Fusion numérique d'informations multi-sources et extraction de connaissances : Application à l'ingénierie du trafic

NOUR-EDDIN EL FAOUZI

Laboratoire d'Ingénierie Circulation-Transports
INRETS-ENTPE
25, avenue François Mitterrand
69675 Bron Cedex.
elfaouzi@inrets.fr

Résumé. La fusion de données permet d'enrichir une série de données parcellaires, souvent issues de sources multiples, en combinant les informations qu'elles contiennent afin d'améliorer la qualité des connaissances extraites de ces données. L'un des objectifs de la fusion de données multi-sources, ainsi définie, est l'extraction de connaissances (KDD).
Dans cet article, plusieurs schémas de fusion de données sont proposés, exploitant à la fois la complémentarité et la redondance de ces informations muti-sources. Ces schémas sont ensuite mis en œuvre pour l'élaboration d'un indicateur de la qualité de la circulation du trafic (temps de parcours) sur la base de données issues de capteurs de trafic traditionnels et de véhicules traceurs (véhicules équipés de capteurs embarqués).

1 Introduction

Les nouvelles technologies de l'information et les progrès qu'ont connu les systèmes de recueil et de collecte de données ont favorisé l'émergence de nouvelles sources de données (nouveaux capteurs de mesure dotés de grande précision, capteurs embarqués à bord de mobile, localisation satellitaire,…. Celles-ci permettent de disposer d'informations de plus en plus riches et complexes, de nature et de fiabilité diverses avec de fortes exigences tant en terme de qualité que de performances. Ainsi, le décideur comme le chercheur se trouve confronter, de plus en plus, à des flux de données sans cesse croissants, souvent hétérogènes, parfois contradictoires dont il est bien difficile de faire la synthèse. La réponse à ce besoin a récemment ouvert un nouveau champ, globalisant les traitements des données disponibles, connu sous les noms de « data mining » (Han et Kamber 2001) et d'*extraction automatique des connaissances* (KDD) (Klosgen et Zytkow 2002). Ce champ se propose de fournir des techniques et des outils d'assistance à l'analyse et à l'extraction automatique des connaissances, prenant simultanément en compte l'ensemble des différentes sources d'informations disponibles.

La fusion de données dont il s'agit dans ce travail doit être distinguée de la fusion de bases de données ou de fichiers (Saporta 2002). Cette dernière consiste à agréger les bases de

