Analysis of Support Vectors Helps to Identify Borderline Patients in Classification Studies

F Schwenker[1], HA Kestler[1,2]

[1]Neural Information Processing, University of Ulm, Germany
[2]Medicine II – Cardiology, University Hospital Ulm, Germany

Abstract

In this work a new approach to the support vector machine (SVM) method is taken. Not in developing a new algorithm, but rather in analyzing the result of the performed classification tasks. The SVM approach provides efficient and powerful classification algorithms. SVM-Classifiers have a few meta parameters to be tuned, are easy to implement, and are trained through optimization of a quadratic cost function, which ensures the uniqueness of the SVM solution. The SVM solution is given through a linear combination of the training samples which are selected by the SVM optimization procedure. This subset of borderline samples close to the decision boundary can be separated into the samples which are misclassified and those samples that are just classified correctly. The potential drawback of the SVMs being restricted to samples from the dataset is at the same time an advantage in medical applications. Here, we applied this approach to a highly selected group of 44 patients with inducible ventricular tachycardia and a group of 51 healthy subjects.

1. Background

High-resolution electrocardiography is used for the detection of fractionated micropotentials, which serve as a noninvasive marker for an arrhythmogenic substrate and for an increased risk for malignant ventricular tachyarrhythmias. Ventricular late potential analysis (VLP) is herein the generally accepted noninvasive method to identify patients with an increased risk for reentrant ventricular tachycardias and for risk stratification after myocardial infarction [1–4]. Techniques commonly applied in this purely time–domain based analysis are signal-averaging, high-pass filtering and late potential analysis of the terminal part of the QRS complex. The assessment of VLP's depends on three empirically defined limits of the total duration of the QRS and the duration and amplitude of the terminal low–amplitude portion of the QRS [5, 6]. Rosenbaum et al. [7] have shown that increased beat–to–beat microvariations of the T–wave, although visually inapparent, are associated with a decreased arrhythmia–free survival. Their method to quantify periodic electrical alternans of the T–wave amplitude has gained growing clinical acceptance as a non–invasive, electrocardiographic risk marker. Previous work of our group showed a significantly higher beat–to–beat variation of the duration of the filtered QRS and an increased total beat–to–beat microvolt variation of both the QRS and the ST-T segment among patients with an increased risk for ventricular tachycardias [8].

In this paper we investigate the importance of these empirically found parameter settings. Here, we concentrate on the so-called support vector machine (SVM) method for linear classifiers [9]. For binary classification problems the aim is to find a decision function $H : \mathbb{R}^d \rightarrow \{0, 1\}$ by using a given set of M predefined training vectors $(x^\mu, y^\mu)_{\mu=1}^M$, with input vector $x^\mu \in \mathbb{R}$ and with $y^\mu \in \{0, 1\}$ (or $y^\mu \in \{-1, 1\}$) as the class label of input x^μ. The paper is organized as follows: In Section 2 support vector learning for linear classifiers is introduced and in Section 3 we describe the group of patients and give details of the data recordings. Finally we present and discuss the results.

2. Linear support vector machines

SVMs were initially developed to classify data points of linear separable data sets [10, 11]. In this case a training set consisting of M examples (x^μ, y^μ), $x^\mu \in \mathbb{R}$, and $y^\mu \in \{-1, 1\}$ can be divided up into two sets by a separating hyperplane. Such a hyperplane is determined by a weight vector $w \in \mathbb{R}^d$ and a bias or threshold $\theta \in \mathbb{R}$ satisfying the separating contraints

$$y^\mu(\langle x^\mu, w\rangle + \theta) \geq 1 \quad \mu = 1, \ldots, M. \tag{1}$$

The distance between the separating hyperplane and the closest data points of the dataset is called the *margin*. Finding the optimal hyperplane with maximal is a quadratic optimization problem with linear constraints.

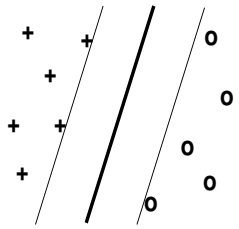

Figure 1. Separating hyperplane with an optimal large margin.

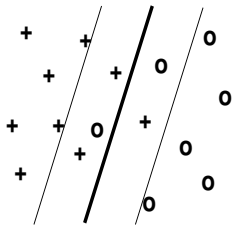

Figure 2. Linear nonseparable classification problem with foru data points within the margin, two of these are misclassified.

