# Advanced Pattern Recognition Techniques for System Monitoring and Diagnosis: A Survey

**Thierry Denœux — Mylène Masson — Bernard Dubuisson**

*Heuristique et Diagnostic des Systèmes Complexes*
*UMR CNRS 6599*
*Université de Technologie de Compiègne*
*BP 20529 - F-60205 Compiègne cedex - France*

RÉSUMÉ. *L'approche par reconnaissance de formes des problèmes de surveillance de systèmes complexes se base sur la connaissance d'un ensemble de mesures effectuées sur le système, et des états de fonctionnement associés. Cette base d'exemples permet de construire une fonction d'un espace de caractéristiques dans un espace de décision, de façon à minimiser le risque de mauvaise classification. Cet article propose une revue des principales techniques de reconnaissance de formes applicables à ce type de problèmes. Les techniques classiques de discrimination sont en général insuffisantes car (1) elles supposent une connaissance a priori de tous les états de fonctionnement, et (2) elle ne prennent pas en compte l'évolution du système. Certains travaux récents ont conduit à proposer de nouvelles méthodes reposant sur des théories non standard de traitement de l'incertitude telles que la logique floue ou la théorie des fonctions de croyance. Ces approches ont permis le développement de règles de décision à la fois plus robustes, adaptatives et prédictives.*

ABSTRACT. *In the feature-based approach to system monitoring and diagnosis, knowledge about the system is assumed to consist exclusively in a database of measurement vectors and associated operating conditions. This data is used to build a mapping from the measurement space onto a decision space, in such a way that the probability of misclassification (or assignment to a wrong state) is minimised. In this paper, the main pattern recognition techniques applicable to this problem are reviewed. Standard statistical techniques are generally not sufficient because (1) they assume a priori knowledge of all system states, and (2) they do not take into account the time evolution of the process under study. Some recent approaches are based on non-standard theories of uncertainty such as fuzzy logic and evidence theory. Specific techniques allow us to make classifiers (1) more robust by taking into account past decisions to establish a diagnostic, (2) adaptive by including incremental procedures for parameter learning and detection of new classes, and (3) predictive by anticipating the evolution of the system.*

MOTS-CLÉS : *Diagnostic, surveillance de systèmes, reconnaissance de formes, discrimination, gestion de l'incertitude, analyse de la décision.*

KEY WORDS : *Fault diagnosis, system monitoring, pattern recognition, classification, uncertainty management, decision analysis.*

## 1. Introduction

In many application areas, it has become increasingly important to monitor the behaviour of complex systems based on multiple measurements. Although the need for such an analysis was first recognised independently in specific domains such as industrial and biomedical engineering [55, 24], it is now well understood that similar problems also arise in other fields such as environmental engineering [73], requiring a unified approach. In general, the task of a monitoring or diagnosis system consists in detecting the departure of a process from normal conditions, to characterise the new process state, and to prescribe appropriate actions.

Diagnosis techniques are usually classified into two main categories. In the model-based approach, sensor signals are considered as the outputs of a dynamic system. Process monitoring is then conducted based on system modelling and validation [40, 2, 54, 27]. Typically, analytical relationships between measurable variables of the system are derived and used to compute residuals which provide the basis for decision-making. Although this approach has proved very effective in many applications, it has two major shortcomings [20]. Firstly, complex technological or natural processes are generally non-linear time-varying systems, which makes it particularly difficult to detect structural changes in the system. Secondly, the available model is often assumed to represent normal operating conditions, and the impact of a departure from these conditions on the model outputs is difficult to predict.

In the feature-based or Pattern Recognition approach, no mathematical model of the process under study is needed [23, 70, 51]. Knowledge about the system is assumed to consist exclusively in a learning set of measurement vectors and associated operating conditions. This data is used to build a mapping from the measurement space onto a decision space, in such a way that the probability of misclassification (or assignment to a wrong state) is minimised. The pattern recognition methodology is usually divided into two stages: feature extraction and classification. Feature extraction consists in finding a parsimonious but informative representation of the process, based on raw measurements. Usually, a large set of candidate features is first computed using signal or image processing techniques, and multivariate statistical procedures are used for selecting a subset of these features or combinations thereof [23, 33, 49]. Once a suitable representation space has been defined, the next step is then to partition this space into decision regions corresponding to assignment to each of the known states, or pattern rejection.

In the sequel, a brief summary of the main classical approaches to statistical pattern recognition is first presented. The central thesis of this paper is that these techniques are not fully adequate for solving some of the most difficult problems encountered in process monitoring and diagnosis applications, because (1) they assume a priori knowledge of all system states, and (2) they do not take into account the time evolution of the process under study. The rest of the paper reviews some recent developments which have attempted to enrich the pattern recognition methodology so as to make it more suitable to diagnosis applications.

## 2. Probabilistic Approach

### 2.1. *Bayes Decision Theory*

In the statistical approach to pattern recognition, we consider a finite number of states of nature or classes $\omega_1, \ldots, \omega_M$. Measurement vectors are assumed to arise from some form of random experiment and are modeled by a random vector $\mathbf{X}$. The probability density of $\mathbf{X}$ in class $\omega_i$ is $f(\mathbf{x}|\omega_i)$, and each state $\omega_i$ appears with prior probability $P(\omega_i)$. The mixture density of $\mathbf{X}$ is then:

$$f(\mathbf{x}) = \sum_{i=1}^{M} P(\omega_i) f(\mathbf{x}|\omega_i). \quad [1]$$

Having observed a realisation $\mathbf{x}$ of $\mathbf{X}$, the posterior probability $P(\omega_i|\mathbf{x})$ can be computed by applying the Bayes theorem:

$$P(\omega_i|\mathbf{x}) = \frac{f(\mathbf{x}|\omega_i)P(\omega_i)}{f(\mathbf{x})}. \quad [2]$$

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. Action $\alpha_i$ is often interpreted as the decision of allocating $\mathbf{x}$ to class $\omega_i$. If, as a result of observing pattern $\mathbf{x}$, we take action $\alpha_i$ while the system under consideration is in state $\omega_j$, we incur a loss $\lambda(\alpha_i|\omega_j)$. The expected loss $R(\alpha_i|\mathbf{x})$ is:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{M} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}). \quad [3]$$

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

$$R(\alpha) = \int_{\mathbb{R}^d} R(\alpha(\mathbf{x})|\mathbf{x}) f(\mathbf{x}) d\mathbf{x} \ . \quad [4]$$

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

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

$$R(\alpha_i|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x}), \quad [5]$$

and the overall risk is the average probability of misclassification. Consequently, the Bayes rule consists in that case in selecting the class with the highest posterior probability. This rule has optimal classification performance in the sense that it minimises 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 of $N$ patterns with known classification is available. The construction of allocation rules based on a limited amount of training data is one of the fundamental problems in statistical pattern recognition. The main approaches are briefly reviewed in the following section.

## 2.2. *Statistical Learning Methods*

Since the 1950s, substantial progress has been achieved in the design of statistical classifiers from empirical data. According to [59], the number of classification methods already published exceeds two hundred. These methods are described in a number of standard textbooks such as [26], [33] and [49].

A useful taxonomy of classification techniques, including statistical and neural network approaches, has been proposed by Lippmann [47]. 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 non parametric methods of density estimation such as the Parzen window approach. Posterior probability classifiers estimate the posterior probability of each class, using simultaneously all the available data. Examples of such methods are the voting $k$-nearest neighbour rule as well as neural network techniques such as multilayer perceptrons [62] and radial basis function networks [58, 35]. From a statistical point of view, these learning procedures essentially perform parameter estimation using either the maximum likelihood or the Bayesian approach [6, 61]. In each case, they consider the training data as an independent, identically distributed sample taken from a well-defined joint probability distribution, an assumption which is not always valid in diagnosis applications. Measurement vectors taken at successive time steps are usually not independent, and the statistical characteristics of each class may change gradually due to system evolution.

The third category of classification methods includes techniques for directly partitioning the feature space into decision regions, using binary indicator outputs. Examples of such boundary forming methods are distance-based neural network classifiers such as Restricted Coulomb Energy [60] or Learning Vector Quantization networks [42, 43], linear classifiers such as the ordinary perceptron, and tree-structured classifiers [9]. A comprehensive survey of neural network models applicable to fault diagnosis has been presented in [70, 69]. An overview of neural network based pattern classifiers is given in [14].

A further distinction can be drawn between top-down and bottom-up approaches. In the top-down (or model-based) approach, a particular classifier is chosen among a predefined family of functions. Parametric classifiers, multilayer perceptrons and Learning Vector Quantization classifiers fall in this category. On the contrary, the form of bottom-up (or data-driven) classifiers is not fixed in advance, but determined by the data. This is the case for Parzen-Window, $k$-NN and tree-structures classifiers, as well as for ontogenic neural networks that adapt their structure during the learning process [44].

## 3. Specific Requirements of Diagnosis Applications

The classical Pattern Recognition methodology summarised here has been successfully applied to a variety of diagnosis problems. Some recent examples include supervision of diesel engine injection [46], predictive diagnosis in the exhaust and discharge systems of reciprocating machines [1], rotating machinery fault diagnosis [72], machining condition monitoring in tapping [21], etc.

In general, conditions for applying such a methodology are (1) the possibility to enumerate in advance all possible states of the system under study, and (2) the availability of data samples corresponding to each of these states. However, as outlined by Dubuisson and Masson [25], there exists a number of particularly complex applications in which one does not know the number of classes or cannot do any recording for some classes. Typically, when monitoring, e.g., the behaviour of a human operator performing a task [56] or the evolution of an environmental system [73, 74], one does not have any precise a priori knowledge of the possible "states" or "classes of behaviour" of the system, which additionally may depend heavily on experimental conditions.

