Load Forecasting using Fixed-Size Least Squares Support Vector Machines

Marcelo Espinoza, Johan A.K. Suykens, and Bart De Moor

K.U.Leuven, ESAT-SCD-SISTA
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium
{marcelo.espinoza,johan.suykens}@esat.kuleuven.ac.be

Abstract. Based on the Nyström approximation and the primal-dual formulation of Least Squares Support Vector Machines (LS-SVM), it becomes possible to apply a nonlinear model to a large scale regression problem. This is done by using a sparse approximation of the nonlinear mapping induced by the kernel matrix, with an active selection of support vectors based on quadratic Renyi entropy criteria. The methodology is applied to the case of load forecasting as an example of a real-life large scale problem in industry, for the case of 24-hours ahead predictions. The results are reported for different number of initial support vectors, which cover between 1% and 4% of the entire sample, with satisfactory results.

1 Introduction

Computing forecasts of a time series requires the estimation of a good initial model. The more information is included in the model, the better the performance of the forecasts, leading to good performance beyond the first step out of sample. Kernel based estimation techniques, such as Support Vector Machines (SVMs) and Least Squares Support Vector Machines (LS-SVMs) have shown to be powerful nonlinear regression methods [14, 1, 15]. Both techniques build a linear model in the so-called feature space where the inputs have been transformed by means of a (possibly infinite dimensional) nonlinear mapping $\boldsymbol{\varphi}$. This is converted to the dual space by means of the Mercer's theorem and the use of a positive definite kernel, without computing explicitly the mapping $\boldsymbol{\varphi}$. The SVM model solves a quadratic programming problem in dual space, obtaining a sparse solution. The LS-SVM formulation, on the other hand, solves a linear system in dual space under a least-squares cost function [13], where the sparseness property can be obtained by sequentially pruning the support value spectrum.

Although the LS-SVM system is solved on its dual form, the problem can be formulated directly in primal space by means of an explicit approximation for the nonlinear mapping $\boldsymbol{\varphi}$. Furthermore, it is possible to compute a sparse approximation by using only a subsample of selected Support Vectors from the dataset in order to estimate a large-scale nonlinear regression problem in primal space, making use of all possible information from the available data sample. Working in primal space gives enough flexibility to apply different techniques from statistics or the traditional system identification framework [9].

In this paper we apply LS-SVM in primal space to the problem of short-term load forecasting, which is an important area of quantitative research [11, 4], in order to build a model that can be used for 24-hours ahead prediction. Usually a load series shows important seasonal patterns (yearly, weekly, intra-daily patterns) that need to be taken into account in the modelling strategy [8]. In our case, the data series comes from a local low voltage substation in Belgium and it contains 28,000 hourly values. This paper is structured as follows. The description of the LS-SVM is presented in Section 2. In Section 3, the methodology for working in primal space is described, with the particular application to a large scale problem. Section 4 presents the problem and describes the setting for the estimation, and the results are reported in Section 5.

2 Function Estimation using LS-SVM

The standard framework for LS-SVM estimation is based on a primal-dual formulation. Given the dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$ the goal is to estimate a model of the form

$$y_i = \boldsymbol{w}^T \boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$, $y \in \mathbb{R}$ and $\boldsymbol{\varphi}(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n_h}$ is a mapping to a high dimensional (and possibly infinite dimensional) feature space. The following optimization problem is formulated:

$$\min_{\boldsymbol{w},b,\boldsymbol{e}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2 \tag{2}$$

$$\text{s.t.} \quad y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i, \quad i = 1, \ldots, N.$$

