ELSEVIER

Maximum margin equalizers trained with the Adatron algorithm

Ignacio Santamaría[a,*], Rafael González[a], Carlos Pantaleón[a], José C. Principe[b]

[a]*DICOM, ETSII y Telecomunicacion, Univ. of Cantabria, Santander 39005, Spain*
[b]*Computational NeuroEngineering Laboratory, Univ. of Florida, Gainesville, USA*

Abstract

In this paper we apply the structural risk minimization principle as an appropriate criterion to train decision feedback and transversal equalizers. We consider both linear discriminant (optimal hyperplane) and nonlinear discriminant (support vector machine) classifiers as an alternative to the linear minimum mean-square error (MMSE) equalizer and radial basis function (RBF) networks, respectively. A fast and simple adaptive algorithm called the Adatron is applied to obtain the linear or nonlinear classifier. In this way we avoid the high computational cost of quadratic programming. Moreover, the use of soft margin (regularized) classifiers is proposed as a simple way to consider "noisy" channel states: this alternative improves the bit error rate, mainly at low SNR's. Furthermore, an adaptive implementation is discussed. Some simulation examples show the advantages of the proposed linear and nonlinear equalizers: a better performance in comparison to the linear MMSE and a simpler structure in comparison to the RBF (Bayesian).
© 2002 Elsevier Science B.V. All rights reserved.

Keywords: Equalization; Support vector machines; Structural risk minimization; Maximum margin classifiers

1. Introduction

Digital communication receivers are often impaired by intersymbol interference when the transmitted symbol sequence goes through a band-limited channel. An equalizer is then used to obtain a reliable data transmission. In the literature, the transversal (see Fig. 1) and the decision feedback structure (see Fig. 2) are the most popular equalizers [24,19]. Both structures can perform a linear or nonlinear filtering of the observations (using a linear or nonlinear discriminant function as shown in Figs. 1 and 2, respectively).

In particular, the optimal nonlinear (Bayesian) equalizer has been developed for transversal and decision feedback structures in [6,9], showing that, for both structures, it can be implemented as a radial basis function (RBF) network.

From a classification point of view, the FIR- and RBF-based equalizers separate the channel states using, respectively, a linear and nonlinear discriminant function, and they both utilize the minimum mean-square error (MMSE) criterion in choosing the optimal weights.

However, it is well known that, in general, the MMSE design does not yield the minimum bit error rate (BER) solution. A more appropriate criterion would be to formulate the problem as a classification task and design a maximum margin classifier by applying the structural risk minimization (SRM)

* Corresponding author. Tel.: +34-42-201552; fax: +34-42-201488.
E-mail addresses: nacho@gtas.dicom.unican.es (I. Santamaría), principe@cnel.ufl.edu (J.C. Principe).
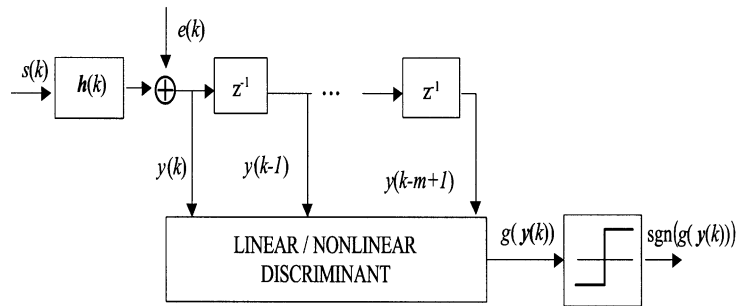
Fig. 1. Schematic diagram of a generic transversal equalizer using a linear or nonlinear discriminant function.
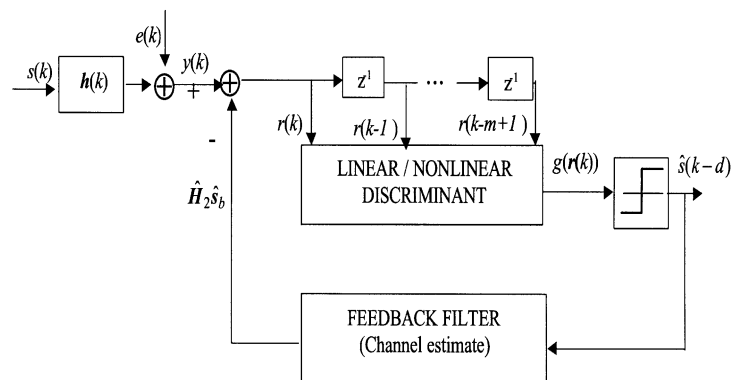

Fig. 2. Schematic diagram of a generic decision feedback equalizer using a linear or nonlinear discriminant function.

principle [27]. The maximum margin classifier can be applied in the input space, providing a linear discriminator that we call the optimal hyperplane (OH), or in feature space implementing a support vector machine (SVM). The architecture of the equalizer is now given in terms of a reduced set of critical training samples known as support vectors (SV).

Recently, some research has been conducted to study the application of SV machines in equalization problems. For instance, in [2,4], the OH was obtained for a DFE showing its potential improvement over the linear MMSE-DFE solution. In [25] an SVM using a polynomial kernel was applied for nonlinear equalization achieving a performance similar to that of neural network-based equalizers. Similar results have been recently found using an SVM with Gaussian kernel [17]. Also, in [7] it is shown that, for high SNR's, the optimal nonlinear discriminant can be approximated using a piecewise linear boundary. This result has

been extended to M-ary pulse amplitude modulation (PAM) signals in [3]. The set of hyperplanes defining the boundary is then obtained by applying the SRM principle.