An illustrative example of such a situation was presented by Peltier and Dubuisson in the context of driving fatigue detection [56]. The aim of this study was to monitor the behaviour of a car driver as perceived from his driving activity, in order to ensure early detection of undesirable physiological states such as hypovigilance or drowsiness. For that purpose, a car was equipped with various of sensors providing information from the man-machine interface, and a pattern vector was computed using relevant features extracted from the signal sensors. Physiological signals (electro-encephalogram, electro-oculogram) were used *a posteriori* to label the pattern vectors. A series of experiments was conducted with various drivers in real-world conditions. In this application, no general-purpose training set could be constructed, because driving behaviour is highly influenced by subjective factors (driving experience, physiological status) as well as external factors (traffic density, weather conditions, etc.). Consequently, a learning set had to be constructed from scratch during each trip. The initial class was considered as the reference state, and training samples representative of new states were incrementally added to the database. Note that, in this application, dangerous states were systematically avoided by warning the driver, and consequently could never be represented in the training set. Another domain in which a similar problem is frequently encountered is the nuclear industry [23]: some faults are known but are extremely difficult to study experimentally for safety reasons, which precludes the creation of a complete training set. Hence, a pattern recognition approach applied to such problems should have the ability to detect novelty in the inputs, a property that is generally absent from conventional pattern classifiers.

Another limitation of general pattern recognition techniques when applied to diagnosis applications is that they fail to take into account the dynamic behaviour of the process under study. In process monitoring problems, gradual transitions from one state to another are frequently observed. As an example, let us consider the problem of continuous monitoring of tool wear and surface finish in milling as described by Zieba and Dubuisson [76]. The purpose of this study was to avoid fatal damage of the machine in case of tool breakage, optimise tool life, and control the quality of products (surface finish and dimensions). In the experiments described in [76], three accelerometers were installed on a cutting machine, and a series of cutting tests were conducted in industrial conditions. Each experiment started with a new tool and was stopped either at the end of tool life or when a degradation of the surface finish was observed. The pattern vector was composed of several parameters extracted from vibration signals using either time or frequency domain analysis. Here again, the first observations corresponding to a new tool were considered as prototypes of a reference class for which a fuzzy membership function was defined (see Section 5). The evolution of the membership degrees computed for subsequent patterns was then monitored using sequential testing procedures, which allowed the early detection of abrupt changes corresponding to surface finish deterioration. Another example of application in which the analysis of system evolution plays a central role is the diagnosis of telephone networks [8]. In the study presented by Boutleux and Dubuisson [8], a fuzzy system approach was used to model the evolution paths between classes in the parameter space, which allowed to follow gradual transitions from the normal state to abnormal ones corresponding to different kinds of network overload conditions.

Generally speaking, analyzing the short term evolution of feature vectors in measurement space may help to increase the robustness of the classifier while making it predictive. It is also important to keep track of the longer-term evolution of the process so as to carry out on-line adaptation of the system to slow changes in the process as well as to the occurrence of new states.

In the following, certain recent endeavours to make the pattern recognition approach applicable to a wider range of diagnosis problems are reviewed. For clarity of presentation, these developments have been classified in three broad categories. In the first category of methods, the probabilistic model underlying the statistical pattern recognition methodology is still considered to be valid, but it is adapted to account for the existence of unknown system states (Section 4). In the second category, the probabilistic formalism for dealing with uncertainty is replaced by another theoretical framework such as fuzzy sets (Section 5) or Dempster-Shafer theory (Section 6). Lastly, the third group of methods is concerned with the introduction of time in the decision process (Section 7).

## 4. Detection of New Classes

### 4.1. *Distance Rejection*

The reject option introduced by Chow [10] consists in postponing decision-making when several classes appear to be almost equally likely, in which case the expected loss of assigning the pattern to a class is higher than some threshold. This situation is easily handled within the framework of Bayes decision theory by considering an action $\alpha_0$ with constant loss $\lambda_0$ for all classes, i.e., $\lambda(\alpha_0|\omega_i) = \lambda_0$, for all $i \in \{1, \ldots, M\}$. If the costs of correct classification and misclassification are set to 0 and 1, respectively, then rejection is decided for pattern $\mathbf{x}$ if

$$\max_i P(\omega_i|\mathbf{x}) < 1 - \lambda_0. \quad [6]$$

To account for the possible existence of new classes, Dubuisson [23, 25] has introduced a second kind of rejection called "distance" rejection, as opposed to the "ambiguity" rejection described above. The idea is to avoid association to one the $M$ known classes of a vector situated in a region of low probability density. A criterion for distance rejection is therefore:

$$f(\mathbf{x}) < C_d, \quad [7]$$

where as before, $f(\mathbf{x})$ denotes the mixture probability density at $\mathbf{x}$ and $C_d$ is a user-defined constant. In general, the true density is unknown and is estimated using certain parametric or non-parametric methods. If the $k$ nearest neighbour method is used, then a distance reject criterion can be based, e.g., on the mean distance between pattern $\mathbf{x}$ and its $k$ nearest neighbors in the training set. In [45], Lengellé et al. have proposed a hybrid model based on a multilayer perceptron for pattern classification and an unsupervised neural network performing density estimation for distance rejection.

Note that, in this form, distance reject does not fit into the framework of Bayes theory in which decisions are based on posterior probabilities only. Rather, the method may be interpreted as a way to detect the inadequacy of the current model involving $M$ classes to the observed data. If too many patterns are found in regions of low probability density (in the current model), then there is strong evidence that the system is in a new state and the model must be updated.

### 4.2. *Bayes Analysis*

A different approach based on Bayes decision analysis has been proposed by Smyth [66] (see also an interesting discussion on this issue in [61, p. 24]). This approach extends the classical Bayesian analysis presented in Section 2.1 by explicitly introducing an $M + 1$th state $\omega_{M+1}$ to cover all possible other states not accounted for in the set $K = \{\omega_1, \ldots, \omega_M\}$ of known states.

The posterior probability of known state $\omega_i, i = 1, \ldots, M$ is computed as:

$$P(\omega_i|\mathbf{x}) = P(\omega_i|\mathbf{x}, K)P(K|\mathbf{x}), \quad [8]$$

where $P(\omega_i|\mathbf{x}, K)$ denotes the probability that the system described by measurement vector $\mathbf{x}$ is in state $\omega_i$, given that it is in one of the known states. This probability may be estimated by a posterior probability classifier, or discriminative model, trained on the available data. Simultaneously, the second term $P(K|\mathbf{x})$ is obtained using the Bayes rule as:

$$P(K|\mathbf{x}) = \frac{f(\mathbf{x}|K)P(K)}{f(\mathbf{x}|\omega_{M+1})P(\omega_{M+1}) + f(\mathbf{x}|K)\sum_{j=1}^{M} P(\omega_j)}, \quad [9]$$

and $P(\omega_{M+1}|\mathbf{x}) = 1 - P(K|\mathbf{x})$. The term $f(\mathbf{x}|K)$ is provided by a generative model or density estimation method such as a Gaussian mixture model or a non-parametric kernel estimator. Two other terms cannot be estimated and need to be fixed by the designer based on his or her state of knowledge concerning the system: the prior probability $P(\omega_{M+1}) = 1 - P(K)$ of being in the unknown state, and the density $f(\mathbf{x}|\omega_{M+1})$ of the observable data given the unknown state. For the latter, a uniform density over a bounded space of feature values may be chosen to reflect the designer's ignorance of the true characteristics of the unknown states.

This method has been applied to on-line failure detection in antenna pointing systems [66].

### 4.3. *Example*

In this section, the decision region obtained by the different decision rules described above are compared on a data set coming from an environmental monitoring application [73, 19]. The data consist in daily measurements of water quality parameters (pH, conductivity, NO3 and NH4 concentrations) performed in the river Seine during two years. The sampling point was located upstream a drinking water production plant. A clustering procedure was used to partition the data into four classes corresponding to distinct states of the river system. A monitoring system under development will process continuous measurements of the same parameters to produce in real-time a diagnosis of water quality. For a better visualisation of the results, the data dimension was reduced to 2 by principle component analysis.

The posterior probabilities of each of the four classes were estimated using a parametric classifier with assumption of normality of the four classes (plug-in quadratic classifier). The maximum posterior probabilities are displayed as grey levels in Figures 1 to 4, together with the decision boundaries induced by different variants of the Bayes rule with 0/1 losses. Figure 1 corresponds to the simple rule of assignment to the class of maximum a posteriori probability. As can be seen from this figure, the decision boundaries are intuitively correct in the region where training patterns are available, but they become far less reliable outside this region. Chow's ambiguity rejection (Eq. 6) and distance rejection (Eq. 7) are demonstrated in Figures 2 and 3, respectively. Clearly, Chow's rule rejects "ambiguous" patterns situated in the vicinity of class boundaries, but does not improve the performance of the classifier in regions of low probability density, a task properly handled by the distance reject option (Figure 3). Results obtained with the Bayesian approach described in Section 4.2 (with assumption of uni-

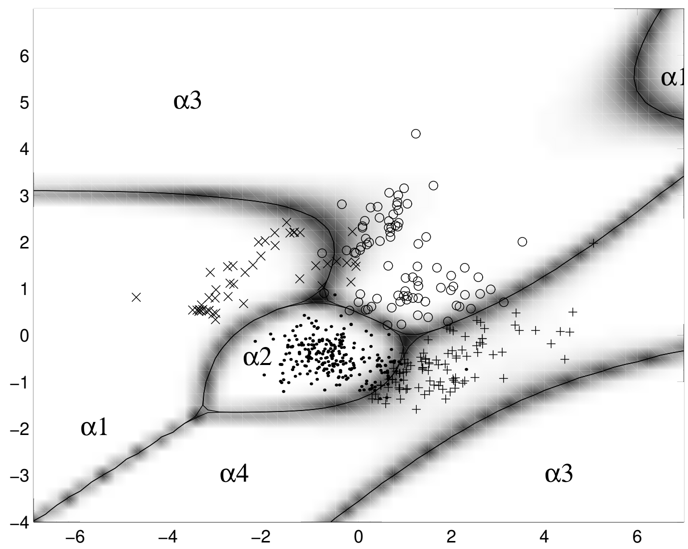


**Figure 1.** *Maximum a posteriori probabilities (grey scales) and decision regions for the Bayes rule with 0/1 losses.*

form probability density for the unknown class) are depicted in Figure 4. The decision regions yielded by this rule are very close to those obtained by combining ambiguity

