Analysis of evidence-theoretic decision rules for pattern classification[1]

Thierry Denœux

Université de Technologie de Compiègne
U.R.A. CNRS 817 Heudiasyc
BP 529 F-60205 Compiègne cedex – France
email: Thierry.Denoeux@hds.utc.fr

[1]To appear in *Pattern Recognition*, 30:7.

Summary

The Dempster-Shafer (D-S) theory of evidence provides a convenient framework for modeling uncertainty in situations where the available evidence is limited or weak. Such situations typically arise in supervised pattern recognition when the training set is small or does not contain samples from all classes. In D-S theory, someone's feeling of uncertainty concerning a certain set of hypotheses is represented by a belief function, that generalizes the classical concept of probability distribution. Recently, several classification methods have been proposed for determining a posterior belief function over a set of classes, given a pattern $\mathbf{x}$ to be classified. However, the problem of making decisions based on a belief function and an arbitrary loss matrix is not simple, since conventional Bayes decision analysis assumes uncertainty to be quantified by a probability distribution.

This paper provides detailed analysis of several decision rules that can be based on a posterior belief function and a quantification of the consequences of each action (assignment to a class or rejection). Different decision strategies are proposed, based on the concepts of upper, lower and pignistic expectation that generalize the usual concept of mathematical expectation underlying Bayes decision theory. The decision rules and the corresponding decision regions in feature space are analyzed under the assumptions of completeness and incompleteness of the training set. As expected, patterns situated close to class boundaries tend to be rejected according to the three decision rules studied. Additionally, the more conservative strategies of upper and pignistic expected loss minimization also lead to the rejection of "atypical" patterns situated far from training samples in feature space. When the training set is not complete, the available options are assignment to one of the known classes, assignment to the unknown class, and rejection. By imposing some restrictions on the loss matrix, it is shown that various configurations of the decision regions can be induced by varying only two parameters corresponding to the costs of rejection and misclassification in the unknown class, respectively.

The analysis is illustrated using real data from a river quality monitoring application.

Abstract

The Dempster-Shafer theory provides a convenient framework for decision making based on very limited or weak information. Such situations typically arise in pattern recognition problems when patterns have to be classified based on a small number of training vectors, or when the training set does not contain samples from all classes. This paper examines different strategies that can be applied in this context to reach a decision (e.g., assignment to a class or rejection), provided the possible consequences of each action can be quantified. The corresponding decision rules are analyzed under different assumptions concerning the completeness of the training set. These approaches are then demonstrated using real data.

Keywords: pattern classification, Dempster-Shafer theory, decision analysis, uncertainty modeling, system diagnosis.

1 Introduction

The main purpose of statistical pattern recognition is the design of decision rules whereby entities, represented by feature vectors, can be assigned to predefined groups of patterns, or *classes*. In the classical Bayesian approach[1], one considers a finite set of actions $\mathcal{A} = \{\alpha_1, \ldots, \alpha_a\}$ and a finite set of classes $\Omega = \{\omega_1, \ldots, \omega_M\}$. Typical actions are assignment to a class and pattern rejection[2]. If, upon consideration of input pattern $\mathbf{x}$, we select action α_i whereas the corresponding entity belongs to class ω_j, we incur a loss $\lambda(\alpha_i|\omega_j)$. If we denote by $P(\omega_j|\mathbf{x})$ the posterior probability of class ω_j given $\mathbf{x}$, then the expected loss associated with the choice of action α_i is:

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

A decision rule D assigns an action $D(\mathbf{x})$ to each feature vector $\mathbf{x}$. Associated to it is an overall risk defined as:

$$R = \int R(D(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad (2)$$

where $p(\mathbf{x})$ denotes the mixture probability density of input vectors, and where the integral extends over the whole feature space. The decision rule with minimal overall risk is the Bayes decision rule, which prescribes the action entailing the smallest expected loss.

In practical situations, the lack of knowledge of the exact posterior probabilities precludes application of the Bayes decision rule. Nevertheless, this rule can be approximated by estimating the posterior probabilities from empirical data, using parametric or non-parametric methods. In each case, the estimated expected losses are assimilated to the real ones, and the action with the lowest estimated expected loss is selected.

Whereas this scheme has proved quite efficient in a wide range of applications, it suffers from severe shortcomings in some situations of practical interest. One such situation is the scarcity of training data. In that case, only poor estimates of the posterior probabilities and of the expected losses can usually be obtained, particularly when the pattern under consideration is very dissimilar from any recorded pattern with known classification. While in some circumstances it is still reasonable in that case to select the action with minimal estimated expected loss, it may sometimes be preferable to reject the pattern when the uncertainty is too high. Note that this type of reject differs essentially from the "ambiguity reject" proposed by Chow[2]. In one case (Chow's reject option), the pattern is rejected because the expected loss of assigning it to a class is higher than the expected loss of rejecting it. This situation is perfectly modeled within the framework of Bayes decision theory. In the other case that is emphasized here, uncertainty results from fundamental ignorance regarding the true expected losses. This type of uncertainty is not easily handled within the framework of Bayes theory, which always assumes precise determination of the posterior probabilities[3].

A second situation, closely linked to the previous one, is the absence of training data belonging to certain classes, or even the lack of knowledge regarding the exact number of classes. Such a situation is typically encountered in system diagnosis applications, because of the impossibility of gathering data corresponding to certain system states, or because the number of possible states is very high(4). To solve this problem, Dubuisson has introduced the "distance reject option" which consists in postponing decision making when the pattern is situated "far" from training samples. This approach has proved very efficient in many diagnosis applications(4). However, it requires the determination of distance thresholds, which cannot easily be done without some underlying model of known and unknown states. Smyth(5) has developed a Bayesian approach based on the estimation of class-conditional densities and posterior probabilities of each class using two distinct pattern recognition models, but the necessity to postulate arbitrarily a "non-informative" density function for the unknown class — e.g. a uniform density over a bounded space of feature values — can also be regarded as a shortcoming of this approach.

As an attempt to bring new answers to the above problems, a novel pattern classification method based on the Dempster-Shafer (D-S) theory of evidence(6) was recently introduced(7). This approach has some similarity with k-nearest neighbor (k-NN) rules(8), in that classification of a feature vector is achieved by considering its nearest neighbors in a set of recorded patterns. However, an important difference resides in the way this evidence is represented and combined. In our approach – hereafter referred to as the evidence-theoretic k-NN rule – the information associated with each neighbor is represented in the form of a basic belief assignment (BBA), which assigns a positive "mass of belief" to each subset of the set of classes. The k BBAs resulting from the consideration of the k nearest neighbors are then combined to yield a *posterior belief function* that summarizes the available evidence and may be interpreted as defining posterior probability intervals for each class.

The last step – decision making – is the main topic of this paper. Extensions of Bayes decision analysis will be proposed, allowing for decision making based on a belief function and an arbitrary loss matrix. Various decision strategies will be analyzed under the assumptions of complete or partial knowledge of possible states of nature. This approach will be demonstrated on a real example.

This paper is organized as follows. Section 2 provides a brief overview of D-S theory and its application to pattern classification. Decision analysis is specifically addressed in Section 3, where the main results are presented. The approach is illustrated using a concrete example in Section 4. Section 5 concludes the paper.

2 Background

2.1 Dempster-Shafer theory

In this section, only the main concepts of D-S theory will be recalled. A complete description can be found in Shafer's book(6). A different but closely related approach (the Transferable Belief Model) has been proposed and justified axiomatically by

Smets[9,10].

Let Θ represent a finite set of hypotheses about some problem domain, called the *frame of discernment*. A piece of evidence that influences our belief concerning these hypotheses induces a basic belief assignment m, defined as a function from 2^Θ to $[0,1]$ verifying:

$$m(\emptyset) = 0 \tag{3}$$

$$\sum_{A \subseteq \Theta} m(A) = 1 \tag{4}$$

The number $m(A)$ can be interpreted as the "mass of belief" that one is willing to commit exactly to hypothesis A (and to none of its subsets) given the available evidence. The subsets A of Θ such that $m(A) > 0$ are called the focal elements of m. The *credibility* $Bel(A)$ and the *plausibility* $Pl(A)$ of $A \subseteq \Theta$ can be computed from m using the following formula:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{5}$$

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B) = 1 - Bel(\bar{A}) \tag{6}$$

with $\bar{A} = \Theta \setminus A$. The quantity $Bel(A)$ can be interpreted as a measure of the extent to which one believes in A, given the evidence pointing to that hypothesis either directly (through $m(A)$) or indirectly (through $m(B)$, for all $B \subseteq A$). The extent to which one doubts hypothesis A is based on the mass of belief committed to $\bar{A}$ or its subsets, and is therefore represented by $1 - Pl(A)$. The belief and plausibility functions can be regarded as defining lower and upper bounds for a set $\mathcal{C}$ of "compatible" probability distributions: $P \in \mathcal{C}$ iif $Bel(A) \leq P(A) \leq Pl(A)$ for all $A \subseteq \Theta$.

An important aspect of D-S theory concerns the aggregation of evidence provided by different sources. If two BBAs m_1 and m_2 induced by distinct items of evidence are such that $m_1(B) > 0$ and $m_2(C) > 0$ for some non-disjoint subsets B and C of Θ, then they are *combinable* by means of Dempster's rule. The *orthogonal sum* of m_1 and m_2, noted $m = m_1 \oplus m_2$, is defined as:

$$m(\emptyset) = 0 \tag{7}$$

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \emptyset} m_1(B) m_2(C)} \qquad \forall A \neq \emptyset \tag{8}$$

The orthogonal sum is commutative and associative. Consequently, the BBA resulting from the combination of several items of evidence does not depend on the order in which the available information is taken in consideration and combined.

Given a function $f : \Theta \mapsto \mathbb{R}$ and a probability distribution P over Θ, the expectation of f relative to P is

$$E(f, P) = \sum_{\theta \in \Theta} f(\theta) P(\theta) \tag{9}$$

The concepts of lower and upper probabilities lead to the definition of upper and lower expectation as[11]:

$$E_*(f) = \min_{P \in \mathcal{C}} E(f, P) \tag{10}$$

$$E^*(f) = \max_{P \in \mathcal{C}} E(f, P) \tag{11}$$

It can be shown that[12]:

$$E_*(f) = \sum_{A \subseteq \Theta} m(A) \min_{\theta \in A} f(\theta) \tag{12}$$

$$E^*(f) = \sum_{A \subseteq \Theta} m(A) \max_{\theta \in A} f(\theta) \tag{13}$$

In the context of decision theory, we can define $f_1(\theta)$ (respectively, $f_2(\theta)$) as the loss incurred if some action α_1 (respectively, α_2) is taken while hypothesis θ is true. If uncertainty is represented by a BBA, then expected losses in the classical sense cannot be uniquely assigned to actions α_1 and α_2. However, decision strategies can still be based on the minimization of the lower or the upper expected loss[3].

