Evaluation of Support Vector Machines and Minimax Probability Machines for Weather Prediction

Stephen Sullivan [*]

UCAR - University Corporation for Atmospheric Research, Boulder, Colorado

[*] *Corresponding author address:* Stephen Sullivan, UCAR - University Corporation for Atmospheric Research, RAL, FL/2, UCAR, 3450 Mitchell Lane, Boulder CO 80301.

E-mail: steves@ucar.edu

ABSTRACT

The paper evaluates two kernel-based methods on the problem of predicting precipitation based on observable variables. The support vector machine (SVM) method finds the two parallel hyperplanes that provide maximal separation of two subsets, excepting outliers. The minimax probability machine (MPM) method finds an optimal separating hyperplane that minimizes the probability of misclassification.

Both SVM and MPM are binary classification methods that can be extended easily to multiclass problems. Both make use of the "kernel trick" to transform a linearly inseparable problem into a higher dimensional space where the problem may be linearly separable.

The paper also investigates the accuracy and Peirce Skill Score (PSS, also known as the Hanssen and Kuipers discriminant) measures resulting from adding derived variables and removing variables. Using cross validation on the training data the accuracy was 70% and PSS 47%. On the final testing data the PSS was 35%.

1. Introduction

The 2008 AMS AI Data Mining Competition (AMS 2008) provided a training set of 847 examples containing 15 independent variables plus an observed result. The testing set contained 363 examples, without the observed result.

The 15 independent variables are shown in Table 1. The dependent variable was "precipitation type" and had three possible values: "none", "frozen", and "liquid".

The remainder of this paper provides sections: **2:** Theory overview, **3:** Method, **4:** Results, **5:** Conclusion

2. Theory overview

a. Classification Problems

Consider a set of observations, the training data, $S = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ where $\mathbf{x}_i \in \mathbf{R}^m$ and $y_i \in Y \subseteq \mathbf{N}$. Let S_k be the set of observations drawn from distribution $F_k, k = 1, \ldots, K$.

In the binary classification problem $K = 2$ and $Y = \{-1, 1\}$; in the multiclass problem $K > 2$ and $Y \subseteq \mathbf{N}$.

We want to find the decision function $f : \mathbf{R}^m \to Y$ that minimizes a loss function such as $L(f) = E((f(\mathbf{x}) - y)^2)$, where E is the expectation operator.

b. Support Vector Machines

The SVM method, (Vapnik (1999), Scholkopf and Smola (2002)) addresses the binary classification problem.

Support vector machines find a linear separation function $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$ via the optimization problem

$$\min_{\mathbf{w},b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x} + b) \geq 1$$

To handle outliers that might make that sets S_1 and S_2 not linearly separable, slack variables ξ_i can be used, leading to:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2} \|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x} + b) \geq 1 - \xi_i \tag{1}$$

The discrimination function in either case is

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b) \tag{2}$$

c. Minimax Probability Machine

The MPM method (Lanckriet et al. 2003) also addresses the binary classification problem.

Given $S_1 \sim F_1$ and $S_2 \sim F_2$ as above, the MPM method finds an optimal separating hyperplane that minimizes the probability of misclassification. The MPM is based on a probability theorem (Marshall and Olkin 1960) recast into an optimization framework by (Popescu and Bertsimas 1999).

Consider a sample having mean $\bar{x}$ and covariance matrix Σ. Given any convex set $V \subseteq \mathbf{R}^m$, we're interested in the maximum, over *all* distributions having mean $\bar{x}$ and covariance Σ, of the probability that an observation $\mathbf{x}$ is in V. The surprisingly simple result is $1/(1+d^2)$ where d is roughly the distance from $\bar{x}$ to the closest point in V. That is,

$$\sup_{\mathbf{x}\sim(\bar{x},\Sigma)} P[\mathbf{x} \in V] = \frac{1}{1+d^2}$$

where

$$d^2 = \inf_{\mathbf{v}\in V} (\mathbf{v} - \bar{x})^T \Sigma^{-1} (\mathbf{v} - \bar{x})$$

Here the supremum is taken over all distributions having mean $\bar{x}$ and covariance matrix Σ.

