Updates for Nonlinear Discriminants

Edin Andelić Martin Schafföner Marcel Katz Sven E. Krüger Andreas Wendemuth
Cognitive Systems Group, IESK
Otto-von-Guericke-University
Magdeburg, Germany
edin.andelic@e-technik.uni-magdeburg.de

Abstract

A novel training algorithm for nonlinear discriminants for classification and regression in Reproducing Kernel Hilbert Spaces (RKHSs) is presented. It is shown how the overdetermined linear least-squares-problem in the corresponding RKHS may be solved within a greedy forward selection scheme by updating the pseudoinverse in an order-recursive way. The described construction of the pseudoinverse gives rise to an update of the orthogonal decomposition of the reduced Gram matrix in linear time. Regularization in the spirit of Ridge regression may then easily be applied in the orthogonal space. Various experiments for both classification and regression are performed to show the competitiveness of the proposed method.

1 Introduction

Models for regression and classification that enforce square loss functions are closely related to Fisher discriminants [Duda and Hart, 1973]. Fisher discriminants are Bayes-optimal in case of classification with normally distributed classes and equally structured covariance matrices [Duda and Hart, 1973][Mika, 2002]. However, in contrast to SVMs, Least Squares Models (LSMs) are not sparse in general and hence may cause overfitting in a supervised learning scenario. One way to circumvent this problem is to incorporate regularization controlled by a continuous parameter into the model. For instance, Ridge regression [Rifkin *et al.*, 2003] penalizes the norm of the solution yielding flat directions in the RKHS, which are robust against outliers caused by e. g. noise. In [Suykens and Vandewalle, 1999] Least-Squares SVMs (LS-SVMs) are introduced which are closely related to Gaussian processes and Fisher discriminants. A linear set of equations in the dual space is solved using e. g. the conjugate gradient methods for large data sets or a direct method for a small number of data. The solution is pruned [De Kruif and De Vries, 2003][Hoegaerts *et al.*, 2004] in a second stage. The close relation between the LS-SVM and the Kernel Fisher Discriminant (KFD) was shown in [Van Gestel *et al.*, 2002]. It follows from the equivalence between the KFD and a least squares regression onto the labels [Duda and Hart, 1973][Mika, 2002] that the proposed method is closely related to the KFD and the LS-SVM. However, the proposed method imposes sparsity on the solution in a greedy fashion using subset selection like in [Billings and Lee, 2002][Nair *et al.*, 2002]. For the SVM case similar greedy approaches exist [G. Cauwenberghs, 2000][Ma *et al.*, 2003].

Especially in case of large data sets subset selection is a practical method. It aims to eliminate the most irrelevant or redundant samples. However, finding the best subset of fixed size is an NP-hard combinatorial search problem. Hence, one is restricted to suboptimal search strategies. Forward selection starts with an empty training set and adds sequentially one sample that is most relevant according to a certain criterion (e. g. the mean square error). In [Nair *et al.*, 2002] an external algorithm which is based on elementary Givens rotations is used to update the QR-decomposition of the reduced Gram matrix in order to construct sparse models. The Gram Schmidt orthogonalization is used in [Billings and Lee, 2002] and [Chen *et al.*, 1991] for the orthogonal decomposition of the Gram matrix. They also apply forward selection in a second step to obtain sparse models. This method is known as Orthogonal Least Squares (OLS). However, the OLS algorithm requires the computation and the storage of the full Gram matrix which is prohibitive for large datasets.

In this paper a very simple and efficient way for constructing LSMs in a RKHS within a forward selection rule with much lower memory requirements is presented. The proposed method exploits the positive definiteness of the Gram matrix for an order-recursive thin update of the pseudoinverse, which reveals to the best of our knowledge a novel kind of update rule for the orthogonal decomposition. The solution is regularized in a second stage using the Generalized Cross Validation to re-estimate the regularization parameter.

The remainder of this paper is organized as follows. In section 2, computationally efficient update rules for the pseudoinverse and the orthogonal decomposition are derived. In section 3, it is shown how the solution may be regularized. In section 4 some experimental results on regression and classification datasets are presented. Finally, a conclusion is given in section 5.

