Proceedings of the VI Brazilian Conference on Neural Networks - VI Congresso Brasileiro de Redes Neurais
pp. 1–6, June 2–5, 2003 - Centro Universitario da FEI, Sao Paulo - SP - Brazil

A Wavelet Network Classifier Applied to Financial Distress Prediction

Roberto Kawakami H. Galvão[1], Victor M. Becerra[2], Magda Abou-Seada[3]
[1]Instituto Tecnológico de Aeronáutica, Div. Engenharia Eletrônica
São José dos Campos – SP, 12228–900
[2]University of Reading, Department of Cybernetics
Reading RG6 6AY, United Kingdom
[3]Middlesex University, Business School
London NW4 4BT, United Kingdom
E-mails: kawakami@ele.ita.br, v.m.becerra@reading.ac.uk, M.Abou-Seada@mdx.ac.uk

Abstract

This work proposes a constructive method for building a wavelet network classifier. The network consists of a wavelet layer, which implements a nonlinear transformation in the input data, and a linear discriminant function, which carries out classification on the basis of the wavelet layer output. The proposed methodology is tested in a financial distress prediction problem involving British companies in the period 1997–2000. In this case study, the wavelet network was found to be a better classifier than a model obtained by linear discriminant analysis and a multi–layer perceptron trained with the Optimal Brain Damage technique.

1. Introduction

A wavelet network is a nonlinear structure that implements input-output mappings as the superposition of dilated and translated versions of a single function, which is localized both in the space and frequency domains [1]. Such structure can approximate any square-integrable function to arbitrary precision, given a sufficiently large number of network elements. This property has been exploited for function approximation and dynamic system identification in a number of works [2],[3],[4],[5], which claim that wavelet networks can be more easily designed and tuned than multi-layer perceptrons and networks of radial basis functions. In fact, efficient construction algorithms have been devised to optimize the structure of the wavelet network [3],[6], and even to adapt it in real time [2],[5].

In the context of classification problems, linear wavelet transformations have been employed to extract features from signals and images as a pre-processing stage for neural network classifiers [7], [8], [9]. However, the use of wavelet networks to implement the nonlinear decision surface of the classifier itself is still incipient.

This paper proposes a constructive algorithm to build a wavelet network classifier. The classifier consists of two blocks: (1) a wavelet layer, which implements a nonlinear transformation in the input data, and (2) a linear discriminant function, which is applied to the output of the wavelet layer. This approach is novel in comparison with usual neural-wavelet classification methodologies [7], [8], in which the wavelet layer is used for a preliminary linear transformation and the classification is carried out by a nonlinear neural network structure.

For illustration, the proposed methodology is applied to the financial distress classification of British companies on the basis of data from 1997–2000. For comparison, a conventional linear discriminant model [10] and a multi–layer perceptron trained with the Optimal Brain Damage algorithm [11] are also employed.

The text is organized as follows. A brief review of wavelet networks is given in section 2. Section 3 presents the proposed method to build a wavelet network classifier. Section 4 describes the financial application example. The classification results are discussed in section 5. Concluding remarks and suggestions for further research are given in section 6.

2. Wavelet Networks

A wavelet network [1], [3] can be regarded as a neural architecture whose activation functions are dilated and translated versions of a single function $\psi \in L^2(\mathbb{R}^d)$, where d is the input dimension[1]. This function, called "mother wavelet", is localized both in the space $(\mathbf{x})$ and frequency $(\boldsymbol{\omega})$ domains in the sense that $|\psi(\mathbf{x})|$ and $|\hat{\psi}(\boldsymbol{\omega})|$ rapidly decay to zero when $\|\mathbf{x}\| \to \infty$ and $\|\boldsymbol{\omega}\| \to \infty$, respectively[2].

As an example, consider the so-called "Mexican Hat" mother wavelet defined as

$$\psi(\mathbf{x}) = (1 - \|\mathbf{x}\|^2) \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right) \qquad (1)$$

which is depicted in Figure 1a for a two-dimensional input. For convenience of visualization, a cross-section of this wavelet is also presented in Figure 1b. This type of multidimensional wavelet is called "radial", because it only depends on the norm of the input vector.

