Componentwise Support Vector Machines for Structure Detection

K. Pelckmans, J.A.K. Suykens, and B. De Moor

K.U.Leuven ESAT-SCD/SISTA, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
{kristiaan.pelckmans,johan.suykens}@esat.kuleuven.ac.be,
http://www.esat.kuleuven.ac.be/sista/lssvmlab

Abstract. This paper extends recent advances in Support Vector Machines and kernel machines in estimating additive models for classification from observed multivariate input/output data. Specifically, we address the question how to obtain predictive models which gives insight into the structure of the dataset. This contribution extends the framework of structure detection as introduced in recent publications by the authors towards estimation of componentwise Support Vector Machines (cSVMs). The result is applied to a benchmark classification task where the input variables all take binary values.

1 Introduction

The theory, methodology and application of Support Vector Machines (SVMs) has gained a mature status in the last decade, see e.g. [16, 3, 12, 13]. This work extends recent advances on primal-dual kernel machines for learning classification rules based on additive models [7] where the primal-dual optimization point of view [2] (as exploited by SVMs [16] and LS-SVMs [14, 13]) is seen to provide an efficient implementation [9]. Although relations exist with results on ANOVA kernels [16, 6], the optimization framework established a solid foundation for extensions towards structure detection similar to LASSO [15] and bridge regression [1] in the context of regression as elaborated in [9, 10]. The key idea was to employ a measure of maximal variation (as defined in the sequel) for the goal of regularization. Extensions towards handling missing values amongst the observed inputs were described in the context of cSVMs in [8].

This paper is organized as follows. Section 2 gives the main result of componentwise SVMs equipped with a measure of maximal variation. Section 3 then illustrates the concept on a UCI benchmark prediction task.

2 Componentwise Support Vector Machines

Given a set of observed input/output data-samples $\mathscr{D} = \{(x_i, y_i)\}_{i=1}^N \subset \mathbb{R}^D \times \mathbb{D}$ where $\mathbb{D} = \{-1, +1\}$. Learning a decision rule then amounts to identifying a function $f : \mathbb{R} \to \mathbb{D}$ such that any data-sample $(x_*, y_*) \in \mathbb{R}^D \times \mathbb{D}$ sampled from the same distribution $P_{XY} : \mathbb{R}^D \times \mathbb{D} \to [0,1]$ underlying the dataset $\mathscr{D}$ deviates minimal from the prediction using the model f. More formally, let F denote the class of admissible functions f. Learning amounts to the task of approximating the minimizer $f^* = \arg\min_{f \in F} \int (y - f(x))^2 dP_{XY}$. The framework of SVMs as given in [16] is adopted.

Definition 1 *(Additive Classifier) Let $x \in \mathbb{R}^D$ be a point with components $x \doteq \left(x^{(1)}, \ldots, x^{(P)}\right)$. Additive classifiers then take a componentwise form [7] defined as*

$$\operatorname{sign}[f(x)] = \operatorname{sign}\left[\sum_{p=1}^{P} f_p\left(x^{(p)}\right) + b\right], \tag{1}$$

with sufficiently smooth mappings $f_p : \mathbb{R}^{D_p} \to \mathbb{R}$ such that the decision boundary is described as in [16, 12]

$$\mathscr{H}_f = \left\{ x_0 \in \mathbb{R}^D \;\middle|\; \sum_{p=1}^{P} f_p\left(x_0^{(p)}\right) + b = 0,\ x_0 \in \mathbb{R}^P \right\}. \tag{2}$$

It is well-known [16] that the distance of any point x the hyper-plane $\mathscr{H}_{f_p}$ is given as

$$d\left(x, \mathscr{H}_f\right) = \frac{|f(x)|}{\|f'(x)\|} \geq \frac{y_i\left(\sum_{p=1}^{P} f_p\left(x^{(p)}\right) + b\right)}{\sum_{p=1}^{P} \|f^{(p)'}(x^{(p)})\|}, \tag{3}$$

as $\left\|\sum_{p=1}^{P} f^{(p)'}\left(x^{(p)}\right)\right\| \leq \sum_{p=1}^{P} \left\|f^{(p)'}\left(x^{(p)}\right)\right\|$ due to the triangle inequality. The optimal separating hyper-plane can be expressed as the model (3) solving

$$\max_{M \geq 0, f_p, b} M \quad \text{s.t.} \quad d(x_i, \mathscr{H}_{f_p}) \geq M. \tag{4}$$

After the change of variables in the function f such that $M\sum_{p=1}^{P} \|f^{(p)'}\| = 1$ and the application of the lower-bound (3), one can write alternatively

$$\begin{aligned}(\hat{f}, \hat{b}) = \arg\min_{f,b} \mathscr{J}(f) = \sum_{p=1}^{P} \left\|f^{(p)'}\right\| \\ \text{s.t.} \quad y_i\left(\sum_{p=1}^{P} f_p\left(x_i^{(p)}\right) + b\right) \geq 1, \ \ \forall i = 1, \ldots, N.\end{aligned} \tag{5}$$

Then the size of the margin is given as $M = 1/\sum_{p=1}^{P} \|f^{(p)'}\|$.

