Ultrasound Obstet Gynecol 2006; **27**: 607–608
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI:** 10.1002/uog.2791

Opinion

Support vector machines versus logistic regression: improving prospective performance in clinical decision-making

In this issue of the Journal De Smet *et al.*[1] describe the use of new models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography. This paper uses standard logistic regression models and compares them with the more advanced least squares support vector machine (LS-SVM) models[2] with linear and radial basis function (RBF) kernels. While classical logistic regression analysis is the standard method used for clinical classification problems[3–5], we believe that this new and more advanced method might still improve performance.

Traditional statistical methods such as multivariate logistic regression intend to build a classification model that fits a set of patients ('training' set) optimally. Unfortunately, this strategy may easily result in a model that fits these training patients too well and is therefore not capable of making good predictions for previously unknown patients ('independent', 'prospective' or 'test' set). This problem is often referred to as overfitting the training patients, and leads to poor generalization to previously unknown patients. Support vector machines (SVMs) are a relatively new method based on the principle of statistical learning theory[6] to solve classification and regression problems. This method tries to learn and generalize well when building a model using a given set of patients. This way, SVMs perform reasonably well on a training set, but not at the expense of performance when making predictions for previously unseen patients.

Logistic regression tries to fit a model as well as possible on the patients of the training set. Even with samples that do not follow the general underlying distribution in the case of outliers, logistic regression fits the training set too well, leading to a substantial number of misclassified patients when applied prospectively. SVMs try to generalize well when building a model using the given set of patients. With SVMs, optimization of the generalization performance is achieved by controlling two terms, i.e. by minimizing the classification error on the training set together with minimizing the complexity of the model. This trade-off is represented by a regularization parameter (γ) in the LS-SVM formulation.

A further disadvantage of logistic regression is that the technique is not able to identify possible non-linear structures in a set of patients. When nonlinear relationships exist, a nonlinear decision boundary may result in a better performance overall. Unlike logistic regression, SVMs are designed to generate more complex decision boundaries. An LS-SVM with a simple linear kernel function corresponds to a linear decision boundary. Instead of a linear kernel, more complex kernel functions, such as the commonly used RBF kernel, can be chosen. An RBF kernel requires optimization of the kernel parameter (σ), which controls the curvature of the decision boundary. Figure 1 shows an example in which using an SVM with an RBF kernel would be more appropriate than would using an LS-SVM with a simple linear kernel. With this more complex decision boundary, the nonlinearity in this set of patients could be better described than would be possible with a linear decision boundary.

LS-SVMs are reformulations of the standard SVMs and qualitatively they are similar. Already they have been used extensively for various classification problems, including medical ones[7]. LS-SVM models can be trained easily using LS-SVMlab vers. 1.5[5,8] for MATLAB, as shown by De Smet *et al.*[1]. These authors aimed to predict the depth of infiltration in endometrial carcinoma based on transvaginal sonography data from 97 training-set patients that included the ultrasound parameters, number of fibroids detected during ultrasound examination, the degree of differentiation of the cancer, the presence of a clear cell component, and the presence of a serous

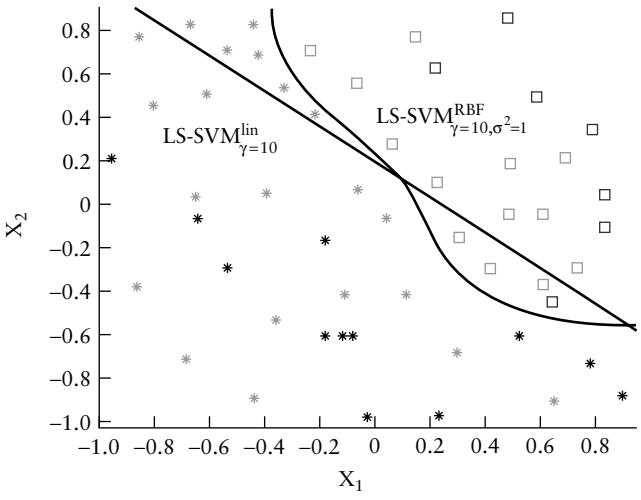

Figure 1 Using least squares support vector machine models (LS-SVMs) with a more complex decision boundary (represented by a radial basis function (RBF) kernel) is more appropriate compared with using LS-SVMs with a simple linear decision boundary (using a linear kernel) in cases of non-linear structures in a set of patients. Training samples are represented in black, and prospective samples in gray. Class 1(∗) and Class 2(□) represent, for example, a set of diseased and a set of disease-free patients respectively. The information known for each of these patients is represented by the two variables X_1 and X_2, which constitute the axes of this plot.

papillary component. Stepwise logistic regression selected the degree of differentiation, the number of fibroids and the endometrial thickness and volume. Subsequently, these variables were used to train a logistic regression model and LS-SVM with linear and RBF kernels. Compared with the area under the receiver–operating characteristics curve (AUC) of the subjective assessment (72%), prospective evaluation of the mathematical models on 76 test-set patients resulted in an equally good AUC for the LS-SVM model with a linear kernel, and in a better AUC (77%) for the LS-SVM model with an RBF kernel, although this difference was not significant. The performance of the standard logistic regression model (66%) was significantly worse, although the training set performance was similar to that of the LS-SVM model with a linear kernel. This shows that the level of overfitting for the standard logistic regression model was higher than was that for the LS-SVM model with a linear kernel (the logistic regression model will also reflect some accidental characteristics of the training set that would not reoccur in an independent set of patients).

