Bayesian Least Squares Support Vector Machines for Classification of Ovarian Tumors

Chuan Lu, Tony Van Gestel, Johan A. K. Suykens, Sabine Van Huffel
Dept. of Electrical Engineering, Katholieke Universiteit Leuven, 3001 Leuven, Belgium
{chuan.lu,tony.vangestel,johan.suykens,sabine.vanhuffel}
@esat.kuleuven.ac.be

Ignace Vergote, Dirk Timmerman
Dept. of Obstetrics and Gynecology, University Hospitals Leuven, 3000 Leuven, Belgium
{ignace.vergote, dirk.timmerman}@uz.kuleuven.ac.be

Abstract

The aim of this study is to develop the Bayesian Least Squares Support Vector Machine (LS-SVM) classifiers, for preoperatively predicting the malignancy of ovarian tumors. We describe how to perform parameter estimation, input variable selection for LS-SVM within the evidence framework. The issue of computing the posterior class probability for risk minimization decision making is addressed. The relation between the LS-SVM model and kernel principal component analysis is also indicated and used for interpretation of the LS-SVM classifiers.

1 Introduction

Ovarian masses are a very common problem in gynecology. The difficulties in early detection of ovarian malignancy result into the highest mortality rate among gynecologic cancers. An accurate discrimination between the benign and malignant tumors before operation is critical to obtain the most effective treatment and best advice, and will influence the outcome for the patient and the medical costs. Several attempts have been made in order to automate the classification process, such as the risk of malignancy index (RMI), logistic regression, neural networks, Bayesian belief networks [1, 3]. In this paper, we focus on the development of least squares support vector machines (LS-SVMs), to preoperatively predict the malignancy of ovarian tumors.

Support vector machines (SVM) have become a state-of-art technique for pattern recognition. The basic idea of the SVM classifier is: map an n-dimensional input vector $x \in \mathbb{R}^n$ into a high n_f-dimensional feature space by the mapping $\varphi(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^{n_f}, x \rightarrow \varphi(x)$, a linear classifier is then constructed in this feature space. Here a least squares version of SVM [4] is considered, in which 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. Also remarkable is that LS-SVM is closely related to kernel Fisher discriminant analysis. In order to achieve a high level of performance with LS-SVMs, the regularization and possible kernel parameters have to be tuned. The Bayesian evidence framework provides a unified theoretical treatment of learning in order to cope with similar problems in neural

networks [6]. Recently the Bayesian method has also been integrated into LS-SVMs, and a numerical implementation was derived. This approach was applied to several benchmark problems, achieving similar results as Gaussian Process Regression and SVMs [7].

The paper is organized as follows. We start in section 2 with a brief review of the LS-SVM classifier and its integration within the Bayesian evidence framework; then we introduce a way to compute the posterior class probabilities for minimum risk decisions making. An input selection scheme and an interpretation of the LS-SVM classifier by kernel PCA are addressed afterwards. In section 3, the application of LS-SVM for predicting the malignancy of ovarian tumors is demonstrated, the performance of models are assessed via Receiver Operating Characteristic (ROC) curve analysis.

2 LS-SVM and Bayesian evidence framework

The LS-SVM classifier $y(x) = \text{sign}\left[w^T\varphi(x) + b\right]$ is inferred from the data $D = \{(x_i, y_i)\}_{i=1}^N$ with binary targets $y_i = \pm 1$, by minimizing the following cost function:

$$\min_{w,b,e} \mathcal{J}_1(w, e) = \mu E_W + \zeta E_D = \frac{\mu}{2} w^T w + \frac{\zeta}{2} \sum_{i=1}^N e_i^2 \quad (1)$$

subject to the equality constraints $e_i = 1 - y_i[w^T\varphi(x_i) + b], i = 1, ..., N$. The regularization and sum of squares error term are defined as $E_W = \frac{1}{2} w^T w$, and $E_D = \frac{1}{2}\sum_{i=1}^N e_i^2$ respectively. The tradeoff between the training error and regularization is determined by the ratio $\gamma = \zeta/\mu$. This optimization problem can be transformed and solved through a linear system in the dual space [4]:

$$\left[\begin{array}{c|c} 0 & Y^T \\ \hline Y & \Omega + \gamma^{-1} I_N \end{array}\right] \left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 1_v \end{array}\right] \quad (2)$$

