ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Bootstrap Feature Selection in Support Vector Machines for Ventricular Fibrillation Detection

F. Alonso Atienza[1], J.L. Rojo Álvarez[1], G. Camps i Valls[2],
A. Rosado Muñoz[2], and A. García Alberola[3] *

1- University Carlos III of Madrid - Dept. of Signal Theory and Communications
Av. Universidad 30, 28911, Leganés, Madrid - Spain

2- University of Valencia - Dept. of Electrical Engineering (GPDS)
Doctor Moliner 50, 46100, Burjassot, Valencia - Spain

3- University Hospital Virgen de la Arrixaca - Lab. of Electrophysiology
Ct. Madrid-Cartagena s/n, 30120, El Palmar, Murcia - Spain

Abstract. Support Vector Machines (SVM) for classification are being paid special attention in a number of practical applications. When using nonlinear Mercer kernels, the mapping of the input space to a high-dimensional feature space makes the input feature selection a difficult task to be addressed. In this paper, we propose the use of nonparametric bootstrap resampling technique to provide with a statistical, distribution independent, criterion for input space feature selection. The confidence interval of the difference of error probability between the complete input space and a reduced-in-one-variable input space, is estimated via bootstrap resampling. Hence, a backward variable elimination procedure can be stated, by removing one variable at each step according to its associated confidence interval. A practical example application to early stage detection of cardiac Ventricular Fibrillation (VF) is presented. Basing on a previous nonlinear analysis based on temporal and spectral VF parameters, we use the SVM with Gaussian kernel and bootstrap resampling to provide with the minimum input space feature set that still holds the classification performance of the complete data. The use of bootstrap resampling is a powerful input feature selection procedure for SVM classifiers.

1 Introduction

Support Vector Machines (SVM) are efficient learning schemes [1], which have been paid special attention during the last years. The SVM classification algorithm has shown an excellent performance in a number of practical applications [2], in terms of minimal classification error probability. In particular, SVM are robust when working with high-dimensional input spaces, such as images or gene expressions [3]. In some applications, not only the best classification is required, but also the quantification of the relative relevance of each of the input space features, as well as the determination of the most reduced set of variables with non-redundant information, is needed. In classical statistics, this twofold task is addressed by linear (and the nonlinear versions) discriminant analysis [4].

*This work has been partially supported by research projects from Guidant Spain and from "Comunidad de Madrid" (GR/SAL/0471/2004).

Several Mercer kernel based versions of linear discriminant analysis have been proposed, such as Fisher discriminant analysis or nonlinear discriminant analysis with kernels [5, 6], which make a discriminant analysis in a reproducing kernel Hilbert space, but often, these approaches do not take into consideration the input space feature reduction stage. Also, a number of procedures have been proposed for input space feature selection in nonlinear SVM classifiers [7], but these methods do not provide with a clear cut-off statistical test.

An advantage of SVM classifiers is its nonparametric nature. Given that their optimizing criterion is the maximum margin in the separating hyperplane, they are not sensitive to the input space feature statistical distribution. According to this property, the use of the (nonparametric version of the) Bootstrap Resampling (BR) [8] for creating sequential procedures of input space variable selection, in connection with SVM classifiers, is an interesting issue. In [9], BR is proposed for tuning the optimal SVM free parameters in reduced data sets. In [10], BR is proposed and tested for comparing the relevance of disjoint subsets of the input feature space. Here, we extend and complete the previous development by proposing the use of BR to provide with a Backward Input Space Selection Procedure (BISSP) based on bootstrap Confidence Intervals (CI). The statistical criterion to be used in here is the difference in Error Probability (P_e) between: (1) the complete model, and (2) a reduced model that only considers a subset of the input space features. This approach lies in eliminating one by one the irrelevant features of the input space, until a subset of only-significant input variables is present. This will ensure that the performance of the final SVM classifier is not significantly different from the complete model trained classifier. The BISSP is tested in a toy example, and then, it is applied to an automatic cardiac Ventricular Fibrillation (VF) detector scheme, presented in [10].

The paper is organized as follows. In the next section, the SVM is briefly revised. Then, the BR-SVM BISSP is presented. Section 4 contains the toy example, and Section 5 introduces the VF discrimination problem and the obtained results. Finally, conclusions are drawn.

