Contribution des mesures d'information à la modélisation crédibiliste de connaissances

Contribution of information measures in evidential knowledge modelling

E. Lefevre, O. Colot, P. Vannoorenberghe et D. de Brucq

Laboratoire **P**erception **S**ystèmes **I**nformation (PSI),

UPRES EA 2120

Université/INSA de Rouen

Place Emile Blondel, BP 08

76131 Mont-Saint-Aignan Cedex

Tél : +33.(0)2.35.52.84.05

Fax :+33.(0)2.35.52.84.83

e-mail : Eric.Lefevre@insa-rouen.fr

Olivier.Colot@insa-rouen.fr

Patrick.Vannoorenberghe@univ-rouen.fr

Denis.Debrucq@univ-rouen.fr

Résumé et mots clef

Dans le cadre de la reconnaissance de formes, plusieurs méthodes de classification ont été développées. Plus récemment, des méthodes utilisant la théorie de Dempster-Shafer ont été mises au point afin de gérer les problèmes liés à la fusion d'informations imparfaites. Nous proposons ici une méthode de discrimination fondée sur l'utilisation de structures de croyance. L'une des principales difficultés de la théorie de l'évidence réside dans la modélisation des connaissances. Afin de pallier ce problème, plusieurs méthodes de modélisation des connaissances à l'aide de fonctions de croyance ont vu le jour, dont celle proposée par A. Appriou [1, 2]. Afin de respecter l'inférence bayésienne dans le cas de la connaissance parfaite des probabilités *a priori*, nous utilisons cette méthode pour initialiser nos fonctions de croyance. Notre contribution réside dans l'utilisation de coefficients de fiabilité attribués à chaque source d'information selon chaque hypothèse afin de modéliser le plus précisément possible l'information disponible. Ces coefficients sont définis par l'intermédiaire d'une mesure de ressemblance entre des approximations de lois de probabilités *a priori* inconnues. Celles-ci sont déterminées par des histogrammes construits à l'aide de critères d'information. Les structures de croyance issues des sources les moins fiables sont alors affaiblies. Ensuite, les informations sont fusionnées à l'aide de l'opérateur de combinaison de Dempster. Des résultats sur des données synthétiques sont proposés afin d'illustrer la méthode.

Théorie de Dempster-Shafer, Fusion d'informations imparfaites, Critères d'information.

Abstract and keywords

Within the framework of pattern recognition, many methods of classification were developed. More recently, techniques using the Dempster-Shafer's theory tried to deal with the problem related to the management of the uncertainty and the data fusion. In this paper, we propose a classification method based on this theory. The main difficulty of this method is the knowledge modelling. To solve this problem, several methods were proposed, in particular by A. Appriou [1, 2]. In order to respect bayesian approach in the case where the a priori probabilities are perfectly known, we use this method to initialize a belief structure. Our contribution lies in the use of reliability factor for each information source according to each hypothesis. These coefficients are defined by a dissimilarity measure between two approximations of unknown probability distributions. These are determined by histograms built by the use of information criteria. The least reliable belief structure are attenuated. Then, we use the Dempster's rule of combination to aggregate the attenuated sources. Results on synthetic data are given in order to illustrate the method.

Dempster-Shafer's Theory, Imperfect Information Fusion, Information Criteria.

1 Introduction

L'un des principaux challenges de la société de l'information concerne la gestion des *informations imparfaites*. En effet, dans de nombreux secteurs applicatifs, il s'agit de classer, décider, surveiller, commander à partir d'informations observées sur le système ou préalablement modélisées [3, 4]. Ces informations peuvent revêtir plusieurs aspects :

- l'aspect *donnée* dans le cadre de la gestion de bases (concept de fouille de données [5, 6, 7]) : données financières, commerciales, médicales,... Ainsi, il peut s'agir d'extraire des variables discriminantes, d'identifier des structures sous-jacentes, d'extraire de la connaissance de manière générale,
- l'aspect *mesure* pour les systèmes multicapteurs [1, 8, 9]. Dans ce contexte, il s'agit alors de fusionner les mesures pour la commande d'un système ou la prise d'une décision.

De plus, le contenu des informations peut se révéler de nature purement numérique mais parfois symbolique (le célèbre exemple : *"Jean est grand"*). Dans ce cas, il est souvent préférable de se ramener à des informations numériques [10, 11]. L'imperfection des informations fait appel à plusieurs concepts. Le premier, généralement bien maîtrisé, concerne l'imprécision des informations. L'incertitude est un second concept à différencier de l'imprécision par le fait qu'il ne fait pas référence au contenu de l'information mais à sa "qualité". Enfin, l'incomplétude peut se rencontrer dans de nombreuses applications. Modéliser des connaissances aussi hétérogènes devient alors rapidement un problème crucial. La communauté scientifique s'attarde d'ailleurs de plus en plus sur ce type de problèmatique (*KDD : Knowledge Discovery in Databases*). Le problème se complique encore lorsqu'il s'agit de fusionner ces informations dans un but de prendre la décision la meilleure au sens d'un critère. Dans cet article, notre propos s'inscrit dans le cadre d'un problème de classification supervisée de données incertaines, généralement dénommé par discrimination ou reconnaissance des formes dans la littérature [12, 13, 14].

Le problème de reconnaissance de formes a tout d'abord été envisagé sous l'approche bayésienne. Emanant de la théorie des probabilités, cette approche repose sur la règle de décision du maximum *a posteriori*. Cette règle permet de fusionner des informations probabilistes provenant de sources indépendantes et de minimiser l'erreur de classification. L'un des inconvénients majeurs de cette technique réside dans l'exigence de la connaissance parfaite des probabilités, et plus particulièrement de la probabilité *a priori*. Malheureusement, lorsque les connaissances sur le problème sont imparfaites, ces probabilités ne sont pas connues avec "certitude". Ces limitations, maintenant bien identifiées [15, 16], ont poussé certains auteurs à développer d'autres approches telles que la théorie des possibilités [17, 18, 19] ou la théorie de l'évidence (théorie de Dempster-Shafer [20]). Ces cadres théoriques permettent la gestion et la fusion de données imparfaites (incertaines, imprécises et incomplétes).

Nous proposons une approche originale de la modélisation de connaissances incertaines et imprécises par l'intermédiaire de la théorie de Dempster-Shafer et de mesures d'information. Notre contribution réside dans l'utilisation de ces mesures

d'information afin d'évaluer la fiabilité et le pouvoir discriminant d'une source d'information. La méthode que nous proposons considère les Q composantes d'un vecteur X' à classer comme autant de sources d'information. Chaque composante est donc assimilée à une source destinée à renforcer l'appartenance du vecteur X' à l'une des hypothèses solution.

L'article est organisé autour de trois sections principales. Nous présenterons dans un premier temps, un rappel des principaux concepts de la théorie de Dempster-Shafer (Section 2). Nous aborderons dans la section 3, la méthode de discrimination proposée en présentant successivement la construction du jeu de masses et la mise en oeuvre de l'affaiblissement de ce jeu de masses. Pour construire cet affaiblissement, notre approche repose sur l'utilisation de critères d'informations (notés *IC*) permettant de construire des histogrammes approchant des lois de probabilité et sur l'utilisation d'une mesure de dissemblance entre lois. La méthode sera finalement illustrée à l'aide de données synthétiques (Section 4).

