J5.1

FEATURE SELECTION OF RADAR-DERIVED TORNADO ATTRIBUTES WITH SUPPORT VECTOR MACHINES

Michael B. Richman*
Budi Santosa
Theodore B. Trafalis

University of Oklahoma, Norman, OK

Tornado circulation attributes/variables derived largely from the National Severe Storms Laboratory Mesocyclone Detection Algorithm (MDA) have been investigated for their efficacy in distinguishing between mesocyclones that become tornadic from those which do not. Using a subset of the MDA variables associated with velocity yields 23 potential predictors. Previous research has shown that the discrimination ability of several of the predictors is not good and the predictor pool has strong associations among subsets of these variables. Despite these drawbacks, applications of artificial neural networks (ANN) and support vector machines (SVM) to the MDA have met with success in predicting correctly pre-tornadic circulations. One of the largest challenges in this regard is to maintain a high probability of detection (POD) while simultaneously minimizing the false alarm rate (FAR).

Both ANN and SVM are non-linear classifiers and, accordingly, the use of linear statistics to screen the predictor pool a priori may not be logically consistent. In this research, the impact of removing individual predictors is examined on the training and testing errors. Results were encouraging as exclusion of specific variables had a notable impact on the ability to distinguish accurately the tornadic from the non-tornadic circulations when viewed from misclassification rates, POD, FAR, and Heidke skill. A key finding is that inclusion of the current month number (1= January, 2 = February, …, 12 = December) in addition to a subset of MDA variables used in SVM is the most accurate set of features tested. This methodology of feature selection outperforms SVM based on the MDA alone, achieving a Heidke skill of 0.844 with a POD of 0.835 and a FAR of 0.135 with more parsimonious models.

1. INTRODUCTION

Accurate detection of tornadoes with ample warning times has been a longstanding goal of severe weather forecasters. With state-of-the-science weather radar, high speed computing and advanced signal processing algorithms, steady progress has been made on increasing the average lead-time of such warnings. An extra minute of lead-time can translate into a number of lives saved. One of the severe weather detection algorithms, created by the National Severe Storms Laboratory (NSSL) and in use at the Weather Surveillance Radar 1998 Doppler (WSR-88D), is the Mesocyclone Detection Algorithm (MDA). This algorithm uses the data stream outputs of the WSR-88D and is designed to detect storm–scale circulations associated with regions of rotation in thunderstorms. The MDA is used by meteorologists as one input in their decision to issue tornado warnings. Marzban and Stumpf (1996) show that the performance of the MDA is improved by NN post-processing of the radar data. Our work will attempt to simplify the redundancies in the MDA, helping to speed up the detection process so that the forecaster can assimilate information from the data set prior to new data streaming in. By identifying patterns associated with tornadoes in a timely fashion, the forecaster can assess the evolution of such patterns.

Kernel based methods, such as SVM, are applied to detect tornado circulations sensed by the WSR-88D radar. These methods do not make any assumptions about the data distributions. Application of the kernel methods is useful to address the problem of nonlinearity of the data in the input space since the data are mapped into a higher dimensional space where there is a high likelihood that the problem becomes linear separable (Vapnik, 1995). By exploiting the kernel method, nonlinear classifiers are found that separate the data into tornado and no tornado cases.

The current problem is a large scale one where a sizeable number of data are required for training. For such problems, the method of decomposition is an important issue to consider for solution efficiency. Hsu and Lin (2002) proposed a decomposition technique to solve the large-scale SVM training problems. The basic algorithm is a simplification of both sequential minimal optimization by Platt (1999) and SVMLight by Joachims (1999). The paper is organized as follows. In section 2, the definition of the problem is provided. In section 3, description of the data is given. In section 4, the basics of the learning machines used and our methodology are discussed. In section 5, the experimental setting is described. Section 6 provides analysis of the results and, finally, section 7 concludes the study.

2. PROBLEM STATEMENT

There are two classes of problems addressed in this research. One is meteorological, relating to tornado warnings, and the other is methodological. The two are intimately entwined for the prediction of tornadoes.

