Orthogonal Least Square with Boosting for Regression

S. Chen[†], X.X. Wang[‡] and D.J. Brown[‡]

[†] School of Electronics and Computer Science
University of Southampton, Southampton SO17 1BJ, U.K.
E-mail: sqc@ecs.soton.ac.uk

[‡] Department of Creative Technologies
University of Portsmouth, Portsmouth PO1 3HE, U.K.
E-mails: xunxian.wang@port.ac.uk, david.j.brown@port.ac.uk

Presented at 5th International Conference on Intelligent Data Engineering and Automated Learning, Exeter, U.K., August 25-27, 2004

Overview

Modeling from data: *generalization, interpretability, knowledge extraction* $\Rightarrow$ All depend on ability to construct appropriate sparse models

○ Existing sparse kernel regression modeling:

1) Orthogonal least squares forward selection construction

2) SVM type kernel modeling techniques

- Kernels position at training input data points with a common kernel variance

○ This contribution considers generalized kernel model with tunable kernel centers and covariance matrices

OLS forward selection: each stage of selection determines a kernel regressor using a guided random search optimization based on boosting

- Enhancing modeling capability with much sparser representation

Generalized Kernel Modeling

- Modeling training data set $\{\mathbf{x}_l, y_l\}_{l=1}^{N}$ with regression model

$$y(\mathbf{x}) = \hat{y}(\mathbf{x}) + e(\mathbf{x}) = \sum_{i=1}^{M} w_i g_i(\mathbf{x}) + e(\mathbf{x})$$

- Generalized kernel

$$g_i(\mathbf{x}) = G\left(\sqrt{(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i)}\right)$$

where $\boldsymbol{\mu}_i$ is kernel center and $\boldsymbol{\Sigma}_i$ diagonal kernel covariance matrix

- Regression model over training set

$$\mathbf{y} = \mathbf{G}\,\mathbf{w} + \mathbf{e}$$

where $\mathbf{y} = [y_1 \cdots y_N]^T$, $\mathbf{w} = [w_1 \cdots w_M]^T$, $\mathbf{e} = [e(\mathbf{x}_1) \cdots e(\mathbf{x}_N)]^T$ and

$$\mathbf{G} = [\mathbf{g}_1\ \mathbf{g}_2 \cdots \mathbf{g}_M] \quad \text{with} \quad \mathbf{g}_k = [g_k(\mathbf{x}_1)\ g_k(\mathbf{x}_2) \cdots g_k(\mathbf{x}_N)]^T$$

Orthogonal Decomposition

○ Orthogonal decomposition

$$\mathbf{G} = \mathbf{P}\mathbf{A}$$

where orthogonal matrix $\mathbf{P} = [\mathbf{p}_1 \ \mathbf{p}_2 \cdots \mathbf{p}_M]$ has orthogonal columns

○ Regression model becomes

$$\mathbf{y} = \mathbf{P}\boldsymbol{\theta} + \mathbf{e}$$

with $\boldsymbol{\theta} = \mathbf{A}\,\mathbf{w} = [\theta_1 \cdots \theta_M]^T$

○ Least squares cost over training set

$$J = \frac{1}{N}\mathbf{e}^T\mathbf{e} = \frac{1}{N}\mathbf{y}^T\mathbf{y} - \frac{1}{N}\sum_{i=1}^{M}\mathbf{p}_i^T\mathbf{p}_i\theta_i^2$$

○ Least squares cost for k-term subset model can be expressed recursively as

$$J_k = J_{k-1} - \frac{1}{N}\mathbf{p}_k^T\mathbf{p}_k\theta_k^2$$

Model Construction

○ Select model terms one by one to incrementally minimize least squares cost

○ Specifically, at k-stage of selection, determine k-th regressor's position $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ by minimizing J_k

$$\min_{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k} J_k\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

○ Procedure stops when

$$J_M < \xi$$

where ξ is a chosen tolerance, ending with an M-term model

○ We propose a guided random search to perform optimization

Alternative criteria, such as leave-one-out test error and optimal experiment design criteria, can be adopted here

Guided Random Search

Consider task of minimizing $f(\mathbf{u})$

Outer Loop: N_G number of generations

Initialization: keep best solution found in previous generation as $\mathbf{u}_1$ and randomly choose rest of population $\mathbf{u}_2, \cdots, \mathbf{u}_{P_S}$

Inner Loop: N_I iterations

- Perform a convex combination

$$\mathbf{u}_{P_S+1} = \sum_{i=1}^{P_S} \delta_i \mathbf{u}_i$$

- Weightings

$$\delta_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{P_S} \delta_i = 1$$

are adopted (boosting) to reflect goodness of $\mathbf{u}_i$

- $\mathbf{u}_{P_S+1}$ replaces worst member in population $\mathbf{u}_i$, $1 \leq i \leq P_S$