2 Théorie de Dempster-Shafer

La théorie de l'évidence fut initialement introduite par Dempster [21] lors de ses travaux sur les bornes inférieure et supérieure d'une famille de distributions de probabilités. A partir de ce formalisme mathématique, Shafer [20] a montré l'intérêt des fonctions de croyance pour la modélisation de connaissances incertaines. L'utilité des fonctions de croyance, comme alternative aux probabilités subjectives, a été démontrée plus tard de manière axiomatique par Smets [22, 23] au travers du *Modèle de Croyance Transférable* fournissant ainsi une interprétation claire et cohérente du concept sous-jacent à la théorie.

2.1 Modélisation des connaissances

Soit Θ l'ensemble des N hypothèses solutions du problème. L'ensemble Θ, appelé *cadre de discernement*, est défini de la manière suivante :

$$\Theta = \{H_1, \cdots, H_n, \cdots, H_N\}. \tag{1}$$

On suppose qu'à chaque vecteur à classer correspond une valeur et une seule dans Θ. Ceci signifie que le cadre de discernement est exhaustif et que les hypothèses sont exclusives. Cette notion est aussi appelée *closed-world* en opposition avec la notion d'*open-world* (cadre non exhaustif) présenté par Smets [24]. On définit une masse de probabilité élémentaire, appelée *masse de croyance*, qui caractérise la véracité d'une proposition $\mathcal{H}$ pour une source d'information S_j ($j = \{1, \ldots, Q\}$) donnée. La masse m_j associée à cette source S_j est alors définie par :

$$m_j : 2^{\Theta} \rightarrow [0, 1] \tag{2}$$

et vérifie les propriétés suivantes :

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

$$\sum_{\mathcal{H}\subseteq\Theta} m_j(\mathcal{H}) = 1. \tag{4}$$

Cette probabilité basique se différencie d'une probabilité au sens classique du terme par le fait que la totalité de la masse de croyance est répartie non seulement sur les hypothèses singletons H_n mais aussi sur les hypothèses combinées $\mathcal{H}$. Cette différence permet d'accorder une partie de la croyance à une proposition et ainsi d'affecter à l'ensemble des hypothèses contenues dans la proposition une croyance à la réalisation de chacune d'entre-elles sans prendre partie pour l'une d'elles précisément. La modélisation issue de la fonction m_j est appelée jeu de masses. Les sous-ensembles $\mathcal{H}$ dont la masse est non nulle sont appelés *éléments focaux*. De plus, l'union de tous les éléments focaux est appelé *noyau*. A partir de la fonction m_j, on définit respectivement les fonctions de *crédibilité* Cr_j et de *plausibilité* Pl_j par :

$$Cr_j(\mathcal{H}) = \sum_{\mathcal{H}'\subseteq\mathcal{H}} m_j(\mathcal{H}') \tag{5}$$

$$Pl_j(\mathcal{H}) = \sum_{(\mathcal{H}\cap\mathcal{H}')\neq\emptyset} m_j(\mathcal{H}') = 1 - Cr_j(\overline{\mathcal{H}}) \tag{6}$$

où $\overline{\mathcal{H}}$ représente l'événement contraire de la proposition $\mathcal{H}$. La crédibilité $Cr_j(\mathcal{H})$ mesure la force avec laquelle on croit en la véracité de la proposition $\mathcal{H}$. La plausibilité $Pl_j(\mathcal{H})$, fonction duale de la crédibilité, mesure l'intensité avec laquelle on ne doute pas de $\mathcal{H}$. La principale difficulté consiste à modéliser les connaissances sur le problème en initialisant de manière adéquate les fonctions de croyance m_j. Cette modélisation dépend généralement de l'application envisagée. Deux principes d'initialisation de fonction de croyance ont été développés. Le premier principe repose sur une analyse de données dans l'espace des caractéristiques [25]. Alors que le second analyse séparément chacune des ces caractéristiques [1].

2.2 Affaiblissement d'un jeu de masses

Un autre aspect de la théorie de l'évidence est la possibilité d'affaiblir le jeu de masse m_j en introduisant un coefficient de confiance pour chaque source d'information S_j. Le but de cette démarche consiste à introduire la notion de fiabilité entre les différentes sources d'information. Le jeu de masses, pondéré par le coefficient de confiance α_j, que nous noterons désormais $m_{(\alpha,j)}$ devient alors :

$$\forall\ \mathcal{H} \in 2^{\Theta} \qquad m_{(\alpha,j)}(\mathcal{H}) = \alpha_j.m_j(\mathcal{H}) \tag{7}$$

$$m_{(\alpha,j)}(\Theta) = 1 - \alpha_j + \alpha_j.m_j(\Theta). \tag{8}$$

Là encore, le problème réside en la détermination des Q coefficients α_j. Appriou [8] préconise de prendre dans le cadre d'un problème de discrimination, comme coefficient de confiance aux sources, les éléments de la matrice de confusion du classifieur utilisé.

2.3 Fusion d'informations

Dans le cadre probabiliste, la fusion est de type bayesien si l'on dispose des probabilités *a priori*. La décision est alors classiquement fondée sur un critère de maximum de probabilité *a posteriori*. En l'absence de connaissance des probabilités *a priori*, on utilise alors la fusion de vraisemblance des sources d'information et la décision repose sur un critère de maximum de vraisemblance. Dans le cas de sources d'information sur des hypothèses composées, on étend le principe précédent [26]. En théorie des sous-ensembles flous [27] comme en théorie des possibilités [19], de nombreux opérateurs de fusion ont été développés. Dans le cadre de la théorie de l'évidence de Dempster-Shafer, la fusion des informations issues de sources distinctes est réalisée en utilisant la *loi de combinaison de Dempster*, aussi appelée somme orthogonale. Celle-ci, qui s'avère commutative et associative, est définie par :

$$\forall\ \mathcal{H} \in 2^{\Theta} \qquad m(\mathcal{H}) = m_1(\mathcal{H}) \oplus \ldots \oplus m_Q(\mathcal{H}) \tag{9}$$

où $\oplus$ représente l'opérateur de combinaison. Dans un cas à deux sources notées S_i et S_j, la combinaison peut se mettre sous la forme :

$$m(\mathcal{H}) = \frac{1}{1-\kappa} \sum_{(\mathcal{H}' \cap \mathcal{H}'') = \mathcal{H}} m_i(\mathcal{H}').m_j(\mathcal{H}'') \tag{10}$$

où κ est défini par :

$$\kappa = \sum_{(\mathcal{H}' \cap \mathcal{H}'') = \emptyset} m_i(\mathcal{H}').m_j(\mathcal{H}''). \tag{11}$$

Dans l'équation (10), le coefficient κ reflète le conflit existant entre les deux sources S_i et S_j. Lorsque ce facteur est égal à 1, les sources sont en conflit total et les informations ne peuvent être fusionnées. Au contraire, lorsque κ est nul, les sources sont en parfait accord. Cette règle de fusion, déduite de la règle de conditionnement [24], a été critiquée dans plusieurs travaux dont [28], en particulier dans le cas de sources en conflit total. Pour pallier cet inconvénient, Dubois et Prade [29] ont été amenés à définir les opérateurs de fusion conjonctive et disjonctive :

$$\forall\ \mathcal{H} \in 2^{\Theta} \qquad (m_i \cap m_j)(\mathcal{H}) = \sum_{\mathcal{H}' \cap \mathcal{H}'' = \mathcal{H}} m_i(\mathcal{H}').m_j(\mathcal{H}'') \tag{12}$$

et

$$\forall\ \mathcal{H} \in 2^{\Theta} \qquad (m_i \cup m_j)(\mathcal{H}) = \sum_{\mathcal{H}' \cup \mathcal{H}'' = \mathcal{H}} m_i(\mathcal{H}').m_j(\mathcal{H}''). \tag{13}$$

