PERGAMON

Neural Networks 13 (2000) 17–19

Neural Networks

www.elsevier.com/locate/neunet

Contributed article

How good are support vector machines?

Š. Raudys*

Institute of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania

Abstract

Support vector (SV) machines are useful tools to classify populations characterized by abrupt decreases in density functions. At least for one class of Gaussian data model the SV classifier is not an optimal one according to a mean generalization error criterion. In real world problems, we have neither Gaussian populations nor data with sharp linear boundaries. Thus, the SV (maximal margin) classifiers can lose against other methods where *more than a fixed number of supporting vectors contribute in determining the final weights* of the classification and prediction rules. A good alternative to the linear SV machine is a specially trained and optimally stopped SLP in a transformed feature space obtained after decorrelating and scaling the multivariate data. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Support vectors; Classification; Generalization; Overtraining

A popular criterion used to design classification and prediction rules is the mean square error (mse). A Fisher standard linear discriminant function (DF) and a standard mse regression are typical examples. One alternative is *the support vector approach* which was originally proposed by Glucksman (1966) and later developed and popularized by a number of authors. In a minimax regression, we minimize the maximal distance of the training-set vectors from a prediction hypersurface. In a maximal margin classifier, we maximize the minimal distance of the training-set vectors from a discriminating hypersurface. In both cases, only a fixed (usually small) number of training-set vectors determine the parameters of the decision rule. Therefore, often this rule is called a support vector (SV) machine. Because in this approach no probability density is estimated, it becomes highly insensitive to the curse of dimensionality. In spite of the fact that for any data model it was never proved that the SV machine is optimal according to the generalization error criterion, some authors consider the support vector approach as “an optimal” (p. 127 of Vapnik 1995), and “very promising” (Tax, de Ridder & Duin, 1997). In this note, we will discuss some advantages and drawbacks of the classical SV approach (the maximal margin classifier) earlier not discussed in the literature.

A theory of statistical decision functions is almost the only one technique which allows us to design *rigorously optimal decision rules*. In classification, a probability of misclassification (a generalization error) is the most popular criterion used to define the optimality. Let $f(\mathbf{x}|\pi_i)$ be a conditional probability density function and q_i be a class prior probability of a vector $\mathbf{x}$ to be classified. Then the optimal classification rule will use DF $g(\mathbf{x}) = \log(q_1 f(\mathbf{x}|\pi_1)/q_2 f(\mathbf{x}|\pi_2))$ to classify vector $\mathbf{x}$ into classes π_1 and π_2. While solving real world problems the densities $f(\mathbf{x}|\pi_1)$ and $f(\mathbf{x}|\pi_2)$ are unknown. In a parametric approach to design the classifier, some assumptions concerning the functional forms of $f(\mathbf{x}|\pi_1), f(\mathbf{x}|\pi_2)$ are made, and the densities are expressed as functions of unknown parameters, i.e. $f(\mathbf{x}|\pi_i) = f(\mathbf{x}|\mathbf{\Theta}, \pi_i)$, where $\boldsymbol{\theta}$ is a vector of the parameters. In a standard linear Fisher classifier, $f(\mathbf{x}|\mathbf{\Theta}, \pi_i)$ is a multivariate Gaussian density, $\mathbf{\Theta}$ is composed from the covariance matrix (CM) and the mean vectors. In the Fisher classifier, we use traditional maximum likelihood estimates of CM and the mean vectors. Such a rule is called a plug-in one, and its optimality is proven only for some particular cases. If CM is proportional to an identity matrix we do not need to estimate it. Then we have an Euclidean distance (nearest means) classifier (EDC). To design EDC we should use only sample maximum likelihood estimates of the mean vectors. In a Bayes approach to design the optimal classification rules, the vector $\mathbf{\Theta}$ is supposed to be a random one. Thus, in the Bayes approach, we formulate an optimality of DF only for *an entity of problems* defined by a prior distribution of vector $\mathbf{\Theta}$. Abramson and Braverman (1962) have shown that for spherical Gaussian vectors $\mathbf{x}$ and spherical Gaussian prior densities of the mean vectors, the optimal decision rule which minimizes the mean generalization error is EDC. Serdobolskij (1983) extended this result and presented more rigorous proof. Thus, for an entity of classification problems defined by the Gaussian prior distribution of a difference in the mean vectors, EDC is optimal and

* Fax: +370-2-729-209.
E-mail address: raudys@das.mii.lt (Š. Raudys).

PII: S0893-6080(99)00097-0

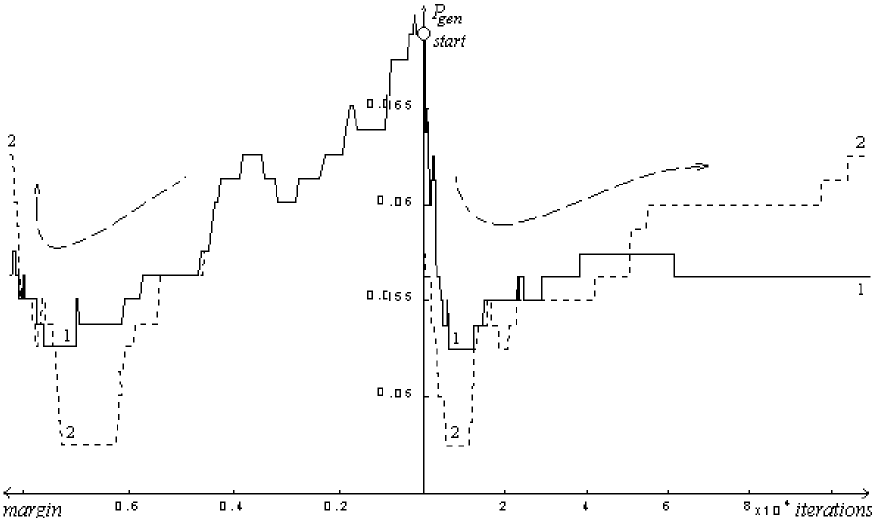

Fig. 1. Use of SLP to classify the vowel data: the generalization error P_N versus the number of iterations and versus the margin in the original (graphs 1) and transformed (graphs 2) feature spaces.

