SIMPLIFYING SUPPORT VECTOR MACHINES FOR REGRESSION ANALYSIS OF HYPERSPECTRAL IMAGERY

Andreas Rabe, Sebastian van der Linden and Patrick Hostert

Humboldt-Universität zu Berlin, Geomatics Lab, Unter den Linden 6, 10099 Berlin, Germany;
[andreas.rabe, sebastian.linden, patrick.hostert]@geo.hu-berlin.de

ABSTRACT

Support Vector Machines for Regression (SVR) proved to perform well. However, they are not preferred in image analysis due to a high number of needed support vectors (SV) and consequently long processing times. We present a method for simplifying the original SVR regression function up to a user-specified degree of accepted performance decrease. We show results for two regression problems: modelling leaf area index and dry vegetation mixing fraction using simulated hyperspectral EnMAP data. In both cases, SVR demonstrate high potential for modelling complex dependencies between hyperspectral reflectance and quantitative targets. By simplifying the original SVR, we observed reduction rates in number of SV in the 86% to 95% range for acceptable degrees of approximation quality. This enables a fast mapping of complete EnMAP scenes.

Index Terms — Support Vector Machines, SVM, approximation, regression, hyperspectral, EnMAP

1. INTRODUCTION

The SVR concept has been shown to be superior to other function estimation approaches. Its ability to model complex, non-linear dependencies in high dimensional feature spaces through the concept of kernel functions and regularization recommends SVR for several remote sensing applications. It was successfully applied for the estimation of biophysical parameters [1], subpixel cover fractions [10] and moisture transport [11]. Beside these direct modelling approaches, SVR is used for image registration [6] and inversion of radiative transfer models [3].

Despite such first promising results SVR are seldom used in remote sensing image analysis. This may, on the one hand, be explained by a lack of available implementations, on the other hand, by high processing times. The latter is explained by the fact that for an accurately parameterized SVR model the number of SV scales linearly with the number of training examples. Smola et al. [7] show that the optimal value for the insensitivity parameter in Vapnik's ε-insensitive loss function is driven by variance around the target function's expectation. For a given ε-loss $\varepsilon \geq 0$, all training vectors with deviation (between reference and estimation) larger than ε become SV and the number of SV is a fraction of the number of training vectors (Fig. 1). In the special case of ε equal zero, also known as laplacian loss function, all training vectors become SV. In comparison to optimal choosen ε-loss, SVR with laplacian loss still leads to good results without tuning ε-loss parameter.

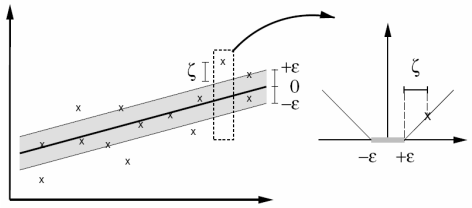

Figure 1 Corresponding loss settings for the ε-insensitive loss function and linear SVR (Smola et al. 1998)

The fact that the number of SV is directly related to the target function's variance can be used to simplify SVR models. Osuna and Girosi [5] propose a method for approximating, and thus simplifying, the decision function of Support Vector Machines for Classification. They achieve reduction rates for the number of SV in different benchmark problems in the 50% to 95% range [5]. The approach can easily be adapted to the SVR case by

1. training the SVR using kernel, regularization and and ε-loss parameters of choice. This step defines the regression function $f(\mathbf{x})$.
2. running a second SVR on the data $\{\mathbf{x}_i, f(\mathbf{x}_i)\}$, where $\mathbf{x}_i$ is an SV obtained in step 1. This will provide a function of the same form, but with fewer SV. In this step, the same kernel as in step 1 together with a high regularization parameter value and the desired ε accuracy is used.

The data in step 2 shows no variance around the target function's expectation, because all original targets are replaced by the SVR estimation $f(\mathbf{x}_i)$ derived in step 1. Under this no-variance condition SVR with nonzero ε-loss lead to sparse models, i.e. less SV. The method can be regarded as a post-training filter with implicitly user driven trade-off between sparseness and approximation accuracy.