2.1 Structure Detection and Maximal Variation

Structure detection as in the case of LASSO and bridge regression in the case of linear parametric models becomes hard to incorporate into non-parametric and kernel methods. A possible approach then is to employ a measure of the contribution of any component which is not expressed directly in terms of the parameters. The following measure was proposed.

Definition 2 *(Maximal Variation) The maximal variation of a function $f_p : \mathbb{R}^{D_p} \to \mathbb{R}$ is defined as*

$$\mathscr{M}_p = \max_{x^{(p)} \in \mathbb{R}^{D_p}} \left|f_p\left(x^{(p)}\right)\right|, \tag{6}$$

for all $x^{(p)} \in \mathbb{R}^{D_p}$. The empirical maximal variation can be defined as

$$\hat{\mathscr{M}}_p = \max_{x_i^{(p)} \in \mathscr{D}} \left| f_p\left(x_i^{(p)}\right) \right|, \tag{7}$$

with $x_i^{(p)}$ denoting the p-th component of the i-th sample of the training set $\mathscr{D}$.

Adopting this definition, it becomes clear that when a certain component f_p finally has a maximal variation $\mathscr{M}_p$ equal to zero, the corresponding variables do not contribute to the learned classifier and may be omitted for the sake of prediction.

2.2 Componentwise Primal-Dual Kernel Classifiers

Consider the model

$$f(x) = \sum_{p=1}^{P} w_p^T \varphi_p\left(x^{(p)}\right) + b, \tag{8}$$

where $\varphi_p(\cdot) : \mathbb{R}^{D_p} \to \mathbb{R}^{n_h}$ denote the potentially infinite dimensional feature map and $w_p \in \mathbb{R}^{n_h}$ is the unknown parameter of the pth component for all $p = 1,\ldots,P$. The following regularized cost-function is considered:

$$\begin{aligned} \min_{w,b,e,t} \mathscr{J}_{\gamma,C}(w,t) &= \gamma \sum_{p=1}^{P} t_p + \frac{1}{2} \sum_{p=1}^{P} w_p^T w_p \\ \text{s.t.} \quad & \begin{cases} y_i \left(\sum_{p=1}^{P} w_p^T \varphi_p\left(x_i^{(p)}\right) + b \right) \geq 1 - e_i & \forall i = 1,\ldots,N \\ \sum_{i=1}^{N} e_i \leq C, \;\; e_i \geq 0 & \forall i = 1,\ldots,N \\ -t_p \leq w_p^T \varphi_p\left(x_i^{(p)}\right) \leq t_p & \forall i = 1,\ldots,N, p = 1,\ldots,P. \end{cases} \end{aligned} \tag{9}$$

The dual problem is given in the following Lemma.

Lemma 1 *(Dual of Componentwise SVM with Maximal Variation) Given the primal problem (9), the dual solution is*

$$\begin{aligned} \max_{\alpha_i,\rho_{ip}^+,\rho_{ip}^-,\lambda} & -\frac{1}{2} \sum_{i,j=1}^{N} \left(\alpha_i y_i + \rho_{ip}^+ - \rho_{ip}^-\right)\left(\alpha_j y_j + \rho_{jp}^+ - \rho_{jp}^-\right) \tilde{\Omega}_{ij}^P + \sum_{i=1}^{N} \alpha_i - \lambda C \\ & s.t. \quad \begin{cases} \sum_{i=1}^{N} y_i \alpha_i = 0 & \\ 0 \leq \alpha_i \leq \lambda & \forall i = 1,\ldots,N \\ \gamma = \sum_{i=1}^{N} (\rho_{ip}^+ + \rho_{ip}^-) & \forall p = 1,\ldots,P \\ \rho_{ip}^+, \rho_{ip}^- \geq 0 & \forall i = 1,\ldots,N, \forall p = 1,\ldots,P, \end{cases} \end{aligned} \tag{10}$$

where $\tilde{\Omega}_{ij}^P = \sum_{p=1}^{P} \tilde{K}_p\left(x_i^{(p)}, x_j^{(p)}\right)$ for all $i,j = 1,\ldots,N$ and where $\tilde{K}_p\left(x_i^{(p)}, x_j^{(p)}\right) = K_p\left(x_i^{(p)}, x_j^{(p)}\right)$. The resulting nonlinear classifier evaluated on a new data point $x_* =$

