Chapter 11

Linear and Nonlinear Preoperative Classification of Ovarian Tumors

Chuan Lu[1], Johan A. K. Suykens[1], Dirk Timmerman[2], Ignace Vergote[2], and Sabine Van Huffel[1]

[1] Dept. of Electrical Engineering, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[2] Dept. of Obstetrics and Gynecology, University Hospitals KU Leuven, 3000 Leuven, Belgium

{chuan.lu, johan.suykens, sabine.vanHuffel}@esat.kuleuven.ac.be
{dirk.timmerman, ignace.vergote}@uz.kuleuven.ac.be

The aim of the chapter is to give an overview of a range of classical and advanced probabilistic learning algorithms for the preoperative ovarian tumor classification based on medical data. The application of the Bayesian approach for (hyper)parameter estimation and computation of the probabilistic output for the models is our focus. We evaluated and compared the performance of various linear and nonlinear models, particularly logistic regression, Bayesian multi-layer perceptrons (MLPs), Bayesian least squares support vector machines (LS-SVMs), and relevance vector machines (RVMs). The experimental results show that the Bayesian models have the potential to obtain a reliable preoperative distinction between malignant and benign ovarian tumors, and to assist the clinician in making a correct diagnosis. A guideline towards robust models for ovarian cancer diagnosis has also been proposed and discussed.

1 Introduction

Ovarian masses are a very common problem in gynecology. In particular, elderly patients in postmenopausal status have a significantly higher risk for ovarian cancer. Detection of ovarian malignancy at an early stage is very important for the survival of the patients. The 5-year survival rate for ovarian cancer when detected at a late clinical stage is only 35% (Ozols *et al.* 2000). In contrast, the 5-year survival for patients with stage I ovarian cancer is about 80% (Vergote *et al.* 2001). However, nowadays 75% of the cases are only diagnosed at an advanced stage (i.e. FIGO stage II-IV), resulting into the highest mortality rate among gynecologic cancers. The treatment and management of different types of ovarian tumors differ greatly. Conservative management or less invasive surgery suffices for patients with a benign tumor; on the other hand, those with suspected malignancy should be timely referred to a gynecologic oncologist. A reliable preoperative discrimination between malignant and benign ovarian tumors is thus critical for obtaining the most effective treatment and best advice, and will influence the outcome for the patient and the medical costs.

In order to assist the preoperative diagnosis, several types of predictive models have been developed previously. There are two major sources from which the models can be built and to extract knowledge and information. One source is the patient data that are collected clinically, the other is the domain knowledge which consists of expert knowledge, literature, and related documents. The data driven models have been developed using various linear and nonlinear techniques (Timmerman *et al.* 1999a, Timmerman *et al.* 1999b, Lu *et al.* 2001). These include logistic regression (LR), multi-layer perceptrons (MLPs), least squares support vector machines (LS-SVMs). Recently, more attention has been given to Bayesian approaches, which make induction and model selection based on probability theory and handle the uncertainty in modelling and prediction in a unified way. Bayesian MLPs and Bayesian LS-SVMs have been demonstrated to be powerful modelling tools for such medical diagnosis task (Lu *et al.* 2003a, Lu *et al.* 2003b).

Bayesian belief networks are graphical (white box) models suitable for representation of the domain knowledge. Thus, belief networks containing the biological models for ovarian tumor classification were constructed based on the knowledge both from the medical experts and the literature (Antal *et al.* 2000). It is also possible to incorporate the prior knowledge into the learning of black box models by using, for example, pseudo-samples generated from the prior belief networks (Antal *et al.* 2003). This hybrid approach is advantageous to the pure data driven approach when the data is noisy and scarce while large amount of domain knowledge is available. In our case, a moderately sized data set is already available, thus here we won't discuss the use of domain knowledge in detail, but concentrate on the data driven black-box modelling.

We give an overview of the development and evaluation of several probabilistic models, including the classical logistic regression, and other computational intelligent methods: Bayesian MLPs, Bayesian LS-SVMs and relevance vector machines (RVMs). Particularly, the application of kernel-based algorithms within the Bayesian framework is our focus for this preoperative ovarian tumor classification problem. The analysis encompasses exploratory data analysis, optimal input variable selection, parameter estimation, and performance evaluation via Receiver Operating Characteristic (ROC) curve analysis.

A series of models using advanced and classical pattern recognition algorithms have been built on 265 training data, and tested on 160 newly collected patient data. The capability of quantifying uncertainty in prediction for models has also been examined. This is essential for medical decision making. Multiple runs of randomized cross validation were also conducted in order to reduce the bias in evaluation of the model performance. Bayesian kernel based classifiers have achieved the best performance in all the experiments. The encouraging results show that the Bayesian kernel based models have the potential to obtain a reliable preoperative distinction between malignant and benign ovarian tumors, and to assist the clinician in making a correct diagnosis.

The rest of the chapter is organized as follows. We start with a review on the conventional and Bayesian modelling techniques. Then the experimental section describes the ovarian tumor data and reports the model development and assessment. Finally, a discussion of the results, conclusions, and guidelines on building robust probabilistic models for the ovarian cancer diagnosis are given.

2 Methods

In this section, we firstly introduce a generic model for regression and classification, followed by an outline of logistic regression, Multi-layer perceptrons, and least squares support vector machines. Then we give an overview of the Bayesian learning approach and how to implement Bayesian evidence methods into the powerful black-box modelling including MLPs, LS-SVMs, and how to derive Bayesian sparse solutions such as in RVMs.

2.1 Generic Modelling and Regularization

Supervised learning infers a functional relation $y \Leftrightarrow f(x)$ from a training data set composed of N independent, identically distributed (i.i.d.) samples $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}^T$. Considering a scalar-valued target function only, in the presence of additive noise, a generic regression model can be written as (MacKay 1995, Tipping 2001)

$$y_n = f(\mathbf{x}_n) + \epsilon_n, \tag{1}$$

where $x_n \in \mathbb{R}^d$, $y_n \in \mathbb{R}$, and ϵ_n is i.i.d. random variable. And ϵ_n is assumed to follow a Gaussian distribution with mean zero and variance σ^2.

Typically, the predictive functions $f(\mathbf{x})$ can be defined in the input space $\mathbf{x}$:

$$f(\mathbf{x}; \mathbf{w}) = \sum_{m=1}^{M} w_m \phi_m(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}). \tag{2}$$

where the output is a linearly weighted sum of M generally fixed basis functions for input $\mathbf{x}$, $\boldsymbol{\phi}(\mathbf{x}) = [\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), ..., \phi_M(\mathbf{x})]^T$. The weights in the model $\mathbf{w}$ are normally estimated by minimizing the regularized cost

function:

$$\mathcal{J}(\mathbf{w}) = E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}). \tag{3}$$

The error term E_D is normally a sum of squared error

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left[y_n - f(\mathbf{x}_n;\mathbf{w})\right]^2. \tag{4}$$

While the weight penalty term is usually the sum of squared weights:

$$E_W(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w}. \tag{5}$$

The regularizer λ determines the amount of regularization, i.e., a trade-off between the model fitness and the smoothness.

Let $\mathbf{y} = (y_1, \cdots, y_N)^T$, and $\mathbf{\Phi}$ the design matrix with element $\mathbf{\Phi}_{nm} = \phi_m(\mathbf{x}_n)$. Then the regularized (penalized) least-squares estimate for the weights is given by $(\mathbf{\Phi}^T\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^T\mathbf{y}$.

In this probabilistic model, the conditional distribution of the target variable given the input variable and the model parameters is again a Gaussian. Thus the likelihood of the data set can be written as (note that $\mathbf{x}$ will be omitted from the conditioning list for notational simplicity):

$$p(\mathbf{y}|\mathbf{w},\sigma^2) = \prod_{n=1}^{N} p(y_n|\mathbf{w},\sigma^2) = \prod_{n=1}^{N}(2\pi\sigma^2)^{-\frac{1}{2}}\exp\left[-\frac{(y_n - f(\mathbf{x}_n;\mathbf{w}))^2}{2\sigma^2}\right]. \tag{6}$$

One can easily verify that, the maximum likelihood estimate for $\mathbf{w}$ actually tends to minimize the negative logarithm of the likelihood $-\log p(\mathbf{y}|\mathbf{w},\sigma^2)$, which is identical to the least-squares solution.

Instead of the regularization, we can also control the complexity of the model via a prior distribution which normally expresses our preference to a smooth model. It is common to assume $p(\mathbf{w}|\alpha) \propto \exp(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w})$ with α being a hyperparameter in a probabilistic framework. As a specific example, we might choose a Gaussian distribution for $p(\mathbf{w}|\alpha)$ of the form

$$p(\mathbf{w}|\alpha) = \left(\frac{\alpha}{2\pi}\right)^{M/2}\exp\left(-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right). \tag{7}$$

In the MAP (maximum a posteriori) approach, the parameters are sought to maximize the posterior probability $p(\mathbf{w}|\mathbf{y}, \alpha, \sigma^2) \propto p(\mathbf{y}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\alpha)$, or to minimize the negative log-posterior cost function $-\log p(\mathbf{y}|\mathbf{w}, \sigma^2) - \log p(\mathbf{w}|\alpha)$ (Bishop 1995). Using (6) and (7), and retaining only the terms dependent on $\mathbf{w}$, we see that maximizing the posterior probability is equivalent to minimizing

$$\mathcal{J}_{\mathrm{MAP}}(\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - f(\mathbf{x}_n;\mathbf{w}))^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{\sigma^2}E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}).$$

Therefore the regularized least-squares approach can be viewed as a special case of the MAP technique, and the regularizer $\lambda = \sigma^2\alpha$.

For binary classification models which apply a logistic sigmoid link function $g(a) = 1/(1+e^{-a})$ to $f(\mathbf{x})$, the computation of the likelihood is based on the Bernoulli distribution:

$$p(\mathbf{y}|\mathbf{w}) = \prod_{n=1}^{N} g(f(\mathbf{x}_n;\mathbf{w}))^{y_n}[1 - g(f(\mathbf{x}_n;\mathbf{w}))]^{1-y_n}, \qquad (8)$$

where the target $y_n \in \{0, 1\}$, and no hyperparameter σ^2 is involved. The priors for the weights are kept the same as for the regression case. Thus the cost function to be minimized turns to be $-\log p(\mathbf{y}|\mathbf{w})) + \alpha E_W$. To find the maximum likelihood or MAP estimates on the basis of the logistic likelihood function, the iteratively reweighted least squares (IRLS) algorithm is often used (Neter *et al.* 1996, Tipping 2001, Bishop and Tipping 2003).