With the application of the Mercer's theorem for the kernel matrix $\boldsymbol{\Omega}$ as $\boldsymbol{\Omega}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{\varphi}(\boldsymbol{x}_j), i, j = 1, \ldots, N$ it is not required to compute explicitly the nonlinear mapping $\boldsymbol{\varphi}(\cdot)$ as this is done implicitly through the use of positive definite kernel functions K. For $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ there are usually the following choices: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i^T\boldsymbol{x}_j$ (linear kernel); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^T\boldsymbol{x}_j + c)^d$ (polynomial of degree d, with c a tuning parameter); $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2/\sigma^2)$ (radial basis function, RBF), where σ is a tuning parameter. From the Lagrangian $\mathcal{L}(\boldsymbol{w}, b, e; \alpha) = \frac{1}{2}\ \boldsymbol{w}^T\boldsymbol{w} + \gamma\frac{1}{2}\sum_{i=1}^N e_i^2 - \sum_{i=1}^N \alpha_i(\boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i - y_i)$, where $\alpha_i \in \mathbb{R}$ are the Lagrange multipliers, the conditions for optimality are given by $\frac{\partial\mathcal{L}}{\partial\boldsymbol{w}} = 0 \to \boldsymbol{w} = \sum_{i=1}^N \alpha_i\boldsymbol{\varphi}(x_i)$, $\frac{\partial\mathcal{L}}{\partial b} = 0 \to \sum_{i=1}^N \alpha_i = 0, \frac{\partial\mathcal{L}}{\partial e_i} = 0 \to \alpha_i = \gamma_i e_i, \qquad i = 1, \ldots, N$, and $\frac{\partial\mathcal{L}}{\partial\alpha_i} = 0 \to y_i = \boldsymbol{w}^T\boldsymbol{\varphi}(\boldsymbol{x}_i) + b + e_i$. By elimination of $\boldsymbol{w}$ and e_i, the following linear system is obtained:

$$\left[\begin{array}{c|c} 0 & \mathbf{1}^T \\ \hline \mathbf{1} & \boldsymbol{\Omega} + \gamma^{-1}\mathbf{I} \end{array}\right] \left[\begin{array}{c} b \\ \hline \boldsymbol{\alpha} \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \boldsymbol{y} \end{array}\right], \tag{3}$$

with $\boldsymbol{y} = [y_1, \ldots, y_N]^T, \boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_N]^T$. The resulting LS-SVM model in dual space becomes

$$y(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b. \tag{4}$$

Usually the training of the LS-SVM model involves an optimal selection of the kernel parameters and the regularization parameter, which can be done using e.g. cross-validation techniques or Bayesian inference.

3 Estimation in Primal Space

In order to work in the primal space, it is required to compute an explicit approximation of the nonlinear mapping $\boldsymbol{\varphi}$. Then, the final estimation of the model can be done using different techniques.

3.1 Nonlinear Approximation in Primal Space

Explicit expressions for $\boldsymbol{\varphi}$ can be obtained by means of an eigenvalue decomposition of the kernel matrix $\boldsymbol{\Omega}$ with entries $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Given the integral equation $\int K(\boldsymbol{x}, \boldsymbol{x}_j)\phi_i(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \lambda_i\phi_i(\boldsymbol{x})$, with solutions λ_i and ϕ_i for a variable $\boldsymbol{x}$ with probability density $p(\boldsymbol{x})$, we can write

$$\boldsymbol{\varphi} = [\sqrt{\lambda_1}\phi_1, \sqrt{\lambda_2}\phi_2, \ldots, \sqrt{\lambda}_{n_h}\phi_{n_h}]. \tag{5}$$

Given the dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^N$, it is possible to approximate the integral by a sample average. This will lead to the eigenvalue problem (Nyström approximation [16])

$$\frac{1}{N}\sum_{k=1}^{N} K(\boldsymbol{x}_k, \boldsymbol{x}_j)u_i(\boldsymbol{x}_k) = \lambda_i^{(s)}u_i(\boldsymbol{x}_j), \tag{6}$$

where the eigenvalues λ_i and eigenfunctions ϕ_i from the continuous problem can be approximated by the sample eigenvalues $\lambda_i^{(s)}$ and eigenvectors u_i as

$$\hat{\lambda}_i = \frac{1}{N}\lambda_i^{(s)}, \hat{\phi}_i = \sqrt{N}u_i. \tag{7}$$