$\left(x_*^{(1)},\ldots,x_*^{(P)}\right)$ *takes the form*

$$\operatorname{sign}\left[\sum_{p=1}^{P}\sum_{i=1}^{N}\alpha_i^{(p)}K_p\left(x_i^{(p)},x_*^{(p)}\right)+b\right], \tag{11}$$

where $\hat{\alpha}_i^{(p)}=\left(\hat{\alpha}_i y_i+\hat{\rho}_{ip}^{+}-\hat{\rho}_{ip}^{-}\right)$ *for all* $i=1,\ldots,N$ *and* $p=1,\ldots,P$ *follow from the unique solution to (10).*

Proof. The dual solution is given after construction of the Lagrangian

$$\begin{aligned}\mathscr{L}_{\gamma,C}(w_p,b,e_i,t_p;\alpha_i,v_i,\rho_{ip}^{+},\rho_{ip}^{-},\lambda)=\mathscr{J}_{\gamma,C}(w_p,t_p)-\sum_{i=1}^{N}v_i e_i\\ -\sum_{i=1}^{N}\alpha_i\left(y_i\left(\sum_{p=i}^{P}w_p\varphi_p\left(x_i^{(p)}\right)+b\right)-1+e_i\right)+\lambda\left(\sum_{i=1}^{N}e_i-C\right)\\ -\sum_{i,p}\rho_{ip}^{+}\left(t_p+w_p^T\varphi_p\left(x_i^{(p)}\right)\right)-\sum_{i,p}\rho_{ip}^{-}\left(t_p-w_p^T\varphi_p\left(x_i^{(p)}\right)\right),\end{aligned} \tag{12}$$

with positive multipliers $0\leq\alpha_i,v_i,\rho_{ip}^{+},\rho_{ip}^{-}$ and $\lambda\geq 0$. The solution is given by the saddle point of the Lagrangian [2]

$$\max_{\alpha_i,v_i,\rho_{ip}^{+},\rho_{ip}^{-},\lambda}\ \min_{w_p,b,e_i,t_p}\ \mathscr{L}_{\gamma,C}. \tag{13}$$

By taking the first order conditions $\frac{\partial\mathscr{L}_{\gamma,C}}{\partial w_p}=0$, $\frac{\partial\mathscr{L}_{\gamma,C}}{\partial b}=0$, $\frac{\partial\mathscr{L}_{\gamma,C}}{\partial e_i}=0$ and $\frac{\partial\mathscr{L}_{\gamma,C}}{\partial t_p}=0$, one obtains the (in)equalities $w_p=\sum_{i=1}^{N}\left(\alpha_i y_i+\rho_{ip}^{+}-\rho_{ip}^{-}\right)\varphi_p(x_i^{(p)})$, $\sum_{i=1}^{N}\alpha_i y_i=0$, $0\leq\alpha_i\leq\lambda$ and $\gamma=\sum_{i=1}^{P}\left(\rho_{ip}^{+}+\rho_{ip}^{-}\right)$. By application of the kernel trick $K_p(x_i,x_j)=\varphi_p(x_i^{(p)})^T\varphi_p(x_j^{(p)})$, the solution to (8) is found by solving the dual problem The primal variables b,e_i and t_p can be recovered from the complementary slackness conditions. □

3 Example: learning logical rules using componentwise SVMs

In order to explore the capabilities of the presented method, the 1984 United States Congressional voting records database available on the UCI benchmark repository is used. This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. A 90%- 95% performance was reported in [11]. Here we explore the capabilities of the described framework to learn a parsimonious decision system from the observed data.

We first elaborate on the issue how to handle the special structure of the input data. Inspired by the method of first order inductive logic programming, one may let the

variable x be mapped on the truth-value as $T(x) = TRUE$ if $x = +1$ and $T(x) = FALSE$ otherwise. Then the following relation holds

$$T(y) = OR(T(x^1), \ldots, T(x^p)) \Leftrightarrow y = \text{sign}\left(\sum_{p=1}^{P} x^p + P - 1\right), \tag{14}$$

where *OR* is the logical OR operation. This motivates the use of the additive model (3). Furthermore note that $\neg T(x) = T(-x)$ holds where $\neg$ denotes the logical negation. The *AND* operator is induced by the use of the following kernel.