2.2 Conventional Linear and Nonlinear Models

Next we briefly explain several specific linear and nonlinear modelling techniques including logistic regression, MLPs and LS-SVMs.

2.2.1 Logistic Regression

Logistic regression is a traditional statistical tool for classification and has gained its popularity in medical data analysis. LR models can output posterior class probabilities and the model parameters are easy to interpret.

For a binary classification problem, it tries to model the conditional probability of the class given its observation:

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}^T\mathbf{x} + b))} \tag{9}$$

where $\mathbf{w}$ and b are the weight vector and intercept of the decision hyperplane respectively, and the data have targets $y_n \in \{0, 1\}$. This model defines a linear relation between the logit of the mean for the response variable and the input variables, which is equivalent to $\text{logit}(\mathcal{P}) = \log(\frac{\mathcal{P}}{1-\mathcal{P}}) = \mathbf{w}^T\mathbf{x} + b$, with $\mathcal{P} = p(y = 1|\mathbf{x})$ denoting the probability for the positive class. In this case, it refers to the probability for a tumor being malignant given the input vector $\mathbf{x}$. Prior class probabilities can be incorporated by adjusting the bias term b. An iteratively re-weighted least squares (IRLS) algorithm such as the Newton-Raphson method can be employed to perform maximum likelihood estimation for LR model fitting (Neter *et al.* 1996).

2.2.2 Multi-layer Perceptrons

A multi-layer perceptron (MLP) creates a nonlinear input-output mapping defined by the layers of summation and elementary nonlinear mappings (Bishop 1995). We concentrate here on one hidden layer MLPs with the hyperpolic tangent ($\tanh$) activation function for the hidden neurons, and with a logistic sigmoidal function as the output activation function. The other typical activation functions such as a threshold function can be chosen as well. The MLP model of our concern can be expressed as:

$$f(\mathbf{x}, \mathbf{w}) = g\left(\sum_{j=0}^{M} w_j^{(2)} \tanh\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i\right)\right), \tag{10}$$

where $\mathbf{w}$ include the weights and the bias parameter, and indices i and j correspond to input and hidden units, respectively. The logistic function is chosen as the output activation function here, since its output value ranges from 0 to 1, and can be interpreted as probability (Bishop 1995).

The training of the feed-forward neural network is often done by an iterative backpropagation procedure (often in relation to advanced local optimization algorithms), until the discrepancy between the target output and actual response is minimized. Traditionally the complexity of the model can be

controlled by weight decay, where the regularized cost function should be minimized.

The well known local minima problem for MLP learning can be tackled by retraining using different reinitalization of the weights. In order to avoid overfitting, one also needs to select an appropriate network architecture, e.g. the number of hidden neurons, types of activation functions, and the regularization hyperparameters.

2.2.3 Least Squares SVMs for Classification

Support Vector Machine (SVM) (Vapnik 1995, Cristianini and Shawe-Taylor 2000) is now a state-of-the-art technique for pattern recognition. LS-SVM is a least squares version of SVM (Suykens and Vandewalle 1999, Suykens *et al.* 2002, Suykens *et al.* 2003). Unlike MLPs, SVMs have attractive features such as good generalization performance, the existence of a unique solution, and strong theoretical background, i.e., statistical learning theory (Vapnik 1995), supporting their good empirical results. The basic idea of the nonlinear SVM classifier is: map a d-dimensional input vector $\mathbf{x} \in \mathbb{R}^d$ into a high (potentially infinite) d_{f}-dimensional feature space by the mapping $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^{d_{\mathrm{f}}} : \mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$. A linear classifier is then constructed in this feature space, while $\boldsymbol{\varphi}(\cdot)$ is implicitly defined by use of kernels. In LS-SVMs, the training is expressed in terms of solving a set of linear equations in the dual space instead of quadratic programming as for the standard SVM case. Moreover, LS-SVM is closely related to Gaussian processes and kernel Fisher discriminant analysis (Van Gestel *et al.* 2002).

The LS-SVM classifier $y(\mathbf{x}) = \mathrm{sign}\left[\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + b\right]$ (in the primal space) is inferred from the data with binary targets $y_i \in \{-1, +1\}$ (-1: benign, $+1$: malignant) by minimizing the cost function:

$$\mathcal{J}(\mathbf{e}, \mathbf{w}, b) = E_D + \lambda E_W = \tfrac{1}{2}\sum_{i=1}^{N} e_i^2 + \tfrac{\lambda}{2}\mathbf{w}^T\mathbf{w}, \qquad (11)$$

subject to the equality constraints $y_i[\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i) + b] = 1 - e_i, i = 1, ..., N$. This optimization problem can be transformed and solved through a linear

system in the dual space (Suykens and Vandewalle 1999, Suykens *et al.* 2002):

$$\left[\begin{array}{c|c} 0 & \mathbf{y}^T \\ \hline \mathbf{y} & \boldsymbol{\Omega} + \lambda^{-1}\mathbf{I} \end{array}\right] \left[\begin{array}{c} b \\ \hline \mathbf{a} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{1} \end{array}\right] \tag{12}$$

with $\mathbf{e} = [e_1, \cdots, e_N]^T$, $\mathbf{a} = [a_1, \cdots, a_N]^T$, and $\mathbf{1} = [1 \cdots 1]^T$. Mercer's theorem is applied to the matrix $\boldsymbol{\Omega}$ with $\boldsymbol{\Omega}_{ij} = y_i y_j \boldsymbol{\varphi}(\mathbf{x}_i)^T \boldsymbol{\varphi}(\mathbf{x}_j) = y_i y_j \, K(\mathbf{x}_i, \mathbf{x}_j)$, where $K(\cdot, \cdot)$ is a chosen positive definite kernel that satisfies Mercer's condition (Mercer 1909). The most common kernels include linear kernels and radial basis function (RBF) kernels, as defined by:

$$\begin{array}{rl} \text{linear kernel} & K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j \\ \text{RBF kernel} & K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / r^2). \end{array}$$

The LS-SVM classifier is then constructed in the dual space as:

$$y(x) = \operatorname{sign}\left[\sum_{i=1}^{N} a_i \, y_i \, K(\mathbf{x}, \mathbf{x}_i) + b\right]. \tag{13}$$

In order to achieve a high performance of LS-SVM models, one still needs to tune the regularization and possible kernel parameters.

In brief, the above conventional modelling techniques often apply error minimization based maximum likelihood or MAP approaches for supervised learning. In order to avoid overfitting, one needs also to control the model complexity, thus to choose the appropriate model structure and hyperparameters such as regularization parameters and possibly the kernel parameters. This is usually done via cross validation (CV). However, we should be aware that CV gives noisy estimates for the model performance, and it easily becomes computationally prohibitive with the increase of the number of (hyper)parameters to be determined.

2.3 Bayesian Learning

Recently, Bayesian methods have become a viable alternative to maximum likelihood or MAP approaches (Bishop 1995, MacKay 1995, Neal

1996, Suykens *et al.* 2002, Tipping 2001). The main advantages of Bayesian methods are:

- Automatic control of the model complexity
- Possibility to use prior information and hierachical models for the hyperparameters
- Predictive distributions for outputs, from which the uncertainty of the prediction can be judged.

The key principle of the Bayesian approach is to construct posterior probability distributions for the parameters and hyperparameters in the models. Predictions can be made based on the posterior (marginal) distributions over all the parameters.

Suppose we have defined a model linear in the parameters by (1) and (2). Given the likelihood and the prior distribution, we can compute the posterior distribution via Bayes' rule:

$$p(\mathbf{w}|\mathbf{y}, \alpha, \sigma^2) = \frac{\text{likelihood} \times \text{prior}}{\text{normalizing factor}} = \frac{p(\mathbf{y}|\mathbf{w}, \sigma^2)p(\mathbf{w}|\alpha)}{p(\mathbf{y}|\alpha, \sigma^2)}.$$

Here, the posterior follows also the Gaussian distribution: $p(\mathbf{w}|\mathbf{y}, \alpha, \sigma^2) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \sigma^2\alpha\mathbf{I})^{-1}\boldsymbol{\Phi}^T\mathbf{y} \quad (14)$$

$$\boldsymbol{\Sigma} = \sigma^2(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \sigma^2\alpha\mathbf{I})^{-1}. \quad (15)$$

The mean of the posterior distribution for the weights $\boldsymbol{\mu}$ is indeed obtained via the MAP method, which results into the same solution as the regularized least squares one with $\lambda = \sigma^2\alpha$. However, the key difference is that in the Bayesian approach the prediction is made by integrating over the distribution of the model parameters $\mathbf{w}$, rather than based on a specific estimate for $\mathbf{w}$. This is the same as taking the average prediction of all the models weighted by their posterior probability, which typically results in improved predictive capability. Specifically, to predict the new output y_* given the new input $\mathbf{x}_*$, a full Bayesian predictive distribution is obtained

by integrating (marginalizing) the predictions of the model with respect to the posterior distribution of the model parameters

$$p(y_*|\mathbf{x}_*, \mathbf{y}) = \int p(y_*|\mathbf{x}_*, \mathbf{w}, \sigma^2) p(\mathbf{w}, \alpha, \sigma^2|\mathbf{y}) d\mathbf{w} d\alpha d\sigma^2. \qquad (16)$$

However, the integrations for getting $p(\mathbf{w}, \alpha, \sigma^2|\mathbf{y})$ or $p(y_*|\mathbf{x}_*, \mathbf{y})$ are nearly always analytically intractable. This necessitates the use of approximation methods for marginalization, such as type II maximum likelihood, Laplace (Gaussian) approximation, variational methods, and Markov Chain Monte Carlo (MCMC) sampling techniques.