2 Update of the Orthogonal Decomposition

In a supervised learning problem one is faced with a training data set $\mathcal{D} = \{\mathbf{x}_i, y_i\}, i = 1 \dots M$. Here, $\mathbf{x}_i$ denotes an input vector of fixed size and y_i is the corresponding target value which is contained in $\mathbb{R}$ for regression or in $\{1, -1\}$ for binary classification. It is assumed that $\mathbf{x}_i \neq \mathbf{x}_j$, for $i \neq j$.

We focus on sparse approximations of models of the form

$$\hat{\mathbf{y}} = \mathbf{K}\boldsymbol{\alpha}. \tag{1}$$

The use of Mercer kernels $k(\cdot, \mathbf{x})$ [Mercer, 1909] gives rise to a symmetric positive definite Gram Matrix $\mathbf{K}$ with elements $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ defining the subspace of the RKHS in which learning takes place. The weight vector $\boldsymbol{\alpha} = \{b, \alpha_1, \ldots, \alpha_M\}$ contains a bias term b with a corresponding column $\mathbf{1} = \{1, \ldots, 1\}$ in the Gram matrix.

Consider the overdetermined least-squares-problem

$$\hat{\boldsymbol{\alpha}}_m = \underset{\boldsymbol{\alpha}_m}{\operatorname{argmin}} \|\mathbf{K}_m \boldsymbol{\alpha}_m - \mathbf{y}\|^2 \tag{2}$$

in the m-th forward selection iteration with the reduced Gram matrix $\mathbf{K}_m = [\mathbf{1}\ \mathbf{k}_1 \ldots \mathbf{k}_m] \in \mathbb{R}^{M \times (m+1)}$ where $\mathbf{k}_i = (k(\cdot, \mathbf{x}_1), \ldots, k(\cdot, \mathbf{x}_M))^T$, $i \in \{1, \ldots, m\}$ denotes one previously unselected column of the full Gram matrix. We denote the reduced weight vector as $\boldsymbol{\alpha}_m = \{b, \alpha_1, \ldots, \alpha_m\} \in \mathbb{R}^{m+1}$ and the target vector as $\mathbf{y} = (y_1, \ldots, y_M)^T$. Among all generalized inverses of $\mathbf{K}_m$ the pseudoinverse

$$\mathbf{K}_m^{\dagger} = (\mathbf{K}_m^T \mathbf{K}_m)^{-1} \mathbf{K}_m^T \tag{3}$$

is the one that has the lowest Frobenius norm [Ben-Israel and Greville, 1977]. Thus, the corresponding solution

$$\hat{\boldsymbol{\alpha}}_m = \mathbf{K}_m^{\dagger} \mathbf{y} \tag{4}$$

has the lowest Euclidean norm.

Partitioning $\mathbf{K}_m$ and $\boldsymbol{\alpha}_m$ in the form

$$\begin{aligned} \mathbf{K}_m &= [\mathbf{K}_{m-1} \mathbf{k}_m] && (5) \\ \boldsymbol{\alpha}_m &= (\boldsymbol{\alpha}_{m-1} \alpha_m)^T && (6) \end{aligned}$$

and setting $\alpha_m = \alpha_{m0} = const$, the square loss becomes

$$L(\boldsymbol{\alpha}_{m-1}, \alpha_{m0}) = \|\mathbf{K}_{m-1} \boldsymbol{\alpha}_{m-1} - (\mathbf{y} - \mathbf{k}_m \alpha_{m0})\|^2. \tag{7}$$

The minimum of (7) in the least-squares-sense is given by

$$\hat{\boldsymbol{\alpha}}_{m-1} = \mathbf{K}_{m-1}^{\dagger} (\mathbf{y} - \mathbf{k}_m \alpha_{m0}). \tag{8}$$

Inserting (8) into (7) yields

$$L(\alpha_{m0}) = \|(\mathbf{I} - \mathbf{K}_{m-1} \mathbf{K}_{m-1}^{\dagger}) \mathbf{k}_m \alpha_{m0} - (\mathbf{I} - \mathbf{K}_{m-1} \mathbf{K}_{m-1}^{\dagger}) \mathbf{y}\|^2 \tag{9}$$

with $\mathbf{I}$ denoting the identity matrix of appropriate size.

