Pattern Classification with Noisy Features

M. Pawlak and D. Siu

Department of Electrical and Computer Engineering, University of Manitoba, Canada;
Motorola Incorporation, Hong-Kong.

Abstract. The paper addresses the problem of pattern classification when distortions are present in the observed data. These are distortions occurring either in the process of classifying a test pattern or during the learning phase of a classifier design. In particular we consider the case when the measurements are corrupted by noise. Additive, multiplicative and generalized noise models are taken into account. A unified framework of the posed problems based on the Bayesian paradigm is given. Learning algorithms stemming from nonparametric curve estimation techniques are derived. We examine the effect of distortions in measurements on the proposed classification algorithms. A class of classification techniques being Bayes risk consistent is derived. The latter result makes use of the idea of deconvolution.

1 Introduction

In many practical problems one often encounters situations where the supplied data are imprecise. This can happen due to a number of reasons, such as a data collection process may be imperfect, a sensor system gathering information may be distorted, data are transmitted through noisy information channel, allocation process is based on subjective opinion of various experts, calibration between sources such as different clinics. These are typical scenarios in applications such as medical diagnosis, remote sensing and communication systems. Such imprecision in feature vectors, although often recognized, has been rarely taken into account in the actual design of classification systems. In fact, relatively little work seems to have been done for the case where the measurements are distorted. We refer to [2]-[10], [12] for some preliminary research on this subject. It is also worth noting that in many studies on artificial neural networks it is customary to add small amount of noise to the training data in order to improve the networks generalization and fault tolerance, see [13], [14] and the references cited therein. It is our task to study the accuracy of classification rules when there is an inherent noise in the measurements. In the case of additive noise model we construct a classification rule which can converge to the optimal Bayes rule when number of training data is increasing. The rate of convergence is also evaluated. Nonparametric deconvolution technique is used to account for errors-in-variables.

1.1 Statistical Classification Problem

In this section we define basic notions of the classification problem in order to laid foundation for our further studies. A pattern recognition problem can be often modeled by the statistical decision-theoretic approach utilizing the Bayesian paradigm, i.e., one wishes to classify a vector $\mathbf{x} \in R^d$ to one of c-classes knowing that the state of $\mathbf{x}$,

denoted θ, takes value in $\Omega = \{\omega_1, ..., \omega_c\}$ with probabilities $\{p(\omega_1), ..., p(\omega_c)\}$, respectively and that $\mathbf{x}$ is a realization of a random vector $\mathbf{X}$ characterized by a conditional distribution $p(\mathbf{x} \mid \theta)$, $\theta \in \Omega$. Thus, the task is to find a measurable mapping $\psi : R^d \to \Omega$ such that the expected loss function $R(\psi) = E\{L(\psi(\mathbf{X}), \theta)\}$, called risk, is minimal. Here $L(\omega_i, \omega_j)$ is the loss incurred by taking action ω_i when the class is ω_j. In this paper we assume, without loss of generality, that $L(\omega_i, \omega_j) = 0$ for $\omega_i = \omega_j$ and $L(\omega_i, \omega_j) = 1$ for $\omega_i \neq \omega_j$ and then $R(\psi) = \mathbf{P}(\psi(\mathbf{X}) \neq \theta)$ is called the probability of error. It is well known that an optimal rule ψ^* (the Bayes rule) which minimizes $R(\psi)$ is of the following form $\psi^*(\mathbf{x}) = \arg\max_{1 \le i \le c} p_i(\mathbf{x})$, where $p_i(\mathbf{x}) = \mathbf{P}(\theta = \omega_i \mid \mathbf{X} = \mathbf{x})$, $i = 1, ..., c$ are the posteriori probabilities. Let R^* denote the Bayes risk, i.e., the risk of the Bayes rule. In practice we rarely have any information about the distribution of the pair $(\theta, \mathbf{X})$, instead there is in our disposal a training set $\eta_n = \{(\theta_1, \mathbf{X}_1), ..., (\theta_n, \mathbf{X}_n)\}$, i.e., a sequence of pairs $(\theta_i, \mathbf{X}_i)$ distributed like $(\theta, \mathbf{X})$, where $\mathbf{X}_i$ is the feature vector and θ_i is its class assignment. An empirical classification rule ψ_n is a measurable function of $\mathbf{X}$ and η_n. It is natural to construct a rule which resembles the Bayes rule, i.e., by replacing $p_i(\mathbf{x})$ by its estimate $p_{in}(\mathbf{x})$. A popular nonparametric classification technique is the kernel classifier being defined as follows