There are several challenges involved in tornado warnings from the meteorological viewpoint. The first one is tornado detection. Of those tornadoes that do occur, the number of tornadoes detected is smaller. The second one is false alarms. Algorithms detect tornado circulations more often than such circulations can be confirmed. The latter is insidious because the warnings have the potential to go unheeded by the public after a series of false alarms. Accordingly, it is

* *Corresponding author address*: Michael B. Richman, University of Oklahoma, School of Meteorology, 100 East Boyd St. Norman, OK 73019; e-mail: mrichman@ou.edu

desirable to develop a statistical learning algorithm that will maximize detection and minimize false alarms.

Prediction of tornadoes is a difficult task owing to the small scale of their circulation and their rapid production in the atmosphere. They can form within minutes and disappear just as quickly. The dynamic nature of this problem requires addressing the time-dependence nature of this application. It requires real time response to observations from radar data. Once the data are collected, algorithms look for signatures of tornadoes in near-real time, since an extra minute of lead-time can translate into a number of lives saved. The incoming radar data stream can be used for dynamic decision making to increase the lead-time in tornado forecasts. However, present day operational radar takes approximately 6 minutes to complete one volume scan. Furthermore, the spatial resolution averages close to ¼ km for Doppler radar velocity. Many tornadoes are smaller than that. Despite the challenges, lead times for tornadoes have increased from a few minutes (a decade ago) to approximately 11 minutes (with current radar), largely due to improvements in algorithms that use the radar data as inputs.

The second research problem is to develop an intelligent system that one can deal with data that have a significant noise component. ANN are considered robust classifiers in terms of input noise. However, the resulting learning optimization problem is nonconvex. An alternative to ANN is SVM where the learning optimization problem is convex. Vapnik (1995) shows, based on statistical learning theory, that SVM have better generalization properties than ANN. SVM is also robust with bounded noise in the input data (Trafalis and Al-Wazzi, 2003).

3. DATA AND ANALYSIS

The MDA data set used for this research is based on the outputs from WSR-88D radar. Any circulation detected on a particular volume scan of the radar can be associated with a report of a tornado. In the severe weather database, supplied by NSSL, there is a label for tornado ground truth that is based on temporal and spatial proximity. If there is a tornado reported between the beginning and ending of the volume scan, and the report is within a reasonable distance of a circulation detection (input manually), then the ground truth value is flagged. If a circulation detection falls within the prediction "time window" of -20 to +6 minutes of the ground truth report duration, then the ground truth value is also flagged. The key idea behind these timings is to determine whether a circulation will produce a tornado within the next 20 minutes, a suitable lead-time for advanced severe weather warnings by the National Weather Service. Any data with the aforementioned flagged values are categorized as tornado cases, with the label set to 1. All other circulations are labeled as 0, corresponding to a no tornado case.

The predictor pool employed in this study consists of 24 variables, of which 23 come from the MDA and are based on Doppler velocity data (Table 1). These variables have been used successfully by Marzban and Stumpf (1996). Additionally, a variable denoting what calendar month a circulation occurs, was added to account for the strong seasonality exhibited by each of the 23 velocity-based variables. The database was stratified into seasons and data were sampled from every season for the training of the learning machines used.

Table 1. List of s, units and ranges used in the Mesocyclone Detection Algorithm.

1. base (m) [0-12000]	13. low-level gate-to-gate velocity difference (m/s) [0-130]
2. depth (m) [0-13000]	14. maximum gate-to-gate velocity difference (m/s) [0-130]
3. strength rank [0-25]	15. height of maximum gate-to-gate velocity difference (m) [0-12000]
4. low-level diameter (m) [0-15000]	16. core base (m) [0-12000]
5. maximum diameter (m) [0-15000]	17. core depth (m) [0-9000]
6. height of maximum diameter (m) [0-12000]	18. age (min) [0-200]
7. low-level rotational velocity (m/s) [0-65]	19. strength index (MSI) weighted by average density of integrated layer [0-13000]
8. maximum rotational velocity (m/s) [0-65]	20. strength index (MSIr) "rank" [0-25]
9. height of maximum rotational velocity (m) [0-12000]	21. relative depth (%) [0-100]
10. low-level shear (m/s/km) [0-175]	22. low-level convergence (m/s) [0-70]
11. maximum shear (m/s/km) [0-175]	23. mid-level convergence (m/s) [0-70]
12. height of maximum shear (m) [0-12000]	