978-1-4244-4687-2/09/$25.00 ©2009 IEEE

In our work we address the following arising questions:

- How is the performance of the simplified SVR affected by the choosen ε-loss?
- What is a good range to search for the optimal ε-loss?
- Which search strategy is most suitable?

Section 2 briefly introduces SVR and Section 3 explains the theory of our methodology. Section 4 sets up a benchmark test to empirically explore the dependency between reduction rate and performance of simplified SVR. Results are shown in Section 5 and discussed in Section 6, including a recommendation for an efficient implementation.

2. SUPPORT VECTOR REGRESSION

The SVR is a universal learning machine for solving function estimation problems [8]. It implements the Structural Risk Minimisation principle, which has been shown to be superior to the traditional Empirical Risk Minimisation principle [8]. Following notation from [2]: given a set of training examples $\{(\mathbf{x}_1,y_1),...,(\mathbf{x}_l,y_l)\}$, such that $\mathbf{x}_i \in R^d$ is an input, called training vector, and $y_i \in R^1$ is a target output, one tries to estimate a function f: $R^d \rightarrow R$ with $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$ minimizing the regularized risk functional

$$\begin{aligned} \min_{\mathbf{w},b,\xi,\xi^*} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i + C\sum_{i=1}^{l}\xi_i^* \\ \text{subject to} \quad & y_i - \mathbf{w}^T\phi(\mathbf{x}_i) - b \le \varepsilon + \xi_i, \\ & \mathbf{w}^T\phi(\mathbf{x}_i) + b - y_i \le \varepsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \ge 0, i = 1,\ldots,l. \end{aligned}$$

The dual is

$$\min_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*} \ \frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T\mathbf{Q}(\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \varepsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*)$$

$$\text{subject to } \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0, \ \ 0 \le \alpha_i, \alpha_i^* \le C, \ \ i = 1,\ldots,l,$$

where $Q_{ij} = K(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ and K is the kernel. Here vectors $\mathbf{x}_i$ are mapped implicitly into a higher (maybe infinite) dimensional feature space. The regularization parameter C determines the trade off between flatness of f and the amount up to which deviations larger than ε are tolerated. ε is the parameter of Vapnik's ε-insensitive loss function $|x|_\varepsilon = \max(0, |x| - \varepsilon)$. SVR regression function is:

$$f(\mathbf{x}) = \sum_{i=1}^{l}(-\alpha_i + \alpha_i^*)K(\mathbf{x}_i,\mathbf{x}) + b.$$

3. METHODS

The approach proposed by Osuna and Gironi [5] requires the user to define a "desired ε accuracy", i.e. ε-loss, during step 2. Although ε is in the same domain as the target, no *explicit* relation between the choosen ε-loss and performance of the simplified SVR exists.

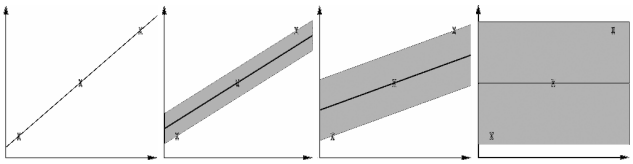

Figure 2 Approximation results for growing ε-loss settings.

Setting ε to 0.1 does not imply that the mean absolute error (MAE), root mean squared error (RMSE) or any other measures, such as Pearson's correlation (R), degrades to the same amount. To explore the highly problem dependent *implicit* relation, we monitor the SVR model performance (error and accuracy) inside a useful range of the ε-domain. To resolve the full spectrum of possible degrees of quality we start from a perfect approximation, i.e. $\varepsilon = 0$, up to a totally degenerated approximation, where the resulting function is a constant, i.e. the training data mean (Fig. 2).
Nevertheless, using the data range as an upper bound for the monitoring process is ineffective, as for example problems may result from a data range stretched by outliers. To overcome this, we introduce the *empirical distribution* of the *deviation* given by the training examples $(\mathbf{x}_i,y_i)$ as a less scale-specific measure of variation: each training vector $\mathbf{x}_i$, with estimated target $f(\mathbf{x}_i)$ derived from the original SVR, has a deviation $|y_i - f(\mathbf{x}_i)|$, contributing to the deviation distribution. On this basis we are able to define the monitoring process by

1. training the SVR using kernel and parameters of choice. This step defines the regression function f.
2. calculating the empirical distribution from the deviations $|y_i - f(\mathbf{x}_i)|$ of the training examples.
3. calculating the distribution percentiles.
4. For each percentile i, apply the method of Osuna and Girosi [5] to simplify the SVR and estimate the performance measures of choice. Use in this step an ε-loss corresponding to the value of the i^{th} percentile.

