ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Multi-class classification of ovarian tumors

Ben Van Calster[1], Dirk Timmerman[2], Antonia C. Testa[3],
Lil Valentin[4], and Sabine Van Huffel[1] *

1- Katholieke Universiteit Leuven - Dept of Electrical Engineering (ESAT-SISTA)
Kasteelpark Arenberg 10, B-3001 Leuven - Belgium

2- University Hospitals Katholieke Universiteit Leuven - Belgium

3- Malmö University Hospital - Sweden

4- Università Cattolica del Sacro Cuore Rome - Italy

Abstract. In this work, we developed classifiers to distinguish between four ovarian tumor types using Bayesian least squares support vector machines (LS-SVMs) and kernel logistic regression. Input selection using rank-one updates for LS-SVMs performed better than automatic relevance determination. Evaluation on an independent test set showed good performance of the classifiers to distinguish between all groups, even though borderline and metastatic tumors were expected to be hard to identify.

1 Introduction

Optimal treatment of ovarian tumors depends on the type of tumor. Adequate treatment decisions have beneficial consequences for the patient, such as improved prognosis. Therefore, an accurate preoperative diagnosis of ovarian tumors is needed. Hitherto, models aimed to distinguish between benign and malignant tumors [1, 2]. However, there are different types of malignant tumors. In this paper, we developed classifiers to distinguish between benign and three types of malignant tumors (primary invasive, borderline, and metastatic) using Bayesian least squares support vector machines and kernel logistic regression. The models output probabilities, which is important because diagnostic uncertainty influences treatment decisions. With respect to input selection, we used a rank-one update method and compared its performance with automatic relevance determination.

This paper is organized as follows. Section 2 describes the used algorithms. Section 3 describes the data set and presents the results. The last section concludes the paper.

2 Classification and input selection algorithms

We used three algorithms for multi-class classification. Using pairwise coupling, we combined 1-versus-1 Bayesian least squares support vector machines (LSSVM-PC) or 1-versus-1 kernel logistic regression models (KLR-PC). Multiclass kernel logistic regression (MKLR) was used as an all-at-once algorithm.

*We thank Fabian Ojeda for constructive discussions. Research supported by GOA-AMBioRICS, CoE EF/05/006, FWO G.0407.02 and G.0302.07, IUAP P6/04, BIOPATTERN (FP6-2002-IST 508803), eTUMOUR (FP6-2002-LIFESCIHEALTH 503094).

Data samples consist of an input vector $\mathbf{x}_n$ ($n = 1, \ldots, N$), and an outcome indicating to which class the sample belongs. The output of all algorithms consists of the 4 posterior class probabilities given $\mathbf{x}$.

2.1 Least squares support vector machines

Least squares support vector machine (LS-SVM) classifiers [3] are obtained by solving a linear set of equations, whereas standard SVMs represent quadratic programming problem. The LS-SVMs classifier in the primal space, $y(\mathbf{x}) = \text{sign}[\mathbf{w}^T\varphi(\mathbf{x}) + b]$, is obtained by solving

$$\min_{\mathbf{w},b,\mathbf{e}} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{2}\gamma \sum_{n=1}^{N} e_n^2 \right), \text{ such that } y_n\Big[\mathbf{w}^T\varphi(\mathbf{x}_n) + b\Big] = 1 - e_n \,,$$

where y_n is the binary class indicator (encoded as -1 vs $+1$), $\mathbf{w}$ the parameter vector, b the bias term, e_n the error variable, and γ the regularization hyperparameter. The mapping $\varphi : \mathbb{R}^q \to \mathbb{R}^r$ maps the input space into a high-dimensional feature space.

