Neural Networks 23 (2010) 887–891

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A new local–global approach for classification

R.T. Peres, C.E. Pedreira *

COPPE-PEE – Engineering Graduate Program and School of Medicine, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil

ARTICLE INFO

Article history:
Received 22 May 2009
Received in revised form 9 January 2010
Accepted 26 April 2010

Keywords:
Bayes classifier
Classification
Local–global
LVQ
SVM
Vector quantization
Pattern recognition
Prototype

ABSTRACT

In this paper, we propose a new local–global pattern classification scheme that combines supervised and unsupervised approaches, taking advantage of both, local and global environments. We understand as global methods the ones concerned with the aim of constructing a model for the whole problem space using the totality of the available observations. Local methods focus into sub regions of the space, possibly using an appropriately selected subset of the sample. In the proposed method, the sample is first divided in local cells by using a Vector Quantization unsupervised algorithm, the LBG (Linde–Buzo–Gray). In a second stage, the generated assemblage of much easier problems is locally solved with a scheme inspired by Bayes' rule. Four classification methods were implemented for comparison purposes with the proposed scheme: Learning Vector Quantization (LVQ); Feedforward Neural Networks; Support Vector Machine (SVM) and *k*-Nearest Neighbors. These four methods and the proposed scheme were implemented in eleven datasets, two controlled experiments, plus nine public available datasets from the UCI repository. The proposed method has shown a quite competitive performance when compared to these classical and largely used classifiers. Our method is simple concerning understanding and implementation and is based on very intuitive concepts.

1. Introduction

A global approach to classification may be pursued by trying to fit a probability density function, or a mixture of distributions, on the observed data. If this density can be estimated, one may end up with a multimodal distribution, each mode possibly corresponding to a class, and after estimating the *a priori* class probabilities, one could apply Bayes' rule (Duda, Hart, & Stork, 2001) to solve the problem in an optimal manner. Unfortunately, as it is well known, for most of real applications it is in general quite hard, or even impossible, to estimate a global probability density function, especially for high dimensionality spaces (Duda et al., 2001; Silverman, 1986). Furthermore, a pure global approach assumes that data is engendered by a phenomenon governed by a global fundamental law and does not take advantage of possible local generative structures. Accordingly, global models are constructed from all the available observations aiming to represent the entire problem space.

In a local classification approach, the aim is instead of constructing a global generative model from all observations, to build up local classification schemes, possibly using just a subset of the sample. The focus moves into partitions of the original problem.

Successful examples of local–global classification approaches are the kernel methods (Shawe-Taylor & Cristianini, 2004), for instance the Support Vector Machine (SVM) procedure (Vapnik, 1998). In SVM, a subset of the observations, the Support Vectors, is employed to determine a somehow optimal separating hyperplane. The local structure arises intrinsically through the kernels; the global framework comes by means of the generated hyperplanes. In Huang, Yang, King, and Lyu (2008), a local–global large margin classifier is proposed. A time series local–global approach can be found in Fariñas, Pedreira, and Medeiros (2004).

A different view comes from Vector Quantization (VQ) (Linde, Buzo, & Gray, 1980; Gray & Neuhoff, 1998) where the key idea is to use the whole sample (global) in an unsupervised environment to generate a partition that may enhance local structures. This results in a quantized approximation of the distribution, using a finite number of prototype vectors. In a supervised context, VQ can be naturally extended to Learning Vector Quantization (LVQ) (Kohonen, 2001), where prototypes location results from an update procedure based on the training dataset. Once the prototypes are set, one may associate one or more of those with each class, and classify an observation by using the nearest-neighbor rule (Duda et al., 2001). Some procedures were proposed to execute LVQ with an appropriately chosen subset of the sample (Pedreira, 2006; Peres & Pedreira, 2009).

Applications of local–global models may be found in diverse areas such as bioinformatics (Kasabov, 2007) or remote sensing imaging (Blanzieri & Melgani, 2008). Here, we propose a new local–global classification scheme that combines supervised and

* Corresponding address: UFRJ, COPPE-PEE, Av. Horácio Macedo, Prédio do CT, Bloco H, 3o andar, Ilha do Fundão, CEP 21941-914, Rio de Janeiro, Brazil. Tel.: +55 21 2562 8627.
E-mail addresses: rperes@lps.ufrj.br (R.T. Peres), pedreira@ufrj.br (C.E. Pedreira).

