Basic Belief Assignment in the Framework of Information Theory

Eric Lefevre, Patrick Vannoorenberghe and Olivier Colot
Laboratoire **P**erception **S**ystèmes et **I**nformation, UPRES EA 2120
Université/INSA de ROUEN
Place Emile Blondel, BP 08
76131 Mont Saint-Aignan Cedex, France.
Phone : +33-(0)2-35-52-84-05, Fax : +33-(0)2-35-52-84-83
E-mail : Patrick.Vannoorenberghe@univ-rouen.fr ; Olivier.Colot@insa-rouen.fr

ABSTRACT : Within the framework of pattern recognition, many methods of classification were developed. More recently, techniques using the Dempster-Shafer's theory or evidence theory tried to deal with the problem related to the management of the uncertainty and data fusion. In this paper, we propose a classification method based on the Dempster-Shafer's theory and information criteria. After an original basic belief assignment method, we introduce an attenuation factor of belief functions based on the dissimilarity between probability distributions. Results on synthetic data sets are given in order to illustrate the proposed methodology.

KEYWORDS : Data Fusion ; Information Criteria ; Dempster-Shafer's Theory.

INTRODUCTION

Data analysis and processing are two important tasks in today's information society. The data management becomes essential when the information is imperfect, that is to say imprecise and uncertain. Traditionally, probability theory, which is inadequate in some cases as well known (see Bezdek (1994)), is used for dealing with imperfect data. In the recent past, other models have been developed for handling imprecise knowledge (theory of fuzzy sets developed by Zadeh (1978)), possibility theory developed by Dubois (1986,1994) or uncertain information (theory of belief functions of Shafer (1976)). In this paper, we deal with a classification method of imperfect data sets using evidence theory proposed by Shafer (1976), Smets (1993) and Smets (1994). Recently, in this context, a new approach using neighborhood information has been developed by Denoeux (1998). Each nearest neighbor of a pattern to be classified is considered as an item of evidence. The resulting belief assignment is also defined as a function of the distance between the pattern and its neighbor. We propose an alternative solution to this classification method in initializing the belief functions using information criteria. This paper is organized as follows. In section entitled DEMPSTER-SHAFER'S THEORY, we introduce notations allowing to describe the Dempster-Shafer's Theory of evidence. Section entitled METHODOLOGY OF CLASSIFICATION PROCESS is devoted to present the proposed methodology. This work is applied to synthetic data (section entitled SIMULATIONS).

DEMPSTER-SHAFER'S THEORY

In this section, a brief overview of the Evidence Theory given in Shafer (1976) is provided. Let Θ represents the set of hypotheses H, called the frame of discernment. The knowledge about the problem induces a basic belief assignment which allows to define a belief function m from 2^{Θ} to $[0,1]$ as :

$$m(\emptyset) = 0 \tag{1}$$

$$\sum_{H_n \subseteq \Theta} m(H_n) = 1. \tag{2}$$

Subsets H_n of Θ such that $m(H_n) > 0$ are called focal elements of m. From this basic belief assignment m, the credibility $Bel(H_n)$ and plausibility $Pl(H_n)$ can be computed using the equations :

$$Bel(H_n) = \sum_{A \subseteq H_n} m(A) \tag{3}$$

$$Pl(H_n) = \sum_{H_n \cap A \neq \emptyset} m(A). \tag{4}$$

The value $Bel(H_n)$ quantifies the strength of the belief that event H_n occurs. These functions (m, Bel and Pl) are derived from the concept of lower and upper bounds for a set of compatible probability distributions. In addition, Dempster-Shafer's theory allows the fusion of several sources using the Dempster's combination operator. It is defined like the orthogonal sum (commutative and associative) following the equation :

$$m(H_n) = m_1(H_n) \oplus m_2(H_n) \oplus .... \oplus m_M(H_n). \tag{5}$$

For two sources S_i and S_j, the aggregation of evidence can be written :

$$\forall\ H_n \subseteq \Theta \qquad m(H_n) = \frac{1}{\mathcal{K}} \sum_{A \cap B = H_n} m_i(A).m_j(B) \tag{6}$$

where $\mathcal{K}$ is defined by :

$$\mathcal{K} = 1 - \sum_{A \cap B = \emptyset} m_i(A).m_j(B). \tag{7}$$

The normalization coefficient $\mathcal{K}$ evaluates the conflict between two sources. An additionnal aspect of the Dempster-Shafer's theory concerns the attenuation of the basic belief assignment m by a coefficient α. The attenuated belief function can be written as :

