Mesure de concordance pour les bases de données évidentielles

Mouna Chebbah*, Arnaud Martin**
Boutheina Ben Yaghlane*

*LARODEC, Université de Tunis, ISG Tunis, Tunisie
Mouna.Chebbah@gnet.tn
boutheina.yaghlane@ihec.rnu.tn
**UMR 6074 IRISA, Université de Rennes1 / IUT de Lannion, France
Arnaud.Martin@univ-rennes1.fr

Résumé. Dans cet article, nous proposons une mesure de concordance d'une source avec les autres sources. Cette mesure pourra servir à réduire l'importance de ses fonctions de masse avant de les combiner afin de trouver un compromis et donc réduire le conflit. Cette mesure sera illustrée par des données réelles.

1 Introduction

Dans le cadre certain, les bases de données permettent de stocker une grande quantité d'information certaine où les valeurs des attributs sont précises. Néanmoins, les données stockées ne sont pas toutes parfaites et certaines, elles sont généralement entachées d'incertitude. Pour aborder le problème de stockage des données imparfaites et incertaines, des bases de données évidentielles [1] ont été proposées par Hewawasam et al. (2005) et Bach Tobji et al. (2008). L'intégration des bases de données évidentielles permet, d'une part, de réduire la quantité d'informations à stocker si on dispose de plus d'une base de données, et d'autre part, d'aider les utilisateurs qu'ils soient humains ou logiciels à la prise de décision en résumant les différentes bases de données évidentielles en une seule base de données intégrée.

En intégrant des bases de données évidentielles, les informations évidentielles qui concernent le même objet sont fusionnées. La prise en considération des données évidentielles fournies par différentes sources hétérogènes lors de la combinaison peut induire l'apparition d'un conflit dû à une contradiction entre ces sources. Il existe différentes manières de résolution du conflit résumées par Martin (2010). L'une de ces méthodes consiste à réduire le conflit avant de combiner en affaiblissant les informations fournies par une source avec son degré de fiabilité.

Dans cet article, nous proposons une nouvelle méthode d'estimation de la concordance d'une source. La concordance d'une source n'est autre que sa fiabilité estimée (qui n'est pas exacte) calculée en se référant à une autre source. En effet, notre méthode est sans aucun *a priori* sur les données réelles ainsi que les fiabilités des sources à part la fiabilité d'une seule

1. Ce sont des bases de données contenant des données représentées par la théorie des fonctions de croyance proposée par Dempster (1967) et Shafer (1976)

source [2]. Nous calculons alors la concordance des sources dont les fiabilités sont inconnues par rapport à la source dont la fiabilité est connue. Ces mesures de concordance serviront par la suite à affaiblir les données évidentielles avant de les combiner afin de prévenir tout conflit.

Le reste de cet article est organisé comme suit : dans la deuxième section nous présentons brièvement les notions de base de la théorie des fonctions de croyance ensuite nous présentons notre méthode d'estimation de la concordance d'une source dans la troisième section, cette méthode a été testée sur des données radar réelles que nous présentons les résultats expérimentaux dans la quatrième et dernière section.

2 Théorie des fonctions de croyance (TFC)

La théorie des fonctions de croyance (*théorie de l'évidence* ou encore *théorie de Dempster-Shafer*) a été introduite par Dempster (1967) et formalisée par Shafer (1976) afin de modéliser des données imparfaites (imprécises et/ou incertaines). *L'ensemble de discernement* $\Omega = \{\omega_1,\ \omega_2,\ \ldots,\ \omega_n\}$ est un ensemble fini de toutes les hypothèses possibles ω_i représentant toutes les valeurs pouvant être la solution d'un problème donné. Soit $2^\Omega = \{A/A \subseteq \Omega\} = \{\emptyset,\ \omega_1, \ldots,\ \omega_n,\ \omega_1 \cup \omega_2,\ \ldots,\ \Omega\}$, l'ensemble de tous les sous-ensembles possibles de Ω.

Une *fonction de masse* m^Ω est une fonction de 2^Ω vers l'intervalle $[0, 1]$ qui affecte à chaque sous-ensemble une valeur entre 0 et 1 représentant sa *masse de croyance élémentaire*. Cette fonction de masse est une représentation des connaissances incertaines et imprécises fournies par un expert (une source, un classifieur, ...). Un élément de 2^Ω ayant une masse strictement positive est *un élément focal*. La masse affectée à un élément focal A représente le degré de croyance élémentaire d'une source à ce que la solution du problème soit incluse ou égale à A. La somme des masses affectées à tous les élément focaux doit être égale à 1.

Dans le cadre de la TFC, il existe un grand nombre de règles de combinaison tel que résumé par Smets (2007). Ces règles sont utilisées pour la combinaison de différentes informations, représentées par des fonctions de masse définies sur le même ensemble de discernement, et fournies par différentes sources. La combinaison permet de résumer différentes fonctions de masse fournies par différentes sources afin d'obtenir une seule fonction de masse.

