Tornado Detection with Support Vector Machines

Theodore B. Trafalis[1], Huseyin Ince[1], and Michael B. Richman[2]

[1] School of Industrial Engineering, University of Oklahoma,
Norman OK 73019, USA,
ttrafalis@ou.edu,
http://www.lois.ou.edu
[2] School of Meteorology, University of Oklahoma,
Norman OK 73019, USA,
mrichman@ou.edu

Abstract. The National Weather Service (NWS) Mesocyclone Detection Algorithms (MDA) use empirical rules to process velocity data from the Weather Surveillance Radar 1988 Doppler (WSR-88D). In this study Support Vector Machines (SVM) are applied to mesocyclone detection. Comparison with other classification methods like neural networks and radial basis function networks show that SVM are more effective in mesocyclone/tornado detection.

1 Introduction

The National Weather Service (NWS)uses several severe weather detection algorithms. One of them is the Mesocyclone Detection Algorithm(MDA). It is based on empirical rule based algorithms and works on the WSR-88D. The skill of the MDA algorithm is rather low. For example, the percentage of observed mesocyclones that were correctly forecast is below fifty percent. Additionally, there are a large number of mesocyclones forecast that do not occur. Moreover, modeling of a complex dynamical system with a closed mathematical expression is not an easy task. Owing to these two factors, there is a need to develop or use new techniques to address the problem.

One of the techniques that does not rely on assumptions about the underlying probability distribution governing the input data is Support Vector Machines(SVM). This is in contrast to other types of models that assume the data follow the normal distribution, like those based on the linear discriminant analysis (DA) method. Furthermore, the SVM classification method is more robust than other techniques such as neural networks (NN) and radial basis networks [14].

The MDA algorithm and SVM methods will be briefly explained in the next two sections. Then a comparison with other techniques is discussed in section 4. Section 5 concludes the paper.

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2660, pp. 289–298, 2003.
© Springer-Verlag Berlin Heidelberg 2003

2 The Mesocyclone Detection Algorithm (MDA)

In this section we present one of the detection algorithms developed by the National Severe Storms Laboratory (NSSL), the MDA (Mesocyclone Detection Algorithm). The data used in this study are intermediate data computed by this algorithm. The NSSL has developed the NSSL MDA for the WSR-88D system to automatically detect and diagnose the Doppler radar radial velocity patterns associated with all storm-scale (1-10km diameter) vortices in thunderstorms, rather than defining the strength thresholds at the very first analysis step used by 88D B9MDA. The first step of the algorithm is to preprocess the Doppler velocity data. Noisy data, such as velocities whose reflectivity values are below a preset threshold (typically 0 - 20 dBZ), are deleted. Next, the NSSL MDA's automated vortex detection techniques set the initial strength thresholds to be much lower, and classifications and diagnosis are performed on the properties of the four-dimensional detections. The algorithm first processes data at the one-dimensional (1D) level; shear segments of cyclonic azimuthal shear are detected. Next, the shear segments are horizontally associated to form two-dimensional (2D) features. The NSSL MDA then uses vertical association to create three-dimensional (3D) detections at the end of each volume scan. Finally, time association and tracking are employed to complete the process. More information about how MDA works can be found in [5,13].

3 Support Vector Machines: A Brief Review

SVM is a learning machine developed by Vapnik [3,14] based on statistical learning theory [15]. In the case of classification [1,3,6,7], we try to find an optimal hyperplane that separates two classes of data points D_ℓ as shown in Figure 1. The objective is to establish an equation of the hyperplane that divides D_ℓ into two sets S_1 and S_2, leaving all the points of the same class on the same side while maximizing the minimum distance between either of the two classes and the resulting hyperplane [3,1]. In order to find the hyperplane that has the maximum margin between the two classes while at the same time minimizes the misclassification error on D_ℓ, one has to solve the following Quadratic Programming (QP) optimization problem:

$$\begin{aligned} \min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i \\ \text{Subject to} \\ y_i \cdot (w \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, i = 1, 2, ..., l \\ \xi_i \geq 0, \end{aligned} \tag{1}$$

where $w \in \Re^d$ is the vector normal to the separating hyperplane, $b \in \Re$ is the offset with respect to the origin, ξ_i are the slack variables that measure the empirical misclassification error, $y = \pm 1$ and C is the regularization parameter [4,

