REGIME DEPENDENT PRECIPITATION TYPE FORECASTING

Tyler McCandless
The Pennsylvania State University Department of Meteorology

1. INTRODUCTION

The ability to determine precipitation type instantaneously across a forecast area would provide forecasters with much more information about current weather conditions. Currently, weather observations are reported once per hour from reporting sites at airports and other official reporting stations. These locations are not fit to a grid with clusters of reporting sites closer to urban areas and less reporting sites in rural areas. This unequal distribution of sites can make determination of precipitation type difficult throughout the forecast area.

The polarimetric radar sends out pulses of radar waves that bounce off of particles in the atmosphere and the energy is reflected back to the radar dish. A computer processes the returned signals and through the use of algorithms can make conclusions about the particles it saw. The polarimetric radar measures such variables as the differential reflectivity, correlation coefficient, linear depolarization ratio, specific differential phase, and cross-polar terms that have yet to be fully studied. Since the polarimetric radar is able to measure more weather variables than Doppler radar, the polarimetric radar has many more applications. A method to determine precipitation type based off of the variables measured by the polarimetric radar would give forecasters an instantaneous view of the current weather and improve short term forecast skill.

The regime dependent nature of the atmosphere has been studied in multiple aspects of weather and climate. In 1986, Brown et al. found that smaller scale, transient eddies may play a regime-dependent role in interactions with atmospheric circulation modes on the scale of persistent anomalies. In 1990, Zwiers and Storch introduced the class of regime dependent autoregressive time series modeling of the Southern Oscillation. In 1998, Paul J. Roebber examined the manner in which degree day forecasters adjust their reliance on particular pieces of forecast information as the large-scale flow pattern evolves into different regimes and found that the weighting of MOS was situation dependent and that forecast skill and value were maintained under large-scale flow regimes in which MOS was less useful through significant adjustment of forecast technique. In 2007, Deloncle et al. used the k-nearest-neighbor classifier and the random forest statistical learning techniques to predict the transition between weather regimes from a three layer quasigeostrophic model. Greybush et al. (2008) showed that the K-means regime-clustering post-processing method produced the lowest MAE in two-meter temperature forecasts when compared with regime regression, a genetic algorithm regime method, and different windowed performance-weighted averages.

In this study, the goal is to use the k-nearest-neighbor classifier to determine precipitation type from the polarimetric radar data. The hypothesis is that there are specific characteristics, or regimes, that are associated with each precipitation type. Therefore, the k-nearest-neighbor classifier is used because this clustering method separates the data based on similarities between various classes. Although each day is not assigned to a specific cluster, or regime, this can be interpreted as a regime identification method because it separates the data by class similarities and dissimilarities.

2. METHODS

This study uses data provided by the 2008 American Meteorological Society Artificial Intelligence Contest. There are 847 observations used in the training dataset and 363 in the testing dataset. The variables in the datasets are: index, latitude, longitude, two-meter air temperature, relative humidity, u-component of the wind, v-component of the wind, freezing level, tilt, range, azimuth, height of radar data, differential reflectivity, cross-correlation coefficient, specific differential phase, and reflectivity. The observed precipitation type is only given in the training data. The observed precipitation type is listed as frozen, liquid, or none. The majority of the polarimetric radar data is complete; however, there are multiple missing differential reflectivity, cross-correlation coefficient, and specific differential phase. The percentage of missing data is shown in Table 1. Due to lack of time and similar percentages of missing values in both the training and testing data, the missing values were not replaced or deleted.

Corresponding author address: Applied Research Laboratory, P.O. Box 30, The Pennsylvania State University, State College, PA 16804-0030, email: tcm5026@only.arl.psu.edu

	Differential Reflectivity	Cross-Correlation Coefficient	Specific Differential Phase
Percentage of Training Cases Missing	12.51%	12.51%	26.45%
Percentage of Testing Cases Missing	8.82%	8.82%	21.76%

Table 1. Percentage of missing values in testing and training data.

The first step in indentifying atmospheric regimes by the polarimetric radar is deciding what variables to use. Three different sets of polarimetric radar data are used in this study. The first method simply uses all possible weather variables. This method takes longer to cluster and has a greater chance classifying instances based off of chance relationships in the data, or overfitting, because there are more possible associations between variables.

The second method uses principal component analysis to determine which variables are the most important in classification. The principal component analysis determined that the first 12 variables added valuable information to the clustering and variables 13, 14, and 15 are insignificant. The first 12 variables are latitude, longitude, two-meter air temperature, relative humidity, u-component of the wind, v-component of the wind, freezing level, tilt, range, azimuth, height of radar data, and differential reflectivity. The elimination of cross-correlation coefficient and specific differential phase are likely due to missing values that skew the actual relationships in the data.

The third method uses an attribute subset evaluator from Weka. Subset evaluators use a subset of attributes to calculate a numeric measure that guides the search. Correlation-based Feature Subset Selection is used to assess the variables, or predictors. This subset evaluator calculates the predictive ability of each attribute individually, as well as the degree of redundancy between them. This creates a set of predictors with high predictive ability and low intercorrelation.