Lors de la combinaison de deux fonctions de masse, un conflit peut apparaître reflétant le degré de désaccord entre les sources. Une des origines du conflit est la non fiabilité d'au moins une des sources. La non fiabilité d'une source peut être réglée par l'affaiblissement de ses fonctions de masse avant la combinaison en utilisant l'opérateur d'affaiblissement proposé par Shafer (1976). Lorsqu'on arrive à quantifier la fiabilité α de chaque source, on peut affaiblir les fonctions de masse associées comme suit :

$$\begin{cases} m^{\Omega,\alpha}(A) = \alpha\, m^\Omega(A) & \forall A \subset \Omega \\ m^{\Omega,\alpha}(\Omega) = (1-\alpha) + \alpha\, m^\Omega(\Omega) \end{cases} \tag{1}$$

avec α le degré de fiabilité de la source et $1 - \alpha$ le facteur d'affaiblissement.

2. L'hypothèse de n'avoir que la fiabilité d'une seule source peut être réelle dans certains domaines tels que la détection de cible, la médecine, ... C'est, par exemple, le cas de plusieurs jeunes médecins et un seul ancien médecin expérimenté qui diagnostiquent une maladie. La fiabilité de l'ancien médecin a déjà été testée mais celles des jeunes médecins sont encore inconnues. C'est également le cas de l'utilisation de plusieurs capteurs pour la détection de cibles dont la fiabilité d'un seul est donnée par le constructeur.

Bases de données évidentielles (BDE) : Une BDE telle que définie par Bach Tobji et al. (2008) est une base de données contenant des données imparfaites (incertaines et/ou imprécises) représentées par les fonctions de masse précédemment décrites.

Formellement, une BDE est une base de données ayant X attributs (colonnes) et Y enregistrements (lignes), chaque *attribut évidentiel* j ($1 \leq j \leq X$) possède un domaine D_j représentant toutes les valeurs de cet attribut : *C'est son ensemble de discernement*. Les valeurs de ces attributs évidentiels sont également évidentielles. *Une valeur évidentielle* V_{ij} de l'enregistrement i ($1 \leq i \leq Y$) pour l'attribut j ($1 \leq j \leq X$) est une fonction de masse $m_{ij}^{D_j}$ précédemment décrite. Les BDE sont utilisées dans différents domaines notamment pour le stockage des fonctions de masse de différents classifieurs présenté par Hewawasam et al. (2005).

3 Mesure de concordance d'une source

L'opérateur d'affaiblissement précédemment décrit permet d'intégrer les fiabilités des sources au sein de leurs fonctions de masse quoique ces fiabilités ne sont pas toujours connues. Ainsi, nous proposons d'estimer les degrés de concordance des sources dont les fiabilités sont inconnues en se référant à la source dont la fiabilité est connue à partir de leurs conflits.

3.1 Quantification du conflit

Martin et al. (2008) considèrent le conflit entre deux sources S_1 et S_2 comme étant la distance séparant leurs fonctions de masse m_1^Ω et m_2^Ω. Dans ce contexte, l'une des distances résumées par Florea et Bossé (2009) peut être utilisée. Nous utilisons la distance de Jousselme et al. (2001) qui quantifie la distance entre deux fonctions de masse m_1^Ω et m_2^Ω comme suit :

$$d(m_1^\Omega, m_2^\Omega) = \sqrt{\frac{1}{2}(m_1^\Omega - m_2^\Omega)^t D(m_1^\Omega - m_2^\Omega)} \tag{2}$$

avec :

$$D(A,B) = \begin{cases} 1 & \text{si } A = B = \emptyset \\ \frac{|A \cap B|}{|A \cup B|} & \forall A, B \in 2^\Omega \end{cases} \tag{3}$$

Nous utilisons cette distance parce qu'elle prend en considération les spécificités des fonctions de masse [3]. La matrice D est définie positive bien que ceci n'a jamais été démontré.

Cette mesure de *conflit relatif* ne peut prendre en considération qu'une seule fonction de masse parmi celles fournies par une même source, or les BDE contiennent un grand nombre de fonctions de masse pour une source donnée. Nous définissons alors, le *conflit absolu* d'une source S_2 par rapport à une autre source S_1 comme étant la moyenne de ses M conflits relatifs calculés à partir de chaque fonction de masse.