2 SVM Classifiers

SVM binary classifier is a sampled-based statistical learning algorithm based on constructing maximum margin separating hyperplanes in a reproducing kernel Hilbert space. A detailed description of SVM can be found, for instance, in [1].

Be $\mathbf{V}$ a set of N observed and labeled data, $\mathbf{V} = \{ (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) \}$, where $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$. Be $\phi(\mathbf{x}_i)$ a nonlinear transformation to a (generally unknown) higher dimensional space $\mathbb{R}^B$, where a separating hyperplane is given by $(\phi(\mathbf{x}_i) \cdot \mathbf{w}) + b = 0$. We know that $K(\mathbf{x}_i, \mathbf{x}_j) = (\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j))$ is a Mercer's kernel, which allows us to calculate the dot product of pairs of vectors transformed by $\phi(\mathbf{x}_i)$ without explicitly knowing the nonlinear mapping. Two often used kernels are the linear, given by $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)$, and the Gaussian, given by $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left\{ -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right\}$.

Under these conditions, the problem is to minimize

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i \tag{1}$$

with respect to $\mathbf{w}, b$, and ξ_i, and constrained to $y_i\big\{\,(\phi(\mathbf{x}_i)\cdot\mathbf{w}) + b\big\} - 1 + \xi_i \geq 0$ and to $\xi_i \geq 0$, for $i = 1,\ldots,N$, where ξ_i represent the losses; C represents a trade-off between margin and losses; and $(\cdot)$ expresses the dot product. By using the Lagrange Theorem, (1) can be rewritten into its dual form, and then, the problem consists of maximizing

$$\sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i y_i \alpha_j y_j K(\mathbf{x}_i,\mathbf{x}_j) \tag{2}$$

constrained to $0 \leq \alpha_i \leq C$ and $\sum_{i=1}^{N}\alpha_i y_i = 0$, where α_i are the Lagrange multipliers corresponding to primal constrains. Gaussian kernel width σ, and parameter C, are free parameters that have to be previously settled. Methods such as cross-validation or bootstrap resampling can be used for this purpose.

3 Bootstrap Feature Selection

In [10], a BR based method for feature selection is proposed, which is here briefly presented according to the principles in [9]. A dependence estimation process between pairs of data in a classification problem, where the data are drawn from a joint distribution $p(\mathbf{x}, y) \to \mathbf{V}$, can be solved using a SVM. The estimated SVM coefficients with the whole data set are $\alpha = [\alpha_1,\ldots,\alpha_N] = s\,(\mathbf{V}, C, \sigma)$, where $s()$ is the operator that accounts for the SVM optimization, and it depends on the data ($\mathbf{V}$) and on the values of C and σ. The empirical risk for the current coefficients is defined as the training error fraction of the machine, $R_{emp} = t\,(\alpha, \mathbf{V})$, where $t()$ is the operator that represents the empirical risk estimation.

A *bootstrap resample* is a new data set obtained from the training set according to the empirical distribution, i.e., it consists of sampling with replacement the observed pairs of data: $\hat{p}(\mathbf{x}, y) \to \mathbf{V}^* = \big\{\,(\mathbf{x}_1^*, y_1^*)\,,\ldots,(\mathbf{x}_N^*, y_N^*)\,\big\}$

Therefore, $\mathbf{V}^*$ contains elements of $\mathbf{V}$ appearing zero, one, or several times. The resampling process is repeated $b = 1,\ldots,B$ times. A partition of $\mathbf{V}$ in terms of resample $\mathbf{V}(b)^*$ is $\mathbf{V} = (\mathbf{V}_{in}^*(b), \mathbf{V}_{out}^*(b))$, where $\mathbf{V}_{in}^*(b)$ is the subset of samples included in resample b, and $\mathbf{V}_{out}^*(b)$ is the subset of non-included samples. SVM coefficients for each resample are given by $\alpha^* = s\,(\mathbf{V}_{in}^*(b), C, \sigma)$. The actual risk estimation for the resample is known as its *bootstrap replicate*, and can be obtained by taking $R^*(b) = t\,(\alpha^*, \mathbf{V}_{out}^*(b))$. Therefore, its normalized histogram for the B resamples approximates the empirical risk density function. A proper choice for B is typically from 50 to 300 resamples.