In this paper we extend these previous works in the following directions: firstly, we systematize the comparisons between the MMSE and the SRM criteria for linear and nonlinear classifiers. Secondly, we consider the application of soft margin SV machines as a way to include a priori information about the noise variance. Finally, to avoid the high computational cost of training SV machines by solving a quadratic optimization problem, we use a fast and simple procedure known as the Adatron algorithm [1,12]. It obtains exactly the SVM solution, but with an exponential rate of convergence in the number of iterations.

By means of some simulation examples we show that the soft margin OH outperforms its MMSE counterpart. On the other hand, if the kernel function to

map the input space into the feature space is chosen as the Gaussian kernel, we have a topology similar to the RBF network, that differs only on the criterion for optimization. The advantage of SVM over RBF equalizers is that an SVM provides a pruned structure since only those training samples or channel states that are important for classification become support vectors. In this way, an SVM equalizer allows the user a tradeoff between complexity and performance.

2. Linear and nonlinear equalization using transversal and decision feedback-structures

In this section we briefly review a general framework to study transversal and decision feedback equalization structures from a classification point of view.

The received signal at the input of the equalizer can be expressed as

$$y(k) = \sum_{i=0}^{n} h_i s(k - i) + e(k), \qquad (1)$$

where the transmitted symbol sequence $s(k)$ is assumed to be an equiprobable binary sequence $\{+1, -1\}$, h_i are the channel coefficients, and the measurement noise $e(k)$ can be modelled as a zero-mean Gaussian with variance σ_n^2.

The transversal equalizer (TE) shown in Fig. 1 is probably the most popular equalization structure: it estimates the value of a transmitted symbol as a (linear or nonlinear) combination of the channel observations, i.e.,

$$\hat{s}(k - d) = \operatorname{sgn}(g(\mathbf{y}(k))), \qquad (2)$$

where $g(.)$ is the linear or nonlinear discriminant function, $\mathbf{y}(k) = [y(k), \ldots, y(k - m + 1)]^T$ is the vector of observations and d is the equalizer delay.

As it was firstly pointed out in [13], the equalization of a digital communications channel can be viewed as a classification problem. To further clarify this idea, let us note that the noiseless vector of channel observations can be expressed using matrix notation as

$$\mathbf{r}(k) = \mathbf{H}\mathbf{s}(k), \qquad (3)$$

where $\mathbf{s}(k) = [s(k), \ldots, s(k - m - n + 1)]^T$ is the vector of transmitted symbols and $\mathbf{H}$ is an $m \times (m + n)$ Toeplitz

channel matrix given by

$$\mathbf{H} = \begin{pmatrix} h_0 & \cdots & h_{nc} & 0 & \cdots & 0 \\ 0 & h_0 & \cdots & h_{nc} & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & h_0 & \cdots & h_n \end{pmatrix} \qquad (4)$$

For a channel with $n + 1$ taps and a TE with m coefficients there are $N = 2^{(m+n)}$ different channel states $\mathbf{r}_i$, which can be obtained as $R = \{\mathbf{r}_i = \mathbf{H}\mathbf{s}_i; i = 1, \ldots, N\}$, where $\mathbf{s}_i$ represents the $2^{(m+n)}$ possible input sequences. For a binary digital signal, R can be partitioned into the following two subsets:

$$R^{(\pm 1)} = \{\mathbf{r}_i, \ s(k - d) = \pm 1\}. \qquad (5)$$

Then, an equalizer must find a classification boundary between $R^{(+1)}$ and $R^{(-1)}$. In particular, the optimal symbol-by-symbol Bayesian equalizer defines a nonlinear boundary $g(\mathbf{r}(k)) = 0$, which is given by

$$g(\mathbf{r}(k)) = \sum_{\mathbf{r}_i \in R^{(+1)}} \lambda_i \exp\left(-\frac{\|\mathbf{r}(k) - \mathbf{r}_i\|_2^2}{2\sigma_n^2}\right)$$

$$- \sum_{\mathbf{r}_i \in R^{(-1)}} \lambda_i \exp\left(-\frac{\|\mathbf{r}(k) - \mathbf{r}_i\|_2^2}{2\sigma_n^2}\right), \qquad (6)$$

where λ_i is the probability of appearance of each channel state, which under the equiprobability assumption is $\lambda_i = 1/N$, since all the channel states have the same probability. On the other hand, the conventional linear transversal equalizer (LTE) defines a linear boundary through a hyperplane

$$g(\mathbf{r}(k)) = \mathbf{w}^T \mathbf{r}(k). \qquad (7)$$

The use of a transversal structure for equalization has several drawbacks: the linear separability of the channel states is not guaranteed [13] and, besides, the LTE suffers from noise enhancement when the channel has zeros close to the unit circle. On the other hand, if we implement a nonlinear boundary through (6), the number of channel states grows as 2^{m+n} and, even for moderate channel lengths, the optimal Bayesian equalizer becomes unfeasible.

