F1.2 Pattern classification

Thierry Denœux

Abstract

Pattern classification consists in assigning entities, described by feature vectors, to predefined groups of patterns. When the statistical characteristics of the problem under consideration are perfectly known, minimal error probability can be achieved by means of the Bayes decision rule. In practice, however, a suboptimal classifier has to be constructed from training data. Several neural network approaches to this problem have been proposed. *Nearest-neighbor* models are based on assessing the similarity between the input pattern and a set of reference patterns with known classification. The *regression* approach consists in predicting category from pattern by minimizing a certain error criterion. In the finite sample case, the definition of the structural complexity of these models is shown to have considerable influence on classification error. Finally, a taxonomy of the main neural network and alternative techniques of pattern classification are presented.

F1.2.1 Introduction

In many application domains such as *character recognition*, *speech understanding*, *medical diagnosis*, *process fault detection* or *financial decision making*, problems arise that consist of classifying entities, represented by feature vectors, into one of several groups of patterns, or *classes*. A classification system is typically composed of two parts (Duda and Hart 1973, Fukunaga 1990). A *preprocessor* transforms raw data produced by sensors or extracted from computer databases into vectors of *observations* or *features*. Features are defined so as to encode in compact form most of the information needed to discriminate between pattern categories. Feature vectors are then passed to a *classifier* that evaluates the evidence presented and makes a decision regarding the class assignment of the entity under consideration.

G1.3, F1.7, G5
G2.8, G6.3

Ever since the pioneering work of Rosenblatt (1958) and Widrow (Widrow and Lehr 1990), a large part of connectionist research has been devoted to the development and theoretical analysis of pattern classifiers having neural-network-like structure and learning capabilities. In recent years, the development of several new models with previously unequaled performance in real-world applications (Rumelhart *et al* 1986, Kohonen 1987) has generated a wave of interest in connectionism and pattern recognition in general. Although this enthusiasm was first considered with some skepticism by researchers in mainstream statistical pattern recognition (Duin 1994), artificial neural networks are now generally seen as particular types of *statistical pattern classifiers* (Schmidt 1993, Werbos 1991).

B6

In the next section, the basic notation and definitions underlying statistical pattern recognition will first be defined. The main neural network approaches to pattern classification will then be described, with an overview of their asymptotic and small-sample properties. In the last section, a taxonomy of statistical and neural network classifiers will be presented.

F1.2.2 Problem description

We consider a finite number M of populations or classes, $\omega_1, \ldots, \omega_M$. An entity of interest is assumed to belong to one and only one of these populations. Each entity is described by a feature vector $\boldsymbol{x} \in \mathbb{R}^d$ which is seen as a realization of a random vector $\boldsymbol{X}$. The probability density function of $\boldsymbol{X}$ in class ω_i is

denoted by $f_X(\boldsymbol{x}|\omega_i)$. Each entity is generally assumed to be drawn from a mixture of the M populations, in proportions $P(\omega_1), \ldots, P(\omega_M)$, respectively, with $\sum_{i=1}^{M} P(\omega_i) = 1$. The mixture density of $\boldsymbol{X}$ is then

$$f_X(\boldsymbol{x}) = \sum_{i=1}^{M} P(\omega_i) f_X(\boldsymbol{x}|\omega_i) . \tag{F1.2.1}$$

$P(\omega_i)$ can be seen as the prior probability that the entity belongs to ω_i. Having observed feature vector $\boldsymbol{x}$, the posterior probability $P(\omega_i|\boldsymbol{x})$ can be computed by applying the Bayes theorem:

$$P(\omega_i|\boldsymbol{x}) = \frac{f_X(\boldsymbol{x}|\omega_i) P(\omega_i)}{f_X(\boldsymbol{x})} . \tag{F1.2.2}$$

If the class-conditional probability distributions and the priors are all known, then an optimal solution to the classification problem is provided by Bayes decision theory. Let us denote by $A = \{\alpha_1, \ldots, \alpha_a\}$ a finite set of actions; α_i is often interpreted as the decision of allocating $\boldsymbol{x}$ to class ω_i. However, other actions such as ambiguity or distance rejection (Chow 1970, Dubuisson and Masson 1993) can also be considered in the analysis.