In our work, type II maximum likelihood approach, i.e. the evidence framework is used. By replacing $p(\alpha, \sigma^2|\mathbf{y})$ with a δ-function at its mode, using $p(\mathbf{w}, \alpha, \sigma^2|\mathbf{y}) \equiv p(\mathbf{w}|\mathbf{y}, \alpha, \sigma^2) p(\alpha, \sigma^2|\mathbf{y})$ and (16), we see that the predictive distribution can be approximated by a distribution at α_{MP} and σ^2_{MP}, which are the "most probable" values (maximizing the posterior) for the hyperparameters α and σ^2, respectively. Note that the δ-function is not necessarily an accurate approximation of $p(\alpha, \sigma^2|\mathbf{y})$. But this leads to the approximation of the integration over the hyperparameters by point estimates at their most probable posterior values. Under the assumption of flat, noninformative priors over $\log \alpha$ and $\log \sigma$, we have $p(\alpha, \sigma^2|\mathbf{y}) \propto p(\mathbf{y}|\alpha, \sigma^2)$, then α_{MP} and σ^2_{MP} can be found by maximizing the marginal likelihood (*evidence*)

$$\begin{aligned} p(\mathbf{y}|\alpha, \sigma^2) &= \int p(\mathbf{y}|\mathbf{w}, \sigma^2) p(\mathbf{w}|\alpha) d\mathbf{w} \\ &= (2\pi)^{-\frac{N}{2}} \ |\, \sigma^2 \mathbf{I} + \alpha^{-1} \mathbf{\Phi} \mathbf{\Phi}^T |^{-\frac{1}{2}} \exp\!\left[-\frac{1}{2} \mathbf{y}^T (\sigma^2 \mathbf{I} + \alpha^{-1} \mathbf{\Phi} \mathbf{\Phi}^T)^{-1} \mathbf{y} \right]. \end{aligned}$$

This is indeed a Gaussian distribution over the N-dimensional data vector $\mathbf{y}$. To determine α_{MP} and σ^2_{MP}, some gradient based optimization methods are often applied. Some relations in the optimality condition can be exploited. As shown in (MacKay 1992), the variance of the distribution for w_m can be estimated by $1/\alpha_{\mathrm{MP}} = \|\mathbf{w}_{\mathrm{MP}}\|^2/\gamma$, where the most probable value $\mathbf{w}_{\mathrm{MP}}$ is the MAP estimate for $\mathbf{w}$ which is also referred to $\boldsymbol{\mu}$ in (14), and γ is the number of effective parameters: $\gamma = M - \alpha \sum_m \mathbf{\Sigma}_{mm}$. Here $\mathbf{\Sigma}$ is the variance-covariance matrix on $\mathbf{w}$ estimated by (15). The quantity

of γ is between 0 and the total number of parameters M. Similarly, in a regression problem with a Gaussian noise model, the variance of the target variable can be estimated by $\sigma^2_{\mathrm{MP}} = 2E_D/(N-\gamma)$.

Having found α_{MP} and σ^2_{MP}, our approximation to the predictive distribution is then a Gaussian:

$$p(y_*|x_*, \mathbf{y}) = \int p(y_*|x_*, \mathbf{w}, \sigma^2_{\mathrm{MP}}) p(\mathbf{w}|\mathbf{y}, \alpha_{\mathrm{MP}}, \sigma^2_{\mathrm{MP}}) d\mathbf{w} = \mathcal{N}(\mu_*, \sigma^2_*)$$

with $\mu_* = f(\mathbf{x}; \boldsymbol{\mu})$, and $\sigma^2_* = \sigma^2_{\mathrm{MP}} + \boldsymbol{\psi}^T \boldsymbol{\Sigma} \boldsymbol{\psi}$, where $\boldsymbol{\psi} = [\phi_1(\mathbf{x}_*), \ldots, \phi_M(\mathbf{x}_*)]$. Intuitively, the mean predictor μ_* is the model evaluated at the posterior mean weights, and the predictive variance σ^2_* is the sum of the variances associated with both the noise process and the uncertainty in the weight estimates.

2.4 Bayesian Model Selection

Suppose that we have a set of hypothesis models $\{\mathcal{H}_k\}_{k=1}^K$ that compete to fit the given data. The model selection here involves mainly the choice of basis functions (e.g. the bandwidth r_k for RBF kernels), which is represented by $\mathcal{H}_k$ with the prior probability of $p(\mathcal{H}_k)$. To be more general, the model hypothesis might describe the selection of input variables, the structure for a neural network, the kernel type and parameters such as the bandwidth of an RBF kernel for SVM or RVM models, etc.. Different models can be compared by examining their posterior $p(\mathcal{H}_k|\mathbf{y}) = p(\mathbf{y}|\mathcal{H}_k)p(\mathcal{H}_k)/p(\mathbf{y})$. Assuming a uniform prior $p(\mathcal{H}_k)$ over all models, the models can be ranked by the marginal likelihood, or model evidence $p(\mathbf{y}|\mathcal{H}_k)$, which can be evaluated using a Gaussian approximation. In this study, the model hypothesis that maximizes the model evidence is chosen from a set of candidates.

Statistical interpretation is also available for the comparison between two models in the Bayesian framework. Bayes factor B_{10} for model $\mathcal{H}_1$ against $\mathcal{H}_0$ from data is defined as $B_{10} = p(\mathbf{y}|\mathcal{H}_1)/p(\mathbf{y}|\mathcal{H}_0)$. Under the assumption of equal model priors, the Bayes factor can be seen as a measure of the

evidence given by the data in favor of a model compared to a competing one. When the Bayes factor is greater than 1, the data favor $\mathcal{H}_1$ over $\mathcal{H}_0$; otherwise, the reverse is true. The rules of thumb for interpreting $2\log B_{10}$ include: the evidence for $\mathcal{H}_1$ is very weak if $0 \leq 2\log B_{10} \leq 2.2$, and the evidence for $\mathcal{H}_1$ is decisive if $2\log B_{10} > 10$, etc., as shown in (Jeffreys 1961).

2.5 Implementation of Bayesian Learning in Black-box Modelling

In this section, we discuss some implementation issues in integrating Bayesian learning with various black-box modelling techniques. This is a review of the work from several researchers. MacKay's evidence framework has provided a unified theoretical treatment of learning in feedforward neural networks (MacKay 1992, Bishop 1995). This evidence framework has also been applied to LS-SVM by Van Gestel and Suykens, achieving similar test set results as Gaussian processes and SVMs in several benchmark problems (Van Gestel *et al.* 2002, Suykens *et al.* 2002). Relevance vector machine, a special case of sparse Bayesian modelling with kernel basis, has been introduced by Tipping recently and appears to yield very sparse solutions and obtain good empirical performance as well (Tipping 2001, Tipping and Faul 2003, Bishop and Tipping 2003).

2.5.1 Bayesian MLPs

In the evidence framework for MLP learning, the posterior distribution for parameters can only be approximated locally by a Gaussian. This is on contrast with the linear parametric model in Section 2.3, where the Gaussian distribution is exact for the posterior of parameters. The approximation in Bayesian MLP modelling takes a Gaussian distribution with the mean being the "most probable" value $\mathbf{w}_{\mathrm{MP}}$, and a covariance matrix: $\mathbf{\Sigma} = H^{-1}$:

$$p(\mathbf{w}|\mathbf{y}, \alpha, \sigma^2) \cong p(\mathbf{w}_{\mathrm{MP}}|\mathbf{y}, \alpha, \sigma^2) \exp\left[-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\mathrm{MP}})^T H (\mathbf{w}-\mathbf{w}_{\mathrm{MP}})\right],$$

where $\mathbf{w}_{\mathrm{MP}}$ is normally obtained by a gradient based optimization approach, and the Hessian H is evaluated at $\mathbf{w}_{\mathrm{MP}}$, $H = -\nabla\nabla \log p(\mathbf{w}|\mathbf{y}, \alpha, \sigma^2)|_{\mathbf{w}_{\mathrm{MP}}}$.

Therefore the predictive distribution becomes $p(\mathbf{y}_*|\mathbf{y}) \approx N(\mu_*, \sigma_*^2)$ with mean $\mu_* = f(\mathbf{x}_*; \mathbf{w}_{\mathrm{MP}})$, and variance $\sigma_*^2 = \sigma_{\mathrm{MP}}^2 + \boldsymbol{\psi}^T H^{-1} \boldsymbol{\psi}$ with $\boldsymbol{\psi} = \frac{\partial f}{\partial \mathbf{w}}|_{\mathbf{x}_*, \mathbf{w}_{\mathrm{MP}}}$.

For classification problems, the *moderated* output can be calculated for MLP with logistic sigmoidal output activation function. The moderated output probability can account for the uncertainty in weights (by using a simple numerical approximation), which is in contrast with the output probability using directly the most probable network output (MacKay 1992).

Moreover, it is a good practice to assign different regularization hyperparameters to the weights and the bias of different layers. For a two layer MLP, for instance, we could use four different values of hyperparameters $\alpha_i, i = 1, \ldots, 4$ for each layer of weights and bias.

2.5.2 Bayesian LS-SVMs

In (Van Gestel *et al.* 2001, Van Gestel *et al.* 2002) the application of the evidence framework to LS-SVMs originated from the feature space formulation, whereas analytic expressions are obtained in the dual space on the three levels of Bayesian inferences. For more computational details, the interested readers are referred to (Suykens *et al.* 2002, Van Gestel *et al.* 2002).

The Bayesian evidence approach first finds implicitly the MAP estimates of model parameters $\mathbf{w}_{\mathrm{MP}}$ and b_{MP}, using conventional LS-SVM training methods, i.e. by solving the linear set of equations in (12) in the dual space.

The inference in Bayesian LS-SVM learning follows the linear regression framework in Section 2.3, and assumes a Gaussian noise model over the target variable. The class probabilities for the classification can be produced in the following way. For a given test case, the conditional class

probabilities $p(\mathbf{x}_*|y_* = \pm 1, \mathbf{y}, \alpha, \sigma^2)$ can be computed using the two Gaussian probability densities for two classes at the most probable value $\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_*)|_{\mathbf{w}_{\mathrm{MP},\pm}} \sim \mathcal{N}(\mu_\pm, \sigma^2_{*,\pm})$. The mean of each distribution $\mu_\pm$ is defined as the class center of the latent output $\mathbf{w}^T_{\mathrm{MP}}\boldsymbol{\varphi}(\mathbf{x})$ in the training set. The variance $\sigma^2_{*,\pm}$ comes from both the target noise and the uncertainty in the parameter $\mathbf{w}$, and can be calculated analogically using the method described in Section 2.3.

By applying Bayes' rule the posterior class probabilities of the LS-SVM classifier can be obtained:

$$p(y_*|\mathbf{x}_*, D, \alpha, \sigma^2) = \frac{p(y)p(\mathbf{x}_*|y_*, \mathbf{y}, \alpha, \sigma^2)}{\sum_{y'=\pm 1} p(y')p(\mathbf{x}_*|y', \mathbf{y}, \alpha, \sigma^2)}, \tag{17}$$

where $p(y)$ corresponds to the prior class probability. The posterior probability could also be used to make minimum risk decisions in case of different error costs. Let c^+_- and c^-_+ denote the cost of misclassifying a case from class '$-$' and '$+$' respectively. One obtains the minimum risk decision rule by formally replacing the prior $p(y)$ in (17) with the adjusted class prior, e.g. $P'(y=1) = P(y=1)c^-_+/(P(y=1)c^-_+ + P(y=-1)c^+_-)$.

