COMPARATIVE ANALYSIS OF IMPORTANCE SAMPLING TECHNIQUES TO ESTIMATE ERROR FUNCTIONS FOR TRAINING NEURAL NETWORKS

Manuel Rosa-Zurera [1]*, Pilar Jarabo-Amores*[1]*, Francisco López-Ferreras* [1]*, José L. Sanz-González*[2]

(1) Departamento de Teoría de la Señal y Comunicaciones,
Escuela Politécnica Superior, Universidad de Alcalá
Campus Universitario, 28805, Alcalá de Henares - Madrid (SPAIN).
email: {manuel.rosa;mpilar.jarabo;francisco.lopez}@uah.es

(2) Departamento de Señales, Sistemas y Radiocomunicaciones,
ETSI de Telecomunicación, Universidad Politécnica de Madrid
Ciudad Universitaria, 28040 - Madrid (SPAIN)
jlsanz@gcs.ssr.upm.es

ABSTRACT

The application of importance sampling to train neural networks which approximates the Neyman-Pearson detector is considered in this paper. A comparative study with two different error functions is carried out. These two error functions are selected to make the Neyman-Pearson detector approximation possible. The importance sampling technique is used to estimate the error function for training. Some results are presented to compare the performance of both approaches to approximate the optimum detector. Furthermore, results show the convenience of using the importance sampling technique for training neural networks, when low probabilities of false alarm are considered.

1. INTRODUCTION

The objective of this paper is to study the application of Importance Sampling (IS) techniques to train neural networks for approximating the Neyman-Pearson detector. This detector maximizes the probability of detection (P_D), while maintaining the probability of false alarm (P_{FA}) lower than or equal to a specified value. The characteristics of such a detector are reflected in its ROC (Receiver Operating Characteristic) curve, that relates P_D to P_{FA} [1].

Neural networks can be used to approximate the optimum Neyman-Pearson detector. Ruck et al. [2], and Wan [3], demonstrated that neural networks training converges on a discriminant function that can be used to implement the minimum probability of error classifier, when the mean squared error is considered as error function. This discriminant function can also be used to implement the Neyman-Pearson detector when the detection threshold is selected for a given P_{FA} value.

Previous works which considered the use of neural networks to approximate the Neyman-Pearson detector have highlighted the poor performance of such detectors for low P_{FA} values [4, 5]. This poor behavior can be explained from the point of view of training. Rare events hardly influence the estimation of the error function for training. So, the approximation of the boundary of the decision regions is not good where rare events occur. Importance sampling techniques can be used to estimate the error function for training.

The application of IS techniques in the training phase of neural network detectors using the Mean-Square (MS) error criterion is explained in [6]. The authors proposed an adaptive importance sampling technique in order to improve MS error estimations in each iteration of the training, by finding a suboptimal probability density function for sampling. In this paper, we extend these results to another error criterion, the *cross entropy error* [7]. A comparative study is carried out for both error functions, highlighting the advantages and drawbacks of each one.

The paper is organized as follows. In Section 2, we explain the fundamentals of the importance sampling techniques. In Section 3, we consider the problem of training neural networks with the considered error criteria for implementing neural detectors which approximate the Neyman-Pearson detector. In Section 4 we present some results which shows the convenience of this approach. Finally, some conclusions are presented for summarizing the results presented in the paper.

This work has been supported by the "Consejería de Educación de la Comunidad de Madrid" (SPAIN), under Project 07T/0036/2003 1

