98-19

Verrelst H., Vandewalle J., De Moor B., Timmerman D., ``Bayesian Input Selection for Neural Network Classifiers", in *Proc. of the Third International Conference on Neural Networksand Expert Systems in Medicine and Healthcare (NNESMED'98)*, Pisa, Italy, Sep. 1998, pp. 125-132., Lirias number: 179277.

Bayesian Input Selection for Neural Network Classifiers

Herman Verrelst*, Joos Vandewalle, Bart De Moor

Department of Electrical Engineering, ESAT-SISTA,
Katholieke Universiteit Leuven,
Kard. Mercierlaan 94,
B-3000 Leuven (Heverlee)
Belgium
Tel.: +32-16-321709, Fax: +32-16-321970
E-mail: Herman.Verrelst@esat.kuleuven.ac.be

Dirk Timmerman

Dept. of Obst. and Gynaec.,
University Hospitals Leuven,
Herestraat 49,
B-3000 Leuven
Belgium

Abstract

In this paper we discuss the use of the Bayesian posterior probability distribution over weight space and Receiver Operating Characteristic curves in a neural network input selection algorithm. The posterior distribution is obtained by combining the likelihood function based on training data and a prior distribution based on expert knowledge. To numerically calculate the marginalisation, Markov Chain Monte Carlo methods are used. We demonstrate the technique on the problem of ovarian cancer classification. The resulting input selection is then used to train a neural network that significantly outperforms the Risk of Malignancy Index, a traditionally used diagnostic aid.

*Corresponding Author

Bayesian Input Selection for Neural Network Classifiers

Herman Verrelst, Joos Vandewalle, Bart De Moor
Dept. of Electr. Eng., ESAT-SISTA, Katholieke Universiteit Leuven, Kard. Mercierlaan 94, B-3000 Leuven, Belgium

Dirk Timmerman
Dept. of Obst. and Gynaec., University Hospitals Leuven, Herestraat 49, B-3000 Leuven, Belgium

In this paper we discuss the use of the Bayesian posterior probability distribution over weight space and Receiver Operating Characteristic curves in a neural network input selection algorithm. The posterior distribution is obtained by combining the likelihood function based on training data and a prior distribution based on expert knowledge. To numerically calculate the marginalisation, Markov Chain Monte Carlo methods are used. We demonstrate the technique on the problem of ovarian cancer classification. The resulting input selection is then used to train a neural network that significantly outperforms the Risk of Malignancy Index, a traditionally used diagnostic aid.

1 Introduction

When applying neural classification techniques (such as multi-layer perceptrons) in the medical field, input selection is interesting from different perspectives. It can for example be seen as a way to keep the number of free parameters in the network small by reducing the dimensionality of the input space. It can also be interesting from a medical perspective. The aim of input selection is then to find a set of features that significantly contribute to a correct classification while avoiding as much redundancy as possible. In this way the input selection tells us something about the underlying medical problem.

When using linear or logistic regression models to the classification task, a basic technique for input selection is significance based forward-backward selection. For such models the parametric optimisation problem (e.g. maximum likelihood) has a unique solution (i.e. unimodality of the cost function). To every selection of inputs, there is a corresponding unique set of parameters that can be used to calculate a performance measure. This measure can then be used to (statistically) calculate the significance of adding or deleting an input to or from the selection set.

It is well known that training algorithms for neural networks suffer from the problem of local minima. So no unique performance measure can be calculated for a certain selection of inputs, as such a measure is inherently dependent on a specific (almost always suboptimal) selection of parameters in the model.

In this paper we present how Bayesian learning techniques can average out the influence of a specific choice of parameters in the neural network model. Based on the data and on prior knowledge, using Bayes rule, a posterior distribution on the weights in the model can be found. In the next sections we will show how for example the area under the ROC-curve can be used as a performance measure that is then averaged out by sampling the posterior distribution with a Markov Chain Monte Carlo method.

In Section 2 we will briefly describe the specific use of the ROC-curves, how the Bayesian posterior is constructed and how the two are combined via marginalisation. Section 3 then reports on some numeric results when applying this techniques to the problem of ovarian cancer classification. Some conclusions will be given in Section 4.

2 Material and methods

2.1 ROC-curves

The ROC-curve is a useful and widely used device for assessing and comparing the value of tests. [2,3] We give a brief description of its definition.