0893-6080/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2010.04.010

unsupervised approaches. We use a VQ unsupervised algorithm to divide the sample in local cells. In this first stage, the whole sample is used and some information on the data distribution is captured in a global mode. In this way, an assemblage of much easier problems is generated and locally solved, in a second stage, with a scheme inspired by Bayes' rule.

2. Methodology and datasets

The proposed approach belongs to the divide-and-conquer family. We create a set of sub-classifiers, applied in sub regions of the space, intending to make the classification task easier in a local level.

Let X be a sample of size m comprised by a set of observations $\{x_1, x_2, \ldots, x_m; x_i \in \Re^n, \forall i = 1, \ldots, m\}$. Let us consider a dichotomous classification environment where each observation x_i is associated to one, out of two, possible classes, C_1 or C_2. We denote $y(x_i)$ the label of observation x_i, being $y(x_i) = 1$ if x_i belongs to C_1, and $y(x_i) = 2$ if x_i belongs to C_2.

The proposed methodology is implemented in two stages. We first partition the observation space into cells using an unsupervised procedure. In a second stage, a supervised classification scheme is applied (locally) in each of the previously generated cells.

The first stage is done by the way of the LBG (Linde-Buzo-Gray) algorithm (Linde et al., 1980). Other algorithms, like for instance the k-means, may produce similar results. The LBG algorithm is an iterative procedure that alternatively conducts two actions, update and partition. It is started by setting the centroid of the sample as the code-vector in the initial iteration. It follows (partition) by adding (and subtracting) a perturbation and consequently splitting this code-vector into two. At each iteration, after completing the partition process, the update operation is done by redefining the cells, allocating each observation to its correspondent closest code-vector (in the sense of minimum squared Euclidean distortion), and updating the current code-vectors to the centroids of its associated observations. This process continues until a pre-established number of cells, let's say r, is reached. In this way, one ends up with r code-vectors $\{p_k \in \Re^n, k = 1, \ldots, r\}$ defining r cells $S_1, \ldots, S_r$.

The second stage consists of applying a classification procedure for each of the generated cells. A trivial situation occurs when a cell is homogeneous, i.e. all (training) observations in this cell belong to the same class. In this case, any testing observations attracted to this cell will be associated to the cell label. Otherwise, if the cell is heterogeneous, we introduce a local scheme inspired by the Bayes classifier (Duda et al., 2001).

Without loss of generality, we focus on an arbitrary testing observation x in cell S_k. Let us consider the two subsets ς_k^1 and ς_k^2 of S_k containing observations of classes C_1 and C_2 respectively:

$\varsigma_k^1 \equiv \{x_i \in S_k | y(x_i) = 1\}$ and $\varsigma_k^2 \equiv \{x_i \in S_k | y(x_i) = 2\}$. The relative frequencies f_k^1 and f_k^2 of classes C_1 and C_2 may now be calculated as

$$f_k^1 = \frac{\#\varsigma_k^1}{\#S_k} \quad \text{and} \quad f_k^2 = \frac{\#\varsigma_k^2}{\#S_k},$$

where # represents cardinality. Note that frequencies f_k^1 and f_k^2 are estimators of the *a priori* probability of classes C_1 and C_2 in cell S_k. The *a priori* probability ratio in S_k may now be estimated as:

$$\pi_k \equiv \frac{f_k^1}{f_k^2}.$$

We may next view the inverse of the distances of an observation x to the class means, as estimatives of the likelihood at x. So, we define

$$\hat{L}_x \equiv \left(d(x, m_k^1)\right)^{-1} / \left(d(x, m_k^2)\right)^{-1} \quad (1)$$

as an estimative for the likelihood ratio $L \equiv p(x|C_1)/p(x|C_2)$, where m_k^1 and m_k^2 are respectively the means of the observations with labels 1 and 2 in cell S_k.

We conclude by proposing the following decision rule for a testing observation x in cell k:

$$\begin{cases} x \to C_1 & \text{if } \hat{L}_x \geq (\pi_k)^{-1} \\ x \to C_2 & \text{otherwise.} \end{cases} \quad (2)$$

The proposed algorithm may be summarized as follows:

1. Segment the (training) sample space into r cells $S_1, \ldots, S_r$ by using the LBG algorithm.
2. Calculate the frequency ratio π and the class means m^1 and m^2 for all heterogeneous cells.
3. If an observation (in the testing set) lies in a homogeneous cell, attribute to this observation the label of this cell.
4. Otherwise, if observation x (in the testing set) lies in a heterogeneous cell, calculate $\hat{L}_x$ (as defined in (1)) and attribute a class in accordance with rule (2).