16,11,15] defining the trade off between margin and empirical error. By assigning a Lagrangian multiplier α_j to each constraint and by introducing the variables $\alpha = (\alpha_1, \alpha_2, ..., \alpha_l)$, the matrix $\mathbf{K}_{ij} = (y_i y_j \mathbf{x}_i \mathbf{x}_j)$, and vectors $e = (1, 1, ..., 1)$, and $y = (y_1, y_2, ..., y_l)$, the dual problem can be formulated in closed form as:

$$\begin{aligned} \max \ & -\alpha^T e + \frac{1}{2}\alpha^T \cdot \mathbf{K} \cdot \alpha \\ & \text{Subject to} \\ & \alpha^T \cdot y = 0 \\ 0 \leq \alpha_j \leq C \quad & \forall j = 1, ..., l. \end{aligned} \tag{2}$$

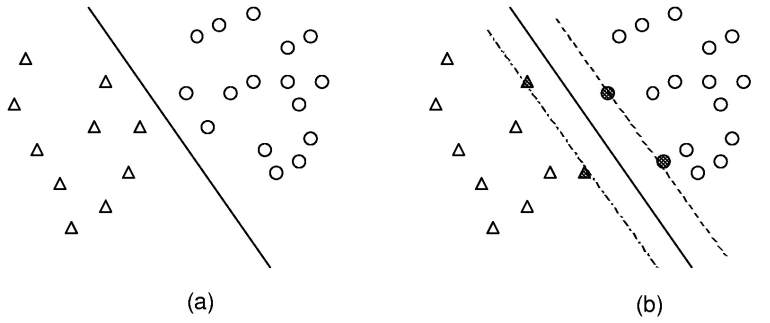

Fig. 1. Separating hyperplane and optimal separating hyperplane. Both solid lines in (a) and (b) separate the two identical sets described by circles and triangles. But the solid line in (b) leaves the closest points (filled circles and triangles) at the maximum distance. The distance between dashed lines in (b) gives the maximum margin.

The above approach can be generalized in the case of nonlinear separating surfaces mapping the input data $\{x_i\}_{i=1}^{l}$ into a higher dimensional feature space through the use of a feature map $\phi : \Re^d \to F$. Then a separating hyperplane (if it exists) can be found in that space.

Our objective is to determine a discriminant function

$$f(x) = sign(w \cdot \phi(x) + b) = \left\{ \begin{matrix} +1 \text{ if } x \in S_1 \\ -1 \text{ if } x \in S_2 \end{matrix} \right\}, \tag{3}$$

where $\phi : \Re^d \to F$ is a map of $\Re^d$ into the feature space F.

Specifically, the corresponding SVM optimization problem is as follows:

$$\begin{aligned} \min \ & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell} \xi_i \\ & \text{Subject to} \\ y_i \cdot (w \cdot \phi(x_i) + b) \geq 1 - \xi_i, & i = 1, 2, ..., l \\ & \xi_i \geq 0. \end{aligned} \tag{4}$$

The solution of the above problem can be expressed as a linear combination of the ϕ-images of the data points [9], i.e.

$$w = \sum_{i=1}^{l} \alpha_i \cdot y_i \cdot \phi(x_i). \tag{5}$$

Therefore, if we define a dot product in the feature space [9] as $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$, then problem 4 can be expressed as in 2 but with $K_{ij} = y_i \cdot y_j \cdot k(x_i, x_j)$.

Combining equation (3) and (5), $f(x)$ can be expressed as follows

$$f(x) = sign\left(\sum_{i=1}^{l} \alpha_i \cdot y_i \cdot \phi(x_i) \cdot \phi(x) + b\right) \tag{6}$$

or equivalently, using the kernel function as

$$f(x) = sign\left(\sum_{i=1}^{l} \alpha_i \cdot y_i \cdot k(x_i, x) + b\right) \tag{7}$$

The important data points are the ones for which $\overline{\alpha_i} > 0$ where $\overline{\alpha}$ denotes the optimal solution of the above problem 1. Those points are called support vectors and provide a sparse representation of the discriminant function. Usually, the number of support vectors is smaller than the number of data points. There are several kernel function to choose, e.g. radial basis function, polynomial function, etc. [16].