Let us assume we have a dichotomic classification problem and that we use a one-hidden-layer feed-forward NN with m inputs x_k^i, n_h hidden neurons with the tanh(.) as activation function, and one output y_k,

$$y_k(\theta) = \sum_{j=1}^{n_h} w_j \tanh(\sum_{i=1}^{m} v_{ij} x_k^i + \beta_j), \tag{1}$$

parameterised by the vector θ consisting of the network's weights w_j and v_{ij} and bias terms β_j. If we train the output $y(x|\theta)$ to be equal to 1 if the input pattern belongs to the targeted class and equal to 0 otherwise, then the ultimate decision is taken by comparing the output to a certain threshold T.

The sensitivity of a classifier is defined as the proportion of malignant cases that are predicted to be malignant and the specificity as the proportion of benign cases that are predicted to be benign. The false positive rate is 1−specificity. When varying the threshold T, the sensitivity and specificity will change. A low threshold will detect almost all malignant cases at the cost of many false positives. A high threshold will give less false positives, but will also detect less malignant cases. Receiver Operating Characteristic (ROC) curves are a way to visualize this relationship. The plot gives the sensitivity versus false positive rate for varying thresholds T.

The proportion of the whole area of the graph which lies below the ROC-curve is a one-value measure of the accuracy of a test.[3] The higher this propor-

tion, the better the test. Since for every set of parameters of the neural network the area under the ROC-curve A_{ROC} is different, we note this dependence as

$$A_{\mathrm{ROC}}(\theta). \tag{2}$$

2.2 Bayesian training: likelihood and prior

As described by Bishop and Mackay[5,6], the Bayesian approach takes into account two types of information. On the one hand we have the data, a set of N samples with measurements $\mathbf{x}_k$ and target values t_k:

$$D = \{\mathbf{X}, \mathbf{T}\} = \{(\mathbf{x}_k, t_k)\}_{k=1\ldots N}.$$

And we have a way to describe how well these data fit into the model via the likelihood function. If we for example assume a Gaussian zero mean, fixed variance (σ^2) noise model on the outputs of the model, the likelihood is

$$\begin{aligned} L(\theta) &= p(\mathbf{T}|\mathbf{X},\theta) \\ &= \tfrac{1}{\sqrt{2\pi}^N}\,(\tfrac{1}{\sigma})^N\,\exp\sum_{k=1}^{N}\frac{(y_k - t_k)^2}{-2\sigma^2}. \end{aligned} \tag{3}$$

On the other hand we have prior knowledge. For example, we know that weights that are too big can be associated with over-fitting. So, we say that for a good performance the weights should be small. This could be written down in a statistical setting as a Gaussian prior with small variance (α^2) around the origin. This gives the prior on the d-dimensional weight vector

$$p(\theta|\mathbf{X}) = \frac{1}{\sqrt{2\pi}^d}\,\frac{1}{\alpha^d}\,\exp\sum_{q=1}^{d}\frac{\theta_q^2}{-2\alpha^2}. \tag{4}$$

These two components are combined to give the posterior distribution on weight space.

$$\begin{aligned} p(\theta|\mathbf{T},\mathbf{X}) &= \frac{p(\mathbf{T}|\mathbf{X},\theta)\;p(\theta|\mathbf{X})}{\int_{\mathbf{R}^d} p(\mathbf{T}|\mathbf{X},\theta)\;p(\theta|\mathbf{X})d\theta} \\ &= C^{\mathrm{st}} \times L(\theta) \times \text{prior}. \end{aligned} \tag{5}$$

This posterior can then be used in an optimisation algorithm (which would in the mentioned examples boil down to sum-squared-error optimisation with regularisation) or it can be used for what is called marginalisation over weight space. In this last approach, the posterior is seen as a weighting that emphasises those weights that optimally fit both data and prior knowledge. In this

paper we apply this marginalisation technique to calculate an average of the area under the ROC curve

$$\langle A_{ROC} \rangle_\theta = \int_{\mathbf{R}^d} A_{\mathrm{ROC}}(\theta)\ p(\theta|\mathbf{T},\mathbf{X})\ d\theta. \tag{6}$$

2.3 Markov Chain Monte Carlo

The integral in Equation 6 can not be calculated analytically, so we have to approximate it in some way. The basic method that is demonstrated here is the Markov Chain Monte Carlo method. In this method the integral is approximated by an average of the area under the ROC-curves over a finite number (M) of samples from weight space $\{\theta_i\}_{i=1\ldots M}$

$$\langle A_{ROC} \rangle_\theta \approx \frac{1}{M} \sum_{i=1}^{M} A_{ROC}(\theta_i),$$

where the θ_i are distributed according to the posterior $p(\theta|\mathbf{T},\mathbf{X})$.

