Classification of Ovarian Tumors Using Bayesian Least Squares Support Vector Machines

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

[1] Dept. of Electrical Engineering, Katholieke Universiteit Leuven, 3001 Leuven, Belgium
{chuan.lu, tony.vangestel, Johan.Suykens, Sabine.VanHuffel}@esat.kuleuven.ac.be
[2] Dept. of Obstetrics and Gynecology, University Hospitals Leuven, 3000 Leuven, Belgium
{dirk.timmerman,ignace.vergote}@uz.kuleuven.ac.be

Abstract. The aim of this study is to develop the Bayesian Least Squares Support Vector Machine (LS-SVM) classifiers for preoperative discrimination between benign and malignant ovarian tumors. We describe how to perform (hyper)parameter estimation, input variable selection for LS-SVMs within the evidence framework. The issue of computing the posterior class probability for risk minimization decision making is addressed. The performance of the LS-SVM models with linear and RBF kernels has been evaluated and compared with Bayesian multi-layer perceptrons (MLPs) and linear discriminant analysis.

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 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][2][3]. In this paper, we focus on the development of Bayesian Least Squares Support Vector Machines (LS-SVMs), to preoperatively predict the malignancy of ovarian tumors.

Support Vector Machines (SVMs) [5] have become a state-of-the-art technique for pattern recognition. The basic idea of the nonlinear SVM classifier and related kernel techniques 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. These kernel-based algorithms have attractive features such as good generalization performance, the

existence of a unique solution, and strong theoretical background, i.e., statistical learning theory [5], supporting their good empirical results. Here a least squares version of SVM [6][7] 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 Gaussian processes and kernel Fisher discriminant analysis [10].

The need of applying Bayesian methods to LS-SVMs for this task is twofold. One is to tune the regularization and possible kernel parameters automatically to their near-optimal values, second is to judge the uncertainty in predictions that is critical in a medical environment. A unified theoretical treatment of learning in feedforward neural networks has been provided by MacKay's Bayesian evidence framework [9][8]. Recently this Bayesian framework was also applied to LS-SVMs, and a numerical implementation was derived. This approach has been applied to several benchmark problems, achieving similar test set results as Gaussian processes and SVMs [10].

After a brief review of the LS-SVM classifier and the Bayesian evidence framework, we will show the scheme for input variable selection and the way to compute the posterior class probabilities for minimum risk decision making. The test set performance of models are assessed via Receiver Operating Characteristic (ROC) curve analysis.

2 Data

The data set includes the information of 525 patients who were referred to a single ultrasonographer 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 benign and malignant adnexal masses [1]. Patients without preoperative results of serum CA 125 levels are excluded from this analysis. The gold standard for discrimination of the tumors were the results of histological examination. Among the available 425 cases, 291 patients had benign tumors, whereas 134 ones had malignant tumors.

The measurements and observations 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 [1][4]. 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 most important variables that were considered. Fig. 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.

Table 1. Descriptive statistics of ovarian tumor data

	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±SD in case of a benign and malignant tumor respectively are reported; for binary variables, the occurrences (%) of the corresponding features are reported.

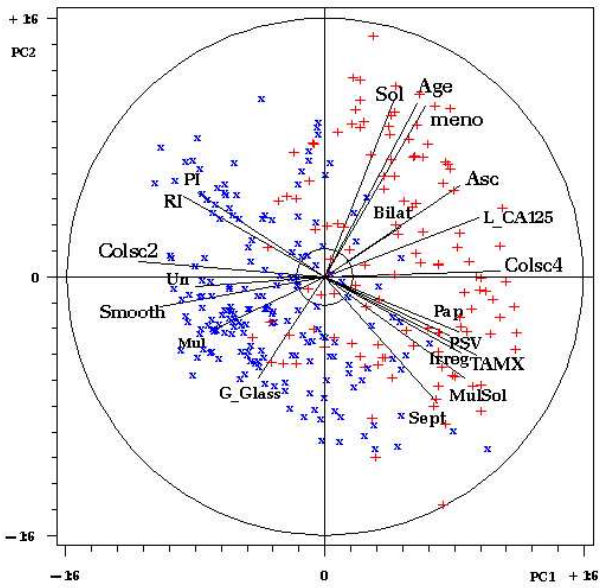


Fig. 1. Biplot of ovarian tumor data ('×'- benign, '+'- malignant), projected on the first two principal components

3 Methods

3.1 Least Squares SVMs for Classification

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 $y_i = \pm 1$ (+1: malignant, −1: benign) by mini-

mizing the following cost function:

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

subject to the equality constraints $y_i[w^T \varphi(x_i) + b] = 1 - e_i, i = 1, ..., N$. 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 [6][7]:

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

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 $K(x_i, x_j) = x_i^T x_j$ and an RBF kernel $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2 / \sigma^2)$. The LS-SVM classifier is then constructed in the dual space as:

$$y(x) = \operatorname{sign}\left[\sum_{i=1}^{N} \alpha_i \, y_i \, K(x, x_i) + b\right].$$

3.2 Bayesian Inference

In [10] 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 the computational details, the interested readers are referred to [10] and [7].