Alternatively, Smets[10,13] has proposed to invoke the "Insufficient Reason Principle" to choose in $\mathcal{C}$ a particular probability distribution allowing for the calculation of probabilistic expectations. It can be argued that, given a lack of information, the mass of belief $m(A)$ should be equally distributed among the elements of A, for all $A \subseteq \Theta$. Application of this principle leads to the concept of *pignistic*[1] probability distribution $BetP$ defined for all $\theta \in \Theta$ as:

$$BetP(\theta) = \sum_{\theta \in A} \frac{m(A)}{|A|} \tag{14}$$

where $|A|$ denotes the cardinality of $A \subseteq \Theta$. Note that there is no contradiction between the choice of D-S theory to represent uncertainty and the introduction of a probability distribution to allow decision making, if one distinguishes two levels: a *credal* level at which evidence is taken in consideration and combined, and a *decision* level which can still be based on probabilistic reasoning and minimization of the expected loss[10].

2.2 Application to pattern classification

Assume that some feature vector $\mathbf{x}$ has to be classified in one of M classes based on a training set of training vectors with known classification[2]. Different solutions to this

[1] The word *pignistic* was coined by Smets from the Latin *pignus* = a bet[10]. In the Transferable Belief Model, this term is used to designate the probability function constructed from a belief function in the context of forced decisions (or bets)[13].

[2] In the following, we shall assume for simplicity the class of each training vector to be known with certainty. The more general situation in which the training set is only imperfectly labeled has been studied elsewhere[7].

problem based on D-S theory have been proposed(14,7,15). In the basic approach(7), the presence of a training pattern $\mathbf{x}^s$ of class ω_q among the k nearest neighbors of a pattern $\mathbf{x}$ to be classified is considered as a piece of evidence that influences our belief concerning the class membership of the entity under consideration. This information is represented by a BBA m^s over the set Ω of classes. A fraction of the unit mass is assigned by m^s to the singleton $\{\omega_q\}$, and the rest is assigned to the whole frame of discernment Ω. The mass $m^s(\{\omega_q\})$ is defined as a decreasing function of the distance d^s between $\mathbf{x}$ and $\mathbf{x}^s$ in feature space:

$$
\begin{aligned}
m^s(\{\omega_q\}) &= \alpha\phi_q(d^s) && (15)\\
m^s(\Omega) &= 1 - \alpha\phi_q(d^i) && (16)
\end{aligned}
$$

where $0 < \alpha < 1$ is a constant, and ϕ_q is a monotonically decreasing function verifying $\phi_q(0) = 1$ and $\lim_{d\to\infty} \phi_q(d) = 0$. An exponential form can be postulated(7) for ϕ_q:

$$
\phi_q(d^s) = \exp(-\gamma_q(d^s)^2) \qquad (17)
$$

γ_q being a positive parameter associated to class ω_q.

The k nearest neighbors of $\mathbf{x}$ can be regarded as k independent sources of information, each one represented by a BBA. This multiple evidence can be pooled by means of Dempster's rule of combination, resulting in a BBA m representing our belief concerning the class membership of $\mathbf{x}$, following the consideration of its k nearest neighbors in the training set. The focal elements of m are singletons of Ω, and Ω itself. Note that the calculation of m involves the combination of k *simple* BBAs(6), and can therefore be performed in linear time with respect to the number M of classes. This evidence-theoretic k-NN rule was shown to have good classification accuracy as compared to the voting and distance-weighted rules(7). A learning algorithm was proposed by Zouhal and Denœux(15) for determining the parameters γ_q in equation 17 by optimizing an error criterion; this improvement was shown to yield further reduction of classification error in most cases.

To reduce computing time, input patterns can also be classified according their distances to a limited number of reference patterns or prototypes(14). For each prototype, a BBA is computed using an expression similar to Equation 17. The corresponding BBAs are then combined using Dempster's rule. This rule can be implemented in a multilayer neural network with specific architecture comprising one input layer, two hidden layers and one output layer. The weight vector, the receptive field and the class membership of each prototype are then determined by gradient descent of an error function.

Once a BBA m has been computed, the credibility $Bel(\{\omega_q\})$, the plausibility $Pl(\{\omega_q\})$ and the pignistic probability $BetP(\{\omega_q\})$ of each class ω_q are respectively equal to:

$$
\begin{aligned}
Bel(\{\omega_q\}) &= m(\{\omega_q\}) && (18)\\
Pl(\{\omega_q\}) &= m(\{\omega_q\}) + m(\Omega) && (19)\\
BetP(\{\omega_q\}) &= m(\{\omega_q\}) + \frac{m(\Omega)}{M} && (20)
\end{aligned}
$$

In the case of $\{0,1\}$ losses (where $\lambda(\alpha_i|\omega_j) = 1 - \delta_{i,j}$) and in the absence of any reject option, it is well known that the Bayes rule leads to selecting the class with maximum posterior probability[1]. Similar strategies within the framework of D-S theory can be shown to consist in maximizing the credibility, the plausibility or the pignistic probability. These strategies will be examined in greater detail in the following section.

3 Decision analysis

3.1 General formulation

In the following development, we assume as usual the finite set Ω to contain all possible states of nature, with $|\Omega| = M$. Upon consideration of input feature $\mathbf{x}$, we contemplate taking one action among a finite set $\mathcal{A}$. Our decision takes into account the losses $\lambda(\alpha|\omega)$ for all $(\alpha,\omega) \in \mathcal{A} \times \Omega$, and a BBA m over Ω obtained from a training set using one of the methods mentioned in the former section.

For an arbitrary BBA m, application of Equations 12 and 13 yields the following expressions for the lower and upper expected loss associated to each possible action $\alpha \in \mathcal{A}$:

$$R_*(\alpha|\mathbf{x}) = \sum_{A \subseteq \Omega} m(A) \min_{\omega \in A} \lambda(\alpha|\omega) \tag{21}$$

$$R^*(\alpha|\mathbf{x}) = \sum_{A \subseteq \Omega} m(A) \max_{\omega \in A} \lambda(\alpha|\omega) \tag{22}$$

The risk with respect to the pignistic probability distribution $BetP$ derived from m is

$$R_{bet}(\alpha|\mathbf{x}) = \sum_{\omega \in \Omega} \lambda(\alpha|\omega) BetP(\{\omega\}) \tag{23}$$

$$= \sum_{\omega \in \Omega} \lambda(\alpha|\omega) \sum_{A \ni \omega} \frac{m(A)}{|A|} \tag{24}$$

$$= \sum_{A \subseteq \Omega} m(A) \frac{1}{|A|} \sum_{\omega \in A} \lambda(\alpha|\omega) \tag{25}$$

Equations 21, 22 and 25 show that the three expectations are linked by the following inequalities:

$$R_*(\alpha|\mathbf{x}) \leq R_{bet}(\alpha|\mathbf{x}) \leq R^*(\alpha|\mathbf{x}) \tag{26}$$

When m is such that $m(A) > 0 \Leftrightarrow A = \Omega$ or $|A| = 1$, Equations 21, 22 and 25 take the following simpler form:

$$R_*(\alpha|\mathbf{x}) = \sum_{\omega \in \Omega} \lambda(\alpha|\omega) m(\{\omega\}) + m(\Omega) \min_{\omega \in \Omega} \lambda(\alpha|\omega) \tag{27}$$

$$R^*(\alpha|\mathbf{x}) = \sum_{\omega \in \Omega} \lambda(\alpha|\omega) m(\{\omega\}) + m(\Omega) \max_{\omega \in \Omega} \lambda(\alpha|\omega) \tag{28}$$

$$R_{bet}(\alpha|\mathbf{x}) \quad = \quad \sum_{\omega\in\Omega} \lambda(\alpha|\omega)m(\{\omega\}) + m(\Omega)\left(\frac{1}{M}\sum_{\omega\in\Omega}\lambda(\alpha|\omega)\right) \tag{29}$$

In the above expressions, the only difference resides in the term associated to the uncommitted mass $m(\Omega)$ in the weighted sum, which is respectively equal to the lowest, the highest and the average loss corresponding to the choice of action α.

In general, application of the principle of expected loss minimization leads to three distinct decision strategies:

- minimization of the lower expected loss, yielding a decision rule D_* defined by:

$$D_*(\mathbf{x}) = \alpha_* \text{ with } R_*(\alpha_*) = \min_{\alpha\in\mathcal{A}} R_*(\alpha|\mathbf{x}) \tag{30}$$

- minimization of the upper expected loss, resulting in decision rule D^*:

$$D^*(\mathbf{x}) = \alpha^* \text{ with } R^*(\alpha^*) = \min_{\alpha\in\mathcal{A}} R^*(\alpha|\mathbf{x}) \tag{31}$$

- minimization of the expected loss relative to $BetP$ (or pignistic expected loss), resulting in decision rule D_{bet}:

$$D_{bet}(\mathbf{x}) = \alpha_{bet} \text{ with } R_{bet}(\alpha_{bet}) = \min_{\alpha\in\mathcal{A}} R_{bet}(\alpha|\mathbf{x}) \tag{32}$$

Obviously, these three decision strategies are not equivalent. For example, let us consider the situation of complete ignorance. The unit mass of belief is then concentrated on Ω ($m(\Omega) = 1$), leading to the following decisions:

$$D_*(\mathbf{x}) \quad = \quad \arg\min_{\alpha\in\mathcal{A}} \min_{\omega\in\Omega} \lambda(\alpha|\omega) \tag{33}$$

$$D^*(\mathbf{x}) \quad = \quad \arg\min_{\alpha\in\mathcal{A}} \max_{\omega\in\Omega} \lambda(\alpha|\omega) \tag{34}$$

$$D_{bet}(\mathbf{x}) \quad = \quad \arg\min_{\alpha\in\mathcal{A}} \frac{1}{M}\sum_{\omega\in\Omega} \lambda(\alpha|\omega) \tag{35}$$

Assuming $\Omega = \{\omega_1, \omega_2, \omega_3\}$, $\mathcal{A} = \{\alpha_1, \alpha_2, \alpha_3\}$ and, say, the following loss matrix:

	α_1	α_2	α_3
ω_1	0.5	2	3
ω_2	2	1	1.1
ω_3	3	2.5	1

we see that $D_*(\mathbf{x}) = \alpha_1$ whereas $D^*(\mathbf{x}) = \alpha_2$ and $D_{bet}(\mathbf{x}) = \alpha_3$.