Consider a hyperplane separating $\bar{x}_1$ and $\bar{x}_2$. V in theorem above can be taken to be the half-space containing $\bar{x}_2$, and an observation from F_1 is misclassified if it falls in V.

Similarly V in theorem above can be taken to be the half-space containing $\bar{x}_1$, and an observation from F_2 is misclassified if it falls in V.

By optimizing the selection of hyperplane, the probability of misclassification is minimized.

d. The Kernel Trick

Many problems are not linearly separable. The "kernel trick" (Mercer (1909), Aizerman et al. (1964), Vapnik (1999)) transforms a problem into a higher dimensional space in which it may be linearly separable. The process involves two steps. First, the optimization problem (eqn 1) and discriminant function (eqn 2) are expressed solely in terms of dot products $(\mathbf{x}_i^T \mathbf{x}_j)$.

Second, a map Φ and companion kernel K are created. The map $\Phi : \mathbf{x}_i \in \mathbf{R}^m \rightarrow \mathbf{y}_i \in H = R^M$ where $M \gg m$ maps the input space into a high dimensional feature space H. In

H, also known as the linearization space, the problem may be linearly separable, with the possible exception of outliers.

The kernel function K satisfies $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{y}_i^T \mathbf{y}_j$ where $\mathbf{y}_i = \Phi(\mathbf{x}_i)$. Since both the SVM and MPM can be expressed entirely in terms of dot products $\mathbf{y}_i^T \mathbf{y}_j$, the SVM and MPM problems in feature space can be expressed in terms of the kernel $K(\mathbf{x}_i, \mathbf{x}_j)$, avoiding all computations in the high dimensional feature space H.

The most common kernel functions are:

linear: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_J)$

polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_J + c)^k$

gaussian: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_J\|)$

e. Multiclass Methods

While SVM and MPM are binary classification methods, they are easily extended to K classes where $K > 2$.

One against all: K SVMs are trained each to discriminate a single class vs all the rest. The decision function chooses the class showing the maximum value of all the SVMs.

One against one: For each of the $K(K-1)/2$ pairs of observed classes, an SVM is trained discriminate between the two classes. The decision function takes the most popular class by vote among all the SVMs.

Binary tree: The observed classes are divided into two subsets, A and B, and an SVM is trained to discriminate between sets A and B. Each subset A, B is divided similarly, recursively, creating a tree of subsets and associated SVMs. The leaf nodes consist of a single class each. The decision function simply executes the SVMs in the tree.

In the present case two SVMs were used. The first distinguished between precipitation = "none" (set A) vs all other (set B). The second SVM distinguished among B members: precipitation = "frozen" (set $B1$) vs "liquid" (set $B2$).

3. Method

In this investigation the gaussian kernel was used for both SVM and MPM, and a binary class division was used to handle the multiclass data.

For each of the SVM configurations described below, 10 random permutations of the training data were used. For each such permutation a 5-fold cross validation was performed. Because the MPM system was slower, each MPM configuration was tested with 5 random permutations and a 5-fold cross validation. Thus each SVM configuration had 50 individual test cases and each MPM configuration had 25.

Each of the individual test cases resulted in a measured accuracy ACC and multiclass Peirce Skill Score PSS, also known as the Hanssen and Kuipers discriminant.

$$ACC = \frac{1}{N} \sum_i c_{ii} \qquad PSS = \frac{1/N \sum_i c_{ii} - 1/N^2 \sum_i f_i o_i}{1 - 1/N^2 \sum_i o_i^2}$$

where

N is the total number observations (number of test cases)

f_i is the number of forecasts in class i

o_i is the number of observations in class i

c_{ij} is the number of forecasts in class i having observations in class j

The test results for each configuration are shown by a box plot in which the box bottom and top represent the 25th and 75th percentiles, and the notched red bar represents the median. The whiskers extend to the most extreme point within $1.5q$ past the box where q is the interquartile distance.