$$Conf_a(S_1, S_2) = \frac{1}{M} \sum_{i=1}^{M} Conf_i(S_1, S_2) \tag{4}$$

3. La matrice D est définie sur 2^Ω et les masses sont calibrées par le coefficient de Jaccard

3.2 Mesure de concordance

Il existe un lien entre le conflit et la concordance puisque le degré de conflit entre deux sources renseigne indirectement sur leur ressemblance. Afin d'illustrer ce lien, nous avons effectué des tests sur des données générées aléatoirement pour deux sources. Des fonctions de masse sont générées en choisissant, aléatoirement, l'ensemble des éléments focaux pour chaque source. Ensuite, nous avons partagé l'intervalle $[0, 1]$ en sous-intervalles continus de même taille. Le nombre de ces sous intervalles est le même que le nombre d'éléments focaux. Enfin, la masse d'un élément focal est un nombre tiré aléatoirement de l'un de ces sous-intervalles et la masse restante (le complément à 1 de la somme de toutes les masses des éléments focaux) est attribuée à l'ignorance totale. Le nombre de sources est fixé à deux puisque dans tous les cas nous calculons le conflit d'une source par rapport à une autre qui sert de référence, c'est la source dont la fiabilité est disponible.

Afin de voir le lien entre les fiabilités des deux sources et leur conflit absolu, les fiabilités des deux sources S_1 et S_2 sont fixées lorsque les fonctions de masse sont générées. La figure 1 est une représentation du conflit de la source S_2 par rapport à la source S_1 tout en variant les différentes fiabilités de ces sources. Cette mesure de conflit absolu est stable car

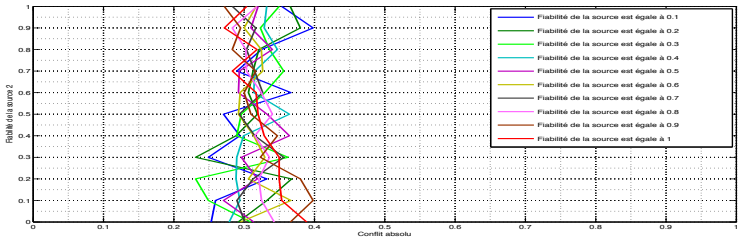

FIG. 1 – *Conflits absolus théoriques entre deux sources S_1 et S_2*

nous obtenons presque les mêmes valeurs et le même graphe en changeant les fonctions de masse aléatoires. Puisque cette mesure est stable, nous proposons une méthode d'estimation de la concordance fondée sur des simulations. Étant données deux BDE contenant des données réelles et appartenant à deux sources différentes S_1 et S_2. La fiabilité de la source S_1 ainsi que la cardinalité du cadre de discernement sont supposées être connues. La méthode d'estimation de la concordance est fondée sur les deux étapes suivantes :

- *Étape 1 : Simulation*. Nous réalisons une simulation du conflit absolu en utilisant *des données générées* sur un cadre de discernement de même cardinalité que celui des données réelles. Lors de cette simulation, nous fixons la fiabilité d'une source $S_1^\star$ au même niveau de fiabilité que S_1 sur des données générées. La fiabilité de $S_2^\star$ est une valeur variable dans $[0, 1]$. Le conflit absolu de $S_2^\star$ est représenté par une courbe en fonction de la fiabilité de $S_1^\star$ qui est fixe et de $S_2^\star$ qui est variable. Le but de cette étape est d'obtenir une courbe similaire aux courbes de la figure 1.
- *Étape 2 : Mesure de la concordance*. Le conflit absolu de S_2 par rapport à S_1 est calculé à partir des fonctions de masse stockées dans les BDE (les données réelles). La concordance de S_2 n'est autre que la fiabilité de $S_2^\star$ correspondant à son conflit absolu à partir de la courbe de la première étape.

4 Expérimentations

Afin de pouvoir illustrer la méthode précédemment décrite sur un exemple réel, nous avons considéré une base de données radar. Ces données ont été recueillies dans la chambre anéchoïque de l'ENSIETA en plaçant une cible (maquette d'avion) et un capteur radar pouvant détecter la cible sous différents points angulaires. Une base de données a été proposée pour l'acquisition et le stockage des signaux par Toumi (2007). Nous considérons ainsi cinq cibles radar différentes (Mirage, F14, Rafale, Tornado, Harrier). Chaque table contient 250 représentations fréquentielles obtenues dans un domaine angulaire d'environ $60°$ et utilisant une bande de fréquence d'environ 6 GHZ. Pour caractériser les cibles, et donc renseigner la base de données, nous avons utilisé trois classifieurs différents : le k-plus proche voisin flou, le k-plus proche voisin crédibiliste et un réseau de neurones. Ces trois classifieurs sont considérés comme des sources, sur lesquelles on déduit des fonctions de masse. Le système d'acquisition et la description des différents classifieurs sont détaillés par Martin et Radoi (2004). Les classifieurs (sources) ont donc fourni 250 fonctions de masse stockées dans trois tables différentes et permettant de classifier les cinq cibles radar différentes.