Given the fact that these different strategies may lead to different decisions, it may be wondered which of them best applies to a given pattern recognition task. This important point is examined in the sequel.

3.2 Decision based on a complete learning set

In this section, we shall assume the learning set to be complete, i.e. to contain patterns from all classes. Situations of this kind typically arise in well-structured domains such as character recognition, in which the classes are all known in advance and instances from each class are easily obtained. If we assume Ω to be composed of M classes $\omega_1, \ldots, \omega_M$, typical actions are then assignment to each class ω_i, noted as α_i, $i \in \{1, \ldots, M\}$ and rejection α_0. To simplify the discussion, the losses will be assumed to be equal to 1 for misclassification, 0 for correct classification and $\lambda_0 \geq 0$ for rejection.

For arbitrary m, application of Equations 21, 22 and 25 yields in that case:

$$
\begin{aligned}
R_*(\alpha_i|\mathbf{x}) &= \sum_{A \subseteq \Omega} m(A) \min_{\omega \in A} \lambda(\alpha_i|\omega) && (36)\\
&= 1 - \sum_{A \ni \omega_i} m(A) && (37)\\
&= 1 - Pl(\{\omega_i\}) && (38)\\
R^*(\alpha_i|\mathbf{x}) &= \sum_{A \subseteq \Omega} m(A) \max_{\omega \in A} \lambda(\alpha_i|\omega) && (39)\\
&= 1 - m(\{\omega_i\}) && (40)\\
&= 1 - Bel(\{\omega_i\}) && (41)\\
R_{bet}(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{M} \lambda(\alpha_i|\omega_j) BetP(\{\omega_j\}) && (42)\\
&= \sum_{j \neq i} BetP(\{\omega_j\}) && (43)\\
&= 1 - BetP(\{\omega_i\}) && (44)
\end{aligned}
$$

for all $i \in \{1, \ldots, M\}$, and

$$
R_*(\alpha_0|\mathbf{x}) = R^*(\alpha_0|\mathbf{x}) = R_{bet}(\alpha_0|\mathbf{x}) = \lambda_0 \qquad (45)
$$

In the special case where the set of focal elements of m is reduced to singletons and the whole frame of discernment, we have:

$$
\begin{aligned}
R_*(\alpha_i|\mathbf{x}) &= 1 - m(\{\omega_i\}) - m(\Omega) && (46)\\
R^*(\alpha_i|\mathbf{x}) &= 1 - m(\{\omega_i\}) && (47)\\
R_{bet}(\alpha_i|\mathbf{x}) &= 1 - m(\{\omega_i\}) - \frac{m(\Omega)}{M} && (48)
\end{aligned}
$$

In that case, $R_*(\alpha_i)$, $R^*(\alpha_i)$ and $R_{bet}(\alpha_i)$ differ only by a constant additive term for $i = 1, \ldots, M$; consequently, these three expectations induce the same ranking of actions $\alpha_1, \ldots, \alpha_M$. However, the conditions for pattern rejection are different:

$$
D_*(\mathbf{x}) = \alpha_0 \quad \Leftrightarrow \quad \max_{j=1,\ldots,M} Pl(\{\omega_j\}) < 1 - \lambda_0 \qquad (49)
$$

$$D^*(\mathbf{x}) = \alpha_0 \quad \Leftrightarrow \quad \max_{j=1,\ldots,M} Bel(\{\omega_j\}) < 1 - \lambda_0 \tag{50}$$

$$D_{bet}(\mathbf{x}) = \alpha_0 \quad \Leftrightarrow \quad \max_{j=1,\ldots,M} BetP(\{\omega_j\}) < 1 - \lambda_0 \tag{51}$$

Since $Bel(\{\omega_j\}) \leq BetP(\{\omega_j\}) \leq Pl(\{\omega_j\})$ for all $j \in \{1, \ldots, M\}$, we have:

$$D_*(\mathbf{x}) = \alpha_0 \Rightarrow D_{bet}(\mathbf{x}) = \alpha_0 \Rightarrow D^*(\mathbf{x}) = \alpha_0 \tag{52}$$

Deeper understanding of these rules can be gained by visualizing the different types of decision (classification or reject) for all possible values of $\mu_1 = m(\Omega)$ and $\mu_2 = \max_{j=1,\ldots,M} m(\{\omega_j\})$, as shown in Figure 1. Since

$$\sum_{i=1}^{M} m(\{\omega_i\}) = 1 - m(\Omega) \tag{53}$$

the joint values of μ_1 and μ_2 are constrained by the following inequalities:

$$\frac{1-\mu_1}{M} \leq \mu_2 \leq 1 - \mu_1 \tag{54}$$

Graphically, all points (μ_1, μ_2) are therefore situated inside the triangle defined by the points of coordinates $(1,0)$, $(0,1/M)$ and $(0,1)$ (Figure 1). The conditions for rejection by D_*, D^* and D_{bet} can be expressed in terms of μ_1 and μ_2 in the following way:

$$D^*(\mathbf{x}) = \alpha_0 \quad \Leftrightarrow \quad \mu_2 < 1 - \lambda_0 \tag{55}$$

$$D_*(\mathbf{x}) = \alpha_0 \quad \Leftrightarrow \quad \mu_1 + \mu_2 < 1 - \lambda_0 \tag{56}$$

$$D_{bet}(\mathbf{x}) = \alpha_0 \quad \Leftrightarrow \quad \frac{\mu_1}{M} + \mu_2 < 1 - \lambda_0 \tag{57}$$

These conditions are given a graphical representation in Figure 1. It is easy to see that rejection by D_* and D_{bet} is possible if and only if $\lambda_0 \leq 1 - 1/M$, whereas rejection by D^* requires only the weaker condition $\lambda_0 \leq 1$. Figure 2 shows that the three decision rules altogether can be seen as partitioning the feature space into four distinct regions with regard to classification and rejection (a finer partition would be obtained by considering the assignment to each individual class). A different view of these regions in feature space will be shown in Section 4 for a particular example.

3.3 Decision based on an incomplete learning set

In some applications, it is difficult or even impossible to build an exhaustive learning set, for one of the following reasons:

1. some classes have very small prior probabilities;

2. collecting samples from some classes would be either too dangerous or too costly;

3. the number of classes is very high;

4. an exhaustive list of all possible states of nature is not available.

A typical application domain in which such problems arise concerns the diagnosis of complex technological or natural systems[4]. In such contexts, the set Ω can be partitioned into a subset K of known classes, and a subset U containing those classes which are not represented in the training set. If U is assumed to contain only *unknown* states of nature, then there is no reason to make a distinction between different elements of U, which can then be considered as a singleton: $U = \{\omega_u\}$. For simplicity, we shall adopt this convention in the rest of the discussion. Nevertheless, the analysis can easily be extended to an arbitrary cardinality of U. We therefore assume Ω to be composed of $M+1$ elements $\Omega = \{\omega_1, \ldots, \omega_M, \omega_u\}$. As before, the possible actions are assignment to one (known or unknown) class, and rejection: $\mathcal{A} = \{\alpha_0, \alpha_1, \ldots, \alpha_M, \alpha_u\}$.

Since by definition the training set does not contain any example from class ω_u, there cannot be any evidence supporting directly the hypothesis that a pattern $\mathbf{x}$ might belong to that class, and consequently we always have $m(\{\omega_u\}) = 0$. However, it is possible to define certain contexts in which action α_u will be selected by one of the above decision rules, as will now be explained.

As before, the losses will be defined as 0 when the pattern under consideration has been correctly classified, 1 when it has been wrongly assigned to one of the known classes, and λ_0 in case of rejection. The consequences of wrongly assigning a pattern to the unknown class ω_u will be assumed to be characterized by a loss $\lambda_1 \geq 0$. This assumption can be interpreted by considering the assignment to ω_u as a particular kind of rejection, the consequences of which are not necessarily equivalent to other misclassification errors. The loss matrix is then:

	α_0	α_1	α_2	$\ldots$	α_{M-1}	α_M	α_u
ω_1	λ_0	0	1	$\ldots$	1	1	λ_1
ω_2	λ_0	1	0	$\ldots$	1	1	λ_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_{M-1}	λ_0	1	1	$\ldots$	0	1	λ_1
ω_M	λ_0	1	1	$\ldots$	1	0	λ_1
ω_u	λ_0	1	1	$\ldots$	1	1	0

Let us now express the different expected values of the losses induced by each possible action, and the corresponding decision rules. For clarity of presentation, only the simplest case will again be considered, in which the BBA m has singletons of Ω as only focal elements, in addition to Ω itself.

Decisions according to D_*

We have:

$$R_*(\alpha_i|\mathbf{x}) = \sum_{j \neq i} m(\{\omega_j\}) = 1 - Pl(\{\omega_i\}) \quad \text{for } i = 1, \ldots, M \tag{58}$$

$$
\begin{aligned}
R_*(\alpha_0|\mathbf{x}) &= \lambda_0 && (59)\\
R_*(\alpha_u|\mathbf{x}) &= \lambda_1 \sum_{j=1}^{M} m(\{\omega_j\}) = \lambda_1(1-m(\Omega)) && (60)
\end{aligned}
$$

As before, the decisions made by D_* can be represented in a diagram as a function of $\mu_1 = m(\Omega)$ and $\mu_2 = \max_{i=1,\ldots,M} m(\{\omega_i\})$. Assignment to one of the known classes is decided if the two inequalities are verified:

$$
\begin{aligned}
\mu_1 + \mu_2 &\geq 1 - \lambda_0 && (61)\\
\mu_1 + \mu_2 &\geq 1 - \lambda_1(1-\mu_1) && (62)
\end{aligned}
$$

Action α_0 is preferred over α_u iif:

$$
\mu_1 \leq 1 - \frac{\lambda_0}{\lambda_1} \tag{63}
$$

which requires $\lambda_0 \leq \lambda_1$. The decision regions are represented in the two situations ($\lambda_0 \leq \lambda_1$ and $\lambda_0 > \lambda_1$) in Figure 3. Note that for actions α_u and α_0 to be possible we must have respectively $\lambda_1 \leq 1 - 1/M$ and $\lambda_0 \leq 1 - 1/M$. The actions that can be taken for different values of λ_0 and λ_1 are represented graphically in Figure 5.