To mitigate these problems we can employ decision feedback. The equalization structure in this case is depicted in Fig. 2. For the DFE, the channel observations (3) can be partitioned as

$$\mathbf{y}(k) = \mathbf{H}_1 \mathbf{s}_f(k) + \mathbf{H}_2 \mathbf{s}_b(k), \qquad (8)$$

where $\mathbf{H}_1$ and $\mathbf{H}_2$ are composed of the first m columns and the last n columns of the channel matrix $\mathbf{H}$. That is, using Matlab notation, $\mathbf{H}_1 = \mathbf{H}(:, 1 : m)$ and $\mathbf{H}_2 = \mathbf{H}(:, m + 1 : m + n_c)$. This partitioning of the channel matrix corresponds to a DFE with feedforward order m, feedback order n (equal to the channel order), and equalizer delay $d = m - 1$: these values are often chosen in conventional DFE's [9]. Now, assuming that the past decisions are correct, the vector of channel states for the DFE (denoted again as $\mathbf{r}(k)$) can be expressed as

$$\mathbf{r}(k) = \mathbf{y}(k) - \mathbf{H}_2 \mathbf{s}_b(k) \qquad (9)$$

and then, as it was shown in [5], the conventional linear DFE is equivalent to a filtering of the "translated" channel states. Again, this filtering can be nonlinear as in (6) or linear as in (7). The main advantages of using a DFE structure are that the number of channel states grows only as 2^m and that now the two classes are always linearly separable [5].

Regardless the use of a transversal or decision feedback equalization structure, typically an MMSE criterion is used to obtain the weights defining the linear or nonlinear discriminant. On the other hand, as it is shown in [6,9,10] the Bayesian detector (6) can be implemented using an RBF network by placing a Gaussian RBF unit at each channel state. Some techniques to reduce the number of relevant channel states via clustering have been recently proposed [14,15].

3. Maximum margin equalizers

3.1. OH and SVM equalizers

In recent years, SV machines have been successfully applied to several classification and pattern recognition problems. As it was described in the previous section, equalization is actually a classification problem. This suggests the interest in training the equalizer using the SRM principle, which allows one to obtain an OH separating two sets of points in the input space or in the feature space, yielding a linear or nonlinear equalizer, respectively.

In non-blind equalization, there are two approaches to use the information provided by the training sequence: we can use a direct approach by training the equalizer directly from the channel observations [25],

or we can follow an indirect approach which uses the training sequence to estimate the channel and then obtain the channel states through (3) for a transversal equalizer or (9) for a DFE. As we will see later, the algorithms used to obtain the OH are memory intensive. For this reason we feel that the indirect approach is more adequate in this context, since the channel states are, in fact, the centers of the channel observation clusters, thus providing somehow a reduction of the training set.

As we have discussed in the previous section, we can formulate our equalization problem as follows: given the training set $I = \{(\mathbf{r}_i, s_i), i = 1, \ldots, N\}$, where $\mathbf{r}_i$ are the channel states and $s_i \in \{+1, -1\}$ are the desired symbols; obtain the classifier that minimizes the BER. The nonlinear equalizer that minimizes the BER is the Bayesian one (6), which becomes unfeasible when the number of channel states is large. On the other hand, LTE's and DFE's that minimize the BER have been also proposed [5,8,18,28]. In these works, the cost function is directly the BER, which is derived and then minimized using gradient descent techniques. Since the BER computation involves all the channel states, some approximation must be made in order to make this approach practical, either by reducing the number of channel states or by estimating the pdf at the output of the equalizer using the Parzen windowing method.

In this work we consider the OH and SVM equalizers as linear or nonlinear approximations to the true minimum BER equalizer. Let us first assume that the channel states are linearly separable (this is always true for a DFE if the feedback order is equal to the channel order, and it also holds true for a LTE if its order and delay are large enough), a reasonable approximation to the minimum BER solution can be obtained by constructing an OH that maximizes the distance between the closest vectors to the hyperplane (i.e., the margin). In [27], it is shown that the weight vector $\mathbf{w}$ and the threshold b of the maximum margin hyperplane can be obtained by minimizing

$$J(\mathbf{w}, b) = \tfrac{1}{2} \|\mathbf{w}\|_2^2, \qquad (10)$$

subject to $s_i(\mathbf{r}_i \mathbf{w} + b) \geqslant 1, i = 1, \ldots, N$. This problem is equivalent to maximize the following quadratic form:

$$W(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j s_i s_j \langle \mathbf{r}_i, \mathbf{r}_j \rangle, \qquad (11)$$

subject to the constraints: $\alpha_i \geqslant 0$, and $\sum_{i=1}^{N} \alpha_i s_i = 0$.

Using matrix notation we can rewrite (11) as

$$W(\mathbf{\Lambda}) = \mathbf{\Lambda}^{\mathrm{T}}\mathbf{1} - \tfrac{1}{2}\mathbf{\Lambda}^{\mathrm{T}}\mathbf{D}\mathbf{\Lambda}, \tag{12}$$

where $\mathbf{1}$ is an N-dimensional unit vector, $\mathbf{\Lambda}$ is a vector with elements α_i, and $\mathbf{D}$ is an $N \times N$ matrix with elements given by $D_{ij} = s_i s_j \langle \mathbf{r}_i, \mathbf{r}_j \rangle$.

The solution of this optimization problem can be expanded in terms of the input patterns and the coefficients α_i as $\mathbf{w} = \sum_i \alpha_i s_i \mathbf{r}_i$, [27,11]. Only the training patterns which lie closest to the hyperplane have $\alpha_i > 0$ and are called support vectors; all others have $\alpha_i = 0$.