2. IMPORTANCE SAMPLING TECHNIQUE

In a binary detection problem we wish to choose between two hypotheses: H_0 and H_1. Suppose $\mathbf{z}$ is the received vector (N-dimensional), and define $D_0 \subseteq \mathbb{R}^N$ and $D_1 \subseteq \mathbb{R}^N$ as the regions where H_0 and H_1 are chosen, respectively. The average probability of error can be expressed as:

$$P_e = P(H_0) \int_{D_1} f(\mathbf{z}|H_0) d\mathbf{z} + P(H_1) \int_{D_0} f(\mathbf{z}|H_1) d\mathbf{z} \tag{1}$$

where $P(H_0)$ and $P(H_1)$ are the prior probabilities of the hypothesis. The integrals in (1) can be simplified as:

$$P_i = \int_{\mathbb{R}^N} I_j(\mathbf{z}) f(\mathbf{z}|H_i) d\mathbf{z}, \quad j \neq i = 0, 1 \tag{2}$$

$I_j(\mathbf{z})$ is the indicator function over Z_j, i.e., $I_j(\mathbf{z}) = 1, \forall \mathbf{z} \in Z_j$, and $I_j(\mathbf{z}) = 0$, otherwise. The maximum likelihood estimator for these integrals is:

$$\hat{P}_i = \frac{1}{M} \sum_{m=1}^{M} I_j(\mathbf{z}_m), \quad j \neq i = 0, 1 \tag{3}$$

where $\mathbf{z}_m$ are independent and identically distributed random variables with probability density function $f(\mathbf{z}|H_i)$, $i = 0, 1$. Note that $\hat{P}_i$ is an unbiased and consistent estimator with variance:

$$var(\hat{P}_i) = \frac{P_i(1 - P_i)}{M} \tag{4}$$

This expression shows that the variance for small P_i is approximately P_i/M. The importance sampling method reduces the variance of the estimate of P_i by generating data from a biased density function $f^*(\mathbf{z}|H_i)$. This density function is chosen such that the observation vector is more likely to come from the important region over which the integral is calculated. Since more errors than expected are generated, the error must be weighted to obtain an unbiased estimate. The resulting estimator for P_i from importance sampling is given as:

$$P_i^* = \frac{1}{M} \sum_{m=1}^{M} W_i(\mathbf{z}_m) I_j(\mathbf{z}_m), \quad j \neq i = 0, 1 \tag{5}$$

If $\mathbf{z}_m$ are independent and identically distributed samples obtained from $f^*(\mathbf{z}|H_i), i = 0, 1$, the estimate of P_i is unbiased when [8]:

$$W_i(\mathbf{z}) = \frac{f(\mathbf{z}|H_i)}{f^*(\mathbf{z}|H_i)}, \forall \mathbf{z} \in Z_j \subset \mathbb{R}^N \quad j \neq i = 0, 1 \tag{6}$$

The variance of the importance sampling estimate is:

$$var(P_i^*) = \frac{E^*[W_i(\mathbf{z}) I_j(\mathbf{z})] - P_i^2}{M} \tag{7}$$

Finally, note that the estimate of P_e is $\hat{P}_e = P(H_0)P_0^* + P(H_1)P_1^*$, being an alternative way of the approach given in [6].

3. IMPORTANCE SAMPLING TECHNIQUE TO ESTIMATE THE ERROR FOR TRAINING NEURAL DETECTORS

D. W. Ruck et al. [2] demonstrated that a multilayer perceptron (MLP) converges to a mean squared-error approximation of the Bayes optimal discriminant function, when trained using the mean squared-error criterion. They study the two-class and multiclass problems, and extended this result to any mean squared-error minimization technique.

For binary detection, they studied a MLP with only one neuron in the output layer. The network was trained to produce 1 when the feature vector was from class H_1 and -1 when the vector was from class H_0. They proved that the neural network output approximates the Bayes optimal discriminant function $g_0(\mathbf{z})$, given in (8) where $\mathbf{z}$ is the feature vector, and $P(H_1|\mathbf{z})$ and $P(H_0|\mathbf{z})$ are the posterior probabilities of the classes.