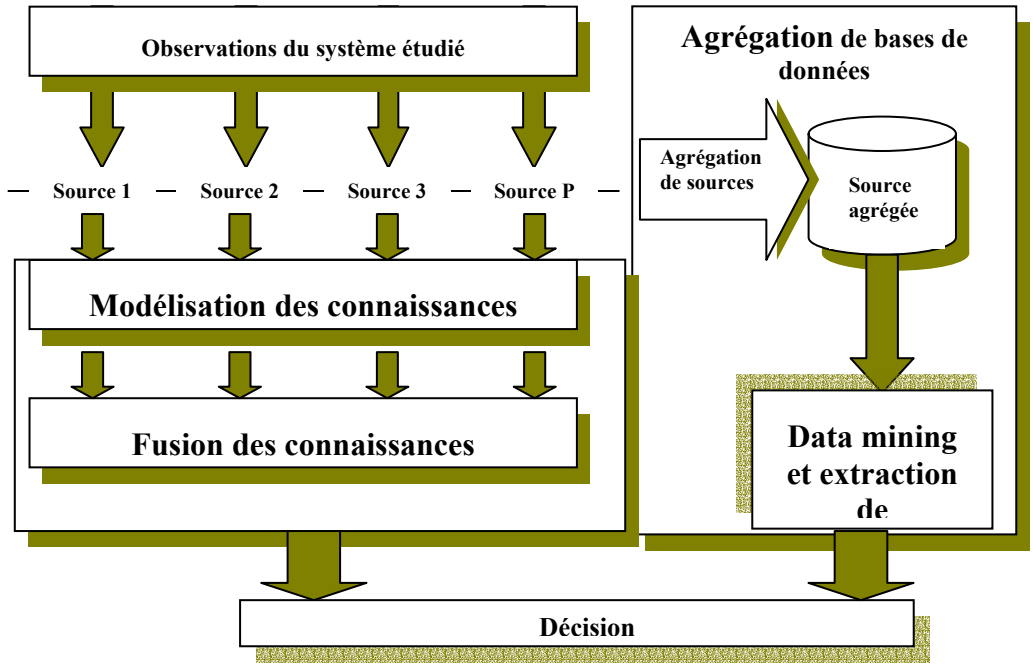

FIG. 1 - *Principes de la fusion de données et de l'agrégation de données*

données disponibles de façon à disposer d'une seule et base, sur laquelle les techniques de fouille de données seront appliquées. En effet, par fusion de données, terme générique d'un ensemble de techniques, nous entendons *le processus qui consiste à combiner au mieux des informations et des connaissances issues d'un ensemble de données multi-sources, éventuellement hétérogènes, afin d'améliorer la qualité des connaissances extraites de ces données* (cf. figure 1). Selon ce schéma, la fusion de données ainsi définie fait partie intégrante du processus d'extraction automatique des connaissances (Dasarathy 2003).

De nombreuses méthodologies mathématiques ont été proposées comme solution potentielle au problème de fusion de données. Parmi les plus abouties, on peut citer la théorie des probabilités, principalement l'approche bayésienne et les réseaux du même nom, la théorie des crédibilités, la théorie des possibilités et ensembles flous et les méthodes d'agrégations multicritères. Parallèlement à ces cadres mathématiques, d'autres techniques relevant de l'informatique avancée ont été utilisées. C'est le cas par exemple des réseaux neuromimétiques et de la cognition artificielle : intelligence artificielle, systèmes experts, systèmes hybrides, etc.

2 Agrégation et fusion de données multi-sources

Dans le cas de données multi-sources, l'extraction et la fusion de connaissances pour l'aide à la décision peuvent être abordées à trois niveaux complémentaires :

– fusion asymétrique dans laquelle les données des diverses sources ne sont pas disponibles simultanément (sources asynchrones conduisant à une incohérence des échantillonnages temporels). Le plus souvent, une des sources est considérée comme principale et les autres comme des sources d'appoint. Cette situation peut s'apparenter à une fusion temporelle.

– agrégation d'estimateurs où chaque source en présence possède des traitements propres (méthodes, modèles) et délivre sa propre décision/estimation. Il s'agit alors d'un problème d'estimation distribuée et d'agrégation d'estimateurs individuels.

– fusion d'informations qui consiste à fusionner les données et les informations multi-sources. Ce type de fusion permet de gérer une multitude d'informations, complémentaires, redondantes,… issues de sources hétérogènes, afin d'obtenir la meilleure connaissance possible de la situation ou du système considéré.

Ce dernier niveau est certainement le plus intéressant car il tire le meilleur parti de la complémentarité et de la redondance des sources. Cependant, cette synergie n'est possible que si l'on est capable d'évaluer assez finement la connaissance contenue dans chacune des sources de données.

Dans cet article, seuls les deux derniers niveaux seront abordés. En effet, dans la mesure où le premier niveau dépend fortement de l'application étudiée.

2.1 Agrégation d'estimateurs

L'idée de combiner des modèles ou des estimateurs au lieu de la sélection d'un seul, obtenu par optimisation d'un critère, est une pratique bien connue en statistique et a engendrée une abondante littérature depuis l'article précurseur de Bates et Granger en 1969. On trouve des applications d'une telle approche dans divers domaines comme par exemple la prévision météorologique (Fraedrich et Leslie 1988), la prévision de la consommation électrique (Smith 1989) et dans des problèmes de macroéconomie (Clemen et Guerard 1989).

Plus récemment, plusieurs contributions ont montré qu'en présence de plusieurs estimateurs ou de modèles il est presque toujours souhaitable de les combiner au lieu d'en sélectionner un et un seul. On consultera à titre d'exemple Wolpert 1992 ; Breiman 1996 dans le cas d'une seule source de données, et EL Faouzi 1997, 1999, 2000b dans le cas de sources multiples.

