Likelihood-based vs Distance-based Evidential Classifiers

Patrick Vannoorenberghe and Thierry Denœux
HEUDIASYC, U.M.R. CNRS 6599
Université de Technologie de Compiègne
BP 20529-F-60205 Compiègne Cedex, France

Abstract— **This paper presents and compares several evidential classifiers, i.e., classification rules based on the Dempster-Shafer theory of evidence. Three methods used in the majority of applications are compared, with emphasis on the techniques used to build belief functions from learning data. The methods are: the consonant method initially introduced by Shafer in the more general context of statistical inference, Appriou's separable method, and the distance-based classifier introduced by Denœux. These models can be derived with two decisions rules, based on the minimization of, respectively, lower and pignistic expected loss. Simulations on synthetic data demonstrate the performance of these techniques and allow to compare the behavior of the proposed models.**

Keywords— **Classification, Dempster-Shafer theory, evidence theory, belief functions.**

I. Introduction

The classification problem consists in assigning an input pattern $\mathbf{x}$ to a class, given a learning set $\mathcal{L}$ composed of n patterns $\mathbf{x}^i$ with known classification. Each pattern in $\mathcal{L}$ is represented by a p-dimensional feature vector $\mathbf{x}^i$ and its corresponding class label ω^i. In the last ten years, several solutions to this problem have been proposed, based on the Dempster-Shafer (DS) theory of evidence [1], [2]. Advantages of these techniques (description of the uncertainty on the prediction, possibility of rejecting a pattern and detecting unknown class) have been demonstrated in numerous papers [3], [4]. In particular, these classifiers are well adapted to applications where the available data come from multiple imperfect information sources (multisensor problems, environmental monitoring, medical diagnosis, classifier combination). The aim of this paper is to present three of the most cited models: two likelihood-based (LB) methods (the original method presented in the Shafer's book [1] and the "separable" method introduced by Appriou [3]), and a distance-based (DB) method introduced by Denœux [5]. These techniques differ by the way in which belief functions are assessed from data. An independent problem concerns the way decisions are made, given a belief function and decision costs; two common approaches to this problem are considered, leading to a total of 6 classification schemes. The paper is organized as follows. The basic concepts of evidence theory are first briefly introduced (Section II), and the three analyzed models are described in Section III. Finally, Section IV gives some experimental results using synthetic data. These simulations allow to understand the differences between the three proposed models in terms of behavior and classification performances.

II. Background

In this section, several concepts of the DS theory of evidence [1] are recalled, which allows to introduce notations used in this paper. Let $\Omega = \{\omega_q, q = 1, \cdots, Q\}$ denote a finite set of possible values for a variable y of interest. A basic belief assignment (bba) m on Ω is defined as a function from 2^Ω to $[0,1]$ verifying $m(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m(A) = 1$. Each subset $A \subseteq \Omega$ such as $m(A) > 0$ is called a focal element of m. From this, a belief function bel and a plausibility function pl are defined, respectively, as $bel(A) = \sum_{B \subseteq A} m(B)$ and $pl(A) = \sum_{A \cap B \neq \emptyset} m(B)$. The quantity $bel(A)$ can be interpreted as a measure of one's belief that hypothesis A is true. The plausibility $pl(A)$ can be viewed as the total amount of belief that could be potentially placed in A. Note that functions m, bel and pl are in one-to-one correspondence [1], and can be seen as three facets of the same piece of information. An α-discounted bba $m_\alpha(.)$ can be obtained from the original bba m as follows:

$$\begin{aligned} m_\alpha(A) &= \alpha m(A) \qquad \forall\, A \subseteq \Omega, A \neq \Omega && (1) \\ m_\alpha(\Omega) &= 1 - \alpha + \alpha m(\Omega) && (2) \end{aligned}$$

with $0 \leq \alpha \leq 1$. The discounting operation is useful when the source of information from which m has been derived is not fully reliable, in which case coefficient α represents some form of metaknowledge about the source reliability, which could not be encoded in m. Two pieces of evidence m_1 and m_2 can be aggregated with the Dempster's rule of combination (orthogonal sum $\oplus$), yielding to a unique belief function m defined as:

$$m(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{\sum_{B \cap C \neq \emptyset} m_1(B) m_2(C)} \quad \forall A \subseteq \Omega. \qquad (3)$$