To solve this quadratic programming problem, several decomposition algorithm have been proposed [12,2,8]. In this paper, an SVM learning decomposition algorithm, SVMTorch [2] is applied to mesocyclone detection since the number of data is large (3768 data points).

4 Experiments

The circulation database used for SVM method was 'truthed' to determine which circulations were associated with reports of tornado events. This was the first step to obtain the actual target value for each observation. Then, the missing observations were removed from the dataset. 23 variables were chosen as inputs. These variables are intermediate outputs of WSR-88D and NSSL mesocyclone algorithms. They are used by the National Weather Service to diagnose mesocyclones during the severe storm warning operations.

In a recent paper [13],the reduction of input variables is suggested by using statistical methods or principal component analysis. On the other hand, some authors suggest that these techniques do not improve the prediction of the tornados [5]. Because of this, all variables will be used in this paper. The definition of the variables and more information can be found in [13,5].

Table 1. Confusion Matrix

		Observed		
		Yes	No	Total
Forecast	Yes	Hits (a)	False Alarm (b)	Forecast Yes
	No	Misses (b)	Correct Negatives (d)	Forecast No
Total		Observed Yes	Observed No	Total

The performance of SVM methods is evaluated by using a suite of forecast evaluation indices based on a contingency table (otherwise also known as a "confusion matrix"). More information about some of these measures can be found in [13,5,10,17]. The confusion matrix is shown in Table 1.

In this definition of the confusion matrix, the Probability of Detection, POD, can be defined as

$$POD = \frac{a}{(a+c)}. \tag{8}$$

POD measures the fraction of observed events that were correctly forecast.Its range is 0 to 1 and a perfect score is 1 (or 100%). Note that POD is sensitive to hits, good for rare events. POD ignores false alarms and can be improved artificially by issuing more "yes" forecasts to increase the number of hits.

False Alarm Rate, FAR, can be defined as

$$FAR = \frac{b}{(a+b)}. \tag{9}$$

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 Critical Success Index, CSI, is defined as

$$CSI = \frac{a}{(a+c+b)}. \tag{10}$$

CSI measures the fraction of observed and/or forecast events that were correctly forecast. Its range is 0 to 1 with a perfect score being 1. CSI is sensitive to hits, penalizes both misses and false alarms. It does not distinguish the source of forecast error and it depends on the climatological frequency of events (worse scores for rarer events) since some hits can occur purely due to random chance.

Accuracy is defined as

$$Accuracy = \frac{(a+d)}{total}. \tag{11}$$

Accuracy measures the fraction of all forecasts that were correct, which makes it a seemingly intuitive measure. The range is 0 to 1 with 1 being best. However, it can be misleading since it is heavily influenced by the most common category, usually "no event" in the case of severe weather.

Bias is defined as

$$Bias = \frac{(a+b)}{(a+c)}. \tag{12}$$

Bias measures the ratio of the frequency of forecast events to the frequency of observed events. The range is from 0 to infinity. A perfect score is 1. Bias indicates whether the forecast system has a tendency to underforecast (bias $<$ 1) or overforecast (bias $>$1) events. It does not measure how well the forecast corresponds to the observations; it measures only relative frequencies.

Probability of False Detection, POFD, is defined as

$$POFD = \frac{b}{(b+d)}. \tag{13}$$

POFD measures the ratio of false alarms to the total number of no observations. The probability of false detection is a measure of inaccuracy with respect to the observations and provides a measure of the extent to which the forecasts provide a false warning for the occurrence of an event. POFD varies from 0 to 1. A perfect score is zero.

Hanssen and Kuipers discriminant (true skill statistic), H-K Skill,is a measure of the improvement of a forecast over some reference forecast (e.g., random forecast).

$$H - KSkill = \frac{a}{(a+c)} - \frac{b}{(b+d)}. \tag{14}$$

H-K Skill measures the ability of the forecast to separate the "yes" cases from the "no" cases. It can also be interpreted as accuracy(events) + accuracy(non-events) - 1. H-K Skill ranges from -1 to 1 with 0 indicating no skill. A perfect statistic value is 1. The advantage of H-K Skill is that it uses all elements in contingency table. It does not depend on climatological event frequency. For rare events, H-K Skill is weighted heavily toward the first term (same as POD).

