Support Vector Regression Algorithms in the Forecasting of Daily Maximums of Tropospheric Ozone Concentration in Madrid

E. G. Ortiz-García, S. Salcedo-Sanz, A. M. Pérez-Bellido, J. Gascón-Moreno and A. Portilla-Figueras

Department Signal Theory and Comunications
Universidad de Alcalá, Madrid, Spain
HAIS 2010 Conference

Index

- Introduction
- SVMr Formulation
- Parameters Search Space
- Air Pollution Monitoring Network of Madrid
- Experiments and Results

Introduction

- Tropospheric Ozone (O_3)
  - Very important pollutant in urban areas
    - Increases the mortality rates
  - Produced by interaction of NOx and VOC
    - Influence of sunlight
- Several works
  - Concentration in a column or in a area
  - Different cities

Introduction

- Support Vector machines regression
  - One of the most important soft computing techniques
  - High quality in regression problems
  - Balance between error aproximation and generalization

$$\min_{w,\xi,\xi^*,b} \left(\frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{l} \xi_i + \xi_i^* \right)$$

$$y_i - \mathbf{w}^T \phi(\mathbf{x_i}) - b \leq \epsilon + \xi_i$$
$$-y_i + \mathbf{w}^T \phi(\mathbf{x_i}) + b \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

SVMr Formulation

- Dual Formulation

$$\max\left(-\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x_i},\mathbf{x_j})-\epsilon\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*)\right)$$

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \qquad \alpha_i,\alpha_i^*\in[0,C]$$

$$y(\mathbf{x})=f(\mathbf{x})+b \qquad f(\mathbf{x})=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)k(\mathbf{x_i},\mathbf{x})$$

Parameters Search Space

- C - Regularization parameter

$$C \leq \frac{y_i^{max} - b - \epsilon}{(1 - \frac{1}{l-1}\sum_{j=1, j\neq i}^{l} K(\mathbf{x_j}, \mathbf{x_i}))}$$

- ϒ - Kernel parameter

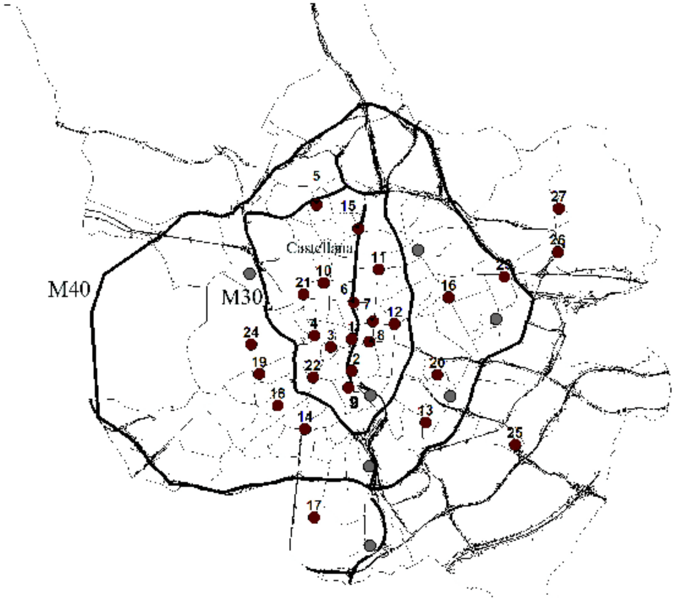
$$\gamma \leq -\frac{log_e(0.001)}{(\frac{1}{l}\sum_{i=1}^{l} \min_{j, i\neq j} d(\mathbf{x_j}, \mathbf{x_i}))^2}$$

- ε – Margin parameter

$$\epsilon < \sigma_y$$

Air Pollution Monitoring Network of Madrid

- Largest in Spain, one of largest in Europe
- 27 stations
- Data from 2002 to 2007
- 6 meteorological stations



Experiments and Results

- General description
  - Daily prediction of maxima concentration
  - Six years 2002-2007
  - 365 samples a year
  - Several train and test by dividing into 5 subsets
  - 30 experiments for statistical tests
    - Kolmogorov-Smirnov normality test
    - T-test
  - 5 chosen stations (highest concentrations)

