Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 12 (2012) 276 – 281

Procedia Computer Science

Complex Adaptive Systems, Publication 2
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2012- Washington D.C.

Adaptive Machine Learning Approaches to Seasonal Prediction of Tropical Cyclones

Michael B. Richman[a]*, Lance M. Leslie[a]

[a]*The University of Oklahoma, Norman, OK, 73072, USA*

Abstract

Tropical cyclones (TCs) are devastating phenomena that cause loss of life and catastrophic damage, owing to destructive winds, flooding rains and coastal inundation from storm surges. Accurate seasonal predictions of TC frequency and intensity are required, with a lead-time appropriate for preemptive action. Current schemes rely on linear statistics to generate forecasts of the TC activity for an upcoming season. Such techniques employ a suite of intercorrelated predictors; however, the relationships between predictors and TCs violate assumptions of standard prediction techniques.

We extend tradition linear approaches, implementing support vector regression (SVR) models. Multiple linear regression (MLR) is used to create a baseline to assess SVR performance. Nine predictors for each calendar month (108 total) were inputs to MLR. MLR equations were unstable, owing to collinearity, requiring variable selection. Stepwise multiple regression was used to select a subset of three attributes adaptive to specific climatological variability. The R^2 for the MLR testing data was 0.182. The SVR model used the same predictors with a radial basis function kernel to extend the traditional linear approach. Results of that model had an R^2 of 0.255 (~ 40% improvement over linear model). Refinement of the SVR to include the Quasi-Biennial Oscillation (QBO) improved the SVR predictions dramatically with an R^2 of 0.564 (~ 121% improvement over SVR without QBO).

Keywords: Prediction, Kernel Methods, Support Vector Regression, Cross-Validation, Tropical Cyclones

1. Introduction

Tropical cyclones (TCs) pose an annual threat to life and property in affected areas. They often make landfall with little or no warning. The American Meteorological Society glossary defines a TC as a disturbance originating over tropical oceans. In [1], a TC is defined as a "warm core cyclonically rotating wind system in which maximum sustained winds are 17 ms^{-1} (40 mph) or greater." The year-to-year variation in the number of TCs and the number of landfalling TCs varies over the TC regions (commonly referred to as TC basins) around the Earth. Obtaining accurate predictions of the number of TCs, with lead times of at least 3 months or more, in an upcoming TC season is of vital importance for the preparation of emergency response services for the areas expected to be affected.

Southern Hemisphere TC basins are not studied as extensively as those in the Northern Hemisphere (e.g. the North Atlantic and the Western North Pacific). Here, the focus is on a subset of the Australian TC region (Fig. 1a). The Northwest Australian TC region (105-135E; 0 to 35S; Fig. 1a) is one of three sub-regions of the Australian TC region. The Northwest Australian TC region is evident as a local maximum in mean annual TC frequency (Fig. 1b, orange and red squares). The annual TC and TC landfalling counts are important because of their impact on the economy of the NW Australian region because TCs impact offshore oil and gas if they remain at sea. However, landfalling TCs cause flooding that impacts negatively local mining and port operations. Owing to the high correlation between the two time series (Fig. 2) and the larger number of total TCs, we focus on that time series herein. In the 41-year period, only 1 year had no landfalling storm.