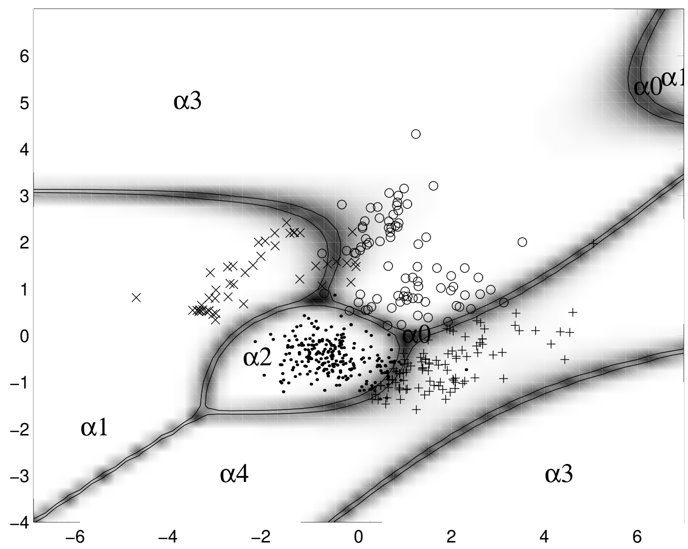


**Figure 2.** *Maximum a posteriori probabilities (grey scales) and decision regions for the Bayes rule with 0/1 losses and Chow's reject option.*

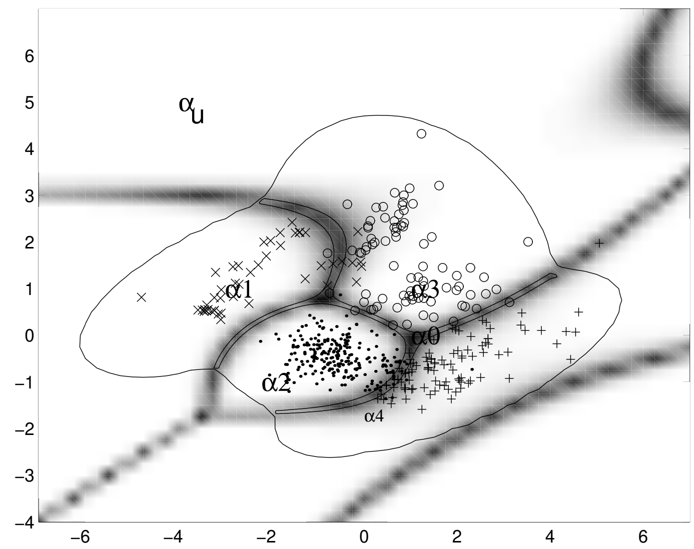


**Figure 3.** *Maximum a posteriori probabilities (grey scales) and decision regions for the Bayes rule with 0/1 losses and distance rejection.*

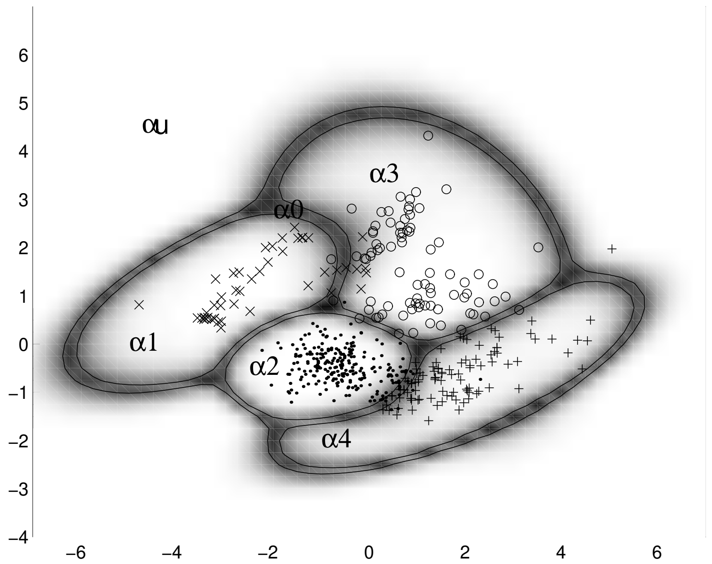


**Figure 4.** *Maximum a posteriori probabilities (grey scales) and decision regions for the Bayes rule with 0/1 losses, taking into account an unknown class.*

and distance rejection (Figure 3), except that the regions of assignment to one of the known classes are now "surrounded" by a region of ambiguity rejection, which corresponds to the situation where the typicality of a measurement vector is dubious. The Bayesian approach thus appears intuitively appealing and, to some extent, more principled than the distance rejection rule. However, the necessity to describe the unknown class in terms of a bounded set of possible feature values and prior densities may be seen as a drawback of this approach, both from theoretical and practical viewpoints.

## 5. Fuzzy Pattern Recognition

### 5.1. *Introduction*

Fuzzy set theory was originally introduced in pattern recognition to provide new solutions to automatic clustering problems [75, 5]. Since then, fuzzy classification and decision methods have attracted an increasing interest in the Pattern Recognition community (see [4] for a collection of landmark articles in this area). In the fuzzy approach, a class – or system state – is no longer modeled as a probability distribution but as a fuzzy subset of the feature space. It is then described by a membership function which quantifies the degree to which an arbitrary pattern may be considered as a representative sample of the class. In this way, each pattern vector $\mathbf{x}$ is assumed to belong to any class $\omega_i$ with a certain grade of membership $\mu_i(\mathbf{x})$. It is properly labelled with a membership vector $(\mu_1(\mathbf{x}), \ldots, \mu_M(\mathbf{x}))^t$.

Obviously, such a model leads to radically different interpretations as compared to the probabilistic one. For example, let us consider a system with two states, a nominal and a faulty one. A pattern that is completely representative of the nominal state is described by membership vector $(1, 0)^t$. Now, let us consider a measurement vector corresponding to a transition from the nominal state to the faulty one. A membership vector of $(0.5, 0.5)^t$ could be associated to this pattern since the state of the system is actually intermediate between a nominal one as a faulty one. This situation should not be confused with the probabilistic context in which the fact that $P(\omega_1|\mathbf{x}) = 0.5$ and $P(\omega_2|\mathbf{x}) = 0.5$ reflects a lack of information to make a reliable decision. In the same way, the occurrence of multiple faults, which is a quite frequent situation in real problems, is easier to model in a fuzzy framework than in a probabilistic one.

It may also be underlined that the fuzzy approach gives a natural solution to the problem of incomplete knowledge about the classes, because the constraint $\sum_{i=1}^{M} \mu_i$ may be easily relaxed. Lastly, the fuzzy context seems also to be well-adapted to the analysis of slow evolutions of the system because of the continuous nature of the membership functions. Applications to tool wear monitoring [76], human car driver performance monitoring [57], or supervision of the state evolution of the French telephone network [8] have shown the potential of fuzzy modelling in system diagnosis.

The design of a fuzzy classifier usually involves two stages: the construction of membership functions and the definition of decision rules. These aspects are briefly discussed in the following sections.

### 5.2. *Construction of Fuzzy Membership Functions*

Unsupervised methods assume that no labelling of the data is available, so the membership values are determined by means of a clustering algorithm. One of the most popular fuzzy clustering methods is the fuzzy $c$-means algorithm (FCM) [5]. Let the learning set be composed of $N$ training patterns $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. FCM provides a fuzzy partition matrix $U = [\mu_i(\mathbf{x}_k)]$ $(i = 1, M; k = 1, N)$ which minimises the following objective function:

$$J = \sum_{k=1}^{N} \sum_{i=1}^{M} \mu_i(\mathbf{x}_k)^m d_i(\mathbf{x}_k)^2, \qquad [10]$$

under the constraints:

$$\mu_i(\mathbf{x}_k) \in [0, 1] \quad \forall i, k, \qquad [11]$$

$$\sum_{i=1}^{M} \mu_i(\mathbf{x}_k) = 1 \quad \forall k, \qquad [12]$$

$$0 < \sum_{k=1}^{N} \mu_i(\mathbf{x}_k) < N \quad \forall i, \qquad [13]$$

where $m$ is a parameter called a fuzzyfier and $d_i(\mathbf{x}_k)$ is the Euclidean distance between the $k$th training pattern $\mathbf{x}_k$ and the cluster centre $\mathbf{v}_i$ of $\omega_i$. The solution of the optimisation problem is given by the two optimality conditions:

$$\mathbf{v}_i = \frac{\sum_{k=1}^{N} \mu_i(\mathbf{x}_k)^m \mathbf{x}_k}{\sum_{k=1}^{N} \mu_i(\mathbf{x}_k)^m} \quad \forall i = 1, M, \qquad [14]$$

$$\mu_i(\mathbf{x}_k) = \frac{1}{\sum_{j=1}^{M} \left(d_i(\mathbf{x}_k)/d_j(\mathbf{x}_k)\right)^{2/(m-1)}}. \qquad [15]$$

The iterative algorithm can be described as follows:

- initialise $U^0$, $t = 0$;
- Repeat
  - $t \leftarrow t + 1$
  - Compute centres $\mathbf{v}_i^t$ using Eq. 14
  - Update partition matrix $U^t$ using Eq. 15

until $||U^t - U^{t-1}|| \le \epsilon$

After convergence, grades of membership for a new observation $\mathbf{x}$ are obtained by replacing $\mathbf{x}_k$ by $\mathbf{x}$ in Eq. 15. Note that, by construction, the normality constraint is applied.

Using supervised methods, it is assumed that expert information is available in the form of either a hard classification or grades of membership for each training pattern. In the first case, a prototypical pattern $\mathbf{p}_i$ is chosen for each class $\omega_i$. The membership function of that class is then defined as a decreasing function of a dissimilarity measure (such as the Euclidean or Mahalanobis distance) between each pattern and the prototype. The shape of that function is chosen according the designer's preference among a lot of possibilities such as an exponential function or the multidimensional $\pi$-function proposed by Pal [53]. Whatever the function choice, this unique prototype approach is well adapted when the shape of the clusters is spherical or elliptical. If it is not the case, a multi-prototype approach may be useful; it is then necessary to aggregate individual membership functions by means of a fuzzy operator which can be chosen in the t-conorm family [22]. For example, if $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{k_i}$ denote the $k_i$ prototypical patterns of class $\omega_i$, the following operator, called bounded sum, can be applied in the following way [48]:

$$\mu_i(\mathbf{x}) = \min\left(1, \sum_{j=1}^{k_i} \mu_{p_j}(\mathbf{x})\right). \qquad [16]$$

In some cases, a fuzzy partition matrix may be directly elicited from an expert. The construction of a membership function may then be viewed as a function approximation problem, given grades of membership at several points in feature space [3]. Milleman [50] has demonstrated an approach to this problem based on a multilayer perceptron, in a study about subjective assessment of car passenger comfort from expert evaluation of vibratory signals.