2.5.3 Relevance Vector Machines

Sparse Bayesian learning is the application of Bayesian automatic relevance determination (ARD) to models linear in their parameters in the form of (2), by which the sparse solutions to the regression or classification tasks can be obtained (Tipping 2001, Bishop and Tipping 2003). Although the sparse Bayesian modelling framework can be applied to general models with arbitrary sets of basis functions, we consider here only one special case, the relevance vector machine (RVM), where the basis functions adopt the kernel representation $\phi_m = K(\mathbf{x}, \mathbf{x}_m)$, the $N \times (N+1)$ design matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1), \boldsymbol{\phi}(\mathbf{x}_2), \ldots, \boldsymbol{\phi}(\mathbf{x}_N)]^T$, with $\boldsymbol{\phi}(\mathbf{x}_n) = [1, K(\mathbf{x}_1, \mathbf{x}_n), \ldots, K(\mathbf{x}_N, \mathbf{x}_n)]^T$. For RVMs, kernels can be any symmetric kernel functions, though kernels in LS-SVMs must be positive definite and satisfy Mercer's condition. This is because RVM is a parameterized kernel model while LS-SVMs have primal-dual representations.

In Section 2.3, we have depicted the Bayesian inference where the hyperparameter α is set to be the same value for all the parameters. In order to encode the preference of sparsity in parameter estimation, one can retain the traditional Gaussian prior for weights, but use M independent hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ to control the prior distribution for $M = N + 1$ weights. The Gaussian prior is modified to be:

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{m=1}^{M} \mathcal{N}(w_m|0, \alpha_m^{-1}), \tag{18}$$

and one can place the previously utilized log-uniform hyperpriors over $\boldsymbol{\alpha}$, which is given by a special case of the Gamma distribution. It is illustrated in (Tipping 2001) and (Bishop and Tipping 2003) that the marginal prior $p(w_m) = \int p(w_m|\alpha_m)p(\alpha_m)d\alpha_m \propto 1/|w_m|$. This is equivalent to the use of a penalty function: $E_W = \sum_m \log |w_m|$, which encourages sparsity. Note that for standard SVMs, sparsity is obtained by Vapnik's ϵ-insensitive loss function.

Having defined the hierarchical prior, Bayesian inference then proceeds as previously. In the regression version, the likelihood is still a Gaussian. The posterior distribution over $\mathbf{w}$ is also Gaussian: $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \sigma^2) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma} = (\sigma^{\mathbf{2}}\boldsymbol{\Phi}^{\mathbf{T}}\boldsymbol{\Phi} + \mathbf{A})^{-\mathbf{1}}$, $\boldsymbol{\mu} = \sigma^{-2}\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\mathbf{y}$, $\mathbf{A} = \mathrm{diag}(\alpha_1, \ldots, \alpha_M)$. The hyperparameters are sought by maximizing the marginal likelihood

$$p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2) = (2\pi)^{-N/2}|\mathbf{C}|^{-1/2} \exp(-\frac{1}{2}\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}),$$

with $\mathbf{C} = \sigma^2\mathbf{I} + \boldsymbol{\Phi}\mathbf{A}^{-1}\boldsymbol{\Phi}^T$. By differentiating $\log p(\mathbf{y}|\boldsymbol{\alpha}, \sigma^2)$, one can get the re-estimation formulas: $\alpha_m^{\mathrm{new}} = \gamma_m/\mu_m^2$, and $(\sigma^2)^{\mathrm{new}} = \|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\mu}\|^2/(N - \sum_{m=1}^{M}\gamma_m)$, where $\gamma_m = 1 - \boldsymbol{\Sigma}_{mm}$ is a measure of effectiveness of parameter w_m.

The optimization of hyperparameters thus can be an iterative procedure using the re-estimation formulas. A more efficient approach which sequentially updates the hyperparameters has also been investigated in (Tipping and Faul 2003) and (Bishop and Tipping 2003). Practically, it has been observed that the optimum value of many of the α_m will be infinite, implying that the corresponding weight estimate will be zero.

In sparse Bayesian classification, the logistic sigmoid link function is utilized and the likelihood is given by the Bernoulli distribution as in (8). The Gaussian approximation is used for computing the weight posterior $p(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha})$ or the marginal likelihood $p(\mathbf{y}|\boldsymbol{\alpha})$. In an iteratively reweighted least squares algorithm, for the current value of $\boldsymbol{\alpha}$, the most probable weights $\widehat{\boldsymbol{\mu}}$ are given by:

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}} &= (\boldsymbol{\Phi}^T \mathbf{B} \boldsymbol{\Phi} + \mathbf{A})^{-1}, \\ \widehat{\boldsymbol{\mu}} &= \widehat{\boldsymbol{\Sigma}} \boldsymbol{\Phi}^T \mathbf{B} \widehat{\mathbf{y}}, \\ \text{and } \widehat{\mathbf{y}} &= \boldsymbol{\Phi} \widehat{\boldsymbol{\mu}} + \mathbf{B}^{-1}(\mathbf{y} - g(\boldsymbol{\Phi} \widehat{\boldsymbol{\mu}})), \end{aligned}$$

where $\mathbf{B} = \mathrm{diag}(\beta_1, \ldots, \beta_n)$, with $\beta_n = g(f(\mathbf{x}_n; \widehat{\boldsymbol{\mu}}))[1 - g(f(\mathbf{x_n}; \widehat{\boldsymbol{\mu}}))]$.

2.6 Input Variable Selection

In medical classification problems, variable selection can have an impact on the economics of data acquisition and the accuracy and complexity of the classifiers, and is helpful in understanding the underlying mechanism that generated the data.

In traditional statistics, information criteria based chi-square tests are often used as measures for model selection. Variable selection can be performed using these criteria and via a heuristic search such as forward, backward and stepwise selection. In forward selection, the variable with the largest score chi-square statistic from the remaining variables is entered into the model if it meets the specified significance level for entry. Backward elimination progressively removes the least significant variable from the model, until all the variables in the model satisfy the critical staying significance level. In stepwise selection variables are entered into and removed from the model in such a way that each forward selection step may be followed by one or more backward elimination steps. The stepwise selection process terminates if no further variable can be added to the model, or if the variable just entered into the model is the only variable removed in the subsequent backward elimination. Stepwise logistic regression has been utilized in this work.

In a Bayesian framework, the models can be ranked by evaluating the pos-

terior probability for the model $p(\mathbf{y}|\mathcal{H}_k)$. By combining the model ranking with the heuristic search method, we can find the optimal subset of variables as well. To start the forward selection (greedy search) based on the evidence, we should define the model type a priori by e.g. giving the kernel type. The procedure starts from zero variables, and chooses each time the variable which gives the greatest increase in the current model evidence. The selection is stopped when the addition of any remaining variables can no longer increase the model evidence.

Rather than using a heuristic search, by use of Bayesian MLP with ARD, one can also obtain the relevance for the sets of candidate variables. However, we met some numerical problems when working on the ovarian tumor data. Moreover, for Bayesian LS-SVMs, given the hyperparameters, the "most probable" estimate of the parameters can be found uniquely by simply solving some linear equations. For RVMs, an iterative algorithm is needed, and some costly matrix inversions are also involved. Thus Bayesian LS-SVM appears faster and more stable than the RVM. So the variable selection based on the Bayesian LS-SVM has been emphasized in this study.

3 Experiments

3.1 Data

The data set includes the information of 525 patients who were referred to a single ultrasonologist at University Hospitals Leuven, Belgium, between 1994 and 1999. These patients have a persistent extrauterine pelvic mass, which was subsequently surgically removed. The study is designed mainly for preoperative differentiation between malignant and benign adnexal masses (Timmerman *et al.* 1999a, Timmerman *et al.* 1999b, Timmerman *et al.* 2000). Patients without preoperative results of serum CA 125 levels are excluded from this analysis. The gold standard for discrimination of the tumors was the result of histological examination. Among the available 425 cases, 291 patients (68.5%) had benign tumors, whereas 134 ones (31.5%) had malignant tumors.

The medical features on demography and some clinical examinations were acquired before operation, including: age and menopausal status of the patients, serum CA 125 levels from the blood test, the ultrasonographic morphologic findings about the mass, color Doppler imaging and blood flow indexing, etc.. The data set contains 27 variables after preprocessing (e.g. color score was transformed into three dummy variables, CA 125 serum level was rescaled by taking its logarithm). Table 1 lists the important variables that were considered.

Table 1. Descriptive statistics for variables of demography, serum marker, color Doppler imaging and morphology.

	Variable (Symbol)	Benign	Malignant
Demographic	Age (Age)	45.6 ± 15.2	56.9 ± 14.6
	Postmenopausal (Meno)	31.0 %	66.0 %
Serum marker	CA 125 (log)(L_CA125)	3.0 ± 1.2	5.2 ± 1.5
CDI	Weak blood flow (Colsc2)	41.2 %	14.2 %
	Normal blood flow (Colsc3)	15.8 %	35.8 %
	Strong blood flow (Colsc4)	4.5 %	20.3 %
	Pulsatility index (PI)	1.34 ± 0.94	0.96 ± 0.61
	Resistance index (RI)	0.64 ± 0.16	0.55 ± 0.17
	Peak systolic velocity (PSV)	19.8 ± 14.6	27.3 ± 16.6
	(Time-averaged)Mean velocity (TAMX)	11.4 ± 9.7	17.4 ± 11.5
B-Mode Ultrasono Graphy	Abdominal fluid (Asc)	32.7 %	67.3 %
	Unilocular cyst (Un)	45.8 %	5.0 %
	Unilocular solid (Unsol)	6.5 %	15.6 %
	Multilocular cyst (Mul)	28.7 %	5.7 %
	Multilocular solid (Mulsol)	10.7 %	36.2 %
	Solid tumor (sol)	8.3 %	37.6 %
Morphologic	Bilateral mass (Bilat)	13.3 %	39.1 %
	Smooth wall (Smooth)	56.8 %	5.8 %
	Irregular wall (Irreg)	33.8 %	73.2 %
	Papillations (Pap)	13.0 %	53.2 %
	Septa$>$ 3mm (Sept)	13.0 %	31.2 %
	Acoustic shadows(Shadows)	12.2 %	5.7 %
Echogenicity	Anechoic cystic content(Lucent)	43.2 %	29.1 %
	Low level echogenicity (Low_level)	12.0 %	19.9 %
	Mixed echogenicity (Mixed)	20.3 %	13.5 %
	Ground glass cyst (G_glass)	19.8 %	8.5 %
	Hemorrhagic cyst (Haem)	3.9 %	0.0 %

Note: for continuous variables, mean$\pm$SD of benign and malignant class respectively, are reported; for binary variables, the occurrences (%) of the corresponding features are reported.