Weight vector $w \in \mathbb{R}^d$ and bias $\theta \in \mathbb{R}$ are determined by minimizing the functional

$$\Phi(w) = \frac{1}{2}\|w\|^2$$

subject to the separating constraints given by Eq 1. Using the Lagrangian formulation and the Kuhn-Tucker conditions this problem can be formulated in it's dual form. Here the functional

$$Q(\alpha) = \sum_{\mu=1}^{M} \alpha_\mu - \frac{1}{2}\sum_{\mu=1}^{M}\sum_{\nu=1}^{M} \alpha_\mu \alpha_\nu y^\mu y^\nu \langle x^\mu, x^\nu \rangle$$

has bo to maximized with subject to the constraints $0 \leq \alpha_\mu$ for all $\mu = 1, \ldots, M$ and $\sum_{\mu=1}^{M} \alpha_\mu y^\mu = 0$. The separating hyperplane with maximal margin is unique and can be expressed by a linear combination of those training examples (so-called support vectors) lying exactly at the margin and has the form

$$H(x) = \sum_{\mu=1}^{M} \alpha_\mu^* y^\mu \langle x, x^\mu \rangle + \theta.$$

where $\alpha_1^*, \ldots, \alpha_M^*$ is the solution of the optimization procedure. Only for the vectors x^μ where the Langrangian mulipliers α^μ have positive values different from zero contributed to the separating hyperplane H, so H is given through a linear combination of the set of support vectors (SVs):

$$H(x) = \sum_{x^\mu \in SVs} \alpha_\mu^* y^\mu \langle x, x^\mu \rangle + \theta.$$

So, finding the optimal separating hyperplane leads to a quadratic programming (QP) problem, which can be solved using standard software packages.

For a dataset that cannot be linearly separated, it would be desirable

1. to separate the data set with minimal error rate
2. to find a hyperplane with maximal margin, so called *soft margin* for the subset of separable data points.

Such a hyperplane is called a *soft margin hyperplane*. In this case a data point x^μ may be we classified correctly but is located very close to the decision boundary and lies within the soft margin. These data points together with the misclassified data points are called the borderline points in this paper. For this nonseparable classification problem positive slack variables $\zeta_1, \ldots, \zeta_M$ are introducted to quantify the borderline data points. These slack variables are used to weaken the separating conditions given in Eq 1:

$$y^\mu(\langle x^\mu, w \rangle + \theta) \geq 1 - \zeta_\mu \quad \mu = 1, \ldots, M. \qquad (2)$$

For a data point x^μ the slack variable ζ^μ is the deviation from the margin border, and slack variables greater than zero correspond to points within the margin, and slack variables greater than one to misclassified data points (see Fig2). It is possible to solve this nonseparable discrimination problem in terms of quadratic optimization by introducing a modified formulation. As for the separable problem, the hyperplane is defined by a weight vector $w \in \mathbb{R}^d$ and bias $\theta\mathbb{R}$ and is defined by minimizing the functional

$$\Phi(w, \zeta) = \frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{\mu=1}^{M} \zeta_\mu$$

subject to the soft constraints

$$y^\mu(\langle x^\mu, w \rangle + \theta) \geq 1 - \zeta_\mu \quad \mu = 1, \ldots, M$$

and a given sufficiently large fixed constant $C > 0$. Transforming this problem in the dual form gives

$$Q(\alpha) = \sum_{\mu=1}^{M} \alpha_\mu - \frac{1}{2}\sum_{\mu=1}^{m}\sum_{\nu=1}^{M} \alpha_\mu \alpha_\nu y^\mu y^\nu \langle x^\mu, x^\nu \rangle$$

which hast to be maximized subject to the constraints

$$0 \leq \alpha_\mu \leq C/M$$

for all $\mu = 1, \ldots, M$ and

$$\sum_{\mu=1}^{M} \alpha_\mu y^\mu = 0.$$

As for the separable case the soft margin hyperplane has a representation in terms of the data points x^μ with $\alpha_\mu > 0$.

