FEATURE SELECTION WITH SUPPORT VECTOR MACHINES APPLIED ON TORNADO DETECTION

Budi Santosa*

ABSTRACT

In this paper, a linear programming support vector machine which is based on L_1-norm is applied to do feature selection in the tornado data set. The data is the ouputs of Weather Surveillance Radar 1998 Doppler (WSR-88D). The approach is evaluated based on the indices of probability of detection, false alarm rate, bias and Heidke skill. 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.

Keywords: classification, detection, feature selection, bayesian neural networks, machine learning, linear programming support vector machines, linear discriminant analysis, performance indices.

.

ABSTRAK

Dalam paper ini, formulasi linear programming support vector machine yang didasarkan pada L_1-norm diaplikasikan untuk melakukan feature selection pada data feature tornado yang merupakan keluaran dari radar. Pendekatan yang dipakai di sini akan dievaluasi dengan menggunakan beberapa parameter yaitu probability of detection, false alarm rate, bias dan Heidke skill. Feature/atribute sirkulasi tornado yang diperoleh dari National Severe Storms Laboratory Mesocyclone Detection Algorithm (Oklahoma, USA) sudah diinvestigasi mengenai kemampuannya untuk membedakan antara mesocyclones yang menjadi tornado dengan yang tidak. Riset-riset sebelumnya menunjukkan beberapa atribut/feature tidak memberikan kontribusi signifikan untuk membedakan antara mesocyclone yang menjadi tornado dan yang tidak. Selain itu, ada asosiasi yang kuat antar atribut secara individu.

Kata kunci: klasifikasi, deteksi, pemilihan fitur, bayesian neural networks, machine learning, linear programming support vector machines, linear discriminant analysis, indeks performansi.

1. INTRODUCTION

A severe weather detection algorithm, created by the National Severe Storms Laboratory and in use at the Weather Surveillance Radar 1998 Doppler (WSR-88D), is the Mesocyclone Detection Algorithm (MDA). This algorithm uses the 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 (Marzban and Stumpf 1996) show that the performance of MDA is improved by ANN post-processing of the radar data.

In this paper, Linear Programming Support Vector Machine (LP-SVM), SVM, Linear Discriminant Analysis (LDA) and Bayesian Neural Network (BNN) are applied to detect tornado circulations sensed by the WSR-88D radar.

LP-SVM is used for feature selection and prediction. Whereas, SVM, LDA and BNN (MacKay 1992a,b) are used for prediction after relevant features are identified from applying LP-SVM. By feature selection, we can identify the most relevant attributes/features for tornado detection and decrease the dimensionality of the data. LP-SVM is used successfully in classification and relevant feature identification in molecular profiling data (Bhattacharyya dkk, 2003). BNN has been applied successfully for tornado detection (Theodore dkk. 2004).

The paper is organized as follows. Section 2 describes the data, whereas, Section 3 describes the basics of the learning machines used and our methodology is discussed. In section 4, the experimental setting is described. Section 5 provides sensitivity analysis of the various learning networks for several forecast evaluation indices. Finally, Section 6 concludes the paper.

2. DATA AND ANALYSIS

The MDA data set used for this research is based on the outputs from the WSR-88D radar. Tornadoes are one of the three categories of severe weather. The others are: hail greater than 1.9 cm in diameter and non-tornadic winds in excess of 25 ms^{-1}. Any circulation detected on a particular volume scan of the radar data can be associated with a report of a tornado. In the severe weather database supplied by NSSL, there are two truth numbers, the first for tornado ground truth, and the second for severe weather

* Department of Industrial Engineering, Institute Technology of Sepuluh Nopember Surabaya
E-mail: budi_s@ie.its.ac.id

ground truth (Marzban and Stumpf 1996). Tornado ground truth is based on temporal and spatial proximity of the circulation to the radar. If there is a tornado reported between the beginning and ending of the volume scan, and the report is within 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 flagged also. The 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 (1). All other circulations are given as 0, corresponding to a no tornado case.