Decisions according to D^*

The upper expected losses are:

$$
\begin{aligned}
R^*(\alpha_i|\mathbf{x}) &= \sum_{j\neq i} m(\{\omega_j\}) + m(\Omega) = 1 - Bel(\{\omega_i\}) && (64)\\
&\quad \text{for } i = 1,\ldots,M\\
R^*(\alpha_0|\mathbf{x}) &= \lambda_0 && (65)\\
R^*(\alpha_u|\mathbf{x}) &= \lambda_1 \sum_{j=1}^{M} m(\{\omega_j\}) + \lambda_1 m(\Omega) = \lambda_1 && (66)
\end{aligned}
$$

In this case, we find that the risk associated to of α_u does not depend on m. Assignment to one of the known classes is therefore decided by D^* whenever $\mu_2 \geq \max(1-\lambda_0, 1-\lambda_1)$. The shape of the decision regions in the (μ_1, μ_2) graph is the same as in Figure 1, with α_u replacing α_0 if $\lambda_1 \leq \lambda_0$. The possible actions for different values of λ_0 and λ_1 are represented graphically in Figure 5.

Decisions according to D_{bet}

The expected losses relative to the pignistic probability distribution $BetP$ are:

$$
\begin{aligned}
R_{bet}(\alpha_i|\mathbf{x}) &= \sum_{j\neq i} m(\{\omega_j\}) + m(\Omega)\frac{M}{M+1} = 1 - BetP(\{\omega_i\}) && (67)\\
&\quad \text{for } i = 1,\ldots,M
\end{aligned}
$$

$$
\begin{aligned}
R_{bet}(\alpha_0|\mathbf{x}) &= \lambda_0 && (68)\\
R_{bet}(\alpha_u|\mathbf{x}) &= \lambda_1 \sum_{j=1}^{M} m(\{\omega_j\}) + m(\Omega)\frac{M\lambda_1}{M+1} && (69)\\
&= \lambda_1(1 - \frac{m(\Omega)}{M+1}) = \lambda_1(1 - BetP(\{\omega_u\}) && (70)
\end{aligned}
$$

Assignment to the known class with highest pignistic probability is therefore decided by D_{bet} if:

$$
\begin{aligned}
1 - \mu_2 - \frac{\mu_1}{M+1} &< \lambda_0 && (71)\\
1 - \mu_2 - \frac{\mu_1}{M+1} &< \lambda_1(1 - \frac{\mu_1}{M+1}) && (72)
\end{aligned}
$$

Graphically, this condition is satisfied if the point (μ_1, μ_2) is situated above the lines BG and CH in Figure 4. Rejection is preferred over assignment to the unknown class if:

$$
\lambda_0 < \lambda_1(1 - \frac{\mu_1}{M+1}) \Leftrightarrow \mu_1 \leq (M+1)(1 - \frac{\lambda_0}{\lambda_1}) \tag{73}
$$

The analysis of the necessary conditions for each action to be possible is slightly more complex than in previous cases. Graphically, the set of values of (μ_1, μ_2) for which α_u is selected is represented in Figure 4 by the intersecting region of triangles (AFD) and (ZEH). This intersection is non-empty iif $\lambda_1 \leq 1$ and:

$$
(M+1)(1 - \frac{\lambda_0}{\lambda_1}) \leq 1 \Leftrightarrow \frac{\lambda_0}{\lambda_1} \geq \frac{M}{M+1} \tag{74}
$$

The set of values of (μ_1, μ_2) for which α_0 is decided is represented by the intersection of triangle (AFD) and the polygon $(BZEO)$. By simple geometrical reasoning, it is straightforward to show that this area is non-empty iif the three conditions are simultaneously satisfied:

$$
\begin{aligned}
\lambda_0 &\leq \lambda_1 && (75)\\
\lambda_0 &\leq \frac{M}{M+1} && (76)\\
\lambda_1 &\geq \frac{\lambda_0}{M(1-\lambda_0)} && (77)
\end{aligned}
$$

These conditions are represented graphically in Figure 5. The region where all $M+1$ decisions are possible is delimited by the lines $\lambda_1 = \lambda_0$ and $\lambda_1 = (M+1)\lambda_0/M$, and by the curve of equation $\lambda_1 = \lambda_0/(M(1-\lambda_0))$.

4 Example

The decision rules described above will now be demonstrated on a real-world pattern recognition problem. The data consists of daily measurements of water quality parameters (pH, temperature, NO3 and NH4 concentrations) performed in the river Seine during two years. The sampling point was located upstream a drinking water production plant. A monitoring system under development will process continuous measurements of the same parameters to make a real-time diagnosis of overall water quality. The data considered here been used for prototyping of the system.

The basic approach consists in assigning measurement vectors to classes corresponding to different states of the river system. A clustering procedure was used to define four distinct classes in the available data. However, a requirement for the final system will be the ability to detect atypical situations and create new classes accordingly.

For better visualization of the results, the data dimension was reduced to 2 by principal component analysis (Figure 6). Our evidence-theoretic neural network model[14] was trained on the 731 input samples with 10 prototypes. In this model, the optimized parameters for each prototype i are a vector of coordinates $\mathbf{p}^i$, degrees of membership $u_1^i, \ldots, u_M^i$ to the M classes (with $\sum_{j=1}^M u_j^i = 1$), and a positive parameter γ^i defining the region of influence of prototype i. For an arbitrary input pattern $\mathbf{x}$, the consideration of prototype i induces a BBA m^i defined as:

$$m^i(\{\omega_q\}) = \alpha u_q^i \exp(-\gamma^i (d^i)^2) \quad \forall q \in \{1, \ldots, M\} \tag{78}$$

$$m^i(\Omega) = 1 - \alpha u_q^i \exp(-\gamma^i (d^i)^2) \tag{79}$$

where α is a parameter and d^i is the Euclidean distance between vectors $\mathbf{x}$ and $\mathbf{p}^i$. The n BBAs $m^1, \ldots, m^n$ are then combined using Dempster's rule to yield a final BBA m with focal elements $\{\omega_1\}, \ldots, \{\omega_M\}$, and Ω. The trained neural network therefore implements a mapping Γ from the feature space onto a $M+1$-dimensional space defined by

$$\Gamma : \mathbf{x} \mapsto (m(\{\omega_1\}), \ldots, m(\{\omega_M\}, m(\Omega)). \tag{80}$$

Contour plots of each component of this mapping are shown in Figures 7 to 11. Given a loss matrix, the output BBA for an arbitrary input pattern $\mathbf{x}$ provides the basis for making decisions according to the rules presented in the previous sections.

To illustrate these rules, we successively place ourselves under two different assumptions.

Assumption 1: Completeness of the training set

First, we assume the four classes found in the initial training data to be representative of all river states to be encountered. The reference set is then $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$, and the set of actions is defined as $\mathcal{A} = \{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4\}$. The loss matrix is defined as in Section 3.2. Decisions rules D_*, D^* and D_{bet} are then defined according to Equations 30 to 32, with R_*, R^* and R_{bet} given by Equations 46 to 48.

The data points and the decision regions in feature space are represented in Figures 12 and 13 for $\lambda_0 = 0.3$ and $\lambda_0 = 0.4$, respectively. As can be seen in these figures, decision rule D_* rejects those patterns which are situated in the neighborhood of class boundaries, which corresponds to the notion of "ambiguity reject" in the terminology introduced Dubuisson[4]. The same patterns are also rejected by D^* and D_{bet}, which additionally reject feature vectors located "far" from training samples. These rules therefore combine the notions of "ambiguity" and "distance" reject. The regions I to IV of Figure 2 are represented in feature space for this example in Figure 14.

Assumption 2: Incompleteness of the training set

In practice, there is no guarantee in this application that the clusters identified in the training set correspond to all possible states of the river. New clusters may appear due to accidental pollution or different weather conditions. A better formalization of the decision problem thus consists in assuming the existence of an additional class ω_u corresponding to all possible states of nature, excluding those already represented in the training set. Possible actions are then contained in $\mathcal{A} = \{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_u\}$. The loss matrix is now defined as in Section 3.3.

The various decision regions for D_* and D_{bet} are represented in Figures 15 and 16 for $\lambda_0 = 0.3$ and two different values of λ_1. Since $\lambda_1 > \lambda_0$, the decision regions for D^* are exactly identical in that case to those depicted in Figure 12. With these new assumptions, we can see that patterns in the vicinity of class boundaries are still rejected by D_* when they are close to training patterns, but are now assigned to the unknown class when they are very dissimilar from previously observed feature vectors. In contrast, decision rule D_{bet} assigns to class ω_u all "atypical" patterns, and rejects those situated in regions of ambiguity between or around data clusters. Note that the different shapes of the reject region for D_{bet} in Figures 15 and 16 can be explained by looking at Figure 4. Consider a point moving from the center of data cluster away from the training set. Its representation in Figure 4 moves along a straight line from point A to point F (μ_1 increases while μ_2 remains at its maximum value). This point crosses the reject region if:

$$1 - (M+1)\frac{\lambda_0}{\lambda_1} < \lambda_0\left(\frac{1}{\lambda_1} - 1\right) \Leftrightarrow \lambda_1 > \frac{M\lambda_0}{M - \lambda_0} \tag{81}$$

With $\lambda_0 = 0.3$, we must therefore have $\lambda_1 > 0.3243$, which is the case in Figure 16 but not in Figure 15.