The Bayesian evidence approach first finds the maximum a posteriori estimates of model parameters w_{MP} and b_{MP}, using conventional LS-SVM training methods, i.e. by solving the linear set of equations in (2) in the dual space in order to optimize (1). Then the distribution over the parameters is approximated using information available at this maximum. The hyperparameters μ and ζ are determined by maximizing the posterior probability of the parameters, which can be estimated using the Gaussian probability at $w_{\mathrm{MP}}, b_{\mathrm{MP}}$.

Different models can be compared by examining their posterior $p(\mathcal{H}_j|D)$. Assuming a uniform prior $p(\mathcal{H}_j)$ over all models, the models can be ranked by the model evidence $p(D|\mathcal{H}_j)$, which can be again evaluated using a Gaussian approximation. The kernel parameters, e.g. the bandwidth parameter σ of the RBF kernel, are chosen from a set of candidates by maximizing the model evidence.

3.3 Model Comparison and Input Variable Selection

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

D is defined as $B_{10} = p(D|\mathcal{H}_1)/p(D|\mathcal{H}_0)$. 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 also shown in Fig. 2 [12].

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 heuristic search strategy for variable selection can be e.g. backward elimination, forward selection, stepwise selection, etc. Here we concentrate on the forward selection (greedy search) method.

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.

3.4 Computing Posterior Class Probability

For a given test case, the conditional class probabilities $p(x|y = \pm 1, D, \mu, \zeta, \mathcal{H})$ can be computed using the two normal probability densities of $w^T\varphi(x)$ for two classes at the most probable value $w_{\mathrm{MP}}^T\varphi(x)$ [10][7]. The mean of each distribution is defined as the class center of the output (in the training set), and the variance comes from both the target noise and the uncertainty in the parameter w. By applying Bayes' rule the posterior class probabilities of the LS-SVM classifier can be obtained:

$$p(y|x, D, \mu, \zeta, \mathcal{H}) = \frac{p(y)p(x|y, D, \mu, \zeta, \mathcal{H})}{\sum_{y'=\pm 1} p(y')p(x|y', D, \mu, \zeta, \mathcal{H})}, \tag{3}$$

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 (3) with the adjusted class prior, e.g. $P'(y{=}1) = P(y{=}1)c_+^-/(P(y{=}1)c_+^- + P(y{=}-1)c_-^+)$.

4 Experiments and Results

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. The proportion of malignant tumors in the training set and test set are both about 1/3. 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 proposed forward procedures. Besides LS-SVM models with linear and RBF kernels, the other considered competitive models include a linear discriminant analysis (LDA) classifier, and a Bayesian MLP classifier as the counterpart of SVMs in neural network modelling.

4.1 Selecting Predictive Input Variables

Selecting the most predictive input variables is critical to effective model development, since it not only helps to understand the disease, but also potentially decreases 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 stabilize 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 variable added to the model at each selection step and the corresponding Bayes factor have been depicted. 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. Ten variables were selected by the LS-SVM with RBF kernels, and they were used to build all the competitive models in the following experiments. Linear kernels have also been tried, but resulted into a smaller evidence and an inferior model performance.

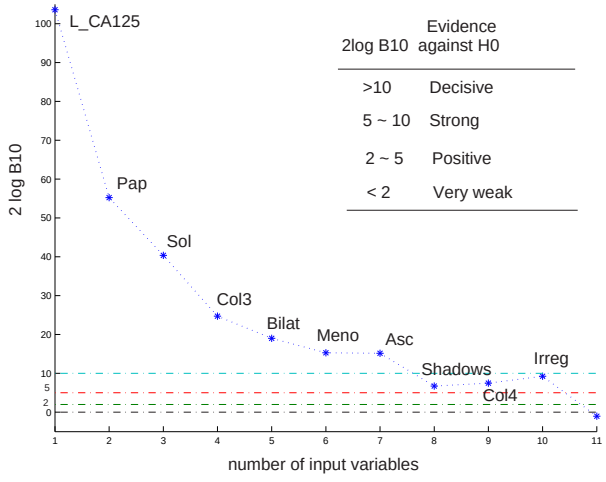


Fig. 2. Evolution of the model evidence during the forward input selection for LS-SVM with RBF kernels

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

4.2 Model Fitting and Prediction

The model fitting procedure for 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 α. 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.