Note that the performance of simplified SVR is sampled along the ε-domain in irregular steps, defined by the percentiles of the deviation distribution and is thus problem dependent. This can be expected to be a clear advantage compared to a fixed linear or exponential step width.

4. BENCHMARK TESTS

Our benchmark tests are performed on two simulated hyperspectral images generated within the pre-flight acitivities of the EnMAP mission. A detailed description on the simulation, including spatial and spectral processing, is given in [4]. Data sets include 1000 x 1000 pixels of 30 m spatial resolution and 273 bands in the spectral range between 420 and 2450 nm.

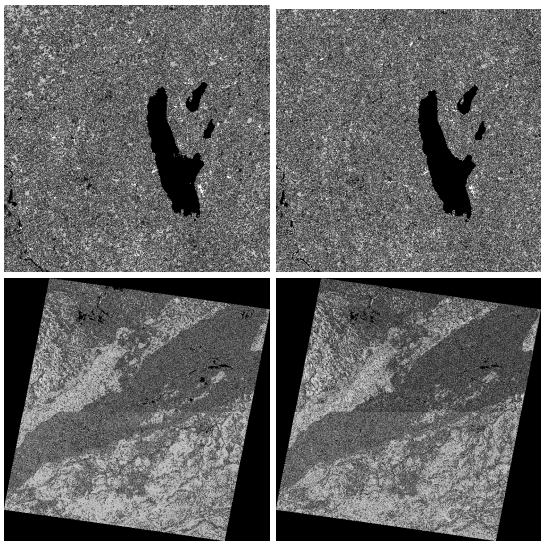

Figure 3 Maps for *DryVeg* (bottom) and *LAI* (top) showing reference (left) and SVR estimation (right)

The first image, Alpine Foreland, is based on Landsat-5 TM data from 06 June 2006 acquired in Southern Germany. Biophysical parameters, e.g. leaf area index (LAI) or chlorophyll content, were derived by radiative transfer models. Using these parameters as input, the EnMAP data was forward simulated [9].

The second image, *Makhtesh Ramon*, was generated by spectral mixing using SPOT-5 data, a mineral distribution map and a spectral data base [4].

By using simulated data, full spatial coverage of quantitative parameters was available. We chose to model LAI values (Alpine Foreland) and dry vegetation (DryVeg) fraction (Makhtesh Ramon) using the respective image. For each image, 1,000 training and 10,000 validation examples were drawn independently from the entire scenes.

Prior to the training, the hyperspectral data was linearly scaled in the range 0 to 1. The SVR were trained using a radial basis function kernel $K(\mathbf{x}_i,\mathbf{x}_j) = \exp(-g\cdot|\mathbf{x}_i - \mathbf{x}_j|^2)$. Best parameters g for kernel and C for regularization in terms of MAE were selected via 2-dimensional grid search using g, $C = 0.1, 1, 10, 100, 1000$. In this step we used $\varepsilon = 0$.

We used the *imageSVM* 2.0 [12] implementation for performing the grid search and the mapping of full EnMAP scenes. *imageSVM* uses the LibSVM [2] tool for solving the SVR optimization problem. We testet different numeric stopping tolerances with 0.0001 and 0.1 to see if faster convergence leads to comparable monitoring results.

As explained in Section 3, we monitored the performance of simplified SVR along the percentile i of the deviations distributen. The performance in terms of MAE and R^2 was estimated from independently drawn validation examples.