Under the assumption of normality of the bba's ($m(\emptyset) = 0$), the use of this rule is possible only if m_1 and m_2 are not totally conflicting, i.e., if there exist two focal elements B and C of m_1 and m_2 satisfying $B \cap C \neq \emptyset$. If it is not the case, solutions exists, such as abandoning the normality assumption, or using other combination rules [6].

Let us assume that we have a bba m on Ω summarizing one's beliefs concerning the value of the unknown variable y, and we have to choose an action among a finite set of actions $\mathcal{A}$. A loss function $\lambda : \mathcal{A} \times \Omega \to \mathbb{R}$ is also assumed to be given, such that $\lambda(a, \omega)$ denotes the loss incurred if one chooses action a and $y = \omega$. Which action should we choose ? Based on rationality arguments, Smets [2] proposes to transform m into a probability function p_m on Ω (called the *pignistic* probability function) defined for all

$\omega \in \Omega$ as: $p_m(\omega) = \sum_{A \ni \omega} \frac{m(A)}{|A|}$, where $|A|$ denotes the cardinality of $A \subseteq \Omega$. In this transformation, the mass of belief $m(A)$ is distributed equally among the elements of A. Based on this probability, we can associate to each $a \in \mathcal{A}$ a risk, defined as the expected loss (relative to p_m) if one chooses action a:

$$R(a) = \sum_{\omega \in \Omega} \lambda(a, \omega) p_m(\omega). \qquad (4)$$

We then choose the action with the lowest risk. Alternatively, the decision process could be based on non-probabilistic extensions of the concept of mathematical expectation [7]. For example, the concept of lower expectation leads to the definition of the lower expected loss as

$$R_*(a) = \sum_{A \subseteq \Omega} m(A) \min_{\omega \in A} \lambda(a, \omega), \qquad (5)$$

which results in a different decision strategy.

In pattern classification, $\Omega = \{\omega_1, \ldots, \omega_Q\}$ is the set of classes, and the elements of $\mathcal{A}$ are, typically, the actions a_q of assigning the unknown pattern to each class ω_q. With 0-1 losses, defined as $\lambda(a_q, \omega_r) = 1 - \delta_{q,r}$ for $q, r \in \{1, \ldots, Q\}$, it can be shown [7] that the minimization of the pignistic risk R leads to choosing the class ω_0 with maximum pignistic probability, whereas the minimization of R_* leads to choosing the class ω_* with maximum plausibility. If an additional rejection action a_0 with constant loss λ_0 is added, then the pattern is rejected if $p_m(\omega_0) < 1 - \lambda_0$ using the first rule, and if $pl(\omega_*) < 1 - \lambda_0$ using the second rule [7].

III. Methods

As remarked in [8], there are two main approaches for building belief functions in classification problems: the LB approach relying on density estimation, and the DB approach in which a bba is directly constructed from the distances to reference pattens. These approaches are briefly described in the sequel.

A. Likelihood-based Methods

Let us assume the class-conditional probability densities $f(\mathbf{x}|\omega_\mathbf{q})$ to be known. Having observed $\mathbf{x}$, the likelihood function is a function from Ω to $[0, +\infty)$ defined as $L(\omega_q|\mathbf{x}) = \mathbf{f}(\mathbf{x}|\omega_\mathbf{q})$, for all $q \in \{1, \ldots, Q\}$. Shafer [1, p.238] proposed to derive from L a belief function on Ω defined by its plausibility function as:

$$pl(A) = \frac{\max_{\omega_q \in A}[L(\omega_q|\mathbf{x})]}{\max_q[L(\omega_q|\mathbf{x})]} \quad \forall A \subseteq \Omega. \qquad (6)$$

In pattern recognition, an application of this method (and a variant thereof) can be found in Ref. [9]. Note that pl defined by (6) is consonant, i.e., its focal elements are nested. For that reason, this first model will be called the "consonant likelihood-based" (CLB) model.