### 5.3. *Fuzzy Classification Rules*

Although the "soft" decisions represented by membership degrees may in some applications be sufficient, a number of classification rules have been proposed to "defuzzify" the decision process. Let $\alpha(\mathbf{x})$ denote the action related to $\mathbf{x}$ chosen in the set $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_M\}$. Action $\alpha_i$ is interpreted as exclusive assignment to class $\omega_i$. The most usual rule consists in selecting the action corresponding to the highest membership degree [52]:

$$\alpha(\mathbf{x}) = \alpha_i \quad \text{if} \quad \mu_i(\mathbf{x}) = \max_{j=1,M} \mu_j(\mathbf{x}). \qquad [17]$$

To include reject options, it is possible to combine this rule with the use of membership thresholds, either fixed *a priori* or computed from the learning set as:

$$\mu_i = \min_{\mathbf{x}_k \in \omega_i} \mu_i(\mathbf{x}_k). \qquad [18]$$

Let us now assume that an extended set $\mathcal{A} = \{\alpha_0, \alpha_d, \alpha_1 \ldots, \alpha_M\}$ of possible actions is considered, where $\alpha_0$ and $\alpha_d$ represent ambiguity and distance rejection, respectively. Let $J(\mathbf{x})$ denote the set of candidate classes for pattern $\mathbf{x}$, defined as:

$$J(\mathbf{x}) = \{i \in \{1, \ldots, M\} | \mu_i(\mathbf{x}) > \mu_i\}. \qquad [19]$$

A reasonable decision rule may then be expressed as follows:

$$\begin{array}{lll}\alpha(\mathbf{x}) = \alpha_i & \text{if} & J(\mathbf{x}) = \{i\}, \\ \alpha(\mathbf{x}) = \alpha_d & \text{if} & J(\mathbf{x}) = \emptyset, \\ \alpha(\mathbf{x}) = \alpha_0 & \text{if} & |J(\mathbf{x})| > 1. \end{array} \qquad [20]$$

A drawback of this simple rule is that ambiguity and distance reject rates are both controlled by the same factor, namely, the membership threshold. To avoid this problem, one could define two thresholds of acceptance and rejection, respectively. Unfortunately, this strategy leads to a prohibitively large number of decision regions for high values of $M$. A solution to this problem was proposed in [28, 32] as the "membership ratio" rule. In this approach, distance rejection and exclusive class assignment are evaluated in the same way as in Eq. 20. However, a new step is introduced to deal with ambiguity rejection: if the cardinality of set $J(\mathbf{x})$ is greater than one, a membership ratio is computed to select a class, or to decide ambiguity rejection. This ratio is defined as:

$$R = \frac{\mu_m(\mathbf{x})}{\mu_p(\mathbf{x})}, \qquad [21]$$

where

$$\mu_p(\mathbf{x}) = \max_{i \in J(\mathbf{x})} \mu_i(\mathbf{x}), \qquad [22]$$

$$\mu_m(\mathbf{x}) = \max_{i \in J(\mathbf{x}) \setminus \{p\}} \mu_i(\mathbf{x}). \qquad [23]$$

This membership ratio may be interpreted as follows:

— if $R$ is close to zero, then $\mu_p(\mathbf{x})$ is much higher than the other grades of membership and action $\alpha_p$ can be selected with high confidence.

— if $R$ is close to one, then one may hesitate between at least two possible actions $\alpha_p$ and $\alpha_m$ and $\mathbf{x}$ should be ambiguity rejected.

To make a decision, the membership ratio is compared to a predefined threshold $T$. If $R > T$ then $\mathbf{x}$ is rejected, otherwise action $\alpha_p$ is selected. In this way, ambiguity and distance reject rates are adjusted independently and a reliable decision is taken.

The membership ratios and decision boundaries for that rule are shown in Figure 5 for the river monitoring data (10 prototypes were generated using the FCM procedures). A comparison of the resulting decision regions with those obtained with the Chow and distance reject options (Figure 3) reveals a strong similarity between these two approaches, on this example. However, as a non parametric approach, the distance-based fuzzy approach would probably perform better in case of classes departing more heavily from Gaussian assumptions.

### 5.4. *Fuzzy Integral Approach*

The theory of fuzzy measures and fuzzy integrals provides the basis for a completely different approach to classification problems [71, 38]. The essential idea in this

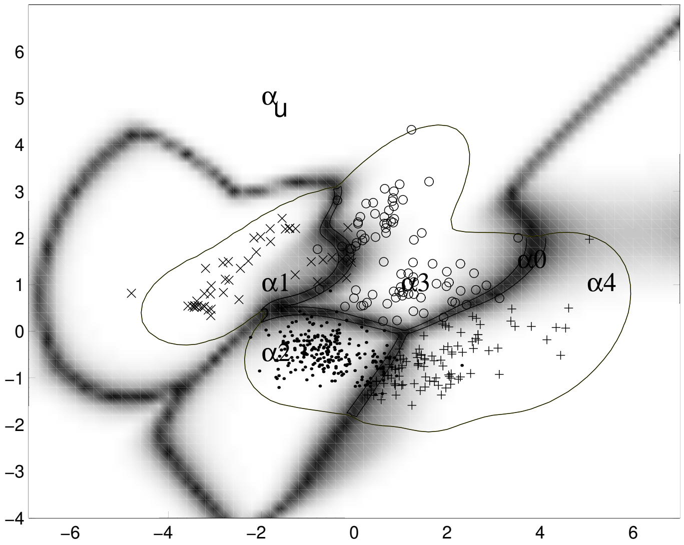


**Figure 5.** *Membership ratio (grey scales) and decision boundaries for Frélicot's membership rule (river monitoring data).*

approach resides in the combination of the information conveyed by various features (or sensors) using a fuzzy integral with respect to a fuzzy measure. Mathematically, a fuzzy measure over a finite space $X$ is a set function $\mu : 2^X \mapsto [0, 1]$ verifying the following axioms:

1. $\mu(\emptyset) = 0, \mu(X) = 1$
2. $\forall A, B \in 2^X, A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$.

The concept of a fuzzy measure is thus very general and has various applications, including uncertainty representation (note that probability, possibility and belief measures are special instances of fuzzy measures) and multicriteria decision-making [37]. In this latter case, $X$ denotes, e.g., a set of attributes or features, and $\mu(A)$ is interpreted as representing the "importance" of the set $A$ of features in a decision problem. The non-additivity of $\mu$ may then be used to model redundancy or, on the contrary, synergy between groups of features [36].

The classification method proposed by Grabisch [38] uses the notion of Choquet integral with respect to a fuzzy measure. The Choquet integral of a positive real function $f$ on $X = \{x_1, \ldots, x_n\}$ is defined as

$$\int f d\mu = \sum_{i=1}^{n} (f(x_i) - f(x_{i-1}))\mu(A_i),$$

where the set of indices have been rearranged in such a way that

$$0 \leq f(x_1) \leq \ldots \leq f(x_n),$$

$f(x_0) = 0$ by convention, and $A_i = \{x_i, x_{i+1}, \ldots, x_n\}$. It is a proper extension of the Lebesgue integral, which is recovered in the case where $\mu$ is an additive measure. An alternative view of a fuzzy integral is also that of an N-place operator $\mathcal{C}_\mu$ with

$$\mathcal{C}_\mu(f(x_1), \ldots, f(x_n)) = \int f d\mu.$$

In the approach described by Grabisch [38], each class is represented by a prototype $\mathbf{p}_k$ characterised by a collection of $d$ fuzzy sets $\mathbf{p}_k^1, \ldots, \mathbf{p}_k^d$ expressing the fuzzy sets of typical values for each attribute in class $\omega_k$. When an unknown sample $\mathbf{x}$ is presented, a degree of matching $\phi_k^i$ between each component $i$ of $\mathbf{x}$ and the each fuzzy set $\mathbf{p}_k^i$ is computed. The degrees of matching concerning each class $\omega_k$ are then merged into a single one by integration with respect to a fuzzy measure $\mu_k$ characteristic of class $\omega_k$:

$$\Phi_{\mu_k}(\omega_k; \mathbf{x}) = \mathcal{C}_{\mu_k}(\phi_k^1, \ldots, \phi_k^d).$$

The unknown pattern is finally assigned to the class with the largest overall degree of matching. Methods for identifying the fuzzy measures $\mu_k$ from the data have been proposed by Grabisch, who presented experimental results demonstrating the good performance of this technique as compared to standard classifiers [38].

Although we are not aware of any application of this approach to diagnosis problems, the theory of fuzzy measures does seem to provide a valuable framework for handling sensor fusion and pattern classification problems, and will probably undergo significant developments in the next years.

## 6. Evidence-Theoretic Pattern Recognition

### 6.1. *Principle Approach*

The Dempster-Shafer (D-S) theory of evidence originated from the concepts of upper and lower probabilities induced by a multi-valued mapping, introduced by Dempster in the 1960s [11]. Shafer [63] showed the advantages of using belief functions for representing someone's degrees of belief, and provided a first formalisation of a theory of belief functions. A new interpretation and an axiomatic justification of this approach was recently proposed by Smets [64], who also clarified the links between representation of belief and decision-making. An application of this theory to pattern recognition was presented in [13, 18] and its usefulness for system diagnosis was discussed in [19].

The evidence-theoretic approach to pattern recognition as introduced in [12] differs radically from the statistical and fuzzy methods described in the previous sections. It is based on a model of the beliefs entertained by a rational agent concerning the class of a new pattern, based on the evidence of a training set of feature vectors with completely or partially known classification. The fundamental concept for representing uncertainty in D-S theory is that of belief structure defined as a function $m$ from the power set $2^\Omega$ of the set $\Omega$ of hypotheses to the interval $[0, 1]$, verifying $m(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m(A) = 1$. Any subset $A$ of $\Omega$ such that $m(A) > 0$ is called a focal element of $m$. The quantity

$m(A)$ may be interpreted as the "mass of belief" that one is willing to commit to $A$ (and to none of its subsets), given the available evidence. Two belief structures $m_1$ and $m_2$ induced by distinct pieces of evidence are said to be combinable if there exist at least two non-disjoint subsets $B$ and $C$ of $\Omega$ such that $m_1(B) > 0$ and $m_2(C) > 0$. The orthogonal sum of $m_1$ and $m_2$, noted $m = m_1 \oplus m_2$, is then defined as $m(\emptyset) = 0$ and

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B) m_2(C)}, \quad [24]$$

for $A \neq \emptyset$. The orthogonal sum is commutative and associative, yielding invariance of the result with respect to the order in which the items of evidence are taken in consideration and combined.