Note that the vector

$$\mathbf{q}_m = (\mathbf{I} - \mathbf{K}_{m-1} \mathbf{K}_{m-1}^{\dagger}) \mathbf{k}_m \tag{10}$$

is the residual corresponding to the least-squares regression onto $\mathbf{k}_m$. Hence, $\mathbf{q}_m$ is a nullvector if and only if $\mathbf{k}_m$ is a nullvector unless $\mathbf{K}$ is not strictly positive definite. To ensure strictly positive definiteness of $\mathbf{K}$, it is mandatory to add a small positive constant ε to the main diagonal of the full Gram matrix in the form $\mathbf{K} \rightarrow \mathbf{K} + \varepsilon \mathbf{I}$. Forward selection may then be performed using this strictly positive definite Gram matrix. In the following $\mathbf{k}_m \neq \mathbf{0}$ is assumed.

The minimum of (9) is met at

$$\hat{\alpha}_{m0} = \mathbf{q}_m^{\dagger} (\mathbf{I} - \mathbf{K}_{m-1} \mathbf{K}_{m-1}^{\dagger}) \mathbf{y} \tag{11}$$

Noting that the pseudoinverse of a vector is given by

$$\mathbf{q}_m^{\dagger} = \frac{\mathbf{q}_m^T}{\|\mathbf{q}_m\|^2} \tag{12}$$

equation (11) may be written as

$$\begin{aligned} \hat{\alpha}_{m0} &= \frac{\mathbf{q}_m^T (\mathbf{I} - \mathbf{K}_{m-1} \mathbf{K}_{m-1}^{\dagger}) \mathbf{y}}{\|\mathbf{q}_m\|^2} && (13) \\ &= \frac{\mathbf{k}_m^T (\mathbf{I} - \mathbf{K}_{m-1} \mathbf{K}_{m-1}^{\dagger})^T (\mathbf{I} - \mathbf{K}_{m-1} \mathbf{K}_{m-1}^{\dagger}) \mathbf{y}}{\|\mathbf{q}_m\|^2}. \end{aligned}$$

The matrix

$$\mathbf{P}_m = \mathbf{I} - \mathbf{K}_{m-1} \mathbf{K}_{m-1}^{\dagger} \tag{14}$$

is an orthogonal projection matrix which implies being symmetric and idempotent and thus equation (13) simplifies to

$$\hat{\alpha}_{m0} = \mathbf{q}_m^{\dagger} \mathbf{y}. \tag{15}$$

Combining (15) with (8) the current weight vector $\hat{\boldsymbol{\alpha}}_m$ may be updated as

$$\hat{\boldsymbol{\alpha}}_m = \begin{bmatrix} \hat{\boldsymbol{\alpha}}_{m-1} \\ \hat{\alpha}_{m0} \end{bmatrix} = \begin{bmatrix} \mathbf{K}_{m-1}^{\dagger} - \mathbf{K}_{m-1}^{\dagger} \mathbf{k}_m \mathbf{q}_m^{\dagger} \\ \mathbf{q}_m^{\dagger} \end{bmatrix} \mathbf{y} \tag{16}$$

revealing the update

$$\mathbf{K}_m^{\dagger} = \begin{bmatrix} \mathbf{K}_{m-1}^{\dagger} - \mathbf{K}_{m-1}^{\dagger} \mathbf{k}_m \mathbf{q}_m^{\dagger} \\ \mathbf{q}_m^{\dagger} \end{bmatrix} \tag{17}$$

for the current pseudoinverse.

Since every projection $\mathbf{q}_m = \mathbf{P}_m \mathbf{k}_m$ lies in a subspace which is orthogonal to $\mathbf{K}_{m-1}$ it follows immediately that $\mathbf{q}_i^T \mathbf{q}_j = 0$, for $i \neq j$. Hence, an orthogonal decomposition

$$\mathbf{K}_m = \mathbf{Q}_m \mathbf{U}_m \tag{18}$$

of the reduced Gram matrix is given by the orthogonal matrix

$$\mathbf{Q}_m = [\mathbf{Q}_{m-1} \mathbf{q}_m] \tag{19}$$

and the upper triangular matrix

$$\mathbf{U}_m = \left[\begin{pmatrix} \mathbf{U}_{m-1} \\ \mathbf{0}_{m-1}^T \end{pmatrix} (\mathbf{Q}_m^T \mathbf{Q}_m)^{-1} \mathbf{Q}_m^T \mathbf{k}_m \right]. \tag{20}$$

In the m-th iteration $\mathcal{O}(Mm)$ operations are required for all these updates. Note that the inversion of the matrix $\mathbf{Q}_m^T \mathbf{Q}_m$ is trivial since this matrix is diagonal. However, the condition number of the matrix $\mathbf{Q}_m$ increases as the number of selected columns m grows. Thus, to ensure numerical stability it is important to monitor the condition number of this matrix and to terminate the iteration if the condition number exceeds a predefined value unless another stopping criterion is reached earlier.