5 Conclusions

The Dempster-Shafer theory of evidence offers a convenient framework for dealing with uncertainty in situations where the available information is limited or weak. Situations of this kind typically arise in pattern recognition when classification of feature vectors must be done based on a small number of training samples, or when the training set does not contain examples from all classes. In this paper, various decision strategies for pattern classification in the context of D-S have been presented. These strategies have been successively analyzed under the assumptions of completeness, and incompleteness of the learning set. When samples from all classes are available, possible actions are assignment to a class and rejection. As expected, patterns situated close to class boundaries tend to be rejected according to the three decision rules studied. Additionally, the more conservative strategies of upper and pignistic expected loss minimization also lead to the rejection of "atypical" patterns situated far from training samples in feature space. When the training set is not complete, the available options are assignment to one of the known classes, assignment to the unknown class, and rejection. By imposing some restrictions on the loss matrix, we have shown that various configurations of the decision regions can be induced by varying only two parameters corresponding to the costs of rejection and misclassification in the unknown class, respectively.

In practice, these theoretical considerations leave the designer of a pattern recognition system with many alternatives. In general, a "liberal" approach would logically lead to considering the lower expected loss of each possible action, while a more "cautious" approach would give more importance to pignistic or upper expected losses. Which strategy should be prefered obviously depends on the application domain. An interesting possibility to be explored in the context of interactive systems would consist in leaving the final decision to the user in case of conflict between different decision rules. The integration of the decision rules described in this paper with unsupervised procedures for detecting new classes based on the analysis of rejected patterns will also be subject to further research.

References

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

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

[3] W. F. Caselton and W. Luo. Decision making with imprecise probabilities: Dempster-shafer theory and application. *Water Resources Research*, 28(12):3071–3081, 1992.

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

[5] 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.

[6] G. Shafer. *A mathematical theory of evidence.* Princeton University Press, Princeton, N.J., 1976.

[7] 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.

[8] B. V. Dasarathy. *Nearest neighbor norms: NN pattern classification techniques.* IEEE Computer Society Press, Los Alamitos, Ca, 1991.

[9] P. Smets. The combination of evidence in the transferable belief model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(5):447–458, 1990.

[10] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–243, 1994.

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

[12] P. Smets. The degree of belief in a fuzzy event. *Information Sciences*, 25:1–19, 1981.

[13] P. Smets. Constructing the pignistic probability function in a context of uncertainty. In M. Henrion, R. D. Shachter, L. N. Kanal and J. F. Lemmer (Editors), *Uncertainty in Artificial Intelligence 5*, Elsevier, 1990.

[14] 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.

[15] 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, September 1995. Springer Verlag.

List of Figures

Biography

Thierry Denœux graduated in 1985 as an engineer from the Ecole Nationale des Ponts et Chaussées in Paris, and earned a PhD from the same institution in 1989. He obtained the "Habilitation à diriger des Recherches" from the Institut National Polytechnique de Lorraine in 1996. Until 1992, he worked as a project manager at LIAC (Laboratoire d'Informatique Avancée de Compiègne), a research center of Lyonnaise des Eaux, where he was in charge of research projects concerning the application of neural networks to forecasting and diagnosis. Dr. Denœux is currently an assistant professor at the Université de Technologie de Compiègne. His research interests include artificial neural networks, statistical pattern recognition, uncertainty reasoning and data fusion.

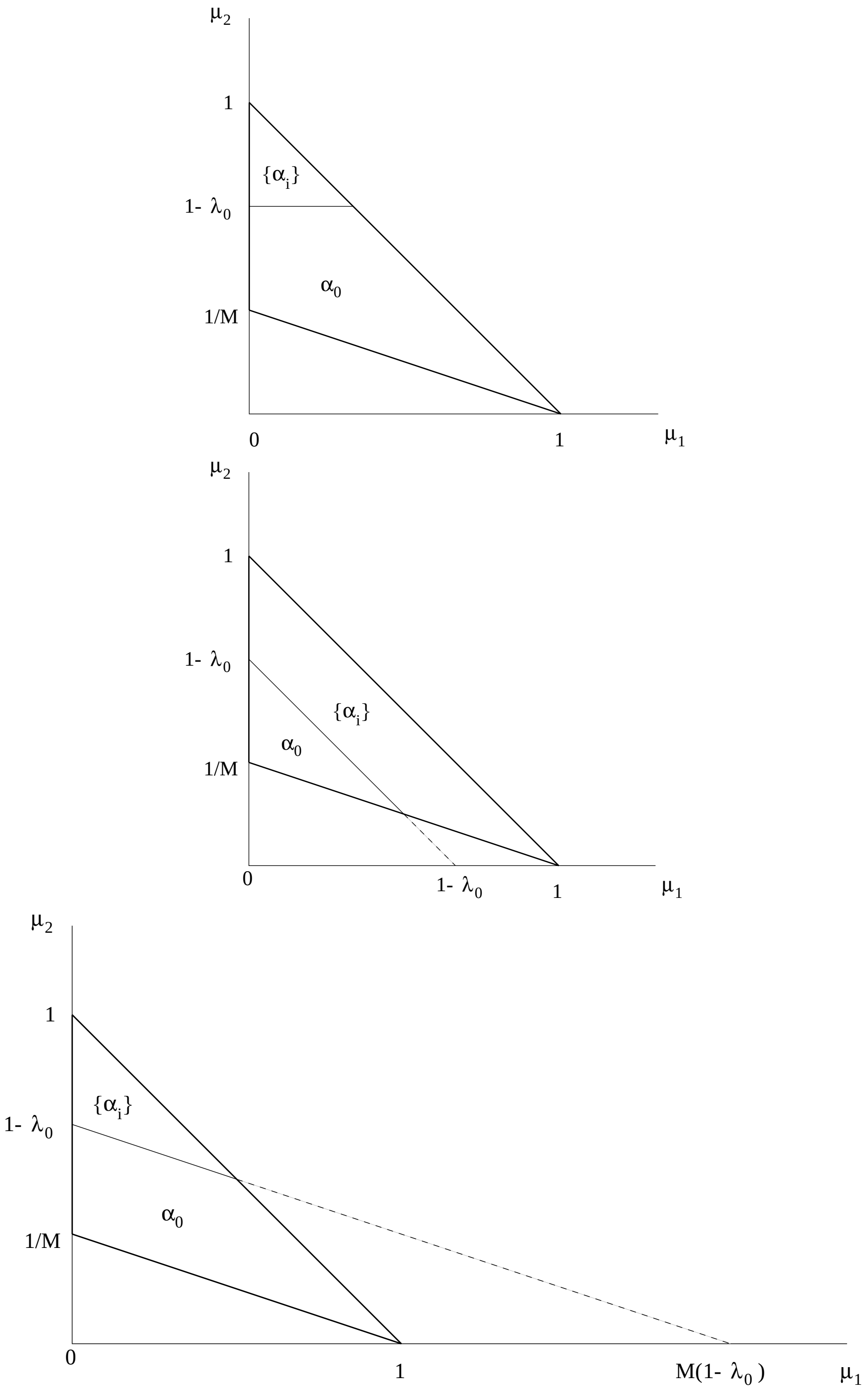

Figure 1: Decisions by D^* (up), D_* (middle) and D_{bet} (down) based on an exhaustive learning set, as a function of $\mu_1 = m(\Omega)$ and $\mu_2 = \max m(\{\omega_i\})$: acceptance $(\{\alpha_i\})$ and reject (α_0).

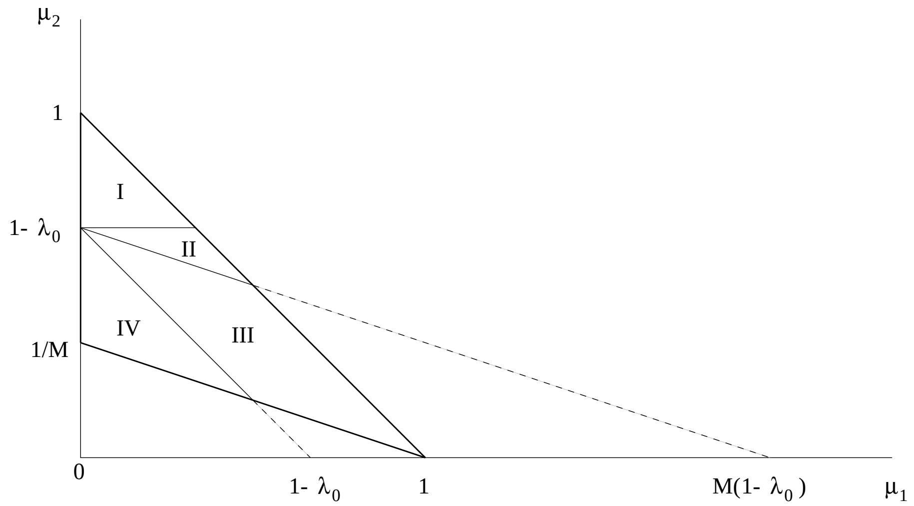

Figure 2: Various decision regions: I: acceptance by D_*, D^* and D_{bet}; II: acceptance by D_* and D_{bet}, reject by D^*; III: acceptance by D_*, reject by D_{bet} and D^*; IV: reject by D_*, D^* and D_{bet}.

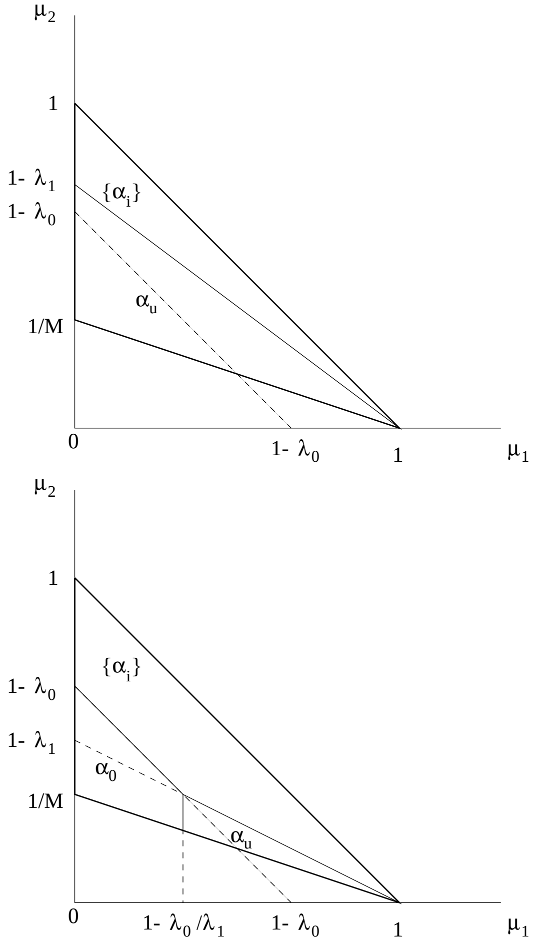

Figure 3: Decisions by D_* based on an incomplete learning set, as a function of $\mu_1 = m(\Omega)$ and $\mu_2 = \max m(\{\omega_i\})$, for $\lambda_1 < \lambda_0$ (up) and $\lambda_1 > \lambda_0$ (down).