with $Y = [y_1 \cdots y_N]^T$, $\alpha = [\alpha_1 \cdots \alpha_N]^T$, $e = [e_1 \cdots e_N]^T$, $1_v = [1 \cdots 1]^T$, and I_N the $N \times N$ identity matrix. Mercer's theorem is applied to the matrix Ω with $\Omega_{ij} = y_i y_j \, \varphi(x_i)^T \varphi(x_j) = y_i y_j \, K(x_i, x_j)$, where $K(\cdot, \cdot)$ is a chosen positive definite kernel that satisfies Mercer condition. The most common kernels include a linear kernel $K(x_1, x_2) = x_1^T x_2$ and an RBF kernel $K(x_1, x_2) = \exp(-\|x_1 - x_2\|_2^2/\sigma^2)$. The LS-SVM classifier is then constructed in the dual space as: $y(x) = \text{sign}\left[\sum_{i=1}^N \alpha_i \, y_i \, K(x, x_i) + b\right]$.

2.1 Probabilistic inferences of LS-SVM within the evidence framework

The evidence framework consists of three levels of inference.

2.1.1 Inference of model parameters (level 1)

The parameter w and bias term b for given value of μ, ζ are inferred from the data D at the first level, by applying Bayes' rule:

$$p(w, b|D, \log\mu, \log\zeta, \mathcal{H}) = \frac{p(D|w, b, \log\mu, \log\zeta, \mathcal{H}) p(w, b|\log\mu, \log\zeta, \mathcal{H})}{p(D|\log\mu, \log\zeta, \mathcal{H})}, \quad (3)$$

where the model $\mathcal{H}$ corresponds to the kernel function K with different kernel parameters such as the width of an RBF kernel σ. The evidence $p(D|\log\mu, \log\zeta, \mathcal{H})$ is a normalizing constant, and will be needed in the next level of inference. Assuming a Gaussian prior over w with variance $1/\mu$, and a separate uniform distribution over b, we come to the prior with the form $p(w, b|\log\mu, \log\zeta, \mathcal{H}) \propto \exp(-\mu E_W)$. Assuming a Gaussian noise on the target variable, the likelihood is equal to $p(D|w, b, \log\zeta, \mathcal{H}) \propto \exp(-\zeta E_D)$, and the parameter ζ defines a noise level (variance) $1/\zeta$. Hence the expression for the first level of inference becomes $p(w, b|D, \log\mu, \log\zeta, \mathcal{H}) \propto \exp(-\mu E_W)\exp(-\zeta E_D) = \exp(-\mathcal{J}_1(w, b))$. Then maximum a posteriori estimates w_{MP} and b_{MP} can be obtained by optimizing (1), i.e. solving the linear set of equations in (2) in the dual space.

2.1.2 Inference of hyperparameters (level 2)

The hyperparameters are determined by maximizing the posterior probability of the parameters, which can be estimated using the Gaussian probability at $w_{\mathrm{MP}}, b_{\mathrm{MP}}$:

$$p(\log\mu, \log\zeta|D, \mathcal{H}) \propto \frac{\sqrt{\mu^{n_f}\zeta^N}}{\sqrt{\det H}} \exp(-\mathcal{J}_1(w_{\mathrm{MP}}, b_{\mathrm{MP}})), \tag{4}$$

with the Hessian $H = \partial^2\mathcal{J}_1(w,b)/\partial[w;b]^2$. The expression for $\det H$ is given by $N\mu^{n_f-N_{\mathrm{eff}}}\zeta\prod_{i=1}^{N_{\mathrm{eff}}}(\mu+\zeta\lambda_{G,i})$, where N_{eff} eigenvalues $\lambda_{G,i}$ are the non-zero eigenvalues of the centered Gram matrix in the feature space and are the solution of the eigenvalue problem $(M\Psi M)\nu_{G,i} = \lambda_{G,i}\nu_{G,i}, i = 1, ..., N_{\mathrm{eff}} \leq N-1$, with the centering matrix $M = (I_N - \frac{1}{N}1_v1_v^T)$ and where Ψ is the $N \times N$ Gram matrix with elements $\Psi_{ij} = K(x_i, x_j)$. Denote $V_G = [\nu_{G,1}, ..., \nu_{G,N_{\mathrm{eff}}}]$, and $\Lambda_G = \mathrm{diag}([\lambda_{\mathrm{G,i}}, ..., \lambda_{\mathrm{G,N_{eff}}}])$.