Based on this approximation, it is possible to compute the eigendecomposition of the kernel matrix $\boldsymbol{\Omega}$ and use its eigenvalues and eigenvectors to compute the ith required component of $\hat{\boldsymbol{\varphi}}(\boldsymbol{x})$ simply by applying (5) if $\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{i=1}^N$ (is a training point), or for any point $\boldsymbol{x}^{(v)}$ by means of

$$\hat{\boldsymbol{\varphi}}_i(\boldsymbol{x}^{(v)}) \propto \frac{1}{\sqrt{\lambda_i^{(s)}}}\sum_{k=1}^{N} u_{ki}K(\boldsymbol{x}_k, \boldsymbol{x}^{(v)}). \tag{8}$$

This finite dimensional approximation $\hat{\boldsymbol{\varphi}}(\boldsymbol{x})$ can be used in the primal problem (2) to estimate $\boldsymbol{w}$ and b.

3.2 Sparseness and Large Scale Problems

It is important to emphasize that the use of the entire training sample of size N to compute the approximation of $\boldsymbol{\varphi}$ will yield at most N components, each one of which can be computed by (7) for all $\boldsymbol{x} \in \{\boldsymbol{x}_i\}_{i=1}^N$. However, if we have

a large scale problem, it has been motivated [14] to use of a subsample of $M \ll N$ datapoints to compute the $\hat{\boldsymbol{\varphi}}$. In this case, up to M components will be computed. External criteria such as entropy maximization can be applied for an optimal selection of the subsample: given a fixed-size M, the aim is to select the support vectors that maximize the quadratic Renyi entropy [7]

$$H_R = -\log \int p(\boldsymbol{x})^2 d\boldsymbol{x}, \tag{9}$$

which can be approximated by using $\int \hat{p}(\boldsymbol{x})^2 d\boldsymbol{x} = \frac{1}{M^2}\mathbf{1}^T \boldsymbol{\Omega}\mathbf{1}$. The use of this active selection procedure can be important for large scale problems, as it is related to the underlying density distribution of the sample. In this sense, the optimality of this selection is related to the final accuracy that can be obtained in the modeling exercise.

3.3 Estimation Technique

Once the nonlinear mapping has been computed using a sparse approximation based on a subsample, the model has to be estimated in primal space. Let us denote by $\boldsymbol{z}_k = \hat{\boldsymbol{\varphi}}(\boldsymbol{x}_k)$ for $k = 1, \ldots N$ and consider the new $\boldsymbol{z}_k \in \mathbb{R}^m$ as the observations for the linear regression

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + b\mathbf{1} + \boldsymbol{e} \tag{10}$$

with $\boldsymbol{e} = [e_1, e_2, \ldots, e_N]^T \in \mathbb{R}^{N\times 1}$ $\boldsymbol{y} = [y_1, y_2, \ \ldots, y_N]^T \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{Z} = [\boldsymbol{z}_1^T; \boldsymbol{z}_2^T; \boldsymbol{z}_3^T; \ldots; \boldsymbol{z}_N^T] \in \mathbb{R}^{N\times m}$. The quantity m is the number of components of $\hat{\boldsymbol{\varphi}}$ that are going to be considered in the estimation [5], with $m \leq M$. For ease of notation, consider the matrix of full regressors $\boldsymbol{Z_F} = [\mathbf{1}\boldsymbol{Z}]$, and the vector of full coefficients $\boldsymbol{\beta_F} = [b, \boldsymbol{\beta}]^T$. The regression (10) can be written as:

$$\boldsymbol{y} = \boldsymbol{Z_F}\boldsymbol{\beta_F} + \boldsymbol{e}. \tag{11}$$

In this paper, we estimate the above equation by Ordinary Least Squares (OLS), where only $m < M$ components of $\hat{\boldsymbol{\varphi}}$ are used, and they are selected by looking at the eigenspectrum of the $M \times M$ kernel matrix $\boldsymbol{\Omega}$ [3].