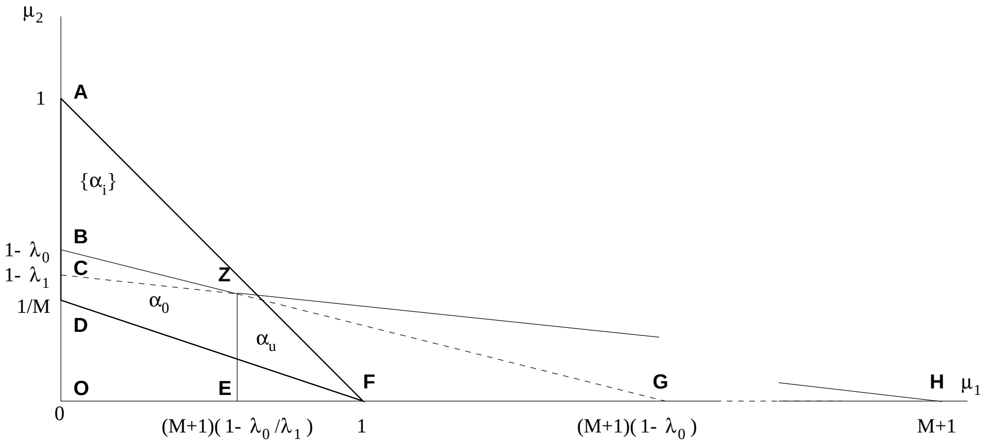

Figure 4: Decisions by D_{bet} based on an incomplete learning set, as a function of $\mu_1 = m(\Omega)$ and $\mu_2 = \max m(\{\omega_i\})$.

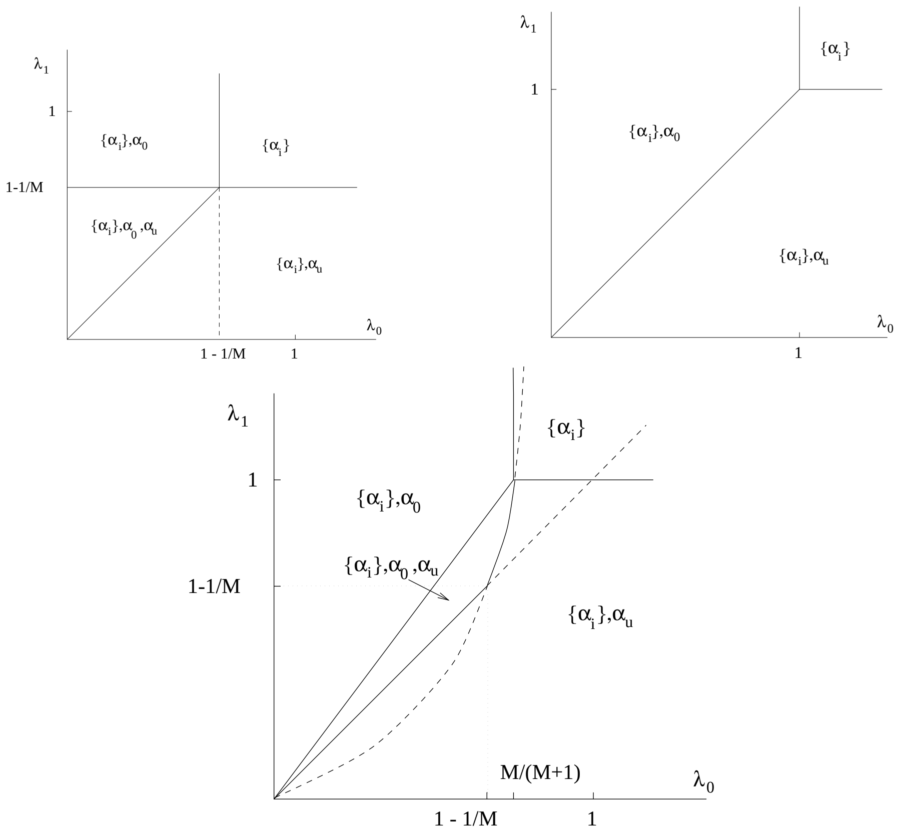

Figure 5: Possible actions by D_* (up-left), D^* (up-right) and D_{bet} (down) based on an incomplete learning set, as a function of λ_0 and λ_1.

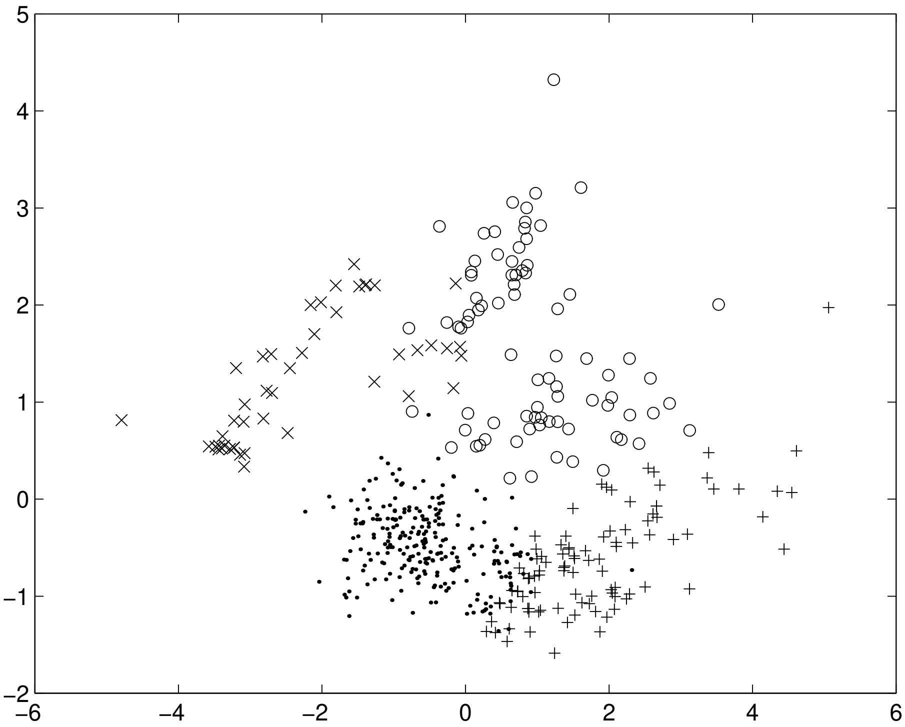

Figure 6: The water quality data. The axes are the first two principal components extracted from the original features. The four classes were determined by a clustering procedure applied to the original features.

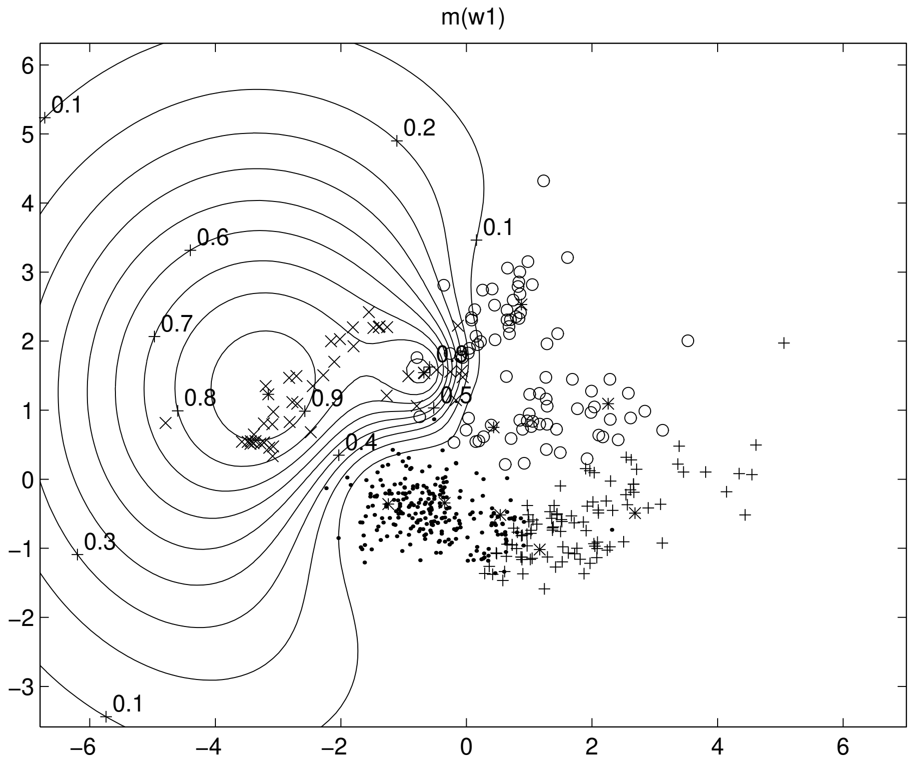

Figure 7: Contours of basic belief number $m(\{\omega_1\})$ in feature space, as learnt from the data by the neural network model. The prototypes are represented by $*$.

