Support Vector Machine Classification using the AI Competition Data Set

Kimberly L. Elmore[1] and Michael B. Richman[2]

1. Introduction

Support vector machines (SVMs) are a classifier that uses optimal separating hyperplanes (Vapnik, 1996). For the two-class case, optimal separating hyperplanes maximize the distance to the closest point from either class. They provide a unique solution to the separating hyperplane problem and they maximize the separating margin between the two classes on the training data which, in theory, leads to better classification performance on the test data.

Given these potentially useful characteristics, the SVM form is investigated here, as it is useful when the different classes are not linearly separable. The generalization of optimal separating hyperplanes to a SVM (or, alternatively a support vector classifier) produces nonlinear boundaries by casting a linear boundary in a large, transformed version of the feature space (Hastie et al, 2001).

The theory behind SVMs is not further developed here, Instead, we outline the method we followed in developing the particular SVM used here. We will discuss what we did, and what we would do differently in future given the observed performance.

2. Missing Data

The first problem encountered is missing data: some predictors in the training data set are missing slightly over 67% of the time. Table 1 provides a summary of the missing values. Table 2 provides the same summary for the testing data set. Of the predictors that are missing, POSH suffers the most and fairs proportionally worse in the testing data set. However, the testing data set has two missing predictors that are not shared by the training data set: Max Azimuthal Shear over the past 120 min, max reflectivity and mean reflectivity.

Table 1:

Predictor	Number of Missing	Proportion of Missing
Lifetime MESH	335	24.7%
Lifetime POSH	802	59.1
Low Level Shear	6	0.4%
MESH	390	28.8%
Max VIL	83	6.1%
Motion East	113	8.3%
Motion South	113	8.3%
POSH	910	67.1%
Max Az Shear	23	1.7%
Max Az Shear 30	4	0.2%
Speed	113	8.3%

There are various ways of dealing with missing data. Extracting cases with missing data is not feasible here (and is typically discouraged), so some sort of imputation is necessary. Some classification algorithms are designed to deal with missing data during the training phase. This means that the imputation is performed internally. The SVM algorithm used here (Chang and Lin, 2001) does not perform internal imputation. Hence, imputation must be performed prior to any training.

Table 2:

Predictor	Number of Missing	Proportion of missing
Lifetime MESH	412	38.5%
Lifetime POSH	740	69.2%
Low Level Shear	15	1.4%
MESH	487	45.6%
Max dBZ	2	0.2%
Max VIL	60	5.6%
Mean dBZ	2	0.2%
Motion East	73	6.8%
Motion South	73	6.8%
POSH	799	74.7%
Max Az Shear 120	31	2.9%
Max Az Shear 30	12	1.1%
Speed	73	6.8%

Data imputation is covered in various sources and is not reviewed here. The method employed for this work uses additive regression, bootstrapping and predictive mean matching. This method incorporates all aspects of uncertainty into by using the bootstrap to approximate the process of drawing predicted values from a full Bayesian predictive distribution. Different bootstrap resamples are used for each of the multiple imputations. A flexible additive model is fit to each sample with replacement from the original data, and this model is used to predict all of the original missing and non-missing values for the target variable.

Multiple imputation often refers to refitting the model using the fitted values as first guesses. This is repeated (usually less than ten times) until the imputed values no longer change. It can also refer to fitting a classifier to each imputation and in an attempt to assess the variability inherent in the classifier based on the imputed data. In this case, we instead generated 100 imputed values for each missing variable and replaced missing variables with the mean of the 100 imputed values. Thus, we had only one model to train, but lacked knowledge of its uncertainty in the face of multiple imputations. We chose a priori to impute all the missing data. This will have ramifications later.

1. Affiliation: Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK.
Corresponding author address: 120 David L. Boren Blvd, Norman, OK 73072
2. Affilaition: School of Meteorology, University of Okalhoma, Norman, OK.

3. Geographic Imbalance

When the latitude and longitude of the training and testing data are plotted, Clearly there is a geographic imbalance (Fig. 1). All of the training data are contained

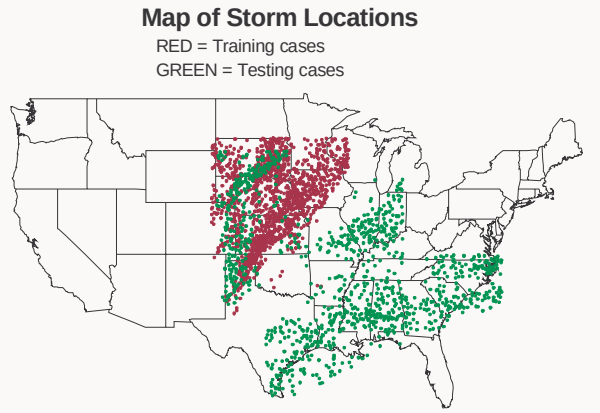

Figure 1. Locations of cases used for training (red) and testing (green).

in the northwest half of an area bounded roughly by the Texas panhandle, central Arkansas, west-central Wisconsin, and the northwest corner of South Dakota. Because of this, latitude and longitude of the cells are a priori retained as predictors.