Finally, the decision function for the OH equalizer is given by

$$f(\mathbf{r}) = \mathrm{sgn}\left(\sum_i \alpha_i s_i \langle \mathbf{r}_i, \mathbf{r} \rangle + b \right). \tag{13}$$

For this particular problem, due to the symmetry of the classes, we know that the border passes through the origin, so that $b = 0$.

The linear classifier can be easily extended to implement a nonlinear decision boundary by replacing the inner product in (11) and (13) by a nonlinear kernel function $K(\mathbf{r}_i, \mathbf{r}_j)$ that satisfies the Mercer condition [27]. For an equalization problem, a suitable nonlinear kernel is a Gaussian RBF

$$K(\mathbf{r}_i, \mathbf{r}_j) = \exp\left(\frac{-\|\mathbf{r}_i - \mathbf{r}_j\|_2^2}{\sigma^2} \right), \tag{14}$$

where σ^2 is a given parameter. In this case, the SVM takes the form of an RBF equalizer

$$f(\mathbf{r}) = \mathrm{sgn}\left(\sum_i \alpha_i s_i K(\mathbf{r}_i, \mathbf{r}) \right). \tag{15}$$

Nevertheless, in a conventional RBF equalizer the units are located at each channel state (or a subset of them [14]), while the SVM selects only those channel states that maximize the margin in the feature space and therefore are relevant for classification purposes. This approach usually leads to a simpler structure.

3.2. Soft margin OH's

The channel states, which are used as support vectors, do not take into account the noise information. Actually, the pdf of the channel observations is a set of Gaussians centered at each of the channel states. In this way, the solution provided by solving (12) can

be considered optimal only for the asymptotic case of SNR $\rightarrow \infty$.

The MMSE solution, on the other hand, takes into account the noise variance through the autocorrelation matrix of the channel observations. When the noise variance increases, the MMSE hyperplane tends to rotate. This difference explains the observation that, at low SNR's, the MMSE solution can achieve a better BER that the OH solution (at least for some channels).

In the following we discuss an alternative to handle these high-noise situations. A first possibility to incorporate the noise into our classification problem could be to train the SV machine directly using the channel observations, $\mathbf{y}_i$, instead of the channel states, $\mathbf{r}_i$ [25,23]. However, this approach would yield a SV machine with a larger number of support vectors.

An alternative solution consists of constructing a soft margin hyperplane by minimizing the following regularized functional:

$$J(\mathbf{w}, b) = \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{N}(\xi_i)^{\beta}, \tag{16}$$

subject to the constraints $s_i(\mathbf{r}_i\mathbf{w} + b) \geqslant 1 - \xi_i$ and $\xi_i \geqslant 0$.

Typically, this soft margin alternative is used to handle situations where all the patterns cannot be correctly classified. Here we use it to improve the performance of the classifier in high-noise situations.

If we consider $\beta = 2$ in (16), the second term minimizes the least-squares errors (LSE). In this case the optimization problem is still quadratic [11]

$$W(\mathbf{\Lambda}) = \mathbf{\Lambda}^{\mathrm{T}}\mathbf{1} - \frac{1}{2}\left(\mathbf{\Lambda}^{\mathrm{T}}\mathbf{D}\mathbf{\Lambda} + \frac{1}{C}\mathbf{\Lambda}^{\mathrm{T}}\mathbf{\Lambda} \right), \tag{17}$$

subject to the constraints: $\mathbf{\Lambda} \geqslant 0$, and $\sum_i \alpha_i s_i = 0$.

From (17) we see that the LSE soft margin reduces to regularize the kernel matrix $\mathbf{D}$ by adding $1/C$ to the elements of the main diagonal. The similarity with the regularization performed in the MMSE solution suggests to select the regularization parameter as $1/C \propto \sigma_n^2$. After an extensive number of simulations, we have found that the optimal value is $1/C_{\mathrm{opt}} = 2^m \sigma_n^2$ (m being the length of the feedforward filter for both the transversal and decision feedback structures).

4. The Adatron algorithm

The computational cost associated to the quadratic programming problems (12) or (17) is one the main drawbacks in applying either the OH or the SVM to practical equalization problems. Several alternative techniques have been proposed in the literature to solve this problem; for instance we could apply an iterated reweighted least square (IRWLS) learning algorithm as proposed in [22,23], or the Adatron algorithm [1,12]. The IRWLS technique requires a matrix inversion at each iteration. Then, even for a problem with a moderate number of channel states, the computational burden of the procedure is high.

In this paper, we use the Adatron algorithm, which unlike the IRWLS is an LMS-like adaptive algorithm. Specifically, at each iteration the Adatron chooses a pattern from the training set and updates the corresponding Lagrange multiplier according to

$$\alpha_i = \alpha_i + \max\{-\alpha_i, \eta(1 - s_i f(\mathbf{r}_i))\}, \tag{18}$$

where $f(\mathbf{r}_i)$ is the output of the SV machine and η is the learning rate. The updates of the Lagrange multipliers are made in accordance with the errors for each pattern presented to the SV machine. Following the convention in the neural networks literature, one complete presentation of the entire training set is called an epoch. Unlike the backpropagation algorithm, however, here we have a quadratic programming problem with a unique global minimum. Therefore, the order of presentation of the input patterns from epoch to epoch plays no special role.

