A classification method based on the Dempster-Shafer's theory and information criteria

E. LEFEVRE
PSI, UNIVERSITE/INSA de ROUEN
Place Emile Blondel, BP 08
76131 Mont-Saint-Aignan Cedex, France

O. COLOT
PSI, UNIVERSITE/INSA de ROUEN
Place Emile Blondel, BP 08
76131 Mont-Saint-Aignan Cedex, France

P. VANNOORENBERGHE
PSI, UNIVERSITE/INSA de ROUEN
Place Emile Blondel, BP 08
76131 Mont-Saint-Aignan Cedex, France

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, we introduce an attenuation factor based on the dissimilarity between probability distributions.***

Keywords: Data Fusion, Dempster-Shafer's theory, Information Criteria, Classification.

1 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 [1], is used for dealing with imperfect data. In the recent past, other models have been developed for handling imprecise knowledge (theory of fuzzy sets [2], possibility theory [3, 4]) or uncertain information (theory of belief functions [5]). In this paper, we deal with a classification method of imperfect data sets using evidence theory [5, 6, 7]. Recently, in this context, a new approach using neighbourhood information has been developed [8]. Each nearest neighbour 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 neighbour. 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 2, we introduce notations allowing to describe the Dempster-Shafer's Theory of evidence. Section 3 is devoted to present the proposed methodology. This work is applied to synthetic and real data (section 4).

2 Dempster-Shafer's Theory

In this section, a brief overview of the Evidence's Theory [5] is provided. Let Θ represents the set of hypotheses H_n, 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 $[0,1]$ such as :

$$m(\emptyset) = 0 \quad (1)$$

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

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 :

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

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

The value $Bel(A)$ quantifies the strength of the belief that event A 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 :

$$m(H_n) = m_1(H_n) \oplus \cdots \oplus m_M(H_n). \quad (5)$$

For two sources S_i and S_j, the aggregation of evidence for a hypothesis $H_n \subseteq \Theta$ can be written :

$$m(H_n) = \frac{1}{\mathcal{K}} \sum_{A \cap B = H_n} m_i(A).m_j(B) \quad (6)$$

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

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

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_j by a coefficient α_j for a source S_j. For all $H_n \subseteq \Theta$, the attenuated belief function can be written as :

$$m_{(\alpha,j)}(H_n) = \alpha_j.m_j(H_n) \quad (8)$$

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

3 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 3.1). The second one consists in attenuating the belief structure by means of a coefficient α_j 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 3.2). Finally, the belief structures defined for each source of information are aggregated in order to decrease significantly the uncertainty for the later classification process (see 3.3).

3.1 Basic Belief Assignment

An important aspect of the classification 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 :

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

that is to say:

$$P(x_j/H_n) = \mathcal{N}(\mu_{(n;j)}, \sigma_{(n;j)}). \quad (11)$$

The pair $(\mu_{(n;j)}, \sigma_{(n;j)})$ represents respectively the mean and the 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 :

$$\mu_{((n,n');j)} \triangleq \frac{\mu_{(n;j)} + \mu_{(n';j)}}{2} \qquad (12)$$

$$\sigma_{((n,n');j)} \triangleq \max(\sigma_{(n;j)}, \sigma_{(n';j)}). \qquad (13)$$

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 $H_n \in 2^\Theta$ depends on the membership probability with respect to :

$$m_j(H_n) = R_j * P(x'_j/H_n). \qquad (14)$$

The coefficient R_j is a normalization factor. It allows to verify the condition given by equation (2). It is defined for a hypothesis $H_n \in 2^\Theta$ as :

$$R_j = \frac{1}{\sum_{H_n \in 2^\Theta} P(x'_j/H_n)}. \qquad (15)$$

3.2 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-sample 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.

3.2.1 Probability density approximation

Let be $A_1 A_2 \ldots A_p \ldots A_q$ 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 $C = B_1 B_2 \ldots B_c$ 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) [9], AIC^* or ϕ^* [10] which are respectively Hannan-Quinn's criterion and Rissanen's criterion. These criteria have the following form :

$$IC(c) = g(c) - \sum_{B \in C} \hat{\lambda}_c \ln \frac{\hat{\lambda}_c(B)}{\nu_c(B)} \qquad (16)$$

where $g(c)$ is a penalty 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 :

$$\forall\, \epsilon \in \omega \qquad f(\lambda, \epsilon) = \frac{d\lambda}{d\nu}(\epsilon). \qquad (17)$$

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.