[1] $L^2(\mathbb{R}^d)$ is the set of square-integrable functions $f : \mathbb{R}^d \to \mathbb{R}$
[2] $\hat{\psi}$ is the Fourier Transform of ψ.

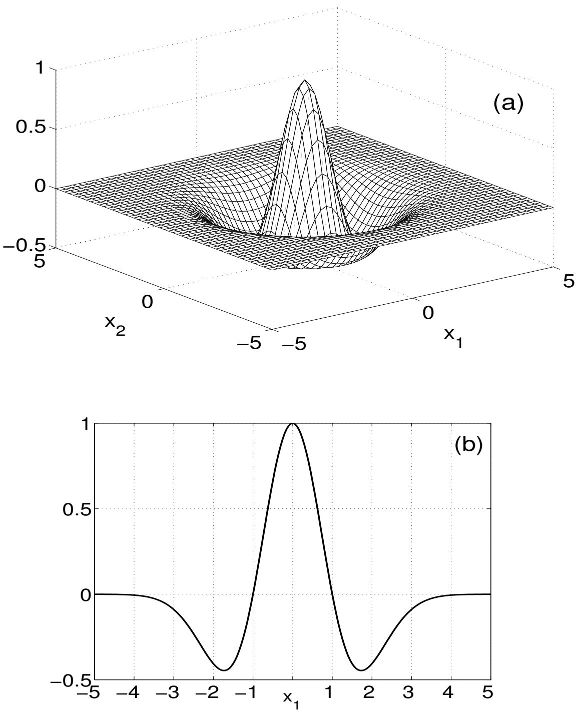

Figure 1: (a) Mexican Hat wavelet with two-dimensional input. (b) Cross-section of the Mexican Hat wavelet.

A wavelet network with one output, d inputs and L nodes can be parameterized as follows [1],[3]:

$$y = \sum_{j=1}^{L} w_j \psi_j(\mathbf{x}) \qquad (2)$$

where $\mathbf{x} = [x_1 \, x_2 \, \cdots \, x_d]^T$ is the vector of inputs. Functions ψ_j, called wavelets, are dilated and translated versions of the mother wavelet function ψ:

$$\psi_j(\mathbf{x}) = a_j^{-d/2} \psi\left(\frac{\mathbf{x} - \mathbf{b}_j}{a_j}\right) \qquad (3)$$

In Equation (3), the dilation parameter $a_j \in \mathbb{R}^*$ controls the spread of the wavelet, whereas the translation parameter $\mathbf{b}_j \in \mathbb{R}^d$ determines its central position. It can be shown that, if pairs $(a_j, \mathbf{b}_j)$ are taken from a grid Λ given by

$$\Lambda = \{(\alpha^m, \mathbf{n}\beta\alpha^m); \; m \in \mathbb{Z}, \mathbf{n} \in \mathbb{Z}^d\} \qquad (4)$$

for convenient values of $\alpha > 1$ and $\beta > 0$, then any function in $L^2(\mathbb{R}^d)$ can be approximated by Equation (2) to arbitrary precision, given a sufficiently large number of wavelets [3],[12]. In this sense, wavelet networks can implement arbitrary nonlinear discriminant functions.

A major advantage of wavelet networks over other neural architectures is the availability of efficient constructive algorithms [2], [3], [5], [6] for defining the network structure, that is, for choosing convenient values for $(m, \mathbf{n})$. Once the structure has been determined, weights w_j can be obtained by linear estimation methods.

3. The proposed constructive method

In this work, a modified version of the constructive method introduced by [3] is proposed as follows. Suppose that M modelling samples of known classification are available in the form of data vectors $\mathbf{x}^k \in \mathbb{R}^d$, $k = 1, \ldots, M$. Then:

1. Normalize the modelling data to fit within the effective support S of the mother wavelet employed. For radial wavelets, S is a hypersphere in $\mathbb{R}^d$ with radius R. For the Mexican Hat, for instance, R can be taken as 5 (see Figure 1b). For computational simplicity, S is approximated as a hypercube inscribed in the hypersphere with edges parallel to the coordinate axis.