When training LS-SVM models with a linear and an RBF kernel, it is necessary to optimize the regularization parameter γ. Using an RBF kernel also requires tuning of the kernel parameter σ. These parameter(s) can be tuned in LS-SVMlab with the 'tunelssvm' function using a 'linesearch' approach for the LS-SVM with a linear kernel (tuning of γ) and a 'gridsearch' approach for the LS-SVM with an RBF kernel (tuning of σ and γ) when optimizing the 'leave-one-out cross-validation' performance on the training set. These parameter settings can be used subsequently when training the definitive model with the 'trainlssvm' function. The 'simlssvm' function allows making predictions for new patients using the previously built model. Since regularization is performed in LS-SVM models, the generalization of this technique on an independent set of patients can be expected to be more optimal than is possible with standard logistic regression.

Although special software is needed (like LS-SVMlab) to train LS-SVM models from a dataset, they can be used easily by clinicians once the parameters of the models have been determined, as shown by De Smet *et al.*[1], using regular software packages that allow implementation of elementary calculations. Both the standard logistic regression model and the LS-SVM with a linear kernel have been stated explicitly as simple equations in four variables and they can be evaluated immediately with a simple calculator, if necessary. Note that, in principle, the number of terms in an LS-SVM model is equal to the number of patients in the training set plus one. However, it is possible to write LS-SVMs with a linear kernel as a simple linear equation in their variables, by rearranging the terms. This is not possible for LS-SVMs with an RBF kernel, although each term in this sum has a very simple form and an LS-SVM model with an RBF kernel can therefore be implemented easily in any software package that allows calculations to be performed simply (e.g. in Microsoft Excel).

To conclude, we have summarized here the two advantages of the recent and more advanced SVMs over traditional logistic regression. Unlike logistic regression, SVMs have means to prevent the model from being sensitive to outliers in the data, resulting in a model that is capable of making good predictions for prospective analyses. Moreover, SVMs are able to cope with nonlinearity in the data by using nonlinear kernel functions instead of a simple linear kernel.

ACKNOWLEDGMENTS

This work was supported by: Research Council KUL: GOA-AMBioRICS, IDO (IOTA Oncology, Genetic networks), several PhD/postdoc and fellow grants; Flemish Government: FWO: PhD/postdoc grants, projects G.0115.01, G.0407.02, G.0413.03, G.0388.03, G.0229.03 and IWT: PhD Grants, STWW-Genprom, GBOU-McKnow, GBOU-SQUAD, GBOU-ANA; Belgian Federal Government: DWTC [IUAP V-22 (2002–2006)]; EU: CAGE; Biopattern.

N. L. M. M. Pochet* and J. A. K. Suykens
Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven (K.U.Leuven), Kasteelpark Arenberg 10, B-3001 Heverlee (Leuven), Belgium
**Correspondence.*
(e-mail: Nathalie.Pochet@esat.kuleuven.be)

REFERENCES

1. De Smet F, De Brabanter J, Van den Bosch T, Pochet N, Amant F, Van Holsbeke C, Moerman P, De Moor B, Vergote I, Timmerman D. New models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography. *Ultrasound Obstet Gynecol* 2006; **27**: 664–671.
2. Suykens JAK, Van Gestel T, De Brabanter J, De Moor BLR, Vandewalle J. *Least Squares Support Vector Machines*. World Scientific: Singapore, 2002.
3. Epstein E, Skoog L, Isberg P, De Smet F, De Moor B, Olofsson P, Gudmundsson S, Valentin L. An algorithm including results of gray scale and power Doppler ultrasound examination to predict endometrial malignancy in women with postmenopausal bleeding. *Ultrasound Obstet Gynecol* 2002; **20**: 370–376.
4. Freedland SJ, Isaacs WB, Mangold LA, Yiu, SK, Grubb KA, Partin AW, Epstein JI, Walsh PC, Platz EA. Stronger association between obesity and biochemical progression after radical prostatectomy among men treated in the last 10 years. *Clin Cancer Res* 2005; **11**: 2883–2888.
5. Timmerman D, Testa AC, Bourne T, Ferrazzi E, Ameye L, Konstantinovic ML, Van Calster B, Collins WP, Vergote I, Van Huffel S, Valentin L. Logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumor Analysis Group. *J Clin Oncol* 2005; **23**: 8794–8801.
6. Vapnik VN. *Statistical Learning Theory*. John Wiley & Sons: New York, 1998.
7. Van Gestel T, Suykens JAK, Baesens B, Viaene S, Vanthienen J, Dedene G, De Moor BLR, Vandewalle J. Benchmarking least squares support vector machine classifiers. *Machine Learning* 2004; **54**: 5–32.
8. LS-SVMlab version 1.5. http://www.esat.kuleuven.be/sista/lssvmlab/.