Starting from axiomatic requirements, Appriou [3] proposed another method based on the construction of Q belief functions $m_q(.)$. The idea consists in taking into account separately each class and evaluating the degree of belief given to each of them. In this case, the focal elements of each bba m_q are the singleton $\{\omega_q\}$, its complement $\overline{\omega_q}$, and Ω. Appriou actually obtained two different models with similar performances [10]. According to Appriou (personnal communication), one of these models seems to be preferable on theoretical grounds, because it is consistent with the generalized Bayes theorem introduced by Smets [6]. This model, hereafter referred to as the Separable Likelihood-based (SLB) method, has the following expression:

$$\begin{aligned} m_q(\{\omega_q\}) &= 0 && (7)\\ m_q(\overline{\omega_q}) &= \alpha_q(1 - R.L(\omega_q|\mathbf{x})) && (8)\\ m_q(\Omega) &= 1 - \alpha_q(1 - R.L(\omega_q|\mathbf{x})), && (9) \end{aligned}$$

where α_q is a coefficient that can be used to model external information such as sensor reliability, and R is a normalizing constant that can take any value in the range $(0, (\max_q(L(\omega_q|\mathbf{x})))^{-1}]$. Parameter R is somewhat arbitrary, but the principle of maximum uncertainty leads to choosing the largest allowed value, which results in the least specific bba. With these Q belief functions and using the Demspter's rule of combination, a unique belief function m is obtained as $m = \bigoplus_q m_q$.

B. Distance-based Method

A totally different approach was introduced by Denoeux [5]. In this method, a bba is constructed directly, using as a source of information the training patterns $\mathbf{x^i}$ situated in the neighborhood of the pattern $\mathbf{x}$ to be classified. If the k nearest neighbors (according to some distance measure) are considered, we thus obtain k bba's that are combined using the Dempster's rule of combination. The initial method was later refined to allow parameter optimization [11], and a neural-network-like version was recently proposed [4]. This version, which will be considered here, uses a set of prototypes that are determined to minimize an error function. Each prototype can be viewed as a piece of evidence that influences the belief concerning the membership class of $\mathbf{x}$. A belief function m^i associated to each prototype i is then defined for all $q \in \{1, \cdots, Q\}$ as:

$$\begin{aligned} m^i(\{\omega_q\}) &= \alpha^i \phi^i(d^i) && (10)\\ m^i(\Omega) &= 1 - \alpha^i \phi^i(d^i) && (11)\\ m^i(A) &= 0 \ \forall A \in 2^\Omega \setminus \{\{\omega_q\}, \Omega\} && (12) \end{aligned}$$

where d^i is the Euclidean distance to the i-th prototype, α^i is a parameter attached to prototype i, and $\phi^i(.)$ is a decreasing function defined as $\phi^i(d^i) = \exp[-\gamma^i(d^i)^2]$. In this expression, γ^i is a positive parameter associated to prototype i. The belief functions m^i for each prototype are then aggregated using the Dempster's rule of combination.

C. Parameter optimization

In the application of the LB methods, the first difficulty concerns the estimation of likelihood functions. Several density estimation can be used, including parametric methods based, e.g., on a Gaussian model, and non parametric kernel methods. In the simulations presented in the sequel, we chose to use a Gaussian mixture model together with the EM algorithm as an estimation technique [12].