In [1], it was proved that the Adatron converges to a maximum margin solution; that is, the minimum of (12) is a fixed point of the adaptive algorithm. Moreover, its convergence rate is exponential with the number of iterations. Finally, it is interesting to remark that the soft margin alternative is still a quadratic programming problem, and that the regularized kernel matrix remains positive definite, therefore the Adatron algorithm can be also used to find the solution in this case.

After each epoch of the Adatron algorithm the margin can be recomputed as

$$M_k = \min_{i \in \{1,\ldots,N\}} s_i f(\mathbf{r}_i), \tag{19}$$

where k denotes the epoch. This value can be used to check the convergence of the Adatron algorithm, for instance we have used the following criterion:

$$\left| M_k - \frac{1}{5} \sum_{j=1}^{5} M_{k-j} \right| \leqslant 10^{-5}. \tag{20}$$

The main advantages of the Adatron algorithm are its conceptual and implementation simplicity. However, it is a memory intensive algorithm, since all the kernel products, $K(\mathbf{r}_i, \mathbf{r}_j)$, must be precomputed and stored (in $\mathbf{D}$). When the number of channel states is large (mainly for transversal equalizers) we should also incorporate a "chunking" technique to handle in an efficient way a large number of training patterns [20].

Finally, using the Adatron algorithm we can propose the following adaptive implementation for the OH or SVM equalizers (valid for transversal or decision feedback implementations):

(1) Use a training sequence to estimate the channel impulse response $\hat{\mathbf{h}}$, and the noise variance $\hat{\sigma}_n^2$.
(2) Using $\hat{\mathbf{h}}$ estimate the channel states as (3) for a transversal structure or (9) if we use decision feedback.
(3) Initialize $\alpha_i = 0$, $i = 1,\ldots,N$, and the learning rate η.
(4) While convergence criterion (20) not true.
   For $i = 1,\ldots,N$.
   (4.1). Calculate update as $\delta_i = \eta(1 - s_i f(\mathbf{r}_i))$.
   (4.2). If $(\alpha_i + \delta_i) > 0$ then $\alpha_i = \alpha_i + \delta_i$, else $\alpha_i = 0$.
   End.
(5) Calculate the new margin as (19).
(6) End while.

Note that if the kernel correlation matrix $\mathbf{D}$ is precomputed and stored, $f(\mathbf{r}_j)$ can be efficiently obtained as

$$f(\mathbf{r}_j) = \langle \alpha \otimes \mathbf{s}, \mathbf{D}_j \rangle, \tag{21}$$

where $\otimes$ denotes elementwise multiplication, and $\mathbf{D}_j$ represents the jth row of matrix $\mathbf{D}$. Furthermore, to consider a soft margin SV machine we just have to regularize the kernel matrix as $\mathbf{D} + (1/C)\mathbf{I}$.

5. Results

The aim of the first example is to compare the performance of the linear MMSE-DFE with the OH using different values of the regularization parameter $1/C$. We send binary symbols through the channel
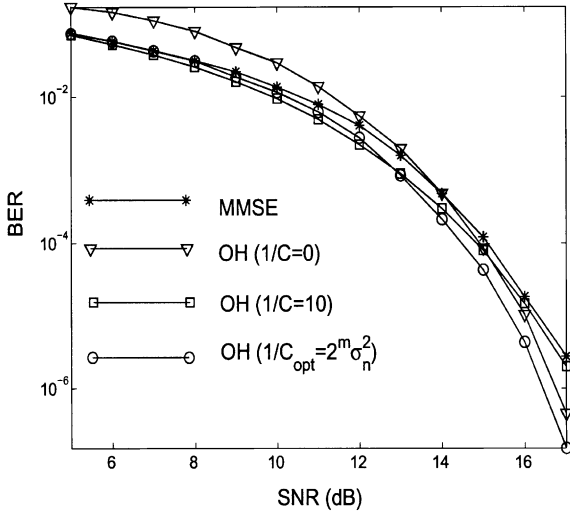
Fig. 3. MMSE versus OH for channel $H(z) = 0.2052 - 0.5131z^{-1} + 0.7183z^{-2} + 0.3695z^{-3} + 0.2052z^{-4}$ using a DFE with $m=4$ (feedforward order), $n=4$ (feedback order) and $d = 3$ (delay).

Fig. 4. MMSE, OH, SVM and Bayesian for channel $H(z) = 0.4 + 0.7z^{-3} + z^{-6} + 0.6z^{-11}$ using a DFE with $m = 8$ (feedforward order), $n = 11$ (feedback order) and $d = 7$ (delay).

$H(z) = 0.2052 - 0.5131z^{-1} + 0.7183z^{-2} + 0.3695z^{-3} + 0.2052z^{-4}$. The structure of the DFE is $m = 4$ (feedforward order), $n = 4$ (feedback order) and $d = 3$ (delay); then, the total number of channel states is 16. To train the OH we use the Adatron algorithm with a learning rate $\eta = 0.05$. Fig. 3 shows the BER curve for the MMSE-DFE and the soft margin OH. At low SNR's the non-regularized solution provides worse results than the MMSE. On the other hand, the soft margin solution with an optimal regularization parameter, $1/C_{opt} = 2^m\sigma_n^2$, is able to rotate the separating hyperplane according to the SNR, thus providing the best results.

