Application of Support Vector Regression to Scatterometer Data

L. Leslie, M. Richman, H. Mansouri, C. Shafer

School of Meteorology
School of Industrial Engineering
The University of Oklahoma

JCSDA 5^{th} Workshop on Satellite Data Assimilation
May 1-2 2007
Marriott Inn and Conference Center
University of Maryland University College
College Park MD

Outline

Brief History

- Vapnik (1974, 1979) created the idea of creating separating hyperplanes with optimal margin.
- In addition, important work in the context of reproducing kernels, related to SV methods, has been done by Wahba and co-workers.

Brief History

In the 1990's three ideas were developed that made support vectors useful to a large number of applications.

- Mercer kernels were used to generalize from optimal hyperplanes to nonplanar decision surfaces. This is done by nonlinearly mapping into some other (possibly high-dimensional) space.
- Optimal margin algorithm is generalized to non-separable problems by the introduction of slack variables in the statement of the optimization problem.
- Different SV classifiers constructed by using different kernels (polynomial , RBF, neural net) extract the same Support Vectors.

Example of the basic idea behind SV regression

Cortes, C.; and Vapnik, V. 1995. Support Vector Networks. Machine Learning 20:273-297.

- In this paper, the question asked was how many inputs will it take to separate a 2-class data set with a second-order polynomial boundary. We can simply employ p inputs of the form $x_1, x_2, \ldots, x_p$, and allow a nonlinear model to map those to a binary output.
- Alternately, we could employ a linear model that, in addition to the p inputs, also includes another p inputs of the form $x_1^2, x_2^2, \ldots, x_p^2$, plus another $\frac{p(p-1)}{2}$ inputs of the form $x_1x_2, x_1x_3, ..., x_px_{p-1}$.
- That amounts to a total of $\frac{p(p+3)}{2}$ inputs, feeding into a linear classifier.

Example of the basic idea behind SV regression

- The data are mapped into a larger-dimensional space where the decision boundary is a linear hyperplane. This is the main idea behind Support Vector Machines. One criterion for fitting the hyperplane through the data is to simply select the hyperplane that maximizes the geometric margin between the vectors of the two classes.
- The main advantage of this choice is that it requires only a portion of the training set, namely those closest to the hyperplane, often referred to as support vectors.
- In practice, SVMs actually construct a linear decision hyperplane first, and then proceed to map the results to a high-dimensional feature space.

Some Applications of Support Vectors to Radar Data I

Trafalis, T.B., A. White, B. Santosa and M.B. Richman, 2002.
Data mining techniques for improved WSR-88D rainfall estimation.
Computers and Industrial Engineering, 43, 775-786, 2002.

Trafalis, T. B., M. B. Richman and B. Santosa, 2002.
Prediction of rainfall from WSR-88D radar using support vector regression.
Intelligent Engineering Systems Through Artificial Neural Networks, ASME Press, 12, 639-644. (Novel Smart Engineering System Design Award).

Trafalis, T.B., B. Santosa and M.B. Richman, 2003.
Prediction of rainfall from WSR-88D radar using kernel-based methods.
International Journal of Smart Engineering System Design, 5, 429-438.

Trafalis, T.B., H. Ince and M.B. Richman, 2003.
Tornado detection with support vector machines.
Computational Science - ICCS 2003, Peter M.A., Sloot, David Abramson, A. Bogdanov, Jack J. Dongarra, A. Zomaya and Yuriy Gorbachev, eds., (Int'l conference Saint Petersburg Russian Federation, Melbourne Australia, June 2-4, 2003 Proceedings), Springer, 202 - 211.

Trafalis, T.B., B. Santosa and M.B. Richman, 2004.
Rule-based support vector machine classifiers applied to tornado prediction.
Computational Science-ICCS 2004, Lecture notes in Computer Science, series LNCS 3036, part III, Springer, 678-684.