2.4 Règles de décision

Une fois la masse résultante m ainsi obtenue, la décision peut alors être prise. Différentes règles de décision ont été définies, les plus courantes étant la règle du maximum de plausibilité et la règle du maximum de crédibilité. A partir des fonctions de croyance, Smets [23, 30] définit une fonction de probabilité appelée *probabilité pignistique*. De manière générale, on définit la fonction de décision δ pour un vecteur X' à classer par :

$$\delta(X') = H_n \quad avec \quad H_n = arg \quad [max_{H_i \in \Theta} \Upsilon(H_i)] \tag{14}$$

où $\Upsilon(.)$ est la fonction de crédibilité, la fonction de plausibilité ou la probabilité pignistique. Une analyse de plusieurs règles de décision basées sur le concept de fonctions de coût est présentée dans [31].

3 Méthodologie

La méthode proposée peut être décomposée en trois étapes. La première, présentée dans le paragraphe 3.1, correspond à l'initialisation du jeu de masses. La seconde partie est relative à l'affaiblissement des structures de croyances à l'aide de coefficients de fiabilité. La méthode d'obtention de ces coefficients est fondée sur des critères d'information (cf. 3.2) permettant de construire un histogramme approchant la loi inconnue d'un processus à partir d'un unique échantillon de T réalisations. Cet histogramme est optimal au sens du critère de maximum de vraisemblance et permet d'utiliser une mesure de dissemblance (distance de Hellinger) entre les distributions de probabilité issues de l'apprentissage et celles issues de la validation et ce pour chaque hypothèse H_n et chaque source S_j. Cette distance permet de construire les coefficients de qualité de l'apprentissage. La dernière étape (cf. 3.3) de la méthodologie est constituée par le processus de fusion des sources d'information et par l'élaboration de la fonction de décision du classifieur.

3.1 Initialisation du jeu de masses

Soit le vecteur X' à Q composantes, avec $X' = [x_1^{'}, \ldots, x_j^{'}, \ldots, x_Q^{'}]^{tr}$, à classer parmi N hypothèses notée H_n. Le cadre de discernement est défini par : $\Theta = \{H_1, \ldots, H_n, \ldots, H_N\}$. Dans notre méthode, chaque composante du vecteur X' sera considérée comme une source d'information notée S_j avec $j \in \{1, \ldots, Q\}$. Pour chacune des hypothèses H_n appartenant au cadre de discernement Θ et pour chacune des sources S_j, nous allons définir un jeu de masses de croyance notée m_{nj}. Ces jeux de masses seront construits à partir de l'un des deux modèles proposés par Appriou [1]. Ces modélisations permettent de conserver les trois propriétés, qui sont : la cohérence avec l'approche bayésienne dans le cas où les probabilités *a priori* sont connues, la séparabilité des hypothèses H_n et la cohérence avec l'association probabiliste des sources. Le modèle 1 proposé

par Appriou [2] est particularisé par :

$$m_{nj}(H_n) = 0 \quad (15)$$

$$m_{nj}(\overline{H}_n) = q_{nj} * \{1 - R_j * p(x_j^{'}/H_n)\} \quad (16)$$

$$m_{nj}(\Theta) = 1 - q_{nj} + q_{nj} * R_j * p(x_j^{'}/H_n) \quad (17)$$

et le modèle 2 :

$$m_{nj}(H_n) = q_{nj} * R_j * p(x_j^{'}/H_n)/\{1 + R_j * p(x_j^{'}/H_n)\} \quad (18)$$

$$m_{nj}(\overline{H}_n) = q_{nj}/\{1 + R_j * p(x_j^{'}/H_n)\} \quad (19)$$

$$m_{nj}(\Theta) = 1 - q_{nj} \quad (20)$$

où R_j est un facteur de normalisation contraint par :

$$R_j \in [0, (max_{n\in[1,N]}\{p(x_j^{'}/H_n)\})^{-1}] \quad (21)$$

où $p(x_j^{'}/H_n)$ représente la densité de probabilité de la mesure $x_j^{'}$ issue de la source S_j sous les différentes hypothèses H_n et où $\overline{H}_n$ représente l'événement contraire de l'hypothèse H_n. Ces densités pourront être déterminées à partir de la connaissance *a priori* de la distribution de probabilité ou à partir d'estimateurs de densité tels que les noyaux. Celles-ci sont donc plus ou moins représentatives des densités réellement rencontrées. Les coefficients q_{nj} caractérisent ici le degré de représentativité. Lorsque les densités sont parfaitement représentatives de l'apprentissage alors les coefficients q_{nj} sont égaux à 1 et dans ce cas les masses de croyance ne sont pas affaiblies. A l'inverse, lorsque la distribution de probabilité est totalement méconnue, ce qui est caractérisé par un coefficient q_{nj} égal à 0, alors les jeux de masses deviennent élément neutre de l'opérateur de combinaison de Dempster. Appriou [1] fixe ces coefficients de fiabilité à 1 lorsque la confiance accordée aux sources est élevée et à 0.9 dans le cas contraire. Une autre approche [32] détermine la valeur de ces coefficients à l'aide d'une fonction linéaire de l'écart-type des données d'apprentissage de chaque hypothèse selon chaque source à valeur dans $[0.9, 1]$. Pour un écart-type minimum, le coefficient est fixé à 1 alors que dans le cas d'un écart-type maximal le coefficient est fixé à 0.9. Dans la section suivante, nous proposons une méthode pour la détermination des coefficients q_{nj} à l'aide de critères d'information.

3.2 Affaiblissement du jeu de masses à l'aide de critères d'information

De manière à qualifier la bonne représentativité de l'apprentissage, nous proposons d'utiliser une mesure d'information qui est calculée pour chaque couple (S_j, H_n). Le but de la démarche est de permettre d'évaluer la fiabilité de l'apprentissage de la source S_j relativement à l'hypothèse H_n. Pour satisfaire ce principe, on calcule la probabilité $p(x_j^{'}/H_n)$ à partir de la base d'apprentissage et une base de validation nous permet d'ajuster le coefficient q_{nj}. Le calcul du coefficient q_{nj} repose donc

sur la ressemblance (ou la dissemblance) entre les lois de probabilité de chaque hypothèse issues de la base d'apprentissage et de la base de validation. Si ces lois sont ressemblantes, alors l'apprentissage de cette densité de probabilité sera considéré comme représentatif, et dans ce cas q_{nj} sera proche de 1. Au contraire, si les lois sont dissemblables, alors l'estimation de la densité $p(x_j^{'}/H_n)$ sera considérée comme peu fiable et le coefficient q_{nj} sera alors proche de 0. Parmi les tests statistiques classiques paramétriques ou non paramétriques, tels que le test du Chi-deux (χ^2) ou le test de Kolmogorov permettant de définir l'adéquation ou la ressemblance statistique entre des lois de probabilités, nous avons arrêté notre choix sur une méthode non paramétrique. Notre démarche est la suivante.