no other classification rule will perform better. This conclusion is valid also for the SV classifiers. Therefore, at least for one class of theoretical data model we can declare: "the SV classifier is not the best machine according to the mean generalization error criterion".

The relative performance of the SV classifier is much easier to comprehend if we remember that *the SV machine can be obtained in a gradient (BP) training of the non-linear single layer perceptrons (SLP).* It was shown (Raudys, 1998) that under special training conditions while training SLP one can obtain EDC just after the first iteration, move further towards a regularized discriminant analysis (DA), to the Fisher classifier, the robust DA and when the training-set error is zero, move towards the SV (maximum margin) classifier. In our experiments, we shifted the training set mean vector into a center of co-ordinates, used the standard sum of squares cost function with the sigmoid activation function and targets either zero or one. For training we applied the total gradient descent training procedure with zero initial weight vector, and in order to obtain the maximal margin quickly, we gradually increased the learning-step η according to a rule $\eta = \eta_0 \times (1 + \epsilon)^t$, where t is a current iteration number and ϵ(is a small positive constant (Raudys, 1998). In addition, much smaller generalization errors can be obtained if prior to training we use a liner data transformation $\mathbf{Y} = \mathbf{D}^{-1/2}\mathbf{T}\mathbf{x}$, where $\mathbf{T}$ and $\mathbf{D}$ are eigenvectors and eigenvalues of the pooled and regularized sample estimate of the covariance matrix (Raudys, 2000).

In principle, it is possible to construct artificial data with sharp linear boundaries of the classes where the SV classifier would be an optimal classifier design strategy according to the generalization error criterion. For an entity of spherically Gaussian pattern classes defined by the spherically Gaussian prior distributions of the difference in the mean vectors, the SV classifier is not the best choice—theory shows that here EDC is the best decision rule.

In real world problems, we have neither Gaussian populations nor data with sharp linear boundaries. Below we present *a characteristic example* with real world data and demonstrate that while training the SLP classifier we obtain a number of different statistical rules involving the support vector rule. The SLP classifier was used to solve the vowel classification problem: two pattern classes, 20 speakers, 28 spectral and cepstral features, the learning-set size $N = 20 + 20$, the test-set size $n = 800$. Fig. 1 shows the generalization error P_N versus the number of iterations (right), and versus the margin (left) in an original feature space (FS)—curves 1, and in the transformed FS—curves 2 ($\eta_0 = 0.001$, $\epsilon = 0.03$). After the first iteration we obtain EDC, then the robust regularized DA, and after 110,000 iterations we have nine support vectors approximately equidistant from the SV hyperplane. In order to approach the maximal margin closely, we need a large number of iterations. Training in the transformed FS is better than in the original FS. Excessive growth in the margin width increases the generalization errors. The minimal generalization error is obtained much before we have the SV classifier.

While analyzing the effectiveness of the data transformation strategies (see e.g. Raudys (2000)), we performed *several thousands* of training experiments with a dozen 18–166-variate real world data sets, various data rotation and scaling methods, learning-set sizes, and different small, randomly chosen learning-sets. In almost all experiments, we observed *an increase in the margin width and overtraining.* Several dozens of exceptions constituted cases with non-zero empirical error, or when the comparatively small number of iterations and a fixed size of the learning-rate parameter did not allow the perceptron to move sufficiently close to maximally possible margins. It means that in many practical situations the non-linear SLP can outperform the SV machine. For this we need to transform data and to have a validation set in order to determine the optimal number of iterations. The SV machine is a unique solution and does not require the validation set. Similar conclusions are valid for the regression obtained by SLP with a special cost function.

References

Abramson, N., & Braverman, D. (1962). Learning to recognize patterns in a random environment. *IRE Transactions on Information Theory*, *IT-8* (5), 58–63.

Glucksman, H. (1966). On improvement of a linear separation by extending the adaptive process with a stricter criterion. *IEEE Transactions on Electronic Computers*, *EC-15* (6), 941–944.

Raudys, Š (1998). Evolution and generalization of a single neurone. I. SLP as seven statistical classifiers. *Neural Networks*, *11* (2), 283–296.

Raudys, Š (2000). Scaled rotation regularization. *Pattern Recognition*, *33*, in press.

Serdobolskij, V. I. (1983). On minimal error probability in discriminant analysis. *Reports of the Academy of Sciences of the USSR*, *270*, 1066–1070.

Tax, D. M. J., de Ridder, D., & Duin, R. P. W. (1997). *Support vector classifiers: a first look*, *Proceedings of ASCI'97, Heijen, NL, 2–4 June* pp. 253-258.

Vapnik, V. N. (1995). *The nature of statistical learning theory*, Berlin: Springer.