3.2.2 Maximum likelihood estimator for a partition Q

Let Q be a partition with q bins and let $\epsilon_1 \ldots \epsilon_N$ 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 :

$$\forall\, p \in \omega \qquad \hat{\lambda}_Q(A_p) = \frac{1}{N} \sum_{\epsilon_k \in A_p} \epsilon_k \qquad (18)$$

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

$$\forall\, \epsilon \in \omega \qquad f(\lambda_Q, \epsilon) = \sum_{A \in Q} \frac{\hat{\lambda}_Q(A)}{\nu(A)} 1_A(\epsilon) \quad (19)$$

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

3.2.3 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 :

$$W(\lambda, \hat{\lambda}) = E_\lambda \left(\psi \left[\frac{f(\hat{\lambda}, \epsilon)}{f(\lambda, \epsilon)} \right] \right) \qquad (20)$$

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 :

$$AIC(c) = \frac{2c-1}{N} - 2 \sum_{B \in C} \hat{\lambda}_c(B) \ln \frac{\hat{\lambda}(B)}{\nu(B)}. \qquad (21)$$

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 :

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

$$AIC^*(c) = \frac{c(1 + \ln N)}{N} - 2 \sum_{B \in C} \hat{\lambda}_c(B) \ln \frac{\hat{\lambda}(B)}{\nu(B)}. \qquad (23)$$

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 [9, 10].

3.2.4 Optimum histogram building process

At first, an initial histogram with $q = Card(Q) = 2.E[\sqrt{N} - 1]$ bins is built giving the partition Q, where $E[]$ denotes the integer part [11]. 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. This initial histogram is made of $16 = 2.E[\sqrt{90} - 1]$ bins. Final histograms according to respectively

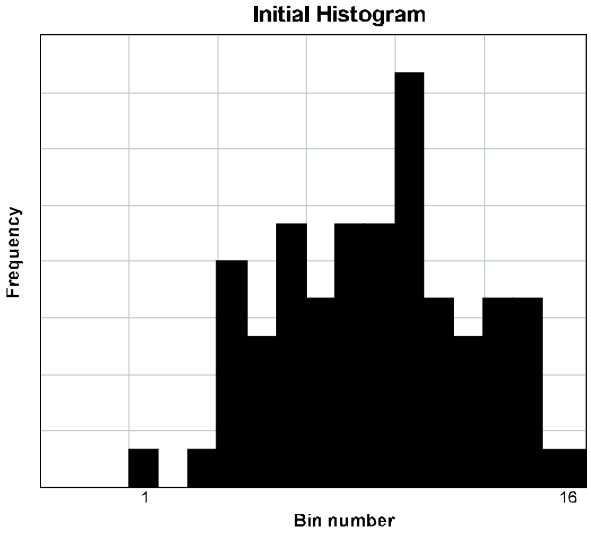

Figure 1: Initial histogram

AIC, AIC^* and ϕ^* are given in figures 2, 3, 4. Figure 5 gives the behaviour of the three