Odds Ratio, OR, is a newer statistic for forecast evaluation.

$$OR = \frac{(a*d)}{(c*b)}. \tag{15}$$

OR measures the ratio of the probability of making a hit to the probability of making a miss or false alarm. For the OR, the range is 0 to infinity, 1 indicates no skill. A perfect score is infinity. OR gives better scores for rarer events.

Comparison of SVM, NN, DA and the rule-based MDA based on these indices are given in Tables 3 and 4. The trade-off value(C), kernel functions and kernel related parameter have to be determined. The trade-off value is set to 1000. The radial basis kernel function is used that is shown in equation (16).

$$k(x, y) = exp(-\frac{1}{2\sigma^2}\|x - y\|^2) \tag{16}$$

We have used different σ's to find the best fit. Table 2 shows the results for different parameters. As it can be seen from the Table 2, the tornado dataset verification indices are very sensitive to this parameter. The results show that accuracy increases up to a σ of 7.25 and then decreases slightly. A similar behavior can be seen for CSI. The POD is relatively insensitive to σ over the range tested, though the highest values occur between σ values of 2.5 and 7.25. The Bias indicates that all of the values lead to underforecasting, though it is minimized at the two lowest values of σ. The odds ratio shows a clear advantage at the two largest σ. Both the FAR and POFD should be lowest for the best results. These measures are lowest for the two largest values of σ. The skill score is high over the σ range 2.5 to 15, with a maximum of 57.13 percent at a σ of 7.5.

Table 2. The Accuracy, CSI, POD, Bias, ODDS, FAR, POFD and H-K Skill for different kernel parameters

σ	ACCURACY (%)	CSI (%)	POD (%)	Bias (%)	ODDS (Ratio)	FAR (%)	POFD (%)	H-K Skill (%)
1.25	83.60	44.65	60.13	94.76	14.08	36.43	9.77	50.36
2.5	85.54	48.70	62.38	90.48	19.54	30.99	7.93	54.45
7.25	88.27	53.46	61.24	75.80	37.44	19.19	4.11	57.13
15	87.98	51.75	58.71	72.15	36.53	18.61	3.79	54.92

Tables 3 and 4 show that the SVM method outperforms the NN, DA and MDA rule based algorithms. For comparison, 10 different training and validation sets are used. The average CSI, POD and FAR are provided.

Since the MDA employs a rule based algorithm, it does not explain the relation between the tornado and the input variables. DA is a classification method with several assumptions such as normality and the homoscedasticity of the distribution. Most of the time those assumptions are violated. The theory behind the NN does not make any assumption about the distribution of the empirical data. Because of this, it outperforms the MDA rule based algorithm and DA. Despite this, the error function for NN is not convex and most of the time, the solution is a local optimum. In contrast, the SVM approach provides the most optimal solution of the methods tested because the SVM training

Table 3. The validation CSI for SVM, NN, MDA and DA for ten different training and validation sets. The values for MDA, DA and NN are taken from [5].

Seed	CSI_{MDA}	CSI_{DA}	CSI_{NN}	CSI_{SVM}
1	26.9	31	36.9	47.78
2	24	29.2	35.7	47.49
3	24.7	28.1	38.3	50.2
4	28.7	27.7	33.6	49.38
5	27.4	30	34.2	47.5
6	28	28.8	32.5	44.93
7	29.9	26.1	33.1	51.62
8	21.3	28.7	29.1	48.19
9	27.7	30.6	37.8	50.5
10	21.5	26.5	31.7	49.41
Average	26.01	28.67	34.29	48.7

Table 4. The POD and the FAR are also shown for SVM and NN. The values for MDA, DA and NN are taken from [5].

Seed	POD_{NN}	FAR_{NN}	POD_{SVM}	FAR_{SVM}
1	51.2	43.10	62.53	33.06
2	50	44.4	60	30.5
3	55	44.3	64.05	30.11
4	58.8	56.1	60.51	27.13
5	50	48.1	60.25	30.81
6	47.5	49.3	59.49	35.26
7	52.5	52.8	68.61	32.42
8	46.2	56	60.51	29.71
9	60	49.5	64.56	30.14
10	47.5	51.3	63.29	30.75
Average	51.87	49.49	62.38	30.99

algorithm is convex. Hence, it outperforms the other three methods. The details of the comparison between these techniques are based on the forecast verification statistics. The results indicate that both the CSI and POD are considerably larger for SVM, whereas the FAR is reduced dramatically to less than half of value found for the currently deployed NN algorithm (Table 2, σ =7.25).