To allow decision-making based on belief structure $m$, Smets [64] has introduced the concept of a "pignistic" probability distribution BetP obtained by distributing the mass $m(A)$ equally to each $\omega \in A$, for all $A \subseteq \Omega$:

$$\text{BetP}(\omega) = \sum_{\omega \in A} \frac{m(A)}{|A|},$$

where $|A|$ denotes the cardinality of $A$. Bayes decision analysis can then be applied by computing expected utilities or losses relative to the pignistic probability distribution. See Smets [64] for a full discussions on the advantages of using belief structures for representing uncertainty and making decisions.

In a classification context, a belief structure representing one's belief concerning the class of a new pattern $\mathbf{x}$ based on training data may be constructed in the following way. Each neighbour $\mathbf{x}_i$ of $\mathbf{x}$ in the training set is considered as an item of evidence that influences one's belief regarding the class membership of $\mathbf{x}$. This evidence is represented by a belief structure over $\Omega$ with two focal elements: the class $\omega_q$ of $\mathbf{x}_i$, and $\Omega$. The fraction of the unit mass assigned to $\{\omega_q\}$ is defined as a decreasing function of the distance between the two vectors. The $k$ belief structures resulting from the consideration of the $k$ nearest neighbours of $\mathbf{x}$ are then combined using Dempster's rule. In [77, 78], a learning algorithm was proposed for optimising the performance of this classification rule. A variant of this method based on a limited number of reference vectors and a connectionist implementation were described in [12, 16].

Note that this approach may easily be extended to handle the case where one's knowledge concerning the class of training patterns is itself affected by uncertainty [13], each training vector being labelled by a (possibly fuzzy) set of classes [79, 80]. This feature may be particularly useful in many diagnosis applications in which a learning base is built a posteriori based on expert judgment. An even more general situation was considered in [17], in which we extended our approach to the case of imprecise (interval-valued or fuzzy) measurement vectors.

### 6.2. *Decision Analysis*

Decisions regarding the classification of pattern $\mathbf{x}$ based on a belief structure $m$ may be made as follows. Let us denote as $\mathcal{A}$ a finite set of actions, including, e.g., assignment to each class and rejection. As before, the loss incurred if one chooses action $\alpha \in \mathcal{A}$ whereas pattern $\mathbf{x}$ belongs to class $\omega \in \Omega$ is denoted by $\lambda(\alpha|\omega)$. Assuming the only focal elements of $m$ to be singletons and $\Omega$, the risk of choosing action $\alpha$ relative to the pignistic probability distribution is:

$$R(\alpha|\mathbf{x}) = \sum_{\omega \in \Omega} \lambda(\alpha|\omega)\text{BetP}(\{\omega\}) \quad [25]$$

$$= \sum_{\omega \in \Omega} \lambda(\alpha|\omega) \left( m(\{\omega\}) + \frac{m(\Omega)}{|\Omega|} \right). \quad [26]$$

The Bayes decision rule then prescribes the action for which the risk is the smallest. Let us now consider two special cases in greater detail [15].

*Case 1:*

The frame $\Omega$ is composed of $M$ classes $\{\omega_1, \ldots, \omega_M\}$ that are all represented in the training set. The actions are $\mathcal{A} = \{\alpha_1, \ldots, \alpha_M\}$, where $\alpha_i$ denotes assignment to class $\omega_i$. The loss is 0 for correct classification and 1 for misclassification. We then have:

$$R(\alpha_i|\mathbf{x}) = \sum_{j \neq i} m(\{\omega_j\}) + \frac{M-1}{M} m(\Omega) \quad [27]$$

$$= 1 - m(\{\omega_i\}) - \frac{m(\Omega)}{M} \quad [28]$$

$$= 1 - \text{BetP}(\{\omega_i\}). \quad [29]$$

Hence, the Bayes rule results in this case in assignment of pattern $\mathbf{x}$ to the class with the largest pignistic probability (or, equivalently, with the largest mass of belief).

*Case 2:*

The frame $\Omega$ is composed of $M$ known classes $\{\omega_1, \ldots, \omega_M\}$ and one unknown class $\omega_u$ representing those states of nature that have not yet been encountered and for which, consequently, no example is available in the training set. In addition to assignment to all known classes, the set of possible actions now includes assignment to the unknown class (or distance rejection), denoted by $\alpha_u$, and ambiguity rejection $\alpha_0$. As before, the loss of wrongly assigning a pattern to one of the known classes is taken equal to 1; $\lambda_0$ and $\lambda_1$ denote the costs of rejection and misclassification in the unknown class, respectively; all other losses are assumed to be equal to 0. The different risks can then be computed as:

$$R(\alpha_i|\mathbf{x}) = 1 - \text{BetP}(\{\omega_i\}) \quad [30]$$

$$\text{for } i = 1, \ldots, M$$

$$R(\alpha_0|\mathbf{x}) = \lambda_0 \quad [31]$$

$$R(\alpha_u|\mathbf{x}) = \lambda_1(1 - \frac{m(\Omega)}{M+1}) \quad [32]$$

$$= \lambda_1(1 - \text{BetP}(\{\omega_u\}). \quad [33]$$

Examples of decision regions induced by the strategy of pignistic risk minimisation in each of these two cases for the river monitoring data are shown in Figures 6 and 7. The results represented in Figure 6 should be compared with those described in Figure 1, which correspond to the same decision space. The decision regions yielded by the evidential approach are more regular and intuitively more satisfactory than those obtained with the Gaussian classifier, although no rigorous comparison can obviously be performed in those regions where no training data is available. Figure 7 should be compared with Figures 3, 4 and 5. The decision regions have some similarity with those yielded by the full Bayesian approach (Figure 4). However, they were obtained with much weaker assumptions.

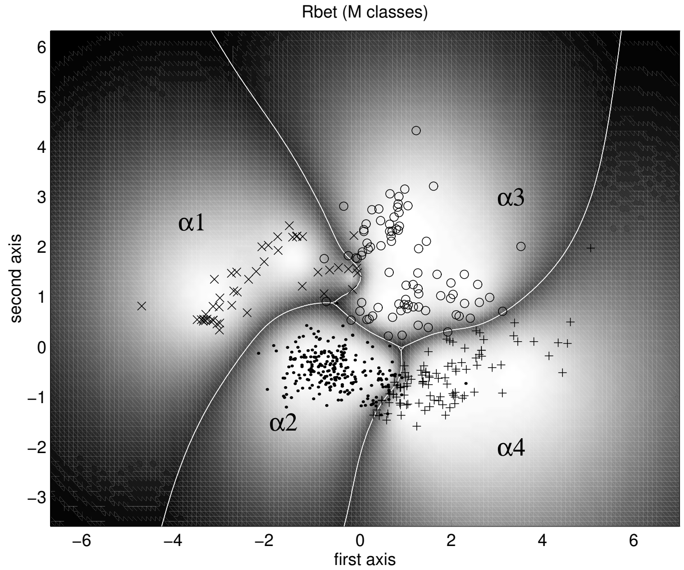


**Figure 6.** *Maximum pignistic probabilities (gray scales) and decision regions for the evidence-theoretic rule with 0/1 losses.*

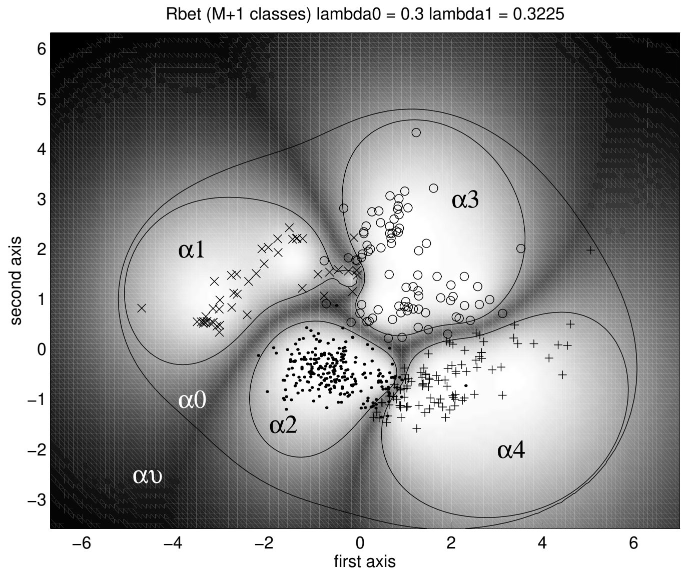


**Figure 7.** *Maximum pignistic probabilities (gray scales) and decision regions for the evidence-theoretic rule with 0/1 losses and unknown class assumption.*

## 7. Sequential Analysis

The techniques presented so far provide instantaneous decisions concerning the classification of a process according to predefined states, based on a single feature vector. This kind of diagnosis may be called "static" in that it does not take into account the temporal behaviour of the process under study. However, the performance a diagnosis system may in some cases be significantly enhanced by explicitly introducing the time evolution of the process in the problem specification, which can be achieved by analysing feature vectors in a sequence. For example, slow changes in the characteristics of some process states, or the occurrence of new states may require the use of an adaptive diagnosis procedure. In some cases, gradual transitions from one state to another also need to be detected and as far as possible anticipated to prevent the occurrence of a dangerous failure. Recently, different approaches have been proposed to make classifiers (1) more robust by taking into account past decisions to establish a diagnostic of the current system state, (2) adaptive by including incremental procedures for parameter learning and detection of new classes, and (3) predictive by anticipating the evolution of the system. These approaches are briefly reviewed in this section.