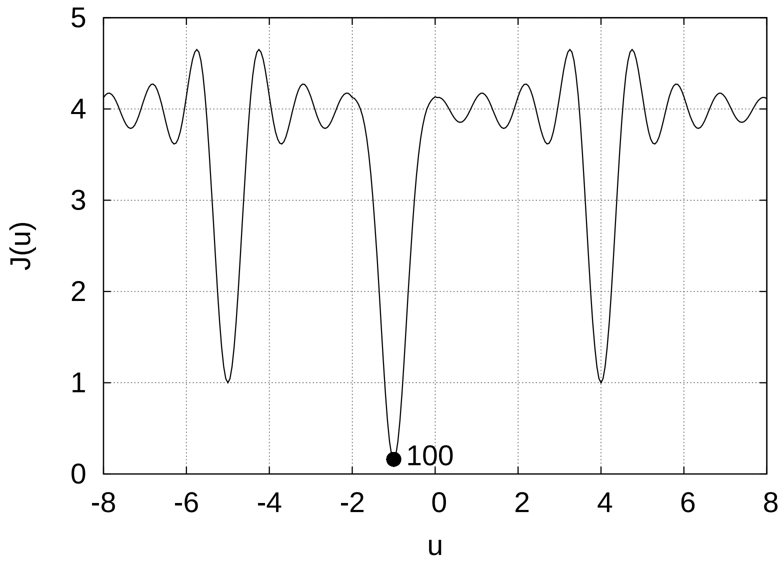
End of *Inner Loop*

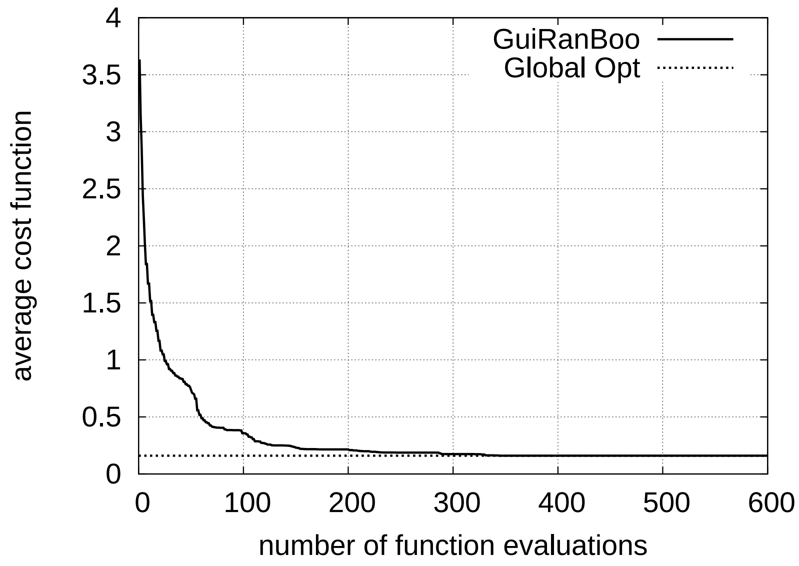
End of *Outer Loop*

Optimization Example

○ Population size $P_S = 6$, number of Inner iterations $N_I = 20$ and number of generations $N_G = 12$

○ 100 random experiments, populations of all 100 runs converge to global minimum





Simple Modeling Example

- 500 points of training data generated from

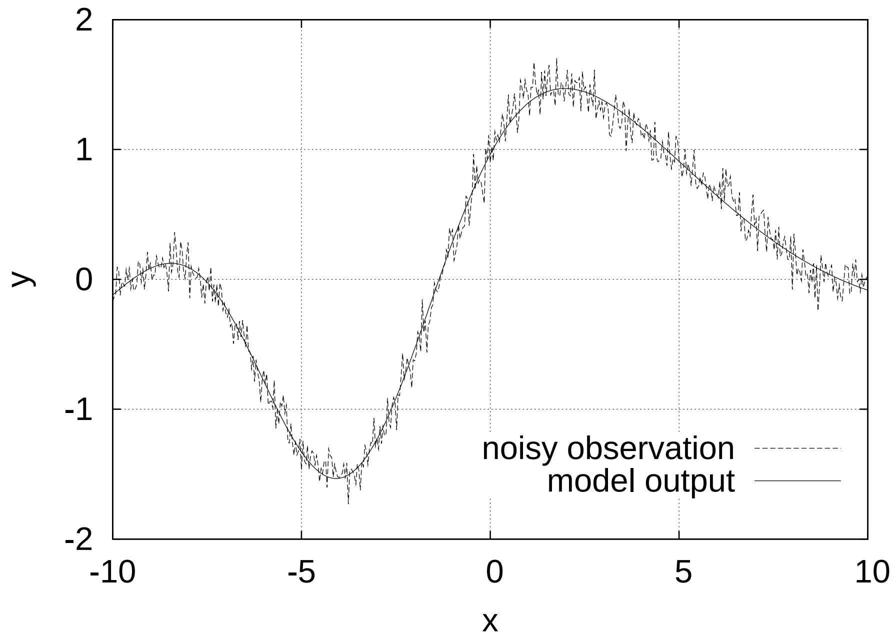
$$y(x) = 0.1x + \frac{\sin x}{x} + \sin 0.5x + \epsilon$$