Four classification methods were implemented for comparison purposes with the proposed scheme: (i) LVQ; (ii) Feedforward Neural Networks (NN), (iii) Support Vector Machine (SVM) and (iv) k-Nearest Neighbors. The NNs were trained with Bayesian Regularization, with 10 initial neurons in the hidden layer and logistic activation function in both, the hidden and the output layers. For SVM we used radial basis function kernel.

2.1. On the datasets

In this sub-section, we briefly describe the datasets used to benchmark the proposed method performance. Besides the proposed algorithm, the four classification procedures, described in the previous sub-section, were implemented in eleven datasets, two controlled experiments, plus nine public available datasets from the UCI[1] repository, namely: Waveform, Letter-B, Statlog, Heart Diseases Diagnosis, Breast Cancer, Ionosphere, Pima Indians, Glass and Lung Cancer.

Experiments 1 and 2 are synthetic data. Experiment 1 consists of two classes divided by a cosine function. We generated 1030 observations of class C_1 and 1027 observations of class C_2 for in-sample, and 1060 observations labeled C_1 and 1041 observations labeled C_2 for out-of-sample. For experiment 2 two classes were generated through a circle and a roll with coincident centers (without superposition). In-sample: 123 observations of class C_1 and 2611 observations of class C_2. Out-of-sample: 127 observations of class C_1 and 2646 observations of class C_2.

The original dataset for experiment 3 had 3 classes of waveforms; here, we tested class 1 against the other two. This dataset is composed of 5000 observations, with 40 input features (3000 used for in-sample and 2000 for the out-of-sample phase).

Experiment 4 concerns letter recognition. The objective is to identify the 26 capital letters in the English alphabet. We set letter B as one class against the other 25 letters. The dataset consists of 20 000 observations with 16 input features, 10 600 of those used for in-sample and 9400 for out-of-sample testing.

The data for experiment 5 comes from the Statlog - landsat satellite. This dataset consists of the multi-spectral values of pixels in 3×3 neighbourhoods in a satellite image. The aim is to classify images, associated with the central pixel in each neighbourhood, given the multi-spectral values. We tested class 1 against the others. There are 6435 observations with 36 input features, 4435 were used for in-sample phase leaving 2000 for the out-of-sample testing.

The dataset for experiment 6, related to the diagnosis of coronary artery disease, was formed by an assemblage of four data sets as in Pedreira, Macrini, and Costa (2005). Each of these four data

[1] http://www.ics.uci.edu/~mlearn/MLSummary.html.

Table 1
Comparative results: out-of-sample correctness rate (parenthesis indicates in-sample) for large size datasets (more than 1000 observations).

	Proposed method		LVQ		K-NN		NN		SVM	
	Correctness (%)	Std	Correctness (%)	Std	Correctness (%)	Std	Correctness (%)	Std	Correctness (%)	Std
Exp 1	98.5 (100)	-	98.7 (99.5)	0.2 0.1	99.1 (99.1)	-	99.4 (99.7)	0.3 0.5	99.0 (99.2)	-
Exp 2	99.1 (99.9)	-	98.9 (99.0)	0.2 0.2	99.5 (99.4)	-	99.7 (100)	0.1 0.0	99.0 (98.7)	-
Exp 3	88.9 (93.9)	-	90.1 (96.9)	0.4 0.4	89.8 (88.2)	-	89.7 (99.6)	0.5 0.1	67.6 (100)	-
Exp 4	98.7 (99.4)	-	98.0 (99.3)	0.1 0.1	99.3 (99.3)	-	98.6 (99.8)	0.2 0.2	99.4 (100)	-
Exp 5	99.0 (100)	-	99.0 (99.6)	0.1 0.1	99.6 (98.9)	-	98.8 (99.9)	0.3 0.1	98.3 (100)	-
Average	96.8 (98.6)	-	96.9 (98.9)	0.2 0.2	97.5 (97.0)	-	97.2 (99.8)	0.3 0.2	92.7 (99.6)	-

Table 2
Comparative results: out-of- sample correctness rate (parenthesis indicates in-sample) for median size datasets (300–1000 observations).