Le problème générique se traduit par la présence de ℓ modèles ou estimateurs $\varphi_1, \varphi_2 \ldots, \varphi_\ell$ d'un paramètre ou d'une fonction f. On supposera par ailleurs que chaque modèle ou estimateur k est construit sur la base d'un échantillon $\left\{y^{(k)}, \mathbf{X}^{(k)}\right\}$ où $y^{(k)}$ sont les réalisations de f (la réponse ou la sortie) et $\mathbf{X}^{(k)}$ le vecteur des variables explicatives (vecteur des entrées). On notera par la suite $\hat{f}_k = \varphi_k(\mathbf{X}^{(k)})$.

Le problème que l'on se propose de résoudre consiste donc à construire un opérateur permettant de combiner les différents estimateurs disponibles pour améliorer les performances globales de l'estimation. Le résultat d'une telle opération sera un nouvel estimateur, obtenu par combinaison des estimateurs multiples, dont la qualité - au sens de l'erreur d'estimation - sera améliorée.

D'un point de vue probabiliste, si l'on note $y = \left[y^{(1)}, y^{(2)}, \ldots, y^{(\ell)} \right]'$, le problème est équivalent à l'estimation de $\mathbb{E}\left\langle y \middle| \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(\ell)} \right\rangle$ sur la base de variables aléatoires $\mathbb{E}\left\langle y \middle| \wp \right\rangle$ où $\wp$ est une $\sigma-$algèbre, non vide, générée à partir de l'ensemble $\left\{ \mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(\ell)} \right\}$.

La technique communément utilisée pour l'agrégation des estimateurs est la moyenne pondérée des estimateurs individuels. Les poids sont généralement, soit fixés, soit calculés par optimisation d'un critère, éventuellement sous contraintes. Les opérateurs de ce type se justifient si l'on assimile les différentes sources à une seule source aléatoire dont les informations à agréger en sont des réalisations. Dans ce cas, l'estimateur agrégé s'exprime comme une moyenne pondérée :

$$\pi = \sum_{k=1}^{\ell} w_k \hat{f}_k \qquad [1]$$

Plusieurs stratégies sont potentiellement utilisables pour l'estimation des pondérations. L'une des plus intéressantes exploite les propriétés statistiques des estimateurs individuels, principalement la variance (EL Faouzi 2000a).

Lorsque tous les estimateurs ou les modèles utilisés fournissent des estimations non-biaisées, afin de s'assurer que l'estimateur agrégé donné par [1] soit lui aussi non-biaisé, on est amené à imposer une contrainte de normalisation aux pondérations. Dans ce cas, les pondérations optimales sont solutions du problème suivant :

$$\begin{cases} \min\limits_{w} \left\| y - \sum\limits_{k=1}^{\ell} w_k \hat{f}_k \right\|^2 \\ \sum\limits_{k=1}^{\ell} w_k = 1 \end{cases} \qquad [2]$$

avec $\|x\| = \mathbb{E}\langle x \rangle$.

La résolution de ce problème aboutie au vecteur suivant :

$$w^* = \left(\mathbb{I}_\ell \Omega^{-1} \mathbb{I}_\ell \right)^{-1} \mathbb{I}_\ell' \Omega^{-1} \qquad [3]$$

où $\mathbb{I}_\ell$ est le vecteur unité de dimension ℓ et Ω est la matrice de variance-covariance des résidus (erreurs des estimateurs individuels).

Si les résidus sont normalement distribués, l'estimateur résultant de l'équation [1] avec les pondérations données par [3] est optimal au sens de la variance minimale des erreurs parmi les estimateurs non-biaisés de f.