Some Applications of Support Vectors to Radar Data II

- Trafalis, T.B., B. Santosa and M.B. Richman, 2005.
  Learning networks in rainfall estimation.
  Computational Management Science, 2, 229-251.
- Trafalis, T.B., B. Santosa and M.B. Richman, 2005.
  Feature selection with linear programming support vector machines and applications to tornado prediction.
  WSEAS Transactions on Computers, 4, 865-873.
- Trafalis, T.B., B. Santosa and M.B. Richman, 2005.
  Learning networks for tornado forecasting: a Bayesian perspective.
  Data Mining VI, WIT Transactions on Information and Communication Technologies, 35, 5-14.
- Son, H-J, T.B. Trafalis and M.B. Richman, 2005.
  Determination of the optimal batch size in incremental approaches: An application to tornado detection.
  Proceedings of International Joint Conference on Neural Networks, IEEE, 2706-2710.
- Trafalis, T.B., M.B. Richman and B. Santosa: 2006.
  Learning networks for tornado detection.
  International Journal of General Systems, 35, 93 - 107.

Introduction to Kernel Methods

- Kernel methods play a major role in Machine Learning.
- They provide a simple framework for manipulating nonlinear relationships.
- Require modest computational resources.

The Mercer Kernels

- A kernel is a continuous symmetric real-valued function defined on compact subsets of $\mathbb{R}^n$, $k : (x, y) \mapsto k(x, y)$.
- A Mercer kernel is a nonnegative definite kernel.
- The domain of a Mercer kernel is called the input space.
- The quantity $k(x, y)$ can be used to represent measures of angle and measures of distance.
- Angles and distances are between inputs mapped in a higher dimensional Hilbert space $\mathcal{H}$.
- The Hilbert space $\mathcal{H}$ is called the feature space.

The Kernel Trick

- Mercer's theorem suggests a particular decomposition of Mercer kernels.
- A Mercer kernel can be expressed as a dot product between two inputs mapped in the feature space, $k(x, y) = \langle \phi(x) \cdot \phi(y) \rangle$.
- Explicit knowledge of the map ϕ and the feature space is not required. The only thing of importance is the kernel itself.

Context of Supervised Learning

- Some problems in complex systems require the determination of some unknown and non-trivial rules for predicting the future system state.
- Algorithms in Machine Learning can automatically search for complex prediction patterns, such as in classification and regression problems.
- Classification and regression problems are supervised learning tasks in the sense that a learning machine is initially trained on known examples before being used to analyze future data coming from the same input source.

Linear Support Vector Regression

- Data points x_i are elements of $\mathbb{R}^n$. Their corresponding targets y_i are in $\mathbb{R}$.
- We need to find a prediction function f such that for each pair data-target (x_i, y_i) we have $f(x_i) \approx y_i$.
- f should be found such that:

$$\begin{array}{rcll} |f(x_i) - y_i| & \leqslant & \xi_i & \text{for every } x_i \\ \xi_i & \geqslant & 0 & \text{for every } x_i \end{array}$$

where ξ_i are slack variables.

Linear Support Vector Regression

- the prediction function f belongs to a class of functions denoted by $\mathcal{F}$ such that:

$$\mathcal{F} := \{x \in \mathbb{R}^n \mapsto \langle w \cdot x \rangle + b : \|w\| \leq B\},$$

  where $B > 0$, $w = \sum_j \alpha_j x_j$, and $\alpha_j \in \mathbb{R}$.
- The constraints become:

$$\begin{aligned} \left|\sum_j \alpha_j < x_j, x_i > +b - y_i\right| &\leqslant \xi_i \quad \text{for every } x_i \\ \xi_i &\geqslant 0 \quad \text{for every } x_i \end{aligned}$$

Nonlinear Support Vector Regression

- Consider the kernel $k : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$,
- Induce a new Hilbert space $\mathcal{H}$ and a map $\phi : \mathbb{R}^n \longrightarrow \mathcal{H}$ such that $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$ for any x and y in $\mathbb{R}^n$
- $\mathcal{F}$ becomes:

$$\mathcal{F} := \{x \in \mathbb{R}^n \mapsto \langle w \cdot \phi(x) \rangle_{\mathcal{H}} + b : \|w\|_{\mathcal{H}} \leq B\}$$

where $w = \sum_j \alpha_j \phi(x_j)$.