If, as a result of observing pattern $\boldsymbol{x}$, we take action α_i while the entity under consideration belongs to class ω_j, we incur a loss $\lambda(\alpha_i|\omega_j)$. The expected loss $R(\alpha_i|\boldsymbol{x})$ is

$$R(\alpha_i|\boldsymbol{x}) = \sum_{j=1}^{M} \lambda(\alpha_i|\omega_j) P(\omega_j|\boldsymbol{x}) . \tag{F1.2.3}$$

A *decision rule* is a function $\alpha : \mathbb{R}^d \mapsto A$ that prescribes an action $\alpha(\boldsymbol{x})$ each time an observation vector $\boldsymbol{x}$ is encountered. The overall risk associated to α is

$$R(\alpha) = \int_{\mathbb{R}^d} R(\alpha(\boldsymbol{x})|\boldsymbol{x}) f_X(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} . \tag{F1.2.4}$$

The decision rule that minimizes the risk can be shown to be the *Bayes rule*, which selects for each vector $\boldsymbol{x}$ the action α_i for which $R(\alpha_i|\boldsymbol{x})$ is minimum.

In the particular case of a zero-one loss function $\lambda(\alpha_i|\omega_j) = 1 - \delta_{ij}$, where δ is the Kronecker symbol, we have

$$R(\alpha_i|\boldsymbol{x}) = 1 - P(\omega_i|\boldsymbol{x}) \tag{F1.2.5}$$

and the overall risk is the average probability of misclassification. Consequently, the Bayes rule consists in this case of selecting the class with the highest posterior probability. This rule has optimal classification performance in the sense that it minimizes the average probability of error.

In practice, however, this rule cannot be applied because the exact posterior probabilities are unknown. However, approximations to that rule can be constructed if a training set $\mathcal{T} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{t}^{(1)}), \ldots, (\boldsymbol{x}^{(\ell)}, \boldsymbol{t}^{(\ell)})\}$ of ℓ patterns with known classification is available; $\boldsymbol{t}^{(i)}$ denotes a vector of M zero-one indicator variables defining the known class of pattern $\boldsymbol{x}^{(i)}$:

$$\boldsymbol{t}_k^{(i)} = 1 \qquad \boldsymbol{x}^{(i)} \in \omega_k \tag{F1.2.6}$$

$$\boldsymbol{t}_k^{(i)} = 0 \qquad \boldsymbol{x}^{(i)} \notin \omega_k . \tag{F1.2.7}$$

The construction of allocation rules based on a limited amount of training data is one of the fundamental problems in statistical pattern recognition and connectionism.

F1.2.3 Neural network classifiers

In the past thirty years, a large number of neural network models have been proposed for performing pattern classification tasks. Although these models are characterized by a variety of architectures and learning rules, most of them can be seen as instances of two main paradigms, the *nearest-neighbor* approach and the *regression* approach, which are summarized in the following sections.

F1.2.3.1 The nearest-neighbor approach

In the nearest-neighbor approach, the most probable classification of an unknown pattern is determined by assessing its similarity with a set of reference vectors or *prototypes* of each class. The pattern is assigned to the class of the nearest prototype. As a consequence, the surface separating the different decision regions is piecewise linear. In such models, learning is essentially a process of prototype formation and adaptation. Two important models in this category are the *restricted Coulomb energy* (RCE) network (Reilly *et al* 1982) and the *learning vector quantization* (LVQ) network (Kohonen 1987). C1.6.3.1 C1.1.5

In the RCE model, each prototype of a given class is characterized by a weight vector and a receptive field size. The learning algorithm combines two mechanisms of prototype formation and receptive field modification. If input $\boldsymbol{x}$ belonging to class ω_j does not fall into the receptive field of any prototype of that class, then a new prototype of class ω_j is created at the location of $\boldsymbol{x}$. If $\boldsymbol{x}$ falls inside the receptive field of some prototype of class $c \neq \omega_j$, then the receptive field of that prototype is reduced so as to exclude $\boldsymbol{x}$. This algorithm has been shown experimentally to be able to resolve class boundaries of arbitrary complexity. However, since no adaptation of prototype vectors is performed, the required number of prototypes may grow very large. Also, the learning process usually becomes unstable in regions where there is a strong overlap between classes. Some improvements to this basic model have been proposed (Reilly *et al* 1982).