$$\psi_n(\mathbf{x}) = \arg\max_{1 \le i \le c} \sum_{j=1}^{n} \mathbf{1}(\theta_j = \omega_i) W\left(\frac{\mathbf{x} - \mathbf{X}_j}{b}\right), \tag{1.1}$$

where $W(\mathbf{x})$ is a kernel function defined on R^d and b is a positive number called bandwidth. The performance of ψ_n is measured by the conditional probability of error $R(\psi_n) = \mathbf{P}(\psi_n(\mathbf{X}) \neq \theta \mid \eta_n)$. The rule ψ_n is said to be Bayes risk consistent if $ER(\psi_n) \to R^*$ as $n \to \infty$. Such an important property is known to hold, in particular, the kernel classification rules regardless of the type of the distribution of $(\theta, \mathbf{X})$, see [16], [17]. Letting the distribution of $(\theta, \mathbf{X})$ to be smooth, the rate of convergence can be also established, i.e., $ER(\psi_n) = R^* + an^{-\alpha}$, where $\alpha > 0$ depends merely on smoothness of the distribution of $(\theta, \mathbf{X})$ and the dimensionality of feature vector. An optimal value for twice differentiable class distributions is $\alpha = 2/(d+4)$.

1.2 Classification with imperfect data

The problem of imperfect measurements is concerned with the lack of availability of $(\theta, \mathbf{X})$, i.e., we rather have its distorted version $(\theta, \hat{\mathbf{X}})$. Two cases can be distinguished: first the whole feature vector $\mathbf{X}$ is distorted, i.e., the pair $(\theta, \hat{\mathbf{X}})$ is in our disposal, second only some features of $\mathbf{X}$ are distorted, i.e., we observe $\hat{\mathbf{X}} = (\mathbf{U}, \hat{\mathbf{V}})$, where $\hat{\mathbf{V}}$ is the distorted part of $\mathbf{X}$. Such an issue has been rarely discussed in the pattern recognition literature and it will be called in this paper as the errors-in features problem. It is worth noting that one can have θ distorted. Such a case is often

referred to as a probabilistic teacher. In the aforementioned situations one wishes to design an optimal classifier knowing the probability characteristics of the true observation $(\theta, \mathbf{X})$ and possibly the distortion process. This is a problem with a full probabilistic information and it is discussed in Section 2. The latter situations form a basis for more realistic situations when there are imperfections in the training set. Hence instead of the sequence η_n we have $\hat{\eta}_n = \left\{(\theta_1, \hat{\mathbf{X}}_1), ..., (\theta_n, \hat{\mathbf{X}}_n)\right\}$, where the distortion concerns only the feature vector $\mathbf{X}$. An extreme case occurs when there is some missing information either in $(\theta, \mathbf{X})$ or in the training set η_n. The absence of θ is often referred to as the unsupervised pattern classification problem. On the other hand for the problem of missing values in the training set η_n we refer to [15] and the references cited therein.

2 Classification of Distorted Pattern Vector

It is often met in practice that instead of the true feature vector $\mathbf{X}$ we can only observe its an inaccurate version $\hat{\mathbf{X}}$. It is clear that a rule minimizing the probability of error $\mathbf{P}(\psi(\hat{\mathbf{X}}) \neq \theta)$ is given by $\hat{\psi}(\hat{\mathbf{x}}) = \arg\max_{1 \leq i \leq c} q_i(\hat{\mathbf{x}})$, where $q_i(\hat{\mathbf{x}})$ is the posteriori probability of $(\theta, \hat{\mathbf{X}})$. Let us assume that θ and $\hat{\mathbf{X}}$ are conditionally independent given $\mathbf{X}$. Under such a condition it is straightforward to show that

$$q_i(\hat{\mathbf{x}}) = \int_{R^d} p_i(\mathbf{x}) p(\mathbf{x} \mid \hat{\mathbf{x}}) d\mathbf{x} \quad , \tag{2.1}$$

where $p(\mathbf{x} \mid \hat{\mathbf{x}})$ is the conditional distribution of $\mathbf{X}$ on $\hat{\mathbf{X}}$.