Le cas particulier intéressant est obtenu lorsque l'on suppose que les estimateurs $\varphi_1, \varphi_2, \ldots, \varphi_\ell$ sont indépendants au sens statistique (absence de corrélation entre estimateurs). Dans ce cas, la matrice de variance-covariance est diagonale $\Omega = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_\ell^2)$ où σ_k^2 est la variance des erreurs du k^{ème} estimateur et l'équation [4] devient :

$$w_k^* = \sigma_k^{-2} \Big/ \sum_{j=1}^{\ell} \sigma_j^{-2} \quad ; \quad k = 1, \ldots, \ell \qquad [4]$$

En conséquence, dans le cas d'estimateurs indépendants, la pondération optimale affectée à chaque estimateur est proportionnelle à la fiabilité de ce dernier, quantifiée ici par l'inverse de la variance des erreurs.

Notons que, dans ce dernier cas contrairement au cas général, toutes les pondérations sont positives ou nulles. En effet, on montre (EL Faouzi 1999) que lorsque les estimateurs sont corrélés positivement, les pondérations sont négatives. Ceci pose la question de l'interprétation de contributions négatives dans l'élaboration de l'estimateur synthétique π.

Afin de s'assurer de la positivité des pondérations, il convient de résoudre le problème [3] sous contraintes de positivité des poids, c'est-à-dire :

$$\begin{cases} \min\limits_{w} \left\| y - \sum\limits_{k=1}^{\ell} w_k \hat{f}_k \right\|^2 \\ \sum\limits_{k=1}^{\ell} w_k = 1 \quad et \quad w_k \geq 0 \end{cases} \qquad [5]$$

Plusieurs algorithmes peuvent être utilisés pour la résolution de ce problème (Lawson et Hanson 1974). L'algorithme retenu pour la recherche de ces solutions repose sur le schéma itératif suivant :

- Initialisation : $k = 1$, $\pi_0^{(1)} = \hat{f}_1$;
- A l'étape *k*, $k = 2, \ldots, \ell$ on injecte l'estimateur *k* dans la somme pondérée :

$$\pi_\mu^{(k)} = \mu_{k-1} \pi_\mu^{(k-1)} + (1 - \mu_{k-1}) \hat{f}_k \qquad [6]$$

Si *n* est le nombre d'observations disponibles, μ_k ; $0 \leq \mu_k \leq 1$ est défini par :

$$\mu_s = \arg\min_{\alpha} \left[\frac{1}{n} \sum_{i=1}^{n} \left(y_i - \pi_\alpha^{(s)} \right)^2 \right] \qquad [7]$$

Au terme de ce processus, on montre aisément que les pondérations obtenues sont positives et normalisées.

REMARQUE. — Sur un plan pratique, notons que la procédure décrite ci-dessus réutilise le même échantillon de référence $\left\{ y^{(k)}, \mathbf{X}^{(k)} \right\}_{k=1}^{\ell}$ qui a servi à la construction des estimateurs individuels $\varphi_1, \varphi_2 \ldots, \varphi_\ell$ pour l'estimation des pondérations. Par conséquent, les valeurs obtenues par [6] et [7] tendent à surestimer les vraies valeurs des poids. Cette limitation est contournable avec des techniques de rééchantillonnage telle la validation croisée (Creven et

Wahba 1979). Dans ce cas, le critère [7] est remplacé par sa version issue de la validation croisée (EL Faouzi 1997).

L'autre difficulté pratique de mise en œuvre de l'approche générale [2] et [3] vient de la nécessité de la matrice de variances-covariances des erreurs Ω, ou tout au moins d'un bon estimateur de celle-ci. Cette dernière matrice est traditionnellement estimée à partir de données historiques en utilisant l'estimateur empirique $\widehat{\Omega}_n$. Opérer de la sorte suppose implicitement les performances des divers estimateurs constantes dans le temps (stationnarité). Lorsqu'une telle hypothèse n'est pas recevable, l'estimation de Ω se doit de tenir compte de cette variabilité temporelle des performances. Une stratégie simple consiste à estimer, à chaque pas de temps, la matrice Ω par l'estimateur empirique pondéré dans lequel les observations récentes ont été favorisées.