This series of weight sets is obtained by defining a Markov Chain in weight space (for example through the Metropolis-Hastings algorithm that is also used in simulated annealing[4]) If the state transition probability of this Markov Chain is well defined and satisfies certain conditions (as reported by MacKay[6] and Neal[7]), then we know that the equilibrium distribution can be tailored to be just the posterior distribution. In this way we know that the tail end of the trajectory of this Markov Chain represents a sample from the posterior. Other, more efficient but inherently the same, techniques are dealt with in the work of Bisshop, MacKay and Neal.[5,6,7]

3 Results

We now demonstrate the above mentioned techniques to the real-world problem of ovarian cancer classification. Due to hardware-improvements (e.g. ultrasound equipment) over the last years, high quality measurements have become available to any well-equipped gynaecologist. Still, experience remains a key factor in the pre-operative discrimination between benign and malignant (i.e. cancerous) ovarian tumours. The goal is therefore to develop a neural network expert system based on recorded examples of correct classification. This neural network would then assist an inexperienced gynaecologist in his diagnosis.

The data were derived from a study group of 191 consecutive patients who were referred to a single institution (University Hospitals Leuven, Belgium)

from August 1994 to August 1996.[8] Table 1 lists the different inputs which were considered, with their mean value or with the relative presence in cases of benign and malignant tumours.

Table 1: Demographic, serum marker, colour Doppler imaging and morphologic indicators. For the continuous valued features the mean for each class is reported. For binary valued indicators, the last two columns give the presence of the feature in both classes e.g. only 2% of malignant tumours had smooth walls.

	N^o	**Indicator**	**Benign**	**Malignant**
Demographic				
	V1	Patient age	49.3	58.3
	V2	Postmenopausal status	40%	70.6%
Serum marker				
	V3	CA125-serum level (log)	2.8	5.2
Colour Doppler				
	V4	Resistency Index	0.64	0.53
	V5	Pulsatility Index	1.4	0.97
	V6	Peak Systolic Velocity	23	29
	V7	Average Maximum Velocity	13	18
Morphology				
	V8	Abdominal fluid	12.1%	52.9%
	V9	Bilateral mass	11.4%	35.3%
	V10	Unilocular cyst	42.1%	5.9%
	V11	Multiloc/solid cyst	16.4%	49.0%
	V12	Smooth wall	58.6%	2.0%
	V13	Irregular wall	32.1%	76.5%
	V14	Papillarities	7.9%	74.5%

To illustrate the straightforward power of the marginalisation approach, we did not incorporate the average area measures in any forward or backward selection strategy. We randomly selected 100 sets of 3 up to 5 inputs. We fixed the architecture of the neural networks to a single hidden layer perceptron with 5 hidden hyperbolic tangent-type neurons. For each of these input selections we calculated the average area under the ROC-curve. For the posterior we used a Gaussian noise model on the outputs of the neural network (fixed variance $\sigma^2 = 0.04$) and a variance $\alpha^2 = 0.01$ in a Gaussian prior on the network weights. For the Markov Chain Monte Carlo method, we took a Gaussian 'new-state-generator' with variance 0.001 and the Metropolis-Hastings acceptance

function. We allowed the first 2000 samples to be thrown away as 'burn-in' and used the subsequent 4000 to calculate the marginalisation. The average acceptance rate was only 65%.

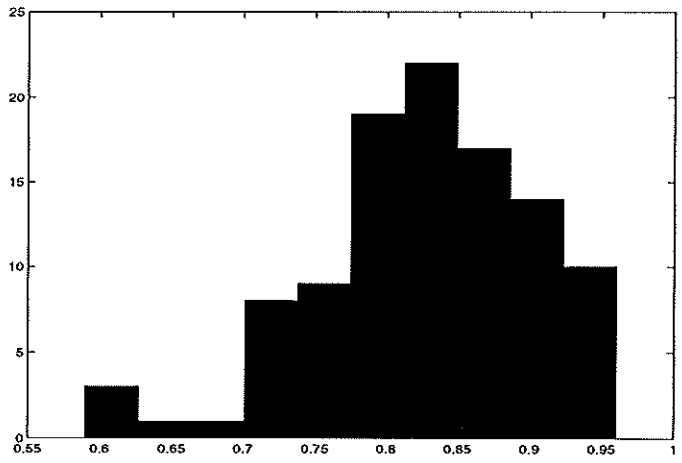


Figure 1: A histogram of Bayesian marginalised area under the ROC-curves for 100 random input selections.

Figure 1 shows the histogram of the resulting averaged areas under the ROC-curves. The inputs that were most present in the top twenty of these 100 selection, were $V14$ papillarities, $V12$ smoothness, $V3$ the CA125 marker level and $V2$ the postmenopausal status.