3. Patients and ECG recordings

High resolution beat-to-beat recordings were obtained from 95 subjects separated into two groups. Group A consisted of 51 healthy volunteers (age 24±4.2 years) without any medication. In order to qualify as healthy, the several risk factors and illnesses had to be excluded. Group B consisted of 44 patients (age 61.2±8.9 years). Inclusion criteria were an inducible clinical ventricular tachycardia (>30 sec) at electrophysiologic study (EPS) with a history of myocardial infarction and coronary artery disease in angiogram, see [8] for more details. Signal-averaged high resolution ECGs were recorded from three orthogonal bipolar leads (X, Y, Z) with sampling rate was 2000 Hz (A/D-resolution: 16 bit; bidirectional filtering with a 40–250 Hz 4-pole Butterworth filter). From the vector magnitude $V = \sqrt{X^2 + Y^2 + Z^2}$ the following three features are derived:

- QRSd: total duration of the filtered QRS
- tRMS: RMS of the terminal 40 msec of the QRS
- LAS: terminal low-amplitude signal of the QRS.

Ventricular late potential analysis (VLP) positive: if 2 of 3 criteria are met: $QRSd > 114msec$, $tRMS < 20\mu V$, and $LAS > 38msec$ [1, 6].

For the beat–to–beat recordings of 30min duration the sampling rate was reduced to 1000Hz. QRS triggering, reviewing of the ECG, and arrhythmia detection was done on a high–resolution ECG analysis platform developed by our group [12]. The three leads were summed into a signal $V = X + Y + Z$. From each recording 250 consecutive sinus beats preceded by another sinus beat were selected for subsequent beat–to–beat variability analysis. In a first step the signals were aligned by maximizing the cross–correlation function [13] between the first and all following beats. Prior to the quantification of signal variability the beats were pre-processed to suppress the main ECG waveform, bringing the beat–to–beat micro-variations into clearer focus. To achieve this, the individual signal was subtracted from its cubic spline smoothed version (spline filtering, spline interpolation through every seventh sample using the not–a–knot end condition) [14]. This method resembles a waveform adaptive, high–pass filtering without inducing phase–shift related artefacts. Next, for each individual beat the amplitude of the difference signal was normalized to zero mean and a standard deviation of 1μV. Beat–to–beat variation of each point was measured as the standard deviation of the amplitude of corresponding points across all 250 beats. For the QRS we used a constant analysis window of 141 ms which covered all QRS complexes of this series [8]. The intra-QRS signal variability index (QVI) was defined as the sum of the standard deviations of corresponding points. Repolarization microvariability inside the ST-T segment (TVI) was measured with the same adaptive spline filtering technique (constant window starting at the QRS-offset of 400ms) but using 3 instead of 21 knots.

4. Results and conclusion

In this investigation the five features QRSd, tRMS, LAS, QVI, and TVI are used as inputs to the linear classifiers trained through support vector learning in order to predict the group status. First, the classification performance is given in terms of *10-fold crossvalidation* results (see Table 1). 10-fold crossvalidation means partitioning the whole data set into 10 disjoint subsets and carrying out 10 training and test runs always using 9 subsets as the training set and testing on the remaining one. The 10-fold crossvalidation simulation was performed 10 times. The difference between subsequent simulations was the random permutation of the data set.

Table 1. Classification results (mean of ten 10-fold crossvalidation runs and standard deviation is given) for the standard 2-of-3 VLP rule and the linear SVM for different input features, in terms of accuracy, sensitivity, and specificity. For the QP optimization the NAG-libary was used. (**A**: standard VLP 2-of-3 rule. Features used inputs in the linear SVM classifiers: **B**: QRSd, tRMS and LAS, **C**: QRSd, **D**: QVI and TVI features, **E**: QRSd, tRMS, LAS, QVI and TVI, **F**: QRSd, QVI and TVI).

	acc [%]	**sensi [%]**	**speci [%]**
A	72.6	63.6	80.4
B	85.47 ± 0.83	68.63 ± 1.79	100.0 ± 0.0
C	83.58 ± 1.07	69.32 ± 1.93	95.88 ± 1.45
D	78.21 ± 0.86	74.09 ± 1.59	81.76 ± 1.32
E	87.79 ± 1.13	77.72 ± 2.08	96.47 ± 1.55
F	89.05 ± 0.74	79.09 ± 0.96	97.64 ± 0.83

Compared to the 2-of-3 rule an increase of accuracy of more than 10%, an increase of sensitivity of about 5%, and an increase of specificity of approximately 20% could be achieved using the linear SVMs. The best performance in terms of accuray and sensitivty was achieved by using QRSd, QVI and TVI as input features (without tRMS and LAS). Although the error rate of the SVM classifiers is significantly be improved, it has to be mentioned that the sensitivity is still too low to qualify as a single screening test.