The effective number of parameters [5, 6] for LS-SVM can be shown to be:

$$\gamma_{\mathrm{eff}} = 1 + \sum_{i=1}^{N_{\mathrm{eff}}} \frac{\zeta_{\mathrm{MP}}\lambda_{G,i}}{\mu_{\mathrm{MP}} + \zeta_{\mathrm{MP}}\lambda_{G,i}} = 1 + \sum_{i=1}^{N_{\mathrm{eff}}} \frac{\gamma_{\mathrm{MP}}\lambda_{G,i}}{1 + \gamma_{\mathrm{MP}}\lambda_{G,i}}. \tag{5}$$

In practice, one can reformulate the optimization problem in μ and ζ into a scalar optimization problem in $\gamma = \zeta/\mu$:

$$\min \mathcal{J}_2(\gamma) = \sum_{i=1}^{N-1} \log[\lambda_{G,i} + \frac{1}{\gamma}] + (N-1)\log[E_W(w_{\mathrm{MP}}) + \gamma E_D(w_{\mathrm{MP}}, b_{\mathrm{MP}})], \tag{6}$$

with $\lambda_{G,i} = 0$ for $i > N_{\mathrm{eff}}$. The expressions for $E_W + \gamma E_D$ and E_W can be given in the dual variables: $E_W(w_{\mathrm{MP}}) + \gamma E_D(w_{\mathrm{MP}}) = \frac{1}{2}y^T M(M\Psi M + \gamma^{-1}I_N)^{-1}My$, and $E_W(w_{\mathrm{MP}}) = \frac{1}{2}y^T MV_G\Lambda_G(\Lambda_G + \gamma^{-1}I_{N_{\mathrm{eff}}})^{-2}V_G^T My$.

This optimal hyperparameter γ is then obtained by solving the optimization problem (6) with gradients using e.g. quasi-newton method. Given the optimal γ_{MP}, one can easily compute μ_{MP} and ζ_{MP} using their relations in the optimum.

2.1.3 Bayesian model comparison (level 3)

Level 3 inference ranks the model by examining its posterior $p(\mathcal{H}_j|D)$. Assuming uniform prior $p(\mathcal{H}_j)$ over all models, then the models can be ranked by their evidence $p(D|\mathcal{H}_j)$, which can be evaluated using a Gaussian approximation, and the expression in the dual space is the following:

$$p(D|\mathcal{H}_j) \propto \sqrt{\frac{\mu_{\mathrm{MP}}^{N_{\mathrm{eff}}}\zeta_{\mathrm{MP}}^{N-1}}{(\gamma_{\mathrm{eff}}-1)(N-\gamma_{\mathrm{eff}})\prod_{i=1}^{N_{\mathrm{eff}}}(\mu_{\mathrm{MP}} + \zeta_{\mathrm{MP}}\lambda_{G,i})}}. \tag{7}$$

2.2 Class probabilities for the LS-SVM classifiers

Given the posterior probability of the model parameters w and b in level 1 inference, we can now integrate over all w and b values so as to obtain the posterior probability $p(y|x, D, \log\mu, \log\zeta, \mathcal{H})$. Typically the training set is unbalanced, and the discriminant shifts toward the prevailing class. Therefore in [7] two error variables corresponding to class '+' and class '−' are introduced: $e_\pm = w^T(\varphi(x) - \hat{m}_\pm)$, where $\hat{m}_+$ and $\hat{m}_-$ are the centers of the positive and negative class respectively. After marginalizing over w the distribution of $e_\pm$ will also be Gaussian, centering around mean $m_{e_\pm}$ with variance $(\zeta_\pm^{-1} + \sigma_{e\pm}^2)$. The expression for the mean is $m_{e_\pm} = w_{\mathrm{MP}}^T(\varphi(x) - \hat{m}_\pm) =$

$\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) - \frac{1}{N_\pm} \sum_{i=1}^{N} \alpha_i y_i \sum_{j \in \mathcal{I}_\pm} K(x_i, x_j)$, where $\mathcal{I}_+$ and $\mathcal{I}_-$ indicate the sets of indices whose corresponding data points have positive and negative labels respectively.