Nous approcherons les lois de probabilité de chaque hypothèse par des histogrammes (Section 3.2.1) construits à l'aide d'un critère d'information que nous noterons $IC(.)$ [33, 34], qui sera fonction du nombre de classes de l'histogramme (Section 3.2.3). Ce dernier pourra être le critère d'Akaïke (AIC) [35], un critère synthétisant AIC et le critère de Rissanen (RIC) [36] ou le critère de Hannan et Quinn [37]. Ces histogrammes, optimaux au sens du maximum de vraisemblance, résument l'information contenue dans chaque source S_j et permettent d'obtenir une estimation optimale de la loi au sens du critère choisi et ce de manière non paramétrique. Nous utilisons ensuite une mesure de dissemblance (Section 3.2.5) entre lois de probabilité (distance de Hellinger) [38] modifiée pour être applicable dans le cas des histogrammes [34] entre chacune des approximations de loi afin de définir le coefficient q_{nj}. Cette mesure entre lois de probabilité offre l'avantage, au contraire de mesures telles que la divergence de Kullback ou la distance de Bhattacharyya [38], de varier entre 0 et 1.

3.2.1 Approximation de loi de probabilité par des histogrammes

Disposant d'un échantillon $\epsilon_1 \dots \epsilon_T$ de taille T d'un processus aléatoire $\mathcal{E}$ de loi de probabilité λ inconnue, que l'on suppose continue par rapport à une loi ν *a priori* donnée, on souhaite effectuer une approximation de λ à l'aide d'un histogramme. Soit Ω l'ensemble des valeurs prises par $\mathcal{E}$. La densité de probabilité f de λ s'exprime à l'aide de la dérivée de Radon-Nicodym par :

$$\forall\, \epsilon \in \Omega \qquad f(\lambda, \epsilon) \triangleq \frac{d\lambda}{d\nu}(\epsilon). \tag{22}$$

La densité f sera approchée à partir du seul échantillon de taille T de $\mathcal{E}$ et d'un histogramme à C classes construit à l'aide de cet échantillon. Il s'agit, dans un premier temps, de déterminer cet histogramme à C classes défini sur une partition $\mathcal{C}$ de Ω.

3.2.2 Estimateur du maximum de vraisemblance pour une partition $\mathcal{C}$

Soit $\mathcal{C}$ une partition à C classes de Ω, soit $\epsilon_1 \dots \epsilon_T$ un T-échantillon d'observation et soit $\lambda_{\mathcal{C}}$ la restriction de λ à la partition $\mathcal{C}$. L'estimateur du maximum de vraisemblance $\hat{\lambda}_{\mathcal{C}}$ de $\lambda_{\mathcal{C}}$ est donné par l'équation suivante :

$$\forall\, p \in \{1, \dots, C\} \qquad \hat{\lambda}_{\mathcal{C}}(A_p) = \frac{\mid A_p \mid}{T} \tag{23}$$

où A_p est une classe de la partition $\mathcal{C}$ et où $\bigcup_{p\in\{1,\dots,C\}} A_p = \mathcal{C}$. Ce résultat provient de l'expression de la densité de $\hat{\lambda}_{\mathcal{C}}$:

$$\forall\, \epsilon \in \Omega \qquad f(\hat{\lambda}_{\mathcal{C}}, \epsilon) = \sum_{A\in\mathcal{C}} \frac{\hat{\lambda}_{\mathcal{C}}(A)}{\nu_{\mathcal{C}}(A)} 1_A(\epsilon) \tag{24}$$

avec $1_A(\epsilon) = 1$ si $\epsilon \in A$ et 0 sinon.

3.2.3 Sélection du nombre de classes d'un histogramme approchant une loi inconnue

L'obtention d'un histogramme optimal est fondé sur l'utilisation d'un critère d'information noté IC. Celui-ci est établi à partir d'une fonction coût de type contraste de Kullback ou distance de Hellinger [34]. En effet, on définit le coût de prendre $\hat{\lambda_{\mathcal{C}}}$ quand λ est la vraie densité de probabilité par :

$$W(\lambda, \hat{\lambda}_{\mathcal{C}}) \triangleq E_\lambda \left(\psi \left[\frac{f(\hat{\lambda}_{\mathcal{C}}, \epsilon)}{f(\lambda, \epsilon)} \right] \right) \tag{25}$$

et le risque moyen par :

$$\overline{W}(\lambda, \hat{\lambda}_{\mathcal{C}}) \triangleq E_\lambda \Big(W(\lambda, \hat{\lambda}_{\mathcal{C}}) \Big) \tag{26}$$

où E_λ est l'espérance mathématique par rapport à λ et ψ une fonction convexe. Selon l'expression de ψ, la fonction risque conduit à divers critères d'information pour établir l'histogramme à C classes, c'est-à-dire pour choisir $\hat{\lambda}_{\mathcal{C}}$ minimisant le risque. Ainsi, si ψ est de type distance de Hellinger [39], nous obtenons relativement à la partition $\mathcal{C}$ le critère défini par :

$$AIC(C) = \frac{2C-1}{T} - 2 \sum_{B_i\in\mathcal{C}} \hat{\lambda}_{\mathcal{C}}(B_i) \ln \frac{\hat{\lambda}_{\mathcal{C}}(B_i)}{\nu_{\mathcal{C}}(B_i)}. \tag{27}$$

On peut voir que cette expression est similaire à la formulation classique du critère d'Akaïke [35], mais adapté ici au contexte de l'histogramme. Si la fonction de coût $W(\lambda, \hat{\lambda})$ est exprimée à l'aide du contraste de KullBack [39], cela conduit à deux nouveaux critères ϕ^* et AIC^* respectivement définis par :

$$\phi^*(C) = \frac{C(1+\ln(\ln T))}{T} - 2 \sum_{B_i\in\mathcal{C}} \hat{\lambda}_{\mathcal{C}}(B_i) \ln \frac{\hat{\lambda}_{\mathcal{C}}(B_i)}{\nu_{\mathcal{C}}(B_i)} \tag{28}$$

et :

$$AIC^*(C) = \frac{C(1+\ln T)}{T} - 2 \sum_{B_i\in\mathcal{C}} \hat{\lambda}_{\mathcal{C}}(B_i) \ln \frac{\hat{\lambda}_{\mathcal{C}}(B_i)}{\nu_{\mathcal{C}}(B_i)}. \tag{29}$$

On reconnaît dans l'équation (28) une formulation semblable à celle du critère d'Hannan et Quinn et dans l'équation (29) une formulation synthétisant les critère d'Akaïke et Rissanen. L'ensemble de ces trois critères peut être synthétisé sous la forme suivante :

$$IC(C) \triangleq g(C) - 2 \sum_{B_i\in\mathcal{C}} \hat{\lambda}_{\mathcal{C}}(B_i) \ln(\frac{\hat{\lambda}_{\mathcal{C}}(B_i)}{\nu_{\mathcal{C}}(B_i)}) \tag{30}$$

où $g(C)$ est un terme de pénalité, qui diffère selon le critère d'information choisi, $\nu_{\mathcal{C}}$ une loi *a priori* qui sera la loi uniforme pour traduire l'absence de connaissance *a priori* sur la structure et C le nombre de classes de la partition $\mathcal{C}$. Ces critères [1] peuvent être utilisés pour sélectionner l'histogramme à C classes approchant la loi inconnue du T-échantillon $\epsilon_1 \ldots \epsilon_T$. Des démonstrations détaillées sont disponibles dans [34].