4. METHODOLOGY

4.1 Support Vector Machine (SVM)

Given a set of data points $\{(x_i, y_i),\ i = 1,..., \ell\}$ with $x_i \in \Re^n$ and $y_i = \pm 1$, SVM consider a problem where a classifier is sought to separate the two classes of points with maximum

margin separation (Figure 1). The SVM formulation can be written as follows (Haykin, 1999),

$$\min_{w,b,\eta} \; C \sum_{i=1}^{\ell} \eta_i + \tfrac{1}{2} \| w \|^2 \qquad (1)$$

st

$$y_i (w x_i + b) + \eta_i \geq 1 \qquad \eta_i \geq 0 \qquad i = 1,\ldots \ell$$

where *C* is a parameter to be chosen by the user, *w* is referring to the vector perpendicular to the separating hyperplane, η_i refers to the misclassification error variables and *b* is the bias of the separating hyperplane. A larger *C* corresponds to assigning a larger penalty to errors. Introducing positive Lagrange multipliers α_i, to the inequality constraints in model (1) we obtain the following dual formulation:

$$\min_{\alpha} \; \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} y_i y_j \alpha_i \alpha_j x_i x_j - \sum_{i=1}^{\ell} a_i$$

$$st \quad \sum_{i=1}^{\ell} y_i \alpha_i = 0 \qquad (2)$$

$$0 \leq \alpha_i \leq C \qquad i = 1,\ldots \ell$$

The solution of the primal problem is then given by $\mathbf{w} = \Sigma \alpha_i y_i \mathbf{x}_i$ where *w* is the vector that is perpendicular to the separating hyperplane. The free coefficient *b* can be found from $\alpha_i (y_i (\mathbf{w} \cdot \mathbf{x}_i + b) - 1) = 0$, for any *i* such that α_i is not zero.

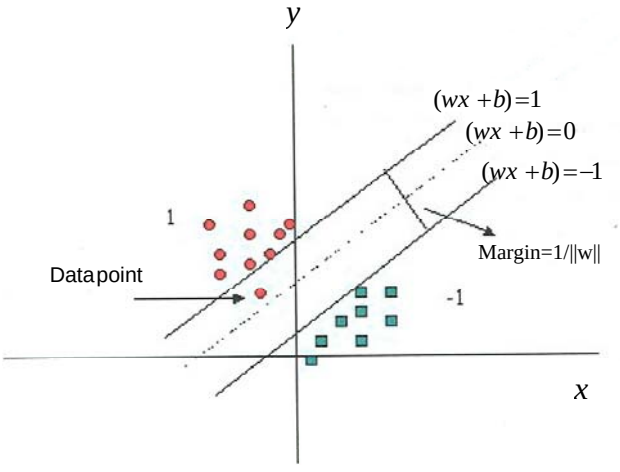


Figure 1. The geometric illustration of SVM.

SVM map a given set of binary labeled training data into a high-dimensional feature space and separate the two classes of data linearly with a maximum margin hyperplane in the feature space. In the case of nonlinear separability, each data point *x* in the input space is mapped into a higher dimensional feature space using a feature map φ. In the new space, the dot product $\langle x, x' \rangle$ becomes $\langle \varphi(x), \varphi(x)' \rangle$. A nonlinear kernel function, *k(x,x')*, can be used to substitute the dot product $\langle \varphi(x), \varphi(x)' \rangle$. The use of a kernel function allows the SVM to operate efficiently in nonlinear high-dimensional feature spaces without being adversely affected by the dimensionality of that space. Indeed, it is possible to work with feature spaces of infinite dimension (Schölkopf and Smola, 2002). Moreover, it is possible to learn in the feature space without knowing the mapping φ and the feature space *F*. The matrix $K_{ij} = \langle \varphi(x_i), \varphi(x_j) \rangle$ is called the *kernel matrix.* In general, the separating hyperplane corresponds to a nonlinear decision boundary in the input space. It can be shown that for each continuous positive definite function $k(x, y)$, there exists a mapping, φ, such that $k(x, y) = \langle \varphi(x), \varphi(y) \rangle$ for all $x, y \in \Re$, where $\Re$ is the input space (Mercer's Theorem).