An important example is when an additive noise is corrupting $\mathbf{X}$, i.e., $\hat{\mathbf{X}} = \mathbf{X} + \varepsilon$, where ε is a random variable being independent of $\mathbf{X}$. In such a case (2.1) takes the following form $q_i(\hat{\mathbf{x}}) = \int_{R^d} p_i(\mathbf{x}) p_\varepsilon(\hat{\mathbf{x}} - \mathbf{x}) d\mathbf{x}$, where p_ε is a density of ε. Hence for the additive noise model the relationship between the class distributions of $\hat{\mathbf{X}}$ and $\mathbf{X}$ is in the form of the convolution operator. In [3] a Gaussian parametric model for $p_i(\mathbf{x})$ and $p(\mathbf{x} \mid \hat{\mathbf{x}})$ in (2.1) has been used yielding an explicit formula for $q_i(\hat{\mathbf{x}})$. In [5]-[7] further accuracy issues under the additive noise distortion model have been examined. In particular it has been shown that $R^* \leq R^*(\hat{\psi}) \leq \hat{R}^*$, where $R^*(\hat{\psi}) = \mathbf{P}(\hat{\psi}(\hat{\mathbf{X}}) \neq \theta)$ is the Bayes risk of corresponding to $\left(\hat{\mathbf{X}}, \theta\right)$ and $\hat{R}^* = \mathbf{P}(\psi^*(\hat{X}) \neq \theta)$ measures the performance of the Bayes rule on noisy feature vectors. The following result [9] concerns the performance of the rule $\hat{\psi}$ tested on noise-free data.

Theorem 1. *Let* $p_i(\mathbf{x})$, $i = 1, ..., c$ *satisfy the Lipschitz condition*

$$| p_i(\mathbf{x}) - p_i(\mathbf{x}_0) | \leq M_i \| \mathbf{x} - \mathbf{x}_0 \|,$$

for some $M_i > 0$, $i = 1, ..., c$.

Let us assume the additive noise distortion model with the noise process being

identically distributed with $\mu_p = E\|\varepsilon\|^p < \infty$, $p > 0$. *Then*

$$R(\hat{\psi}) - R^* \le (p+1)\left((M/p)^p \mu_p\right)^{1/(p+1)},$$

where $R(\hat{\psi}) = \mathbf{P}(\hat{\psi}(\mathbf{X}) \neq \theta)$ *and* $M = \max_{1 \le i \le c} M_i$.

This result reveals that faster the tails of p_ε tend to zero more accurate is $\hat{\psi}$.

A general distortion model of the form $\tilde{\mathbf{X}} = \mathbf{T}(\mathbf{X}, \varepsilon)$ can also be taken in consideration with ε being a random variable independent from $(\theta, \mathbf{X})$ and $\mathbf{T}$ is a transform defined on the domain of the random pair $(\mathbf{X}, \varepsilon)$. Furthermore the above result can be extended to partial distortion models.

Yet another important extension is the case when no probabilistic information about the class distributions is given but instead the noise-free training set η_n is available. On the other hand, only distorted version $\tilde{\mathbf{X}}$ of $\mathbf{X}$ is observed. In this case one can use the aforementioned classification rules with the estimated class distributions. For instance the kernel classification rule for the additive noise distortion model takes the following form

$$\hat{\psi}_n(\tilde{\mathbf{x}}) = \arg\max_{1 \le i \le c} \sum_{j=1}^{n} \mathbf{1}(\theta_j = \omega_i) W\left(\frac{\mathbf{x} - \mathbf{X}_j}{b}\right) p_\varepsilon(\tilde{\mathbf{x}} - \mathbf{X}_j).$$

It can be proved that this rule has a limit risk smaller than the standard plug-in rule for classifying the distorted pattern $\tilde{\mathbf{x}}$.

3 Classification with Distorted Training Set

Clearly the most challenging problem arises when there are some imperfections in the learning set. This is issue is addressed in this section. We confine our discussion to kernel classification rules although it seems to be possible to extend our results to other nonparametric classification rules. Furthermore we assume that the pattern $\mathbf{X}$ is to be classified is noise free. Hence one wishes to classify the pattern $\mathbf{X}$ based on the distorted training set $\tilde{\eta}_n = \left\{(\theta_1, \tilde{\mathbf{X}}_1), \ldots, (\theta_n, \tilde{\mathbf{X}}_n)\right\}$, where $\tilde{\mathbf{X}}_i$ is the distorted version of $\mathbf{X}_i$. This data can be used to estimate the conditional probabilities $q_i(\mathbf{x})$ in (2.1). This yields the following version of the kernel classification rule:

$$\hat{\psi}_n(\mathbf{x}) = \arg\max_{1 \le i \le c} \sum_{s=1}^{n} \mathbf{1}(\theta_s = \omega_i) W\left(\frac{\mathbf{x} - \tilde{\mathbf{X}}_s}{b}\right) \tag{3.1}$$