2.2 Fusion de données

2.2.1 Fusion de données

Initialement développée dans un cadre militaire, la fusion de données est de plus en plus utilisée dans une multitude d'applications issues de multiples domaines. A titre d'exemple, citons le traitement d'images (Bloch et Maître 2002), la surveillance de systèmes complexes (Waltz et Llinas 1990), et enfin en ingénierie de la circulation routière (EL Faouzi 2003).

Dans un processus de fusion, quatre phases principales sont enchaînées successivement, chacune correspond à un ou plusieurs traitements des données à fusionner (EL Faouzi 2000a).

– *Représentation homogène et recalage des informations pertinentes* : cette étape consiste à rechercher un espace de représentation commun, dans lequel les différentes données à fusionner renseignent sur une même entité. La fusion s'effectuera dans cet espace.

– *Modélisation des connaissances* : cette étape, essentielle au processus de fusion, consiste à modéliser et la connaissance apportée par chaque source au travers de mesures de vraisemblance (probabilistes, crédibilistes,…).

– *Fusion* : c'est l'opération de fusion proprement dite. Les mesures de vraisemblance sont combinées selon une règle de combinaison propre au cadre théorique retenu.

– *Décision par choix d'une stratégie* : la fusion doit permettre de choisir l'état le plus crédible, au sens d'un certain critère, parmi toutes les alternatives possibles. En ce sens, la fusion de données aboutit bien souvent à une classification. Le critère de décision dépend du cadre théorique et de l'objectif visé.

Le formalisme théorique exploré dans le cadre de ce travail est celui de la théorie des crédibilités, dite aussi théorie de l'évidence ou théorie des croyances. Elle a été introduite par Dempster en 1967 puis étendue par Shafer en 1976 et considérée comme une extension de la théorie des probabilités aux situations d'ignorance (Dempster 1968).

2.2.2 Théorie des crédibilités

Dans cette théorie, on représente les caractéristiques ou les états du système étudié sous forme d'un ensemble fini de propositions ou d'hypothèses *mutuellement exclusives*, appelé *cadre de discernement* $\mathfrak{X}=\{u_1,u_2,\ldots,u_\tau\}$, $\tau \geq 1$. On suppose que $\mathfrak{X}$ est exhaustif de telle sorte que la vérité se trouve obligatoirement dans $\mathfrak{X}$ (principe du monde fermé).

On définit sur l'ensemble des parties de $\mathfrak{X}$, noté $\mathfrak{P}(\mathfrak{X})$, une mesure de croyance notée m et appelée *masse de croyance.* Cette mesure associée à chaque $A \in \mathfrak{P}(\mathfrak{X})$, $m(A) \in [0;1]$ avec $m(\varnothing) \neq 0$ et $\sum m(A)=1$, représente la croyance associée à la réalisation de A sans pour autant être capable de partager sa valeur sur les éléments de $\mathfrak{X}$ formant A. Les éléments $A \in \mathfrak{P}(\mathfrak{X})$ dont $m(A) \neq 0$ sont appelés *éléments focaux*.

La principale caractéristique qui différencie cette théorie du cadre probabiliste (bayésien) classique est sa capacité d'affecter une masse (probabilité) non seulement à chaque hypothèse simple (élément de $\mathfrak{X}$), mais également à des unions d'hypothèses (hypothèses composées), sans faire appel aux axiomes de probabilités.

Deux autres mesures sont associées à chaque $A \in \mathfrak{P}(\mathfrak{X})^*$: la *crédibilité* et la *plausibilité*. Elles sont définies respectivement par :

$$\mathrm{Cr}(A)=\sum_{B:B\subset A} m(B) \qquad [8]$$

$$\mathrm{Pl}(A)=\sum_{B:B\cap A\neq\varnothing} m(B)=1-\mathrm{Cr}(\bar{A}) \qquad [9]$$

$\bar{A}$ désigne le complémentaire de A dans $\mathfrak{X}$.