There are three specific kernel functions usually used in the SVM literature: polynomial, radial basis function, and tangent hyperbolic (Haykin, 1999). In this research, only polynomial kernel functions ($k(x, y) = (x^T y + 1)^p$, where *p* is the degree of the kernel) are investigated.

4.2 Forecast Evaluation Indices for Tornado Detection

In the detection paradigm, the forecast results are assessed by using a suite of forecast evaluation indices based on a contingency table (otherwise also known as a "confusion matrix"). The confusion matrix is defined in Table 2.

TABLE 2. Confusion matrix.

			Observed	
		Yes	No	Total
Predicted	Yes	Hits (a)	False alarm (b)	Forecast Yes
	No	Misses (c)	Correct negative (d)	Forecast No
Total		Observed Yes	Observed No	

The cell counts (a, b, c, d) from the confusion matrix can be used to form forecast evaluation indices (Wilks, 1995). In this definition of the confusion matrix, one such index is the Probability of Detection, POD, which is defined as a/(a+c). POD measures the fraction of observed events that were forecast correctly. Its range is 0 to 1 and a perfect score is 1 (or 100%). Note that POD is sensitive to hits, therefore, good for rare events. However, POD ignores false alarms and it can be

improved artificially by issuing more "yes" forecasts to increase the number of hits.

False Alarm Rate, FAR, is defined as b/(a+b). FAR measures the fraction of "yes" forecasts in which the event did not occur. Its range is 0 to 1, and 0 is a perfect rate. FAR is sensitive to false alarms and it ignores misses. It can be improved artificially by issuing more "no" forecasts to reduce the number of false alarms.

The concept of skill is one where a forecast is superior to some known reference forecast (e.g., random chance). Skill ranges from –1 (anti-skill) to 0 (no skill over the reference) to +1 (perfect skill). Heidke's skill is commonly utilized in meteorology since it uses all elements in the confusion matrix and works well for rare event forecasting (e.g. tornadoes) (Doswell et al., 1990). Heidke's Skill is defined as 2(ad-bc)/[(a+b)(b+d)+(a+c)(c+d)].

5. EXPERIMENTS

The data were split into two sets: *training* and *testing*. The cases used for training are different from those used in the testing set.

The SVM, experiments were performed in the MATLAB environment. OSU SVM Classifier Matlab Toolbox by Ma et al. (2003) was used to run experiments of SVM for classification. These codes are MATLAB versions of Chih-Chung Chang and Chih Jen Lin's LIBSVM algorithm (Chang and Lin, 2003).

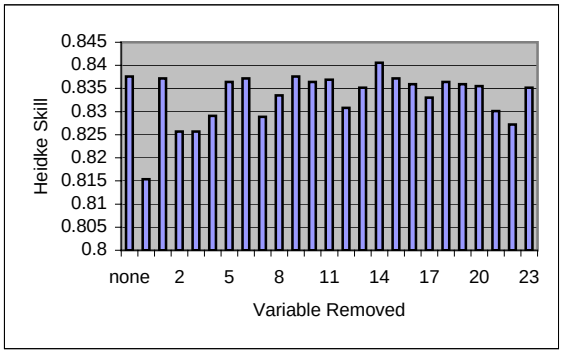
The data structure consists of an *m* by *n* matrix, where *m* refers to observations and *n* refers to variables as listed in Table 1. For the training set, m is equal to 749 and for the testing set m is equal to 18202, owing to the percentage of tornado events used herein (2%). The data are preprocessed before each method is applied. For each column, each data point (observation) was divided by the norm of the column.