The type of machine-learning algorithm used in this study is the k-nearest-neighbor instance-based learning. An instance-based learning method uses the instances themselves to represent what is learned, rather than inferring a rule set or decision tree and storing it instead. In this type of instance-based learning, the real work is done when it is time to classify a new instance; therefore, because instance-based learning defers the work for as long as possible, it is defined being lazy. A nearest-neighbor classification method compares a new instance to an existing one using a distance metric to assign the class to a new one. In the k-nearest-neighbor method, the distance-weighted average of k of the closest neighbors is used to assign the new instance. Although the k-nearest-neighbor classification does not distinctly define a cluster for each instance, the k-nearest-neighbor classification can be thought of as assigning precipitation type depending on the most similar regime. It has also been shown that the k-nearest-neighbor method can identify transitions between regimes (Deloncle et al 2007). The k-nearest neighbor machine learning algorithm is trained on each of the three different subsets of variables and predicts the precipitation type for the testing data.

Woodcock and Engel (2005) argued that the combination of forecasts can improve prediction upon a single more complex algorithm. Therefore, after classifying each instance as liquid, frozen, or none in the testing data, the mode of the forecasts is used as the deterministic forecast. The mode of the forecasts is used because the forecast is nominal. For the three different forecasts per index, only one had three different predictions. For that index, a forecast of frozen is applied because frozen was the most common observed values in the training dataset.

3. RESULTS

The goal of this study is to determine precipitation type from polarimetric radar data. The metric used is the multi-categorical form of Peirces's Skill Score (PSS), which is also referred to as the true skill statistic or the Hanssen and Kuipers discriminant,. The PSS is a value between negative one and positive one. Zero indicates no skill over a baseline forecast, which in this case is climatology. A score of one indicates a perfect forecast. A bootstrap analysis was performed and 95% bootstrap BCa confidence intervals with 1000 replicates was calculated. The lower value for the k-nearest-neighbor classifier is calculated to be 0.204, the mean value is 0.2842, an upper value of 0.3730, and an official Pierce Skill Score of 0.28510.

Lower	Mean	Upper	PSS
0.204	0.2842	0.3730	0.28510

Table 2. Peirce Skill Score results.

4. CONCLUSIONS

The results show that there is forecast ability in the k-nearest-neighbor classifier. A value of 0.28510 for the Pierce Skill Score is above zero, which is a no skill forecast, and below one, which is a perfect skill forecast. A permutation test (3000 permutations) for a significant difference (at 95%) showed that there is statistical significant between a PSS of 0.28510 and zero. This means that there is statistically significant difference

between climatology and the k-nearest-neighbor classifier. Although the results show that there is skill in the k-nearest-neighbor classifier, there are more advances that could improve upon this forecast. First, addressing the issue of missing forecast would likely improve forecast ability. Although the differential reflectivity, cross-correlation coefficient, and specific differential phase had numerous missing values, the available values were important for forecasting. In the analysis with the Correlation-based Feature Subset Selection, these variables showed correlation coefficient values with the observed much higher than all but four other variables. I would hypothesize that a full dataset of these variables would have much higher correlation coefficients and likely add significant value to the k-nearest-neighbor classifier. A method to impute the missing values by finding correlations between variables would likely help improve the classifier. Second, a combination of more variables sets would likely improve the results. In this study there are three different sets of variables used. Combining more subsets of variables could also add value to the forecast. Third, a weighted average scheme would also improve the forecast. In this study, the mode of the forecasts was used; however, using a weighted average could add value to a dataset that had a better skill score on the training dataset. Finally, a combination of multiple instance based classifiers could improve precipitation type forecast ability.

ACHKNOWLEDGMENTS – The author wishes to thank Dr. Sue Ellen Haupt for insightful discussions on this project. The author also wishes to thank Dr. Kim Elmore for his work putting together the Artificial Intelligence Contest as well as calculating the results. The research was funded in part by the PSU Applied Research Laboratory Honors Program.

References:

Deloncle, A., R. Berk, F. D'Andrea, and M. Ghil, 2007: Weather Regime Prediction Using Statistical Learning. *J. Atmos. Sci.*, **64**, 1619–1635.

Greybush, S.J., S.E. Haupt, and G.S. Young, 2008: The Regime Dependence of Optimally Weighted Ensemble Model Consensus Forecasts of Surface Temperature. *Wea. Forecasting*, **23**, 1146–1161.

Brown, P.S., J.P. Pandolfo, and A.R. Hansen, 1986: Circulation Regime-Dependent Nonlinear Interactions during Northern Hemisphere Winter. *J. Atmos. Sci.*, **43**, 476–485.

Zwiers, F., and H. Von Storch, 1990: Regime-Dependent Autoregressive Time Series Modeling of the Southern Oscillation. *J. Climate*, **3**, 1347–1363.

Roebber, P.J., 1998: The Regime Dependence of Degree Day Forecast Technique, Skill, and Value. *Wea. Forecasting*, **13**, 783–794.

Wilks, D.S., 2006: *Statistical methods in the atmospheric sciences*, 2nd ed., Academic Press, 626 pp.

Witten, I. H., and E. Frank., 2005: Data Mining: *Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Woodcock, F. and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting* **20,** 101-111.