Now we consider a reduced version of the observed data ($\mathbf{W}_u$), in which the u^{th} variable is removed in all the available observations. If we perform a parallel resampling procedure according to the complete set resampling,

$$\hat{p}(\mathbf{x}, y) \rightarrow \mathbf{W}_u^* = \left\{ (\mathbf{x}_1^*, y_1^*), \ldots, (\mathbf{x}_N^*, y_N^*)_{|\text{uremoved}} \right\} \tag{3}$$

then the bootstrap replication of the actual risk in the incomplete model can be obtained as $R_u^*(b) = t\,(\alpha^*, \mathbf{W}_{out}^*(b))$, so the statistic $\triangle\mathbf{R}^*(\mathbf{b}) = \mathbf{R}_\mathbf{u}^*(\mathbf{b}) - \mathbf{R}^*(\mathbf{b})$ can be replicated at each resample, and it represents an estimate of the loss in actual risk in the uncomplete model. For a set of variables $U = \{u_1, \ldots, u_r\}$, this statistic also represents the estimated loss due to the information in the removed variables. An adequate risk measurement in a classification task is the classification error probability $(P_e^*(b))$.

Note that complex interactions among the input variables can be expected whenever a nonlinear model is built, such as collinearity (for the nonlinear case, co-information or redundant information), irrelevant or noisy variables, and subsets of variables being relevant only when interacting among them. These situation have been widely studied in the discriminant analysis literature. With this BR approach, there is not a statistic associated to the CI, but still it is possible to propose a backward procedure. In order to eliminate one variable at each step, we can remove the variable either with the widest or with the smallest CI overlapping zero. The procedure for variable selection, then, is as follows:

1. Start with all the input variables. Calculate the B resamples and their corresponding actual risk minimization for (a) the complete input space $P_e^*(b)$ and (b) the incomplete model $P_{e,u}^*(b)$.
2. Compute the statistic $\triangle P_e^*(b) = P_{e,u}^*(b) - P_e^*(b)$ and find the 95% CI.
3. If there is any variable with CI overlapping 0:
    - remove the variable with wider CI overlapping 0, or
    - remove the variable with smaller CI overlapping 0.
4. Finish whenever every variable has a not zero-overlapping CI.

It should be noted that after fixing σ and C by using a cross-validation technique (N-fold,$N = 5$), SVM is retrained for every reduced model resampling.

4 Toy Example

The first experiment deals with feature selection problem for a synthetic set of data. This set consists of eight features $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_8)$, in which the first two variables define an XOR problem (and consequently they label the problem) with $N(0,1)$. From $\mathbf{x}_3$ to $\mathbf{x}_5$ different noisy variables are introduced (Gaussian $N(0,2)$, Uniform $U(0,1)$ and Rayleigh $R(1)$ noise, respectively). Collinearity is introduced with $\mathbf{x}_6 = \mathbf{x}_1 + 3\mathbf{x}_2 + N(0,2)$ and $\mathbf{x}_7 = \mathbf{x}_1 - \mathbf{x}_2 + N(0,2)$, while $\mathbf{x}_8 = (\mathbf{x}_1)^2 \cdot (\mathbf{x}_2)^2$ implements a nonlinear combination of these variables.

We test our BISSP for this data set, using B = 50 resamples and applying the two criteria, i.e, wider CI and smaller CI. Better results, in terms of P_e, are obtained in the case of wider CI, as can be seen in Table 1. Figure 1 represents the error probability statistic ($\triangle P_e^*$) histograms of the variables$\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_7$ in 3 different steps of the BISSP (wider CI criterion). As the procedure evolves, the relevance of the variables to be selected increases, so the the statistic becomes larger and positive. Non zero-overlapping variables, like $\mathbf{x}_7$, whose statistic result in a "negative" histogram are eliminated, since they are even increasing the P_e.