Of the 24 candidate variables, reduction of this number was achieved by removing each variable. The impact of the removal was noted in the testing data through the forecast evaluation indices. Any variable resulting in an increase in skill or POD was noted and combinations of these variables were used to determine if further increases could be achieved. Similarly, for those variables resulting in a lower FAR, removal was used one at a time and then in pairs, triplets, etc. until all possible subsets were examined. Such a process has been used successfully in regression analysis (Draper and Smith 1998). By testing on POD, FAR and Heidke skill, the SVM could be tailored to optimize a specific portion of the forecast.

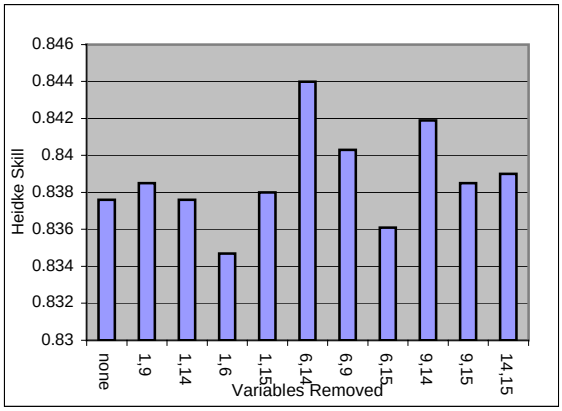
6. RESULTS

Figures 3 through 5 depict different performance aspects of the aforementioned forecast evaluation indices for each index.

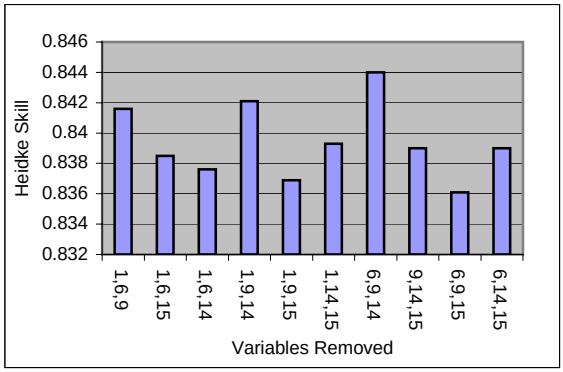
Assessment of a successful forecast was made through examination of Heidke Skill. The full variable SVM and with single variables removed in shown in Fig. 3a. The full model (with no variables removed) has a skill of 0.8355. Removal of certain variables, such as month number, reduces skill substantially; therefore, calendar month would not be a candidate for discarding. Conversely, removal of variable 14 increases the skill and should be removed. All variables that did not lead to a decrease in skill when removed were tested further. These included the following subset of MDA variables: 1 (mesocyclone base height), 6 (height of maximum diameter), 9 (height of maximum rotational velocity), 14 (maximum gate to gate velocity difference), and 15 (height of maximum gate to gate velocity difference). For all of the single variables tested, removal of variable 14 led to the best skill (0.841). When all possible pairs of the subset were tested (Fig. 3b), the pair of 6 and 14 led to the most improvement with a skill of 0.844. When all possible triplets of the subset were tested (Fig. 3c), the highest skill remained at 0.844 for removal of variables 6, 9 and 14. Such a model would be preferable as it is more parsimonious. When all subsets of 4 variables were removed (Fig. 3d), the skill remained at 0.844 by omitting MDA variables 1, 6, 9 and 14. The same result was found for removal of all five predictors (Fig 3d). Therefore, this would seem to be the most desirable model as it was most compact.



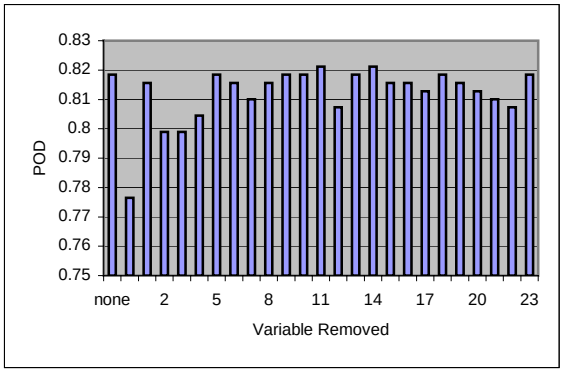
(a)