As remarked by Bastière [8], there is no general technique for evaluating the discounting coefficients α_q in the separable method. In this paper, we propose to use the same approach as used by Denoeux [4] for the DB method, i.e., minimizing the following error criterion:

$$E(\alpha) = \sum_{i=1}^{n} \sum_{q=1}^{Q} (p^i(\omega_q) - u_q^i)^2 \qquad (13)$$

where u_q^i is the class indicator of pattern $\mathbf{x^i}$ ($u_q^i = 1$ if $\omega^i = \omega_q$), and $p^i(\omega_q)$ is the pignistic probability of ω_q for vector $\mathbf{x^i}$. In the same manner, it is possible to define an error criterion based on the plausiblity function E_* where p is replaced with pl. These different techniques associated to the two decision rules presented in the Section 2 are demonstrated in the sequel.

IV. Results

For the following simulations, a learning set $\mathcal{L}$ was generated using 3 classes containing 50 bidimensional vectors each. Each vector $\mathbf{x}$ from class q was generated by first drawing a vector $\mathbf{z}$ from a Gaussian $f(\mathbf{z}|\omega_\mathbf{q}) \sim \mathcal{N}(\mu_\mathbf{q}, \mathbf{\Sigma_q})$, and applying a non linear transformation $\mathbf{z} \mapsto \mathbf{x} = \exp(0.3\,\mathbf{z})$ to obtain non-Gaussian data. The means of the 3 Gaussian distributions were taken as: $\mu_1 = (-1,-1)'$, $\mu_2 = (1,2)'$, $\mu_3 = (-1.5,2)'$ and the variance matrices were of the form $\Sigma_q = D_q A D_q'$ with

$$A = \begin{pmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{3}/3 \end{pmatrix} \quad D_q = \begin{pmatrix} \cos\theta_q & -\sin\theta_q \\ \sin\theta_q & \cos\theta_q \end{pmatrix}$$

and $\theta_1 = \pi/3$, $\theta_2 = \pi/2$, $\theta_3 = -\pi/3$.

A. Decision regions

The decision regions for the CLB and DB methods, with the two decision rules are shown in Figures 1 and 2 (the decision regions for the SLB method are somewhat similar to those of the CLB method, and are consequently not shown here for lack of space). In these figures, mixture component centers and prototypes are represented as asterisks ($*$). For the LB methods, likelihood functions were estimated using a Gaussian mixture model with $k = 2$ modes per class, and the parameters were estimated by the EM algorithm [12]. For the DB method, we chose by analogy two prototypes per class whose locations were initialized using the c-means algorithm. Concerning the separable method, parameters α_q were fixed at the following values: $\alpha_1 = 0.4$, $\alpha_2 = \alpha_3 = 0.9$. The value of the rejection cost λ_0 was set at 0.4. The specific form of the belief functions for the CLB and SLB methods impose that $\max_q pl(\{\omega_q\}) = 1$. For this reason, only the DB method allows to reject patterns using the maximum plausibility decision rule. As can be seen from these figures, both the inference method and the decision rule have a dramatic influence on the shape of the decision regions.

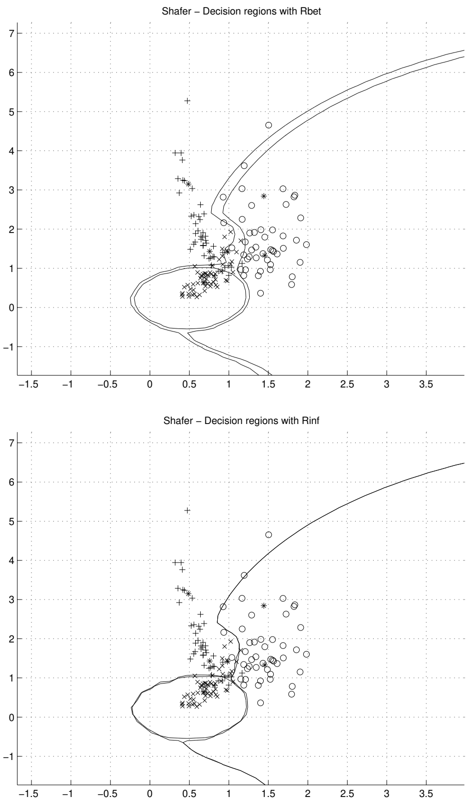

Fig. 1. Decision regions for the CLB Method (Shafer) with R (up) and R_* (down) for rejection loss $\lambda_0 = 0.4$, ($\omega_1 = \times$, $\omega_2 = \circ$, $\omega_3 = +$)