Figure 1 shows the biplot generated by the first two principal components of the data set, visualizing the correlation between the variables, and the

relations between the variables and classes. In particular, a small angle between two variables such as (Age, Meno) points out that those variables are highly correlated; the observations of malignant tumors (indicated by '+') have relatively high values for variables Sol, Age, Meno, Asc, L_CA125, Colsc4, Pap, Irreg, etc; but relatively low values for the variables Colsc2, Smooth, Un, Mul, etc. The biplot reveals that many variables are correlated, implying the need of variable selection. On the other hand, quite a lot of overlap between the two classes can also be observed, suggesting that the classical linear techniques might not be enough to capture the underlying structure of the data.

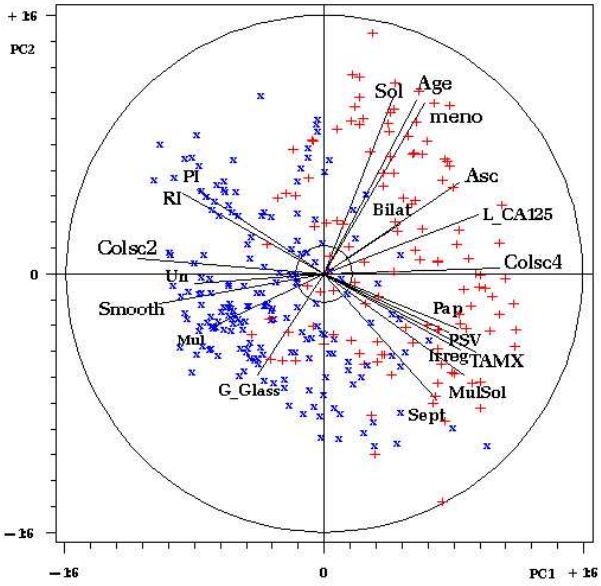


Figure 1. Biplot of ovarian tumor data ('×'- benign, '+'- malignant), projected on the first two principal components.

3.2 Experimental Settings

Our experiments consist of two parts. One part is the so called "temporal validation". In these experiments, the data set is split according to the time scale: the data from the first treated 265 patients (collected between 1994

and 1997) are taken as training set, 160 of the remaining data (collected between 1997 and 1999) are used as test set. In this way, we are trying to mimic a kind of prospective evaluation of our models, which is important before using the models in clinical practice. The proportion of malignant tumors in the training set and in the test set is both about $1/3$. However, such hold out cross validation is somehow biased and depends on the splitting of the training and test sets (Altman *et al.* 2000). Hence in the second part of the experiment, we try to reduce the bias by randomizing the data set and repeat the hold out cross-validation 30 times. The generalization performance of the models will be evaluated by averaging. The proportion of the malignant cases has been kept to be the same (about $1/3$) for all the realization of the training and test sets.

All the input data have been normalized using the mean and variance estimated from the training data. Several competitive models are built and evaluated using the same variables selected from the procedures including stepwise logistic regression and forward selection based on the Bayesian LS-SVM model evidence. Various models have been developed and evaluated using the following probabilistic modelling techniques: logistic regression, Bayesian MLPs, Bayesian LS-SVM with linear and RBF kernels, and the relevance vector machines with linear and RBF kernels. We also show the performance for non-probabilistic models such as standard SVMs and LS-SVMs, and we can see why the probabilistic output is needed for our cancer diagnosis tasks.

All the experiments have been done in matlab, except for the explorative data analysis and stepwise logistic regression which were conducted in SAS. Several matlab toolboxes have been employed, for example Netlab [1] was used for logistic regression and Bayesian MLP modelling, the LS-SVMlab [2] for standard and Bayesian LS-SVM modelling, the SparseBayes V1.0 [3] for fast sequential sparse Bayesian modelling after some modifications.

[1] http://www.ncrg.aston.ac.uk/netlab/

[2] http://www.esat.kuleuven.ac.be/sista/lssvmlab/

[3] http://research.microsoft.com/mlp/RVM/SparseBayesV1.00.tar.gz

3.3 Selecting Predictive Input Variables

Consistent with our "temporal" validation, we selected the variables based on only the 265 training data collected between 1994 and 1997. Firstly, we adapted the forward selection which tries to maximize the evidence of the LS-SVM classifiers with either linear or RBF kernels. In order to stabilize the selection, the three variables with the smallest univariate model evidence are first removed. Then the selection was limited to the remaining 24 candidate variables. Figure 2 shows the evolution of the model evidence during variable selection using RBF kernels. The Bayes factor for the univariate model is obtained by comparing it to a model with only a random variable, the other Bayes factors are obtained by comparing the current model to the previously selected models. The variable added to the model at each selection step and the corresponding Bayes factor have been depicted. We ended up with ten variables based on LS-SVM with RBF kernels. The order of selection for the ten variables are the following : *L_CA125, Pap, Sol, Colsc3, Bilat, Meno, Asc, Shadows, Colsc4, Irreg*, and we denote this subset of variables as MODEL1.

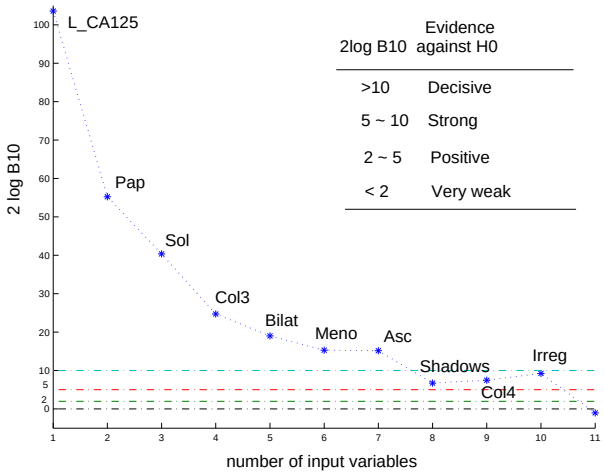


Figure 2. Evolution of the model evidence during forward input selection for LS-SVM with RBF kernels.

LS-SVMs with linear kernels have also been tried for variable selection, but resulted into a smaller evidence and an inferior model performance. Therefore, no discussion will be given on the models based on the subset of variables selected using linear kernels.

We also performed classical stepwise logistic regression, and only eight variables (denoted as MODEL2) are needed for the final logit model to obtain a satisfactory goodness of fit (compared to the full model that uses all the variables): *L_CA125, Pap, Sol, Asc, smooth, Col3, Unsol, Shadows*. This model doesn't include the variable Meno, which is important though according to the medical expert, since postmenopausal patients have a higher risk for ovarian cancer than premenopausal patients. However, in order to have a fair comparison for the variable selection method, i.e., without considering the expert knowledge, we keep using this subset of variables in this study.

In practice, many other models have been built by trying different combination of the variables and taking into account the opinion of the experts. Other types of variable selection procedures based on logistic regression have also been tried, such as backward, forward selection, and the score selection based on the likelihood score (chi-square) statistic for various model sizes. Here we won't describe all those models. Instead, we pay more attention to the models selected with least human interactions. At the end, the experts will check the automatically selected variables and see whether they are meaningful and interpretable.

3.4 Model Fitting and Prediction

For MLP models, we use MacKay's Bayesian MLP classifier (MacKay 1992, Bishop 1995), which is limited to one hidden layer with two hidden neurons, with hyperbolic tangent activation function for the hidden layer, and logistic activation function for the output layer. Other models with various number of hidden neurons were also tried, but not reported here due to their smaller evidence and inferior performance on the test set. Because of the existence of multiple local minima, the MLP classifier was trained

10 times with different initialization of the weights, and the one with the highest evidence was chosen.

The model fitting procedure for Bayesian LS-SVM classifiers has two stages. The first is the construction of the standard LS-SVM model within the evidence framework. Sparseness can be imposed to LS-SVMs at this stage in order to improve the generalization ability, by iteratively pruning, e.g. the 'easy' cases which have negative support values in $\mathbf{a}$. At the second stage, all the available training data will be used to compute the output probability, indicating the posterior probability for a tumor to be malignant.

As to relevance vector machines, given the kernel type and corresponding kernel parameter (e.g. the bandwidth r for RBF kernels), the fast sequential learning algorithm was used for training (Tipping and Faul 2003). In our experiments, for each iterative optimization step, ten relevance values (corresponding to ten different kernel bases) are randomly selected as candidates for modification. This procedure was repeated until convergence. We used the same approach to handle the same local minima problem as in the MLP learning: choosing the parameters with the highest model evidence from multiple (five in this case) runs of the optimization process.

In risk minimization decision making, different error costs should be considered in order to reduce the expected loss. Since misclassification of a malignant tumor is very serious, we need a model with high sensitivity while keeping an acceptable low false positive rate. The unbalance of the training data will normally influence the predicted outcome of the models, which minimize the total errors regardless of misclassification cost. That is, the resulting model will tend to give a higher probability for the prevalent class, which is the class of benign tumors in this case. Thus, the often used probability cutoff level 0.5 will lead to a low sensitivity for malignancy, and is therefore not suitable for cancer diagnosis. Without incorporating the prior costs, an acceptable cutoff level would be less than 0.2 or so, in order to reach a sensitivity of 80%. Hence, the adjusted prior for the malignant class was used in the experiments, which is intuitively set to $2/3$, higher than that of the benign class (set to $1/3$).

The same adjusted class priors have been combined for the computation of the posterior output for all models under consideration. For Bayesian LS-SVM models, this is done via (17). For models using logistic output activation functions, such as LRs, MLPs and RVM models, we incorporate the $2:1$ ratio of the adjusted prior class probability between the malignant and benign cases, into the computation of the probability outcome. That is, by correcting the bias term b_0 in the logit function as follows: $b = b_0 - \log(N_+/N_-) + \log(2/1)$, where N_+ and N_- denote the number of malignant and benign cases in the training set respectively.

3.5 Model Evaluation

The model performance is assessed by ROC analysis. Unlike the classification accuracy, ROC is independent of class distributions or error costs, and has been widely used in the biomedical field (Provost *et al.* 1998). The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for the different cutoff values of a diagnostic test. Here the *sensitivity* and *specificity* is the correct classification rate for the malignant and benign class, respectively. The area under the ROC curve (AUC) can be statistically interpreted as the probability of the classifier to correctly classify malignant cases and benign cases (Hanley and McNeil 1982). In this work, the AUC was obtained by a nonparametric method based on the Wilcoxon statistic, using the trapezoidal rule, to approximate the area (Hanley and McNeil 1982). The method proposed in (DeLong *et al.* 1988) was used to compute the variance and covariance of the nonparametric AUC estimates derived from the same set of cases. AUC can be used for comparing two different ROC curves. Related performance measures include also positive predictive value (PPV) and negative predictive value (NPV). The PPV is the proportion of all positive tests that are true positive; the NPV is the proportion of all negative tests that are true negative. The predictive value can help in interpreting the test result for an individual.