The variance from the target noise $\zeta_\pm^{-1}$ for different classes may differ, and can be approximated by: $\zeta_\pm^{-1} = \sum_{j \in \mathcal{I}_\pm} e_{\pm,j}^2 / (N_\pm - \gamma_{\text{eff}} \frac{N_\pm}{N})$.

The additional variance $\sigma_{e_\pm}^2 = [\varphi(x) - \hat{m}_\pm]^T Q_{11} [\varphi(x) - \hat{m}_\pm]$ is due to the uncertainty in the parameters w, and where Q_{11} is the upper left $n_f \times n_f$ block of the covariance matric $Q = H^{-1}$. Let $\theta(x) = [K(x, x_1) \cdots K(x, x_N)]^T$; and define $1_+, 1_- \in \mathbb{R}^N$ as the vector with element zero or one, for $i = 1, ..., N$, $1_{\pm,i} = 1$ if $y_i = \pm 1$, otherwise $1_{\pm,i} = 0$. The derived expression for the variance in the dual space is the following [7]:

$$\begin{aligned} \sigma_{e\pm}^2 &= \frac{1}{\mu} K(x, x) - \frac{2}{\mu N_\pm} \sum_{i \in \mathcal{I}_\pm} K(x, x_i) + \frac{1}{\mu N_\pm^2} \sum_{i,j \in \mathcal{I}_\pm} K(x_i, x_j) - \\ & \frac{\zeta}{\mu} (\theta^T(x) - \frac{1}{N_\pm} 1_\pm^T \Psi) M (\mu I_N + \zeta M \Psi M)^{-1} M (\theta(x) - \frac{1}{N_\pm} \Psi 1_\pm). \quad (8) \end{aligned}$$

Thus the conditional probabilities can be computed as:

$$p(x|y = \pm 1, D, \log \mu, \log \zeta, \log \zeta_\pm, \mathcal{H}) = (2\pi(\zeta_\pm^{-1} + \sigma_{e_\pm}^2))^{-\frac{1}{2}} \exp(-\frac{m_{e_\pm}^2}{2(\zeta_\pm^{-1} + \sigma_{e_\pm}^2)}). \quad (9)$$

By applying Bayes' rule the posterior class probabilities of the LS-SVM classifier are obtained (for notation simplicity, $\log \mu, \log \zeta, \log \zeta_\pm, \mathcal{H}$ are dropped in this expression):

$$p(y|x, D) = \frac{p(y) p(x|y, D)}{P(y = 1) p(x|y = 1, D) + P(y = -1) p(x|y = -1, D)}, \quad (10)$$

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 trick to combine the posterior probability with the different error costs is by replacing $p(y)$ in (10) with the adjusted class prior, e.g. $P'(y{=}1) {=} P(y{=}1) c_+^- / (P(y{=}1) c_+^- + P(y{=}{-}1) c_-^+)$.

2.3 Input variable selection

In the Bayesian framework, given the likelihoods of the models $\mathcal{H}_0$ and $\mathcal{H}_1$, two models can be compared by the ratio of posterior probabilities: $p(\mathcal{H}_i|D) = \frac{p(D|\mathcal{H}_1) p(\mathcal{H}_1)}{p(D|\mathcal{H}_0) p(\mathcal{H}_0)} = \frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)} B_{10}$, where $B_{10} = \frac{p(D|\mathcal{H}_1)}{p(D|\mathcal{H}_0)}$ is the Bayes factor for model $\mathcal{H}_1$ against $\mathcal{H}_0$ from data D. If equal priors are assigned to the models, the posterior odds ratio then equals the Bayes factor, which 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 also shown in Fig. 2 [10].

In the context of the Bayesian evidence framework, the posterior likelihood (or evidence) of the models $p(D|\mathcal{H}_j)$ are computed with (7) on level 3 inference. A higher $p(D|\mathcal{H}_1)$ over $p(D|\mathcal{H}_0)$ means the data favor $\mathcal{H}_1$ to $\mathcal{H}_0$. Therefore, given a certain type of kernel for the model, we propose to select the input variables according to the model evidence $p(D|\mathcal{H}_j)$.