Table 1 Statistics and selected parameters for trained SVR

	Performance		**Selected Parameters**		**Full Scene Mapping**
	MAE	R^2	g	C	Time [min]
LAI	0.1601	0.9855	0.1	1000	62
DryVeg	0.0264	0.9623	0.1	10	66

Table 2 Statistics of simplified SVR: faster full scene mapping under acceptable performance decrease.

	Performance decrease	**Percentile**	**Number of SV**	**Full Scene Mapping***
LAI	R^2 - 0.01	76	107	6.7 (9.3)
	R^2 - 0.05	92	60	3.8 (16.6)
	MAE + 0.1	68	136	8.5 (7.3)
	MAE + 0.5	92	60	3.8 (16.6)
DryVeg	R^2 - 0.01	72	75	4.9 (13.3)
	R^2 - 0.05	94	50	3.3 (20.2)
	MAE + 0.01	70	76	5.0 (13.2)
	MAE + 0.05	96	41	2.7 (24.4)

* time in minutes, speed up factor in brackets

5. RESULTS

Table 1 shows statistics for trained SVR. The estimated performance in terms of R^2 is always above 0.96. For the DryVeg case, the MAE of estimated dry vegetation fraction is under 3 % and the estimated LAI is estimated with an MAE of around 0.16. Because SVR was trained with $\varepsilon = 0$, all training vectors become SV. Visual comparison to the reference after mapping complete EnMAP scenes showed no distinctive differences (Fig. 3). Figure 4 visualizes the dependency between performance of simplified SVR and ε-loss. The range of sampled values for ε-loss, which are corresponding to the percentiles of the deviation distribution, resolves the full spectrum of possible approximation quality. At the lower bound, the 0^{th} percentile leads to a perfect approximation with neither performance decrease nor reduction in SV. This situation complements at the 100^{th} percentile, where the number of SV is near to zero and performance is highly degenerated.

In between this range, the performance and the number of SV are both decreasing monotonically. Note that the performance of simplified SVR is comparable to that of original SVR for a long range, while the number of SV was significantly reduced already. Different degrees of accepted performance decrease, leading to simplified SVR, are up to 24 times faster in mapping full EnMAP scene (Tab. 2).