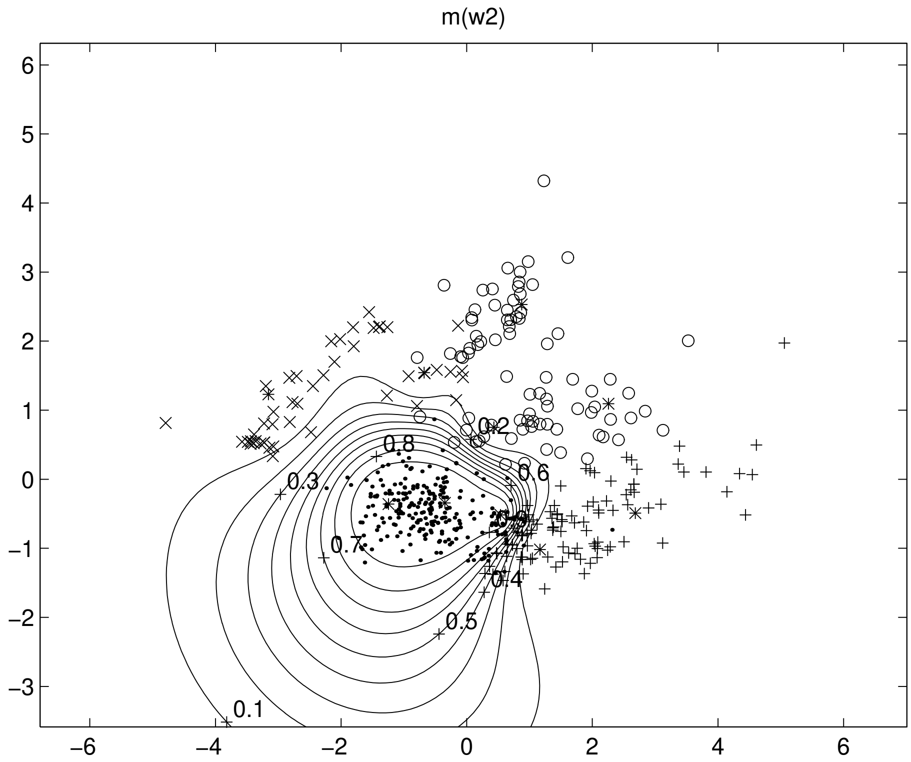

Figure 8: Contours of basic belief number $m(\{\omega_2\})$ in feature space, as learnt from the data by the neural network model. The prototypes are represented by $*$.

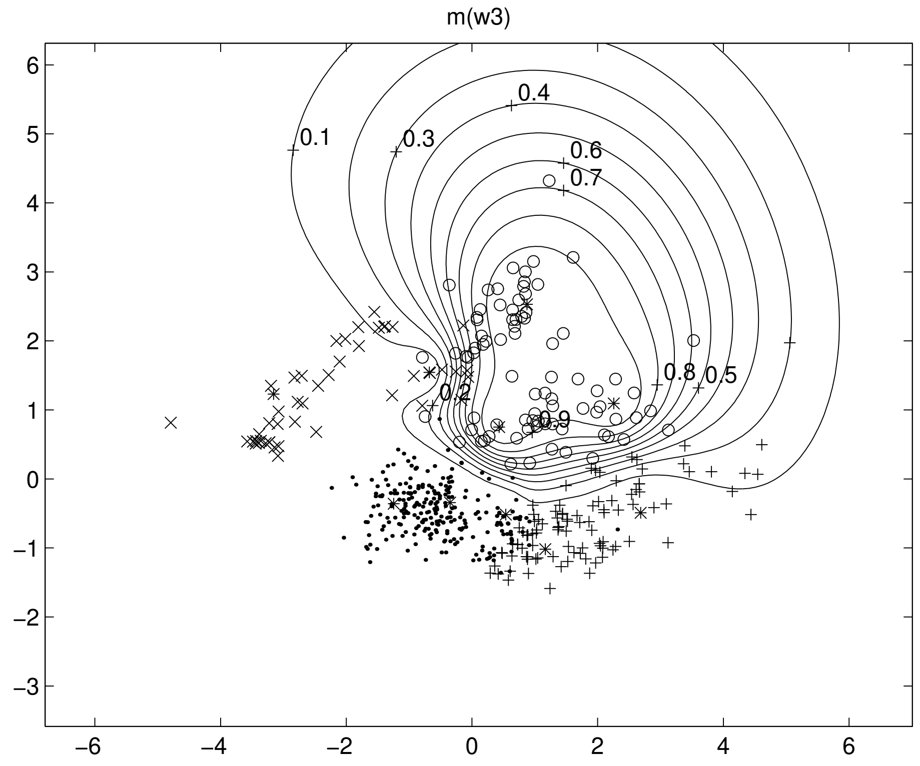

Figure 9: Contours of basic belief number $m(\{\omega_3\})$ in feature space, as learnt from the data by the neural network model. The prototypes are represented by $*$.

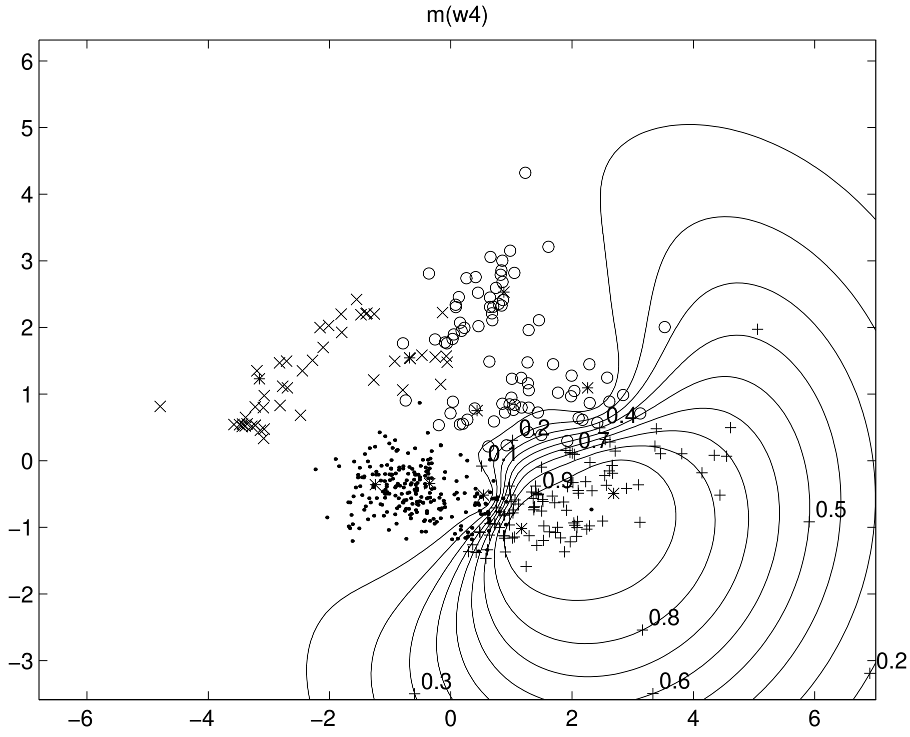

Figure 10: Contours of basic belief number $m(\{\omega_4\})$ in feature space, as learnt from the data by the neural network model. The prototypes are represented by $*$.

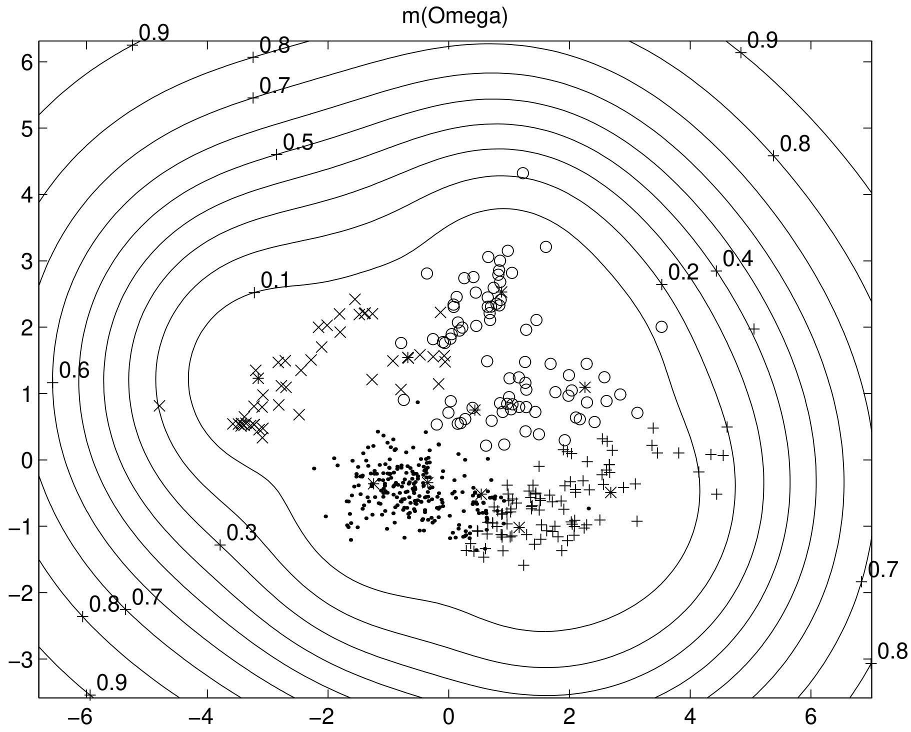

Figure 11: Contours of basic belief number $m(\Omega)$ in feature space, as learnt from the data by the neural network model. The prototypes are represented by $*$.

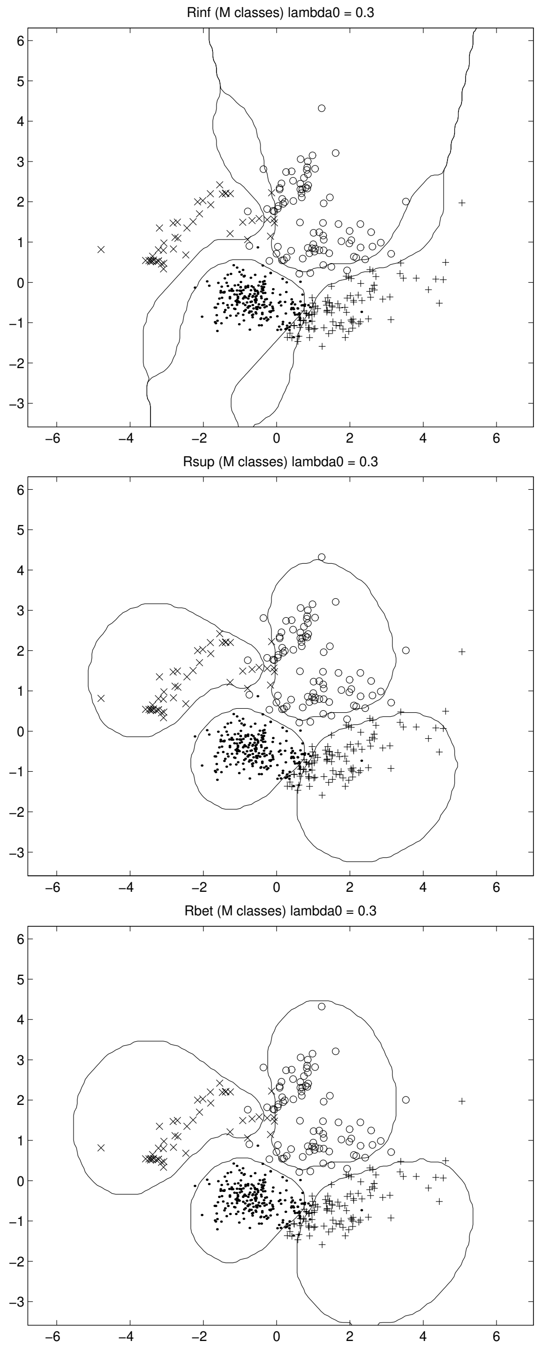

Figure 12: Decision regions for the water quality data, with assumption of completeness of the learning set: results with D_* (up), D^* (middle) and D_{bet} (down), for $\lambda_0 = 0.3$.