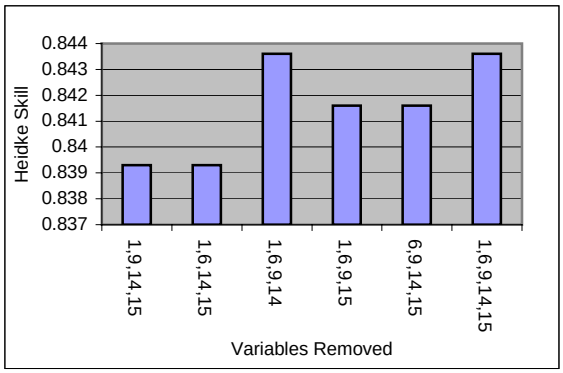
(b)



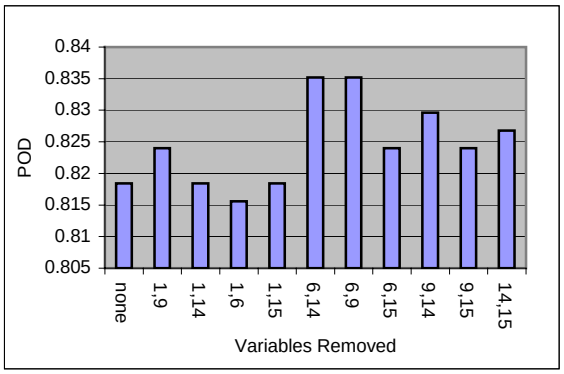
(c)



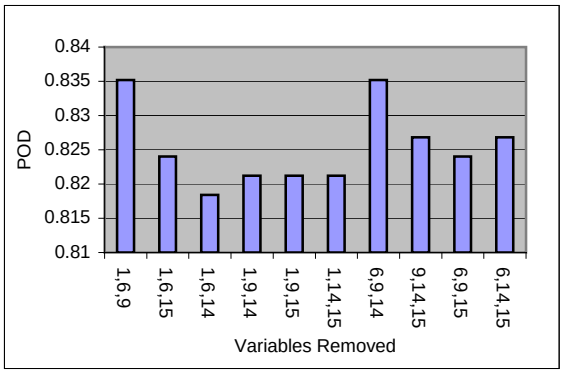
(a)



(d)



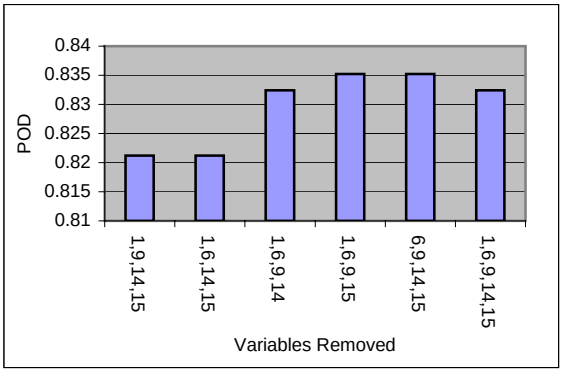
(b)



(c)

Figure 3. Skill for removal of no variables, month number and MDA variables 1 – 23.

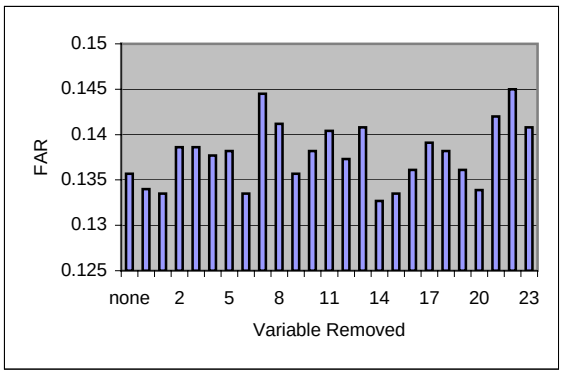
Next, the removal of these variables was examined in more detail for their impact on POD and FAR. Results on POD for removing single variables can be seen in Fig. 4a. Note that removal of MDA variable 11 or 14 increased POD over the full model. In these experiments, the same subset of five variables was removed in pairs, triplets, etc. in an all-subsets approach. The results of removing the various pairs (Fig. 4b) indicate variables 6 and 9 or 6 and 14 increased the POD from 0.818 to 0.835. When triplets were removed (Fig 4c), variables 1, 6 or 9 or 6, 9 and 14 resulted in the same increase in POD and for the doubles. However, the removal of triplets results in a more parsimonious model. For the set of four variables removed (Fig. 4d), either variables 1, 6, 9 and 15 or 6, 9, 14 and 15 yielded the same high POD. Note that when all five variables were removed, the POD decreased slightly.