3 Regularization and Selection of Basis Centers

The goal of every forward selection scheme is to select the columns of the Gram matrix that provide the greatest reduction of the residual. Methods like basis matching pursuit [Mallat and Zhang, 1993], order-recursive matching pursuit [Natarajan, 1995] or probabilistic approaches [Smola and

Schölkopf, 2000] are several contributions to this issue. In [Nair *et al.*, 2002], forward selection is performed by simply choosing the column that corresponds to the entry with the highest absolute value in the current residual. The reasoning is that the residual provides the direction of the maximum decrease in the cost function $0.5\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} - \boldsymbol{\alpha}^{\mathbf{T}}\mathbf{y}$, since the Gram matrix is strictly positive definite. The latter method is used in the following experiments. But note that the derived algorithm may be applied within any of the above forward selection rules. In the following we will refer to the proposed method as Order-Recursive Orthogonal Least Squares (OROLS).

Consider the residual

$$\tilde{\mathbf{e}}_m = \mathbf{y} - \hat{\mathbf{y}}_m = \mathbf{y} - \mathbf{Q}_m\tilde{\boldsymbol{\alpha}}_m \tag{21}$$

in the m-th iteration. The vector $\tilde{\boldsymbol{\alpha}}_m$ contains the orthogonal weights.

The regularized square residual is given by

$$\begin{aligned}\tilde{E}_m &= \tilde{\mathbf{e}}_m^T\tilde{\mathbf{e}}_m + \lambda\tilde{\boldsymbol{\alpha}}_m^T\tilde{\boldsymbol{\alpha}}_m \\ &= \mathbf{y}^T\tilde{\mathbf{P}}_m\mathbf{y}\end{aligned} \tag{22}$$

where λ denotes a regularization paramter. The minimum of (22) is given by

$$\begin{aligned}\tilde{\mathbf{P}}_m &= \mathbf{I} - \mathbf{Q}_m(\mathbf{Q}_m^T\mathbf{Q}_m + \lambda\mathbf{I}_m)^{-1}\mathbf{Q}_m^T \\ &= \tilde{\mathbf{P}}_{m-1} - \frac{\mathbf{q}_m\mathbf{q}_m^T}{\lambda + \mathbf{q}_m^T\mathbf{q}_m}.\end{aligned} \tag{23}$$

Thus, the current residual corresponding to the regularized least squares problem may be updated as

$$\begin{aligned}\tilde{\mathbf{e}}_m &= (\tilde{\mathbf{P}}_{m-1} - \frac{\mathbf{q}_m\mathbf{q}_m^T}{\lambda + \mathbf{q}_m^T\mathbf{q}_m})\mathbf{y} \\ &= \tilde{\mathbf{e}}_{m-1} - \mathbf{q}_m\frac{\mathbf{y}^T\mathbf{q}_m}{\lambda + \mathbf{q}_m^T\mathbf{q}_m}.\end{aligned} \tag{24}$$

The orthogonal weights

$$(\tilde{\boldsymbol{\alpha}}_m)_i = \frac{\mathbf{y}^T\mathbf{q}_i}{\lambda + \mathbf{q}_i^T\mathbf{q}_i}, \quad 1 \le i \le m. \tag{25}$$

can be computed when the forward selection is stopped. The original weights can then be recovered by

$$\hat{\boldsymbol{\alpha}}_m = \mathbf{U}_m^{-1}\tilde{\boldsymbol{\alpha}}_m \tag{26}$$

which is an easy inversion since $\mathbf{U}_m$ is upper triangular.