Definition 3 *(Logic 'AND' Kernel) Let π_p be a nonempty set of indices $1, \ldots, D$ for all $p = 1, \ldots, P$ and let D_p denotes the number of different indices in π_p. Let the feature space mapping of the pth component be defined as*

$$\varphi_p(x) = I_{AND}\left(x^{\pi_p(1)}, \ldots, x^{\pi_p(D_p)}\right), \tag{15}$$

where the indicator function $I_{AND}(\cdot)$ is $+1$ if all arguments equal $+1$ and -1 otherwise. The corresponding Mercer kernel becomes

$$K_p(x_i, x_j) = I_{AND}\left(x_i^{\pi_p(1)}, \ldots, x_i^{\pi_p(D_p)}\right) I_{AND}\left(x_j^{\pi_p(1)}, \ldots, x_j^{\pi_p(D_p)}\right). \tag{16}$$

For this example, a maximal order of $D_p = 2$ is used in order to keep the computations tractable. The datset was divided in disjunct training dataset ($N = 250$), validation set (of size 100) and test set (of size 85). The pameters γ and C where tuned minimizing the misclassification rate on the validation set. A Monte Carlo simulation was conducting resulting in mean testset performance of 96.24% with a one sigma bound as $96.24\% \pm 1.2\%$. The vote of congressmen pro or contra the democration candidate is in most predictors of the sample proportional to their vote pro the resolutions of (3) cost-sharing of water-projects (4) inversely proportional to the adoption of the budget-resolution and (4) proportional to the vote concerning immigration.

4 Conclusions

This paper studied the estimation of additive classifiers by componentwise SVMs. The measure of maximal variation was used to perform structure detection. An example was elaborated where the task is to infer logical rules from binary observations by use of the additive model structure and the use of a logical AND kernel.

Acknowledgments. This research work was carried out at the ESAT laboratory of the KUL. Research Council KU Leuven: Concerted Research Action GOA-Mefisto 666, GOA-Ambiorics IDO, several PhD/postdoc & fellow grants; Flemish Government: Fund for Scientific Research Flanders (several PhD/postdoc grants, projects G.0407.02, G.0256.97, G.0115.01, G.0240.99, G.0197.02, G.0499.04, G.0211.05, G.0080.01, research communities ICCoS, ANMMM), AWI (Bil. Int. Collaboration Hungary/ Poland), IWT (Soft4s, STWW-Genprom, GBOU-McKnow, Eureka-Impact, Eureka-FLiTE, several PhD grants); Belgian Federal Government: DWTC IUAP IV-02 (1996-2001) and IUAP V-10-29 (2002-2006) (2002-2006), Program Sustainable Development PODO-II (CP/40); Direct contract research: Verhaert, Electrabel, Elia, Data4s, IPCOS. JS is an associate professor and BDM is a full professor at K.U.Leuven Belgium, respectively.

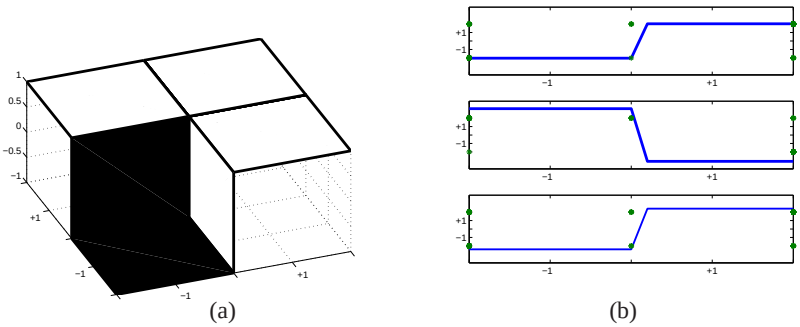

Fig. 1. (a) *Indicator function of the function* $y = \text{sign}\left(x^1 + x^2 + 1\right)$ *and* **(b)** *The tuned predictor is a function of the votes on vote (3), (4) and (11).*

References

1. A. Antoniadis and J. Fan. Regularized wavelet approximations (with discussion). *Jour. of the Am. Stat. Ass.*, 96:939–967, 2001.
2. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
3. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
4. L.E. Frank and J.H. Friedman. A statistical view of some chemometric regression tools. *Technometrics*, 35:109–148, 1993.
5. W.J. Fu. Penalized regression: the bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, 7:397–416, 1998.
6. S. R. Gunn and J. S. Kandola. Structural modelling with sparse kernels. *Machine Learning*, 48(1):137–163, 2002.
7. T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990.
8. K. Pelckmans, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Maximal variation and missing values for componentwise support vector machines. Technical report, SCD - ESAT - KULeuven, Leuven, 2005.
9. K. Pelckmans, I. Goethals, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Componentwise least squares support vector machines. Chapter in *Support Vector Machines: Theory and Applications*, L. Wang (Ed.), Springer, 2004, *In press*.
10. K. Pelckmans, J.A.K. Suykens, and B. De Moor. Building sparse representations and structure determination on LS-SVM substrates. *Neurocomputing, in press*, 2005.
11. J.C. Schlimmer. *Concept acquisition through representational adjustment*. PhD thesis, Department of Information and Computer Science, University of California, Irvine, CA, 1987.
12. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
13. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.
14. J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.
15. R.J. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society*, 58:267–288, 1996.
16. V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.