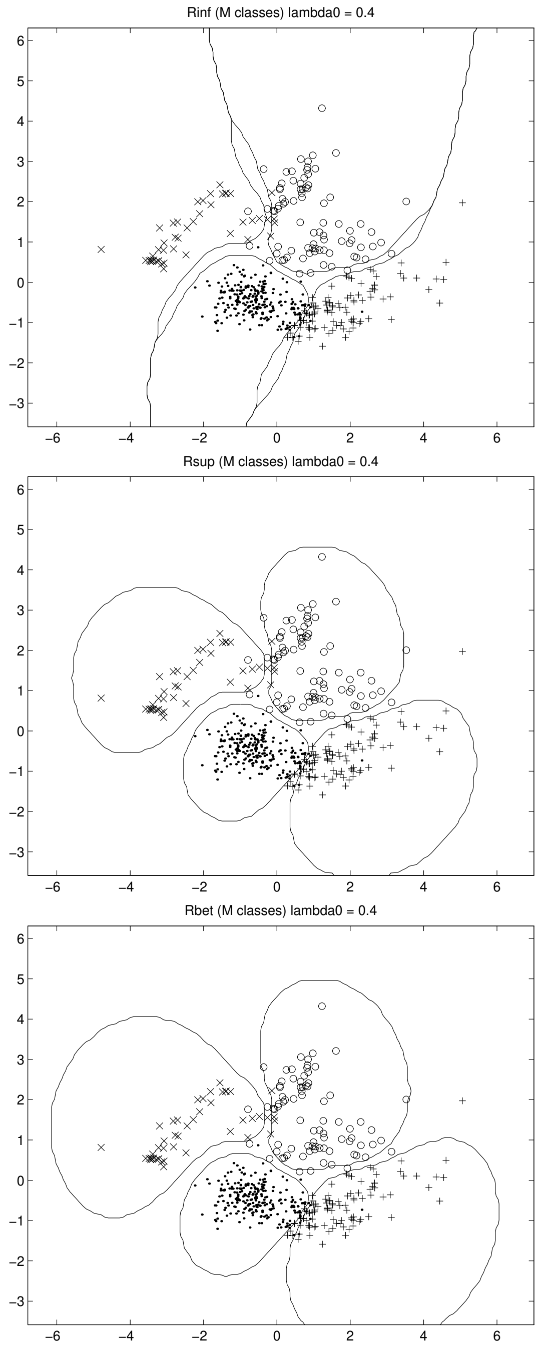

Figure 13: Decision regions for the water quality data, with assumption of completeness of the learning set: results with D_* (up), D^* (middle) and D_{bet} (down), for $\lambda_0 = 0.4$.

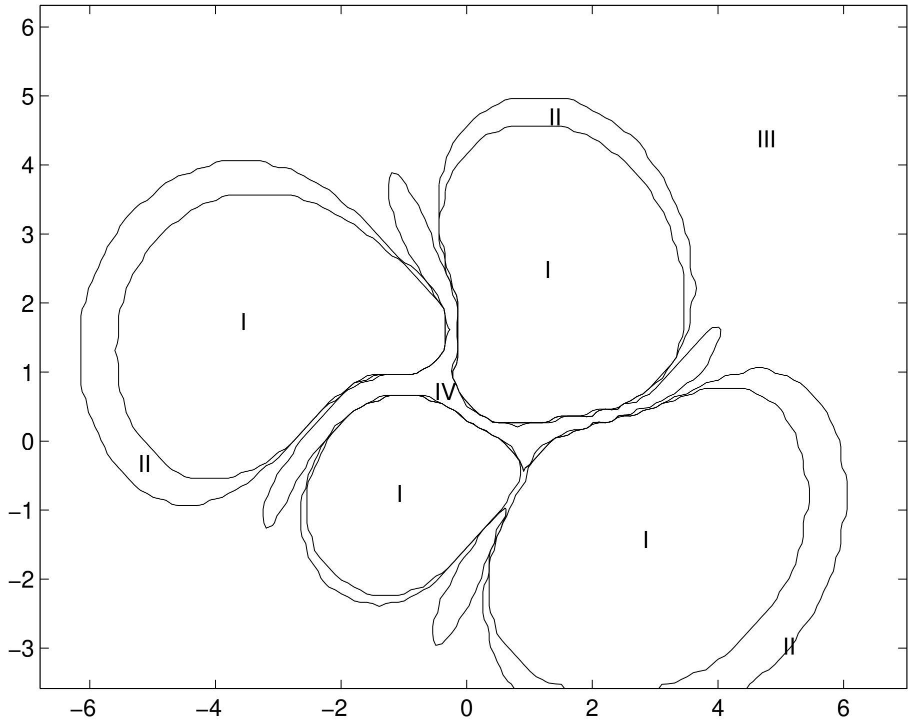

Figure 14: Decision regions in feature space: I: acceptance by D_*, D^* and D_{bet}; II: acceptance by D_* and D_{bet}, reject by D^*; III: acceptance by D_*, reject by D_{bet} and D^*; IV: reject by D_*, D^* and D_{bet}.

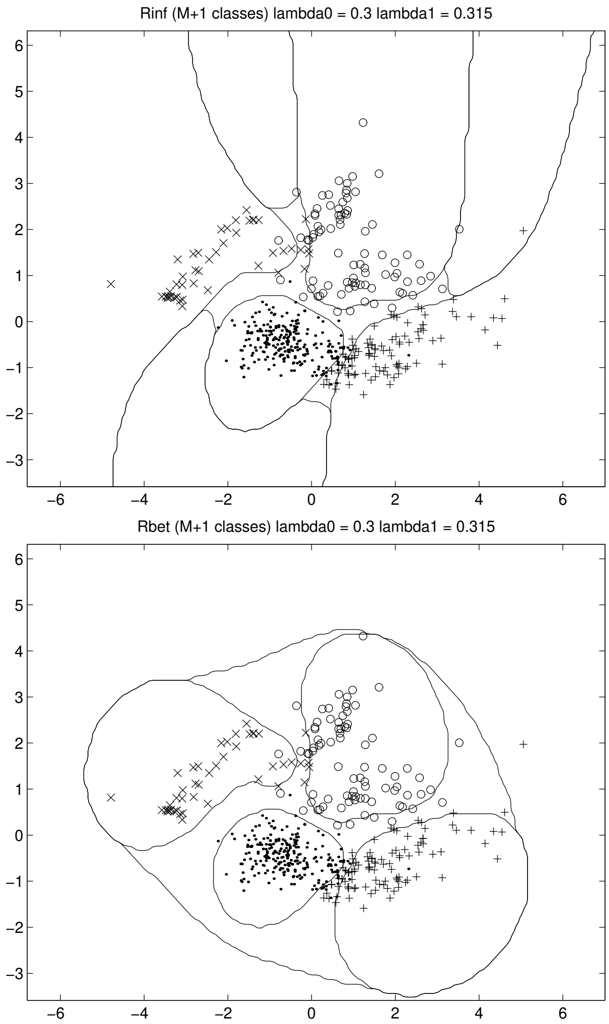

Figure 15: Decision regions for the water quality data, with assumption of incompleteness of the learning set: results with D_* (up) and D_{bet} (down), for $\lambda_0 = 0.3$ and $\lambda_1 = 0.315$.

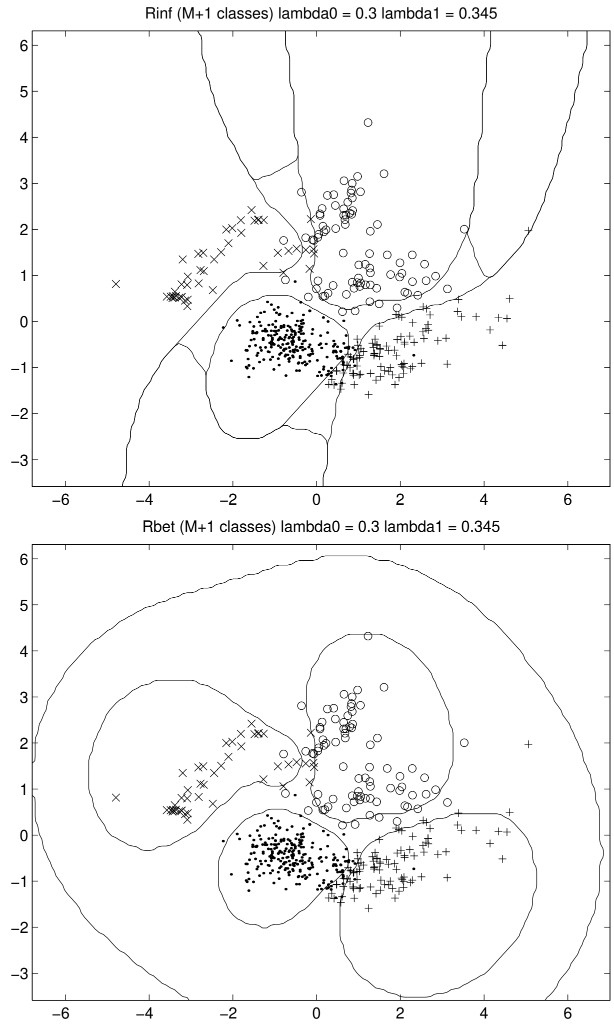

Figure 16: Decision regions for the water quality data, with assumption of incompleteness of the learning set: results with D_* (up) and D_{bet} (down), for $\lambda_0 = 0.3$ and $\lambda_1 = 0.345$.