By taking the Lagrangian, the classifier can be reformulated in the dual space as $y(x) = \text{sign}\big[\sum_{n=1}^{N} \alpha_n y_n K(\mathbf{x}, \mathbf{x}_n) + b\big]$, where $\alpha_1, \ldots, \alpha_N$ are the support values. The positive definite kernel function $K(\mathbf{x}, \mathbf{z}) = \varphi(\mathbf{x})^T\varphi(\mathbf{z})$ allows us to work in the feature space without explicitly constructing it. Popular kernels are the linear kernel to build classifiers with a linear decision boundary in the input space, and the nonlinear radial basis function (RBF) kernel. The RBF kernel has parameter σ to denote the kernel width. For standard LS-SVMs, typically no support values are zero, such that the sparseness property of SVMs is lost. Due to the 2-norm in the LS-SVM cost function, α_n equals γe_n, resulting in negative support values for easy training samples. In this work, we were able to impose sparseness by repeatedly deleting samples with a negative support value.

To obtain class probabilities, we formulated the LS-SVM model in a Bayesian framework [4]. An advantage is that both γ and the kernel parameters (if any) are tuned automatically.

2.2 Kernel logistic regression

Kernel logistic regression (KLR) is a method that is closely related to SVMs. However, KLR directly results in probabilistic output, and can be readily extended to multi-class KLR (MKLR). Recently, an LS-SVM framework for solving MKLR was derived [5]. This method starts from a regularized version of MLR that is solved by optimizing the penalized negative log likelihood function using iteratively regularized re-weighted least squares. By mapping the input space into a high-dimensional feature space using a positive definite kernel, and by applying a model with the structure of an LS-SVM in each iteration, a kernel version of MLR is obtained using an iteratively re-weighted LS-SVM method. We tuned hyperparameters using five-fold cross-validation (5CV).

2.3 Pairwise coupling

Pairwise coupling (PC) combines 1-versus-1 probabilities into multi-class probabilities. In the method adopted in this paper [6], one solves

$$\pi_k = \sum_{l=1,k\neq l}^{K} \frac{\pi_k + \pi_l}{K-1} p_{kl}, \ \forall k, \ \text{with} \ \sum_{k=1}^{K} \pi_k = 1, \ \pi_k \geq 0\,,$$

where $\pi_1, \ \ldots, \ \pi_K$ are the multi-class probabilities to be estimated, and p_{kl} is the probability obtained by applying the k-versus-l binary classifier.

2.4 Input selection

2.4.1 Automatic relevance determination (ARD)

For Bayesian LS-SVMs, ARD is implemented by inserting a diagonal weight matrix in the kernel function [7]. The weights are optimized in the Bayesian model. The initial choice for is taken to be $[1, \ ..., \ 1]/\sigma$, with σ being the optimal RBF kernel parameter for the unweighted Bayesian model. Inputs with low weights can be dropped from further analysis.

2.4.2 Rank-one updates for LS-SVMs (R1U)

Selecting inputs by repeatedly adding the input that gives the best gain in performance based on leave-one-out cross-validation (LOO-CV) is, in its original form, computationally intensive. When using LS-SVMs, this strategy can be speeded up for two reasons. First, the LS-SVM model structure allows for the fast computation of model performance measures based on LOO-CV [8]. Second, the LS-SVM model can be updated using rank-one adjustments in the kernel matrix such that adding an input does not require the re-computation of the model [9]. This method is very fast for performing greedy input selection (i.e. forward or backward), but is only available for linear kernel LS-SVMs.

3 Results

3.1 Data

We used the data set from Phase I of the International Ovarian Tumor Analysis (IOTA) group, containing data from 1066 women collected at nine centers across Europe. There were 75% benign, 16% primary invasive, 5% borderline, and 4% metastatic tumors. The training set consisted of 754 women (71%), the test set of the remaining 312. This split was stratified for tumor type and center. Data collected involved demographic information, cancer history, pain during examination, and a large amount of morphologic and blood flow measurements describing the tumor. Submission of complete data was obligatory for all measurements.