2. For each sample $\mathbf{x}^k$ in the modelling set, find I^k, the index set of wavelets whose effective supports contain $\mathbf{x}^k$:

   $$I^k = \{(m, \mathbf{n}) \text{ s.t. } \mathbf{x}^k \in S_{m,\mathbf{n}}, \\ m_{min} \leq m \leq m_{max}\}, \; k = 1, \ldots, M \quad (5)$$

   where $S_{m,\mathbf{n}}$ is a hypercube centered in $\mathbf{n}\beta\alpha^m$ with edges $\alpha^m R\sqrt{2}$. As a rule of thumb, set the minimum and maximum scale levels to $m_{min} = -1$ and $m_{max} = +1$, respectively. In the present application, these settings proved to be adequate but, in the general case, more scale levels may be added until the performance of the wavelet network is considered satisfactory.

3. Determine the pairs $(m, \mathbf{n})$ which appear in at least two sets I^{k_1} and I^{k_2}, $k_1 \neq k_2$. These are the wavelets whose effective support include at least two samples. This step is different from the algorithm described in [3], which allows for wavelets with effective supports containing only one sample. Since such wavelets introduce oscillations between neighbor data points, they should be excluded from the modelling process to avoid overfitting problems.

4. Let L be the number of wavelets obtained above. For simplicity of notation, replace the double index $(m, \mathbf{n})$ by a single index $j = 1, \ldots, L$.

5. Apply the L wavelets to the M modelling samples and gather the results in matrix form as

   $$\mathbf{\Psi} = \begin{bmatrix} \psi_1(\mathbf{x}^1) & \psi_1(\mathbf{x}^2) & \cdots & \psi_1(\mathbf{x}^M) \\ \psi_2(\mathbf{x}^1) & \psi_2(\mathbf{x}^2) & \cdots & \psi_2(\mathbf{x}^M) \\ \vdots & \vdots & \cdots & \vdots \\ \psi_L(\mathbf{x}^1) & \psi_L(\mathbf{x}^2) & \cdots & \psi_L(\mathbf{x}^M) \end{bmatrix} \quad (6)$$

Notice that each sample is now represented by L wavelet outputs (a column of $\mathbf{\Psi}$), instead of d variables. Since the mapping $\mathbf{x} \longmapsto \mathbf{\Psi}(\mathbf{x})$ is a nonlinear transformation, patterns which were not linearly separable in the x-variable domain may be so in the domain of wavelet outputs. However, many wavelets resulting from steps 1-4 may be redundant or may not convey useful discriminating information. Thus, the next step consists in determining which wavelets or, alternatively, which rows of $\boldsymbol{\Psi}$ are the most relevant for the classification task.

At this point, the algorithm of [3], which was proposed in an estimation framework, makes use of the information available in the dependent variable (the variable to be estimated). However, in the present case, which is aimed at classification, a different approach must be used. For simplicity, a two-class problem will be assumed, without loss of generality.

Before describing the proposed approach for wavelet selection, two definitions are needed.

Definition 1 (Fisher Discriminant Criterion). Let $\boldsymbol{\psi}$ be a vector containing elements of class 1 and of class 2. The Fisher Discriminant Criterion of $\boldsymbol{\psi}$ is defined as

$$F(\boldsymbol{\psi}) = \frac{[\mu_1(\boldsymbol{\psi}) - \mu_2(\boldsymbol{\psi})]^2}{[\sigma_1(\boldsymbol{\psi})]^2 + [\sigma_2(\boldsymbol{\psi})]^2} \tag{7}$$

where $\mu_1(\boldsymbol{\psi})$ and $\sigma_1(\boldsymbol{\psi})$ are respectively the mean and standard deviation of the elements of $\boldsymbol{\psi}$ associated to class 1. In the same manner, $\mu_2(\boldsymbol{\psi})$ and $\sigma_2(\boldsymbol{\psi})$ are defined for the elements associated to class 2. The Fisher Discriminant Criterion is a measure of how well the two classes are separated in $\boldsymbol{\psi}$.