Nonlinear Support Vector Regression

- By kernelizing the previous constraints we obtain:

$$\begin{array}{rcll} \left|\sum_j \alpha_j k(x_j, x_i) + b - y_i\right| & \leqslant & \rho_i + \xi_i & \text{for every } x_i \\ \xi_i & \geqslant & 0 & \text{for every } x_i \end{array}$$

- The problem is formulated for α, ξ and b as variables.
- The SVM literature proposes an objective function that reduces the slack variables and the expected value of $|f(x_i) - y_i|$. Hence, we need to minimize the quantities $|b|$, $|\xi|$, and $\|w\|_{\mathcal{H}}$.

Nonlinear Support Vector Regression

- The quadratic programming that needs to be solved is:

$$\begin{aligned} \text{minimize} \quad & \alpha^T \mathbf{K}\alpha + C\xi^T\xi + b^2 \\ \text{subject to} \quad & |\mathbf{K}\alpha + b\mathbf{1} - y| \leq \xi \end{aligned}$$

- where $C > 0$, $(\mathbf{K})_{ij} = k(x_i, x_j)$, and y is the vector which entries are the y_i's.
- The optimal solution (α^*, b^*) of this problem gives the following prediction function:

$$f : x \mapsto \sum_j \alpha_j^* k(x_j, x) + b^*.$$

The vectors x_j for which the values α_j^* are nonzero are called support vectors.

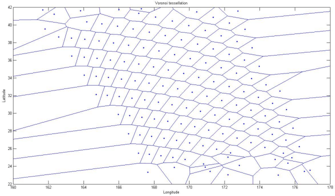
Voronoi Tessellation

- Due to the limitation imposed by the number of data points, we cannot apply SVR on the whole data set.
- Applying a Voronoi tessellation on the data set.
- Apply the SVR in each cell.

Voronoi Tessellation:
For any discrete set S of points in Euclidean space
and for almost any point x, there is one point of S closest to x.

Voronoi Tessellation

- The set of all points closer to a point c of S is the interior of a convex polytope (in some cases unbounded) called the Dirichlet domain or Voronoi cell for c.
- The set of such polytopes tessellates the whole space, and is the Voronoi tessellation corresponding to the set S (figure 1).



WindSat Data

- WindSat is a polarimetric microwave radiometer developed and built by NRL.
- The primarily objective of WindSat is to measure the ocean surface wind vector (speed and direction).
- The secondary objective is the measurements of sea surface temperature, rain rate and water vapor.
- Each orbit takes about 2 hours.
- During each orbit, over 120,000 observations are collected.

Data Set

- EDR Data from January 01, 2005 inside the region (127W,145E) and (23N,42N).
- After we took off all data points that have some missing information, we end up with 13500 points.
- The SVR allows only 1 output, hence, we need to find 2 prediction functions; one for the speed and one for the direction.
- SVR needs to be applied on continuous data. However, the direction is modulo 2π. Therefore, we need to transform the wind speed and direction to the U-V components.

Experiments

- The input vector contains longitude, latitude, SST, water vapor, cloud liquid water, rain rate. The target is either the U-component or the V-component.
- Before, applying the SVR, we needed to change the origin for the longitude to get a continuous transition from 180E to 180W.
- Linear $K(x_i, x_j) =< x_i, x_j >$. RBF $K(x_i, x_j) = e^{-\frac{\left\|x_i - x_j\right\|^2}{\sigma^2}}$.