$$g_0(\mathbf{z}) = P(H_1|\mathbf{z}) - P(H_0|\mathbf{z}) \tag{8}$$

The mean squared-error between the network output, $F(\mathbf{z})$, and the desired outputs, is given by (9). E_s is the sample mean error calculated for a set of n pre-classified feature vectors, and Z_1 and Z_0 are the sets of all possible feature vectors for class H_1 and H_0, respectively ($Z_0 \cup Z_1 = \mathbb{R}^N$, $Z_0 \cap Z_1 = \varnothing$, $\mathbb{R}^N$ being the input space).

$$E_m = \lim_{n\to\infty} \frac{E_s}{n} = \lim_{n\to\infty} \frac{1}{n} [\sum_{\mathbf{z}\in Z_1} (F(\mathbf{z})-1)^2 + \sum_{\mathbf{z}\in Z_0} (F(\mathbf{z})+1)^2] \tag{9}$$

Using the Strong Law of Large Numbers, applying the Bayes formula and rearranging terms, (9) converts into (10).

$$E_m = \int_{\mathbb{R}^N} (F(\mathbf{z}) - g_0(\mathbf{z}))^2 f(\mathbf{z}) d\mathbf{z} + \big[1 - \int_{\mathbb{R}^N} g_0^2(\mathbf{z}) f(\mathbf{z}) d\mathbf{z}\big] \tag{10}$$

If the training set represents a reasonable approximation to the input space, although the network is trained for minimizing E_s, E_m will be minimized. Since the term in square brackets in expression (10) is independent of the network architecture or parameters, minimizing E_m is equivalent to minimizing expression (11). So, the network output is an

approximation of the Bayes optimal discriminant function in the mean squared-error sense.

$$E = \int_{\mathbb{R}^N} (F(\mathbf{z}) - g_0(\mathbf{z}))^2 f(\mathbf{z}) d\mathbf{z} \tag{11}$$

This expression represents the weighted squared error between $F(\mathbf{z})$ (the neural network function), and $g_0(\mathbf{z})$, the optimum Bayessian discriminant function, which can implement the Neyman-Pearson detector. The weighting function is the probability density function of the patterns which constitute the input to the detector. In actual situations, the training set does not represent $f(\mathbf{z})$ in the overall input space. For example, it is not likely that the training set contains high energy noise patters which give rise to outputs greater than the threshold which implements the Neyman-Pearson detector for low P_{FA} values. So, E_s is not a good approximation to E_m in the region which gives rise to low P_{FA} values. To overcome this drawback, importance sampling can be used. E_s/n is the maximum likelihood estimate of E_m. Importance sampling techniques can be applied to evaluate this expression [6]:

$$E_s^* = \frac{1}{n}[\sum_{\mathbf{z}\in Z_1} (F(\mathbf{z})-1)^2 W(\mathbf{z}) + \sum_{\mathbf{z}\in Z_0} (F(\mathbf{z})+1)^2 W(\mathbf{z})] \tag{12}$$

$W(\mathbf{z}) = \frac{f(\mathbf{z})}{f^*(\mathbf{z})}$ is the importance sampling weighting function. If $f(\mathbf{z})$ depends on a parameter vector θ, $f^*(\mathbf{z})$ can be obtained with (13) for a given parameter vector θ:

$$f^*(\mathbf{z}) = P(H_0) f(\mathbf{z}; \theta_\mathbf{0}|H_0) + P(H_1) f(\mathbf{z}; \theta_\mathbf{1}|H_1) \tag{13}$$

Importance sampling techniques can also be applied to other error functions, for example, the cross entropy error criterion. For the cross entropy error criterion to be applied the neural network output must be in the interval $(0, 1)$.