The procedure performs a forward selection (greedy search), starting from zero variables, and choosing 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.

2.4 Interpretation of the LS-SVM classifier using kernel PCA

In medical pattern recognition, it is required to be able to interpret the models. Here we attempt to interpret the LS-SVM classifiers in the high dimensional feature space by kernel principal component analysis (kPCA). The idea is to first transform the $\varphi(x)$ in the kernel induced feature space into a score vector $z = V^T(\varphi(x) - \hat{m}_\varphi)$, where $\hat{m}_\varphi = \frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)$, V are the projection vectors, corresponding to the principal components (PCs) found by applying PCA to the covariance matrix of the centered vectors $[\varphi(x_1) - \hat{m}_\varphi, ..., \varphi(x_N) - \hat{m}_\varphi]$. Since the feature space is typically unknown, z scores will be computed in the dual space using the kernel trick [8].

Now we check the relevance of the z scores with the estimation function of LS-SVM classifier, which intrinsically performs ridge regression for target $y = \pm 1$ using regularization parameter γ: $\tilde{y}(x) = w^T\varphi(x) + b = w^T V V^T(\varphi(x) - \hat{m}_\varphi) + b + w^T\hat{m}_\varphi = \tilde{w}^T z + \tilde{b}$, using $VV^T = I_N$, the regression coefficient vector $\tilde{w} = V^T w$ and bias term $\tilde{b} = b + w^T\hat{m}_\varphi$. If we take all nonzero z scores into account, no information of the data will be lost, so we can estimate the parameter $\tilde{w}$ by performing a ridge regression over the transformed data $\{(z_i, y_i)\}_{i=1}^{N}$. If we use the same γ as the LS-SVM, then the output of the two estimation functions will be equivalent. Based on this relation, the data points and the separating hyperplane in the feature space $w^T\varphi(x) + b = 0$ can be visualized by projecting them on different pairs of PCs in the feature space. In this supervised problem, PCs with large correlation to the target will be selected for visualization. Details for computation and benchmark examples can be seen in [8].

One can also superimpose the explanatory variables on the graph, forming into some kind of nonlinear biplot [11]. For the kth variable, pseudosamples can be generated by setting e.g. the data mean as the starting point, varying gradually the value of kth variable while fixing the others; then a trajectory can be obtained by tracing the projection of the pseudosamples onto the pairs of PCs in the feature space. In case of linear kernels, those trajectories can be called variable axis.

3 Application

3.1 Data

The data set includes the information of 525 patients who were referred to the University Hospitals Leuven, Belgium, between 1994 and 1999 [1]. Patients without preoperative results of serum CA 125 levels are excluded from this analysis. Among the available 425 cases, 291 patients had benign tumors, whereas 134 ones had malignant tumors. After preprocessing, e.g. CA 125 serum level was rescaled by taking its logarithm, the data set contains 27 variables. Table 1 lists the most important variables that were considered. Shown in Fig. 1 is 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. The data set is split according to the time scale: the data from the first treated 265 patients are taken as training set, 160 of the remaining data are used as test set. The proportion of malignant tumors in the training set and test set are both about $1/3$. Thanks to the Bayesian methods implemented here, no validation set is needed during training.

3.2 Selecting predictive input variables

Selecting the most predictive input variables is critical to effective model development, it can not only assist the understanding of the disease, but also potentially decrease the measurement cost for the future. Here we adapt the forward selection which tries to maximize the evidence of the LS-SVM classifiers with either linear or RBF kernels. In order to sta-

Table 1: Demographic, serum marker, color Doppler imaging and morphologic variables

	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	Normal blood flow (Colsc3)	15.8 %	35.8 %
	Strong blood flow (Colsc4)	4.5 %	20.3 %
Morphologic	Abdominal fluid (Asc)	32.7 %	67.3 %
	Bilateral mass (Bilat)	13.3 %	39.1 %
	Solid tumor (Sol)	8.3 %	37.6 %
	Irregular wall (Irreg)	33.8 %	73.2 %
	Papillations (Pap)	13.0 %	53.2 %
	Acoustic shadows(Shadows)	12.2 %	5.7 %

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

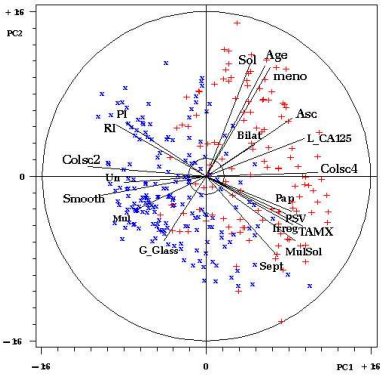


Figure 1: Biplot of ovarian tumor data('×'- benign, '+'- malignant), projected on the first two PCs.