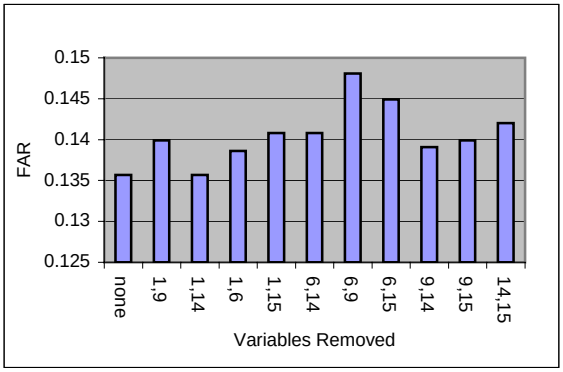
(d)

Figure 4. POD for removal of no variables, month number and MDA variables 1 – 23.

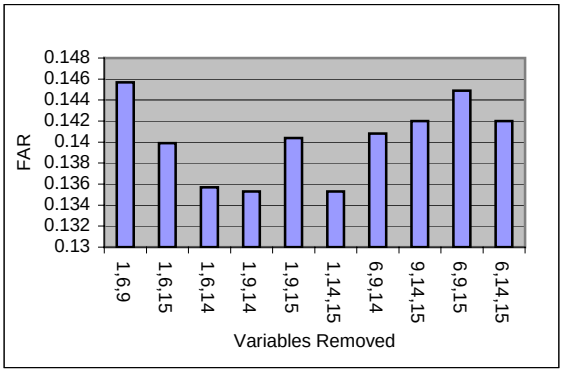
The FAR was examined as another arbiter of success. Results for removing single variables (Fig. 5a) had different results from either those for either skill or POD. Excluding the month number lowered the FAR. Therefore, the month number resulted in an overforecasting bias. However, when all pairs of variables identified previously were tested (Fig. 5b), results indicated that, in all but one case, the FAR increased. However, when triplet subsets were removed (Fig. 5c), the FAR decreased for sets 1, 9 and 14 or 1, 14 and 15. Finally, when subsets of four variables were removed (Fig. 5d), the FAR decreased to the same (lower) value for variables 1,9, 14 and 15 or 1, 6, 14 and 15.



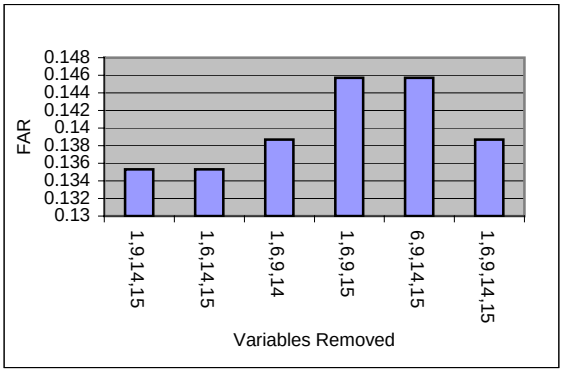
(a)



(b)



(c)



(d)

Figure 5. FAR for removal of no variables, month number and MDA variables 1 – 23.

Taken collectively, the skill and POD results are similar and suggest that at least 4 of the 23 MDA variables can be discarded safely. However, if the goal is to minimize

FAR (at the expense of lower detection and skill), then the subsets of variables to be removed do not overlap completely with those that increase skill and POD.