The LVQ model introduced by Kohonen (1987, 1990) essentially differs from the previous one in that the number of prototypes is fixed, but their weight vectors are continuously updated in the course of the learning process by a *competitive learning* mechanism. Upon presentation of input vector $\boldsymbol{x}$ of class ω_j, the nearest prototype i is selected. If that prototype belongs to class $c^{(i)}$, its weight vector $\boldsymbol{p}^{(i)}$ is updated as

$$\boldsymbol{p}^{(i)} \leftarrow \boldsymbol{p}^{(i)} + \eta(t)(\boldsymbol{x} - \boldsymbol{p}^{(i)}) \qquad \text{if} \quad c^{(i)} = \omega_j \tag{F1.2.8}$$

$$\boldsymbol{p}^{(i)} \leftarrow \boldsymbol{p}^{(i)} - \eta(t)(\boldsymbol{x} - \boldsymbol{p}^{(i)}) \qquad \text{if} \quad c^{(i)} \neq \omega_j \tag{F1.2.9}$$

where $\eta(t)$ is a time-decreasing scalar parameter ($0 < \eta(t) < 1$). After training, the prototype vectors acquire values such that classification using the nearest-neighbor principle approximates the Bayes rule with zero-one costs. Variants of this basic scheme have been proposed by Kohonen (1990) and others (e.g. Poirier and Ferrieux 1991).

Simulations performed with both models (RCE and LVQ) on a simple two-class problem are reported in figure F1.2.1. The LVQ algorithm can be seen to yield a smoother decision boundary with a comparatively smaller number of neurons, as a result of prototype adaptation during training.

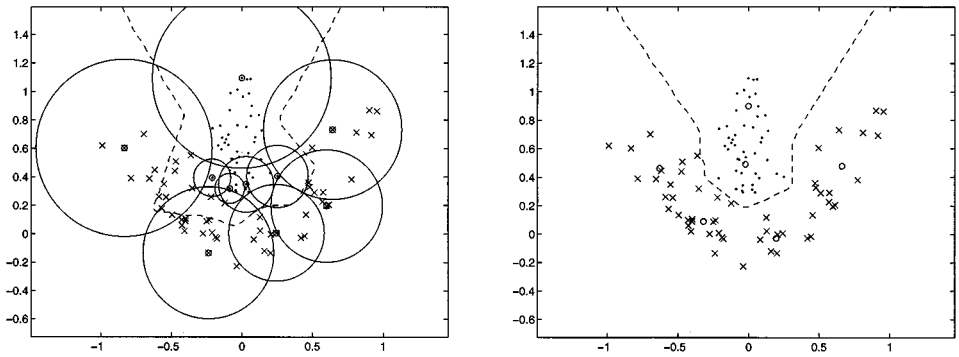

Figure F1.2.1. Prototypes (○) and decision boundaries (– – –) obtained by RCE (left) and LVQ (right) networks in a two-class problem. The receptive fields of RCE prototypes are indicated as circles.

Neural network classifiers based on the nearest-neighbor approach have the advantage of being fast during both training and operation. Experimentally, they are generally found to offer good performance as compared to other, more computationally demanding methods (Kohonen *et al* 1988). Although it is conjectured that the methods relying on competitive learning allow approximation to the Bayes rule for large sample sizes (Kohonen 1990), the determination of the quality of this approximation is a difficult theoretical problem. When classification is performed by considering the nearest neighbor *among samples*, the asymptotic error rate is known to be bounded between the Bayes error and twice the Bayes error (Cover and Hart 1967). This result can be seen as a heuristic justification of the good performance of nearest-neighbor techniques in large-sample problems.

F1.2.3.2 The regression approach

Classification by regression is certainly the most popular approach in the field of artificial neural networks. A regression classifier attempts to predict category from pattern by minimizing a measure of expected error between output and target patterns (Thomas and Mitiche 1994).

More precisely, let us denote the input–output function implemented by a neural network with specified architecture by

$$F : \mathbb{R}^d \times \mathbb{R}^W \mapsto \mathcal{O} \tag{F1.2.10}$$

$$(\boldsymbol{x}, \boldsymbol{w}) \rightarrow F(\boldsymbol{x}, \boldsymbol{w}) \tag{F1.2.11}$$

where $\mathcal{O}$ is the set of possible output values and $\boldsymbol{w}$ is the vector of weights of size W.