Definition 2 (Condition Number associated to a set of vectors). Let $A = \{\boldsymbol{\psi}_j,\ j = 1, \ldots, J\}$ be a set of M-dimensional row vectors ($M > J$). The condition number associated to A is defined as the condition number (ratio between the maximum and the minimum singular values) of a matrix $\boldsymbol{\Delta}$ built as

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\psi}_1 \\ \boldsymbol{\psi}_2 \\ \vdots \\ \boldsymbol{\psi}_J \end{bmatrix} \tag{8}$$

The condition number can be used as a measure of the collinearity (or "redundancy") between the vectors in A [13]. In fact, if the vectors are linearly dependent, the condition number is infinite and, if they form an orthonormal set, the condition number equals one.

The proposed algorithm for wavelet selection chooses rows from $\boldsymbol{\Psi}$ in a stepwise manner, starting from the one which displays the largest Fisher Discriminant Criterion and adding a new row at each iteration.

a) (Initialization) Let A be the set of row vectors still available for selection and B the set in which the selected vectors are stored. Initially, $A = \{\boldsymbol{\psi}_j,\ j = 1, \ldots, L\}$, where $\boldsymbol{\psi}_j = [\psi_j(\mathbf{x}^1)\ \psi_j(\mathbf{x}^2)\ \cdots\ \psi_j(\mathbf{x}^M)]$, and $B = \emptyset$.

b) (Preliminary pruning) Eliminate from A all vectors whose norm is lower than $\kappa \max_j(\|\boldsymbol{\psi}_j\|)$ for a fixed $0 < \kappa < 1$.

c) (First selection) Determine the vector in A which displays the largest Fisher Discriminant Criterion. Move this vector from A to B.

d) (Collinearity prevention) Remove from A all vectors which, if added to B, result in a condition number larger than a fixed threshold $\chi > 0$. If all vectors in A are eliminated then stop.

e) (Selection) For each of the remaining vectors in A, obtain a linear discriminant model (see Appendix) by using this vector in conjunction with the vectors in B. Evaluate the Fisher Discriminant Criterion of the model output. Determine the vector which leads to the largest Fisher Discriminant Criterion and move it from A to B.

f) Return to step (d).

Notice that redundancy is avoided in step (d), whereas step (e) selects the vector that displays the better synergy with those already selected. The selection procedure described above can be regarded as a growing network construction algorithm, since it starts with a single-wavelet network, which is augmented with a new wavelet at each iteration. Parsimony considerations can be used to select an optimal size for the network, as it will be shown in the section of Results and Discussion.

4. Application Example: Financial Distress Prediction

A company is said to be insolvent or under financial distress if it is unable to pay its debts as they become due, which is aggravated if the value of the firm's assets is lower than its liabilities. Once a company has become insolvent there are several courses of action covered by the relevant laws. Not all of these courses of action necessarily mean the end of a company or its business activity. The primary objective is to recover as much of the money owed to creditors as possible.

Distress prediction models are used by a large number of parties, which include lenders, investors, regulatory authorities, government officials, auditors and managers [14]. The development of prediction models for financial distress started in the late 1960's [15], [16] and continues to this day. Financial distress models are usually based on ratios of financial quantities, rather than absolute values. By deflating statistics by size, the use of ratios allows a uniform treatment of different firms.

A number of works have found neural network classifiers to provide better results than linear models in case studies based on American [17], [18], [19], [20], [21] and British [22], [23] firms.