The predictor pool employed in this study consists of two data sets. The first one has month number and 17 attributes based on Doppler velocity data (Table 2). These same attributes have been used successfully by Marzban and Stumpf (Marzban and Stumpf 1996) in their work on post-processing radar data. The second one has 34 attributes, which contains MDA attributes and "near storm environment" (NSE) attributes as additional attributes (features) derived from Doppler (Lakshmanan et al. 2005).

3. METHODOLOGY

3.1 Support Vector Machines (SVM)

Consider a problem with two classes. Data, $\{(x_1,y_1),.., (x_N,y_N)\}$,consist of *N* example vectors, $x_i \in R^p$. The label, $y_i \in \{+1,-1\}$, indicates whether the example vector x_i is equated with class 1 or with class 2. In SVM, a classifier is sought to separate two classes of points. The SVM formulation can be written as follows (Haykin 1999),

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

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

where *C* is a parameter to be chosen by the user, *w* is the vector perpendicular to the separating hyperplane, *b* is the offset and the η_i are referring to the slack variables for possible infeasibilities of the constraints. A larger *C* corresponds to assigning a larger penalty to errors.

3.2 Linear programming Support Vector Machines (LP-SVM)

Consider a problem with two classes. Data, $\{(x_1,y_1),.., (x_N,y_N)\}$,consist of *N* example vectors, $x_i \in R^p$. The label, $y_i \in \{+1,-1\}$, indicates whether the example vector x_i is equated with class 1 or with class 2. In LP-SVM, we seek a hyperplane, $wx + b = 0$, that separates the two class of points, where w is a weight vector in R^p, *b* is an offset term in *R*. A classifier is the hyperplane which satisfies the *N* inequalities $y_i(w^Tx_i + b) \geq 0$ $\forall i \in \{1,.., N\}$. The learning problem is to estimate the optimal weight vector $w^*$and offset b^*. Given this hyperplane, a vector *x* is assigned to a class based on the sign of the corresponding decision function. If sign$(w^{*T}x + b^*) = +1$, *x* is identified with class 1, otherwise, it is assigned to class -1. The problems of classification and relevant feature identification can be solved concurrently by considering a sparse hyperplane, one for which the weight vector *w* has few non-zero elements. Recall that the class of a vector *x* is assigned according to the sign of , where *z* is defined as

$$z = w^T x + b = \sum_{p=1}^{P} w_p x_p + b = \sum_{w_p \neq 0} w_p x_p + b \quad ....(2)$$

If a weight vector element is zero, $w_p = 0$, then feature *p* in the example vector does not decide the class of *x* and is thus "irrelevant". Only a feature for which the element is non-zero, $w_p \neq 0$, contributes to sign(*z*) and is thus useful for discrimination. Accordingly, the problem of defining a small number of relevant features can be thought of as synonymous with identifying a sparse hyperplane. The procedure of learning a sparse hyperplane can be formulated as an optimization problem. Minimizing the L_0 norm of the weight vector, $||w||_0$, minimizes the number of non-zero elements. The L_0 norm is defined as $||w||_0$= number of $\{p: w_p \neq 0\}$. Unfortunately, minimizing an L_0 norm is NP-hard. However, a tractable, convex approximation is to replace the L_0 norm with the L_1 norm (Donoho and Huo 1999). Minimizing the L_1 norm of the weight vector, $||w||_1$, minimizes the sum of the absolute magnitudes of the elements and sets most of the elements to zero. The L_1 norm is

$$||w||_1 = \sum_{p=1}^{P} | w_p |$$

The learning optimization problem becomes

$$\min_{w,b} \ \|w\|_1$$

$$st. \quad y_i\,(wx_i + b) \ge 1 \qquad i = 1,\dots N \qquad ....(3)$$

$||w||_1 = \sum_{p=1}^{P} |w_p|$, where $|w_p| =$ sign$(w_p)w_p$.