4 Practical Example: Short-Term Load Forecasting

In this section the practical application is described, in term of the problem context, methodological issues and results.

4.1 Description and Objective

The main goal in load forecasting to estimate a model that can capture all the dynamics between possible explanatory variables for the load, in order to produce forecasts from one hour ahead, up to one day ahead [10]. In the literature there is a broad consensus about important explanatory variables: past values of the load, weather information, calendar information, etc. For further details, the interested reader is referred to [4, 5]. Our objective is to show the application of the large-scale methodology described in the previous section, which is a practical approach to exploit the information contained in a large load time series.

4.2 Data and Methodology

The data consists of a 3.5-years load series from a particular substation in Belgium, thuis containing more than 28,000 datapoints. We define our training set to consist of 26,000 datapoints. The model formulation to be tested is a nonlinear ARX specification, with the following structure: an autoregressive part of 164 lagged load values, covering a complete week; temperature-related variables measuring the effect of temperature on cooling and heating requirements [2]; and calendar information in the form of dummy variables for month of the year, day of the week and hour of the day [4]. This leads to a set of 211 explanatory variables. The procedure for modelling with the large-scale fixed-size LS-SVM methodology can be described in the following steps:

1. Normalize inputs and outputs to have zero mean and unit variance
2. Select an initial subsample of size M
3. Build the M-size kernel matrix and compute its eigendecomposition
4. Build the nonlinear mapping approximation for the rest of the data
5. Estimate a linear regression in primal space
6. Estimate the nonlinear mapping approximation for a test data
7. Use the regression estimates with the test data nonlinear mapping to produce the out of sample forecast

In this example, tuning of the hyperparameter σ is performed by 10 fold crossvalidation (CV) in the training sample. We keep the value of σ that minimizes the out-of-sample mean squared error (MSE). To illustrate the effect of increasing sizes of M, the above methodology is tested for sizes of $M = 200, 400, 600, 800$ and 1000 support vectors, selected with the quadratic entropy criterion. It is important to emphasize that we are using between 1% and 4% of the available dataset to build the nonlinear mapping for the entire sample. Values of M larger than 1000 are possible, as the only constraint in this approach is the available computational resources at hand.

To compare the results of the nonlinear models, we estimate a linear regression with the same set of initial inputs. After estimation, the inputs that are not statistically significant are removed on a stepwise selection process based on the $t-$statistics, yielding a final set of 153 regressors.

5 Results

In this section the results of the fixed-size LS-SVM methodology applied to the load modelling problem are reported, for the training procedure, the selection of support vectors and the out of sample performance.

5.1 Training Performance

The results for the computed MSE in a crossvalidation basis, and the equivalent result for the linear model, are shown in Table 1 using the selected σ. It is also reported the number of support vectors m that are used for building the nonlinear mapping on each case.

Estimation	MSE (CV)	Selected components m
Linear	0.043	153
M=200	0.032	166
M=400	0.022	327
M=600	0.017	416
M=800	0.016	462
M=1000	0.015	496

Table 1. CV performance of fixed-size LS-SVM with M support vectors.

Model	MSE test (one-hour-ahead)	MSE test (24-hours-ahead)
Linear	0.038	0.042
M=200	0.027	0.031
M=400	0.019	0.025
M=600	0.016	0.023
M=800	0.014	0.021
M=1000	0.014	0.021

Table 2. Performance on a test set for different forecasting horizons

5.2 Support Vector Selection

It is possible to compare the performances for a model estimated with a random selection of support vectors awith respect to the same model estimated with a quadratic Renyi entropy selection starting from the same random selection. Both models can be compared in terms of performance on the same test set. Figure 1 show the comparison for the results after 20 random initial selections, in which the model is either estimated after quadratic entropy selection taking the random selection as the initial starting point (Case I), or it is estimated directly (Case II). In all tests it has been used $M = 200$. Clearly starting from different random selections, the entropy-based selection yields lower dispersion in the errors.