4. Results

a. Software and Performance

The SVM tests used the C++ system libsvm (Chang and Lin 2008). The MPM tests used a combination of the author's own code and parts of Lanckriet's MPM system (Lanckriet et al. 2003). The MPM tests were written in the Octave mathematical language (Eaton 2008). Since C++ is a compiled language while Octave is interpreted, the SVM tests ran approximately 1200 times faster than the MPM tests. Because of the slow speed of the MPM system, the MPM system was tested only on the parametrization study below.

b. Parameterization

The performance of SVM using the Table 1 variables as a function of γ and C is shown in Figures 1 (accuracy) and 2 (PSS). The best configuration was near $\gamma = 0.3$ and $C = 10$, giving median accuracy of 69% and PSS of 46% in the cross validation tests. The performance of MPM as a function of γ is shown in Figure 3. The best performance was at $\gamma = 0.007$, giving median accuracy of 67% and PSS of 42% in the cross validation tests. For SVM and MPM, larger γ values decrease the radius of the influence of the training points. For SVM, larger values of C give more weight to outliers.

c. Additional Wind and Radar Variables

In an attempt to make the wind and radar information more readily accessible to the machine learning system, two sets of derived independent variables were tested: wind direction and speed, derived from urel and vrel; and radar-derived latitude and longitude, derived from range and azimuth.

The performance of the SVM method with no, one, or both sets of additional variables is shown in Figure 4. These tests showed resulted in negligible changes in accuracy and PSS.

d. Omitted Variables

Individual variables were tested to determine if omitting them would improve results (Figure 5). Omitting tilt, azimuth, RhoHV, or Kdp provided a marginal increase in accuracy and an increase of roughly 1% in PSS.

e. Categorized Values

In an attempt to make the data more accessible to the SVM method, the urel and vrel fields were each divided into 10 categories by value, resulting in 100 new variables to express urel and vrel jointly (Figure 6). Thus a given urel,vrel combination (say urel = 3.2, vrel = 6.5) would be represented by one of the new variables having value 1, and the remaining 99 being 0. Similarly, windDir and windSpeed were jointly divided, creating 100 new variables. Finally, windDir, windSpeed, and temperature were jointly divided, creating 1000 new variables. Dividing variables into categories decreased SVM accuracy by 4 to 7% and decreased PSS by 8 to 12%.

5. Conclusion

This paper has given an overview of two kernel-based methods, support vector machines and minimax probability machines, and their use in statistical weather prediction.

Using cross validation tests on the training data, the SVM method showed 2 % better accuracy and 4 % better PSS than the MPM method. The SVM implementation was approximately 1200 times faster than the MPM, making further experimentation with the SVM method more feasible.

Tests to omit independent variables and to add derived variables proved fruitful in increasing the prediction performance.

The cross validation tests on training data showed an accuracy of 70% and PSS of 47%. On the final testing data the PSS was 35%.

Acknowledgments.

The author thanks John Williams and Jennifer Abernethy for their thoughtful advice, Kim Elmore for his management of the competition, and the AMS Committee on Artificial Intelligence for organizing the competition.

REFERENCES

Aizerman, M., E. Braverman, and L. Rozonoer, 1964: Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, **25**, 821–837.

AMS, 2008: 2008 artificial intelligence competition. [Available online at http://www.nssl.noaa.gov/ai2008/], [Available online at http://www.nssl.noaa.gov/ai2008/].

Chang, C. and C. Lin, 2008: Libsvm – a library for support vector machines. Tech. rep., National Taiwan University. URL `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

Eaton, J., 2008: *GNU Octave Manual Version 3*. Network Theory Ltd., 568 pp.

Lanckriet, G., L. E. Ghaoui, C. Bhattacharyya, and M. Jordan, 2003: A robust minimax approach to classification. *Journal of Machine Learning Research*, **3**, 555–582.

Marshall, A. and I. Olkin, 1960: Multivariate chebyshev inequalities. *Annals of Mathematical Statistics*, **31**, 1001–1014.

Mercer, J., 1909: Functions of positive and negative type and their connection with the theory of integral equations. *Transactions of the Royal Society London Series A*, **209**, 415–446.