 doi:10.1016/j.procs.2012.09.069

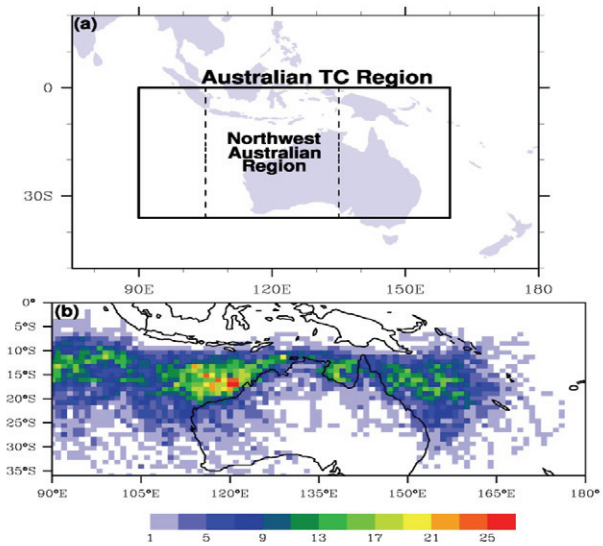

Fig. 1. (a) Geographical location of the Australian tropical cyclone region. Dashed lines in the inner box show NW Australian TC region. (b) Frequency counts of TCs for 1 degree squares, *1970-2010.*

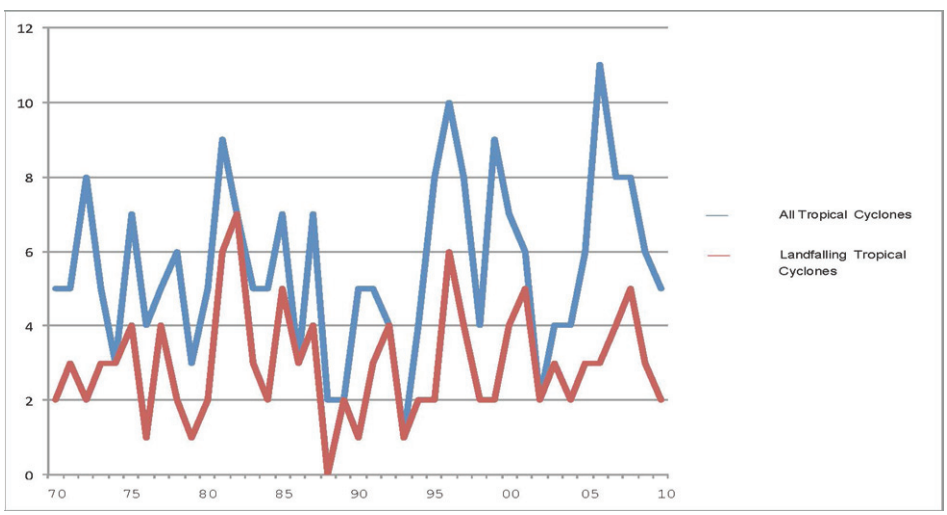

Fig. 2. Annual TC count for years 1970 – 2010 (blue) and annual TC landfalling count (red).

Several schemes for predicting TC seasonal frequencies have been developed for the entire Australian TC region by [2,3] with annual TC frequencies linearly regressed with Darwin's winter pressures. In [4], a MLR scheme is applied, with large scale climate mode predictors (e.g., El Niño Southern Oscillation (ENSO)).

As mentioned above, statistical forecasting of TCs is most often accomplished for this region using linear regression with predictors known to modulate TC formation and track. In the western Australian region, this linear approach offers only a modest level of skill. The technique makes assumptions, such as linear relationship between the predictors / response variable and independent predictors, that are often violated. In this research, we extend the general approach by employing support vector regression (SVR) to allow for nonlinear relationships. Additionally, through a process of stepwise regression and all-subsets regression, a subset of predictors is selected that reduces the number and intercorrelation among the predictors.

2. Data and Methodology

2.1 Data

The Australian region has observations of TCs dating back to the beginning of the 19th century, and reliable observations increased through the middle part of the 20th century, with the onset of the satellite era around 1970 [5]. The data sets for the predictors (Niño1, 2, 4, MODOKI, QBO, SAM) all are freely available from online websites (e.g. http://www.bom.gov.au; http://www.esrl.noaa.gov/psd/data/gridded/).