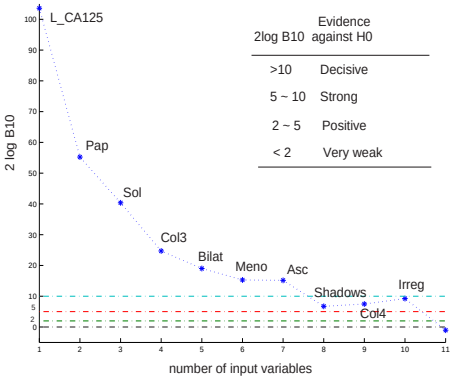


Figure 2: Evolution of the model evidence during the forward input selection for LS-SVM with RBF kernels.

bilize the selection, the three variables with the smallest univariate model evidence are first removed. Then the selection starts from the remaining 24 candidate variables. Fig. 2 shows the evolution of the model evidence during the input selection using RBF kernels. The Bayes factor for the univariate model was derived 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. Ten variables were selected by the LS-SVM with RBF kernels. Linear kernels have also been tried but result into a smaller evidence and an inferior model performance.

Compared to the variables selected by a stepwise logistic regression based on the whole dataset (which should be optimistic) [3], the new identified subset based only on the 265 training data includes only 2 more variables, however still gives a comparable performance on the test set.

3.3 Model fitting, prediction and interpretation

The model fitting procedure has two stages, the first is the construction of the standard LS-SVM classifier within the evidence framework. Sparseness can be imposed to LS-SVM at this stage, by iteratively pruning, e.g. the 'easy' cases which have negative support values.

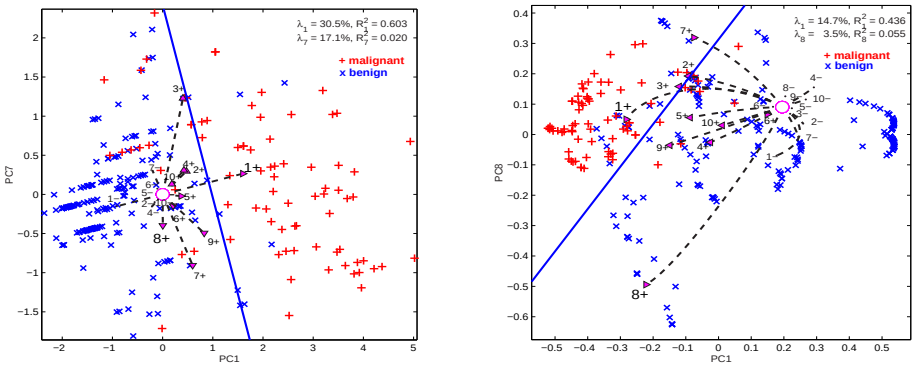


Figure 3: Interpretation of LS-SVM with linear (left) and RBF kernel (right) by kPCA.

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.

In risk minimization decision making, different error costs are considered in order to reduce the expected loss. Since misclassification of a malignant tumor is very serious, the adjusted prior for the malignant class in the following experiments is intuitively set to $2/3$, higher than that of the benign class $1/3$.

In order to interpret the LS-SVM classifier, the 10 dimensional training data are projected to pairs of PCs in the feature space, with linear and RBF kernel respectively, as depicted in Fig. 3. The approximated decision boundary (solid line) using the ith and jth PCs is equal to $\tilde{w}(i)z_i + \tilde{w}(j)z_j + \tilde{b} = 0$. The dashed line shows the trajectories of the variables, concurrent at the center of the data (denoted as a circle); the pseudosample is bounded by the minimum and the maximum value of the corresponding variable in the training set (the positive end of the trajectory is indicated by a triangle). Interpretation can be done through observing the class distributions, the length of trajectories, and the correlations between the trajectories and classes. In the biplot for LS-SVM with RBF kernels, e.g., variable 1 (L_CA125) exhibits to be the most discriminative, while variable 8 (Shadows) seems contribute little to the distinction as it is almost parallel to the decision boundary.