3.2.4 Construction de l'histogramme optimal

Initialement, un histogramme à $C_{init} =\mid \mathcal{C} \mid= 2 \times [\sqrt{T} - 1]$ classes de même pas (même largeur) est construit sur la partition $\mathcal{C}$, où $[\,]$ est la partie entière. Le choix du nombre initial de classes est conforme à ce qui est préconisé dans [40] et utilisé dans [34]. Ensuite, une partition à $(C_{init} - 1)$ classes est considérée. Pour chaque fusion parmi les $(C - 1)$ fusions possibles de deux classes adjacentes de l'histogramme à C classes, le critère $IC(C - 1)$ est calculé. Le choix de la meilleure fusion est guidé par la minimisation de la quantité $IC(C - 1)$. Ceci fait, la recherche de la meilleure partition à $(C - 2)$ classes est menée selon le même principe que précédemment. Finalement, l'histogramme à C classes tel que $IC(C)$ est minimum pour $C \in \{1, \ldots, C_{init}\}$ est retenu. On note $\mathcal{C}_{opt}$ cette sous-partition optimale à C_{opt} classes définissant la meilleure estimation $\hat{\lambda}_{C_{opt}}$ de la loi inconnue λ. La figure FIG. 1 montre un histogramme initial construit à l'aide d'un échantillon de taille $T = 90$ généré aléatoirement selon une loi gaussienne. L'histogramme final obtenu respectivement par l'utilisation des critères AIC, AIC^* et ϕ^* est donné en figure FIG. 3. La figure FIG. 2 donne le comportement des trois critères en fonction du nombre de classes. On constate que les critères AIC^* et ϕ^* donnent le même histogramme final. Le critère AIC donne quant à lui un histogramme final disposant d'un nombre de classes plus élevé. Cette différence tient au type de convergence des différents critères d'information [34]. On notera que le critère AIC tend à induire une sur-paramétrisation, inconvénient déjà montré dans la littérature sur des problèmes de sélection de modèles par ce type de critère.

[1] Pour des raisons de convergence, il est préférable d'utiliser soit le critère AIC^* soit le critère ϕ^*.

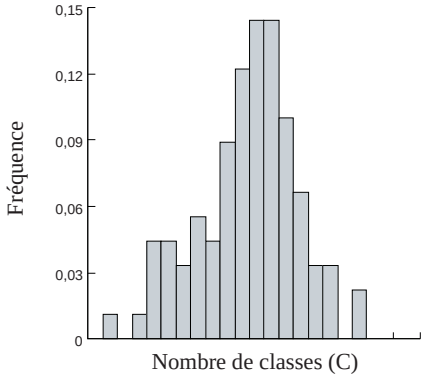

FIG. 1: Histogramme original.

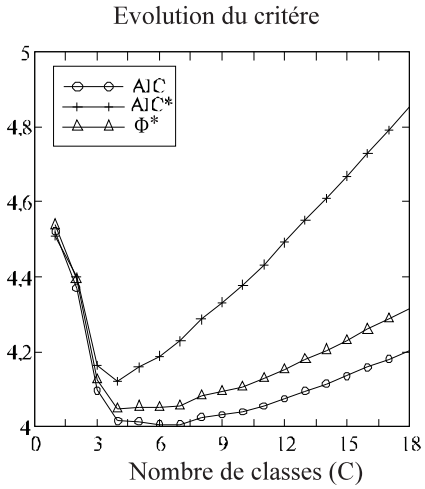

FIG. 2: Evolution des critères d'information en fonction du nombre de classe.

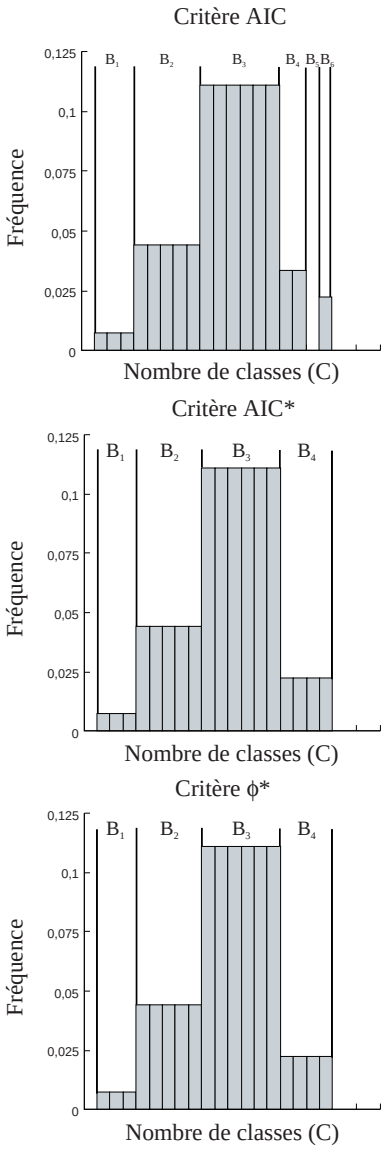

FIG. 3: Histogrammes optimaux selon le critère d'information.

3.2.5 Calcul du coefficient de confiance

La dissemblance entre deux estimations de lois peut être calculée en introduisant la distance de Hellinger [34, 38]. La partition optimale des données de la source S_j sous l'hypothèse H_n, $\mathcal{C}_{opt}^j/H_n$ à C_{opt}^j/H_n classes est tout d'abord construite à partir de la base d'apprentissage et de la base de validation. Une fois obtenue, on en déduit $\hat{\lambda}^a_{C_{opt}^j/H_n}$ l'estimation de la loi sur $\mathcal{C}_{opt}^j/H_n$ de la base d'apprentissage et $\hat{\lambda}^v_{C_{opt}^j/H_n}$ l'estimation de la loi sur $\mathcal{C}_{opt}^j/H_n$ de la base de validation . La distance de Hellinger entre ces deux estimations de lois est donnée par [34] :

$$Hell(\hat{\lambda}^a_{C_{opt}^j/H_n}, \hat{\lambda}^v_{C_{opt}^j/H_n}) \triangleq 1 - \sum_{i=1}^{C_{opt}^j/H_n} \sqrt{\hat{\lambda}^a_{C_{opt}^j/H_n}(B_i) \times \hat{\lambda}^v_{C_{opt}^j/H_n}(B_i)}. \tag{31}$$

Lorsque les estimations de lois sont proches, alors la distance tend vers 0. Au contraire, si les lois sont fortement dissemblables, alors la distance est proche de 1. Le sens de variation de cette distance est l'inverse de celui du coefficient de confiance q_{nj} que nous avons défini au début de la section 3. Nous prendrons alors le coefficient q_{nj} égal à :

$$q_{nj} = 1 - Hell(\hat{\lambda}^a_{C_{opt}^j/H_n}, \hat{\lambda}^v_{C_{opt}^j/H_n}). \tag{32}$$

.