In the case of *multilayer perceptrons* (MLPs) (Rumelhart *et al* 1986) with one hidden layer and a logistic activation function in the hidden layer, the kth component $F_k(\boldsymbol{x}, \boldsymbol{w})$ of output vector $F(\boldsymbol{x}, \boldsymbol{w})$ is defined as C1.2

$$F_k(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=1}^{N_2} w_{kj}^{(2)} \sigma\left(\sum_{i=1}^{d} w_{ji}^{(1)} x_i + \theta_j^{(1)}\right) + \theta_k^{(2)} \tag{F1.2.12}$$

where $w_{ji}^{(1)}$ is the weight from input unit i to hidden unit j, $\theta_j^{(1)}$ is the bias of hidden unit j, $w_{kj}^{(2)}$ is the weight from hidden unit j to output unit k, $\theta_k^{(2)}$ is the bias of output unit k, N_2 is the size of the hidden layer, and σ is a sigmoid function.

In the case of *radial basis function* (RBF) networks (Poggio and Girosi 1988, Girosi 1994), the output from hidden unit j is defined as a function of the Euclidean distance between input $\boldsymbol{x}$ and a prototype vector $\boldsymbol{p}^j$. As in the previous model, output units compute a weighted sum of the outputs from the hidden layer. The output $F_k(\boldsymbol{x}, \boldsymbol{w})$ from unit k is given by C1.6.2

$$F_k(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=1}^{N_2} w_{kj} \exp\left(-\frac{1}{2\sigma_j^2} \|\boldsymbol{x} - \boldsymbol{p}^{(j)}\|^2\right) \tag{F1.2.13}$$

where w_{kj} is the weight from hidden unit j to output unit k, σ_j is a parameter defining the size of the receptive field of prototype j, and N_2 is defined as above.

An important distinction between MLPs and RBF networks concerns the nature of the internal representation of input patterns. In MLPs, an input signal may activate an arbitrary number of hidden units, resulting in a *distributed* representation. In RBF networks, one input predominantly activates the hidden unit with the closest weight vector, which creates a *local* representation. From this point of view, RBF networks are related to the nearest-neighbor classifiers described in the previous section. A comparison of both models on the same two-class problem as above is shown in figure F1.2.2.

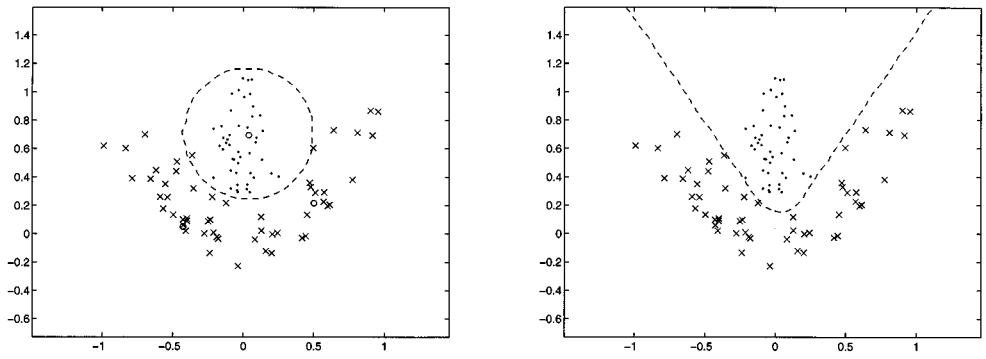

Figure F1.2.2. Decision boundaries (– – –) obtained by an RBF network with three prototype units (left) and by an MLP with two hidden units (right) in a two-class problem.

Both MLPs and RBF networks share the fundamental property of being universal approximators; that is, given enough hidden units, they can approximate any continuous mapping with arbitrary accuracy (Poggio and Girosi 1988, Hornik 1991).

In the MLP and RBF network models, training is performed by optimizing the performance on a training set $\mathcal{T} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{t}^{(1)}), \ldots, (\boldsymbol{x}^{(\ell)}, \boldsymbol{t}^{(\ell)})\}$, using some iterative procedure (Rumelhart *et al* 1986).