	Proposed method		LVQ		K-NN		NN		SVM	
	Correctness (%)	Std	Correctness (%)	Std	Correctness (%)	Std	Correctness (%)	Std	Correctness (%)	Std
Exp 6	79.6 (91.5)	2.4 0.8	76.1 (88.2)	3.9 0.8	78.9 (79.1)	2.6 0.8	74.9 (96.9)	4.2 0.4	75.0 (100)	5.1 0.0
Exp 7	96.8 (99.8)	2.3 0.3	96.6 (98.6)	2.5 0.3	97.5 (97.4)	2.3 0.4	94.9 (99.8)	3.5 0.3	95.9 (100)	3.1 0.0
Exp 8	90.3 (96.4)	3.6 0.7	86.9 (93.9)	6.5 1.5	83.8 (84.7)	5.7 1.1	88.0 (100)	4.2 0.0	89.5 (100)	8.0 0.0
Exp 9	72.3 (94.3)	5.8 1.0	68.3 (87.9)	5.8 1.0	73.2 (73.2)	4.3 0.8	70.1 (70.6)	6.6 3.4	68.4 (100)	4.1 0.0
Average	84.8 (95.5)	3.5 0.7	82.0 (92.2)	4.7 0.9	83.4 (83.6)	3.7 0.8	82.0 (91.8)	4.6 1.0	82.2 (100)	5.1 0.0

sets is individually available in the UCI[2] machine learning repository. The data was collected from the Cleveland Clinic Foundation; the Hungarian Institute of Cardiology; the V.A. Medical Center, and the Zurich University Hospital. All the original databases have 76 features but only 13 of them are actually relevant (Detrano et al., 1989). The goal is to predict angiographic disease status concerning narrowing in major vessels. After missing data elimination, the dataset ended up with 740 patients and 10 input features.

Experiment 7 dataset, provided by the University of Wisconsin Hospitals, is related to an application in diagnosing breast mass cytology. The features are nine cytological characteristics of benign or malignant breast fine-needle aspirates (uniformity of cell shape, uniformity of cell size, clump thickness, bare nuclei, cell size, normal nucleoli, clump cohesiveness, nuclear chromatin, and mitosis). All the features assume discrete values between 1 and 10. The two output classes are "benign" or "malignant" tumor. After removing the observations with missing values, the dataset ended up with 683 observations.

The dataset for experiment 8 refers to data collected by a radar system in Goose Bay, Labrador. This system consists of a phased array of 16 high-frequency antennas with a total transmitted power on the order of 6.4 kW. The targets were free electrons in the ionosphere. "Good" radar returns are those showing evidence of some type of structure in the ionosphere. "Bad" returns are those that do not; their signals pass through the ionosphere. Received signals were processed using an autocorrelation function whose arguments are the time of a pulse and the pulse number. There were 17 pulse numbers with observations described by 2 features per pulse number resulting in 34 input features in 351 observations.

Experiment 9 is the Pima Indians diabetes dataset. All patients are females, at least 21 years old of Pima Indian heritage. There are 768 observations with 8 input features, 500 being negative and 268 positive for diabetes.

The Glass Identification, experiment 10 dataset, is used in forensic science for the identification of 'windows' or 'non-windows' glass. In the original dataset, four different types of window glass and three different types of non-window glass are provided; here we used window against non-window glass. This dataset is composed of 9 input features and 214 observations.

The Lung Cancer dataset, used for experiment 11 was originally constituted by 3 types of pathological lung cancers with 56 input features, we used pathology 1 against the others. This dataset has just 27 observations after removing the missing values, 8 from C_1 and 19 from C_2.

3. Results

Performance results for the proposed method, and comparatively for LVQ, Layered Neural Networks, SVM and K-NN, for all the used datasets are condensed in Tables 1–3. These tables include in-sample and out-of-sample correctness rates.

The number of prototypes plays of course a key role in local–global schemes, and performance is expected to be affected by this choice. As it classically occurs with unsupervised procedures, the 'best' number of prototypes cannot be generally established. Notice that the local–global balance may be also related to the ratio of the number of prototypes and the size of the sample. Note that, an unbalanced large quantity of prototypes in a sample with few observations would most likely result in overfitting. On the other hand, a low proportion of prototypes, in relation to sample size, would hardly be able to provide an appropriate model for these observations. As the quantity of prototypes increases, the number

[2] http://www.ics.uci.edu/~mlearn/MLSummary.html.

Table 3
Comparative results: out-of-sample correctness rate (parenthesis indicates in-sample) for small size datasets (less than 300 observations).