Finally, we want to illustrate how the structure of SVM classifiers can be utilize to analyze the input features spaces. In Fig.3 the Lagrangian multipliers (LM) α_μ for the five different feature sets (see caption of Table1) are given by a 95×5-matrix. A dark entry encodes a

large positive value of α_μ, and a white entry stands for $\alpha_\mu = 0$, so the dark entries are the borderline samples. We found 40 borderline samples with the standard signal averaged features and 48 borderline samples with the beat-to-beat variability features, and only 26 samples for the intersection of both features. We conclude that static and dynamic parameters derived form the high-resolution ECG give a complementary view on the cardiac conduction characteristics of these paptient groups.

Training a linear SVM leads to a quadratic optimization problem that can easily be solved by scientific software packages (e.g. `matlab`), whereas the standard learning rules for neural networks are derived from highly nonlinear target functions with many disadvantages (e.g. local minima). This is an important advantage of SVMs in applications where a classifier system has to be built by a non-expert in the field of classifier design. Through SVM training a subset of samples is identified, an useful feature of the SVM method, particularly in medical applications.

Figure 3. The 95×5-matrix of the Lagrangian multipliers $\alpha_1 \ldots . \alpha_{95}$ for five different input feature sets. B,C,D,E,F denote the five different feature sets (see Table 1).

Acknowledgement

We would like to thank Dr. Martin Höher and Professor Vinzenz Hombach for providing the medical data and an inspiring research enviroment, and Professor Günther Palm for many fruitful discussions.

References

[1] Simson M. Use of Signals in the Terminal QRS Complex to Identify Patients with Ventricular Tachycardia after Myocardial Infarction. Circulation 1981;64(2):235–242.

[2] Gomes J, Winters S, Martinson M, Machac J, Stewart D, Targonski A. The prognostic significance of quantitative signal-averaged variables relative to clinical variables, site of myocardial infarction, ejection fraction and ventricular premature beats. JACC 1989;13:377–384.

[3] Höher M, Hombach V. Ventrikuläre Spätpotentiale – Teil I Grundlagen. Herz Rhythmus 1991;3(3):1–7.

[4] Höher M, Hombach V. Ventrikuläre Spätpotentiale – Teil II Klinische Aspekte. Herz Rhythmus 1991;3(4):8–14.

[5] Breithardt G, Borggrefe M. Pathophysiological mechanisms and clinical significance of ventricular late potentials. Eur Heart J 1986;7:364–385.

[6] Breithardt G, Cain M, El-Sherif N, Flowers N, Hombach V, Janse M, Simson M, Steinbeck G. Standards for analysis of ventricular late potentials using high resolution or signal-averaged electrocardiography. Eur Heart J 1991;12:473–80.

[7] Rosenbaum D, Jackson L, Smith L, Garan H, Ruskin J, Cohen R. Electrical alternans and vulnerability to ventricular arrhythmias. N Engl J Med 1994;330(4):235–41.

[8] Kestler HA, Wöhrle J, Höher M. Cardiac vulnerability assessment from electrical microvariability of the high-resolution electrocardiogram. Medical Biological Engineering Computing 2000;38:88–92.

[9] Cortes C, Vapnik V. Support vector networks. Machine Learning 1995;20:273–297.

[10] Schölkopf A, Burges C, Smola A. Advances in Kernel Methods — Support Vector Learning. MIT Press, 1998.

[11] Vapnik V. Statistical Learning Theory. John Wiley and Sons, 1998.

[12] Ritscher D, Ernst E, Kammrath H, Hombach V, Höher M. High-resolution ecg analysis platform with enhanced resolution. Computers in Cardiology 1997;24:291–94.

[13] van Bemmel J, Musen M (eds.). Handbook of Medical Informatiks. Springer Verlag, 1997.

[14] de Boor C. A Practical Guide to Splines. Springer Verlag, 1978.

Address for correspondence:

Friedhelm Schwenker
Dept. of Neural Information Processing / University of Ulm
D-89069 Ulm / Germany
tel.: ++49-731-5024159
fschwenker@neuro.informatik.uni-ulm.de