	Kernel	Attributes
2 L	Linear	Longitude and latitude
2 RBF	RBF	Longitude and latitude
All L	Linear	Long., lat., SST, water vapor, cloud liquid water, & rain rate
All RBF	RBF	Long., lat., SST, water vapor, cloud liquid water, & rain rate

Table: The set of experiments

Results

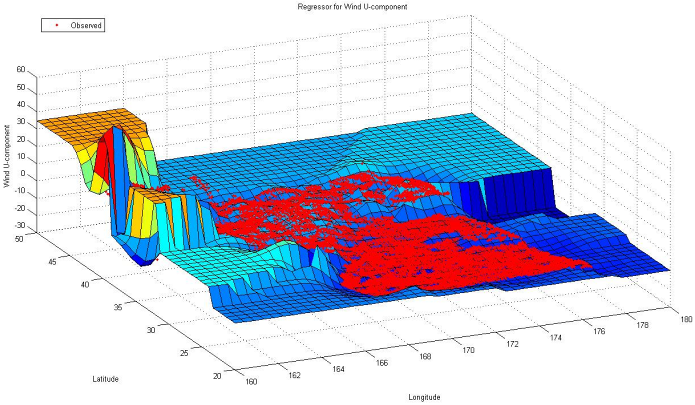
U and V predictions' errors

	U-component		V-component	
	MAE	MSE	MAE	MSE
2 L	2.03	12.70	1.93	7.77
All L	2.06	12.84	1.96	7.82
2 RBF	1.10	5.28	0.86	2.84
All RBF	1.15	20.99	.72	2.70

Table: U and V predictions' errors

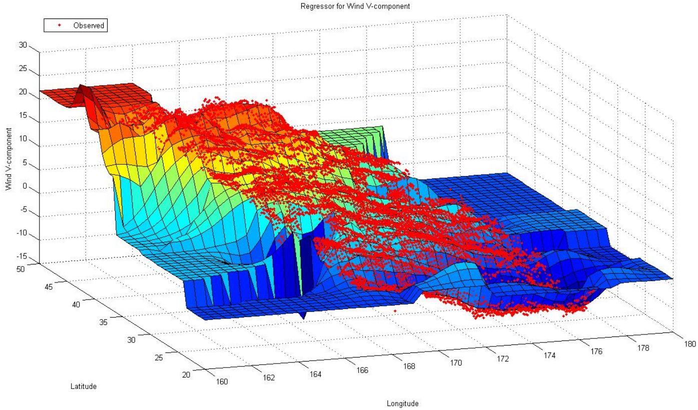
Results

The SVR solution for the U-component-



Results

The SVR solution for the V-component-



Results

The Contour Plot for the Observed U-component-

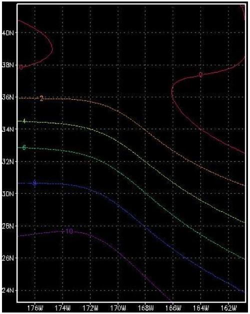


Figure: The Contour Plot for the Observed U-component -All RBF-

Results

Results -The Contour Plot for the Predicted U-component-

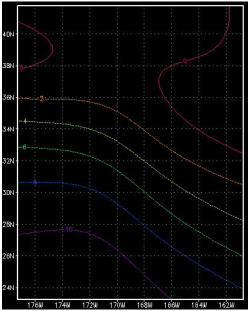


Figure: The Contour Plot for the Predicted U-component -All RBF-

Results

The Contour Plot for the Observed V-component-

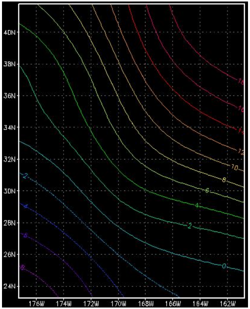


Figure: The Contour Plot for the Observed V-component -All RBF-

Results

The Contour Plot for the Predicted V-component-

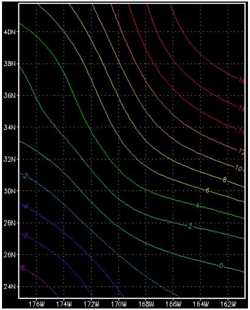
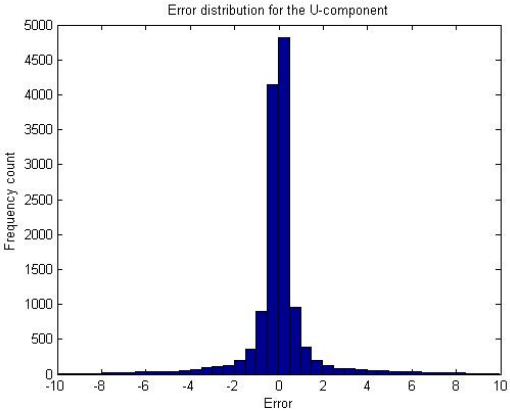