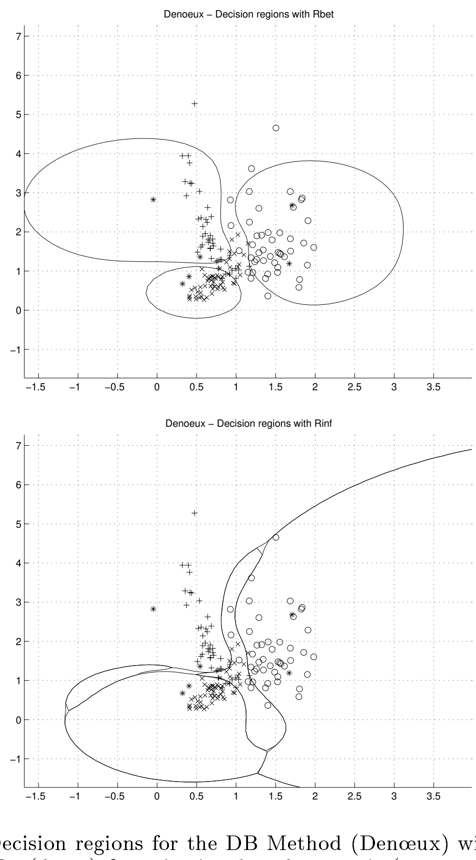

Fig. 2. Decision regions for the DB Method (Denœux) with R (up) and R_* (down) for rejection loss $\lambda_0 = 0.4$, ($\omega_1 = \times$, $\omega_2 = \circ$, $\omega_3 = +$)

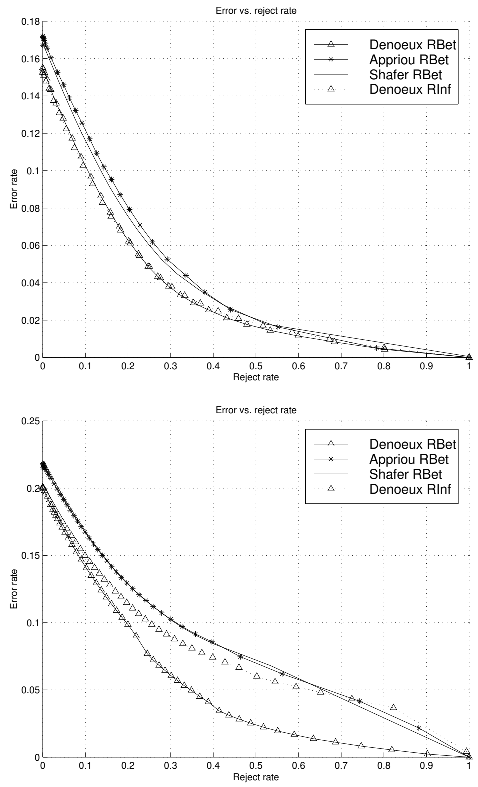

Fig. 3. Test error rate vs. rejection rate for the three methods with the two decision rules without (up) and with outliers (down)

B. Performance comparison

To compare the performances of the 3 models, a test set $\mathcal{T}$ was generated using the same distribution as $\mathcal{L}$ with $15,000$ samples. The experiment was repeated ten times with independent training sets. The number of components in the mixture model (for the LB methods) and the number of prototypes (for the DB method) were optimized using a cross-validation set. The upper part of Fig. 3 shows the error rate vs. the reject rate for the 3 methods and the 2 decision rules. For the CLB and SLB methods associated to the maximum plausibility decision rule, there are no rejected patterns. For this data set, all the proposed models obtain comparable performances. However, the DB model yields lower error rates as compared to the LB model without rejection. Moreover, if the classes have different prior probabilities, this gain is further increased.