For desired outputs 1 for input vectors from hypothesis H_1 and 0 for input vectors from hypothesis H_0, the function to be minimized during training is expressed in (14).

$$\frac{E_s}{n} = -\frac{1}{n}[\sum_{\mathbf{z}\in Z_1} log[F(\mathbf{z})] + \sum_{\mathbf{z}\in Z_0} log[1-F(\mathbf{z})]] \tag{14}$$

When the training set becomes dense, expression (14) comes to (15) if the Strong Law of Large Numbers is applied:

$$E_m = -P(H_1) \cdot \int_{\mathbb{R}^N} f(\mathbf{z}|H_1) \cdot log(F(\mathbf{z})) d\mathbf{z} - P(H_0) \cdot \int_{\mathbb{R}^N} f(\mathbf{z}|H_0) \cdot log(1-F(\mathbf{z})) d\mathbf{z} \tag{15}$$

Taking into consideration that $F(\mathbf{z}) \in (0, 1)$, minimizing E_m is equivalent to minimize the following expression:

$$-P(H_1) \cdot f(\mathbf{z}|H_1) \cdot log[F(\mathbf{z})] - P(H_0) \cdot f(\mathbf{z}|H_0) \cdot log[1-F(\mathbf{z})] \tag{16}$$

Expression (15) can be written as (17):

$$E_m = -\int_{\mathbb{R}^N} \Big\{ P(H_1|\mathbf{z}) \cdot log(F(\mathbf{z})) + (1-P(H_1|\mathbf{z})) \cdot log(1-F(\mathbf{z})) \Big\} f(\mathbf{z}) d\mathbf{z} \tag{17}$$

Again, $f(\mathbf{z})$ works as a weighting function, which indicates that low probability events hardly influence on training. In order to improve training for these low probability events, importance sampling can be included to estimate E_m during training. An estimate of E_m is presented in expression (18):

$$E = -[\sum_{\mathbf{z}\in Z_1} log[F(\mathbf{z})] W_1(\mathbf{z}) + \sum_{\mathbf{z}\in Z_0} log[1-F(\mathbf{z})] W_0(\mathbf{z})] \tag{18}$$

In this paper, we propose the following importance sampling weighting functions:

$$W_1(\mathbf{z}) = \frac{f(\mathbf{z}|H_1)}{f^*(\mathbf{z}|H_1)}, \forall \mathbf{z} \in \mathbb{R}^N \tag{19}$$

$$W_0(\mathbf{z}) = \frac{f(\mathbf{z}|H_0)}{f^*(\mathbf{z}|H_0)}, \forall \mathbf{z} \in \mathbb{R}^N \tag{20}$$

Being $f^*(\mathbf{z}|H_1)$ and $f^*(\mathbf{z}|H_0)$ the suboptimal importance sampling densities, that are usually obtained from the original $f(\mathbf{z}|H_i), i = 0, 1$ varying one of the parameters the likelihood functions depends on.

The objective of this paper is to compare neural detectors obtained from applying importance sampling techniques to estimate the mean squared error or the cross entropy error to train neural networks. Both error criteria allow to approximate the Neyman-Pearson optimum detector.

4. RESULTS

Results comparing neural networks which approximate the Neyman-Pearson detector for both error criteria, the mean squared error and the cross entropy error are presented in this section. The standard error-back-propagation algorithm has been used for training all the neural networks.

The problem of detecting Gaussian signals in Gaussian interference is considered. This problem arise, for example, when detecting Swerling I radar targets in white Gaussian noise. In this section, the ROC curves of the neural detectors

trained with these two criteria, and with and without the use of importance sampling techniques, are represented.

Figures 1 and 2 represent the ROC curves of the neural detectors trained with the mean squared error criterion, with and without the use of importance sampling, respectively. Figure 1 shows the results obtained with MLPs of eight neurons in the hidden layer and sixteen inputs. The neural networks are trained with different values of the Training Signal to Noise Ratio (TSNR), ranging from 0dB to 15dB. The SNR value for testing is 3dB. Figure 2 shows the results obtained with the same MLP structure, using importance sampling to estimate the mean squared error. Black lines correspond to the ideal Neyman-Pearson detector, color lines correspond to neural detectors trained with different signal to noise ratios.