4.1. Data set employed in this study

Financial data from 29 failed and 31 continuing British corporations were used in this study. The data set covers the period between 1997 and 2000. The variables employed are financial ratios commonly found in the literature: x_1 (working capital/total assets), x_2 (accumulated retained profit/total assets), x_3 (profit before interest and tax/total assets), x_4 (book value of equity/book value of total liabilities), x_5 (sales/total assets). Here the only difference from Altman's choice [16] is the use of the book value of equity, rather than the market value of equity, to calculate x_4. Tables with the complete data set are available in [24].

The companies were divided into a modelling set (21 failed and 21 continuing firms) and a validation set (8 failed and 10 continuing firms).

5. Results and Discussion

5.1. Linear classifier

A linear classifier was initially built by linear discriminant analysis (see Appendix) using the five ratios described above. The resulting model is given by:

$$Z = 3.31x_1 + 0.93x_2 - 2.09x_3 + 0.47x_4 + 0.14x_5 \quad (9)$$

with cut–off value $z_c = 1.51$. A company is classified as continuing if $Z > z_c$ and as failed otherwise.

An inspection of (9) reveals that the 3^{rd} coefficient does not have a logical value. In fact, it should be positive, because x_3 relates profit before interest and tax with total assets, and the higher the value of this ratio, the less likely it is that the firm is in financial trouble (all other factors kept constant). The source of this problem can be understood by calculating the coefficient of multiple correlation [25] of each ratio with respect to the other four. For ratio x_i, this coefficient is defined as $\sigma(\hat{x_i})/\sigma(x_i)$, where $\sigma(\cdot)$ stands for the sample standard deviation and $\hat{x_i}$ is the least-squares estimate of x_i obtained from the other ratios. The values obtained for x_1, x_2, x_3, x_4, x_5 were 0.96, 0.75, 0.97, 0.56, 0.41, respectively. As can be seen, x_3 is the ratio that can be better predicted from the other four. Thus, it is the most likely source of collinearity problems, which are known to deteriorate the prediction ability of linear discriminant models [26].

In order to circumvent the collinearity problem, a new classifier was calculated with four inputs (x_1, x_2, x_4 and x_5), rather than five. The resulting model is as follows:

$$Z = 0.22x_1 + 0.38x_2 + 0.47x_4 + 0.12x_5 \quad (10)$$

with cut–off value $z_c = 0.75$. As can be seen, all coefficients display positive values, which is in agreement with the financial interpretation of the model. Moreover, by re-evaluating the coefficients of multiple correlation after the exclusion of x_3, the values obtained for x_1, x_2, x_4, x_5 were 0.58, 0.60, 0.42, 0.31, which shows that the collinearity effects are much smaller than in the previous case. Interestingly, ratio x_1 no longer exhibits a large coefficient of multiple correlation, which suggests that x_1 was primarily correlated with x_3.

The classifier in eq. (10) yields 9 errors on the modelling set and 4 errors on the validation set, as shown in Table 1. It is worth noting that, if ratio x_3 is not discarded, the number of validation errors increases to 7, which corroborates the financial and statistical reasonings presented above.

Table 1: Results, linear classifier using four ratios. Error types: I (failed company classified as continuing) or II (continuing company classified as failed)

Data set	Type I errors	Type II errors	Total errors	Percent accuracy
Modelling	2	7	9	79 %
Validation	0	4	4	78 %

5.2. Multi–layer perceptron

A multi–layer perceptron with one output, one hidden layer with 10 neurons and hyperbolic tangent activation function in all neurons was used. The company is classified as continuing if the network output is positive and as failed otherwise. The inputs to the network were the five financial ratios x_1, x_2, x_3, x_4, x_5.

The network was trained with the Levenberg–Marquardt method [27], with randomly generated initial weights. In order to alleviate overfitting problems, which might result from an over-parameterized structure, the Optimal Brain Damage pruning technique [11] was employed. This technique employs 2^{nd}-derivative information (Hessian of the cost function with respect to the network parameters) to identify synaptic weights whose elimination will result in the smallest increase in the output error. Pruning was performed by removing one connection at a time and then retraining the network [28].