In our second example, we use the following multipath channel: $H(z) = 0.4 + 0.7z^{-3} + z^{-6} + 0.6z^{-11}$, and a DFE with $m = 8$, $d = 7$ and $n = 11$. In this case, we have $2^8 = 256$ channel states. Fig. 4 shows the results obtained with the linear MMSE-DFE, the OH with $1/C = 0$, the soft margin OH with $1/C_{opt} = 2^m\sigma_n^2$, the SVM using a Gaussian kernel with $\sigma^2 = 2$ and the optimal Bayesian equalizer, which is implemented through an RBF.

For this example the total number of the channel states is 256, from which 92 and 152 become SV's for the OH and the SVM, respectively. On the other hand,
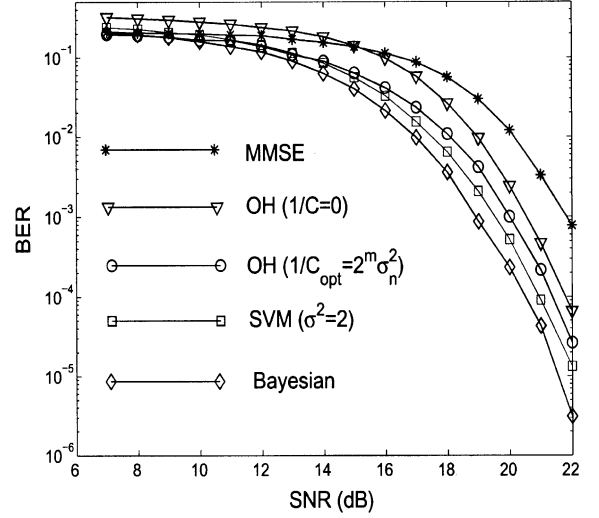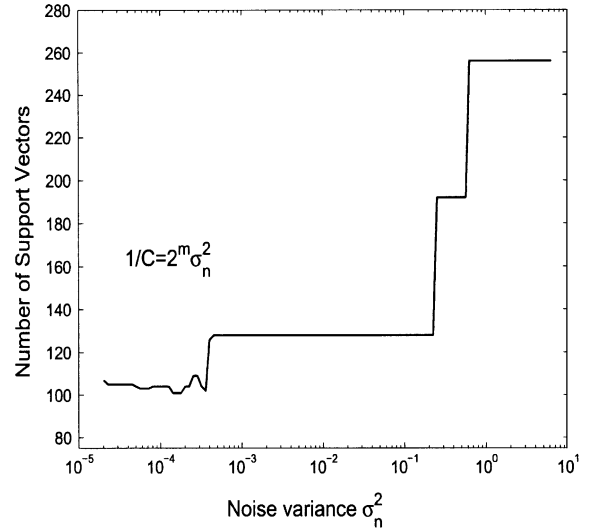

Fig. 5. Number of support vectors vs σ_n^2 for the soft margin OH. The channel is $H(z) = 0.4 + 0.7z^{-3} + z^{-6} + 0.6z^{-11}$ and we use a DFE with $m = 8$ (feedforward order), $n = 11$ (feedback order) and $d = 7$ (delay).

the number of SV's for the soft margin OH varies with the SNR, since the regularization parameter is chosen as $1/C_{opt} = 2^m\sigma_n^2$ (Fig. 5 shows the number of SV's for the OH with respect to σ_n^2). Finally, the
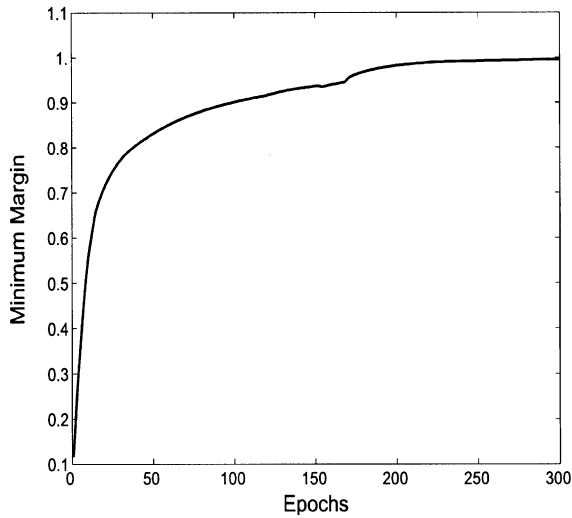
Fig. 6. Evolution of the margin with the number of epochs for the Adatron algorithm.

Fig. 7. MMSE, OH and SVM for channel $H(z) = 0.6 + z^{-1} + 0.5z^{-2} + 0.2z^{-3}$ using a LTE with $m = 7$ (feedforward order) and $d = 4$ (delay).

Bayesian equalizer uses the total number of channel states.

The performance of the SVM depends on the Gaussian kernel size, σ^2, in the following way: if the kernel size is chosen as the noise variance, all the channel states become support vectors and the SVM coincides with the Bayesian equalizer. As long as the kernel size decreases, the number of support vector (as well as the performance and the computational cost) also decrease. In the limit, when a very large kernel size is used, the border between classes become almost linear and the SVM reduces to the OH without regularization term. In this way, by using SVM's with different values of σ^2, we can tradeoff complexity for performance. In particular, for the chosen kernel size $\sigma^2 = 2$, the performance of the SVM is similar to the Bayesian equalizer with approximately half the complexity: 152 Gaussian kernels for the SVM instead of 256 for the Bayesian equalizer. Finally, an example of the convergence of the Adatron algorithm for this channel is depicted in Fig. 6.