Performance is assessed by computing the mean of some error measure between the classifier output and target values. Different output coding schemes and error measures have been proposed. Typically, the desired output for training vector $\boldsymbol{x}^{(i)}$ is taken as $\boldsymbol{t}^{(i)}$, and the error for that pattern is defined as $\|\boldsymbol{t}^{(i)} - F(\boldsymbol{x}^{(i)}, \boldsymbol{w})\|^2$. The empirical performance on the training set is then

$$J_\ell(\boldsymbol{w}) = \frac{1}{\ell} \sum_{i=1}^{\ell} \|\boldsymbol{t}^{(i)} - F(\boldsymbol{x}^{(i)}, \boldsymbol{w})\|^2 . \tag{F1.2.14}$$

During training, one seeks a weight vector $\boldsymbol{w}_\ell$ solution of the problem:

$$\min_{\boldsymbol{w}} J_\ell(\boldsymbol{w}) . \tag{F1.2.15}$$

However, in most cases, the ultimate goal of learning is in fact to minimize the overall performance for any possible input vector, which can be measured by

$$J(\boldsymbol{w}) = E(\|\boldsymbol{T} - F(\boldsymbol{X}, \boldsymbol{w})\|^2) \tag{F1.2.16}$$

where $\boldsymbol{X}$ is a random input vector and $\boldsymbol{T}$ is the corresponding random target vector. For large ℓ, $J_\ell(\boldsymbol{w})$ can be seen as an approximation to $J(\boldsymbol{w})$, and $\boldsymbol{w}_\ell$ approximates the solution $\boldsymbol{w}^*$ to

$$\arg\min_{\boldsymbol{w}} J(\boldsymbol{w}) . \tag{F1.2.17}$$

If the training set is now seen as a realization of a random sample

$$((\boldsymbol{X}^{(1)}, \boldsymbol{T}^{(1)}), \ldots, (\boldsymbol{X}^{(\ell)}, \boldsymbol{T}^{(\ell)})) \tag{F1.2.18}$$

then $J_\ell(\boldsymbol{w})$ and $\boldsymbol{w}_\ell$ become realizations of random variables $\hat{J}_\ell(\boldsymbol{w})$ and $\hat{\boldsymbol{w}}_\ell$, respectively. White (1989) discusses conditions on which the sequence of real-valued random variables $\hat{\boldsymbol{w}}_\ell$ converges, in some strict mathematical sense, to $\boldsymbol{w}^*$.

So far, we have assumed performance to be assessed by a measure of the distance between desired and obtained output patterns. Intuitively, a classifier whose outputs are close to target values for each $\boldsymbol{x}$ can be expected to have low error probability. As the number ℓ of training vector becomes infinitely large, it is interesting to study the relationships of this approach with the Bayes rule. This has been done by many authors (White 1989, Hampshire and Pearlmutter 1991, Lee *et al* 1991, Thomas and Mitiche 1994). The main result is that $\boldsymbol{w}^*$ minimizes

$$\int_{\mathbb{R}^d} \|E(\boldsymbol{T}|\boldsymbol{x}) - F(\boldsymbol{x}, \boldsymbol{w})\|^2 \, \mathrm{d}\boldsymbol{x} . \tag{F1.2.19}$$

By definition of $\boldsymbol{T}$, $E(T_j|\boldsymbol{x}) = P(T_j = 1|\boldsymbol{x}) = P(\omega_j|\boldsymbol{x})$. Consequently, $F_j(\boldsymbol{w}^*, \boldsymbol{x})$ is a mean-squared approximation to the posterior probability $P(\omega_j|\boldsymbol{x})$. A classifier trained by minimization of the mean-squared error criterion therefore approximates the Bayes rule asymptotically in ℓ. This result has been extended to other error functions by Hampshire and Pearlmutter (1991) and to more general output coding schemes by Thomas and Mitiche (1994). Note, however, that the quality of this approximation depends on the architecture of the network under consideration, as well as on the training procedure employed, which may not be able to reach a global minimum of the error function.

F1.2.3.3 Small-sample problems

As remarked by Raudys and Jain (1991a), the asymptotic classification error of a regression classifier, assuming a perfect training algorithm, cannot be increased by introducing new hidden units. If the classifier is made more complex, this can only result in a closer approximation to the Bayes classification rule. This remark also applies, to some extent, to nearest-neighbor classifiers, since increasing the number of prototypes results in a closer approximation to the 1-NN classifier, which is known to have near-optimal asymptotic performance.