Lorsque tous les éléments focaux sont des singletons (éléments de $\mathfrak{X}$), ces deux fonctions coïncident et sont égales à la masse, *i.e.* $\mathrm{Cr}(A)=\mathrm{Pl}(A)=m(A)$. On retrouve ainsi le cadre bayésien classique. Autrement, on a les inégalités suivantes :

$$\mathrm{Cr}(A) \leq m(A) \leq \mathrm{Pl}(A) \qquad [10]$$

Dans la situation d'informations multi-sources, chaque source k ; $k=1,\ldots,\ell$ fournit un vecteur de masse m_k ; $k=1,\ldots,\ell$ qu'elle associe aux éléments $A \in \mathfrak{P}(\mathfrak{X})^*$. La fusion de données via la théorie des croyances revient à combiner ces vecteurs de masse m_k ; $k=1,\ldots,\ell$ en utilisant la règle dite *orthogonale* de Dempster :

$$m(A)=\left(m_1 \oplus m_2 \oplus \cdots \oplus m_\ell\right)(A) \propto \sum_{A_1\cap A_2\cap\ldots\cap A_\ell=A} m_1(A_1)m_2(A_2)\cdots m_\ell(A_\ell) \qquad [11]$$

Si les éléments focaux $A_1, A_2, \ldots, A_\ell$ ont des intersections non vides, les deux expressions de part et d'autre du signe de proportionnalité $\propto$ dans [11] sont égales. Dans le cas contraire, *i.e.* $A_1, A_2, \ldots, A_\ell$ sont mutuellement disjoints, il existe des conflits ou des contradictions entre sources, car $m(\varnothing) \neq 0$. La masse affectée à l'ensemble vide constitue une mesure de conflits entre sources et fournit une indication sur l'exhaustivité du cadre de discernement $\mathfrak{X}$. Cette quantité est donnée par :

$$\kappa=\sum_{A_1\cap A_2\cap\ldots\cap A_\ell=\varnothing} m_1(A_1)m_2(A_2)\cdots m_\ell(A_\ell) \qquad [12]$$

Pour préserver le principe du monde fermé, la masse associée à l'ensemble vide est alors répartie sur les éléments focaux non vides via la normalisation de la règle de Dempster [11] :

$$m(A)=\left(m_1\oplus m_2\oplus\cdots\oplus m_\ell\right)(A)=\frac{1}{(1-\kappa)}\sum_{A_1\cap A_2\cap\ldots\cap A_\ell=A} m_1(A_1)m_2(A_2)\cdots m_\ell(A_\ell) \qquad [13]$$

Une fois les masses combinées, la décision finale, i.e. l'alternative à retenir, est celle qui maximise la crédibilité [8], la plausibilité [9] ou la probabilité pignistique [14] (Smets 1989). Ce dernier critère consiste à répartir la masse placée sur chaque proposition composée sur les propositions élémentaires qui la composent.

$$\mathrm{P_G}\left(A\right)=\sum_{A\subseteq B;B\in\Re(\mathfrak{X})}\frac{m(B)}{\mathrm{Card}(A)} \qquad [14]$$

REMARQUE. — Dans les applications pratiques, la valeur de $\kappa\in[0,1]$ définie par [12] permet d'évaluer la qualité de la combinaison. Ainsi, lorsque cette valeur est trop grande, l'utilisation de la règle de combinaison normalisée [13] est sujette à caution. De par le fort conflit entre les sources, on peut décider de ne pas combiner les informations des sources disponibles, sauf à disposer d'indications sur la fiabilité de ces dernières.

3 Mise en œuvre opérationnelle

Les techniques d'agrégation et de fusion que nous venons d'exposer sont appliquées au problème d'estimation du temps de parcours sur un axe routier. Dans cette étude, les données de trois sources ont été recueillies au cours d'une campagne de mesures qui s'est déroulée pendant huit jours ouvrables à raison de trois heures par jour :

– Source 1 : Données macroscopiques sur le trafic, fournies par des capteurs de trafic traditionnels (une douzaine), agrégées sur une période de six minutes (240 observations). Préalablement à toute fusion ou agrégation, les données de cette source ont été converties en temps de parcours via un algorithme de conversion (Bonvalet et Robin-Prevallée 1987).