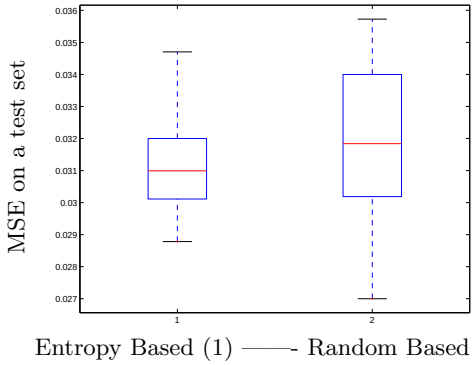

Fig. 1. Box-plot of the MSE in a test set for models estimated with entropy-based (1) and random (2) selection of support vectors. Results for 20 repetitions.

5.3 Out of Sample Performance

The estimated models are tested in a one-hour-ahead basis and a 24-hours-ahead prediction with iterative simulations running for the next 24 hours. Averages of the MSE for each case are reported on Table 3. An example of one-hour-ahead prediction for the case $M = 400$ and 24-hours-ahead prediction for $M = 800$ are shown in Figure 2.

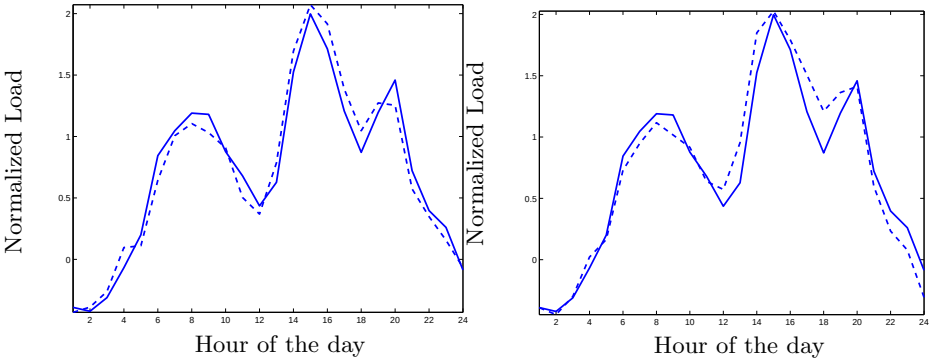

Fig. 2. Out of sample forecasts (dashed lines) and actual values (full line) for a selected 24-hours period, for the case of one-hour-ahead with $M = 400$ (Left); and 24-hours-ahead with $M = 800$ (Right).

6 Conclusion

The methodology of fixed-size LS-SVM has been applied to short-term load forecasting. This paper has shown that it is possible to build a large scale nonlinear regression model from a dataset with 26,000 samples using different subsamples as support vectors, with satisfactory results. The results show that the nonlinear regressions in primal space improve their accuracy with larger values of M. The maximum value of M to be used depends on the computational resources at hand, and it also depends on the underlying distributional properties of the dataset. In this context, it was shown that quadratic entropy active selection of support vectors leads to performances which are less disperse as those obtained by random selection of support vectors. The forecasting performance is very good, with MSE's around 2% for both forecasting modes (one-hour-ahead and 24-hours-ahead) when the nonlinear mapping is computed using 800 selected support vectors, which represent 3% of the sample available. Further research on a more dedicated definition of the initial input variables should lead to further improvements.