	Correctness				
	Proposed method (%)	LVQ (%)	K-NN (%)	NN (%)	SVM (%)
Exp 10	94.9 (98.9)	94.4 (97.4)	94.4 (94.4)	91.1 (96.4)	93.9 (100)
Exp 11	77.8 (99.6)	77.8 (98.1)	85.2 (84.6)	85.2 (100)	70.4 (100)
Average	86.4 (99.3)	86.1 (97.8)	89.8 (89.5)	88.2 (98.2)	82.2 (100)

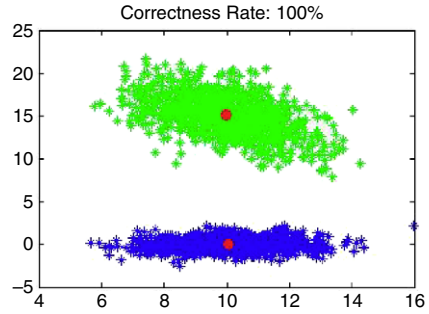

Fig. 1. Toy example: simple structure.

of observations allocated in each cell diminishes, small areas in the observation space are 'populated' by prototypes, making the analysis more local. In the limit, if one assigns one prototype to each observation, the approach would turn completely local. On the other hand, by setting a single prototype to the whole dataset, the approach would be totally global, representing the observations by its mean. In this way, the number of prototypes one employs is directly associated to the level of details in data one is aiming to capture. Accordingly, in problems involving more complexity, it may be indicated to allow for more prototypes. As an illustration, let us consider the toy example in Figs. 1 and 2. Simulation in Fig. 1 is extremely simple and the problem, not surprisingly, may be solved with just 2 prototypes. In the slightly more complex situation in Fig. 2, one notices that performance enhances as more prototypes are added, more details in data are captured by the introduction of new prototypes. This association, of the convenient number of prototypes with complexity, is in fact the same one has when dealing with statistical models over or under parameterization. In our case, this association is, by the nature of the proposed method, clearly linked with a more global or local focus.

For the prototype methods, the LVQ and the proposed scheme, we implemented four different ratios of number of prototypes and size of the sample: 1/3; 1/5; 1/7 and 1/9. For LVQ, half of the prototypes were allocated for each class.

Note that by construction of the LBG, as used here, the prototypes are set as powers of 2 (2, 4, 8, 16, etc.). For all experiments, we use the maximum power of 2 that is less than established ratio of the sample size. For instance, in a dataset with 1000 observations, the number of experimented prototypes would be 256, 128, 128, and 64 (for ratios 1/3; 1/5; 1/7 and 1/9).

The experiments were divided in three groups concerning their sample sizes and distinct strategies were adopted for the out-of-sample performance evaluation in each of these groups. Group 1, experiments 1 to 5, comprises the larger datasets, with more than 1000 observations. Group 2, experiments 6 to 9, are the ones with 300 to 1000 observations and group 3, experiments 10 and 11, are the smaller ones, with less than 300 observations.

For the larger datasets in group 1 the observations are divided into in-sample and out-of-sample sets. For the other two groups of experiments we applied *k*-fold and leave-one-out schemes as described in the sequence.

For group 1 experiments the indicator of performance was set as follows: For LVQ we used the average of 10 prototypes initializations; For NN, we used the average of ten weight initializations; For SVM, we tested for kernel width and constant C for the following values: 2^0, $2^{0.5}$, 2, 2^2, 2^3, 2^4, 2^5; For K-NN, 3, 5, 7 and 9 neighbors were implemented.

For experiments in group 2, we used a 10-fold cross-validation. Accordingly, for each dataset, a 10-fold disjoint partition of the observations was created, and for each of the methods and for each parameter settings the in-sample phase was performed with 9 of the 10 folds. The remaining fold was used to calculate the out-of sample performance. This procedure was repeated 10 times, each of those with a different fold as a test set and the average correctness of these 10 runs is calculated.

Finally, for the smaller datasets in group 3 we used the classical leave-one-out procedure (Duda et al., 2001) that corresponds to the *k*-fold cross-validation using *k* equal to the number of available observations.

The reported out-of-sample performance in Tables 1–3 refers, in each case, to the correctness rate in the out-of-sample set corresponding to the configuration of parameters that reached the highest performance for the in-sample set. The standard deviation, in Table 1, refers to the average of the experiments, and in Table 2 to *k*-fold. For group 3 (Table 3), we implemented a leave-one-out procedure.