$$\forall\ H_n \in 2^\Theta \qquad m_{(\alpha,j)}(H_n) = \alpha_j.m_j(H_n) \tag{8}$$

$$m_{(\alpha,j)}(\Theta) = 1 - \alpha_j + \alpha_j.m_j(\Theta). \tag{9}$$

METHODOLOGY OF CLASSIFICATION PROCESS

The proposed methodology can be decomposed in three steps. The first one corresponds to the basic belief assignment based on analysis of the learning set (see section Basic Belief Assignment). The second one consists in attenuating the belief structure by means of a coefficient α derived from the Hellinger's distance between probability distributions. This one has a lower bound equal to 0 and an upper bound equal to 1. This distance allows to estimate the similarity between two probability distributions and, in particular to check if the gaussian assumption is correct (see subsection Belief function attenuation). Finally, the belief structures defined for each source of information are aggregate in order to decrease significantly the uncertainty for the later classification process (see subsection Information sources aggregation and decision)

BASIC BELIEF ASSIGNMENT

An important aspect of the discrimination concerns learning knowledge using data. In evidence theory, this problem leads to initialize the belief functions m. We make the hypothesis that the data extracted from one information source S_j among M sources can be represented as a gaussian distribution. This assumption is obtained by means of the study of the learning database defined as : $\mathcal{X} = \{\mathcal{X}_{(n;1)}, \ldots, \mathcal{X}_{(n;M)}\}$ where $\mathcal{X}_{(n;j)} = \{X_{(n;j)}\}$ represents the set of vectors X_n classified in the hypothesis H_n. For the value x_j, we determine the membership probability according to the hypothesis as :

$$P(x_j/H_n) = \frac{1}{\sigma_{(n;j)}\sqrt{2\pi}} e^{\frac{-(x_j - \mu_{(n;j)})^2}{2\sigma^2_{(n;j)}}} = \mathcal{N}(\mu_{(n;j)}, \sigma_{(n;j)}) \tag{10}$$

The pair $(\mu_{(n;j)}, \sigma_{(n;j)})$ represent respectively the mean and standard deviation computed after the learning step for each hypothesis H_n and each source S_j. In addition, we compute a third gaussian distribution representing the conjunction of the two hypotheses. This new distribution has the following mean and standard deviation :

$$\mu_{(n,n');j} = \frac{\mu_{(n;j)} + \mu_{(n';j)}}{2} \tag{11}$$

$$\sigma_{(n,n');j} = \max(\sigma_{(n;j)}, \sigma_{(n';j)}). \tag{12}$$

This assumption allows to generate the belief functions. Let X' a M component vector to be classify with $X' = [x'_1, \ldots, x'_M]^t$. The belief given for each hypothesis H_n depends on the membership probability with respect to :

$$\forall\ H_n \in 2^\Theta \qquad m_j(H_n) = R_j * P(x'_j/H_n) \tag{13}$$

The coefficient R_j is a normalization coefficient. It allows to verify the condition given by equation (2). It is defined as :

$$\forall\ H_n \in 2^\Theta \qquad R_j = \frac{1}{\sum_{H_n \in 2^\Theta} P(x'_j/H_n)}. \tag{14}$$

BELIEF FUNCTION ATTENUATION

After this learning step, the main idea is to resume the information contained in each source S_j by means of an optimum histogram computed on the set $\bigcup_{i \in H_n} \mathcal{X}_{(i;j)}$ in the sense of the maximum likelihood and of a mean square cost. This histogram will be used in order to establish the relevance of a source of information. First, we have to build an approximation of the unknown probability distribution with only the N-samples given in each source. That is done by means of a histogram building which is led by the use of an information criterion. We will see that different information criteria initially designed for model selection can be used.

Probability density approximation

Let be $A_1 A_2 \dots A_p \dots A_q$ an initial partition Q of an unknown distribution λ with $q = Card(Q)$. The aim is to approximate λ with a histogram built on a subpartition $C = B_1 B_2 \dots B_c$ of Q with c bins such as $c \leq q$. The probability distribution $\hat{\lambda}_C$ built with C is an optimum estimation of λ according to a cost function to define. C results from an information criterion called IC issued from the basic Akaike's information criterion (AIC) proposed in Colot (1993), AIC^* or ϕ^* as proposed in Colot (1994) which are respectively Hannan-Quinn's criterion and Rissanen's criterion. These criteria have the following form :

$$IC(c) = g(c) - \sum_{B \in C} \hat{\lambda}_c \ln \frac{\hat{\lambda}_c(B)}{\nu_c(B)} \tag{15}$$

where $g(c)$ is a penalty term which differs from one criterion to another one. Let us note ε a random process of a probability distribution λ supposed absolutely continuous to an *a priori* given probability distribution ν. Let ω be the set of all values taken by ε. The probability density f of λ is given by the Radon-Nycodim's derivative such as :

$$\forall\ \epsilon \in \omega \qquad f(\lambda, \epsilon) = \frac{d\lambda}{d\nu}(\epsilon). \tag{16}$$