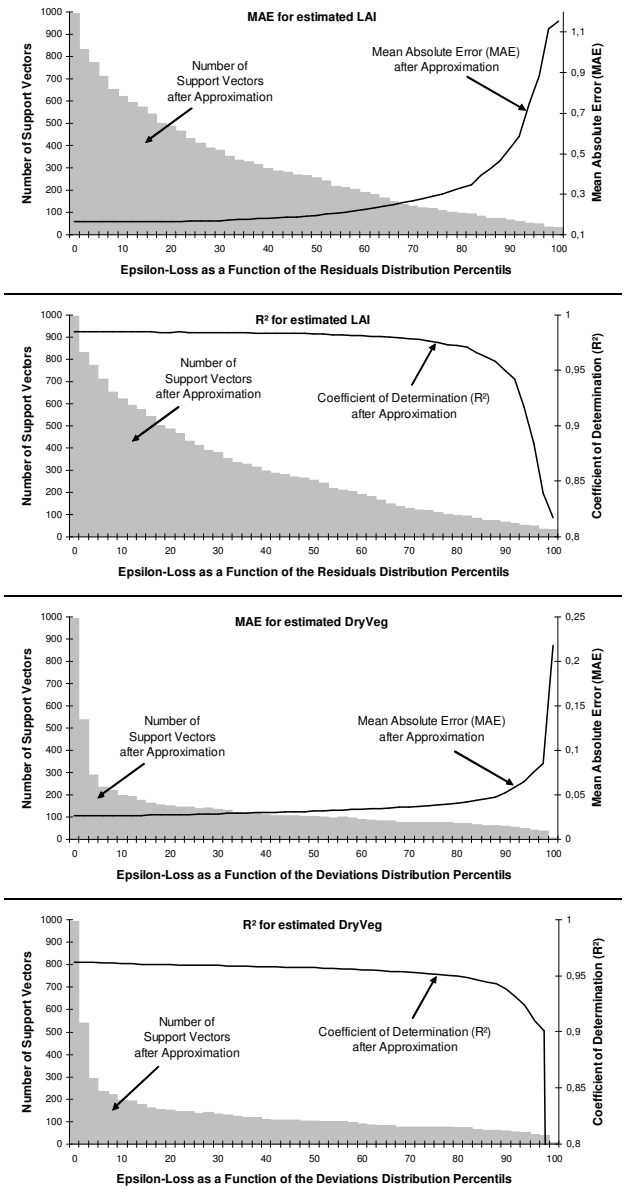

Figure 4 Dependency between performance (MAE and R^2) of simplified SVR and ε-loss. Note that ε-loss is sampled along the percentiles of the deviation distribution.

6. CONCLUSION

We presented a method and empirical results for monitoring the performance of simplified SVR by adapting the approach of Osuna and Girosi for SVC. It shows that for growing values of ε-loss, which we extract from the percentiles of the deviation distribution of the original SVR, the performance decreases monotonically between perfect and highly degenerated approximation. For practical and efficient implementation purposes we recommend a binary search for optimal ε-loss on values corresponding to the percentiles, with user-defined accepted performance decrease. We empirically found that faster convergence can be achieved by (i) using the same regularization parameter *C* for SVR training and approximation and (ii) relaxing the stopping tolerance for solving the SVR optimation problem during the search without affecting the solution (results not shown).

7. ACKNOWLEDGEMENTS

This work was supported by the German Federal Ministry of Economic Affairs and Technology within the EnMAP mission. EnMAP data was provided by K. Segl, L. Guanter (both GeoForschungsZentrum German Research Centre for Geosciences, Postdam) and H. Bach, Vista GmbH, Munich.

8. REFERENCES

[1] Camps-Valls, G., L. Bruzzone, et al. (2006). "Robust support vector regression for biophysical variable estimation from remotely sensed images." *IEEE Geosc. Rem. Sens. Let.* 3: 339-343.

[2] Chang, C.-C. and C.-J. Lin (2001), "LIBSVM : a library for support vector machines." *Software available at* http://www.csie.ntu.edu.tw/~cjlin/libsvm

[3] Durbha, S., R. King, et al. (2007). "Support vector machines regression for retrieval of leaf area index from multiangle imaging spectroradiometer." *Remote Sens. Env.* 107(1-2): 348-361.

[4] Guanter, L., K. Segl, et al. (2009), "Simulation of Optical Remote-Sensing Scenes With Application to the EnMAP Hyperspectral Mission." *IEEE Trans. on Geosci. and Remote Sens.*, accepted.

[5] Osuna, E. and F. Girosi (1998). "Reducing the run-time complexity of Support Vector Machines." *Proceedings of the International Conference on Pattern Recognition.*

[6] Peng, D. Q., J. Liu, et al. (2006). "Transformation model estimation of image registration via least square support vector machines." *Pattern Recognition Letters* 27(12): 1397-1404.

[7] Smola, A., N. Murata, et al. (1998). "Asymptotically Optimal Choice of ε-Loss for Support Vector Machines." *Proceedings of the Eighth Intern. Conference on Artificial Neural Networks.*

[8] Vapnik, V., *The Nature of Statistical Learning Theory.* Springer, 1999

[9] Verhoef, W. and H. Bach (2007). "Coupled soil-leaf-canopy and atmosphere radiative transfier modeling to simulate hyperspectral multi-angular surface reflectance and TOA radiance data." *Remote Sens. Environ.* 109(2): 166-182.

[10] Walton, J. T. (2008). "Subpixel urban land cover estimation: Comparing Cubist, Random Forests, and support vector regression." *Photogr. Eng. Remote Sens.* 74(10): 1213-1222.

[11] Xie, X. S., W. T. Liu, et al. (2008). "Spacebased estimation of moisture transport in marine atmosphere using support vector regression." *Remote Sensing of Environment* 112(4): 1846-1855.

[12] Rabe, A., van der Linden, S. (2009). imageSVM 2.0. *Software available at* www.hu-geomatics.de.