Figure: The Contour Plot for the Predicted V-component -All RBF-

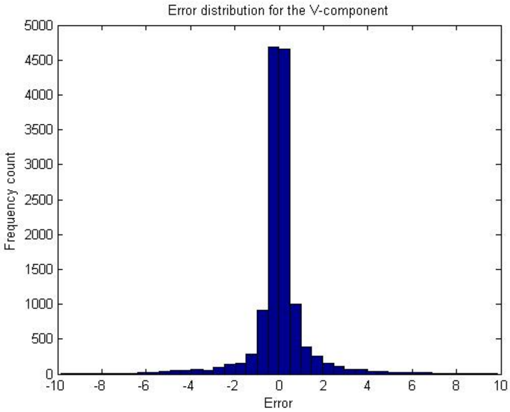
Results

Error distribution for the U-component-



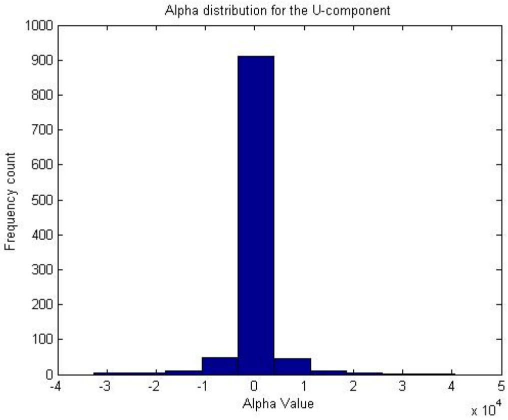
Results

Error distribution for the V-component-



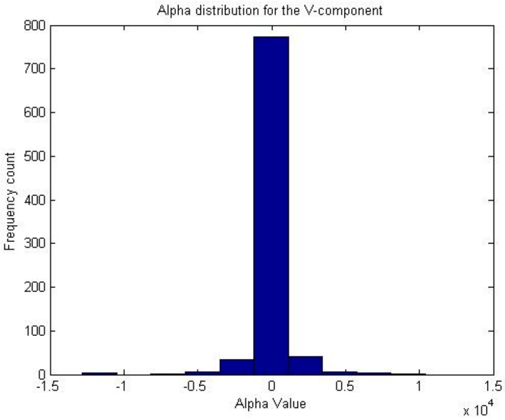
Results

Alpha distribution for the U-component-



Results

Alpha distribution for the V-component-



Conclusion and Future Plan

- Fewer than 8% of the data points (1000 support vectors out of 13540 data points) are needed to predict the U and V for the other data points.
- The support vectors reconstruct the ocean surface wind vector field with high accuracy (Correlation rate over 93%).
- Developed a MATLAB Toolbox able to achieve training speeds reaching 3000 vectors per seconds, which is roughly 11 millions vectors per hour. The code can be even faster if it is vectorized or pipelined.
- To use the points selected by the SVM and assimilate those into numerical weather prediction models.
- To use different data sets from different satellites.

Project Summary

Application of Support Vector Regression to Scatterometer Data

PIs and Co-PIs: M. Richman and L. Leslie

NWP Center Collaborators: JCSDA

Accomplishments

- **Fewer than 8% of the data points (1000 out of 13540) are needed to predict the U and V for the other data points**.
- **The support vectors reconstruct the ocean surface wind vector field with high accuracy (Correlation rate over 93%).**

Future Plan

- **To use the points selected by the SVM and assimilate those into numerical weather prediction models.**
- **To use different data sets from different satellites.**