3.3 Fusion et décision

Dans le cadre de la méthodologie proposée, nous utilisons la loi de combinaison de Dempster [20] afin de fusionner les N jeux de masses obtenus pour chacune des sources. On note alors les jeux de masses résultants m_j définis par :

$$m_j(.) = \bigoplus_{n \in [1,N]} m_{nj}(.). \tag{33}$$

Le jeu de masses, noté m, résultant de la fusion des Q structures de croyance m_j est lui aussi obtenu à l'aide de la somme orthogonale de Dempster :

$$m(.) = \bigoplus_{j \in [1,Q]} m_j(.). \tag{34}$$

Une fois le jeu de masses résultant obtenu, la décision d'affectation du vecteur X' à l'hypothèse H_n peut être prise à l'aide d'une fonction de décision répondant à l'équation (14).

4 Résultats expérimentaux

La méthode que nous avons présentée précédemment a été testée sur des données synthétiques afin d'évaluer l'apport, en terme de discrimination, des coefficients de fiabilité des sources d'information relativement à une hypothèse. Pour cela,

nous considérons les différents exemples introduit par Appriou [1]. Dans un premier temps, nous verrons l'efficacité des coefficients de fiabilité dans le cas de la fusion d'un capteur incertain avec un capteur sûr en situation de contexte évolutif (Section 4.1). Ensuite, nous présenterons une situation à deux capteurs et deux données incertaines (Section 4.2).

4.1 Deux capteurs et une donnée incertaine

Nous considérons un problème de discrimination à deux hypothèses H_1 et H_2. Le but de ce test est de montrer l'efficacité des coefficients de fiabilité déterminés à l'aide des mesures d'information afin de pallier l'incertitude d'un capteur de bonne qualité en lui associant un capteur de moins bonne qualité mais sûr. Les distributions $p(x'_j/H_n)$ apprises par les capteurs S_j sont supposées être des lois normales telles que :

$$p(x'_1/H_1) = p(x'_2/H_1) = \mathcal{N}(0,1), \tag{35}$$

$$p(x'_1/H_2) = \mathcal{N}(2,1), \tag{36}$$

$$p(x'_2/H_2) = \mathcal{N}(6,1). \tag{37}$$

Les mesures x'_j relevées dans la réalité sont générées à partir de lois normales différentes telles que :

$$p(x'_1/H_1) = p(x'_2/H_1) = \mathcal{N}(0,1), \tag{38}$$

$$p(x'_1/H_2) = \mathcal{N}(2,1), \tag{39}$$

$$p(x'_2/H_2) = \mathcal{N}(S,1). \tag{40}$$

La base de validation nous permettant de définir les valeurs de coefficients de fiabilité est définie sur le même modèle que la base de test. Les résultats en terme de taux de classification sont représentés sur la figure FIG. 4. D'après cette figure, nous pouvons constater que la démarche de calcul de coefficient de fiabilité présentée ici permet d'obtenir un gain de classification notable par rapport à une valeur de coefficient fixée à 0.9 lorsque la base d'apprentissage n'est plus représentative des données de la base de test. Lorsque l'apprentissage représente correctement les données de test $(S \approx 6)$, les performances obtenues sont alors identiques dans les trois cas. La figure FIG. 5 représente l'évolution des coefficients de fiabilité déterminés à l'aide de critères d'information. Nous constatons que le coefficient q_{22} accordé au capteur S_2 sous l'hypothèse H_2 varie en fonction du signal S. Ce coefficient est proche de 0, lorsque l'apprentissage n'est pas fiable $S \approx 0$ et proche de 1 dans le cas contraire. Les autres coefficients de fiabilité sont constants en fonction du signal S et proche de 1 car l'apprentissage est représentatif du contexte.

4.2 Deux capteurs et deux données incertaines

Nous considérons de la même manière que précédemment un problème à deux capteurs et deux hypothèses. Les deux capteurs sont identiques et possède un bon pouvoir discriminant. Toutefois, l'apprentissage de l'hypothèse H_2 pour les deux

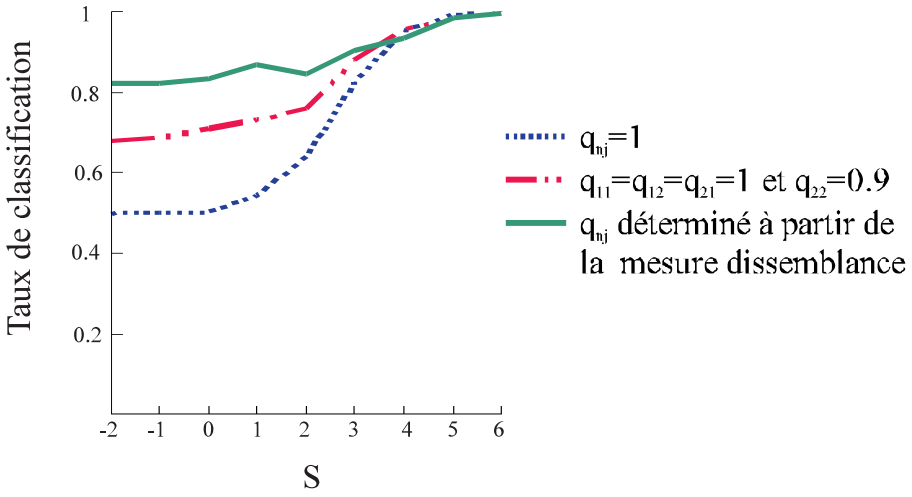

FIG. 4: Evolution du taux de classification en fonction du signal S

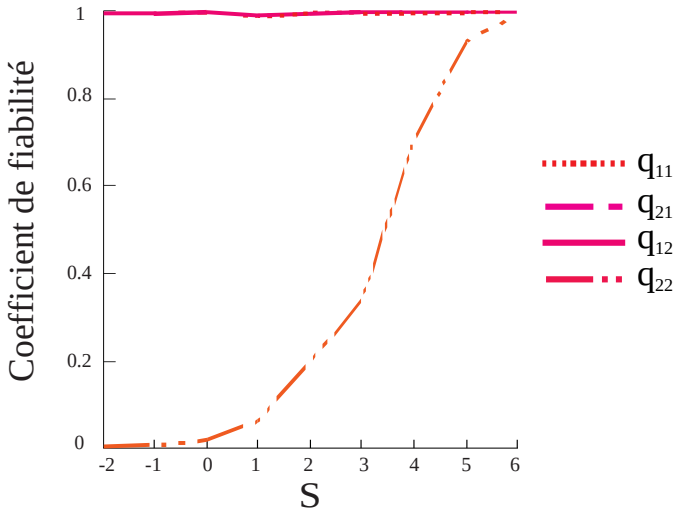

FIG. 5: Evolution des coefficients de fiabilité déterminés à l'aide de la distance de Hellinger en fonction du signal S

capteurs est douteux car le contexte lié à cette hypothèse a évolué. L'intérêt d'une telle configuration est de voir l'apport, en terme de taux de classification, des coefficients de confiance dans le cadre de la fusion de capteurs peu fiables. De manière pratique, les données d'apprentissages sont simulées à l'aide de lois gaussiennes telles que :

$$p(x_1'/H_1) = p(x_2'/H_1) = \mathcal{N}(0,1), \tag{41}$$

$$p(x_1'/H_2) = p(x_2'/H_2) = \mathcal{N}(6,1). \tag{42}$$

Les mesures réellement simulées satisfont en revanche :