Criterion	P_e (Complete model)	P_e (Reduced model)	Selected Variables
Wider CI	0.05|0.06	0.05|0.04	$\mathbf{x}_1$, $\mathbf{x}_2$
Smaller CI	0.05|0.07	0.08|0.07	$\mathbf{x}_1$, $\mathbf{x}_6$

Table 1: BISSP criteria comparison (TRAIN | TEST)

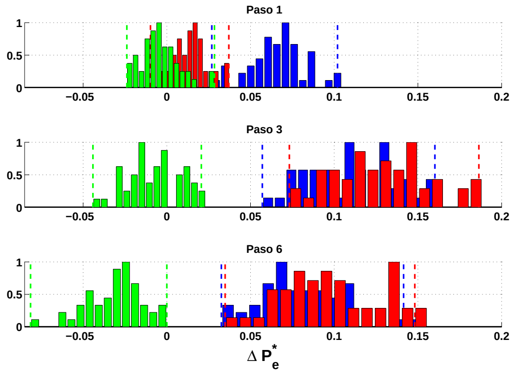

Fig. 1: $\triangle P_e{}^*$ histograms of the variables $\mathbf{x}_1$ (blue), $\mathbf{x}_2$ (red) and $\mathbf{x}_7$ (green).

5 VF Feature Selection

The second experiment presents a direct application of the problem of feature selection. In [11], 29 records from AHA and MIT/BIH databases were characterized by 27 time(t), frequency(f) and time-frequency(tf) parameters. We presented in [10] a BR technique for selecting a set of variables (time, frequency or time-frequency) that better discriminate VF. Here, we extend this approach to the BISSP previously described. Table 2 shows the mean values and the CI of $\triangle P_e^*(\%)$ for the selected variables after the BISSP. For the completed input space $P_e = 0.02$, while for the reduced data set $P_e = 0.07$. It is remarkable that the error probability for the excluded variables resulted to be $P_e = 0.06$, showing that this problem contains a lot of redundant information. However, the selected input space ensures a good performance of the FV detection while keeping a smaller dimensionality.

Selected Variable	$\triangle\mathbf{P}_\mathbf{e}^*(\%)$
curve (f)	2.28 [1.17|3.66]
qtel (f)	1.85 [0.89|3.42]
maximfreq (f)	1.80 [0.84|3.71]
tmy (tf)	1.65 [0.43|3.09]
mdl8 (t)	1.61 [0.51|2.36]
ct8 (t)	1.60 [0.51|2.72]
dispersion (tf)	1.42 [0.28|2.73]
area (tf)	1.38 [0.47|2.61]
tsnz (tf)	1.30 [0.37|2.40]
pmxfreq (f)	1.22 [0.05|2.08]
minfreq (f)	1.17 [0.14|2.13]

Table 2: Mean $\triangle P_e^*(\%)$ and CI for the selected variables.

6 Conclusions

In this study we have presented a novel feature selection procedure by using BR techniques for SVM classifiers. The reduced selected input space provides with a good performance of the detector, while reducing the dimensionality.

References

[1] V. Vapnik, *The Nature of Statistical Learning Theory.* New York: Springer–Verlag, 1995.

[2] L. Wang, *Support Vector Machines: Theory and Applications.* Springer-Verlag, 2005.

[3] Y. Lee and C.-K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinformatics*, vol. 19, pp. 1132 – 1139, June 2003.

[4] T. Hastie and R. Tibshirani, "Discriminant analysis by Gaussian mixtures," *Journal of the Royal Statistical Society series B*, vol. 58, pp. 158–176, 1996.

[5] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller, "Fisher discriminant analysis with kernels," 1999.

[6] V. Roth and V. Steinhage, "Nonlinear discriminant analysis using kernel functions technical report nr," 1999.

[7] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for SVMs," in *NIPS*, 2000, pp. 668–674.

[8] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap.* Chapman&Hall, 1998, vol. 57.

[9] J. L. Rojo-Álvarez, A. Arenal-Maíz, and A. Artés-Rodríguez, "Support vector black-box interpretation in ventricular arrhythmia discrimination," *IEEE Eng. in Med. and Biol.*, vol. 21, pp. 27–35, 2002.

[10] F. Alonso-Atienza, G. Camps-Valls, A. Rosado-Muñoz, and J. Rojo-Álvarez, "Selección de características en máquinas de vectores soporte para la discriminación automática de fibrilación ventricular," in *CASEIB*, 2005, pp. 435–438.

[11] A. Rosado, A. Serrano, M. Martínez, E. Soria, J. Calpe, and M. Bataller, "Detailed study of time-frequency parameters for ventricular fibrillation detection," in *ESEM*, 1999, pp. 379–380.