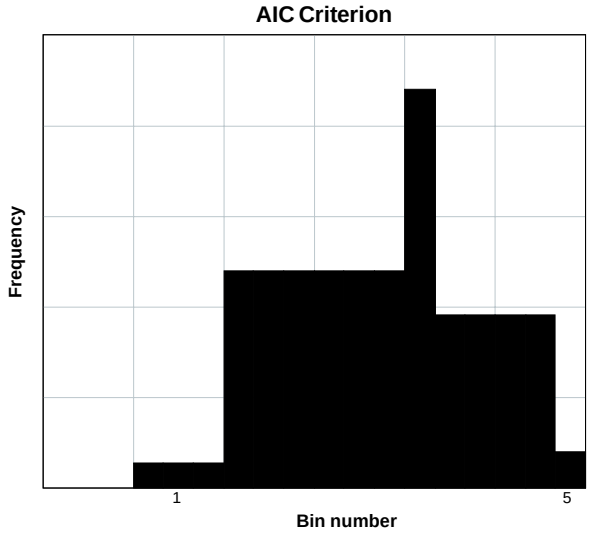

Figure 2: Optimum histogram with AIC

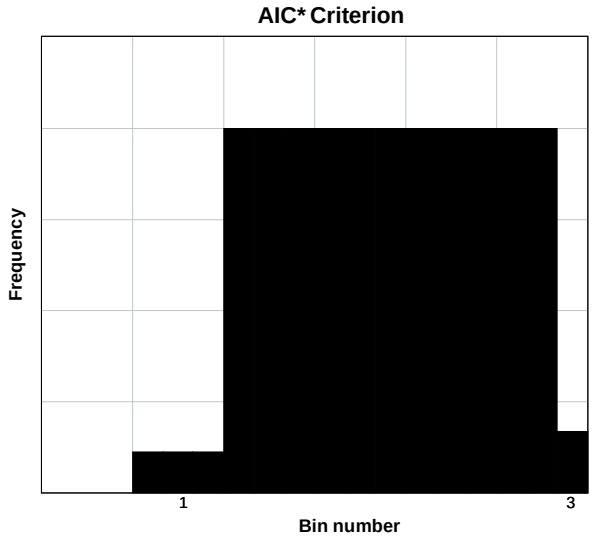

Figure 3: Optimum histogram with AIC^*

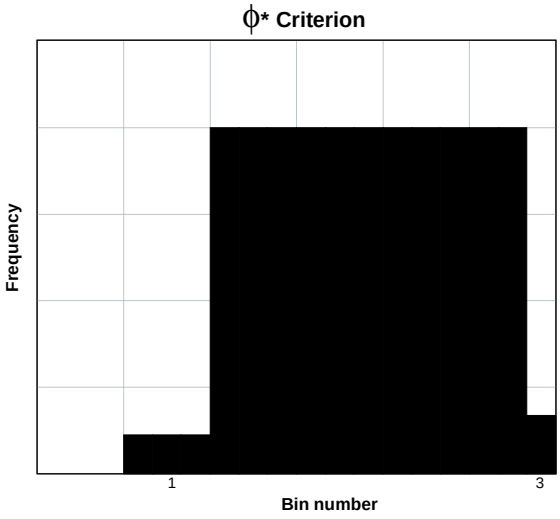

Figure 4: Optimum histogram with ϕ^*

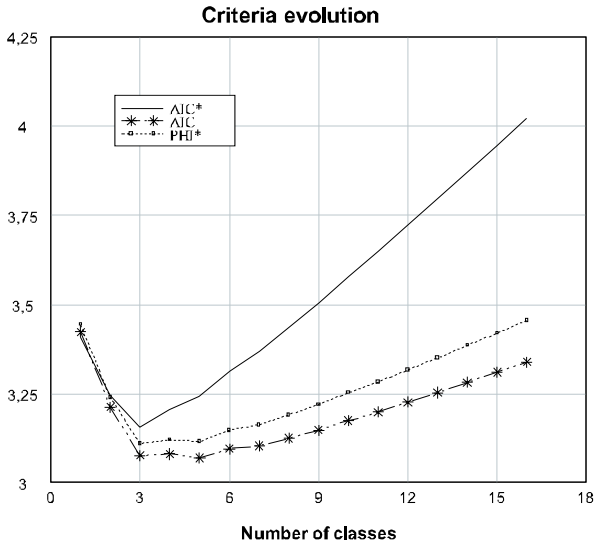

Figure 5: Criteria evolutions

criteria. It can be seen that AIC^* and ϕ^* give the same final histogram. AIC gives a final histogram with an upper bin number. This difference is linked to the type of convergence for each information criterion [10].

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 $H_{n'}$.

3.3 Information sources aggregation and decision

We attenuate the belief structures according to the equations (8) and (9) where α_j is the

Hellinger's distance :

$$\alpha_j = \psi\left[\frac{f(\hat{\lambda},\epsilon)}{f(\lambda,\epsilon)}\right] = 4\sum_{B\in C}\lambda(B)\left[\sqrt{\frac{\hat{\lambda}(B)}{\lambda(B)}}-1\right]^2. \quad (24)$$

The information sources S_j 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 :

$$\delta(X', H_n) = n \text{ iif } H_n = \arg\max_{H_i\in 2^\Theta}(Bel(H_i)) \quad (25)$$

4 Simulations

The proposed method has been applied to several sets of artificial and real data in order to perform an evaluation of the algorithm.

4.1 Synthetic data

In this section, we present results obtained on synthetic data. For the simulations, we have generated three gaussian distributions such as : $\mu_1 = [1,1,1]^t$ and $\sigma_1^2 = 1$; $\mu_2 = [-1,1,0]^t$ and $\sigma_2^2 = 4$; $\mu_3 = [0,-1,1]^t$ and $\sigma_3^2 = 3$. The first 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 [12]. The results are given in the two following tables for the first learning set.

For the method proposed by Zouhal, the good classification rate is of 59.16% and 62.50% for our method. According to the second learning set, we get the following results (see tables 3 and 4).