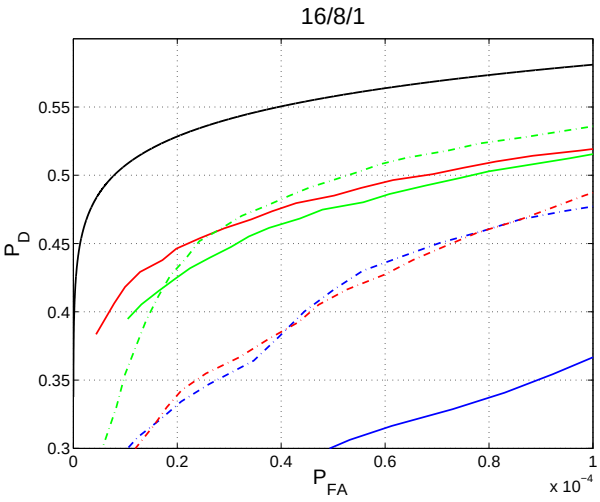

Fig. 1. ROC curves obtained with the mean squared error criterion without using importance sampling for training.

On the other hand, results that compare neural networks trained with and without importance sampling techniques to minimize the cross entropy error function are presented in figures 3 and 4, which represent the ROC curves of the neural detectors trained with and without the use of importance sampling, respectively. Again, neural networks with eight neurons in the hidden layer and sixteen inputs are considered. The neural networks are also trained with different Training Signal-to-Noise Ratio (TSNR) values, ranging from 0dB to 15dB, and the SNR value for testing is 3dB.

Results show that the use of importance sampling techniques for training highly improves neural network performance, using both the mean squared error and the cross entropy criteria. Besides, it can be observed that the cross entropy criteria is more suitable for implementing neural detectors which approximate the Neyman-Pearson detector.

Without using importance sampling techniques, the performance of neural networks trained using the cross entropy

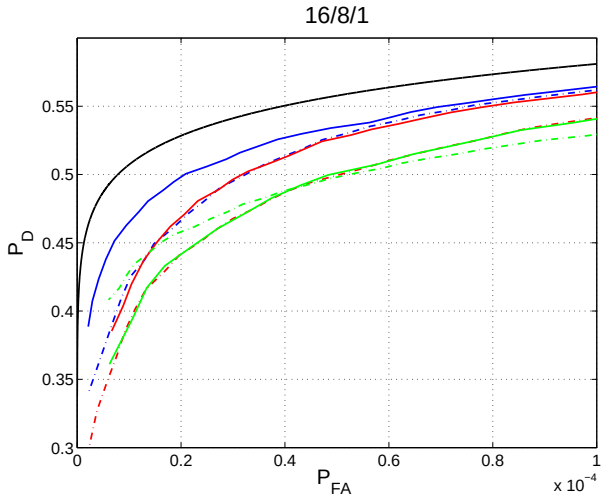

Fig. 2. ROC curves obtained with the mean squared error criterion using importance sampling for training.

error is clearly better than the performance of those trained using the mean squared error. Although for the cross entropy error there is a reduction of the dependence on TSNR, the probability of detection for almost all the considered TSNRs is significantly lower than the Neyman-Pearson detector one.

When considering the mean squared error, the performance of the neural detectors improves when importance sampling techniques are used. The results are only slightly better than those obtained with the cross entropy error without using importance sampling.

Finally, the best results are obtained using the cross entropy error in combination with importance sampling techniques.

5. CONCLUSIONS

In this paper, the application of importance sampling techniques to train neural networks which approximates the optimum Neyman-Pearson detector is considered. Two error functions are used throughout the paper, the mean squared error and the cross entropy error. These two error functions are selected to make the Neyman-Pearson detector approximation possible.

The results presented in the paper show that the cross entropy criterion is more suitable than the mean squared error criterion to approximate the Neyman-Pearson detector. Furthermore, the application of importance sampling techniques to estimate the error functions highly improves the performance of the implemented neural detectors.