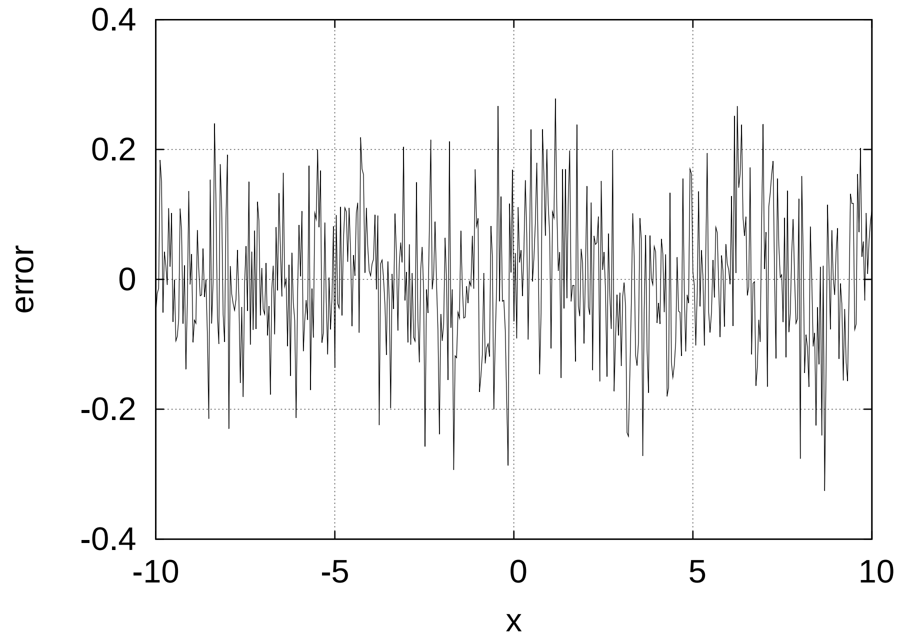
where $x \in [-10,\ 10]$ and ϵ Gaussian white noise of variance 0.01

- Generalized Gaussian kernel used, modeling accuracy set to $\xi = 0.012$:

regression step k	mean μ_k	variance σ_k^2	weight w_k	MSE J_k
0	–	–	–	0.8431
1	2.6911	4.2480	2.3527	0.3703
2	-4.0652	2.1710	-2.5197	0.0339
3	3.0314	2.0059	-1.0609	0.0172
4	-4.1771	1.0909	0.8982	0.0151
5	-1.9783	64.0000	0.1190	0.0129
6	6.6853	0.3894	0.1548	0.0118

Simple Modeling Example (continue)





Noisy training data $y(x)$, model output $\hat{y}(x)$ and modeling error $e(x) = y(x) - \hat{y}(x)$

Engine Data Modeling

◯ Modeling relationship between fuel rack position (input $u(t)$) and engine speed (output $y(t)$) for a Leyland TL11 turbocharged, direct injection diesel engine operated at low engine speed

◯ Data set contains 410 pairs of input-output samples (u_i, y_i), modeled as $y_i = f_s(\mathbf{x}_i) + \epsilon_i$ with $\mathbf{x}_i = [y_{i-1} \; u_{i-1} \; u_{i-2}]^T$; First 210 data points for training and last 200 points for testing