Figure 2 displays the mean square error MSE in the modelling set as a function of the number of remaining parameters (weights and biases) in the network. This graph suggests that a good trade-off between model complexity and classification performance is obtained at the inflection point indicated by an arrow (22 parameters). Table 2 presents the classification performance of the pruned network on the modelling and validation sets.

Table 2: Results, multi–layer perceptron. Error types: I (failed company classified as continuing) or II (continuing company classified as failed)

Data set	Type I errors	Type II errors	Total errors	Percent accuracy
Modelling	0	0	0	100 %
Validation	0	3	3	83 %

By comparing Tables 1 and 2, it can be seen that the multi–layer perceptron is a better classifier than the linear model. Moreover, the removal of ratio x_3 was not

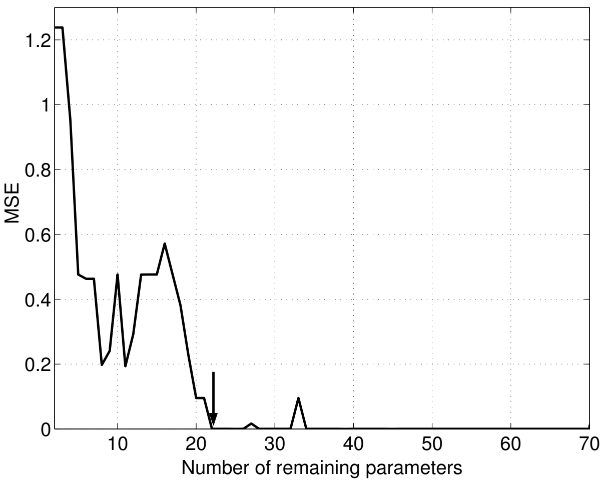

Figure 2: Pruning record of the Optimal Brain Damage training. The arrow indicates the inflection point used to select the number of parameters.

required in this case, which suggests that the neural network may be more robust to collinearity problems than the linear discriminant model.

However, it is worth mentioning that the validation performance of the multi–layer perceptron depended on the initial weights used in its training (with the modelling and validation sets kept the same). This is due to the presence of local minima in the error surface. In fact, by repeating the training procedure several times starting from different sets of weights, it was verified that in some cases, the trained network yielded 4 validation errors instead of 3. The network that gives the best validation performance is usually chosen, as this indicates a better generalisation ability.

5.3. Wavelet network

A Mexican Hat wavelet with grid parameters $\alpha = 2$ and $\beta = R/2 = 2.5$ was employed. This choice of β results in a maximum overlap of 25% for wavelets at a given scale level.

Steps 1 and 2 of the algorithm described in section 3 resulted in 991 wavelets, a number which was reduced to 457 by Step 3. Of these, 258 were discarded in step (b) of the selection process with $\kappa = 10^{-3}$. The use of a threshold $\chi = 10$ for the condition number caused the algorithm to stop after selecting 11 wavelets.

Figure 3 displays the number of classification errors on the modelling set as a function of the number of wavelets added to the network during the selection process. This graphic suggests the choice of six wavelets, according to the principle that, given models with similar prediction capabilities, the one with fewest parameters should be favored (Parsimony Principle). Interestingly, these wavelets are equally distributed among the three scale levels employed ($m = -1, 0, +1$), which justifies the importance of using a multi-scale network structure.

As shown in Table 3, the six-wavelet model leads to 2 errors on the validation set. When compared to the multi–layer perceptron obtained above, this wavelet network is less successful in discriminating the training patterns, but its performance on the validation set is better. The reason may lie in the fact that, although the wavelet network has 43 parameters (6 wavelets × (1 scale + 5 translations) + 6 weights + 1 cut-off value), only 7 of them (weights and cut-off value) are real-valued. Thus, on the overall, the wavelet network structure is simpler then the multi–layer perceptron, which had 22 real-valued parameters.