3.5.1 Results from Temporal Validation

Table 2 reports the performance of different models on the test set. The performance of the Risk-of-Malignancy index (RMI), a widely used score system (calculated as the product of the CA 125 level, a morphologic score, and a score for the menopausal status), is also shown as a reference (Jacobs *et al.* 1990). The AUC and its corresponding standard error (SE), which are independent of the cutoff value, are reported in the second column. The four performance measures, including the accuracy, sensitivity, specificity, PPV and NPV, were calculated at different decision levels. Except for RMI which is not a probabilistic model, all other decision levels are actually probability cutoff levels. For RVM models, the numbers between parentheses in the first column indicate the number of support (or relevance) vectors (N_{SV}) used for prediction. For LS-SVM classifiers, N_{SV} refers to the number of support vectors used in the first stage of model fitting other than in prediction, since all the training data have been used in prediction (when computing the probability distributions of the two classes over the latent output) in this study. Note that the decision levels we used in the experiments are considered 'good' according to our goal. They lead to a high sensitivity and low false positive rate on the training set; those decision levels, which result into a high accuracy but a too low sensitivity or specificity, are considered unacceptable in this context.

The ROC curves for the series of linear models and nonlinear models have been depicted in Figure 3. The ROC curve for RMI on the test set is shown in both figures for comparison.

All the competitive models, both linear and nonlinear models, seem to perform much better than RMI. Let us consider now the case in which the models are given the same set of input variables. The LR models do not perform as well as the other linear and nonlinear models. This can be partially explained by the fact that no regularization has been used in LR modelling. No significant difference can be observed among the other models including the Bayesian MLPs, the Bayesian LS-SVMs with linear and RBF kernels, and the RVMs with linear and RBF kernels.

Table 2. Comparison of the temporal validation performance on the **test set** ($N_{\mathrm{train}} = 265, N_{\mathrm{test}} = 160$).

Model Type (N_{SV})	AUC $\pm SE$	Decision Level	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
RMI	0.8733	100	78.13	74.07	80.19	65.57	85.86
	± 0.0298	**75**	**76.88**	**81.48**	**74.53**	**61.97**	**88.76**
LR1	0.9111	0.5	81.25	74.07	84.91	71.43	86.54
	± 0.0246	0.4	80.63	75.96	83.02	69.49	87.13
		0.3	80.63	77.78	82.08	68.85	87.88
		0.2	**80.63**	**81.48**	**80.19**	**67.69**	**89.47**
MLP1	0.9174	0.5	82.50	77.78	84.91	72.41	88.24
	± 0.0224	0.4	83.13	81.48	83.96	72.13	89.90
		0.3	**81.87**	**83.33**	**81.13**	**69.23**	**90.53**
LS-SVM1$_{\mathrm{Lin}}$	0.9141	0.5	82.50	77.78	84.91	72.41	88.24
(118)	± 0.0236	0.4	81.25	77.78	83.02	70.00	88.00
		0.3	**81.87**	**83.33**	**81.13**	**69.23**	**90.53**
LS-SVM1$_{\mathrm{RBF}}$	0.9184	0.5	84.38	77.78	87.74	76.36	88.57
(97)	± 0.0225	0.4	83.13	81.48	83.96	72.13	89.90
		0.3	**84.38**	**85.19**	**83.96**	**73.02**	**91.75**
RVM1$_{\mathrm{Lin}}$	0.9196	0.5	83.75	81.48	84.91	73.33	90.00
(5)	± 0.0231	0.4	83.13	85.19	82.08	70.77	91.58
		0.3	**83.13**	**85.19**	**82.08**	**70.77**	**91.58**
RVM1$_{\mathrm{RBF}}$	0.9195	0.5	84.38	85.19	83.96	73.02	91.75
(5)	± 0.0222	0.4	83.75	85.19	83.02	71.88	91.67
		0.3	**83.75**	**87.04**	**82.08**	**71.21**	**92.55**
LR2	0.9106	0.5	82.50	74.07	86.79	74.07	86.79
	± 0.0229	0.4	80.63	77.78	82.08	68.85	87.88
		0.3	**80.63**	**79.63**	**81.13**	**68.25**	**88.66**
		0.2	77.50	83.33	74.53	62.50	89.77
MLP2	0.9144	0.5	83.75	79.63	85.85	74.14	89.22
	± 0.0225	**0.4**	**80.63**	**79.63**	**81.13**	**68.25**	**88.66**
		0.3	78.75	83.33	76.42	64.29	90.00
LS-SVM2$_{\mathrm{Lin}}$	0.9144	0.5	84.38	79.63	86.79	75.44	89.32
(106)	± 0.0224	0.4	80.63	79.63	81.13	68.25	88.66
		0.3	**80.00**	**83.33**	**78.30**	**66.18**	**90.22**
LS-SVM2$_{\mathrm{RBF}}$	0.9158	0.5	84.38	79.63	86.79	75.44	89.32
(97)	± 0.0224	0.4	83.13	79.63	84.19	72.88	89.11
		0.3	**79.37**	**83.33**	**77.36**	**65.22**	**90.11**
RVM2$_{\mathrm{Lin}}$	0.9158	0.5	83.75	75.93	87.74	75.93	87.74
(4)	± 0.0222	**0.4**	**80.63**	**83.33**	**79.25**	**67.16**	**90.32**
		0.3	79.37	85.19	76.42	64.79	91.01
RVM2$_{\mathrm{RBF}}$	0.9147	0.5	81.25	79.63	82.08	69.35	88.78
(10)	± 0.0225	**0.4**	**80.63**	**79.63**	**81.13**	**68.25**	**88.66**
		0.3	79.37	87.04	75.47	64.38	91.95

The 'best' results of each model obtained at a certain decision level are indicated in bold. A number is added to the name of the classifier in order to distinguish the subsets of variables in use: '1' stands for MODEL1 and '2' for MODEL2. Subscripts 'RBF' and 'Lin' indicate the kernel type that is used.

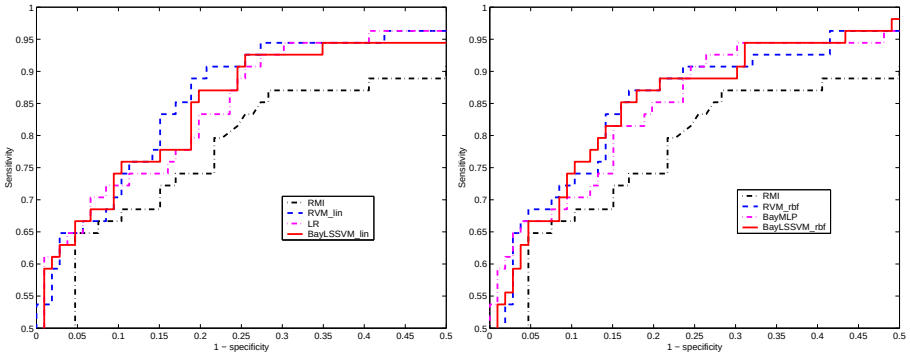


Figure 3. ROC curves on the test set. Different types of linear (left) and nonlinear (right) models using MODEL1 (ten variables) and the RMI as a reference model have been compared.

The sparsity of the RVM models deserves our special attention. The parsimony of the models can not only improve the generalization capability but also increase the interpretability of the model. Only five data points are needed for computing the outcome of RVM models both with linear and RBF kernels given MODEL1.

As to the input variable selection methods, forward selection based on the RBF LS-SVMs performed better than stepwise logistic regression. This is quite obvious, since models using the 10 variables of MODEL1 obtained better performance compared to using MODEL2. As we mentioned previously, the lack of menopausal status for a patient might be one reason why the models with MODEL2 do not generalize as well as with MODEL1.

We also tried to check the ability of our classifiers in rejecting uncertain test cases which need further examination by a human expert. The discrepancy (in absolute value) between the posterior probability and the cutoff value reflects the uncertainty of the prediction: the smaller the discrepancy, the larger the uncertainty. The performance of the models has been re-evaluated after rejecting a certain number of the most ‘uncertain’ test cases. It is shown in Figures 4 and 5, how the rejection of uncertain cases can improve the performance of the probabilistic models. The performance

with rejection for standard SVM models and standard LS-SVM models is also depicted for comparison. Here the hyperparameters for both standard SVM and LS-SVM models were selected using 5-fold CV. Due to the absence of the probability outcome for these non-probabilistic models, the rejection is based on the absolute value of the predictive outcome (as the discrepancy between zero and the latent output value) of the model. If this value reflects the uncertainty in prediction, the similar trend should be observed, that is, the performance of the model should increase as the fraction of test cases to be rejected increases.

Figure 4 shows the AUC-rejection curves for all the linear and nonlinear models using MODEL1, respectively. In both cases, it is observed that all probabilistic models keep getting higher performance after rejecting more and more uncertain cases, resulting into some sort of smoothly increasing curves. Some models such as linear Bayesian LS-SVM, linear RVM and Bayesian MLP and RVM with RBF kernels can even reach AUC of 0.95 after rejecting 18% of uncertain cases. On the contrary, the curves for non-probabilistic models of SVM and LS-SVM models appear to be bumpy and even keep decreasing in some regions. Moreover, probabilistic models obtain much higher AUCs compared to the others. This indicates that the posterior probability of the models does provide a way of quantifying uncertainty, which is essential for decision making in medical diagnosis.

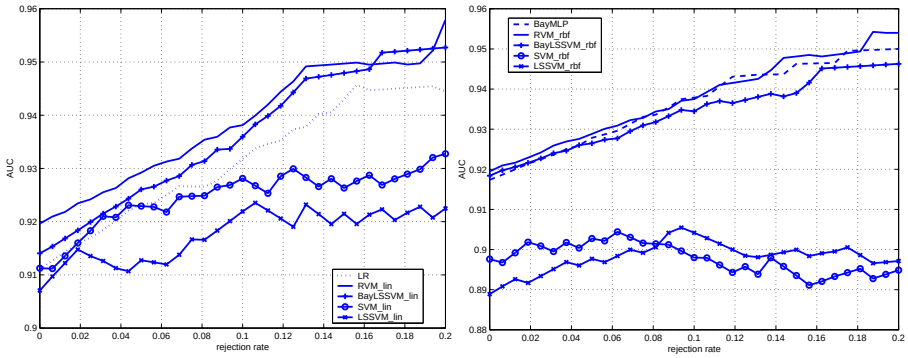


Figure 4. Test AUC vs. increasing rejection rate for different linear (left) and nonlinear (right) models.