5 Conclusions

We have applied a novel approach to mesocyclone and tornado detection. In the forecasting of tornadoes, accuracy or detection, length of lead time and a low false

alarm rate are crucial elements for success. If mesocyclones are predicted with an algorithm containing these attributes, that will help to minimize the loss of life. Comparison of the four methods (MDA, DA, NN, SVM) has been performed. Currently, the WSR-88D Doppler radar uses the MDA rule based algorithm and NN. These existing algorithms have moderate detection probabilities and moderate false alarm rates. The moderate FAR is particularly insidious, as it tends to lull the public into a false sense of complacency concerning tornado warnings. The SVM algorithm is the most accurate algorithm in terms of the highest values of CSI, POD and the lowest of FAR (less than half the FAR of the NN technique for one SVM model tested). In order to improve the capability of WSR-88D Doppler radar to detect mesocyclones, this work has established that the SVM algorithm can be used successfully. Accordingly, SVM should be tested more fully to determine if these results generalize well in other situations and geographical locations. We are currently investigating refining SVM modeling to improve these results with data being assimilated from an array of radars.

Acknowledgment. This research is partially funded by NSF EIA-0205628. The authors would like to thank Greg Stumpf for providing the datasets and Alexander Malyscheff for helping with the final manuscript.

References

1. Burges, C.J.C.: A Tutorial on Support Vector Machines for Pattern Classification. Data Mining and Knowledge Discovery **2**(2) (1998) 121–167
2. Collobert, R., Bengio, S.: SVMTorch: Support vector machines for large-scale regression problems. Journal of Machine Learning Research **1** (2001) 143–160
3. Cortes, C., Vapnik, V.: Support Vector Networks. Machine Learning **20** (1995) 273–297
4. Girosi, F.: An Equivalence Between Sparse Approximation and Support Vector Machines. Neural Computation **10**(6) (1998) 1455–1480
5. Marzban, C. , Stumpf, G.J.: A Neural Network for Tornado Prediction Based on Doppler Radar-Derived Attributes. Journal of Applied Meteorology **35**(5) (1996) 617–626
6. Osuna, E., Freund, R., Girosi, F.: Training Support Vector Machines: An Application to Face Detection. Proc. Computer Vision and Pattern Recognition '97 (1997) 130–136
7. Pontil, M., Verri, A.: Properties of Support Vector Machines. Technical Report, Massachusetts Institute of Technology, Artificial Intelligence Laboratory (1997)
8. Rifkin, R.: SvmFu a Support Vector Machine Package. http://five-percent-nation.mit.edu/PersonalPages/rif/SvmFu/index.html (2000)
9. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press, Cambridge, Massachusetts (2002)
10. Stephenson, D.B.: Use of the "Odds Ratio" for Diagnosing Forecast Skill. Weather and Forecasting **15**(4) (2000) 221–232
11. Evgeniou, T., Pontil, M., Poggio, T.: Regularization Networks and Support Vector Machines. Advances in Computational Mathematics **13** (2000) 1–50

12. Trafalis, T.B., Evgeniou, T., Ince, H.: Hierarchical Methods for Training Support Vector Machines with Very Large Datasets. In: Proceedings of the 30th International Conference on Computers and Industrial Engineering, Tinos Island, Greece (2002)
13. Trafalis, T.B., White, A., Fras, A.: Data Mining Techniques for Tornadic Pattern Recognition. In: C.H. Dagli, A.L. Buczak, J. Ghosh, M. Embrechts, O.Ersoy, and S. Kercel, editors, Intelligent Engineering Systems Through Artificial Neural Networks **10** ASME (2000) 455–460
14. Vapnik, V.: Estimation of Dependencies Based on Empirical Data. Springer Verlag (1982)
15. Vapnik, V.: Statistical Learning Theory. Wiley (1998)
16. Wahba, G.: Splines Models for Observational Data. Series in Applied Mathematics **59** SIAM (1990)
17. Wilks, D.S.: Statistical methods in the Atmospheric Sciences. Academic Press (1995)