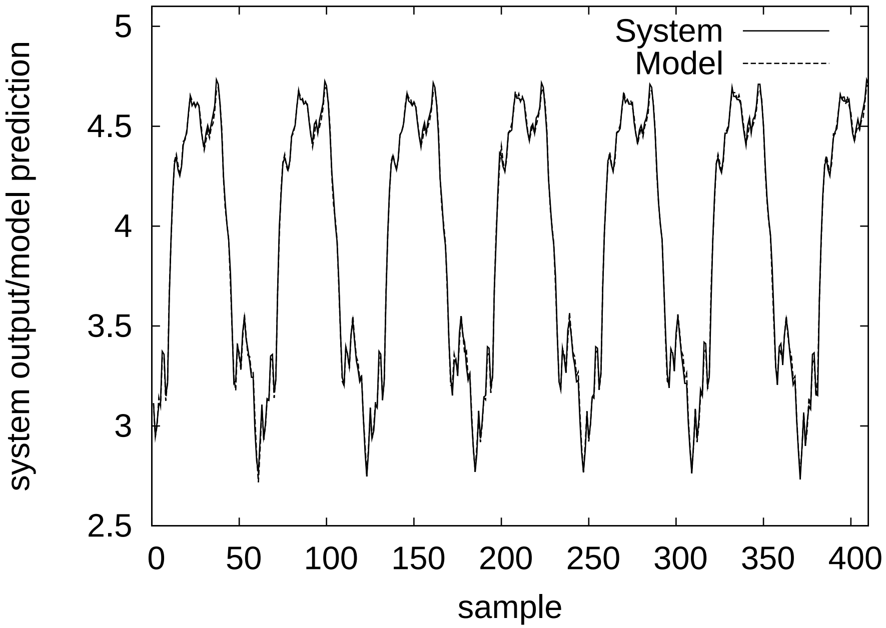
◯ Generalized Gaussian kernel used, modeling accuracy set to $\xi = 0.00055$:

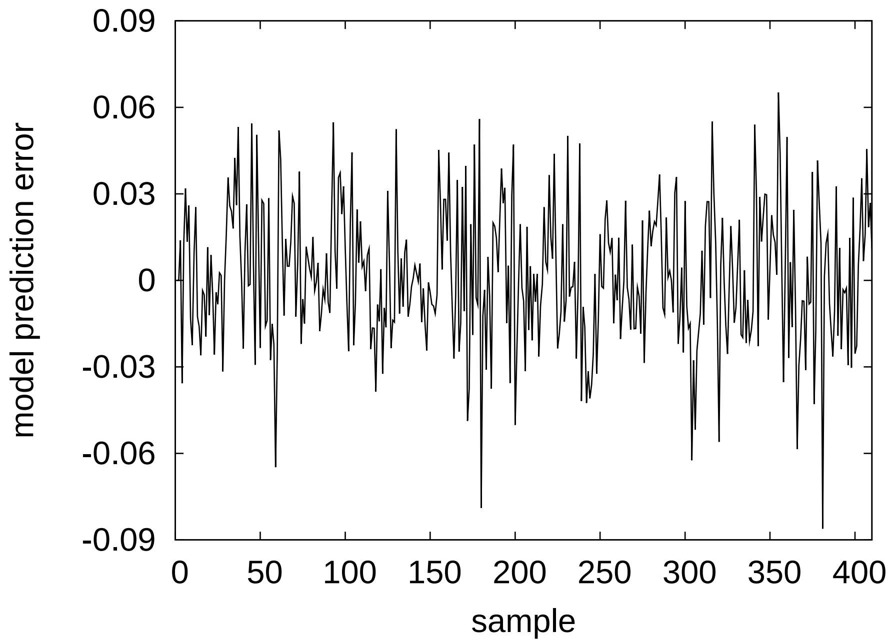
step k	mean vector $\boldsymbol{\mu}_k$			diagonal covariance $\boldsymbol{\Sigma}_k$			weight w_k	MSE $J_k \times 100$
0		–			–		–	1558.9
1	5.2219	5.5839	5.6416	7.3532	21.0894	22.4661	6.0396	0.3866
2	4.2542	5.2741	4.1028	1.8680	10.0863	49.8826	-1.2845	0.1311
3	3.8826	5.1707	6.3200	0.1600	0.1600	64.0000	-0.1539	0.0996
4	2.3154	3.2544	5.4897	0.9447	0.3329	11.7564	-0.1433	0.0913
5	4.0673	4.4276	3.5963	0.1608	18.3731	0.2207	0.1945	0.0740
6	2.3663	3.2377	5.1376	0.1754	0.9317	0.1600	0.9658	0.0547

Test MSE: 0.000573

◯ To achieve same modeling accuracy for this data set, existing state-of-art kernel regression techniques required at least 22 regressors

Engine Data Modeling (continue)





Noisy training data y_i, model output $\hat{y}_i$ and modeling error $e_i = y_i - \hat{y}_i$

Conclusions

- A novel construction algorithm has been proposed for parsimonious regression modeling based on OLS algorithm with boosting
- Proposed algorithm has ability to tune center and diagonal covariance matrix of individual regressor to incrementally minimize training mean square error
- A guided random search method has been developed to append regressors one by one in an orthogonal forward regression procedure
- Our method offers enhanced modeling capability with very sparse representation