The probability density f is approximated from N samples (ϵ_k) of ε by means of a histogram with c bins obtained with these N values. An optimum histogram to approximate the unknown probability distribution λ is obtained in two steps. The first one consists in merging two contiguous bins in a histogram with c bins among the $(c-1)$ possible fusions of two bins. This is made by minimizing the IC criterion. The second one consists in finding the "best" histogram with c bins. The optimum histogram with $c = c_{opt}$ bins is the one which minimizes IC.

Maximum likelihood estimator for a partition Q

Let Q be a partition with q bins and let $\epsilon_1 \dots \epsilon_N$ be a N-observation sample and let be λ_Q the probability distribution according to Q. The maximum likelihood estimator $\hat{\lambda}_Q$ of λ_Q is given by the following equation :

$$\forall\ p \in \omega \qquad \hat{\lambda}_Q(A_p) = \frac{1}{N} \sum_{\epsilon_k \in A_p} \epsilon_k \tag{17}$$

where A_p is a bin of the partition Q. This result derives from the density expression of λ_Q :

$$\forall\ \epsilon \in \omega \qquad f(\lambda_Q, \epsilon) = \sum_{A \in Q} \frac{\hat{\lambda}_Q(A)}{\nu(A)} 1_A(\epsilon) \tag{18}$$

with $1_A(\epsilon) = 1$ if $\epsilon \in A$ and 0 otherwise.

Selection of the bin number of a histogram

The obtaining of the optimum histogram is based on the use of an information criterion IC which gives the number of bins optimal thanks to a cost function based on the Kullback's contrast or the Hellinger's distance.

We define the cost to take $\hat{\lambda}$ when λ is the true probability density by :

$$W(\lambda, \hat{\lambda}) = E_\lambda \left(\psi \left[\frac{f(\hat{\lambda}, \epsilon)}{f(\lambda, \epsilon)} \right] \right) \tag{19}$$

where E_λ is the mathematical expectation according to λ and ψ is a convex function. According to the expression of ψ the cost function leads to different information criteria to choose the histogram with c bins. So, if ψ is the Hellinger's distance we get :

$$AIC(c) = \frac{2c-1}{N} - 2 \sum_{B \in C} \hat{\lambda}_c(B) \ln \frac{\hat{\lambda}(B)}{\nu(B)}. \tag{20}$$

It can be seen that it is identical to the classical Akaike's information criterion. If the cost function $W(\lambda, \hat{\lambda})$ is expressed according to the KullBack's contrast, we obtain two new criteria such as :

$$\phi^*(c) = \frac{c(1 + \ln(\ln N))}{N} - 2 \sum_{B \in C} \hat{\lambda}_c(B) \ln \frac{\hat{\lambda}(B)}{\nu(B)} \tag{21}$$

$$AIC^*(c) = \frac{c(1 + \ln N)}{N} - 2 \sum_{B \in C} \hat{\lambda}_c(B) \ln \frac{\hat{\lambda}(B)}{\nu(B)}. \tag{22}$$

These criteria can be used to select the optimum histogram with c bins to approximate the unknown probability density of a N-sample. Detailed demonstrations are available in Colot (1993) and in Colot (1994).

Optimum histogram building process

At first, an initial histogram with $q = Card(Q) = 2 \times E[\sqrt{N} - 1]$ bins is built giving the partition Q, where $E[\,]$ denotes the integer part. Then, a partition with $(q-1)$ bins is considered. For each possible fusion of two contiguous bins among $(q-1)$ the criterion $IC(q-1)$ is computed. The choice of the best fusion is made according to the minimization of $IC(q-1)$.When it is done, we look for the best partition with $(q-2)$ bins according to the same rule. Finally, the histogram with c bins such as $IC(c)$ for $c \in \{1, \ldots, q\}$ is retained. Figure 1 shows an initial histogram built with a N-sample ($N = 90$) randomly generated according to a gaussian distribution with mean equal to 0 and with a variance equal to 1 and the final histogram according to AIC^*. Figure 2 gives the behaviour of the three criteria. It can be seen that AIC^* and ϕ^* give the same final bin number. AIC gives a final histogram with an upper bin number. This difference is linked to the type of convergence for each information criterion Colot (1994).The optimum histogram is computed on the set $\bigcup_{i \in H_n} \mathcal{X}_{(i;j)}$. Once this histogram is obtained, we use the Hellinger's distance between the approximated distribution $\hat{\lambda}_C$ computed on the set $\mathcal{X}_{(n;j)}$ and the approximated distribution $\hat{\lambda'}_C$ computed on the set $\mathcal{X}_{(n';j)}$. This distance gives a dissimilarity between the two probability densities that is to say the ability of the source to distinguish the two hypotheses H_n and $H_{n'}$.