Similar results have been found for transversal equalizers: for instance, Fig. 7 shows the results obtained when we equalize the channel $H(z) = 0.6 + z^{-1} + 0.5z^{-2} + 0.2z^{-3}$ using a transversal structure with $m = 7$ (feedforward order) and $d = 4$ (delay). The poor performance of the conventional
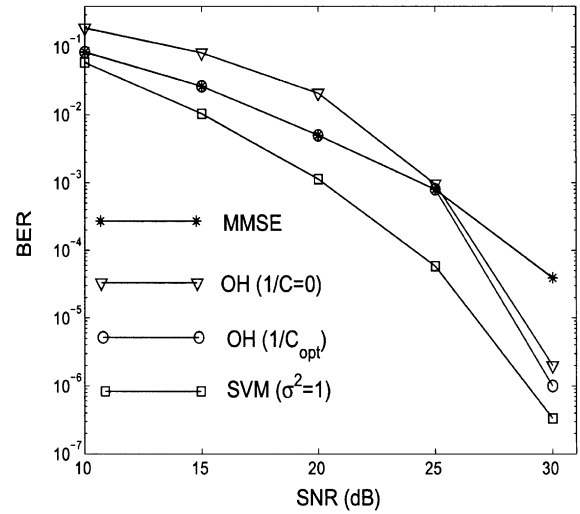
OH at low SNR's is more noticeable in this example. As in the previous examples the soft margin OH solves this problem.

In all the previous examples the equalizer has been trained using the true noiseless channel states. In a practical situation, of course, the channel states must be estimated using a training sequence. To illustrate this point, in our final example we compare the performance of the adaptive versions of the linear MMSE-DFE and the OH. We use a different channel: $H(z) = 0.35 + 0.8z^{-1} + z^{-2} + 0.8z^{-3}$, and the structure of the DFE is $m = 4$, $n = 3$ and $d = 3$. The channel states and the noise variance are estimated using a training sequence of 150 symbols. For this particular channel, due to the location of the channel states, the performance of the OH-DFE is not critical with respect to $1/C$. Fig. 8 shows the values obtained for the OH-DFE with C_{opt}: its performance is clearly better than the MMSE-DFE.

6. Conclusions

Taking into account that equalization is in fact a classification problem, we have applied the structural risk minimization (SRM) principle to obtain
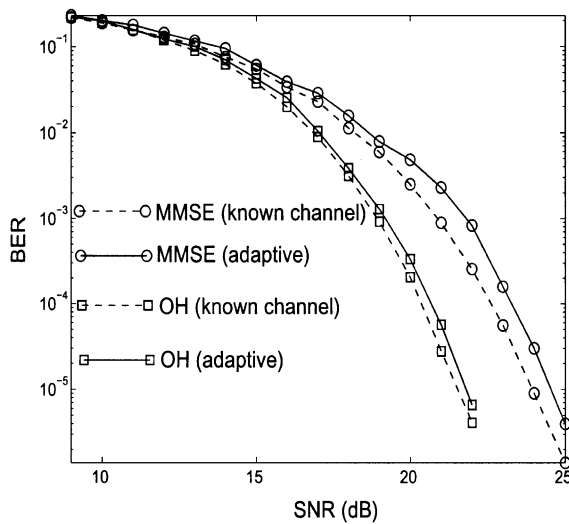
Fig. 8. MMSE versus OH for channel $H(z) = 0.35 + 0.8z^{-1} + z^{-2} + 0.8z^{-3}$ using a DFE with $m = 4$, $n = 3$ and $d = 3$. (Dashed lines) Estimated channel states using 150 training symbols. (Solid lines) Noiseless channel states.

maximum margin classifiers for transversal as well as decision feedback equalizers. Instead of quadratic programming, we train the optimal hyperplane (OH) and support vector machine (SVM) equalizers using an iterative algorithm called the Adatron. We have shown that, for some channels the OH solution can achieve a large improvement over the minimum mean-square error equalizers. When we train the equalizer by classifying the channel states it is important to include somehow in the training process information about the noise variance. For linear maximum margin equalizers, we have shown that this information can be easily taken into account by considering a (regularized) soft margin OH. On the other hand, in comparison with the optimal Bayesian equalizers (implemented through an radial basis function network), the SVM uses only those channel states that are closest to the maximum margin hyperplane in the feature space, and then it yields a reduced structure with only a slight degradation in performance.

The extension of SRM-based equalizers to multilevel modulation (thus yielding a multiclass classification problem [26,3]) is an interesting line for further research. Another avenue for research aims at finding a true sample-by-sample (on-line) training of SVM-based equalizers: some ideas in this direction have been recently proposed in [21,16].

Acknowledgements

References

[1] J.K. Anlauf, M. Biehl, The AdaTron: an adaptive perceptron algorithm, Europhys. Lett. 10 (1989) 687–692.

[2] S. Chen, S. Gunn, C.J. Harris, Decision feedback equaliser design using support vector machines, IEE Proc. Visual Image Signal Process. 147 (3) (June 2000) 213–219.

[3] S. Chen, L. Hanzo, B. Mulgrew, Decision-feedback equalization using multiple-hyperplane partitioning for detecting ISI-corrupted M-ary PAM signals, IEEE Trans. Commun. 49 (5) (May 2001) 760–764.