$$p(x_1'/H_1) = p(x_2'/H_1) = \mathcal{N}(0,1), \tag{43}$$

$$p(x_1'/H_2) = \mathcal{N}(2,1) \tag{44}$$

$$p(x_2'/H_2) = \mathcal{N}(S,1). \tag{45}$$

Nous pouvons alors constater que le premier capteur ne reconnaît pas correctement l'hypothèse H_2. De plus, la capacité du deuxième capteur à détecter cette même hypothèse est dépendante d'un signal S, représentant une évolution possible du contexte sous l'hypothèse H_2. L'évolution du taux de classification en fonction du signal S est représentée sur la figure FIG. 6. Nous pouvons constater que, dans le cas d'un apprentissage non adapté ($S \approx -2$), les taux de bonne classification obtenus en affaiblissant ($q_{nj} < 1$) les informations issues des source S_1 et S_2 pour l'hypothèse H_2 sont meilleurs que dans le cas non affaibli ($q_{nj} = 1$). De plus, les coefficients obtenus à l'aide de la distance de Hellinger permettent d'obtenir de meilleurs résultats que dans le cas d'un coefficient q_{nj} fixe égal à 0.9. Dans ce test, la différence entre les performances, obtenues avec des coefficients de fiabilité déterminés de façon arbitraire et celles obtenues à l'aide de coefficients issus de critère d'information, est moins importante que dans le cas du premier test (Section 4.1). Ceci résulte d'un affaiblissement plus important de la source S_1 sous l'hypothèse H_2 ($q_{21} < 0.9$) (figure FIG. 7). Enfin, dans le cas d'un apprentissage fiable ($S \approx 6$), les taux de classification sont identiques quelque soit la valeur des coefficients de qualité. Les variations des

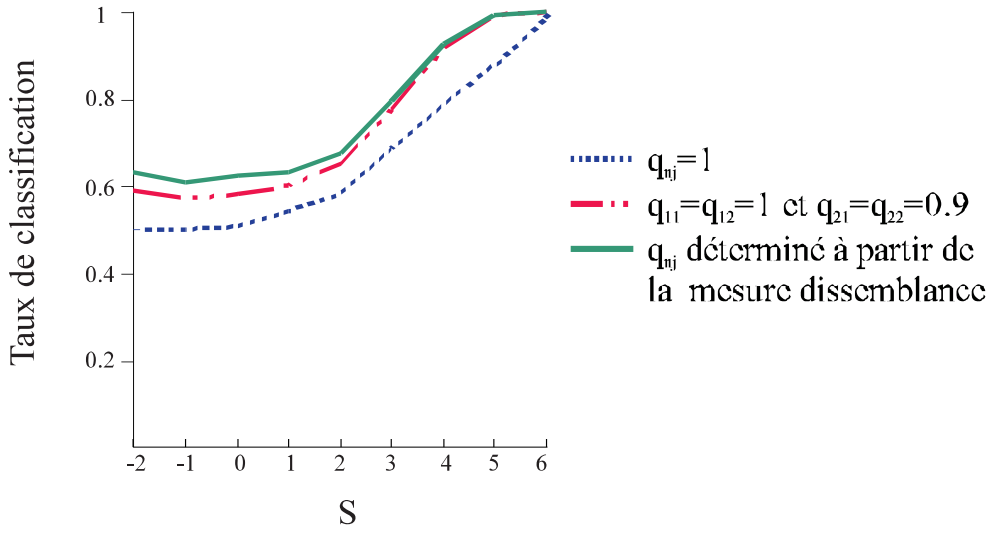

FIG. 6: Evolution du taux de classification en fonction du signal S

coefficients q_{nj} obtenus à l'aide de la distance d'Hellinger sont représentées en fonction du signal S sur la figure FIG. 7. Nous constatons que le coefficient q_{22} évolue de manière croissante en fonction du signal S. En effet, la qualité des informations issues du capteur augmente lorsque le signal S se rapproche de l'apprentissage. Les autres coefficients sont constants en fonction de ce signal. Les coefficients q_{11} et q_{12} qui s'appliquent à chacune des sources relativement à l'hypothèse H_1 sont égaux à 1, car les mesures réelles sont proches de l'apprentissage. Le coefficient q_{21} est égal à 0.2, car le premier capteur reconnaît mal l'hypothèse H_2. Globalement, la méthode permet d'adapter des coefficients d'affaiblissement et d'atteindre, dans tous les cas, des taux de classification supérieurs à ceux obtenus avec un coefficient d'affaiblissement fixe ($q_{nj} = 0.9$).

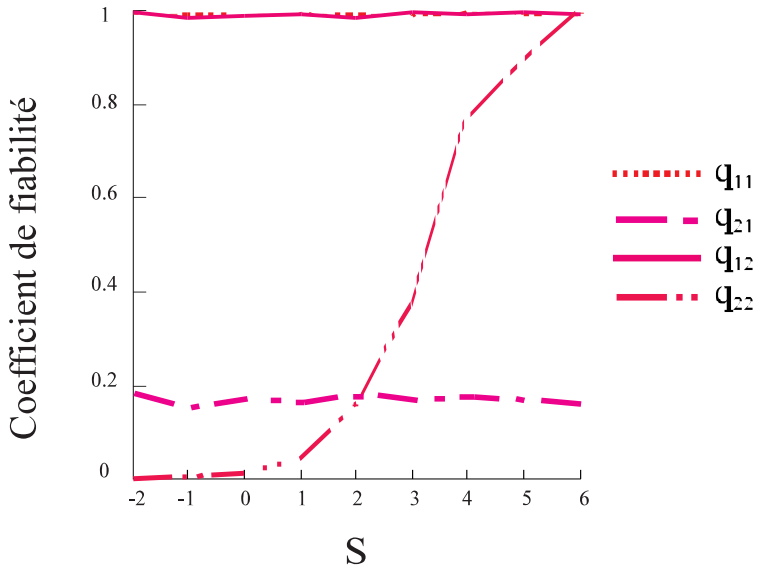

FIG. 7: Evolution des coefficients de qualité déterminé à l'aide de la distance de Hellinger en fonction du signal S

5 Conclusion et perspectives

Nous avons présenté une méthode de calcul de coefficients de fiabilité accordés à une source dans le cadre de la discrimination fondée sur la théorie de Dempster-Shafer. La méthode consiste, après avoir obtenu les jeux de masses à l'aide des densités de probabilité définies par apprentissage, à déterminer, à l'aide d'une distance *ad hoc*, un paramètre permettant de connaître le degré de fiabilité q_{nj} d'une source d'information S_j relativement à une hypothèse H_n, ceci afin d'affaiblir les sources les moins fiables de manière adaptée. Des tests effectués sur des données synthétiques montrent l'apport de l'affaiblissement par une telle approche. Nos travaux futurs porteront sur la règle de décision. Actuellement, la décision n'est prise que sur des hypothèses singletons. Ceci nous amène à prendre une décision précise, qui peut être au détriment de la fiabilité. La prise de décision sur des hypothèses composites et l'introduction d'une classe de rejet permettraient d'améliorer les résultats de classification.