In each iteration one chooses the $\mathbf{q}_i$ which corresponds to the highest absolute value in the current residual and adds it to $\mathbf{Q}_{m-1}$. It is possible to determine the number of basis functions using crossvalidation or one may use for instance the Bayesian Information Criterion or the Minimum Description Length as alternative stopping criteria. Following [Gu and Wahba, 1991] and [Orr, 1995] it is possible to use the Generalized Cross Validation (GCV) as a stopping criterion. We will now summarize the results. For details see [Orr, 1995].

The GCV is given by

$$GCV_m = \frac{1}{M}\frac{\|\tilde{\mathbf{P}}_m\mathbf{y}\|^2}{\Big((1/M)\ \text{trace}(\tilde{\mathbf{P}}_m)\Big)^2}. \tag{27}$$

Minimizing the GCV with respect to λ gives rise to a re-estimation formula for λ. An alternative way to obtain a re-estimation of λ is to maximize the Bayesian evidence [MacKay, 1992].

Differentiating (27) with respect to λ and setting the result to zero gives a minimum when

$$\mathbf{y}^T\tilde{\mathbf{P}}_m\frac{\partial\tilde{\mathbf{P}}_m\mathbf{y}}{\partial\lambda}\text{trace}(\tilde{\mathbf{P}}_m) = \mathbf{y}^T\tilde{\mathbf{P}}_m^2\mathbf{y}\frac{\partial\text{trace}(\tilde{\mathbf{P}}_m)}{\partial\lambda}. \tag{28}$$

Noting that

$$\mathbf{y}^T\tilde{\mathbf{P}}_m\frac{\partial\tilde{\mathbf{P}}_m\mathbf{y}}{\partial\lambda} = \lambda\tilde{\boldsymbol{\alpha}}_m^T(\mathbf{Q}_m^T\mathbf{Q}_m + \lambda\mathbf{I}_m)^{-1}\tilde{\boldsymbol{\alpha}}_m \tag{29}$$

equation (28) can be rearranged to obtain the re-estimation formula

$$\lambda := \frac{[\partial\text{trace}(\tilde{\mathbf{P}}_m)/\partial\lambda]\mathbf{y}^T\tilde{\mathbf{P}}_m^2\mathbf{y}}{\text{trace}(\tilde{\mathbf{P}}_m)\tilde{\boldsymbol{\alpha}}_m^T(\mathbf{Q}_m^T\mathbf{Q}_m + \lambda\mathbf{I}_m)^{-1}\tilde{\boldsymbol{\alpha}}_m} \tag{30}$$

where

$$\frac{\partial\text{trace}(\tilde{\mathbf{P}}_m)}{\partial\lambda} = \sum_{i=1}^{m}\frac{\mathbf{q}_i^T\mathbf{q}_i}{(\lambda + \mathbf{q}_i^T\mathbf{q}_i)^2}. \tag{31}$$

The forward selection is stopped when λ stops changing significantly.

The computational cost for this update is $\mathcal{O}(m)$. The OROLOS algortihm is summarized in pseudocode in Algorithm 1.

4 Experiments

To show the usefulness of the proposed method empirically, some experiments for regression and classification are performed. In all experiments the Gaussian kernel

$$k(\mathbf{x}, \mathbf{x}^{'}) = exp\left(-\frac{\|\mathbf{x} - \mathbf{x}^{'}\|^2}{2\sigma^2}\right) \tag{32}$$

is used. The kernel parameter σ is optimized using a 5-fold crossvalidation in all experiments. For the classification experiments the one-vs-rest approach is used to obtain a multi-class classification hypothesis.

4.1 Classification

For classification, 5 well-known benchmark datasets were chosen. The USPS dataset contains 256 pixel values of handwritten digits as training and testing instances.

The letter dataset contains 20000 labeled samples. The character images were based on 20 different fonts and each letter within these fonts was randomly distorted to produce a dataset of unique stimuli. For this dataset no predefined split for training and testing exist. We used the first 16000 instances for training and the remaining 4000 instances for testing.

Optdigits is a database of digits handwritten by Turkish writers. It contains digits written by 44 writers. The training set is generated from the first 30 writers and digits written by the remaining independent writers serve as testing instances. The database was generated by scanning and processing forms to obtain 32 × 32 matrices which were then reduced to 8 × 8.