Now we focus on the models of our interest - probabilistic models. More details of the performance at the cutoff level of 0.5 could be seen in Figure 5, where the sensitivity and the false positive rate have been plotted against the rejection rate. We clearly see how the sensitivity slowly rises with the increase of the rejection rate, while the false positive rate decreases with a slightly higher speed. This shows that the decrease in false alarm rate contributes more to the improvement of the overall model performance.

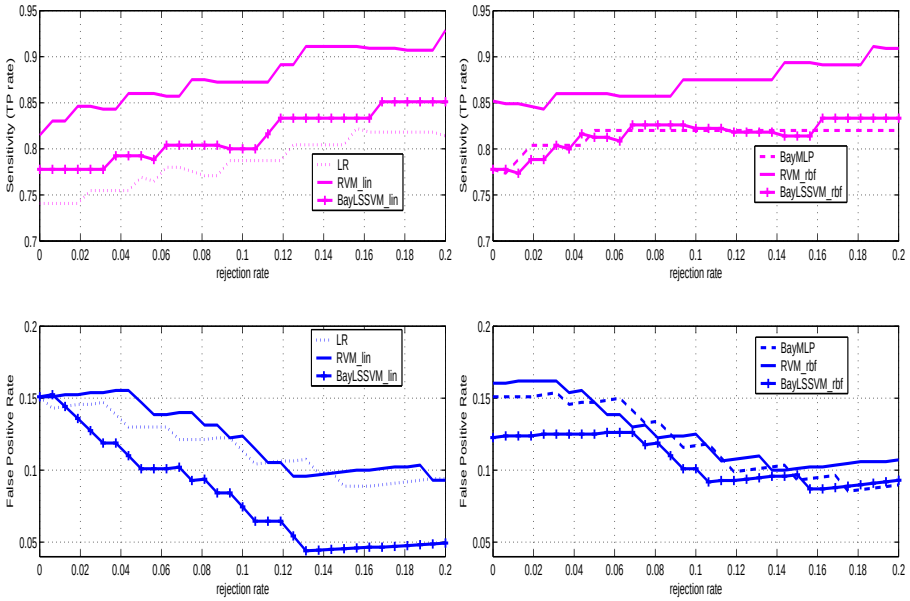


Figure 5. Test sensitivity and false positive rate vs. increasing rejection rate for different linear (left) and nonlinear (right) models (at 0.5 probability cutoff level).

3.5.2 Results from Multiple Runs of Randomized Cross Validation

In order to see the performance of our model with less bias, we performed another set of experiments in which the holdout cross-validation based on random splitting of training and test set was repeated 30 times. The average performance is considered as the model performance in the ideal situation

where the training and test set have similar distributions.

Table 3 presents the averaged performance of the competing models from the 30 runs of randomized cross validation by using MODEL1 and MODEL2. The average of AUC ($\overline{\mathrm{AUC}}$), the corresponding standard deviation (SD, derived from 30 AUC values) are reported. The number between the parentheses indicates the mean of the number of support vectors or relevance vectors ($\overline{N_{\mathrm{SV}}}$).

From this experiment, an increase of the validation performance is observed, which is mainly due to the randomization of the training set and test set. Among all the models, RMI has the worst performance. LS-SVM1$_{\mathrm{RBF}}$ obtained the best averaged performance, with mean AUC=0.9424. However, mixed performance were obtained by both the linear models and nonlinear models. No conclusions can be drawn that nonlinear models are superior to the linear models (or models using linear kernels), or vice versa. It is also difficult to tell whether MODEL1 gives better results than MODEL2.

4 Discussion

In this section, we would like to discuss several issues related to the development and evaluation of our diagnostic models in clinical practice.

- *Comparison with human investigators.* The diagnostic performance of our models has not yet reached the same level as the very experienced human expert (Lu *et al.* 2003a). The human expert made his diagnosis based on all the clinical data and measurements of the patients before operation (Timmerman *et al.* 1999c). His diagnosis on the 160 newest patients results in a sensitivity of 81.48% and specificity 93.40%. Whereas for Bayesian MLPs, Bayesian LS-SVMs and RVMs with linear kernels, in the case of using MODEL1, the specificity is around 85% when the sensitivity is at the same level of 81.48%. However, we still see the potential value of these mathematical models for less experienced human investigators in helping them to make the correct preoperative

Table 3. Averaged performance on the **test set** from 30 runs of randomized cross validation ($N_{\mathrm{train}} = 265, N_{\mathrm{test}} = 160$).

Model Type ($\overline{N_{\mathrm{SV}}}$)	$\overline{\mathrm{AUC}}$ $\pm SD$	Decision Level	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
RMI	0.8882	100	82.65	81.73	83.06	68.89	90.96
	$\pm$0.0284	**80**	**81.10**	**83.87**	**79.85**	**65.61**	**91.63**
LR1	0.9397	0.6	84.94	86.33	84.30	71.64	93.24
	$\pm$0.0209	**0.5**	**83.29**	**89.33**	**80.55**	**67.81**	**94.43**
		0.4	81.94	91.60	77.55	65.16	95.38
MLP1	0.9409	0.6	84.44	87.20	83.18	70.40	93.55
	$\pm$0.0198	**0.5**	**82.15**	**90.73**	**78.24**	**65.64**	**94.96**
		0.4	79.63	93.27	73.42	61.64	96.08
LS-SVM1$_{\mathrm{Lin}}$	0.9405	0.5	84.31	87.40	82.91	70.09	93.62
(150.2)	$\pm$0.0199	**0.4**	**82.77**	**90.47**	**79.27**	**66.61**	**94.88**
LS-SVM1$_{\mathrm{RBF}}$	0.9424	0.5	84.85	86.53	84.09	71.46	93.31
(137.1)	$\pm$0.0207	**0.4**	**83.52**	**90.00**	**80.58**	**67.98**	**94.71**
RVM1$_{\mathrm{Lin}}$	0.9411	0.6	84.81	86.93	83.85	71.21	93.49
(4.6)	$\pm$0.0199	**0.5**	**83.06**	**90.00**	**79.91**	**67.29**	**94.71**
		0.4	81.15	92.67	75.91	63.81	95.87
RVM1$_{\mathrm{RBF}}$	0.9407	0.6	84.96	86.53	84.24	71.60	93.32
(6.1)	$\pm$0.0195	**0.5**	**83.13**	**89.67**	**80.15**	**67.39**	**94.55**
		0.4	81.25	92.20	76.27	64.04	95.61
LR2	0.9399	0.6	86.06	88.47	84.97	73.17	94.29
	$\pm$0.0175	**0.5**	**83.31**	**91.40**	**79.64**	**67.44**	**95.44**
		0.4	80.83	94.20	74.76	63.18	96.66
MLP2	0.9404	0.6	85.81	88.47	84.61	72.65	94.26
	$\pm$0.0181	**0.5**	**82.94**	**92.27**	**78.70**	**66.68**	**95.82**
		0.4	79.60	94.20	72.97	61.53	96.57
LS-SVM2$_{\mathrm{Lin}}$	0.9413	0.6	86.71	88.07	86.09	74.44	94.13
(149.7)	$\pm$0.0173	**0.5**	**84.60**	**91.00**	**81.70**	**69.59**	**95.30**
		0.4	82.35	93.47	77.30	65.40	96.37
LS-SVM2$_{\mathrm{RBF}}$	0.9376	0.6	86.33	87.00	86.03	74.12	93.63
(145.2)	$\pm$0.0198	0.5	84.63	90.00	82.18	69.90	94.85
		0.4	82.42	92.67	77.76	65.72	95.98
RVM2$_{\mathrm{Lin}}$	0.9419	0.6	85.98	89.13	84.55	72.69	94.57
(4.8)	$\pm$0.0172	**0.5**	**82.85**	**92.60**	**78.42**	**66.42**	**95.97**
		0.4	79.96	94.47	73.36	61.89	96.74
RVM2$_{\mathrm{RBF}}$	0.9405	0.6	85.23	89.40	83.33	71.18	94.63
(6.2)	$\pm$0.0185	**0.5**	**82.67**	**92.80**	**78.06**	**66.13**	**96.07**
		0.4	79.87	94.80	73.09	61.80	96.93

The 'best' results of each model obtained at a certain decision level are indicated in bold. A number is added to the name of the classifier in order to distinguish the subsets of variables in use: '1' stands for MODEL1 and '2' for MODEL2. Subscripts 'RBF' and 'Lin' indicate the kernel type that is used.

prediction of the outcome. It is remarkable that the expert had examined ultrasonographically more than 5000 patients, while the models were trained using only 265 patients. Moreover, unlike the human expert, the

mathematical models have not accessed some clinical features, e.g., the medical history, family history, genetic factors, and the whole image of transvaginal sonography.

- *Incorporation of Prior Knowledge.* There are still some steps that might help in improving the performance of the models. One is to incorporate the domain knowledge into the learning of the black-box models. However, further investigation and validation of such a hybrid approach for this ovarian tumor classification problem is still needed.

- *On Bayesian Modelling.* Other types of priors for the model parameters can be used. The posterior probability of the model (the marginalization) can be computed in a more accurate way, using MCMC sampling and variational methods (Neal 1996, Bishop and Tipping 2000).

- *Dealing with missing values.* Missing data occur often in reality. For example, the level of CA 125, though the best known ovarian tumor marker till now, is not always available due to various reasons such as the cost. We could build another series of models without considering this variable. A more general solution to the incomplete data problem is *imputation*, i.e. to replace the missing value by a plausible value using, e.g., the EM algorithm (Dempster *et al.* 1977). Multiple imputation (Rubin 1987) could also be applied, where the missing values are estimated by a set of simulated values. According to our preliminary test results, this could be a valuable tool to handle such problems in cancer diagnosis.

- *Uncertainty in Input Variables.* This might influence the variable selection and model selection in general. Some variables (such as CA 125 level) are quite objective; whereas, some are subjective or semi-subjective (such as color score), and some are noisy. There is a confidence and reliability measure associated with each variable or measurement. The Bayesian approach we considered so far has taken into account the uncertainty in the model parameters and the output, however, it relies on the assumption that the input is known exactly. How to integrate the uncertainty being present in the input variables into the

variable selection and prediction remains an open research question.