Problem (4) can be viewed as a special case of minimizing a weighted L1 norm, $\min_{\mathrm{w}} \sum_{p=1}^{P} a_p\,|w_p|$, in which the vector of weighting coefficients, *a,* is a unit vector with $a_p = 1$, $\forall p \in \{1,..,P\}$. In other words, all features are presumed to be equally good relevant feature candidates. Prior knowledge about the (un)importance of feature *p* can be encoded by specifying the value of a_p. If the data are not linearly separable, misclassification can be accounted for by adding a non-negative slack variable η_i to each constraint and introducing a weighted penalty term to the objective function

$$\min_{w,b} \ \|w\|_1 + C\sum_{i=1}^{N} \eta_i \qquad ....(4)$$

$$st. \quad y_i\,(wx_i + b) + \eta_i \ge 1 \quad \eta_i \ge 0 \quad i = 1,..,N$$

The term $C\sum_{i=1}^{N} \eta_i$ is an upper bound on the number of misclassifications. The parameter *C* represents a trade off between misclassification and sparseness. The value of *C* can be chosen more systematically via cross validation. Problem (2) can be recast as a linear programming problem by introducing extra variables u_p and v_p where $w_p = u_p - v_p$ and $|w_p| = u_p + v_p$. These variables are the *p*th elements of $u, v \in R^P$. The L1 norm becomes $\|\mathrm{w}\|_1 = \sum_{p=1}^{P}(u_p + v_p) = u + v$ and the problem can be rewritten in a standard form as follows:

$$\min_{u,v,b} \ u + v + C\sum_{i=1}^{N} \eta_i$$

$$st \quad y_i\,((u - v)x_i + b) + \eta_i \ge 1 \qquad \eta_i \ge 0 \ \ i = 1,..,N$$

$$u_p, v_p \ge 0 \ \ \forall p \in \{1,..,P\} \qquad ....(5)$$

Problem (5) is a linear programming problem.

3.3 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 or a "confusion matrix", see Table 1.

Table 1. 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.

Bias is defined as (a+b)/(a+c). 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.

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)].

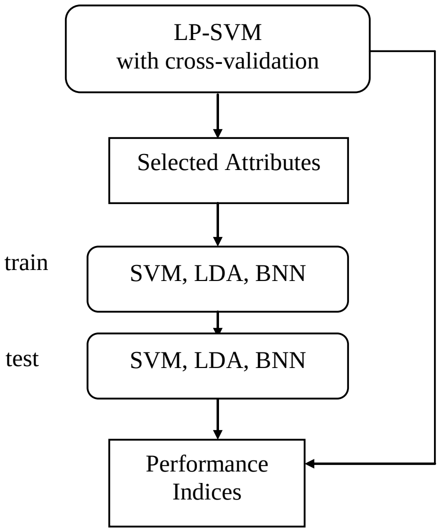

Fig. 1. The procedure schema for attribute/feature selection and tornado prediction.

Table 2. List of features selected. Column 3 includes month number and MDA (1 to 17). Column 4 adds NSE features (18 to 34).

	Attributes	Selected from MDA	Selected from MDA & NSE
1	Month	*	**
2	Meso range	*	**
3	Meso depth	*	**
4	Meso strength rank	*	**
5	Meso low-level diameter	*	**
6	Meso maximum diameter	*	**
7	Meso height of maximum diameter	*	
8	Meso low-level rotational velocity	*	**
9	Meso maximum rotational velocity	*	
10	Meso low-level shear		**
11	Meso height of maximum shear	*	
12	Meso maximum gate-to-gate velocity difference	*	**
13	Meso height of maximum gate-to-gate velocity difference	*	**
14	Meso core depth	*	**
15	Meso age (min)	*	**
16	Meso relative depth		**
17	Meso low-level convergence	*	**
18	Meso mid-level convergence		**
19	V-component of estimated storm motion vector (north-relative)		**
20	Estimated 0-3 km storm relative helicity		**
21	Downdraft CAPE (DCAPE) for a parcel 1 km above ground		**
22	DCAPE for the parcel at 0 Celsius		**
23	LFC (Level of Free Convection) in the lowest 100 mb		**
24	EHI (Energy-Helicity Index) in the lowest 100 mb		**
25	Magnitude of the storm-relative flow for the 0-2 km above ground layer		**
26	Magnitude of the storm-relative flow for the 9-11 km above ground layer		**
27	BRN shear		**
28	Mean shear through a specified depth		**
29	Average Mixing Ratio in 0-3 km layer		**
30	Average Mixing Ratio in 0-6 km layer		**
31	9-11 km storm-relative flow		**
32	4-6 km storm-relative flow		**
33	Normalized most-unstable parcel CAPE		**
34	Most-unstable parcel CAPE from sfc to 3 km above ground		**