In order to determine if selected removal of variables let to statistically significant increases in skill, the testing data set was bootstrap resampled 30 times. Selected results of the experiments are shown in Figure 6. Boxplots suggest that the gain in skill through removal of variables 6 and 14 yielded skill that overlapped the interquartile range (IQR) of the solution for no variables removed. However, removal of variables 6 and 14 solution was significantly more skillful than removal of variables 1,6,9,and 14 or 1,6,9,14, and 15 as the IQR did not overlap.

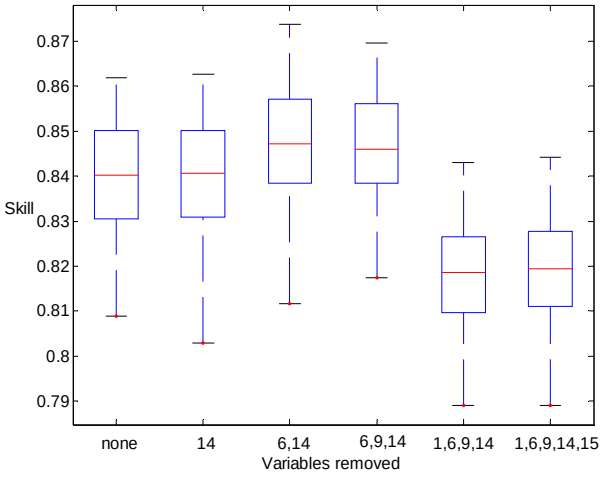


Figure 6. Boxplots of Heidke Skill for removal of selected variables.

7. CONCLUSIONS

As with any ranking scheme, the process of examining all possible subsets of skill-related variables for model improvement has positive and negative aspects. This is a form of data mining and it is possible in the process to lose sight of the goal to find a model that has physically meaning variables and is parsimonious. When the variables to be removed are examined, one discovers that the majority of those are related to height of the base or of the height of rotation. It is noteworthy that four of the five variables are height related. Such height variables do not add information that the SVM can process into more skillful forecasts of those mesocyclone circulations that remain nontornadic compared to those that become tornadic. It is equally significant that the SVM model can forecast those circulations that become tornadic from those that do not with the high amount of skill and low FAR found, particularly given the realistically low percentage of tornadoes (2 percent) in these experiments.

ACKNOWLEDGMENTS

The present work has been partially supported by the NSF grant EIA-020568.

REFERENCES

Chang C.C. and C.J. Lin, 2003: *LIBSVM: A Library for Support Vector Machines* http://www.csie.ntu.edu.tw/-cjlin/libsvm

Doswell,III, C.A., R. Davies-Jones, and D. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables, *Weather and Forecasting,* **5** 576-585.

Draper, N.R. and H. Smith, 1998: *Applied Regression Analysis*.3rd Ed. John Wiley and Sons, Inc., New York.

Haykin, S., 1999: *Neural Networks: A Comprehensive Foundation,* 2nd edition, Prentice-Hall, Uppersaddle River, NJ.

Hsu, C.W. and C.J. Lin, 2002: A simple decomposition method for support vector machines, *Machine Learning*, **46**, 291-314.

Joachims, T., 1999: Making large-scale SVM learning practical, *Advances in Kernel Methods – Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola (eds.), MIT Press.

Ma, J., Y. Zhao, and S. Ahalt, 2003: *OSU SVM Classifier Matlab Toolbox*, available at http://www.ece.osu.edu/~maj/osu_svm/

Marzban, C. and G.J. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar - derived variables, *Journal of Applied Meteorology*, **35**, 617-626.

Trafalis, T.B., B. Santosa and M.B. Richman, 2003: Tornado detection with kernel-based methods, *Intelligent Engineering Systems Through Artificial Neural Networks*, C.H. Dagli, A.L. Buczak, J. Ghösh, M. Embrechts and O. Ersoy, (eds.), ASME Press, **13**, 677-682.

Trafalis, T.B. and S.A. Alwazzi, 2003: Robust optimization in support vector machine training with bounded errors, *Proceedings of International Conference in Neural Networks, IJCNN 2003,* Portland, Oregon, IEEE Press.

Vapnik, V.N., 1995: *The Nature of Statistical Learning Theory*, Springer, New York.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*, Acad emic Press, London.