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

Table 1: Results of method [12]

	Classified		
Presented	C_1	C_2	C_3
C_1	81	7.5	11.5
C_2	29.5	43.5	27
C_3	31	16	53

Table 2: Results of our method

	Classified		
Presented	C_1	C_2	C_3
C_1	87	3.5	9.5
C_2	33	42.5	24.5
C_3	27	15	58

4.2 Real data

A second database concerns a set of 16 characteristics extracted from 122 images of dermatological lesions. Details concerning the features can be found in [13]. The database is composed of 101 naevi (no pathological lesions) and 21 melanoma. Final results are presented in the following tables (Tables 5 and 6). The proposed method allows to obtain 81.1% of good classification towards 75.7% for the method presented in [12].

5 Conclusion

In this paper, we have presented an original method of classification using both information criteria and Dempster-Shafer's theory. The proposed methodology consists in initializing the belief functions with probability densities obtained by learning. By means of informa-

Table 3: Results of method [12]

	Classified		
Presented	C_1	C_2	C_3
C_1	78.5	4.5	17
C_2	29.5	47	23.5
C_3	26	26	48

Table 4: Results of our method

	Classified		
Presented	C_1	C_2	C_3
C_1	83.5	6.5	10
C_2	38	41	21
C_3	27.5	16	56.5

Table 5: Results of method [12]

	Reality	
Decision	Naevus	Melanoma
Naevus	99	1
Melanoma	47.6	52.4

tion criteria, we determine the attenuation of the belief assignment based on the dissimilarity between probability distributions. Results on artificial and real data demonstrate the effectiveness of the proposed method. Concerning the real-world data (diagnosis in dermatology), tests on a larger base are processing at this time. Future work is concerned with analysis of several decision rules using uncertainty measures proposed by Klir [14, 15].

6 Acknowledgements

The authors would like to acknowledge the support of the Normandy regional council. The co-operation of P. Joly of the clinic of dermatology of Charles Nicolle's hospital of Rouen in supplying diagnosis on melanoma and acquisition of image lesions, is also gratefully acknowledged. Finally, authors would like to express their appreciation to Professor D. De Brucq, whose comments led to a significant improvement of the paper.

Table 6: Results of our method

	Reality	
Decision	Naevus	Melanoma
Naevus	86.1	13.9
Melanoma	23.8	76.2

References

[1] J.C. Bezdek. Fuzziness vs. probability - the n-th round. *IEEE Trans. on Fuzzy Systems*, 2(1):1–42, 1994.

[2] L.A. Zadeh. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1:3–28, 1978.

[3] D. Dubois. Belief structures, possibility theory and decomposable confidence measures on finite sets. *Comput. Artif. Intell.*, 5(5):403–416, 1986.

[4] D. Dubois, J. Lang, and H. Prade. Automated reasoning using possibilistic logic: Semantics, belief revision, and varaible certainty weights. *IEEE Trans. on Knowledge and Data Engineering*, 6:64–71, 1994.

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

[6] P. Smets. Belief functions: The disjonctive rule of combination and the generalized bayesian theorem. *Int. J. of Approximate Reasoning*, 9:1–35, 1993.

[7] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66:191–234, 1994.

[8] T. Denoeux. Analysis of evidence-theory decision rules for pattern classification. *Pattern Recognition*, 30(7), 1998.

[9] O. Colot. *Apprentissage et détection automatique de changements de modèles - Application aux signaux électro-encéphalographiques.* PhD thesis, Université de Rouen, 1993.

[10] O. Colot and al. Information criteria and abrupt changes in probability laws. In M. Holt, C. Cowan, P. Grant, and W. Sandham, editors, *Signal Processing VII: Theories and Applications*,

pages 1855–1858. EUSIPCO'94, September 1994.

[11] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaïke Information Criterion Statistics.* KTK Scientific Publishers, Tokyo, 1986.

[12] L. M. Zouhal and T. Denoeux. An adaptative k-NN rule based on dempster-shafer theory. In *Proc. Of the 6th Intern. Conf. On Comput. Analys. Of Imag. and Pattern*, pages 310–317. ICAIP'95, Springer Verlag, September 1995.

[13] O. Colot and Al. *A Colour Image Processing Method for Melanoma Detection.* Lecture Notes in Computer Science, 1496. Springer-Verlag, October 1998.

[14] G.J. Klir and T.A. Folger. *Fuzzy Sets, Uncertainty, and Information.* Prentice Hall P T R, Englewood Cliffs, New Jersey 07632, 1988.

[15] Z. Wang and G.J. Klir. *Fuzzy Measure Theory.* Plenum Publishing Corporation, 233, Spring Street New York, 1992.