4. Discussion and final remarks

The proposed method has shown a quite competitive performance when compared to the classical and largely used classifiers: LVQ, NN, SVM and *k*-Nearest Neighbors. Simplicity and easy implementation may be key factors in the choice of a classification procedure. Our method is simple concerning understanding and implementation and is based on very intuitive concepts.

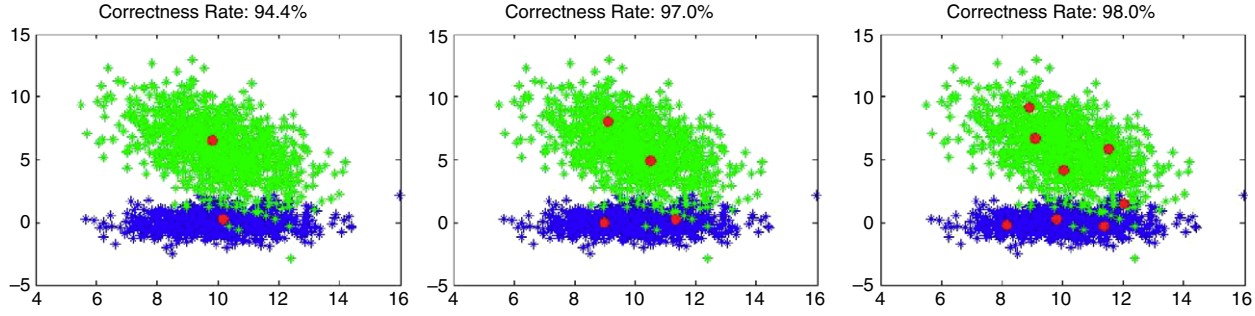

Fig. 2. Toy example: less simple case.

Although the proposed scheme reached comparatively better performances for the smaller and median size datasets, for the large size ones it ended up just 0.7% behind (concerning the average correctness for the 5 experiments) against the best result (k-NN). Our method reached the best performance among all for the median size experiments and was ranked third in the few observations group.

The chosen number of prototypes reflects how local the procedure is and obviously impact on performance. This number should be related to the size of the sample and to the complexity of the problem the dataset emerged from. If a large number of prototypes is set for a given level of complexity, the model would get too local and would probably miss the structure of the data by fitting specificities of the training dataset. On the other hand, a short number of prototypes would imply in a less-than-it-should parameterized model, not allowing for a proper representation of the data structure. Here, we just intended to guide how this specificity versus complexity issue is reflected in the context of the proposed method. This concern is of course a general one and appears in different forms in all methods.

References

Blanzieri, E., & Melgani, F. (2008). Nearest neighbor classification of remote sensing images with the maximal margin principle. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(6), 1804–1811.

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *American Journal of Cardiology*, *64*(5), 304–310.

Duda, R. O., Hart, P. E., & Stork, G. (2001). *Pattern recognition* (2nd ed.). US: Wiley.

Fariñas, M. S., Pedreira, C. E., & Medeiros, M. C. (2004). Local–global neural networks: a new approach for nonlinear time series modelling. *JASA - Journal of the American Statistical Association*, *99*(468), 1092–1107.

Gray, R. M., & Neuhoff, D. L. (1998). Quantization. *IEEE Transactions on Information Theory*, *44*, 2325–2384.

Huang, K., Yang, H., King, I., & Lyu, M. R. (2008). Maxi–min margin machine: learning large margin classifiers locally and globally. *IEEE Transactions on Neural Networks*, *19*(2), 260–272.

Kasabov, N. (2007). Global, Local and personalized modeling and pattern discovery in bioinformatics: an integrated approach. *Pattern Recognition Letters*, *28*, 673–685.

Kohonen, T. (2001). *Self-organizing maps* (3rd ed.). Springer.

Linde, Y., Buzo, A., & Gray, R. M. (1980). An algorithm for vector quantizer design. *IEEE Transactions on Communications*, *COM-28*(1), 84–95.

Pedreira, C. E. (2006). Learning vector quantization with training data selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(1), 157–162.

Pedreira, C. E., Macrini, L., & Costa, E. S. (2005). *Proceedings from IEEE-INNS-ENNS international joint conference on neural networks*, Montreal.

Peres, R. T., & Pedreira, C. E. (2009). The generalized risk zone: Observations selection for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(7), 1331–1337.

Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

Silverman, B. W. (1986). *Monographs on statistics and applied probability*: *Vol. 26*. *Density estimation for statistics and data analysis*. Chapman & Hall/CRC.

Vapnik, V. N. (1998). *Statistical learning theory*. New York: Wiley.