For MLP models, we use MacKay's Bayesian MLP classifier [9][8], which is limited to one hidden layer with two hidden neurons, with hyperbolic tangent activation function for the hidden layer, and sigmoidal 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.

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 $1/3$. The same adjusted class priors have been combined for the computation of the posterior output for all the compared models.

4.3 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. 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 [11].

Fig. 3 and Table 2 report the performance of different models on the test set, the performance of 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 listed as a reference. All the competitive models perform much better than RMI, among which the LS-SVM model with RBF kernels achieves the best performance. The performance of Bayesian MLP is comparable to Bayesian LS-SVM with RBF kernels.

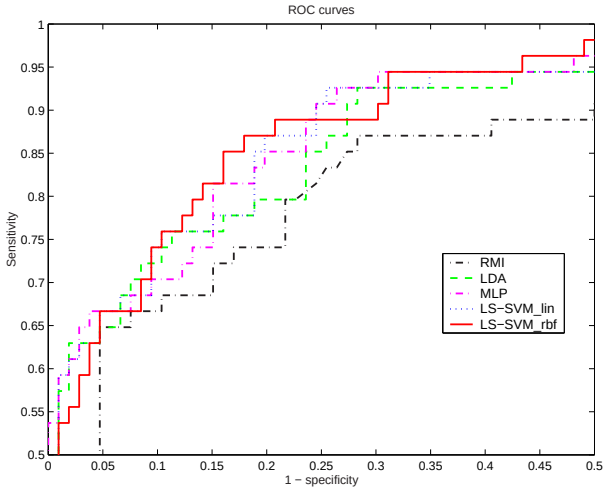


Fig. 3. ROC curves from different models on the test set

Table 2. Comparison of the model performance on the test set

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**
MLP	0.5	82.50	77.78	84.91
(0.9174)	0.4	83.13	81.48	83.96
	0.3	**81.87**	**83.33**	**81.13**
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**

We also try to check the ability of our classifiers in rejecting uncertain test cases which need further examination by a human expert. The discrepancy 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 reevaluated after rejecting a certain number of the most 'uncertain' test cases, and RBF LS-SVM model keeps giving the best

results. Table 3 shows how the rejection of the uncertain cases can improve the performance of the RBF LS-SVM classifier.

Table 3. Classification performance with rejection

Model Type	Reject	AUC	Acc(%)	Sens(%)	Spec(%)
LS-SVM$_{RBF}$	5%(8/160)	0.9343	87.50	82.61	89.80
	10%(16/160)	0.9420	88.97	83.72	91.40

Note: the cutoff probability level is set to 0.5.

5 Conclusions

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 variable selection and computation of posterior class probability 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 part of the International Ovarian Tumor Analysis (IOTA) project, which is a multi-center study on the preoperative characterization of ovarian tumors based on artificial intelligence models [4]. Future work include the application of our models to the multi-center data on a larger scale, and possibly further subclassify the tumors.

Acknowledgments

S. Van Huffel is a full professor with KU Leuven, Belgium. J.A.K. Suykens is a postdoctoral researcher with FWO Flanders and a professor with KU Leuven. T. Van Gestel is a postdoctoral researcher with FWO Flanders. 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 and IDO/99/03, and the FWO projects G.0407.02 and G.0269.02.

References

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

2. C. Lu, J. De Brabanter, S. Van Huffel, I. Vergote, D. Timmerman, "Using artificial neural networks to predict malignancy of ovarian tumors," in *Proc. 23rd Annu. Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, Istanbul, Turkey, October 25-28, 2001, Paper 4.2.2-6.
3. P. Antal, H. Verrelst, D. Timmerman, Y. Moreau, S. Van Huffel, B. De Moor, I. Vergote, "Bayesian networks in ovarian cancer diagnosis: potentials and limitations," in *Proceeding of the 13th IEEE Symposium on Computer-Based Medical Systems (CBMS 2000)*, Houston, TX, 2000, 103-109.
4. D. Timmerman, L. Valentin, T.H. Bourne, W.P. Collins, H. Verrelst, I. Vergote, "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, 2000.
5. V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
6. J.A.K. Suykens, J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293-300, 1999.
7. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*. Singapore: World Scientific, 2002.
8. C.M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, 1995).
9. D.J.C. MacKay, "The evidence framework applied to classification networks," *Neural Computation*, vol. 4, no. 5, pp. 698-741, 1992.
10. T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, J. Vandewalle, "A Bayesian framework for Least Squares Support Vector Machine classifiers, Gaussian processes and kernel Fisher discriminant analysis," *Neural Computation*, vo. 15, no.5, pp. 1115-1148, 2002.
11. J.A. Hanley, B. McNeil, "The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve," *Radiology*, vol. 143, pp. 29-36, 1982.
12. H. Jeffreys, *Theory of Probability*. New York: Oxford University Press, 1961.