If we now do an optimisation on the posterior probability, using these inputs, while keeping all other settings the same, we get a neural network who's ROC curve is shown in Figure 2, together with the ROC-curve of the well-known an widely used Risk of Malignancy Index (RMI) due to Jacobs et al.[1] The 4 selected inputs are recognised by experienced gynaecologists as being indeed the most relevant ones. The quality of the input selection is also corroborated by the fact that they are also used in the traditional RMI, but, although the RMI uses the same and even more inputs than the neural network, we see that the latter's performance is much better: area $95.3 \pm 2.2\%$ versus $88.9 \pm 3.5\%$. The significance of this difference, calculated following Hanley and McNeil[2] is $p = 0.017$.

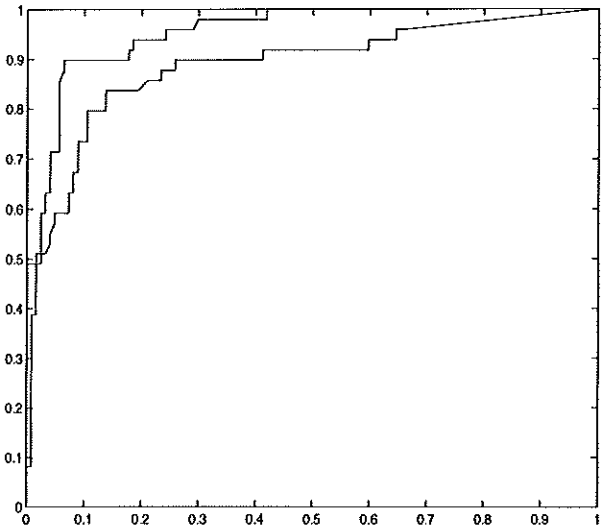


Figure 2: ROC-curves of the neural network with inputs V2, V3, V12, V14 and of the Risk of Malignancy Index. The curves, calculated for the total database of 191 patients, span respectively 95.3% and 88.9% of the total region.

4 Conclusions

In this paper we have shown how Bayesian learning techniques and ROC-curves can be combined to overcome the problem of local minima when one tries to assess the overall performance of a certain input selection. Via simple, but not necessarily the most efficient, techniques (such as Gaussian likelihood and Metropolis Hastings Markov Chain Monte Carlo methods) we have shown the principles of this approach. We have shown the real life applicability on the problem of ovarian tumour classification. The obtained results there on selection of the most relevant inputs were not only confirmed by experienced gynaecologists, they also resulted in a neural network that significantly outperformed a traditionally used method (RMI) to assist gynaecologists in their diagnosis.

Acknowledgements

This research work was carried out at the ESAT laboratory and the ICNN of the K.U.Leuven, in the following frameworks: the Programme on Interuni-

versity Poles of Attraction, initiated by the Belgian State, Prime Minister's Office-OSTC (IUAP P4-02 and IUAP P4-24), a Concerted Action Project MIPS (Modelbased Information Processing Systems) of the Flemish Community, the FWO Research Communities: ICCoS and the FWO (Fund for Scientific Research - Flanders) project G.0262.97. The scientific responsibility rests with its authors.

References

1. Jacobs I., Oram D., Fairbanks J., et al., "A risk of malignancy index incorporating CA 125, ultrasound and menopausal status for the accurate preoperative diagnosis of ovarian cancer," *Br. J. Obs. Gyn.*, Vol. 97, pp. 922-29, 1990
2. Hanley J.A., McNeil B., "A method of comparing the areas under the receiver operating characteristics curves derived from the same cases," *Radiology*, Vol. 148, pp. 839-843, 1983
3. Swets J.A., "Measuring the accuracy of diagnostic systems," *Science*, Vol. 240, pp. 1285-1293, 1988
4. Kirkpatrick S., Gelatt C.D., Vecchi M., "Optimization by simulated annealing," *Science*, Vol. 220, pp. 621-680, 1983.
5. Bishop C., *Artificial Neural Networks for Pattern Recognition*, OUP, Oxford, 1996
6. MacKay D., *Bayesian Methods for Adaptive Models*, PhD Thesis, California Institute of Technology, 1991
7. Neal R., "Probabilistic inference using Markov Chain Monte Carlo Methods," Tech. Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, 1993
8. Timmerman D., *Ultrasonography in the assessment of ovarian and tamoxifen-associated endometrial pathology*, Acta Biomedica Lovaniensia, Kath. Univ. Leuven, 1997