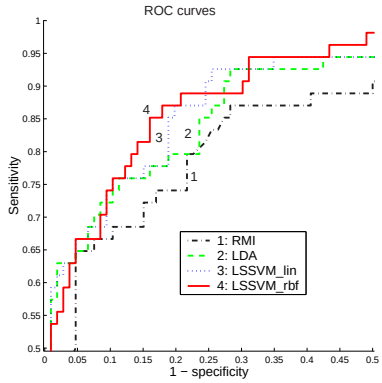
3.4 Model evaluation

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. ROC curve is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) for the different cutoff of a diagnostic test. Here the *sensitivity* and *specificity* is the correct classification rate for 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 [9].

Fig. 4 reports the performance of different models on the test set. The compared models include linear discriminant (LDA), and LS-SVM with linear and RBF kernels; the performance of RMI, a widely used score system, is listed as a reference. The LS-SVM with RBF kernels achieves the best performance.

4 Conclusion

In this paper, we have discussed the application of Bayesian LS-SVM classifiers to predict the malignancy of the ovarian tumors. Within the evidence framework, the hyperparameter tuning, input selection and computation of posterior class probability for risk minimization



Model Type (AUC)	Cutoff value	Accuracy (%)	Sensitivity (%)	Specificity (%)
RMI	100	78.13	74.07	80.19
(0.8733)	**75**	**76.88**	**81.48**	**74.53**
LDA	0.5	84.38	75.93	88.68
(0.9034)	0.4	83.13	75.93	86.79
	0.3	**81.87**	**77.78**	**83.96**
LS-SVM$_{\text{Lin}}$	0.5	82.50	77.78	84.91
(0.9141)	0.4	81.25	77.78	83.02
	0.3	**81.88**	**83.33**	**81.13**
LS-SVM$_{\text{RBF}}$	0.5	84.38	77.78	87.74
(0.9184)	0.4	83.13	81.48	83.96
	0.3	**84.38**	**85.19**	**83.96**

Figure 4: Comparison of the model performance on the test set via ROC analysis.

decision making can be conducted in a unified way. Our results demonstrate that the LS-SVM 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. This work is one part of the International Ovarian Tumour Analysis (IOTA) project, which is a multi-center study on the preoperative characterization of ovarian tumors based on artificial intelligence models [2]. Future work will apply our method to the multi-center data in a larger scale, and possibly further subclassify the tumors.

Acknowledgments

This paper presents research results of the Belgian Programme on Interuniversity Poles of Attraction (IUAP V-10-29), of the Concerted Research Action (GOA) projects of the Flemish Government MEFISTO-666, of the IDO/99/03 project (K.U.Leuven) and of the FWO (Fund for Scientific Research Flanders) project G.0407.02. JS is a postdoctoral researcher with the National Fund for Scientific Research FWO - Flanders.

References

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

[2] D. Timmerman, et al. Measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group, *Ultrasound Obstet Gynecol* (2000) 16:500-505.

[3] C. Lu, J. De Brabanter, S. Van Huffel, I. Vergote, D. Timmerman, Using artificial neural networks to predict malignancy of ovarian tumors, in *Proc. of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (Istanbul, Turkey, October 25-28, 2001), paper 4.2.2-6, 4 pp.

[4] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural Processing Letters* (1999) 9(3):293-300.

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

[6] D.J.C. MacKay, The evidence framework applied to classification networks, *Neural Computation* (1992) 4(5): 698-741.

[7] T. Van Gestel, J.A.K. Suykens, et al. A Bayesian framework for least squares support vector machine classifiers, *Neural Computation* (2002) 15(5):1115-1148.

[8] T. Van Gestel, *From linear to Kernel Based Methods in Classification, Modelling and Prediction*, PhD thesis, K.U.Leuven (Leuven, Belgium, 2002).

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

[10] H. Jeffreys, *Theory of Probability* (Oxford University Press, New York, USA, 1961).

[11] J.C. Gower and D.J. Hand, *Biplots* (Chapman & Hall, London, 1996).