Experiments and Results

- Dependence with solar radiation and temperature

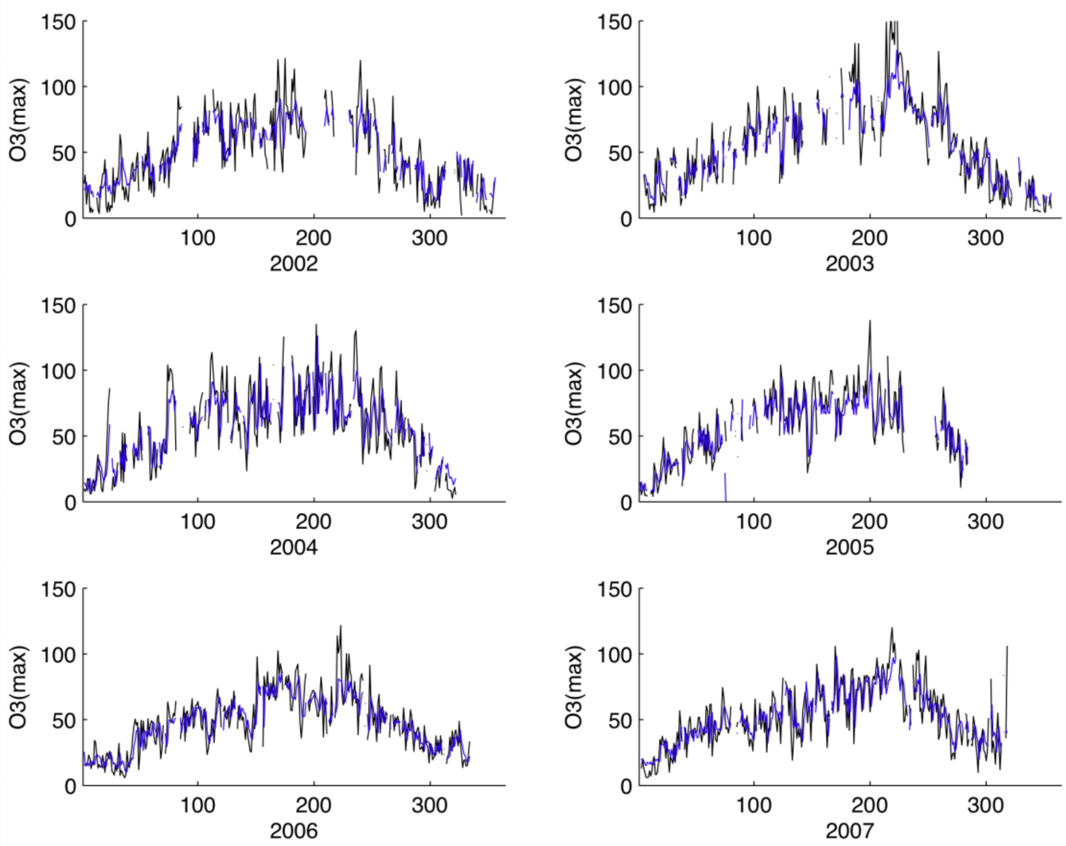
	None		Solar radiation		Temperature		Both	
Station	Mean	Std	Mean	Std	Mean	Std	Mean	Std
5	17.56	4.80	17.53	4.59	17.82	4.19	17.68	4.22
9	15.69	4.06	15.61	4.18	15.78	4.17	15.83	4.13
10	17.38	4.91	17.13	4.83	17.39	4.50	17.13	4.49
14	16.84	3.72	16.53	3.11	17.01	4.11	16.87	3.88
24	17.29	4.01	17.00	3.79	17.23	3.66	17.04	3.74

	Solar radiation		Temperature		Both	
Station	P-value	W-L-T	P-value	W-L-T	P-value	W-L-T
5	0.80*	15-15-0	0.26*	15-15-0	0.69*	19-11-0
9	0.65*	16-14-0	0.62*	15-15-0	0.53*	16-14-0
10	0.04*	21-9-0	0.97*	17-13-0	0.09*	19-11-0
14	0.22*	17-13-0	0.56*	15-15-0	0.92*	17-13-0
24	0.02*	18-12-0	0.71*	14-16-0	0.06*	18-12-0

* t-test $\alpha = 0.05$

Experiments and Results

- Dependence with solar radiation and temperature



Experiments and Results

- Comparison SVMr versus MLP
  - Multilayer Perceptron
    - Number of neurons from 6 to 20
    - Levenberg-Marquardt (20 repetitions)
    - Hold-out validation

Experiments and Results

- Comparison SVMr versus MLP

	MLP		SVMr		SVMr vs MLP	
Station	Mean	Std	Mean	Std	t-test	W-L-T
5	34.60	14.75	17.53	4.59	0.00*	29-1-0
9	32.90	16.12	15.61	4.18	0.00*	29-1-0
10	34.99	15.97	17.13	4.83	0.00*	28-2-0
14	31.58	14.13	16.53	3.11	0.00*	28-2-0
24	33.28	15.26	17.00	3.79	0.00*	29-1-0

* t-test α =0.05

Thank you for your attention!!