INFORMATION SOURCES AGGREGATION AND DECISION

We attenuate the belief structures according to the equation (9) where α is the Hellinger's distance. The information sources S_j for $j = 1$ to m are then aggregated using the Dempster's combination rule (see equation (5)). Finally, the decision is made by assigning the vector X' to the hypothesis H_n with the maximum credibility. The decision rule is based on the decision function δ which assignes a vector X' to the hypothesis H_n following :

$$\delta(X', H_n) = n \text{ iif } H_n = \arg \max_{H_i \in 2^\Theta} (Bel(H_i)) \tag{23}$$

SIMULATIONS

The proposed method has been applied to several sets of artificial data in order to perform an evaluation of the algorithm. To illustrate the method, we give in Figure 1 and Figure 2 the results of the bin fusion with the AIC^* criterion.

We present results obtained on synthetic data. For the simulations, we have generated three gaussian distributions such as : $\mu_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$ and $\sigma_1^2 = 1$; $\mu_2 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ and $\sigma_2^2 = 4$; $\mu_3 = \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix}$ and $\sigma_3^2 = 3$. The first

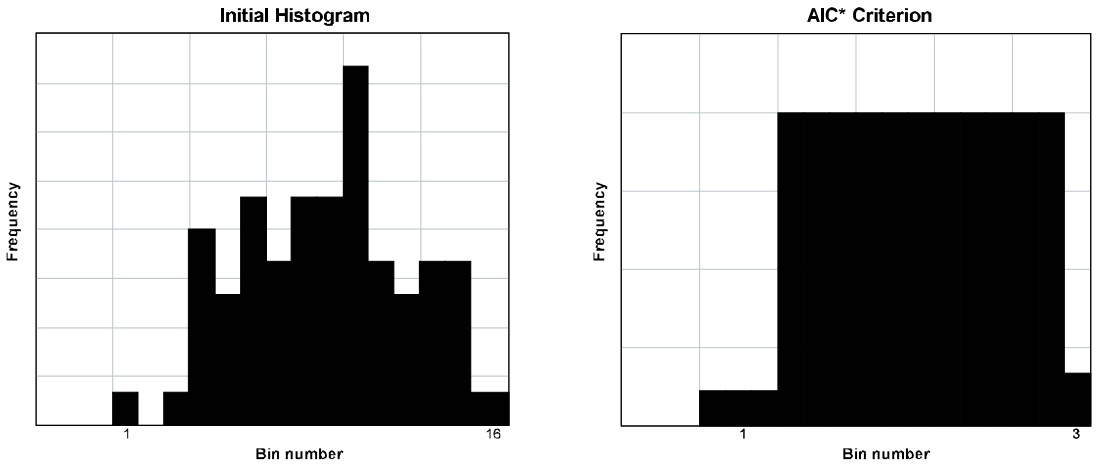

Figure 1: Initial histogram and optimum histogram for AIC^*

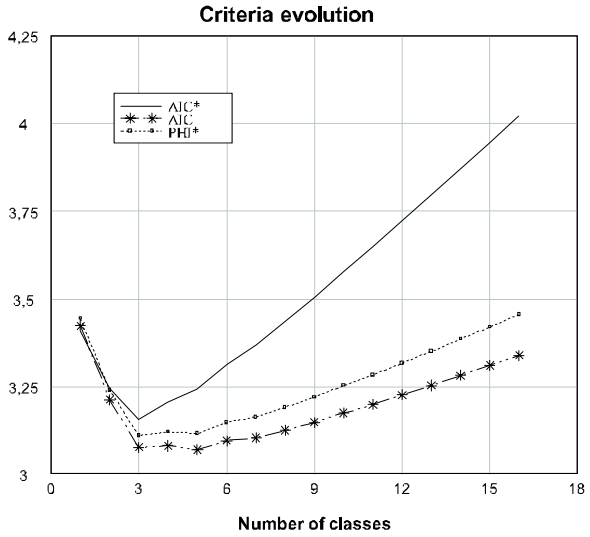

Figure 2: Criteria evolutions

learning set is made of $N = 90$ elements (30 for each class) and the second one is made of $N = 200$ elements (70 elements in the first class, 50 elements in the second class and 80 elements in the third class). The test base is made of 600 elements. Our method is compared to the method proposed in Zouhal (1995) but without optimized belief structure parameters. The results are given in the two following tables (see tables 1 and 2) for the first learning set. For the method proposed by Zouhal, the good classification rate is of 59.16% and 62.50%