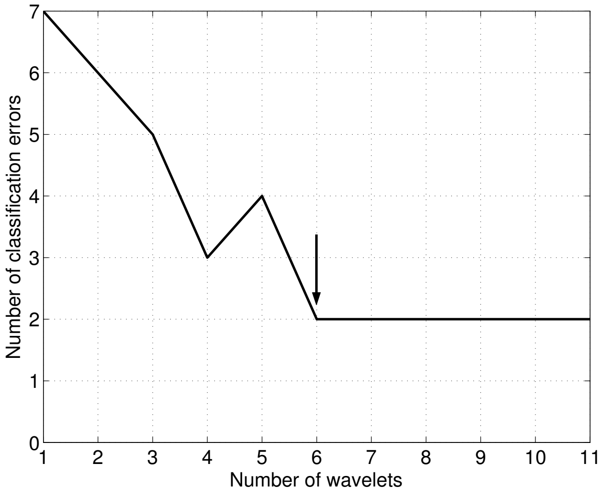

Figure 3: Modelling errors as a function of the number of wavelets employed. The arrow indicates the inflection point used to select the number of wavelets.

Table 3: Results, wavelet netwok. Error types: I (failed company classified as continuing) or II (continuing company classified as failed)

Data set	Type I errors	Type II errors	Total errors	Percent accuracy
Modelling	1	1	2	95 %
Validation	0	2	2	89 %

Notice that, unlike the multi–layer perceptron training, the construction of the wavelet network does not exhibit reproducibility problems, because it does not involve a random generation of initial synaptic weights.

6. Conclusions

This paper proposed a constructive method for building a wavelet network classifier. Tests were carried out in a financial distress prediction problem using data from British firms. The results showed that both the wavelet network and a conventional multi–layer perceptron yielded better results than a classifier obtained by linear discriminant analysis. However, when compared to the multi–layer perceptron, the wavelet network displayed the following advantages:

1. Better generalization ability, which can be attributed to a more parsimonious model structure.

2. Reproducibility of the training outcome, since the construction of the wavelet network does not require the random generation of an initial set of weights.

A possibility not explored in this paper is subjecting the wavelet network to a backpropagation training after selecting the wavelets to be included in the model. This training could possibly improve the network performance by fine-tuning the scale and translation parameters. However, the advantage of having a simple structure, mostly parameterized by integer values, would be lost.

Finally, it is worth noting that the wavelet selection algorithm, which was developed for a two-class problem, can be easily extended to the multi-class problem. Suffice it to use a multi-class Fisher Discriminant Criterion, defined as the ratio of a between-class and a within-class scatter measures [29].

Acknowledgments

The first author acknowledges the support of FAPESP under grant 00/09390-6.

Appendix: Linear Discriminant Analysis

The following linear discriminant function $Z : \mathbb{R}^n \rightarrow \mathbb{R}$ can be derived for the case of binary classification [10]:

$$Z(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{S}^{-1} \mathbf{x} \tag{11}$$

where $\mathbf{x} = [x_1\, x_2\, \cdots\, x_n]^T$ is a vector of n classification variables, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^n$ are the sample mean vectors of each group, and $\mathbf{S}_{n \times n}$ is the common sample covariance matrix. Equation (11) can also be written as:

$$Z = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \mathbf{w}^T \mathbf{x} \tag{12}$$

where the vector of coefficients $\mathbf{w} = [w_1\, w_2\, \cdots\, w_n]^T$ is obtained as $\mathbf{w} = \mathbf{S}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

The cut-off value z_c for classification can be calculated as $z_c = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{S}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$. A given vector $\mathbf{x}$ should be assigned to population 1 if $Z(\mathbf{x}) > z_c$, and to population 2 otherwise.

References

[1] Q. Zhang and A. Benveniste. Wavelet networks. *IEEE Trans. Neural Networks*, 3(6):889–898, 1992.

[2] M. Cannon and J.-J. E. Slotine. Space-frequency localized basis function networks for nonlinear system estimation and control. *Neurocomputing*, 9:293–342, 1995.

[3] Q. Zhang. Using wavelet network in nonparametric estimation. *IEEE Trans. Neural Networks*, 8(2):227–236, 1997.