- *Data Splitting in Model Evaluation.* The splitting of the data set into a training and test set according to the time scale is more natural in clinical practice. There is a danger for changes in the patient population over time, however. For instance, the patients of the old group (recruited before 1997) are elder in average than the patients of the new group (recruited after 1997). Consequently the old group has a higher proportion of post-menopausal patients than in the new group. It is well known that the level of CA 125 can better predict the presence of cancer in post-menopausal patients, compared to that in pre-menopausal patients. This implies that it is harder to predict correctly the malignancy of the tumors in the new patient group. A random splitting of test and training set leads to a more equilibrated distribution of the patient data over both sets other than for a random variation, and is thus a weak procedure and less stringent (Altman *et al.* 2000).

- *Multi-center Study.* The temporal validation performance of our Bayesian models is quite encouraging though not perfect. Since the underlying mechanism of the disease is not yet clear, a statistically reliable model is only possible based on a sufficiently large amount of training examples. This is one motivation for the International Ovarian Tumor Analysis (IOTA) project (Timmerman *et al.* 2000). More than 1000 patient data from more than ten centers located in different countries have been collected. Based on this large data set, mathematical models can be developed for not only preoperative classification of benign and malignant ovarian tumors, but also subclassification of the tumors (e.g. borderline malignant, endometrioma). The variation between centers in outcomes of histology and the performance of the models will also be assessed. Then another 1000 patient data will be collected for future prospective validation.

5 Conclusions: Towards Robust Ovarian Cancer Diagnostic Models

In this paper, we have discussed the application of various probabilistic models to predict the malignancy of the ovarian tumors preoperatively. Particularly the LS-SVM and RVM classifiers with Bayesian learning have been advocated. Within the Bayesian evidence framework, the hyperparameter tuning, input variable selection and computation of posterior class probability can be conducted in a unified way. More flexibility has been given to the modelling by use of (either linear or nonlinear) kernels. Compared with RVMs, Bayesian LS-SVMs have the advantages in training speed and global optimization in its conventional training step, though the model of RVM is usually very sparse. Unlike the two types of kernel based models, extra hyperparameters need to be tuned for the MLP modelling such as the structure of the neural network. Our results demonstrate that both the Bayesian LS-SVM models and RVM models have the potential to obtain a reliable preoperative distinction between benign and malignant ovarian tumors, and to assist the clinicians for making a correct diagnosis.

Now, we summarize the model development for ovarian cancer diagnosis, and highlight some important issues that should be emphasized during the data modelling.

First of all, medical classification is an integral part of the medical decision making process, where one should consider not only the classification accuracy, but also many other factors, such as the uncertainty in model selection and parameter estimation, different misclassification cost and the variance in the distribution of the patient data with time and environment. Remember that the reasoning is done with the presence of uncertainty and the risk should be minimized when making a decision. In working towards a more robust medical diagnostic model, the following issues are worth examining.

First, instead of emphasizing the predicting accuracy only, other medical

diagnostic measures should be considered, such as the use of ROC curves, sensitivity, specificity, positive predictive value, etc. Instead of just presenting the black and white classification output, probabilistic output values should be provided.

Second, the patient data should be adequately preprocessed. Data screening, explorative data analysis with some visualization is useful in getting insights into the data and choosing the modelling techniques. Selection of input variables could help in enhancement of the model performance and better understanding of the underlying medical problems.

Third, a control system should be used to decide on the best model to use for a certain case. We can not claim that just one of the models is the best diagnostic model for all the cases. But we may conclude that a certain class of models should be used for a certain type of patients. The different models can be combined into committee machines. In particular, in a Bayesian framework, the posterior probabilities can be obtained via model averaging.

The art of model building and model selection deserves a lot of attention. The model structure, the hyperparameters such as the amount of regularization, and parameters for the kernels need to be tuned. The Bayesian approach, which handles the uncertainty in modelling and prediction, is advocated for this purpose.

It is remarkable that different criteria exist for the academic world versus the clinical world. Also, to build a good system we need sufficient data and experiments. A multi-center study is essential for collecting enough data and getting diversity in the patients. Clinically, more confidence can be gained if we test the models in a multi-center environment.

Acknowledgments

This research is supported by the projects of Belgian Federal Government IUAP IV-02 and IUAP V-22, of the Research Council KUL MEFISTO-

666, IDO/99/03 and IDO/02/009, and the FWO projects G.0407.02 and G.0269.02. C. Lu is also supported by a KUL doctoral fellowship.

References

Altman, D.G., Royston, P. (2000), "What do we mean by validating a prognostic model?" *Statistics in Medicine*, vol. 19, pp. 453-473.

Antal, P., Verrelst, H., Timmerman, D., Moreau, Y., Van Huffel, S., De Moor, B., Vergote, I. (2000), "Bayesian networks in ovarian cancer diagnosis: potentials and limitations," *Proceedings of the 13th IEEE Symposium on Computer-Based Medical Systems (CBMS2000)* , pp. 103-109.

Antal, P., Fannes, G., Timmerman, D., De Moor, B., Moreau, Y. (2003), "Bayesian applications of belief networks and multilayer perceptrons for ovarian tumor classification with rejection," *Artif Intelligence in Medicine*, vol. 29, pp. 39-60.

Bishop, C.M. (1995), *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford.

Bishop, C.M., and Tipping, M.E. (2000), "Variational relevance vector machines". In: C. Boutilier and M. Goldszmidt (Eds.), *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, pp. 46-53.

Bishop, C.M., and Tipping, M.E. (2003), "Bayesian regression and classification". In Suykens, J.A.K. *et al.* (Eds.), *Advances in Learning Theory: Methods, Models and Applications*, Vol. 190. IOS Press, pp. 267-288.

Cristianini, N., Shawe-Taylor, J. (2000), *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK.

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

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977) "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society Series B*, vol. 39(1), pp. 1-38.

Hanley, J.A., McNeil, B. (1982), "The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve," *Radiology* vol. 143, pp. 29-36.

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

Jeffreys H. (1961), *Theory of Probability*. Oxford University Press, New York: 1961.

Lu, C., De Brabanter, J., Van Huffel, S., Vergote, I., Timmerman, D. (2001), "Using artificial neural networks to predict malignancy of ovarian tumors," *Proceedings of the 23rd International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC01)*, CD-ROM.

Lu, C., Van Gestel, T., Suykens, J.A.K., Van Huffel, S., Vergote, I., Timmerman, D. (2003a), "Preoperative prediction of malignancy of ovarium tumor using least squares support vector machines," *Artificial Intelligence in Medicine*, vol. 28(3), pp. 281-306.

Lu, C., Van Gestel, T., Suykens, J.A.K., Van Huffel, S., Vergote I., Timmerman D. (2003b), "Classification of ovarian tumor using Bayesian least squares support vector machines," *Proceedings of the 9th Conference on Artificial Intelligence in Medicine Europe(AIME03)*, pp. 219-228.

MacKay, D.J.C. (1995), "Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks," *Network: Computation in Neural Systems*, vol. 6, pp. 469-505.

MacKay, D.J.C. (1992), "The evidence framework applied to classification networks," *Neural Computation*, vol. 4(5), pp. 698-741.

Mercer, J. (1909), "Function of positive and negative type and their connection with the theory of integral equations," *Philosophical Trans. of the Royal Society London*, vol. A209, pp. 415-446.

Neal, R.M. (1996), *Bayesian Learning for Neural Networks*, Lecture Notes in Statistics, vol. 118, Springer, New York.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W. (1996), *Applied Linear Statistical Models*, fourth edition, McGraw-Hill/Irwin, Chicago.

Ozols, R.F., Rubin, S.C., Thomas, G.M., Robboy, S.J. (2000), "Epithelial Ovarian Cancer," in: Hoskins, W.J., Perez, C.A., Young, R.C., eds., *Principles and practice of gynecologic oncology*, Lippincott Williams and Wilkins, Philadelphia, pp. 981-1058.

Provost, F., Fawcett, T., Kohavi, R. (1998), "The case against accuracy estimation for comparing induction algorithms," in: Shavlik, J., eds., *Proceedings of the 15th International Conference on Machine Learning (IMLC-98)*, pp. 445-453.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc, New York.

Suykens, J.A.K., Vandewalle, J. (1999), "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9(3), pp. 293-300.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J. (2002), *Least Squares Support Vector Machines*, World Scientific, Singapore.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J. (2003), "Support vector machines: least squares approaches and extensions," in Suykens, J.A.K. *et al.* (Eds.), *Advances in Learning Theory: Methods, Models and Applications*, vol. 190, IOS Press, pp. 155-177.

Timmerman, D., Bourne, T.H., Tailor, A., Collins, W.P., Verrelst, H., Vandenberghe, K., Vergote, I. (1999), "A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: The development of a new logistic regression model," *Am J Obstet Gynecol*, vol. 181, pp. 57-65.

Timmerman, D., Verrelst, H., Bourne, T.H., De Moor, B., Collins, W.P., Vergote, I. and Vandewalle, J. (1999), "Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses," *Ultrasound Obstet Gynecol*, vol. 13, pp. 17-25.

Timmerman, D., Schwärzler. P., Collins, W.P., Claerhout, F., Coenen, M., Amant, F., Vergote, I., Bourne, T.H. (1999), "Subjective assesment of adnexal masses with the use of ultrasonography: an analysis of interobserver variability and experience," *Ultrasound Obstet Gynecol*, vol. 13, pp. 11-16.

Timmerman, D., Valentin, L., Bourne, T.H., Collins, W.P., Verrelst, H., Vergote, I. (2000), "Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group," *Ultrasound Obstet Gynecol*, vol. 16, pp. 500-505.

Tipping, M.E. (2001), "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211-244.

Tipping, M.E. and Faul, A. (2003), "Fast marginal likelihood maximisation for sparse Bayesian models," *Proceedings of Artificial Intelligence and Statistics '03*.

Van Gestel, T., Suykens, J.A.K., Baestaens, D-E., Lamrechts, A., Lanckriet, G., Vandaele, B., De Moor, B., Vandewalle, J. (2001), "Financial time series prediction using least squares support vector machines within the evidence framework," *IEEE Trans. on Neural Networks*, vol. 12(4), pp. 809-821.

Van Gestel, T., Suykens, J.A.K., Lanckriet, G., Lambrechts, A., De Moor, B., Vandewalle, J. (2002), "A Bayesian framework for least squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis," *Neural Computation*, vol. 14(5), pp. 1115-1148.

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

Vergote, I., De Brabanter, J., Fyles, A., Bertelsen, K., Einhorn, N., Sevelda, P., Gore, M.E., Kaern, J., Verrelst, H., Sjovall, K., Timmerman, D., Vandewalle, J., Van Gramberen, M., Trope, C.G. (2001), "Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma," *Lancet*, vol. 357(9251), pp. 176-82.