2.2 Methodology

Several steps were required to preprocess the data. As the data were not measured in the same metric, all data were standardized to set the mean to zero and the standard deviation to one. The data were divided into training and testing sets, of 28 and 12 years, respectively, to allow for testing generalization properties of predictor equations. Extensive investigation, using these data, suggested that 28 years of training data were necessary to achieve stable prediction equations. The testing block of 12 years was large enough to insure a rigorous assessment of skill by avoiding large sampling error. Data were available for each calendar month; however, since the Australian TC season spans December through April, proper forecasting requires that the training period does not overlap the testing period. Moreover, using data at lead times too far in advance of the TC season would increase the risk of identifying chance relationships. Hence, May through September data were used to form the prediction equations.

In SVR, each input (observation, x) is mapped into an m-dimensional feature space. A model, $f(x, \omega)$, is then constructed in the feature space, given by:

$$f(x,\omega) = \sum_{j=1}^{m} \omega_j \, g_j(x) + b$$

where $g_j(x), j = 1, \ldots, m$ represents a set of nonlinear transformations and b is the bias term. Since the data are standardized, the bias term is omitted.

SVR operates in high dimensional feature space using ε -insensitive loss while simultaneously reducing model complexity by minimizing $\| \omega \|^2$ using slack variables, ξ_i, ξ_i^*, $i = 1, \ldots, n$ to measure the deviation of the training data outside the ε -insensitive area. Hence, the SVR model can be defined as minimizing the function

$$\frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

$$s.t. \begin{cases} y_i - f(x_i,\omega) \le \varepsilon + \xi_i^* \\ f(x_i,\omega) - y_i \le \varepsilon + \xi_i \\ \xi_i^*, \xi_i \ge 0, i = 1, \ldots n \end{cases}$$

This optimization is transformed to a kernel approach by the solution

$$f(x) = \sum_{i=1}^{n_{SV}} (\alpha_i - \alpha_i^*) \, K(x_i - x)$$

$$\text{s.t. } 0 \le \alpha_i^* \le C, 0 \le \alpha_i \le C$$

where n_{SV} is the number of support vectors and the kernel function is

$$K(x, x_i) = \sum_{j=1}^{m} g_j(x) g_j(x_i).$$

The quality of the prediction depends on finding values of C, ε the kernel selection and parameters that lead to good generalization. All calculations used Weka [6]. Given the large number of potential predictors, initial variable selection was employed using forward stepwise regression, reducing the number of predictors from 48 to 6. Those 6 predictors were subject to an all-subsets regression and all-subsets SVR and the generalization properties on the testing data were noted. The regression with the optimal generalization prediction had three predictors, MODOKI (Jun.), SAM (Aug.) and NINO 4 (Sep.). These three

predictors had correlations of ≤0.38. Linear, polynomial and radial basis function kernels were evaluated for SVR. After substantial experimentation with 406 kernels, the radial basis function kernel was selected, with gamma = 0.73 and C =1 having optimal generalization properties.

Since the QBO predictor was not available for the first ten years of the training data, a second experiment used a subset of the last 19 years of the testing data. For the second experiment, the SVR used a gamma=0.12.

3. Results

3.1 Baseline: Multiple Linear Regression

After variable selection, the MLR generalized best with three predictors (Table 1, Fig. 3), MODOKI, SAM and NINO4. These three predictors had p-values (F-test) of 0.004, 0.003 and 0.04, respectively in the training data and the regression equation had a p-value of 0.0005. There were 28 observations in the training data. The prediction equation was: 5.8239 -1.1866 * MODOKI_6 -0.6155 * SAM_8 -1.3181 * NINO4_9 with R^2=0.182. The MAE was 1.44 and the RMSE was 1.71. The testing data had 12 years and the resulting predictions are summarized in Fig. 5.

Table 1. Predictors in TC prediction experiments.

Predictors MLR and SVR Experiment 1	Predictors SVR Experiment 2
MODOKI (Jun.)	NINO1+2 (May)
SAM (Aug.)	MODOKI (Jun.)
NINO4 (Sep.)	QBO (Aug.)
	NINO4 (Sep.)