It can be shown that the probability of error $\mathbf{P}\left(\hat{\psi}_n(\mathbf{X}) \neq \theta \mid \tilde{\eta}_n\right)$ of $\hat{\psi}_n$ tends to $\mathbf{P}\left(\hat{\psi}(\mathbf{X}) \neq \theta\right) = R(\hat{\psi})$ as $n \to \infty$, in probability where $\hat{\psi}$ is defined as follows $\hat{\psi}(\mathbf{x}) = \arg\max_{1 \le i \le c} q_i(\mathbf{x})$. The result of Theorem 1 in Section 2 reveals that $R(\hat{\psi}) \ge R^*$, i.e., the rule defined in (3.1) is not Bayes risk consistent.

To illustrate this phenomenon let us consider the two-class classification problem with equal prior probabilities and class-conditional densities being Gaussian, i.e., $p(\mathbf{x} \mid \omega_1) \propto N_d((1,...1)^T; \mathbf{I}))$, $p(\mathbf{x} \mid \omega_2) \propto N_d((3,...3)^T; \mathbf{I})$. The additive distortion model is used, i.e., $\tilde{\mathbf{X}} = \mathbf{X} + \varepsilon$, where $\varepsilon \propto N_d((0,...,0)^T; \tau^2\mathbf{I})$. Figure 1 shows the classification boundary (along with testing samples) of the kernel classifier ($W(\mathbf{x})$ is assumed to be the Gaussian kernel and b is selected optimally) for $d = 2$ with τ^2=0 (no noise) and τ^2 =1. The training set of the size $n = 200$ was used. Figure 2 plots the classifier risk as a function of the noise variance τ^2. The deterioration of the classifier accuracy is clearly revealed.

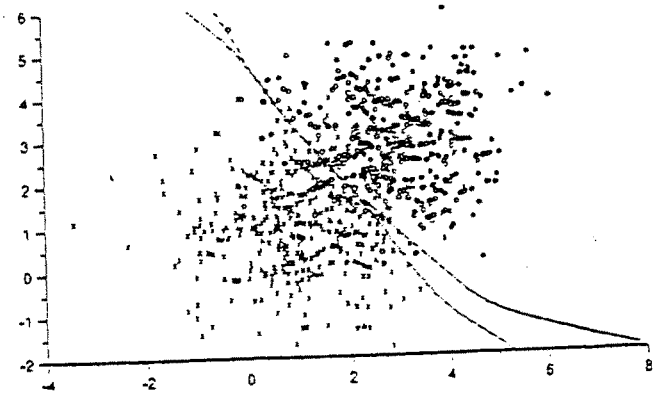

Fig. 1. Classification boundary of the kernel classifier for noise-free and noisy data.

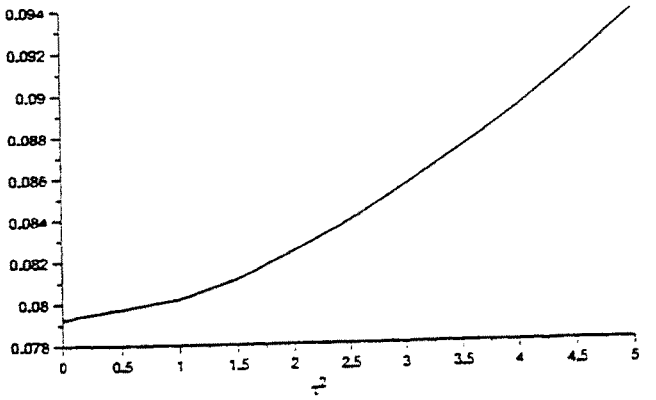

Fig.2. The probability of error versus τ^2.