In practice, however, one is always in a situation where only a finite number of training samples is available. In such a case, numerical simulations reveal the existence of the so-called *peaking phenomenon* (Raudys and Jain 1991a). As the complexity of the classifier increases, classification error initially drops,

then attains a minimum, and then begins to increase. Intuitively, this is due to the fact that inexact estimation of additional parameters increases classification error. At some point, this effect becomes larger than the gain resulting from greater flexibility of the classifier. For that reason, the design of patterns classifiers with optimal complexity is of the utmost importance in practical applications, and the development of *heuristic methods* for automatic determination of a near-optimal number of hidden neurons has been the subject of very intensive research. A variety of techniques have been proposed, which can be categorized as relying on *destructive*, *constructive* or *direct* strategies. In the *destructive* approach, the complexity of the network is gradually reduced either by penalizing complexity through addition of a bias term to the error function, or by pruning the least relevant units in the course of the training process (Reed 1993). In the *constructive* strategy, a small initial network is gradually expanded until the task is considered to be solved. Examples of such techniques are described in Fahlman and Lebiere (1990, Hirose *et al* (1991), Lengellé and Denœux (1992, 1996), Platt (1991), Lee (1992). The *direct* approach consists in using prior information, acquired through preprocessing or readily available from domain knowledge, to design a neural network that can then be further trained using a standard learning procedure such as *backpropagation* (Sethi 1990, Denœux and Lengellé 1993, Karouia *et al* 1995).

B2.10

C1.2.3

In all cases, the classification error of the classifier has to be either estimated, or derived from theoretical considerations. Raudys and Jain (1991b) discuss several methods of error estimation including the resubstitution, hold-out, cross-validation and bootstrap methods. The hold-out method consists in dividing the available data into a training set and a test set used for independent error estimation. This method provides an unbiased error estimate, but it has the disadvantage of preventing the use of all the data for the learning process. The cross-validation and bootstrap methods are more efficient, but also more computationally demanding.

As an alternative, an idea of the generalization performance of a classifier can sometimes be gained as a result of some kind of theoretical analysis. Recent investigations based on the *Vapnik–Chervonenkis theory* (Vapnik 1982) and the PAC learning model (Valiant 1984) have led to the derivation of bounds for the true and estimated classification errors of minimum empirical error classifiers (Baum and Haussler 1989, Anthony 1994, Kraaijveld 1993). However, these results are based on a worst-case analysis and lead to very pessimistic estimates of the number of samples needed to train a classifier (Raudys 1994). Nevertheless, some of the most recent results are already applicable with approximations as an initial aid to neural network design (Holden and Niranjan 1994). Further improvements are expected from the consideration of specific input distributions and training algorithms in this analysis (Kraaijveld 1993).

B3.5.2.2

F1.2.4 Alternative approaches

Since the 1950s, substantial progress has been achieved in the design of statistical classifiers from empirical data. According to Raudys and Jain (1991b), the number of classification methods already published exceeds two hundred. These methods are described in a number of standard textbooks such as Duda and Hart (1973), Fukunaga (1990) and McLachlan (1992).

A useful taxonomy of classification techniques, including statistical and neural network approaches, has been proposed by Lippmann (1994). Pattern classifiers can be seen as belonging to three main categories. *Probability density function* classifiers estimate class-conditional probability densities separately for each class. They include parametric normal classifiers with different forms of covariance matrices, and nonparametric methods of density estimation such as the Parzen-window approach. *Posterior probability classifiers* estimate the posterior probability of each class, using all the available data simultaneously. Examples of such methods are MLPs and RBF networks, and the voting k-NN rule. The third category of classification method includes techniques for directly partitioning the feature space into decision regions, using binary indicator outputs. Examples of such *boundary forming* methods are the nearest-neighbor methods, such as RCE or LVQ networks, and tree-structured classifiers (Breiman *et al* 1984).

A further distinction can be drawn between *model-based* and *data-driven* approaches. In the model-based approach, a particular classifier is chosen among a predefined family of functions, or *model*. Parametric classifiers, MLPs and LVQ classifiers fall in this category. In contrast, the form of data-driven classifiers is not fixed in advance, but determined by the data. This is the case for Parzen-window, k-NN and tree-structure classifiers, as well as for *ontogenic neural networks* that adapt their structure during the learning process.