## 7.1. *Improvement of Classification Reliability*

Until now, we have assumed that all the information needed to make a decision concerning the state of the system at time $t$ was contained in a single measurement vector $\mathbf{x}(t)$. However, observations in a diagnosis system are often collected sequentially, and the quantity of available information about the system increases with time. We hereafter present two probabilistic approaches that attempt to exploit the additional information contained in sequences of feature vectors: sequential tests and Markov modelling.

### 7.1.1. *Sequential Tests*

Assume that we have at our disposal a sequence of $m$ feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m$ regarded as a realisation of $m$ independent and identically distributed random vectors [33]. For simplicity, we place ourselves in the case where the number of classes is equal to 2, and we consider the following hypotheses:

$$\begin{array}{ll} H_0: & \text{the } \mathbf{x}_i \text{ are taken from } f(\mathbf{x}|\omega_1), \\ H_1: & \text{the } \mathbf{x}_i \text{ are taken from } f(\mathbf{x}|\omega_2). \end{array}$$

Let $s$ denote the log-likelihood ratio:

$$\begin{aligned} s &= \ln \frac{f(\mathbf{x}_1, \ldots, \mathbf{x}_m|\omega_2)}{f(\mathbf{x}_1, \ldots, \mathbf{x}_m|\omega_1)} \\ &= \sum_{i=1}^{m} \left[ \ln \frac{f(\mathbf{x}_i|\omega_2)}{f(\mathbf{x}_i|\omega_1)} \right] \\ &= \sum_{i=1}^{m} \Lambda(\mathbf{x}_i). \end{aligned}$$

where $\Lambda(\mathbf{x}_i)$ denotes the log-likelihood ratio for observation $\mathbf{x}_i$. A statistical decision procedure consists in comparing $s$ to some threshold, for instance $s_0 = \ln \frac{P(\omega_1)}{P(\omega_2)}$ for the Bayes classifier, and to accept $H_1$ if $s > s_0$. The mathematical expectation and variance of $s$ are given by:

$$\begin{aligned} \mathrm{E}(s|\omega_i) &= \sum_{j=1}^{m} \mathrm{E}[\Lambda(\mathbf{x}_j)|\omega_i)] = m\eta_i, \\ \mathrm{Var}(s|\omega_i) &= \sum_{j=1}^{m} \mathrm{Var}[\Lambda(\mathbf{x}_j)|\omega_i)] = m\sigma_i. \end{aligned}$$

$\eta_i$ et $\sigma_i$ being the expectation and variance of $\Lambda(\mathbf{x})$ conditionally on $\omega_i$. The expectation of $s$ thus increases proportionally to $m$, whereas its standard deviation increases proportionally to $\sqrt{m}$; hence, the two classes become increasingly separable as $m$ increases, and the probability of error becomes smaller.

A slightly different approach consists in postponing decision-making until $s$ becomes greater than some threshold; this is the principle of Wald's procedure:

$$
\begin{array}{lcl}
s_m \leq a & \Rightarrow & \text{assign the } \mathbf{x}_i \text{ to } \omega_1 \\
a < s_m < b & \Rightarrow & \text{collect one additional observation } \mathbf{x}_{m+1} \\
s_m \geq b & \Rightarrow & \text{assign the } \mathbf{x}_i \text{ to } \omega_2
\end{array}
$$

The performance of such a procedure may be assessed according to several criteria: the probabilities of error of the first and second kind $\alpha$ et $\beta$, and the expected number $m$ of observations needed to make a decision. As $a$ and $b$ increase, the error rates decrease, but $m$ becomes larger. To fix $a$ and $b$ for given $\alpha$ and $\beta$, one may use the following approximations [33]:

$$
\begin{aligned}
a &\approx -\ln \frac{1-\alpha}{\beta}, \\
b &\approx -\ln \frac{\alpha}{1-\beta}.
\end{aligned}
$$

It may be shown that Wald's sequential test terminates with probability one, and that it minimises the mean number of observations needed to reach given error rates $\alpha$ and $\beta$.

### 7.1.2. *Markov Modelling*

In [67, 68], a pattern recognition approach and a Markov model are combined to improve the reliability and accuracy of the decision procedure. The state of the system at time $t$, say $\omega(t)$, is modelled by a first-order Markovian process associated to a transition matrix $A = [a_{ij}]$ with:

$$a_{ij} = P\left(\omega_j(t) | \omega_i(t-1)\right). \qquad [34]$$

These matrix coefficients are derived from prior knowledge concerning the system failure modes. Let $\Phi_t = \{\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_0\}$ denote all the patterns collected up to time $t$. An algorithm is proposed for recursively computing posterior probabilities given all the observations:

$$
\begin{cases}
P(\omega_j(t)|\Phi_t) = \dfrac{1}{H(t)} \dfrac{P(\omega_j(t)|\mathbf{x}_t)}{P(\omega_j)} \displaystyle\sum_{i=1}^{c} a_{ij} P(\omega_i(t-1)|\Phi_{t-1}) \\
P(\omega_j(1)|\Phi(0)) = P(\omega_j) \quad \forall j = 1, M
\end{cases}
\qquad [35]
$$

where $H(t)$ is a normalisation constant, $P(\omega_j)$ the prior probability of $\omega_j$, and $P(\omega_j|\mathbf{x}_t)$ the posterior probability of $\omega_j$ given observation $\mathbf{x}_t$ which can be computed from the Bayes decision rule. The main advantage of the Markov model is to reduce the effects of isolated errors in smoothing the state estimate. An application to a real fault monitoring problem allowed a comparison between this approach and neural and parametric ones [67, 68]. The results clearly demonstrated a significant improvement in false alarm rate.

### 7.2. *Classifier Adaptation*

Complete knowledge about the process under normal and faulty conditions is, in practice, seldom available at the beginning of the monitoring process. New collected data should then be used to complete the understanding of the system. A first approach consists in updating the classifier parameters as new observations have been assigned to known classes [65]. Slight drifts of a known mode may be accounted for in such a way. This procedure is often applied on-line since no heavy computation is usually needed.

The introduction of distance rejection or novelty detection is a more difficult problem but a key point in diagnosis, as stressed earlier. In fact, rejected points are generally representative samples of unknown classes and the decision space must be extended. Different approaches have been proposed.

The off-line approach is the most classical one. It consists in applying a clustering algorithm when a predefined number of patterns have been rejected. Statistical [41], fuzzy [5] or evidence-theoretic [19] clustering procedures provide a large variety of tools for this purpose. The exhibited clusters must be analysed from a physical point of view and periodically updated by splitting and merging methods. As new classes are detected and validated, a new learning step is started and the previous decision rule is improved. This procedure allows to start with minimal prior knowledge, essentially concerning the normal state.

Figure 8 illustrates this strategy applied to the river quality monitoring problem described in [19]. The data have the same origin than those described in Section 4.3, but two original variables (pH and conductivity) instead of the first two principle axes are displayed. The results obtained with the evidence-theoretic method described in [19] at different time steps are illustrated in Figures 8 (a) to (f). The initial data consisted of the first 30 measurements. At time $t = 76$, 30 patterns have been distance rejected (a) and a clustering procedure is run, yielding one new class (b). At $t = 164$, 30 new patterns have again been rejected (c), but no class is created (d). A third class appears at $t = 186$ (e-f). The procedure was then iterated until $t = 731$ (end of second year) without creation of any new classes. Close examination of the results by domain experts led to the interpretation of the three clusters as corresponding to two nominal states and one particular form of pollution.

This general approach is quite efficient when there is a sudden change from one state to another, but it may fail in case of slow evolution. This problem has been studied by Boudaoud et al. [7], who proposed to use a diagnosis system composed of two modules. The first one is dedicated to the sequential elaboration of fuzzy membership functions using an agglomerative clustering algorithm. The second one is acting as a supervisor that decides whether or not to start a learning phase in the first module, according to the dynamic behaviour of the system. In [56], an intermediate strategy is proposed. The idea is to create new learning sets as soon as a change is detected. First, a fuzzy membership function of the nominal state is learnt using a radial basis function network. Then, the authors suggest that a new steady state of the system is characterised by patterns situated close to each other both in space and in time. Membership-rejected

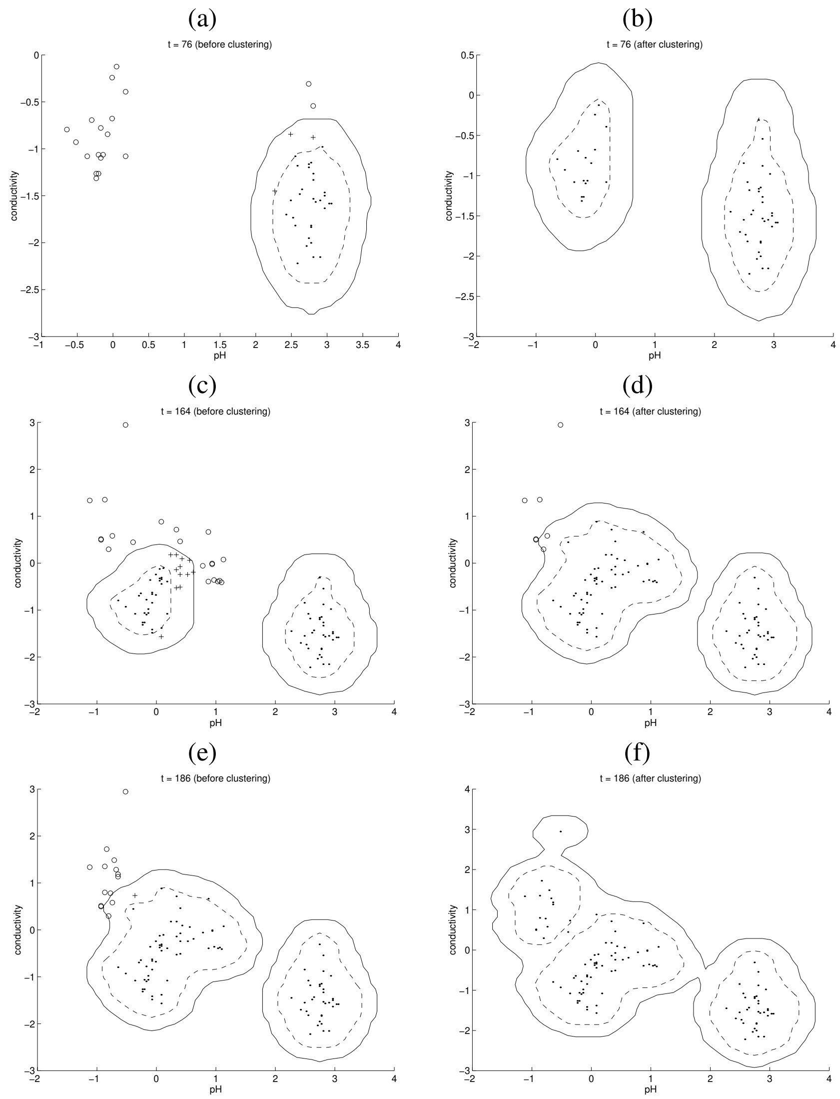


**Figure 8.** *Simulation results with water quality data. Ambiguity and distance rejected points are represented by + and o, respectively. The dotted lines delimit the regions of assignment to a known class; the boundaries of distance rejection regions are indicated as solid lines.*

patterns are gradually clustered according to a spatial and temporal proximity criterion. As soon as the size of a new group is sufficient, an additional network is trained and the decision space is extended with a new class. This approach has been applied to monitor the behaviour of a human car driver.