The use of importance sampling techniques for training neural networks in order to approximate the Neyman-Pearson detector using the mean squared error gives rise

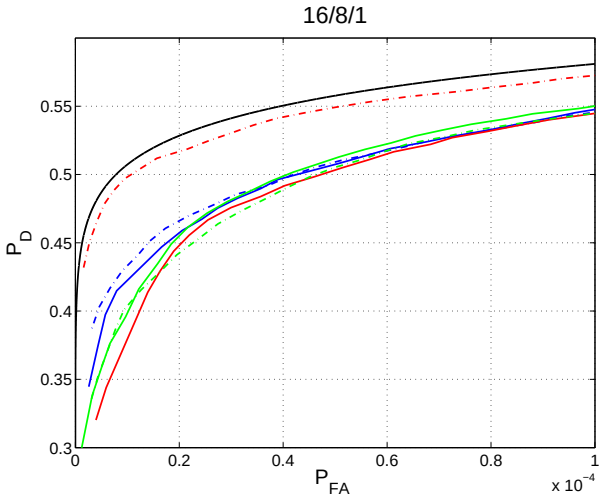

Fig. 3. ROC curves obtained with the cross entropy criterion without using importance sampling for training.

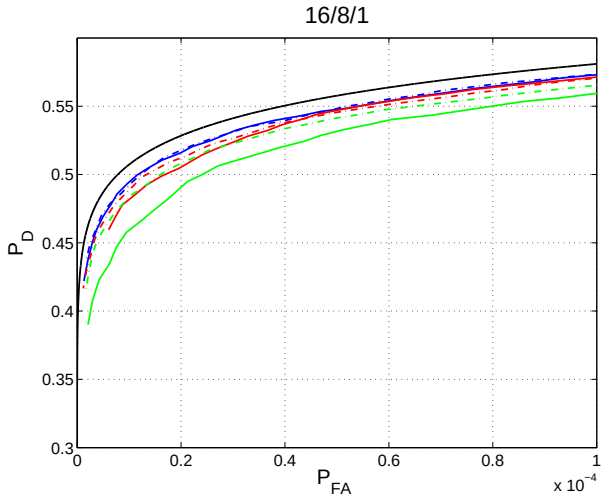

Fig. 4. ROC curves obtained with the cross entropy criterion using importance sampling for training.

to results similar to those obtained with the neural netwok-based detectors trained with the cross entropy error without using importance sampling.

The best results are obtained using the cross entropy error and importance sampling. The neural networks trained in such a way can implement a very good approximation of the Neyman-Pearson detector.

6. REFERENCES

[1] H.L. Van Tress, *Detection, Estimation, and Modulation Theory*, vol. 1, Wiley, 1968.

[2] D.W. Ruck, S.K. Rogers, M. Kabrisky, M.E. Oxley, and B.W. Suter, "The multilayer perceptron as an aproximation to a bayes optimal discriminant function," *IEEE Transactions on Neural Networks*, vol. 1, no. 4, pp. 296–298, December 1990.

[3] E.A. Wan, "Neural network classification: a bayesian interpretation," *IEEE Transactions on Neural Networks*, vol. 1, no. 4, pp. 303–305, Diciembre 1990.

[4] P.P. Gandhi and V. Ramamurti, "Neural networks for signal detection in non-gaussian noise," *IEEE Transactions on signal processing*, vol. 45, no. 11, pp. 2846–2851, Noviembre 1997.

[5] D. Andina and J.L. Sanz-Gonzlez, "Comparison of a neural network detector vs. neyman-pearson optimal detector," *Proceedings of the 1996 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6, pp. 3573–3576, 1995.

[6] J.L. Sanz-González, D. Andina, and J. Seijas, "Importance sampling and mean-square error in neural detector training," *Neural Processing Letters*, vol. 16, no. 6, pp. 256–276, December 2002.

[7] C. M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, Gran Bretaa, 1 edition, 1995.

[8] G. C. Orsak, "A note on estimating false alarm rates via importance sampling," *IEEE Transactions on Communications*, vol. 41, no. 9, pp. 1275–1277, Septiembre 1993.