[4] G. P. Liu, S. A. Billings, and V. Kadirkamanathan. Nonlinear system identification using wavelet networks. *Int. J. Systems Science*, 31(12):1531–1541, 2000.

[5] C. Souza Jr., E. M. Hemerly, and R. K. H. Galvao. Adaptive control for mobile robot using wavelet network. *IEEE Trans. Systems, Man, and Cybernetics, Part B*, 32(4):493–504, 2002.

[6] K.-C. Kan and K.-W. Wong. Self-construction algorithm for synthesis of wavelet networks. *Electronic Letters*, 34(20):1953–1955, 1998.

[7] H. H. Szu, B. Telfer, and S. Kadambe. Neural network adaptive wavelets for signal representation and classification. *Optical Engineering*, 31(9):1907–1916, 1992.

[8] H. Dickhaus and H. Heinrich. Classifying biosignals with wavelet networks - a method for noninvasive diagnosis. *IEEE Eng. Med. Biol. Magazine*, 15(5):103–111, 1996.

[9] R. K. H. Galvao and T. Yoneyama. Improving the discriminatory capabilities of a neural classifier by using a biased-wavelet layer. *Int. J. Neural Systems*, 9(3):167–174, 1999.

[10] D. Morrison. *Multivariate Statistical Methods*. McGraw-Hill, New York, 1990.

[11] Y. Kun, J. Denker, and S. Solla. Optimal brain damage. In D. Touretzky, ed., *Advances in Neural Inf. Proc. Syst.*, p. 598–605. Morgan Kaufmann, San Mateo, 1990.

[12] I. Daubechies. *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[13] C. L. Lawson and R. J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, 1974.

[14] G. Foster. *Financial Statement Analysis*. Prentice-Hall, London, 1986.

[15] W. Beaver. Financial ratios as predictors of failure. *Empirical Research in Accounting*, 5:71–111, 1966.

[16] E. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *J. Finance*, 23(4):589–609, 1968.

[17] M. Odom and R. Sharda. A neural network model for bankruptcy prediction. In *Int. Joint Conf. Neural Networks*, vol. II, p. 163–167, San Diego, California, 1990.

[18] K. Tam and M. Y. Kiang. Predicting bank failures: a neural network approach. *Applications of Artificial Intelligence*, 4:265–282, 1990.

[19] K. Tam and M. Y. Kiang. Managerial applications of neural networks. *Management Science*, 38:926–947, 1992.

[20] P. Coats and L. Fant. Recognizing financial distress patterns using a neural network tool. *Financial Management*, 22:142–155, 1993.

[21] R. L. Wilson and R. Sharda. Bankruptcy prediction using neural networks. *Decision Support Systems*, 11:545–557, 1994.

[22] Y. Alici. Neural networks in corporate failure prediction: the UK experience. In A. Refenes, Y. Abu-Mostafa, and J. Moody, ed., *Neural Networks in Financial Engineering*. World Scientific, London, 1996.

[23] E. Tyree and J. Long. Assessing financial distress with probabilistic neural networks. In A. Refenes, Y. Abu-Mostafa, and J. Moody, ed., *Neural Networks in Financial Engineering*. World Scientific, London, 1996.

[24] R. K. H. Galvao, V. M. Becerra, and M. Abou-Seada. Variable selection for financial distress classification using a genetic algorithm. In *Proc. IEEE Congress on Evolutionary Computation, Honolulu*, p. 2000–2005, 2002.

[25] M. Ezekiel and K. A. Fox. *Methods of Correlation and Regression Analysis*. John Wiley, New York, 1959.

[26] T. Naes and B. H. Mevik. Understanding the collinearity problem in regression and discriminant analysis. *J. Chemometrics*, 15(4):413–426, 2001.

[27] P. Gill, W. Murray, and M. Wright. *Practical Optimization*. Academic Press, London, 1981.

[28] M. Norgaard. Neural Network Based System Identification Toolbox. Technical Report 00-E-891, Technical University of Denmark, Dept. Automation, 2000.

[29] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, New York, 1973.