The problem which emerges now is how to construct a learning rule from the distorted data which can be Bayes risk consistent. This requires the inversion of the integral equation given in (2.1), provided that the noise characteristic (quantified by the conditional distribution of $\tilde{\mathbf{X}}$ on $\mathbf{X}$) is known. In the case of the additive noise model this is equivalent to the problem of deconvolution. The Fourier transform technique can be used here to design a classifier with the required property [9], [10]. Hence, let $\phi_W(\mathbf{t})$ and $\phi_\varepsilon(\mathbf{t})$ be the Fourier transforms of the kernel function $W(\mathbf{x})$ and the probability density p_ε, respectively. Let us also define the following kernel function:

$$\tilde{W}_b(\mathbf{x}) = (2\pi)^{-d} \int_{R^d} \exp(-j\mathbf{t} \bullet \mathbf{x}) \frac{\phi_W(\mathbf{t})}{\phi_\varepsilon(\mathbf{t})} d\mathbf{x}, \tag{3.2}$$

where $\mathbf{t} \bullet \mathbf{x}$ is the scalar product of the vectors $\mathbf{t}$ and $\mathbf{x}$. Note that $\tilde{W}_b(\mathbf{x}) = W(\mathbf{x})$ when there is no noise present in the measurements.
The following kernel classifier can be defined in the case of the additive noise model:

$$\tilde{\psi}_n(\mathbf{x}) = \arg\max_{1 \le i \le c} \sum_{s=1}^{n} \mathbf{1}(\theta_s = \omega_i) \tilde{W}_b\left(\frac{\mathbf{x} - \tilde{\mathbf{X}}_s}{b}\right). \tag{3.3}$$

In order to see the potential benefits of this new rule let us plot its classification error and the error of the rule $\hat{\psi}_n$ in (3.1) which ignores the noise problem. Figure 3 plots the error versus n for the data as in Figure 1. It is seen that $\tilde{\psi}_n$ greatly outperforms $\hat{\psi}_n$.

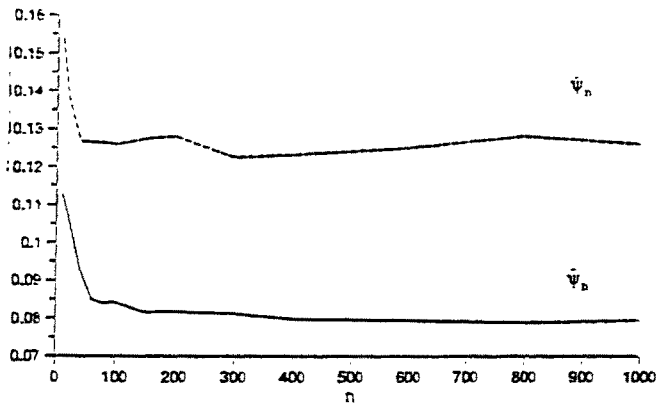

Fig.3. The probability of error for the classifiers $\hat{\psi}_n$ and $\tilde{\psi}_n$ versus n.

As far as the asymptotic properties of the rule in (3.3) the following result can be proved.

Theorem 2. *Assume that* $\phi_\varepsilon(\mathbf{t})$ *satisfies*

(a) $\| \mathbf{t} \|^{\gamma(d)} | \phi_\varepsilon(\mathbf{t}) | \to a(d)$ *as* $\| \mathbf{t} \| \to \infty$ *for some* $\gamma(d) > 0$ *and* $a(d) > 0$.

Assume that the kernel function $W(\mathbf{t})$ *satisfies*

(b) $\int_{R^d} \| \mathbf{t} \|^{2\gamma(d)} | \phi_W(\mathbf{t}) |^2 d\mathbf{t} < \infty$

(c) $\int_{R^d} \| \mathbf{t} \|^2 \, W(\mathbf{t}) \, d\mathbf{t} < \infty$ *and* $\int_{R^d} t_j W(\mathbf{t}) \, d\mathbf{t} = 0, \; j = 1, \ldots, d.$

Let $p_j(\mathbf{x}), \; j = 1, \ldots, c$ *and* $p(\mathbf{x})$ *be twice differentiable.*
If

$$b \approx n^{-1/(4+d+2\gamma(d))}$$

then

$$E\left\{ \mathbf{P}(\tilde{\psi}_n(\mathbf{X}) \neq \theta \mid \hat{\eta}_n) \right\} = R^* + cn^{-2/(4+d+2\gamma(d))}, \tag{3.4}$$

for some $c > 0$.