6 Remerciements

Les auteurs tiennent à remercier le Professeur T. Denoeux, du laboratoire Heudiasyc de l'Université de Technologie de Compiègne, pour ses remarques et ses conseils qui leur ont permis d'améliorer cet article, ainsi que les rapporteurs qui par leurs commentaires riches et constructifs, ont permis de préciser plusieurs points de cet article.

Références

[1] A. Appriou, “Probabilités et incertitude en fusion de données multi-senseurs,” *Revue Scientifique et Technique de la Défense*, vol. 11, pp. 27–40, 1991.

[2] A. Appriou, “Multisensor signal processing in the framework of the theory of evidence,” in *Application of Mathematical Signal Processing Techniques to Mission Systems*, pp. (5–1)(5–31), Research and Technology Organization (Lecture Series 216), November 1999.

[3] I. Bloch and H. Maître, “Fusion de données en traitement d’images : Modèles d’informations et décisions,” *Traitement du Signal*, vol. Numéro Spécial : Fusion de données, no. 11, pp. 435–446, 1994.

[4] I. Bloch, “Some aspects of Dempster-Shafer evidence theory for classification of multi-modality medical images taking partial volume effect into account,” *Pattern recognition Letters*, vol. 17, pp. 905–919, 1996.

[5] G. Cooper and E. Herskovits, “A bayesian method for constructing bayesian belief networks from databases,” in *Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence*, (Los Angeles), pp. 86–94, Morgan Kaufman, 1991.

[6] U. Fayyad and P. Smyth, “Image database exploration : Progress and challenge,” in *Proceedings of the 1993 AAAI Workshop on Knowledge Discovery in Database*, 1993.

[7] D. Heckerman, D. Geiger, and D. Chickering, “Learning bayesian networks : The combination of knowledge and statistical data,” *Machine Learning*, vol. 20, pp. 197–243, 1995.

[8] A. Appriou, “Formulation et traitement de l’incertain à l’analyse multi-senseurs,” in *14ème Colloque GRETSI*, pp. 951–954, 1993.

[9] S. Chauvin, *Evaluation des théories de la décision appliquées à la fusion de capteurs en imagerie satellitaire*. PhD thesis, Université de Nantes, 1995.

[10] T. Carron, *Segmentation d’images couleur dans la base teinte-luminance-saturation : approche numérique et symbolique*. PhD thesis, Université de Savoie, 1995.

[11] J. Desachy, L. Roux, and E. Zahzah, “Numeric and symbolic data fusion : A soft computing approach to remote sensing images analysis,” *Pattern Recognition Letters*, vol. 17, pp. 1361–1378, 1996.

[12] M. Degroot, *Optimal Statistical Decisions*. New York : McGraw-Hill, 1970.

[13] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. New-York : John Wiley & Sons, 1973.

[14] B. Dubuisson, *Diagnostic et Reconnaissances de Formes*. Hermes, 1990.

[15] D. Dubois and H. Prade, *Decison Evaluation Methods under Uncertainty and Imprecision*, vol. 310 of *Lecture Notes in Economics and Mathematical Systems : Combining Fuzzy Imprecision with the Probabilistic Uncertainty in Decision Making*, pp. 48–65. Berlin : Springer-Verlag, j. kacprzyk and m. federizzi ed., 1987.

[16] J. Bezdek, "Fuzziness vs. probability - the n-th round," *IEEE Trans. on Fuzzy Systems*, vol. 2, no. 1, pp. 1–42, 1994.

[17] D. Dubois and H. Prade, "The use of fuzzy numbers in decision analysis," in *Fuzzy Information and Decision Processes* (M. Gupta and E. Sanchez, eds.), (New-York), pp. 309–321, North-Holland, 1982.

[18] D. Dubois, "Belief structures, possibility theory and decomposable confidence measures on finite sets," *Comput. Artif. Intell.*, vol. 5, no. 5, pp. 403–416, 1986.

[19] D. Dubois and H. Prade, *Possibility Theory : An Approach to Computerized Processing of Uncertainty*. New-York : Plenum Press, 1988.

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

[21] A. Dempster, "Upper and lower probabilities induced by multivalued mapping," *Annals of Mathematical Statistics*, vol. AMS-38, pp. 325–339, 1967.

[22] P. Smets, "What is Dempster-Shafer's model ?," in *Advances in the Dempster-Shafer Theory of Evidence* (R. Yager, M. Fedrizzi, and J. Kacprzyk, eds.), pp. 5–34, Wiley, 1994.

[23] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[24] P. Smets, "The combination of evidence in the transferable belief model," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 12, no. 5, pp. 447–458, 1990.

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

[26] A. Nifle and R. Reynaud, "Un argument pour le choix entre décision pignistique et maximum de plausibilité en théorie de l'évidence," in *Seizième Colloque GRETSI*, (Grenoble), pp. 1411–1414, 1997.

[27] L. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets and Systems*, vol. 1, pp. 3–28, 1978.

[28] R. Yager, "On the Dempster-Shafer framework and new combination rules," *Information Sciences*, vol. 41, pp. 93–138, 1987.

[29] D. Dubois and H. Prade, "A set-theoric view of belief functions : Logical operations and approximations by fuzzy sets," *International Journal of General Systems*, vol. 12, pp. 193–226, 1986.

[30] P. Smets, "Constructing the pignistic probability function in a context of uncertainty," in *Uncertainty in Artificial Intelligence 5* (M. Henrion, R. D. Schachter, L. Kanal, and J. Lemmer, eds.), (Amsterdam), pp. 29–40, North-Holland, 1990.

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

[32] S. Mathevet, “Application de la théorie de l’évidence à la combination de segmentations en régions,” in *Dix-Septième Colloque GRETSI (Vannes)*, pp. 635–638, Septembre 1999.

[33] D. de Brucq, “Characterization of the optimal number of classes for a histogram with Hellinger’s distance,” in *Signal Processing IV : Theories and Applications* (J. Lacoume, A. Chehikian, N. Martin, and J. Malbos, eds.), pp. 1117–1120, EUSICOP’88, September, 1988.

[34] O. Colot, C. Olivier, P. Courtellemont, A. El-Matouat, and D. de Brucq, “Information criteria and abrupt changes in probability laws,” in *Signal Processing VII : Theories and Applications* (M. Holt, C. Cowan, P. Grant, and W. Sandham, eds.), pp. 1855–1858, EUSIPCO’94, September 1994.

[35] H. Akaike, “Information theory and an extension of the maximum likelihhod principle,” in *Proc. Of the 2nd Int. Symp. Of Info. Theory* (B. Petrov and F. Csaki, eds.), (Budapest), pp. 267–281, 1973.

[36] J. Rissanen, T. Speed, and B. Yu, “Density estimation by stochastic complexity,” *IEEE Trans. on Information Theory*, vol. 58, no. 2, pp. 315–323, 1992.

[37] E. Hannan, “The estimation of the order of an ARMA process,” *Annals of Statistics*, vol. 8, no. 5, pp. 1071–1081, 1980.

[38] M. Basseville, “Distance measures for signal processing and pattern recognition,” Rapport de recherche interne 899, INRIA, Rennes, 1988.

[39] A. El-Matouat and C. Olivier, “Sélection du nombre de classe d’un histogramme et contraste de Kullback,” in *25ème Journées Internationales de Statistiques*, pp. 193–196, ASU, 1993.

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