4. EXPERIMENTS

In the experiments, the data are split into two sets: *training* and *testing*. For the training set, the ratio between tornado and non-tornado observations is about the same. In the *testing* sets, the ratio is 2%. The cases used for training are different to those used in the testing set. The same *training* and *testing* sets are applied to all methods. The SVM, LP-SVM, LDA (Heijden et al. 2004) and BNN (Sigurdsson 2002) experiments are performed in the MATLAB environment. First, LP-SVM is applied to identify the relevant features. MINOS solver (Murtagh 1998) is used to solve the linear programming problem resulting from an LP-

SVM formulation. The experiments apply cross-validation techniques to find the best trade off parameter cost *C*. After the relevant features are identified, then the data set with those features is applied on SVM and LDA (see Fig. 1). These are done both for data sets with MDA features and MDA and NSE attributes.

5. RESULTS

The results are presented in Tables 2 to 6. Table 2 presents the list of selected features/attributes obtained after running LP-SVM on the data. The original numbers of MDA features is 23. While, the original NSE consists of 84 features.

Table 3. Performance indices using MDA features.

	LP-SVM	LDA	SVM	BNN
Bias	0.986	0.9497	0.9441	0.9972
POD	0.8296	0.8212	0.8045	0.8603
FAR	0.1586	0.1353	0.1479	0.1373
Heidke Skill	0.8322	0.8393	0.8242	0.8588

Tables 3 to 6 show the performance indices for all evaluated methods using month number and MDA (Table 3), whereas Table 4 does this for a subset of the attributes. Table 5 has month number, MDA and NSE attributes while Table 6 has a subset of these attributes after running LP-SVM. Results in Table 3 indicate that BNN is the best overall solution based on the highest POD, bias closest to one and highest skill.

Table 4. Performance indices using selected features from MDA.

	LP-SVM	LDA	SVM	BNN
POD	0.8296	0.8184	0.824	0.919
FAR	0.1586	0.1556	0.1783	0.1387
Bias	0.986	0.9693	1.0028	1.067
Heidke Skill	0.8322	0.8279	0.8193	0.8869

Table 4 indicates that the best performance is given by BNN. In this table, BNN shows the highest Skill and POD and the lowest FAR. The Bias given by BNN is not as close to one as that given by SVM.

Table 5. Performance indices using MDA & NSE features.

	LP-SVM	LDA	SVM	BNN
POD	0.8547	0.0028	0.6816	0.243
FAR	0.2214	0	0.1029	0.9818
Bias	1.0978	0.0028	0.7598	13.3184
Heidke Skill	0.811	0.0055	0.7707	-0.0027

Table 5 indicates that LDA and BNN produce very poor results, since LDA underforecasts tremendously whereas BNN overforecasts badly. FAR reaches the lowest value for LDA, but the aforementioned bias invalidates the solution. LP-SVM performs best in this data set (Table 5). The performance of LP-SVM is still worse than that given by BNN with selected attributes shown in Table 4. The NSE attributes selected relate well to fast storm movement, vigorous rotation, strong updrafts, good shear profiles and robust inflow and outflow at the bottom and top of a storm.

Table 6. Performance indices using selected features from MDA & NSE.