Hence the rule $\tilde{\psi}_n$ utilizing the deconvolution approach is able to converge to the Bayes classifier with a certain rate depending on the smoothness of p_ε. In fact, the assumption (a) in Theorem 3 describes the smoothness of the density p_ε of the noise process, i.e., larger values of $\gamma(d)$ correspond to smoother p_ε. Consequently the rate of convergence is slower for smooth densities of the noise process. This is due to the fact that the deconvolution with smooth measurement error is intrinsically difficult. Furthermore for $\gamma(d) = 0$ we can recover the rate $O\left(n^{-2/(4+d)}\right)$ for the noise-free data case.
The case of normal measurement error is particularly important. It is known that in this case $\phi_\varepsilon(\mathbf{t})$ has an exponential decay as $\| t \| \to \infty$, i.e., the condition (a) of Theorem 2 does not hold. Nevertheless, it can be shown that in this case we have

$$E\left\{ \mathbf{P}(\tilde{\psi}_n(\mathbf{X}) \neq \theta \mid \hat{\eta}_n) \right\} = R^* + c(\log(n))^{-1}. \tag{3.5}$$

Let us note that this very slow rate is independent of the dimensionality of the feature vector.
It is important to note, however, that all the aforementioned results have been obtained from the upper bounds. One can conjecture that optimal rates can be much better than those given in (3.4) and (3.5).

5 Concluding Remarks

The classification problem in the presence of distorted data has been presented. It has been observed that efficient solutions in this situation lead to the inverse problem, i.e., one wishes to infer about the true model from the distorted one. Suggestions have

been made how to design nonparametric classifiers from imperfect data. A number of issues remain open, e.g., optimal convergence rates of the proposed classification schemes, small sample properties, the choice of classifiers parameters. Furthermore, an important problem of assessing the performance of a classifier derived from distorted data remains to be addressed.

References

1.W.Greblicki, Learning to recognize patterns with a probabilistic teacher, *Pattern Recognition*, **12**, pp.159-164, 1980.

2. T.Farago and L.Gyorfi, On the continuity of the error distortion function for multiple-hypothesis decisions, *IEEE Trans. Inform. Theory*, **21**, pp. 458-460, 1975.

3. J.Aitchison and I.J.Lauder, Statistical diagnosis from imprecise data, *Biometrika*, **66**, pp.475-483, 1979.

4. W.H.Tsai and K.S. Fu, A pattern deformational model and Bayes error-correcting recognition system, *IEEE Trans. on Systems, Man, and Cybernetics*, **12**, pp.745-756, 1979.

5. B.B.Chaudhuri, Bayes' error and its sensitivity in statistical pattern recognition in noisy environment, *International Jour.System Sci.*, **13**, pp.559-570, 1982.

6. B.B.Chaudhuri, C.A.Murthy and D.Duttamajumder, Bayes' error probability for noisy and imprecise measurement in pattern recognition, *IEEE Trans. on Systems*, Man, and Cybernetics, **13**, pp.89-94,1983.

7. S.D.Gupta, Pattern recognition when feature variables are subject to error, *Sankhya*, Ser.B, **51**, pp.287-294, 1989.

8. G.Lugosi, Pattern classification from distorted sample, *Problems of Control and Information Theory*, **20**, pp.1-9, 1991.

9. M.Pawlak, Pattern classification and deconvolution problems, *Technical Report* , 1998.

10. M.Pawlak, Learning with noisy patterns; discrete feature case, *Technical Report* , 1998.

11. J.Fan, On the optimal rates of convergence for nonparametric deconvolution problems, *Ann. Statistics*, **19**, pp.1257-1272, 1991.

12. A.R.Webb, Functional approximation by feedforward networks: a least-squares approach to generalization, *IEEE Trans. on Neural Networks*, vol. 5, pp.363-371, 1994.

13. R.Reed, R.J.Marks and S.Oh, Similarities of error regularization, sigmoid gain scaling, smoothing, and training with jitter, *IEEE Trans. on Neural Networks*, vol. 6, pp.529-538, 1995.

14. L.Holmstrom and P.Koistinien, Using additive noise in backpropagation training, *IEEE Trans. on Neural Networks*, vol. 3, pp.24-381, 1992.

15. M.Pawlak, Kernel classification rules from missing data, *IEEE Trans. Inform. Theory*, **39**, pp. 979-988, 1993.

16. L. Devroye, L. Gyorfi and G.Lugosi, *A Probabilistic Theory of Pattern Recognition*, Wiley, 1996.

17. W. Greblicki and M. Pawlak, Necessary and sufficient conditions for Bayes risk consistency of a recursive kernel classification rule, *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 408-412, 1987.