flow within papillary projection (0-1)	×	×	×
Number of papillary projections (0-4)	×		
Acoustic shadows (0-1)		×	×
Max. diameter of solid component (mm)	×	×	×
Unilocular tumor (0-1)		×	×
Multilocular tumor + solid component (0-1)	×		
Entirely solid tumor (0-1)	×	×	×

E. Model evaluation

Original test set – Test set results of the models are given in Table 2, test set ROC curves in Figure 2. Also, Figure 3 shows the distribution of y for each outcome in the test set. Performance is very good for all models. The results indicate that the classification problem under study is fairly linear and no high level non-linear modeling seems necessary. This was already apparent from the CV results that suggested only two hidden neurons in the MLP. BMLP11-2a has a test set AUC

of 0.942 (SE = 0.018). BPER11 even has a test set AUC of 0.947 (0.016). BMLP11-2b's test set AUC is only 0.933 (SE = .017). Statistical comparison of these AUC values did not yield significant differences between these models. The sensitivity and specificity goals (90% and 75%) are met.

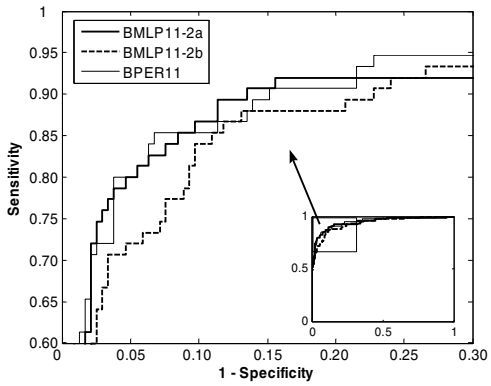
Resampling – Table 2 presents the results of the 100 test set AUC values. The median AUC for the BMLP11-2b algorithm (based on ARD2 instead of ARD10) is nearly 1% higher than that for the BMLP11-2a algorithm. Since performance is very high already, this increase can be clinically relevant. The BMLP11-2b and BPER11 algorithms perform similar with a median AUC of almost 0.94.

TABLE 2
MODEL PERFORMANCE RESULTS

Original test set		
Model	**AUC (SE)**	**Sens – Spec – PPV – NPV**
BMLP11-2a	.942 (.018)	92 – 81 – 60 – 97 %
BMLP11-2b	.933 (.017)	92 – 75 – 53 – 97 %
BPER11	.947 (.016)	95 – 76 – 55 – 98 %
Resampling		
Model type	**Median AUC (IQR)**	**Median Sens – Spec – PPV – NPV**
BMLP11-2a	.930 (.022)	91 – 80 – 58 – 96 %
BMLP11-2b	.939 (.017)	91 – 78 – 57 – 96 %
BPER11	.938 (.021)	92 – 77 – 56 – 97 %

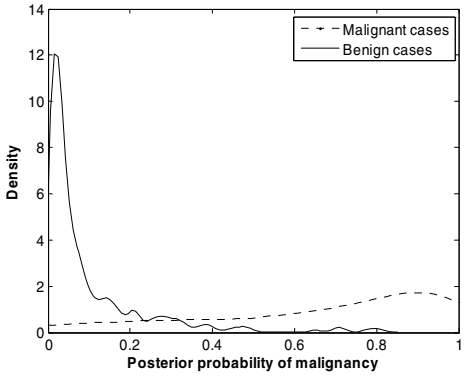
A widely used model in clinical practice is the Risk of Malignancy Index (RMI) [9], a fairly simple scoring system based on the CA125 tumor marker, menopausal status and ultrasound results. The test set AUC for RMI (AUC = 0.870, SE = 0.028) could be computed only on the test set cases with CA125 information (n = 236, 76%). On this test set, the Bayesian models' AUC is 5.9 to 7.2% higher than that of the well-known RMI ($p < 0.01$). When cut-offs are applied to make hard predictions, the RMI has much lower sensitivity. Moreover, our Bayesian models do not use CA125 measurements which are expensive and time-consuming.

FIGURE 2
TEST SET ROC CURVES



An advantage of BMLP11-2b is that it does not use the level of intratumoral blood flow that, although useful in predicting malignancy, is a subjective score given by the expert. The inter-expert variability herein can affect model performance. On the other hand, BMLP11-2b uses the presence of pelvic pain. This also is subjective.

FIGURE 3
TEST SET DISTRIBUTION OF y FOR BMLP11-2A



V. CONCLUSION

The Bayesian MLPs developed in this paper have good performance. The analyses suggest that the classification problem under study appears not highly non-linear. Interestingly, the size of the hidden layer for the ARD input selection analyses may affect model performance. Future research could focus on this finding.

REFERENCES

[1] I. Vergote, J. De Brabanter, A. Fyles, K. Bertelsen, N. Einhorn, P. Sevelda, et al., "Prognostic factors in 1545 patients with stage I invasive epithelial ovarium carcinoma : Importance of degree of differential and cyst rupture in predicting relapse," *Lancet,* vol. 357, pp. 176-182, 2001.

[2] D. Timmerman, L. Valentin, T. H. Bourne, W. P. Collins, H. Verrelst, I. Vergote, "Terms, definitions and measurements to describe the sonographic features of adnexal tumors: A consensus opinion from the International Ovarian Tumor Analysis (IOTA) group," *Ultrasound Obst Gyn*, vol. 16, pp. 500-505, 2000.

[3] D. Timmerman, A. C. Testa, T. Bourne, E. Ferrazi, L. Ameye, M. L. Konstantinovic, et al., "Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: A multicenter study be the International Ovarian Tumor Analysis Group," *J Clin Oncol*, Vol. 23, pp. 8794-8801, 2005.

[4] C. M. Bishop, *Neural networks for pattern recognition*. Oxford University Press, 1995.

[5] D. J. C. MacKay, "Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks," *Network - Comp Neural*, vol. 6, pp. 469-505, 1995.

[6] I. T. Nabney, *NETLAB. Algorithms for pattern recognition*. London: Springer, 2002.

[7] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359-366, 1989.

[8] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, pp. 837-845, 1988.

[9] I. Jacobs, D. Oram, J. Fairbanks, J. Turner, C. Frost, and J. G. Grudzinskas, "A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer," *BJOG,* vol. 97, pp. 922-929, 1990.