Because storm characteristics differ depending on geographic regions, training data sets should ideally cover the same area that the testing data cover. SUch an approach results in a more robust classifier. However, having a training data set in a different region is one way to test the general applicability of a classifier.

4. Model Development

Once the missing data are imputed, a model is generated and tested. SVMs may use various kernels. The radial basis function (RBF) kernel is used here. The RBF nonlinearly maps samples into a higher-dimensional space, has fewer numerical difficulties, and unlike a linear kernel can handle cases when the relationship between classes and attributes is nonlinear. In addition both the linear and sigmoid kernels are special cases of the RBF (Lin and Lin 2003). Using the RBF, the classifier is a hyperplane in the high-dimensional feature space, but nonlinear in the original input space. All data are scaled to mean zero and unit variance prior to fitting the model.

The RBF requires the choice of two hyperparameters, C and γ. C is considered to be a cost function weight while γ is a radius (in some high dimension) that may also be thought of as a smoothing parameter.

This approach uses a gridded search, where C and γ are varied over a grid of values and some interesting characteristic is mapped as a surface spanning the ranges of C and γ (Fig. 2).

While the minimum is near the edge of the displayed surface, as γ (labelled "gamma") decreases, error rates rise rapidly for all C (labeled "cost"), making the surface difficult to visualize. As C increases, error rates slowly increase. Overall, there is a minimum in error rates in a broad trough for γ ~ 0.06-0.10. Bootstrap confidence

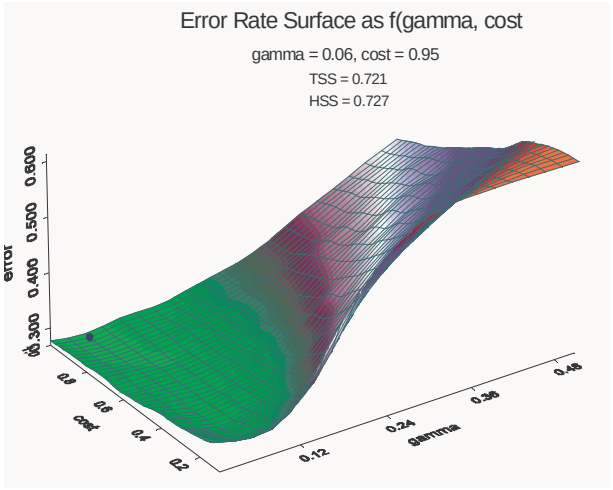

Figure 2. Error rate surface for values of C ("cost") and γ ("gamma") spanning ranges of [0.01, 0.51] and [0.1, 1.0] respectively. The surface shows the error (hit rate) for SVM models using various C and γ. The minimum error rate occurs at the dot.

intervals have not been computed for this surface. Hence there is likely to be a neighborhood for which the error rates that are statistically indistinguishable for a range of C and γ. Hiedke and True (or Pierce) skill statistics (WIlks 2006) are also plotted shown on Fig. 2.

Other data scalings were tried, such as scaling the data to the interval [-1, 1] and even no scaling, but these either produced much poorer results obvious over-fitting (perfect skill scores).

5. Results and Lessons Learned

Results are relatively poor for this particular model. The overall error rate on the test data is 0.666, compared to 0.738 for the baseline results, and the TSS = 0.494 compared to 0.583 for the baseline results.

Appropriate predictor selection is the most likely key to any improvement. For a better SVM implementation, the first suspected problem is in the imputation of missing data. Experience has shown one of us (Richman) that if more than about 25% to 30% or the data are missing, imputation can give poor results. That appears to be the case here, as imputation results in very different distributions (not shown) for the imputed values vs. the observed values for the most often missing predictors (POSH and Lifetime POSH) when conditioned upon storm category.

From an imputation standpoint, the safest course is to not impute these two variables and omit them from the model While imputing these missing values independently for each storm category might seem a promising approach there is 1) be no way to do the same for the testing data, and 2) POSH is missing for 98% of the nonsevere cases in the training data. Imputation in such cases is essentially hopeless. Lifetime POSH mimics POSH in the degree of missing data.

Another approach might be to convert POSH or Lifetime POSH to a categorical variable that indicates if it is missing, as the lack of a POSH (or Lifetime POSH) value seems to be a good indicator that the storm is at least nonsevere.

No tests for model simplification/variable reduction were performed due to time constraints. While costly in time, in the real world such tests are mandatory: some variables simply do not help and can even degrade model performance. Given the geographical imbalance of the training data, Latitude and longitude should be tested to determine their usefulness. Some intelligent choices must be in this regard as a blind, brute force approach that tries every possible combination of predictors is unlikely to be a wise use of time or resources.

6. References

Chang, C.-C. and C.-J. Lin, 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Hastie, T, R. Tibshirani, and J. Friedman, 2001: *The Elements of STatistical Learning.* Springer, Noew York, NY, 533 pp.

Lin, H.-T. and C.-J. Lin, 2003. A study on sigmoid kernels for SVM and the trainingof non-PSD kernels by SMO-type methods. Technical report, Department of Computer Science, National Taiwan University.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, 736 pp.

Wilks, D S., 2006: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 627 pp.