Notre but est d'intégrer ces trois tables en combinant les 250 fonctions de masse fournies par chaque source (classifieur). Les fiabilités calculées à partir des matrices de confusion par le taux de bonne reconnaissance du k-plus proche voisin flou, du k-plus proche voisin crédibiliste et du réseau de neurones sont respectivement égales à 0.948, 0.972 et 0.812.

Afin de tester notre méthode, nous faisons l'hypothèse que nous disposons uniquement de la fiabilité du k-plus proche voisin flou (le choix est arbitraire, on pourra choisir n'importe quelle autre source). Lors de la première étape de simulation, nous avons généré le graphe de la figure 2. Ensuite, dans la deuxième étape nous avons mesuré la concordance du k-plus

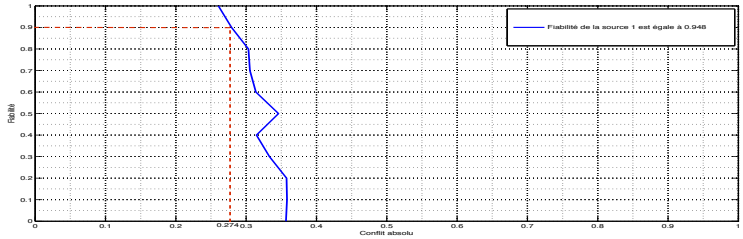

FIG. 2 – *Conflits absolus de $S_2^\star$ en fonction de sa fiabilité calculés à partir de* 250 *fonctions de masse générées aléatoirement sur un ensemble de discernement de cardinalité* 5

proche voisin crédibiliste et du réseau de neurones qui sont respectivement égale à 1 et 0.9 . Nous constatons que ces valeurs de concordance sont très proches des fiabilités réelles.

La dernière étape consiste à affaiblir les fonctions de masse proportionnellement à ces valeurs de concordance avant la phase de combinaison lors de l'intégration des BDE.

5 Conclusion

Dans cet article, nous avons proposé une nouvelle mesure, la concordance, permettant d'assurer un compromis entre des sources lors de l'intégration de leurs BDE. En effet, un conflit

peut apparaître en combinant les fonctions de masse stockées dans ces BDE. Ce conflit pourra être prévenu en affaiblissant ces fonctions de masse avec les fiabilités de leurs sources. Notre méthode suppose l'absence de tout *a priori* sur les données réelles et les fiabilités des sources à part la fiabilité d'une seule source. Nous proposons alors, une mesure de concordance estimant la fiabilité d'une source en se référant à une autre source dont la fiabilité est connue.

Remerciements Ce travail a été validé grâce aux données radar fournies par le laboratoire de recherche E^3I^2, EA3876, ENSIETA (Brest, France).

Références

Bach Tobji, M.-A., B. Ben Yaghlane, et K. Mellouli (2008). A new algorithm for mining frequent itemsets from evidential databases. In *IPMU'2008*, Malaga, Spain, pp. 1535–1542.

Dempster, A. P. (1967). Upper and Lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics 38*, 325–339.

Florea, M. C. et E. Bossé (2009). Crisis management using Dempster Shafer theory : Using dissimilarity measures to characterize sources' reliability. In *C3I for Crisis, Emergency and Consequence Management*, Bucharest, Roumania.

Hewawasam, K., K. Premaratne, S. Subasingha, et M.-L. Shyu (2005). Rule mining and classification in imperfect databases. In *Int. Conf. on Information Fusion*, Philadelphia, USA, pp. 661–668.

Jousselme, A.-L., D. Grenier, et E. Bossé (2001). A new distance between two bodies of evidence. *Information Fusion 2*, 91–101.

Martin, A. (2010). Le conflit dans la théorie des fonctions de croyance. In *Actes Extraction et gestion des connaissances (EGC'2010)*, Hammamet, Tunisia, pp. 655–666.

Martin, A., A.-L. Jousselme, et C. Osswald (2008). Conflict measure for the discounting operation on belief functions. In *Int. Conf. on Information Fusion*, Cologne, Germany, pp. 1003–1010.

Martin, A. et E. Radoi (2004). Effective ATR Algorithms Using Information Fusion Models. In *Int. Conf. on Information Fusion*, Stockholm, Sweden, pp. 161–166.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton University Press.

Smets, P. (2007). Analyzing the combination of conflicting belief functions. *Information Fusion 8*, 387–412.

Toumi, A. (2007). *Intégration des bases de connaissances dans les systèmes d'aide à la décision : Application à l'aide à la reconnaissance de cibles radar non-coopératives*. Ph. D. thesis, Université de Bretagne Occidentale, ENSIETA, Brest.

Summary

In this paper, we suggest a new measure, named concordance, in order to replace source's reliability when this reliability is unknown. This measure was tested on real radar data and can be used to discount belief functions before the combination step when integrating EDB.