C1.7, C2.4

This multiplicity of classification techniques obviously poses a serious problem to the practitioner. Many comparative studies have been made to assess the strengths and weaknesses of various methods

(e.g. Tsoi and Pearson 1991, Ng and Lippmann 1991, Brown *et al* 1993, Blue *et al* 1994). In general, comparable error rates are achieved by several techniques, provided they are properly tuned. As noted by Ng and Lippmann (1991), the selection of a classifier for a particular task should primarily be guided by practical considerations such as training and classification time, and memory storage requirements. Neural network classifiers usually offer a good compromise between performance and practical applicability.

References

Anthony M 1994 Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants *Technical Report* NC-TR-94-3 Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK

Baum E B and Haussler D 1989 What size net gives valid generalization *Neural Comput.* **1** 151–60

Blue J L, Candela G T, Grother P J, Chellappa R and Wilson C L 1994 Evaluation of pattern classifiers for fingerprint and OCR applications *Patt. Recog.* **27** 485–501

Breiman L, Friedman J H, Olshen R A and Stone C J 1984 *Classification and Regression Trees* (Belmont, CA: Wadsworth)

Brown D E, Corruble V and Pittard C L 1993 A comparison of decision tree classifiers with backpropagation neural networks for multimodal classification problems *Patt. Recog.* **26** 953–61

Chow C K 1970 On optimum recognition error and reject tradeoff *IEEE Trans. Inform. Theory* **16** 41–6

Cover T M and Hart P E 1967 Nearest neighbor pattern classification *IEEE Trans. Inform. Theory* **13** 21–7

Denœux T and Lengellé R 1993 Initializing back-propagation networks with prototypes *Neural Networks* **6** 351–63

Dubuisson B and Masson M 1993 A statistical decision rule with incomplete knowledge about classes *Patt. Recog.* **26** 155–65

Duda R O and Hart P E 1973 *Pattern Classification and Scene Analysis* (New York: Wiley)

Duin R P W 1994 Superlearning and neural network magic *Patt. Recog. Lett.* **15** 215–7

Fahlman S E and Lebiere C 1990 The cascade-correlation learning architecture *Advances in Neural Information Processing Systems 2* ed D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 524–32

Fukunaga K 1990 *Introduction to Statistical Pattern Recognition* 2nd edn (Berlin: Academic)

Girosi F 1994 Regularization theory, radial basis functions and networks *From Statistics to Neural Networks* ed V Cherkassky, J H Friedman and H Wechsler (Berlin: Springer) pp 166–87

Hampshire J B and Pearlmutter B 1991 Equivalence proof for multilayer perceptron networks and the Bayesian discriminant function *Connectionist Models, Proc. 1990 Summer School* ed D S Touretzky, J L Elman, T J Sejnowski and G E Hinton (San Mateo, CA: Morgan Kaufmann) pp 159–72

Hirose Y, Yamashita K and Hijiya S 1991 Back-propagation algorithm which varies the number of hidden units *Neural Networks* **4** 61–6

Holden S B and Niranjan M 1994 On the practical applicability of VC dimension bounds *Technical Report* CUED/F-INFENG/TR155 Cambridge University Engineering Department, Cambridge CB2 1PZ, UK

Hornik K 1991 Approximation capabilities of multilayer feedforward networks *Neural Networks* **4** 251–7

Karouia M, Lengellé R and Denœux T 1995 Performance analysis of a MLP weight initialization algorithm *Proc. ESANN'95 European Symp. on Artificial Neural Networks* (Brussels: De facto) pp 347–52

Kohonen T 1987 *Self Organisation and Associative Memory* 2nd edn (Berlin: Springer)

——1990 The self-organizing map *Proc. IEEE* **78** 1464–80

Kohonen T, Barna G and Chrisley R 1988 Statistical pattern recognition with neural networks: benchmarking studies *Proc. ICNN'88 Int. Conf. on Neural Networks* vol I (IEEE Computer Society Press) pp 61–8

Kraaijveld M A 1993 Small sample behavior of multi-layer feedforward network classifiers: theoretical and practical aspects *PhD Thesis* Delft University, Delft, The Netherlands

Lee D-S, Srihari S N and Gaborski R 1991 Bayesian and neural network pattern recognition: a theoretical connection and empirical results with handwritten characters *Artificial Neural networks and Statistical Pattern Recognition* ed I K Sethi and A K Jain (Amsterdam: Elsevier) pp 89–108