3.2 Step one: comparison of input selection methods

Using the training data, we applied three input selection methods to all six 1-versus-1 classification problems: ARD with linear kernel (ARD-lin), ARD with RBF kernel (ARD-RBF), and R1U. Both ARD methods were applied to 36 initial inputs, and the 16 with lowest weight were deleted. In the second and third ARD run, the worst 8 and 4 inputs were dropped, respectively. The final run ranked the remaining 8 inputs. Because dropping inputs may influence the input ranking, this approach was preferred over a single ARD run. For R1U, the LOO-AUC (LOO area under the ROC curve) was used in each step to determine which input to add. The value of γ was re-tuned in each step by selecting the value with optimal LOO-AUC performance.

The results of the three methods were compared as follows. Per binary problem, we considered 12 Bayesian LS-SVM models. We considered a Bayesian LS-SVM with linear kernel using inputs from either ARD-lin, ARD-RBF, or R1U, and a Bayesian LS-SVM with RBF kernel using inputs from ARD-RBF. For each of these, we used the 3, 5, and 8 best inputs as determined by the input selection method, thus resulting in 12 models overall. The performance of these models was determined using 20 runs of 5CV on the training set, yielding 100 validation AUCs per model that were summarized by their mean. We combined, by averaging, the results for all binary problems into one summarizing plot (Figure 1a). The R1U method gave the best results.

3.3 Step two: determine final input sets for each binary problem

We need good input sets for all binary problems while keeping the overall number of inputs small. Therefore, we selected 16 promising inputs based on the R1U ranking, and ran R1U again to rank these 16. Then, 20 runs of 5CV were used to determine how many of the most important units were to be chosen. This extensive analysis resulted in final input sets for each binary problem (Table 1). Three to five inputs were selected per binary problem, 11 in total.

1-versus-1 problem	Inputs
Ben vs Inv	ascites, max. diam. of solid component (maxsol), irregular internal cyst wall (irreg), age, acoustic shadows (5 inputs)
Ben vs Bor	age, history of ovarian cancer (historyOC), max. diam. of lesion, number of papillary projections, blood flow in pap. projections (5)
Ben vs Met	ascites, maxsol, irreg, entirely solid tumor (solid) (4)
Inv vs Bor	ascites, maxsol, irreg, age, bilateral tumors (5)
Inv vs Met	ascites, solid, historyOC (3)
Bor vs Met	ascites, maxsol, solid (3)

Ben = Benign, Inv = Primary invasive, Bor = Borderline, Met = Metastatic

Table 1: Selected inputs for each 1-versus-1 problem.

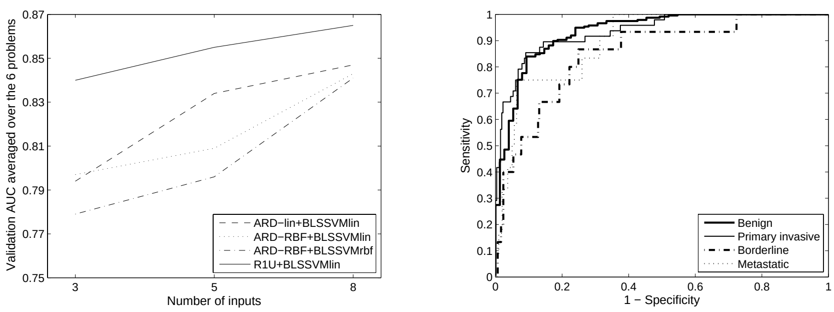
3.4 Step three: Model training and evaluation

As mentioned earlier, we focus on three algorithms to construct classifiers: LSSVM-PC, KLR-PC, and MKLR. The linear kernel was used for these models. The comparison of the different input selection methods suggested that purely linear approaches may perform best (Figure 1a). The MKLR model is an 'all-at-once' method (as opposed to a combination of 1-versus-1 models). Therefore, we used all 11 inputs from Table 1 for the MKLR classifier. Both other models, on the contrary, used these inputs only where they were deemed necessary (cf. Table 1).