3.2 Support Vector Regression – Experiment 1

The SVR used Sequential Minimal Optimization (SMO) that has an analytical solution. Therefore, SVM is optimized without requiring a QP solver, making the process efficient and scalable [7]. The build time for the model was approximately 0.01 seconds. The same variables were used as in the MLR experiment (Table 1, Fig. 3). After experimentation, it was found that a radial basis function kernel with C=1 and gamma = 0.73 provided the optimal generalization by maximizing the R^2 and minimizing both the MAE and RMSE on the testing data. In comparison to the MLR, the SVR R^2 (.255) had a 40.1% improvement, whereas the MAE (1.41) and RMSE (1.68) improved by modest amounts of 2.1% and 1.8%, respectively. The SVR predictions are shown in Fig. 5.

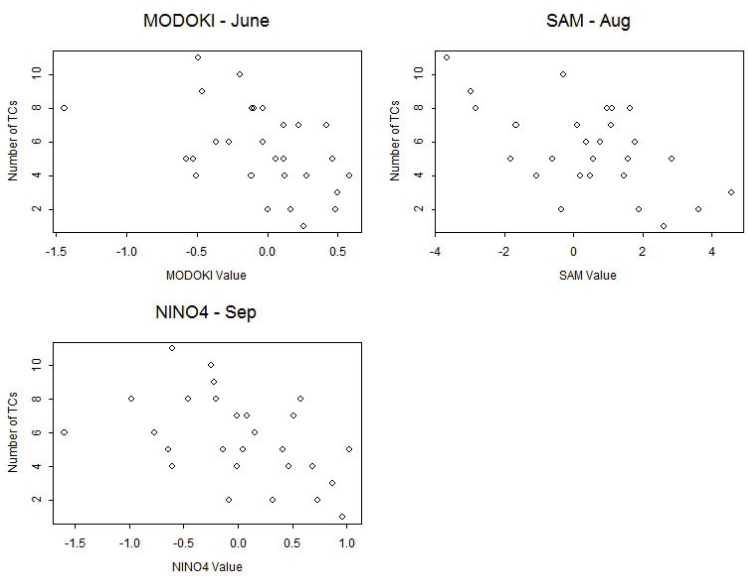

Fig. 3. Plots of the predictor values (x-axis) versus the response variables (y-axis) for MLR and SVR experiment 1.

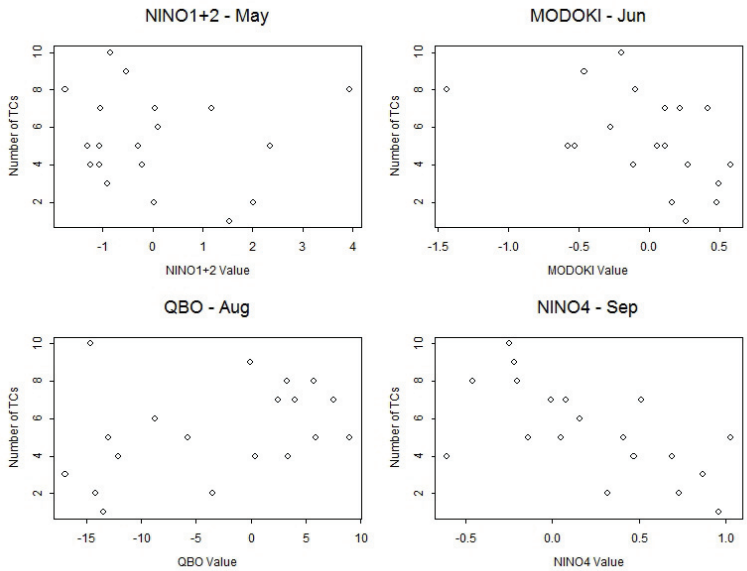

Fig. 4. Plots of the predictor values (x-axis) versus the response variables (y-axis) for SVR experiment 2.