Popescu, I. and D. Bertsimas, 1999: Optimal inequalities in probability theory: A convex optimization approach. Tech. Rep. HD28 .M414 no.4083-99, MIT Sloan School of Management. URL `http://hdl.handle.net/1721.1/2755`.

Scholkopf, B. and A. Smola, 2002: *Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 626 pp.

Vapnik, V., 1999: *The Nature of Statistical Learning Theory*. 2d ed., Springer, 314 pp.

List of Tables

Table 1. Original independent variables

Label	Description
lat	latitude
lon	longitude
tmpc	2 m air temperature (Celsius)
relh	relative humidity (percent)
urel	u (East-West) wind (W is pos)
vrel	v (North-South) wind (S is pos)
frzl	freezing level (m MSL)
tilt	radar elevation angle (degrees)
range	range from radar (km)
azimuth	radar azimuth (deg, 0 is North)
hgt	height of radar data (km)
Zdr	differential reflectivity (dB)
RhoHV	cross-correlation coefficient
Kdp	specific diff phase (deg km-1)
Z	reflectivity (dBZ)

List of Figures

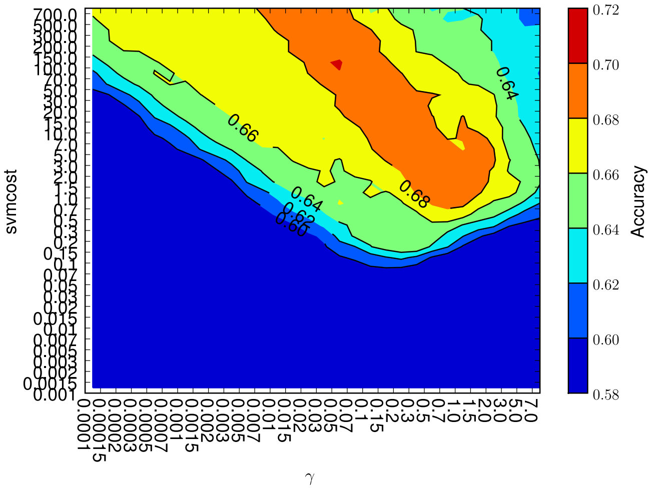

FIG. 1. SVM accuracy vs γ and cost C

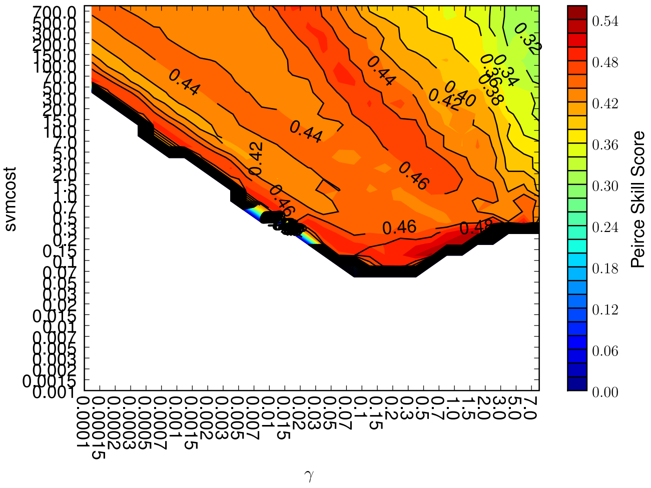

FIG. 2. SVM Peirce skill score vs γ and cost C

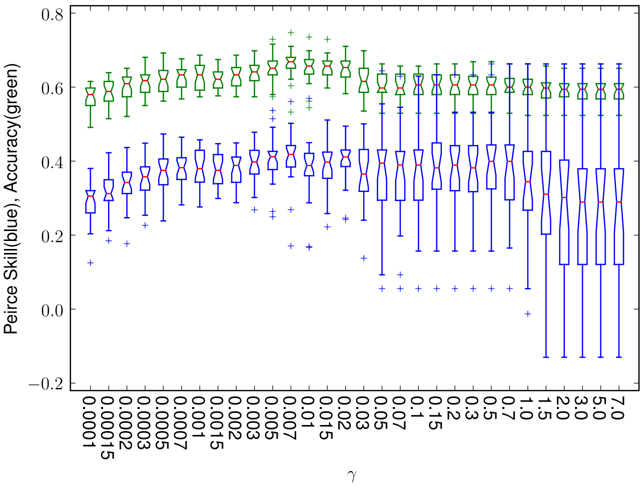

FIG. 3. MPM accuracy and Peirce skill score vs γ

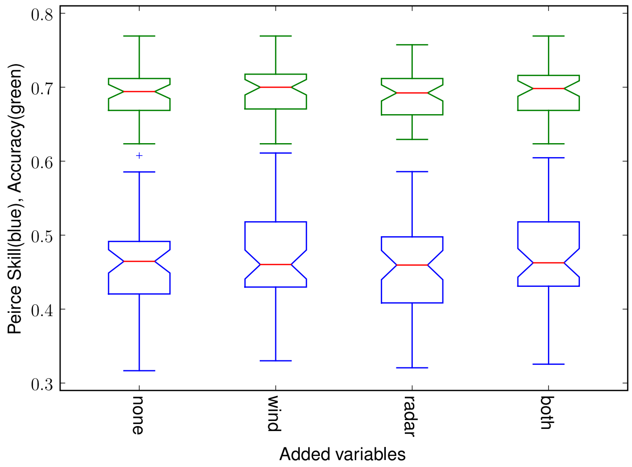

Fig. 4. Comparison of SVM on original variables ("none"), using additional windSpeed and WindDir variables ("wind"), using additional radarLat and radarLon variables ("radar"), and using both sets of additional variables ("both").

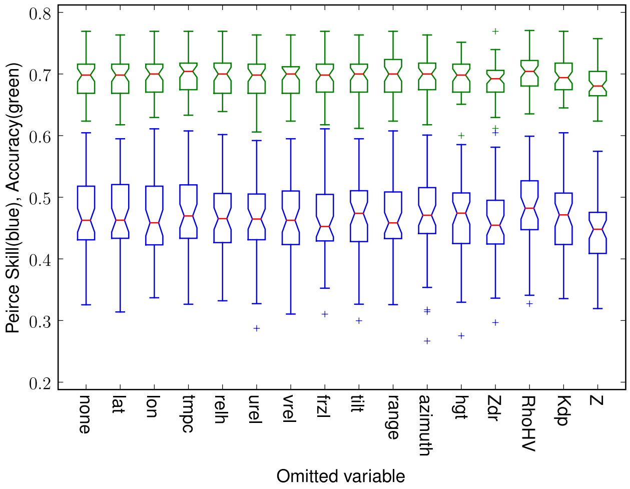

Fig. 5. Comparison of SVM on original variables ("none") with SVM omitting single variables

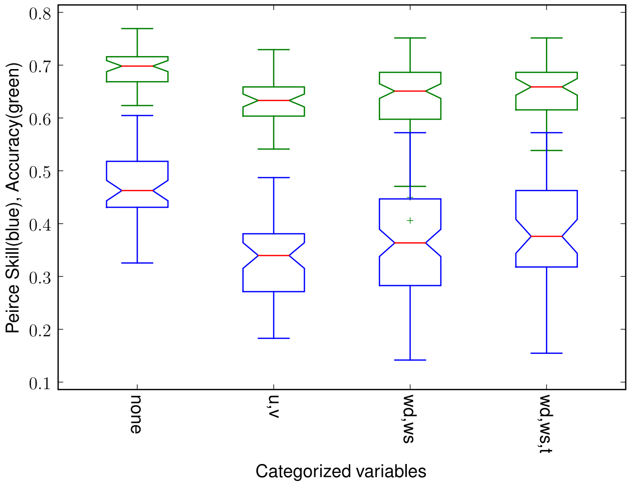

Fig. 6. Comparison of SVM on original variables ("none") with SVM categorizing urel and vrel ("u,v"), or windSpeed and windDir ("wd,ws"), or windSpeed and windDir and temperature ("wd,ws,t")