Table 2 presents the test set performance of the classifiers by showing the test set AUC and the estimated probability that a classifier is the best of the three (P_1). This probability is obtained by drawing 1000 bootstrap data sets from the test set, and computing the proportion of data sets for which the classifier had the best AUC. All classifiers had good performance. Benign and primary invasive tumors were identified most accurately, but also metastatic and borderline tumors were identified very well. KLR-PC performs well for identifying borderline tumors, MKLR for metastatic tumors. The ROC curves for the KLR-PC classifier are shown in Figure 1b.

Model	Benign	Primary invasive	Borderline	Metastatic
LSSVM-PC	0.934 − 0.16	0.938 − 0.55	0.838 − 0.08	0.906 − 0.23
KLR-PC	0.934 − 0.25	0.936 − 0.17	0.850 − 0.87	0.896 − 0.01
MKLR	0.936 − 0.59	0.933 − 0.28	0.826 − 0.05	0.915 − 0.76

Table 2: Test set performance (AUC and P_1).



(a) (b)

Fig. 1: Comparison of ARD and R1U (a) and test set ROC curves for KLR-PC (b).

4 Discussion and conclusions

In the present work, classifiers were constructed that were able to accurately distinguish between four ovarian tumor types. Most notably, the good performance of the models for borderline and metastatic tumors is promising. These groups are small, and were expected to be difficult to separate from other groups. The classifiers were based on rank-one update input selection, a fast method to perform forward or backward input selection for LS-SVM classifiers using LOO-CV. An interesting aspect of the method is that it can easily use AUC information to select inputs. This is important if one wishes to construct classifiers with optimal discriminatory power. At least for our application, this method performed better than automatic relevance determination. A drawback is that the rank-one update method is only suited for linear kernels.

A few important analyses are not reported due to lack of space, or are still due. First, it is essential to investigate whether the predicted probabilities are calibrated. Second, comparing the algorithms rather than the classifiers using repeated training-test splits is useful too, even though clinicians have more interest in classifiers. Third, since we aim to distribute good classifiers to clinicians for use in gynaecological practice, it is necessary to evaluate the classifiers prospectively in the same and in other centers.

References

[1] C. Van Holsbeke, B. Van Calster, L. Valentin, et al., External validation of mathematical models to distinguish between benign and malignant adnexal tumors, *Clinical Cancer Research*, 13:4440-4447, 2007.

[2] B. Van Calster, D. Timmerman, C. Lu, et al., Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods, *Ultrasound in Obstetrics and Gynecology*, 29:496-504, 2007.

[3] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, et al., *Least squares support vector machines*, World Scientific, Singapore, 2002.

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

[5] P. Karsmakers, K. Pelckmans and J.A.K. Suykens, Multi-class kernel logistic regression: a fixed size implementation. In *Proceedings of the* 20^{th} *International Joint Conference on Neural Networks*, pages 1756-1761, Orlando (USA), 2007.

[6] T.-F. Wu, C.-J. Lin and R.C. Wen, Probability estimates for multi-class classification by pairwise coupling, *Journal of Machine Learning Research*, 5:975-1005, 2004.

[7] T. Van Gestel, J.A.K. Suykens, B. De Moor, et al., Automatic relevance determination for least squares support vector machines classifiers. In *Proceedings of the* 9^{th} *European Symposium on Artificial Neural Networks*, pages 13-18, Bruges (Belgium), 2001.

[8] G.C. Cawley, Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proceedings of the* 19^{th} *International Joint Conference on Neural Networks*, pages 2970-2977, Vancouver (Canada), 2006.

[9] F. Ojeda, J.A.K. Suykens and B. De Moor, Variable selection by rank-one updates for least squares support vector machines. In *Proceedings of the* 20^{th} *International Joint Conference on Neural Networks*, pages 2283-2288, Orlando (USA), 2007.