[4] S. Chen, C.J. Harris, Design of the optimal separating hyperplane for the decision feedback equalizer using Support Vector Machines, in: Proceedings of the ICASSP 2000, Istanbul, Turkey, June, 2000.

[5] S. Chen, B. Mulgrew, E.S. Chng, G.J. Gibson, Space-translation properties and the minimum-BER linear-combiner DFE, IEE Proc. Commun. 145 (5) (October 1998) 316–322.

[6] S. Chen, B. Mulgrew, P.M. Grant, A clustering technique for digital communications channel equalization using radial basis function networks, IEEE Trans. Neural Networks 4 (4) (July 1993) 570–578.

[7] S. Chen, B. Mulgrew, L. Hanzo, Asymptotic bayesian decision feedback equalizer using a set of hyperplanes, IEEE Trans. Signal Process. 48 (12) (December 2000) 3493–3500.

[8] S. Chen, B. Mulgrew, L. Hanzo, Adaptive least error rate algorithm for neural network classifiers, in: Neural Networks for Signal Processing XI—Proceedings of the 2001 IEEE Workshop, Falmouth, MA, USA, September 2001, pp. 223–232.

[9] S. Chen, B. Mulgrew, S. McLaughlin, Adaptive Bayesian equalizer with decision feedback, IEEE Trans. Signal Process. 41 (9) (September 1993) 2918–2926.

[10] J. Cid-Sueiro, A. Artés-Rodríguez, A.R. Figueiras-Vidal, Recurrent radial basis function networks for optimal symbol-by-symbol equalization, Signal Process. 40 (1994) 53–63.

[11] C. Cortes, V. Vapnik, Support-Vector networks, Mach. Learning 20 (1995) 273–297.

[12] T. Frieß, N. Cristianini, C. Campbell, The kernel-Adatron algorithm: a fast and simple learning procedure for support vector machines, in: J. Shavlik (Ed.), Machine Learning:

Proceedings of the 15th International Conference, Morgan Kauffman Publishers, San Francisco, 1998.

[13] G.J. Gibson, S. Siu, C.F.N. Cowan, The application of nonlinear structures to the reconstruction of binary signals, IEEE Trans. Signal Process. 39 (8) (August 1991) 1877–1884.

[14] J. Lee, C. Beach, N. Tependelenlioglu, A practical radial basis function equalizer, IEEE Trans. Neural Networks 10 (2) (March 1999) 450–455.

[15] A. Lyhyaoui, M. Martinez, I. Mora, M. Vazquez, J.L. Sancho, A.R. Figueiras-Vidal, Sample selection via clustering to construct support vector-like classifiers, IEEE Trans. Neural Networks 10 (6) (November 1999) 1474–1481.

[16] M. Martínez, J.L. Sancho, C. Bousoño, A.R. Figueiras-Vidal, Channel equalization via sample selection, Proceedings of the International COST 254 Workshop Intell. Commun. Technol. Appl. Emphasis Mobile Commun., Neuchatel, Switzerland, May 1999.

[17] B. Mitchinson, R.F. Harrison, Digital communications channel equalization using the kernel Adaline, IEEE Trans. Commun. 50 (4) (April 2002) 571–576.

[18] B. Mulgrew, S. Chen, Stochastic gradient minimum-BER decision feedback equalisers, in: Proc. IEEE Symposium on Adaptive Systems for Signal Processing, Communications and Control, Lake Louise, AB, Canada, October 2000, pp. 93–98.

[19] B. Mulgrew, C.F.N. Cowan, Adaptive Filters and Equalizers, Kluwer Academic Publishers, Boston, 1988.

[20] E. Osuna, R. Freund, F. Girosi, An improved training algorithm for support vector machines, in: J. Principe, L. Gile, N. Morgan, E. Wilson (Eds.), Neural Networks for Signal Processing VII—Proceedings of the 1995 IEEE Workshop, New York, 1997, pp. 276–285.

[21] F. Pérez-Cruz, A. Artés-Rodríguez, Adaptive SVC for nonlinear channel equalization, in: Proceedings of the EUSIPCO 2002, Vol. II, Toulouse, France, Sept: 2002, pp. 45–48.

[22] F. Pérez-Cruz, A. Navia-Vázquez, P.L. Alarcón-Diana, A. Artés-Rodríguez, Support vector classifier with hyperbolic tangent penalty function, Proceedings of the ICASSP 2000, Istanbul, Turkey, June, 2000.

[23] F. Pérez-Cruz, A. Navia-Vázquez, P.L. Alarcón-Diana, A. Artés-Rodríguez, SVC-based equalizer for burst TDMA transmissions, Signal Process. 81 (2001) 1681–1693.

[24] S.U.H. Qureshi, Adaptive equalization, Proc. IEEE 73 (1985) 1349–1387.

[25] D.J. Sebald, J.A. Bucklew, Support vector machine techniques for nonlinear equalization, IEEE Trans. Signal Process. 48 (11) (November 2000) 3217–3226.

[26] D.J. Sebald, J.A. Bucklew, Support vector machines and the multiple hypotheses test problem, IEEE Trans. Signal Process. 49 (11) (November 2001) 2865–2872.

[27] V. Vapnik, The Nature of Statistical Learning Theory, 1995, Springer, New York.

[28] C.C. Yeh, J.R. Barry, Approximate minimum bit-error rate equalization for binary signals, in: Proceedings of the International Conference on Communications, Vol. 2, Montreal, Canada, 1997, pp. 1095–1099.