Algorithm 1 Order-Recursive Orthogonal Least Squares (OROLS)

```
Require: Training data $\mathbf{X}$, labels $\mathbf{y}$, kernel
  Initializations: $\lambda \leftarrow 0$, $m \leftarrow 1$, $\mathbf{K}_1 = \mathbf{1}$, $\mathbf{K}_1^\dagger = \frac{1}{M}\mathbf{1}^T$, $\mathbf{Q}_1 = \mathbf{1}$, $\mathbf{U}_1 = [1]$, $I = \{1, \dots, M\}$, $I_{opt} = \{\}$

  while $\lambda$ changes significantly and $\mathbf{Q}_m$ is not illconditioned do
     Update $\tilde{\mathbf{e}}_m$

     find the index $i_{opt}$ of the entry of $\tilde{\mathbf{e}}_m$ with the highest absolute value

     $I_{opt} \leftarrow \{I_{opt}, i_{opt}\}$

     $I \leftarrow I \setminus \{i_{opt}\}$

     Compute $\mathbf{k}_{i_{opt}}$
     Compute $\mathbf{q}_{i_{opt}}$
     $\mathbf{K}_m \leftarrow [\mathbf{K}_{m-1}\mathbf{k}_{opt}]$

     $\mathbf{Q}_m \leftarrow [\mathbf{Q}_{m-1}\mathbf{q}_{opt}]$

     Update $\mathbf{K}_m^\dagger$ and $\mathbf{U}_m$ using $\mathbf{k}_{opt}$ and $\mathbf{q}_{opt}$

     Update $\lambda$

     $m \leftarrow m + 1$

  end while
  return $\hat{\boldsymbol{\alpha}}_m, I_{opt}$
```

Pendigits contains pen-based handwritten digits. The digits were written down on a touch-sensitive tablet and were then resampled and normalized to a temporal sequence of eight pairs of (x, y) coordinates. The predefined test set is formed entirely from written digits produced by independent writers.

The satimage dataset was generated from Landsat Multi-Spectral Scanner image data. Each pattern contains 36 pixel values and a number indicating one of the six classes of the central pixel.

The caracteristics of the datasets are summarized in table 1. The results can be seen in table 2. Especially for the optdigits and pendigits datasets OROLS appears to be significantly superior compared with SVMs. The performance on the remaining 3 datasets is comparable with SVMs.

DATA SET	# CLASSES	# TRAINING	# TESTING
USPS	10	7291	2007
LETTER	26	16000	4000
OPTDIGITS	10	3823	1797
PENDIGITS	10	7494	3498
SATIMAGE	6	4435	2000

Table 1: Datasets used for the classification experiments.

DATA SET	*SVM*	*OROLS*
USPS	4.3	4.4(10)
LETTER	2.75	2.61(4.3)
OPTDIGITS	2.73	1.11(10.2)
PENDIGITS	2.5	1.66(1.9)
SATIMAGE	7.8	8.2(7.5)

Table 2: Test errors in % on 5 benchmark datasets. The one-vs-rest approach is used. Average fraction of selected basis centers in % within parantheses.

4.2 Regression

For regression, we first perform experiments on a synthetic dataset based on the function $\text{sinc}(x) = \sin(x)/x, \ x \in (-10, 10)$ which is corrupted by Gaussian noise. All training and testing instances are chosen randomly using a uniform distribution on the same interval. The results are illustrated in figures 1-3 and table 3.

Additionally, the two real world datasets Boston and Abalone, which are available from the UCI machine learning repository, are chosen. The hyperparameters are optimized in a 5-fold crossvalidation procedure. For both datasets, random partitions of the mother data for training and testing are generated (100 (10) partitions with 481 (3000) instances for training and 25 (1177) for testing for the Boston and Abalone dataset, respectively). All continuous features are rescaled to zero mean and unit variance for both Abalone and Boston. The gender encoding (male / female /infant) for the Abalone dataset is mapped into $\{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$. The Mean Squared Error (MSE) of OROLS is compared with a forward selection algorithm based on a QR-decomposition of the Gram matrix [Nair *et al.*, 2002]. The results in table 4 show that the MSE is improved significantly by OROLS. In contrast to OROLS the QR method uses an external algorithm with reorthogonalization for the update of the orthogonal decomposition. We observed that our update scheme which is to the best of our knowledge a novel update needs not to be reorthogonalized as long as the Gram matrix has full rank. This improvement of accuracy could be one reason for the good performance of OROLS. Furthermore, it should be noted that the best performance of OROLS for the Boston dataset is quite favourable compared with the best performance of SVMs (MSE 8.7 ± 6.8) [Schölkopf and Smola, 2002].