3.3 Support Vector Regression – Experiment 2

The same SVR algorithm as in experiment 1 was used with an additional variable, the Quasi-Biennial Oscillation (QBO) Index of stratospheric winds. The complication in using the QBO Index is that it was not calculated or available for the first 9 years of the training data. Therefore, rather than 28 years of training data, we are now limited to 19 years. Additionally, by using the same detailed process for variable selection, a new El Niño region to the east of the previous one is selected (Table 1 and Fig. 4). Since the linear model (Section 3.2) has considerably lower skill than the SVR model, there was no testing of the linear model in this experiment. There were 90 kernel evaluations and the radial basis function with C=1 and gamma = 0.12 was found to maximize the R^2 and minimize the MAE and RMSE in the testing data. When SVR experiment 2 is compared to SVR experiment 1, the effect of adding the QBO improved the R^2 (.564) by ~ 121%, whereas the MAE (0.97) and RMSE (1.19) improved by 31.2% and 29.2%, respectively. The predictions are shown in Fig. 5.

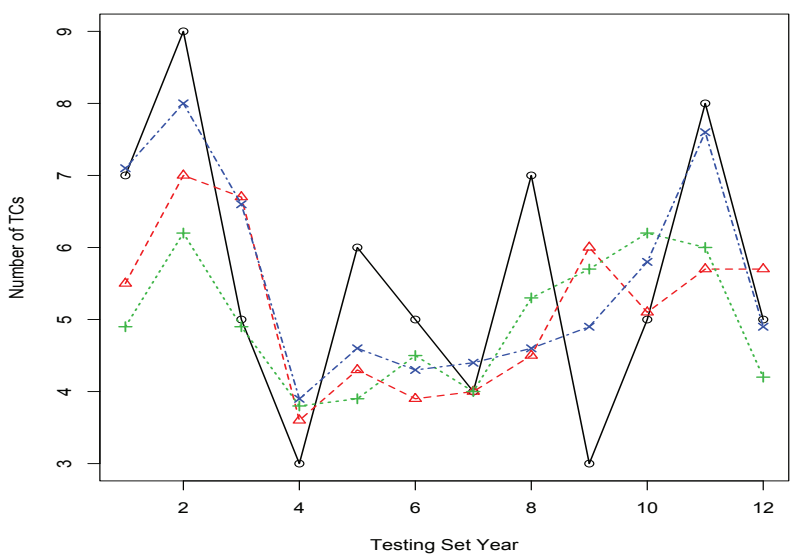

Fig. 5. Number of TCs for the 12 year prediction window (o, black line), MLR model predictions (Δ, red dashed line), SVR experiment 1 predictions (+, green dotted line) and SVR experiment 2 predictions (x, blue dash-dotted line).

4. Summary and Conclusions

SVR has been applied successfully to a challenging problem of predicting the number of tropical cyclones in the northwest portion of Australia. Compared to MLR, nonlinear SVR improved the prediction accuracy, measured by R^2, by ~ 40%. Furthermore, by incorporating a new variable in SVR, the prediction accuracy improved by ~ 121% over the original SVR. Such a large improvement in prediction implies that creating a SVR model could provide valuable information to protect life and property. For future research, the training data should be updated every 5-10 years to account for low frequency variability. In this case, the cross-validation scheme would require updating.

References

1. Gray, W. M., 1968: Global view of the origin of tropical disturbances and storms. *Mon. Wea. Rev.*, **96,** 669–700.
2. Nicholls, N., 1979: A possible method for predicting seasonal tropical cyclone activity in the Australian region. *Mon. Wea. Rev.*, **107**, 1221–1224.
3. Nicholls, N., 1985: Predictability of interannual variations of Australian seasonal tropical cyclone activity. *Mon. Wea. Rev.*, **113,** 1144–1149.
4. Goebbert, K. H., L. M. Leslie, 2010: Interannual Variability of Northwest Australian Tropical Cyclones. *J. Climate*, **23**, 4538–4555.
5. Holland, G. J., 1981: On the quality of the Australian tropical cyclone database. *Aust. Meteor. Mag.*, **29,** 169–181.
6. Hall, M. E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, 2009: The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, **11**, 10-18.
7. Platt, J., 1998: Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds., MIT Press, Cambridge, MA, 185–208.