	LP-SVM	LDA	SVM	BNN
POD	0.8547	0.8492	0.7235	0.8603
FAR	0.2214	0.2083	0.119	0.1873
Bias	1.0978	1.0726	0.8212	1.0587
Heidke Skill	0.811	0.8157	0.7908	0.8324

Table 6 indicates that BNN produces the best results, except for FAR. The improvement in skill from Table 5 to Table 6 is striking and makes a strong case for feature selection. SVM and LP-SVM improve only slightly, which indicates these techniques are robust with respect to the variance of the attributes.

6. CONCLUSIONS

Feature selection of radar-derived velocity data using LP-SVM has significantly improved the probability of detection of tornadoes and lowers the false alarm rate, compared to the raw mesocyclone detection algorithm or mesocyclone detection algorithm & near storm environment features.

Based on the performance improvements of BNN over LDA, LP-SVM or SVM, further research should be persued. Feature selection using LP-SVM is important for tornado detection using a Bayesian network. The high level of skill shown is an improvement over previous research in terms of predictability and could result in considerable reduction in loss of life if implemented operationally.

REFERENCES

Bishop, C.M. (1995), **Neural Networks for Pattern Recognition**, University Press, Oxford.

Bhattacharyya, C., Grate, L.R., Rizki, A., Radisky, D., Molina, F.J., Jordan, M.I., Bissell, M.J., and Mian, I.S. (2003), ‘Simultaneous

Classification and Relevant Feature Identification in High-dimensional Spaces: Application to Molecular Profiling Data', *Signal Processing*, Vol. 83, Issue 4, pp. 729-743.

Donoho, D. and Huo, X. (1999), 'Uncertainty Principles and Ideal Atomic Decomposition', *Technical Report*, Statistics Department, Stanford University, http://www-stat.stanford.edu/~donoho/reports.html.

Doswell, C.A. III, Davies-Jones, R. and Keller, D. (1990), 'On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables', *Weather and Forecasting*, Vol. 5, pp. 576-585.

Haykin, S. (1999), **Neural Networks: A Comprehensive Foundation**, 2nd Edition, Prentice-Hall, Upper Saddle River, NJ.

Heijden, F., Duin, R.P.W., Ridder, D. and Tax, D.M.J. (2004), **Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB**, John Wiley and Sons Ltd., West Sussex, England.

Lakshmanan, V., Stumpf, G. and Witt, A. (2005), 'Neural Network for Detecting and Diagnosing Tornadic Circulations using the Mesocyclone Detection and Near Storm Environment Algorithms', *21st Int'l Conference on Information Processing Systems*, Amer. Meteo. Soc., San Diego, CD-ROM, pp. J5.2.

MacKay, D. (1992a), 'A Practical Bayesian Framework for Backpropagation Networks', *Neural Computation*, Vol. 4, pp. 448-472.

MacKay, D. (1992b), 'The Evidence Framework Applied to Classification Networks', *Neural Computation*, Vol. 4, pp.720-736.

Marzban, C. and Stumpf, G. J. (1996), 'A Neural Network for Tornado Prediction Based on Doppler Radar-Derived Attributes', *Journal of Applied Meteorology*, Vol. 35, pp. 617-626.

Murtagh, B.A. and Saunders, M.A. (1998), **MINOS 5.5 USER'S GUIDE**, Technical Report SOL 83-20R, Revised July 1998, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford.

Sigurdsson, S. (2002), **Binary Neural Classifier**, Version1.0,http://mole.imm.dtu.dk/toolbox/ann/.

Theodore, B.T., Santosa, B. and Richman, M.B. (2004), 'Bayesian Neural Networks for Tornado Detection', *WSEAS Transaction on Systems*, Vol. 3, Issue 10, pp. 3211-3216.

Wilks, D.S. (1995), **Statistical Methods in the Atmospheric Sciences**, Academic Press, London.

Diterima: 12 April 2006

Disetujui untuk diterbitkan: 13 Desember 2006