Acknowledgments. This work is supported by grants and projects for the Research Council K.U.Leuven (GOA- Mefisto 666, GOA- Ambiorics, several PhD/ Postdocs & fellow grants), the Flemish Government (FWO: PhD/ Postdocs grants, projects

G.0211.05, G.0240.99, G.0407.02,G.0197.02,G.0141.03,G.0491.03,G.0120.03,G.0452.04, G.0499.04, ICCoS, ANMMM; AWI; IWT: PhD grants, GBOU (McKnow) Soft4s), the Belgian Federal Government (Belgian Federal Science Policy Office: IUAP V-22; PODO-II (CP/ 01/40), the EU (FP5- Quprodis; ERNSI, Eureka 2063- Impact; Eureka 2419- FLiTE) and Contracts Research / Agreements (ISMC /IPCOS, Data4s, TML, Elia, LMS, IPCOS, Mastercard). J. Suykens and B. De Moor are an associated professor and a full professor at the K.U.Leuven, Belgium, respectively. The scientific responsibility is assumed by its authors.

References

1. Cristianini, N., and Shawe-Taylor, J., *An introduction to Support Vector Machines.* Cambridge University Press, 2000.
2. Engle, R., Granger, C.J., Rice, J., and Weiss, A. "Semiparametric Estimates of the Relation Between Weather and Electricity Sales," *Journal of the American Statistical Association*, Vol.81, No.394, 310-320, 1986.
3. Espinoza, M., Suykens, J.A.K. and De Moor B. "Least Squares Support Vector Machines and Primal Space Estimation," in Proc. of the *IEEE 42nd Conference on Decision and Control*, Maui, USA, 2003, pp. 3451-3456.
4. Espinoza, M., Joye, C., Belmans, R., and De Moor, B. "Short Term Load Forecasting, Profile Identification and Customer Segmentation: A Methodology based on periodic time series," Internal Report 04-141, ESAT-SISTA, K.U.Leuven, 2004.
5. Espinoza, M., Suykens, J.A.K. and De Moor, B. "Fixed Size Least Squares Support Vector Machines: A Large Scale Application in Electrical Load Forecasting," Internal Report 03-92, ESAT-SISTA, K.U.Leuven, 2003.
6. Fay, D., Ringwood, J., Condon, M. and Kelly, M. "24-h Electrical Load Data-A Sequential or Partitioned Time Series?," *Neurocomputing* 55, 469-498, 2003.
7. Girolami, M. "Orthogonal Series Density Estimation and the Kernel Eigenvalue Problem," *Neural Computation* 14(3), 669-688, 2003.
8. Hylleberg, S. *Modelling Seasonality.* Oxford University Press, 1992.
9. Ljung, L. *Systems Identification: Theory for the User.* Prentice Hall, 1999.
10. Mariani, E. and Murthy, S.S. *Advanced Load Dispatch for Power Systems.* Advances in Industrial Control, Springer-Verlag, 1997.
11. Ramanathan, R., Engle, R., Granger, C.W.J., Vahid-Aragui, F., Brace, C. "Short-run Forecasts of Electricity Load and Peaks," *Int. J. of Forecasting* 13, 1997.
12. Steinherz, H., Pedreira, C. and Castro, R. "Neural Networks for Short-Term Load Forecasting: A Review and Evaluation," *IEEE Trans. Power Syst.*, Vol.16. No.1, 2001.
13. Suykens, J.A.K., Vandewalle, J. "Least Squares Support Vector Machines Classifiers," *Neural Processing Letters* 9, 293-300, (1999).
14. Suykens J.A.K.,Van Gestel T., De Brabanter J., De Moor B.,Vandewalle J., *Least Squares Support Vector Machines.* World Scientific, 2002, Singapore.
15. Vapnik, V. *Statistical Learning Theory.* John Wiley & Sons, New York, 1998.
16. Williams, C.K.I and Seeger, M. "Using the Nyström Method to Speed Up Kernel Machines," in T.Leen, T.Dietterich, V.Tresp (Eds.), *Proc. NIPS'2000*, vol 13., MIT press.