### 7.3. *Evolution Analysis and Prognosis*

Different approaches have been proposed to follow the evolution of the system state and to propose a predictive diagnostic.

#### 7.3.1. *Heuristic Approach*

A heuristic method was proposed by Grenier [39] for providing a qualitative description of the dynamic behaviour of a system using a small set of production rules. The goal of this method is to derive a diagnosis from the membership functions of $l$ consecutive patterns $\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \ldots, \mathbf{x}_{t_l}$ and from their variations between two time steps. Let $\Delta\mu_i(\mathbf{x}_t)$ be the change in membership of pattern $\mathbf{x}(t)$ to class $i$:

$$\Delta\mu_i(\mathbf{x}_t) = \mu_i(\mathbf{x}_t) - \mu_i(\mathbf{x}_{t-1}). \qquad [36]$$

These are some examples of rules proposed by Grenier for describing the evolution of the system in symbolic terms:

1. **If** there is only one class $\omega_i$ such that
   $\mu_i(\mathbf{x}_t) \geq \mu_i$ for all $t \in \{t_1, t_2, \ldots, t_l\}$,
   **Then** the system is said to remain in steady state $\omega_i$.
2. **If** there is only one class $\omega_i$ such that we have
   $\mu_i(\mathbf{x}_t) \geq \mu_i$ for several $t \in \{t_1, t_2, \ldots, t_l\}$, and
   $\Delta\mu_i(\mathbf{x}_t) > 0$ for all $t \in \{t_1, t_2, \ldots, t_l\}$,
   **Then** the system is said to evolve towards state $\omega_i$.
3. **If** for one class $i$ we have
   $\mu_i(\mathbf{x}_t) \geq \mu_i$ for several $t \in \{t_1, t_2, \ldots, t_l\}$, and
   $\exists t \quad \Delta\mu_i(\mathbf{x}_t) > 0$ and $\exists t' \quad \Delta\mu_i(\mathbf{x}_{t'}) < 0$,
   **Then** the system is said to oscillate near $\omega_i$.

Although a lot of other rules have been defined for facing different situations, this method remains sensitive to noise and is limited to short-term prognosis.

#### 7.3.2. *Kalman Filtering*

Gana [34] has shown that the prediction problem for slowly varying systems could be solved using a Kalman filter to predict the trajectory of patterns in feature space. A sequence of patterns is assumed to be a realisation of a discrete time stochastic process represented by the following linear state-space model:

$$\begin{cases} \mathbf{x}_{k+1} = A_k \mathbf{x}_k + \mathbf{w}_k \\ \mathbf{z}_k = \mathbf{x}_k. \end{cases} \qquad [37]$$

The transition matrix $A_k$ is slowly varying and is estimated. Using Kalman filtering equations, one is able to predict, given the set of observations $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k$, a sequence of patterns $\{\hat{\mathbf{x}}_{k+i|k}, i = 1, h\}$. The prognosis consists in applying a decision rule to the predicted vectors as shown in Figure 9.

Another approach was proposed by Frélicot [30, 29]. Because the dimension of the feature space is often higher than the dimension of the decision space, the idea is to make a prediction in a fuzzy decision space using the following non-linear state-space model:

$$\left\{ \begin{array}{l} \mathbf{x}_{k+1} = A_k \mathbf{x}_k \\ \mu_k = \mu(\mathbf{x}_k) + \mathbf{w}_k \, . \end{array} \right. \qquad [38]$$

The membership vector is then considered as the measurement vector, and the state vector is the pattern vector. An extended Kalman filter provides a sequence of fuzzy membership vectors $\{\hat{\mu}_{k+i|k}, i = 1, h\}$. A predictive diagnostic is directly produced by applying a fuzzy decision rule to the predicted membership vectors (see Section 5). The prognosis scheme is schematically depicted in Figure 10.

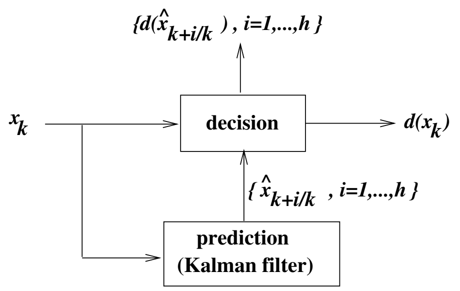


**Figure 9.** *Prognosis in feature space.*

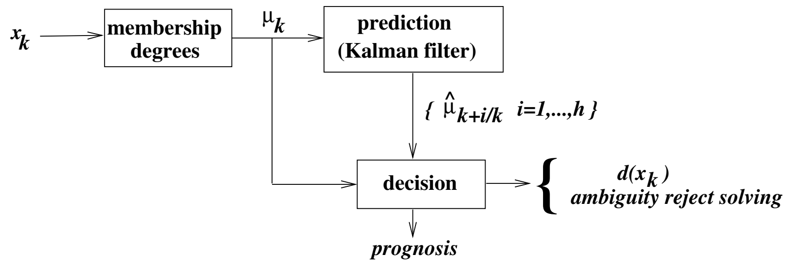


**Figure 10.** *Prognosis in decision space.*

Note that such a prediction procedure provides additional information that may be used for refining the decisions made on the basis of individual feature vectors [31]. For example, if pattern $\mathbf{x}(t)$ was ambiguity rejected, but future vectors are expected to be assigned to class $\omega_i$, then assignment to that class may be anticipated and already applied to $\mathbf{x}(t)$.

## 8. Concluding Remarks

This paper has attempted to review some of the latest developments in the application of the pattern recognition methodology for the monitoring and diagnosis of complex systems. As opposed to the model-based approach, these techniques do not require any analytic model of the process under study but only assume the availability of representative measurements from at least some of the operating conditions of the process under study. However, in many situations, it is not possible to draw a complete list of possible system states, and a lot of research has been been devoted to the development of decision procedures based on incomplete and uncertain information. Different approaches have been proposed based on various theoretical frameworks including Probability theory, Fuzzy Logic and Dempster-Shafer theory. The comparison of these theories as means of representing various kinds of uncertainty is still a much debated question and is clearly out of the scope of this paper. From a practical point of view, most of these techniques have demonstrated a remarkable performance and great flexibility in system diagnosis applications. The potentiality of various approaches for providing efficient solutions to such difficult problems as the fusion of diverse and heterogeneous sensor data will have to be assessed in future research.

## Bibliographie


[1] O. Bardou and M. Sidhamed. Early detection of leakages in the exhaust and discharge systems of reciprocating machines by vibration analysis. *Mechanical Systems and Signal Processing*, 8(5):551–570, 1994.

[2] M. Basseville. Detecting changes in signals and systems: A survey. *Automatica*, 24:309–326, 1988.

[3] R. Bellman, L. Kalaba, and L. A. Zadeh. Abstraction and pattern classification. *J. Math. Anal. Appl.*, 13:1–7, 1966.

[4] J. C. Bezdek and S. K. Pal. *Fuzzy models for pattern recognition*. IEEE Press, Piscataway, NJ, 1992.

[5] J.C Bezdek. *Pattern Recognition with fuzzy objective function algorithm*. Plenum Press, 1981.

[6] C. M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, Oxford, 1995.

[7] N. Boudaoud, M.H. Masson, and B. Dubuisson. One-line diagnosis of a technological system: a fuzzy pattern recognition approach. In *13th IFAC Congress*, San Francisco, USA, August 1996.

[8] E. Boutleux and B. Dubuisson. Detection and supervision of the state evolution of the French telephone network. In *12th Int. Conference on Systems Science*, volume 4, pages 696–701, Vancouver, Canada, October 1993.

[9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.

[10] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inform. Theory*, IT-16:41–46, 1970.

[11] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, AMS-38:325–339, 1967.

[12] T. Denœux. An evidence-theoretic neural network classifier. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 3, pages 712–717, Vancouver, October 1995.

[13] T. Denœux. A $k$-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Transactions on Systems, Man and Cybernetics*, 25(05):804–813, 1995.

[14] T. Denœux. Chapter F1.1: Classification. In E. Fiesler and R. Beale, editors, *Handbook of Neural Computation*. Oxford University Press and Institute of Physics Publishing, 1996.

[15] T. Denœux. Analysis of evidence-theoretic decision rules for pattern classification. *Pattern Recognition*, 30(7):1095–1107, 1997.

[16] T. Denœux. A neural network classifier based on Dempster-Shafer theory. Technical Report 97/45, Université de Technologie de Compiègne, Heudiasyc Laboratory, September 1997.

[17] T. Denœux. Reasoning with imprecise belief structures. Technical Report 97/44 (`http://www.hds.utc.fr/~tdenoeux/`), Université de Technologie de Compiègne, Heudiasyc Laboratory, September 1997.

[18] T. Denœux. Application du modèle des croyances transférables en reconnaissance de formes. *Traitement du Signal (In press)*, 1998.

[19] T. Denœux and G. Govaert. Combined supervised and unsupervised learning for system diagnosis using Dempster-Shafer theory. In P. Borne et al., editor, *CESA'96 IMACS Multiconference. Symposium on Control, Optimization and Supervision*, volume 1, pages 104–109, Lille, July 1996.

