Machine learning techniques to analyze extra-tropical transition

Israel Vaughn, J. Scott Tyo, and Elizabeth Ritchie
ivaughn@optics.arizona.edu
University of Arizona

Introduction

Extratropical cyclones in the Western Pacific cause a great deal of damage annually. Generally, a tropical cyclone (TC) will track southerly into the mid-latitudes, where it interacts with mid-latitude weather, and subsequently re-intensifies about 51% of the time. When this occurs, the resulting extratropical cyclone is asymmetric, and has a large area of effect. The transition itself, extratropical transition (ET), and the subsequent dissipation or intensification of the extratropical cyclone is difficult to predict (in the figure below there is no difference in surface pressure prior to ET).

Data

Due to the large physical size of the atmosphere and the difficulty of collecting data for many variables (e.g., wind speed, temperature, etc.) a numerical physics model is backfitted to measured data that is obtained from various sources. The model outputs are then used as the “truth” for input into our machine learning algorithms. We are currently using the GFS-FNL data for a storm centered volume of ±25° in latitude, ±30° in longitude, 1000-100hPa in pressure for 7 different variables, and times at 6 hour intervals from -72 hours to +72 hours. Our data set consists of 108 storms that underwent ET from 2000-2008; with each storm having a vector of length 5,945,121. Although we are currently investigating other machine learning methods, our primary results have been accomplished via the SVM algorithm [1].

Data Issues

The low number of labeled data (108 storms) in very high dimensions (n = 5, 945, 121) causes over-training ("curse of dimensionality"). A priori feature extraction is used to mitigate over-training.

Classification

Support vector machines (SVMs) project the data into an infinite dimensional (Hilbert) space, and then find the "best" hyperplane that separates the classes in the infinite space. The hyperplane in the infinite dimensional dual space corresponds to some arbitrarily shaped surface in the data space. Functional analysis then allows

\[ \sum_{i=1}^{n} \alpha_i \phi(x_i) - \sum_{i=1}^{n} \alpha_i y_i = 0 \leq y_i \alpha_i \leq C \]

to be minimized over \( \alpha \) (a quadratic programming problem), where the inner product \( \phi(x_i) \phi(x_j) \) is the kernel function, the \( y_i \in \{+1, -1\} \)s are the labels, and \( C \) is the box constant. The data is then input into the SVM for training. The training is accomplished using the k-folds method, with a validation test set withheld from the training data.

Results from SVM

The dimensionality of the input vectors was reduced using correlation-based feature selection (CFS) developed by Hall before input into the SVM. For the -72 hour time point, the best ROC curve was obtained by using 29 dimensional vectors (from Hall CFS) from equivalent potential temperature, \( \theta_v \), volumes (1 for each storm) for the SVM training, and then running the resultant SVM classifier on the withheld test set. \( \theta_v \) can be thought of as a quantity proportional to energy, and it is computed from the raw outputs of the FNL data.

ET Chaotic?

The results from the SVM classification indicate that the ET process is not chaotic on a ±72 hour time scale. This highlights that the dynamics and physics of these systems are not well understood, and further research is needed to understand the time scales at which predictability breaks down, i.e. at what space and time scales is the system chaotic? We intend to explore this question by using machine learning, specifically RML [2], to extract the intrinsic features of the data manifold.

References


Funding

This work was supported by the NSF Division of Atmospheric and Geospace Sciences, Award # ATM0730079