To demonstrate the robustness of these methods, the test set $\mathcal{T}$ was then contaminated with $1,500$ outliers with uniform distribution and random class labels. The lower part of Fig. 3 presents the error rates of the different methods as functions of the rejection rates. The most robust decision rule seems to be the DB method with the maximum pignistic probability rule. This observation is easily explained by the shapes of the decision regions.

V. Conclusion

This paper has focused on pattern recognition techniques based on the DS theory of evidence. Three models and two decision rules have been presented and discussed, and a method for evaluating parameters (discounting coefficients) of the LB models has been introduced. From experimental results, we can draw several conclusions:

- The output belief functions take very different forms from the 3 methods studied (more or less specific, consonant or not); consequently, the uncertainty related to the prediction is not represented in the same manner by the 3 models.
- All the proposed models (except LB methods with the maximum plausibility decision rule) obtain comparable performances in the case of "standard" data; however, the DB method associated to pignistic risk minimization seems to be more robust to outliers than the other methods.

Although these conclusions cannot be blindly generalized to all classification tasks, they seem to be sufficiently explicit to guide the choice of a model. An important remark concerns the application of such techniques to multisensor data fusion (military applications for example). In such applications, each sensor is considered as an independent source of information. The three proposed techniques can be applied by considering separately each sensor (associated with a confidence coefficient) and combining the information with the Dempster's rule of combination. As far as we know, only the DB method can cope with a learning set composed of samples with partially known labels [13].

References

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

[2] P. Smets and R. Kennes, "The Transferable Belief Model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[3] A. Appriou, "Uncertain data aggregation in classification and tracking processes," in *Aggregation and Fusion of imperfect information* (B. Bouchon-Meunier, ed.), pp. 231–260, Heidelberg: Physica-Verlag, 1998.

[4] T. Denoeux, "A neural network classifier based on Dempster-Shafer theory," *IEEE Transactions on Systems, Man and Cybernetics A*, vol. 30, no. 2, pp. 131–150, 2000.

[5] T. Denoeux, "A k-nearest neighbour classification rule based on Dempster-Shafer theory," *IEEE Transactions on Systems Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[6] P. Smets, "Belief functions: The disjunctive rule of combination and the Generalized Bayesian Theorem," *International Journal of Approximate Reasoning*, vol. 9, pp. 1–35, 1993.

[7] T. Denoeux, "Analysis of evidence-theoretic decision rules for pattern classification," *Pattern Recognition*, vol. 30, no. 7, pp. 1095–1107, 1997.

[8] A. Bastière, "Methods for multisensor classification of airbone targets integrating evidence theory," *Aerospace Science and Technology*, no. 6, pp. 401–411, 1998.

[9] H. Kim and P. Swain, "Evidential reasoning approach to multisource-data classification in remote sensing," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 8, pp. 1257–1265, 1995.

[10] S. Fabre, A. Appriou, and X. Briottet, "Presentation and description of two classification methods using data fusion based on sensor management," *Information Fusion*, vol. 2, pp. 49–71, 2001.

[11] L. Zouhal and T. Denoeux, "An evidence-theoretic k-nn rule with parameter optimization," *IEEE Transactions on Systems, Man and Cybernetics C*, vol. 28, pp. 263–271, May 1998.

[12] G. J. M. Lachlan and T. Krishnan, *The EM Algorithm and Extensions*. New-York: Wiley, 1997.

[13] T. Denoeux and L. Zouhal, "Handling possibilistic labels in pattern classification using evidential reasoning," *Fuzzy Sets and Systems*, vol. 122, no. 3, pp. 47–62, 2001.