[20] R. Du, M. A. Elbastawi, and S. M. Wu. Automated process monitoring, Part 1: Monitoring methods. *Journal of Engineering for Industry*, 117:121–132, 1995.

[21] R. Du, M. A. Elbastawi, and S. M. Wu. Automated process monitoring, Part 2: Applications. *Journal of Engineering for Industry*, 117:133–141, 1995.

[22] D. Dubois and H. Prade. *Fuzzy Sets and Systems: Theory and Applications*. Academic Press, New York, 1980.

[23] B. Dubuisson. *Diagnostic et Reconnaissance des Formes*. Hermès, Paris, 1990.

[24] B. Dubuisson and P. Lavison. Surveillance of a nuclear reactor by use of a pattern recognition methodology. *IEEE Trans. Syst. Man Cybern.*, 10:603–609, 1980.

[25] B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(1):155–165, 1993.

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

[27] P. M. Frank. Fault diagnosis in dynamic systems using analytical and knowledge-based redundancy: A survey and some new results. *Automatica*, 26:459–474, 1990.

[28] C. Frélicot. *Un système adaptatif de diagnostic prédictif par reconnaissance des formes floues*. PhD thesis, Université de Technologie de Compiègne, Compiègne, 1992.

[29] C. Frélicot. A fuzzy-based adaptive pronostic system. *Journal Européen des Systèmes Automatisés*, 30(2-3):281–299, 1996.

[30] C. Frélicot and B. Dubuisson. K-step ahead prediction in fuzzy decsion space - application to prognosis. In *IEEE International Conference on fuzzy Systems*, San Diego, USA, 1992.

[31] C. Frélicot and B. Dubuisson. A posteriori ambiguity reject solving in fuzzy pattern classification using a multi-step predictor of membership vectors. In *Uncertainty in intelligent systems*. B. Bouchon-Meunier, L.Valverde, R. Yager, Elsevier Science, 1993.

[32] C. Frélicot, M.H Masson, and B. Dubuisson. Reject options in fuzzy classification rules. In *Proc. EUFIT'95*, volume III, pages 1495–1464, Aachen, Germany, August 1995.

[33] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1990.

[34] K. Gana. *Suivi d'évolution et aide au pronostic en maintenance de systèmes industriels*. Thèse de docteur-ingénieur, Université de Valenciennes, 1987.

[35] F. Girosi. Regularization theory, radial basis functions and networks. In J. H. Friedman V. Cherkassky and H. Wechsler, editors, *From Statistics to Neural Networks*, pages 166–187. Springer Verlag, Berlin, 1994.

[36] M. Grabisch. The representation of importance and interaction of features by fuzzy measures. *Pattern Recognition Letters*, 17:567–575, 1996.

[37] M. Grabisch. Fuzzy measures and integrals: a survey of applications and recent issues. In D. Dubois, H. Prade, and R. R. Yager, editors, *Fuzzy information engineering. A guided tour of applications*, pages 507–529. John Wiley, New York, 1997.

[38] M. Grabisch and J.-M. Nicolas. Classification by fuzzy integral: performance and tests. *Fuzzy sets and systems*, 65:255–271, 1994.

[39] D. Grenier. *Méthode de détection d'évolution. Application à l'instrumentation nucléaire*. Thèse de docteur-ingénieur, Université de Technologie de Compiègne, 1984.

[40] R. Iserman. Process fault detection based on modeling and estimation methods: A survey. *Automatica*, 20:387–404, 1984.

[41] A.K. Jain. *Algorithm for clustering data*. Prentice Hall, New Jersey, 1988.

[42] T. Kohonen. *Self organisation and associative memory*. Springer Verlag, Berlin, 1987.

[43] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[44] R. Lengellé and T. Denœux. Training MLPs layer by layer using an objective function for internal representations. *Neural Networks*, 9:83–97, 1996.

[45] R. Lengellé, Y. Hao, N. Schaltenbrand, and T. Denœux. Ambiguity and distance rejection in multilayer neural networks. In C. H. Dagli, S. R. T. Kumara, and Y. C. Shin, editors, *Intelligent Engineering Systems Through Artificial Neural Networks*, pages 299–304, New-York, 1991. ASME Press.

[46] S. Leonhardt, C. Ludwig, and R. Schwarz. Real-time supervision for diesel engine injection. *Control Engineering Practice*, 3(7):1003–1010, 1995.

[47] R. P. Lippmann. Neural networks, Bayesian a posteriori probabilities, and pattern classification. In J. H. Friedman V. Cherkassky and H. Wechsler, editors, *From Statistics to Neural Networks*, pages 83–104. Springer-Verlag, Berlin, 1994.

[48] M.H. Masson, B. Dubuisson, and C. Frélicot. Conception d'un module de reconnaissance des formes floues pour le diagnostic. *RAIRO-APII-JESA*, pages 319–341, 1996.

[49] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley and Sons, New York, 1992.

[50] S. Millemann and R. Lengellé. Fuzzy supervised membership estimation using a multilayer perceptron. In *Proc. EUFIT'95*, volume I, pages 538–542, Aachen, Germany, August 1995.

[51] G. Mourot, S. Bousghiri, and J. Ragot. Pattern recognition for diagnosis of technological systems: a review. In *Int. Conf. on Systems, Man and Cybernetics*, volume 5, pages 275–281, Le Touquet, IEEE,October 1993.

[52] S.K. Pal. Fuzzy sets and decision making approaches in vowel and speaker recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 625–629, 1977.

[53] S.K. Pal. Fuzzy tools for the management of uncertainty in pattern recognition, image analysis, vision and expert systems. *Int. Journal of Systems Science*, 22(3):511–549, 1991.

[54] R. J. Patton, P. M. Frank, and R. N. Clark. *Fault Diagnosis in Dynamic Dystems*. Prentice Hall, 1989.

[55] L. F. Pau. Diagnosis of equipment failure by pattern recognition. *IEEE Trans. Reliability*, 3:202–208, 1974.

[56] M.-A. Peltier and B. Dubuisson. A human state detection system based on a fuzzy approach. In *Tooldiag'93, Int. Conf. on Fault Diagnosis*, pages 645–652, Toulouse, April 1993.

[57] M.A. Peltier and B. Dubuisson. A fuzzy clustering algorithm based on the k-nearest neighbors rule for the detection of evolution. In *IEEE Int. Conference on Systems, Man and Cybernetics*, volume 4, pages 696–701, Le Touquet, France, October 1993.

[58] T. Poggio and F. Girosi. A theory of networks for approximation and learning. Technical Report A.I. Memo No. 1140, M.I.T., 1988.

[59] S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264, 1991.

[60] D. L. Reilly, L. N. Cooper, and C. Elbaum. A neural model of category learning. *Biological Cybernetics*, 45:35–41, 1982.

[61] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[62] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. McClelland, editors, *Parallel Distributed Processing*. MIT Press, Cambridge, MA, 1986.

[63] G. Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, N.J., 1976.

[64] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66:191–243, 1994.

[65] A. Smolarz. Un algorithme de discrimination adaptatif avec rejet. *APII*, 21:449–474, 1987.

[66] P. Smyth. Detecting novel fault conditions with hidden Markov models and neural networks. In E. S. Gelsema and L. N. Kanal, editors, *Pattern Recognition in Practice IV*, pages 525–536. Elsevier, Amsterdam, 1994.

[67] P. Smyth. Detecting novel fault conditions with hidden markov models and neural networks. In *Pattern Recognition in Practice IV*, pages 525–536. Elsevier Science, 1994.

[68] P. Smyth. Hidden Markov models for fault detection in dynamic systems. *Pattern Recognition*, 27:149–164, 1994.

[69] T. Sorsa and H. N. Koivo. Application of artificial neural networks in process fault diagnosis. *Automatica*, 29(4):843–849, 1993.

[70] T. Sorsa, H. N. Koivo, and H. Koivisto. Neural networks in process fault monitoring. *IEEE Trans. Syst. Man Cybern.*, 21(4):815–825, 1991.

[71] H. Tahani and J. M. Keller. Information fusion in computer vision using the fuzzy integral. *IEEE Transactions on Systems, Man and Cybernetics*, 20(3):733–741, 1990.

[72] Y. Tinghu, Z. Binglin, and H. Ren. A neural network methodology for rotating machinery fault diagnosis. In *Tooldiag'93, Int. Conf. on Fault Diagnosis*, pages 170–178, Toulouse, April 1993.

[73] T. Trautmann, M. Clément, T. Denœux, and T. Wittig. Application of intelligent techniques to river quality monitoring. In *Proceedings of EUFIT'95*, volume 3, pages 1603–1609, Aachen, August 1995.

[74] T. Trautmann and T. Denœux. Comparison of dynamic feature map models for environmental monitoring. In *Proceedings of ICNN'95*, volume 1, pages 73–78, Perth, Australia, IEEE, November 1995.

[75] L. A. Zadeh. Fuzzy sets and their application to pattern classification and cluster analysis. Technical Report UCB/ERL M-607, University of California, Berkely, 1977.

[76] S. Zieba and B. Dubuisson. Tool wear monitoring and diagnosis in milling using vibration signals. In *SafeProcess'94*, pages 696–701, Espoo, Finland, June 1994.

[77] L. M. Zouhal and T. Denœux. An adaptive $k$-NN rule based on Dempster-Shafer theory. In *Proc. of the 6th Int. Conf. on Computer Analysis of Images and Patterns (CAIP'95)*, pages 310–317, Prague, Springer Verlag, September 1995.

[78] L. M. Zouhal and T. Denœux. A comparison between fuzzy and evidence-theoretic $k$-NN rules for pattern recognition. In *Proceedings of EUFIT'95*, volume 3, pages 1319–1325, Aachen, August 1995.

[79] L. M. Zouhal and T. Denœux. Reconnaissance de formes floues par la théorie de Dempster et Shafer. In *Rencontres Francophones sur la Logique Floue et ses Applications*, pages 3–8, Nancy, Cépaduès, December 1996.

[80] L. M. Zouhal and T. Denœux. Generalizing the evidence-theoretic $k$-NN rule to fuzzy pattern recognition. In *Proceedings of the Second International Symposium on Fuzzy Logic and Applications ISFL'97*, pages 294–300, Zurich, ICSC Academic Press, February 1997.