%	**Classified** C_1	**Classified** C_2	**Classified** C_3
C_1 **presented**	81	7.5	11.5
C_2 **presented**	29.5	43.5	27
C_3 **presented**	31	16	53

1: Results of method Zouhal (1995)

%	**Classified** C_1	**Classified** C_2	**Classified** C_3
C_1 **presented**	87	3.5	9.5
C_2 **presented**	33	42.5	24.5
C_3 **presented**	27	15	58

2: Results with our method

for our method. According to the second learning set, we get the following results (see tables 3 and 4). For the

%	**Classified** C_1	**Classified** C_2	**Classified** C_3
C_1 **presented**	78.5	4.5	17
C_2 **presented**	29.5	47	23.5
C_3 **presented**	26	26	48

3: Results of method Zouhal (1995)

%	**Classified** C_1	**Classified** C_2	**Classified** C_3
C_1 **presented**	83.5	6.5	10
C_2 **presented**	38	41	21
C_3 **presented**	27.5	16	56.5

4: Results of our method

method proposed by Zouhal, the good classification rate is of 57.83% and 60.33% for our method.

CONCLUSION

In this paper, we have presented an original method of discrimination using both information criteria and Dempster-Shafer's theory. The proposed methodology consists in intializing the belief functions with probability densities obtained by learning. By means of information criteria, we determine the attenuation of the belief assignment based on the dissimilarity between probability distributions. Results on artificial data demonstrate the effectiveness of the proposed method. Concerning real-world data, a study is engaged concerning the melanoma detection in dermatology as an help to skin cancer diagnosis. Future work is concerned with analysis of several decision rules using uncertainty measures proposed by Klir (1988) or Wang (1992).

REFERENCES

Bezdek, James C., 1994, "Fuzziness vs. Probability. The N-th Round", IEEE Trans. on Fuzzy Systems 2, pp. 1-42.

Colot, Olivier, 1993, "Apprentissage et dtection automatique de changements de modles - Application aux signaux lectro-encphalographiques", PhD Thesis, University of Rouen.

Colot, Olivier; Olivier, Christian; Courtellemont, Pierre; El-MAtouat, Abdelaziz, 1994, "Information Criteria and Abrupt Changes in Probability Laws", Signal Processing VII : Theories and Applications, EUSIPCO'94, pp. 1855-1858.

Denoeux, Thierry, 1998, "Analysis of Evidence Theory Decision Rules for Pattern Classification", Pattern Recognition 30, pp.

Dubois, Didier, 1986, "Belief Structures, Possibility Theory and Decomposable Confidence Measures on Finite Sets", Comput. Artificial Intelligence 5, pp. 403-416.

Dubois, Didier; Lang, J; Prade, Henri, 1994, "Automated Reasoning Using Possibilistic Logic : Semantics, Belief Revision, and Variable Certainty Weights", IEEE Trans. on Knowledge and Data Engineering 6, pp. 64-71.

Klir, George J.; Folger, Tina A., 1988, "Fuzzy Sets, Uncertainty and Information", Prentice Hall PTR, Englewood Cliffs, New Jersey 07632.

Shafer, Glenn, 1976, "A Mathematical Theory of Evidence", Princeton University Press.

Smets, Philippe, 1993, "Belief Functions : The Disjonctive Rule of Combination and the Generalized Bayesian Theorem",International Journal of Approximate Reasoning, pp. 1-35.

Smets, Philippe; Kennes, R, 1994, "The Transferable Belief Model", Artificial Intelligence 66, pp. 191-234.

Wang, Zhenyuan; Klir, George J., 1992, "Fuzzy Measure Theory", Plenum Plublishing Corporation, New York.

Zadeh, Lofti A., 1978, "Fuzzy Sets as a Basis for a Theory of Possibility", Fuzzy sets and Systems 1, pp. 3-28.

Zouhal, Lalla Meriem; Denoeux, Thierry, 1995, "An Adaptative k-NN Rule Based on Dempster-Shafer Theory", Proc. of 6th International Conference on Computerized Analysis of Images and Pattern, ICAIP'95, Springer-Verlag, pp 310-317.