METHOD	RMSE
SVM	0.0519
RVM	0.0494
OROLS	0.0431

Table 3: Average RMSE for the sinc experiment. 50 / 1000 randomly generated points are used for training / testing. The standard deviation of the Gaussian noise is 0.1 in all runs. The results are avereged over 100 runs.

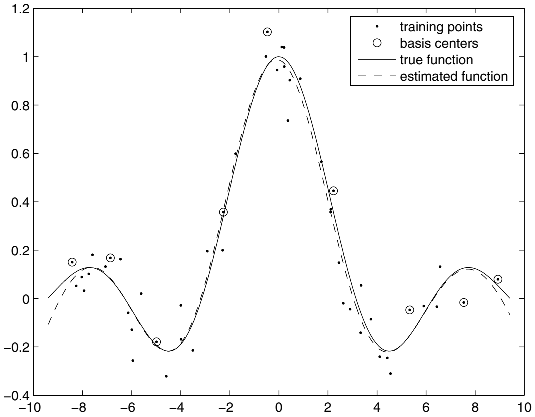

Figure 1: Example fit to a noisy sinc function using 50 / 1000 randomly generated points for training / testing. The standard deviation of the Gaussian noise is 0.1. The Root Mean Square Error (RMSE) is 0.0269 in this case. 9 points are selected as basis centers.

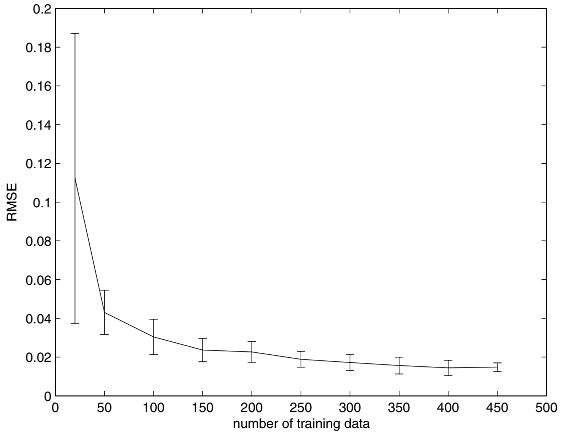

Figure 2: RMSE of fits to a noisy sinc function w. r. t. different training set sizes. 1000 randomly generated points are used for testing. The standard deviation of the Gaussian noise is 0.1 in all runs. The results are avereged over 100 runs for each size.

Dataset	QR	OROLS
Boston	8.35±5.67	7.9±3.28(26)
Abalone	4.53±0.29	4.32±0.17(10.8)

Table 4: Mean Square Error (MSE) with standard deviations for the Boston and Abalone dataset using different methods. Average fraction of selected basis centers in % within parantheses.

5 Conclusion

A computationally efficient training algorithm for orthogonal least squares models using Mercer kernels is presented.

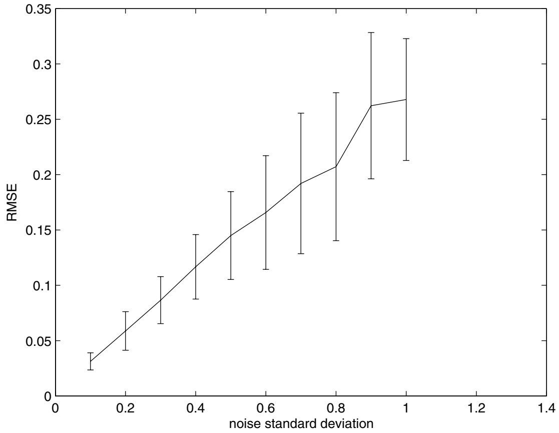

Figure 3: RMSE of fits to a noisy sinc function w. r. t. different noise levels. 100 / 1000 randomly generated points are used for training / testing. The results are avereged over 100 runs for each noise level.