Lee S 1992 Supervised learning with Gaussian potentials *Neural Networks for Signal Processing* ed B Kosko (Englewood Cliffs, NJ: Prentice-Hall) pp 189–227

Lengellé R and Denœux T 1992 Optimizing multilayer networks layer per layer without back-propagation *Artificial Neural Networks II* ed I Aleksander and J Taylor (Amsterdam: North-Holland) pp 995–8

——1996 Training multilayer perceptrons layer by layer using an objective function for internal representations *Neural Networks* **9** 83–97

Lippmann R P 1994 Neural networks, Bayesian *a posteriori* probabilities, and pattern classification *From Statistics to Neural Networks* ed V Cherkassky, J H Friedman and H Wechsler (Berlin: Springer) pp 83–104

McLachlan G J 1992 *Discriminant Analysis and Statistical Pattern Recognition* (New York: Wiley)

Ng K and Lippmann R P 1991 A comparative study of the practical characteristics of neural networks and conventional pattern classifiers *Advances in Neural Information Processing Systems* vol 3 ed R L Lippman, J E Moody and D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 970–6

Platt J C 1991 Learning by combining memorization and gradient descent *Neural Information Processing 3* ed R P Lippman, J E Moody and D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 714–20

Poggio T and Girosi F 1988 A theory of networks for approximation and learning *Technical Report* AI Memo No 1140 MIT

Poirier F and Ferrieux A 1991 DVQ: dynamic vector quantization—an incremental LVQ *Artificial Neural Networks* vol 2 ed T Kohonen, M Mäkisara, O Simula and J Kangas (Amsterdam: Elsevier) pp II-1333–1336

Raudys S J 1994 Why do multilayer perceptrons have favorable small sample properties *Pattern Recognition in Practice IV* ed E S Gelsema and L N Kanal (Amsterdam: Elsevier) pp 287–98

Raudys S J and Jain A K 1991a Small sample problems in designing artificial neural networks *Artificial Neural networks and Statistical Pattern Recognition* ed I K Sethi and A K Jain (Amsterdam: Elsevier) pp 33–50

——1991b Small sample size effects in statistical pattern recognition: recommendations for practitioners *IEEE Trans. Patt. Anal. Machine Int.* **13** 252–64

Reed R 1993 Pruning algorithms: a survey *IEEE Trans. Neural Networks* **4** 740–7

Reilly D L, Cooper L N and Elbaum C 1982 A neural model of category learning *Biol. Cybern.* **45** 35–41

Rosenblatt F 1958 The perceptron: a probabilistic model for information storage and organization in the brain *Psychol. Rev.* **65** 386–408

Rumelhart D E, Hinton G E and Williams R J 1986 Learning internal representations by error propagation *Parallel Distributed Processing* ed D E Rumelhart and J McClelland (Cambridge, MA: MIT Press)

Schmidt W F 1993 Neural pattern classifying systems *PhD Thesis* Delft University, Delft, The Netherlands

Sethi I K 1990 Entropy nets: from decision trees to neural networks *Proc. IEEE* **78** 1605–13

Thomas D S and Mitiche A 1994 Asymptotic optimality of pattern recognition by regression analysis *Neural Networks* **7** 313–20

Tsoi A C and Pearson R A 1991 Comparison of three classification techniques CART C4.5 and multi-layer perceptrons *Advances in Neural Information Processing Systems* vol 3 ed R L Lippman, J E Moody and D S Touretzky (San Mateo, CA: Morgan Kaufmann) pp 963–9

Valiant L G 1984 A theory of the learnable *Commun. ACM* **27** 1134–42

Vapnik V N 1982 *Estimation of Dependences Based on Empirical Data (Springer Series in Statistics)* (Berlin: Springer)

Werbos P J 1974 Beyond regression: new tools for prediction and analysis in the behavioral sciences (Cambridge, MA: Harvard University)

——1991 Links between artificial neural networks and statistical pattern recognition *Artificial Neural networks and Statistical Pattern Recognition* ed I K Sethi and A K Jain (Amsterdam: Elsevier Science) pp 11–31

White H 1989 Learning in artificial neural networks: a statistical perspective *Neural Comput.* **1** 425–64

Widrow B and Lehr M A 1990 30 years of adaptive neural networks: perceptrons, madaline and backpropagation *Proc. IEEE* **78** 1415–42