The pseudoinverse is updated order-recursively and reveals the current orthogonal decomposition of the reduced Gram matrix within a forward selection scheme. The Generalized Cross Validation serves as an effective stopping criterion and allows to adapt the regularization parameter in each iteration. Extensive empirical studies using synthetic and real-world benchmark datasets for classification and regression suggest that the proposed method is able to construct models with a very competitive generalization ability. The advantage of the proposed method compared with e. g. SVMs is its simplicity. Sparsity is achieved in a computationally efficient way by constuction and can hence better be controlled than in the SVM case where a optimization problem is to be solved. Furthermore, in contrast to SVMs OROLS allows an easy incorporation of multiple kernels, i e. the kernel parameters may be varied for different training instances in order to obtain more flexible learning machines. This possibility is not examined here and may be an interesting direction for future work. A further step for future work could be the development of the proposed algorithm for tasks like dimensionality reduction or online-learning.

References

[Ben-Israel and Greville, 1977] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. Wiley, 1977.

[Billings and Lee, 2002] S. A. Billings and K. L. Lee. Nonlinear fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm. *Neural Networks*, 15:263–270, 2002.

[Chen *et al.*, 1991] S. Chen, C. F. N. Cowan, and P. M. Grant. Orthogonal least squares learning for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.

[De Kruif and De Vries, 2003] B. J. De Kruif and T. J. A. De Vries. Pruning error minimization in least squares support vector machines. *IEEE Transactions on Neural Networks*, 14(3):696–702, 2003.

[Duda and Hart, 1973] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. John Wiley and Sons, 1973.

[G. Cauwenberghs, 2000] T. Poggio G. Cauwenberghs. Incremental and decremental support vector machine learning. In *Advance in Neural Information Processing Systems (NIPS 2000)*, 2000.

[Gu and Wahba, 1991] C. Gu and G. Wahba. Minimizing gcv/gml scores with multiple smoothing parameters via the newton method. *SIAM Journal on Scientific and Statistical Computing*, 12(2):383–398, 1991.

[Hoegaerts *et al.*, 2004] L. Hoegaerts, J. A. K. Suykens, J. Vanderwalle, and B. De Moor. A comparison of pruning algorithms for sparse least squares support vector machines. In *Proceedings of the 11th International Conference on Neural Information Processing (ICONIP 2004)*, Calcutta, India, Nov 2004.

[Ma *et al.*, 2003] J. Ma, J. Theiler, and S. Perkins. Accurate on-line support vector regression. *Neural Computation*, 15:2683–2703, 2003.

[MacKay, 1992] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[Mallat and Zhang, 1993] S. Mallat and Z. Zhang. Matching pursuit in a time-frequency dictionary. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[Mercer, 1909] J. Mercer. Functions of positive and negative type and their connections to the theory of integral equations. In *Philos. Trans. Roy. Soc.*, pages A 209:415–446, London, 1909.

[Mika, 2002] S. Mika. *Kernel Fisher Discriminants*. PhD thesis, Technical University Berlin, 2002.

[Nair *et al.*, 2002] P. Nair, A. Choudhury, and A. J. Keane. Some greedy learning algorithms for sparse regression and classification with mercer kernels. *Journal of Machine Learning Research*, 3:781–801, 12 2002.

[Natarajan, 1995] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal of Computing*, 25:227–234, 1995.

[Orr, 1995] M. Orr. Regularisation in the selection of radial basis function centres. *Neural Computation*, 7:606–623, 1995.

[Rifkin *et al.*, 2003] R. Rifkin, G. Yeo, and T. Poggio. Regularized least squares classification. In J. A. K. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, editors, *Advances in Learning Theory: Methods, Models and Applications*, volume 190 of *NATO Science Series III: Computer and Systems Sciences*, chapter 7, pages 131–154. IOS Press, Amsterdam, 2003.

[Schölkopf and Smola, 2002] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[Smola and Schölkopf, 2000] A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 911–918. Morgan Kaufmann, 2000.

[Suykens and Vandewalle, 1999] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.

[Van Gestel *et al.*, 2002] T. Van Gestel, J. A. K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. A bayesian framework for least squares support vector machine classifiers, gaussian processes and kernel fisher discriminant analysis. *Neural Computation*, 14(5):1115–1147, May 2002.