– Source 2 : Un échantillon de 160 observations, agrégées sur six minutes, de temps de parcours réalisés par des véhicules équipés de capteurs embarqués (véhicules traceurs).

– Source 3 : Les temps de parcours recueillis sur le principe d'enquête minéralogique entre les entrées-sorties de l'axe étudié (230 observations agrégées sur six minutes).

Les deux premières sources, que l'on nommera «capteurs» et «traceurs» respectivement, ont été utilisées pour fournir des informations sur le temps de parcours à fusionner. Les temps de parcours fournis par la dernière source servent de référence à l'évaluation.

3.1 Agrégation d'estimateurs

Dans le cadre, chacune des deux premières sources fournit un estimateur de temps de parcours, notés TP1 et TP2, qui sont ensuite agrégés via les deux schémas suivants :

– schéma d'agrégation de type variance minimum en distinguant le cas reposant sur l'hypothèse d'indépendance des estimateurs (SA.1) et le cas général (SA.2)

– schéma d'agrégation avec contraintes de positivité des pondérations (SA.3).

L'évaluation des performances des deux estimateurs repose sur une procédure apparentée à la validation croisée. Cette procédure consiste à estimer l'erreur d'estimation (erreur moyenne quadratique RMSE) avec un échantillon test.

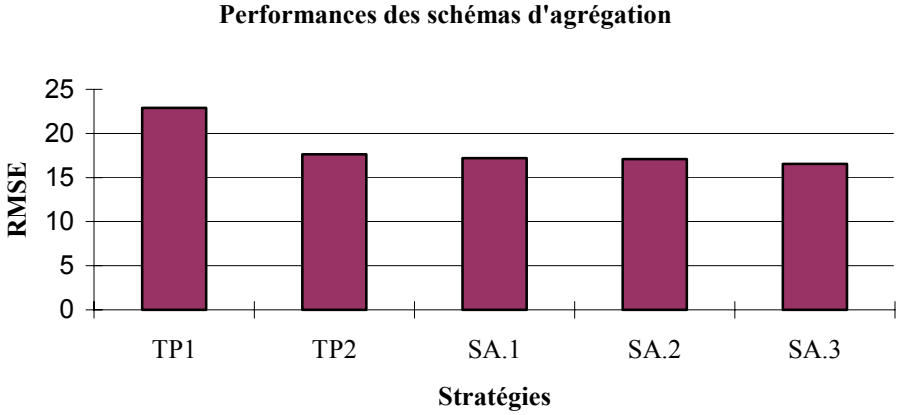

FIG. 2 - *Performances des schémas d'agrégation d'estimateurs.*

Ces résultats montrent la propension des estimateurs résultant de la fusion à l'amélioration de la qualité globale de l'estimation. En effet, globalement, on obtient de meilleures performances qu'avec les estimateurs individuels. L'amélioration est plus nettement marquée avec la stratégie SA.3.

Il est à relever que les résultats donnés par les stratégies (SA.1) et (SA.2), bien que comparables, laisse entrevoir un léger avantage à l'agrégation supposant l'indépendance des deux estimateurs. Deux explications sont envisageables. La première est la légitimité de l'hypothèse d'indépendance. Cette dernière peut être acceptée, en première approximation, lorsque les estimateurs proviennent de sources multiples. La seconde est la grande sensibilité des performances de la stratégie (SA.2) à la qualité d'estimation de la matrice de variances-covariances des erreurs Ω ; estimation qui requiert des données en grand nombre.

3.2 Fusion crédibiliste

L'estimation du temps de parcours sous forme de problème de fusion pouvant être traité par la théorie des croyances impose une discrétisation des temps de parcours. Pour cela, nous avons procédé à un découpage préalable en classes de valeurs de temps de parcours. La définition du cadre de discernant est obtenu par découpage en K classes $\mathfrak{T} = \{\mathfrak{h}_1, \mathfrak{h}_2, \ldots, \mathfrak{h}_k\}$ des temps de parcours de référence.

Afin de tester les performances de la fusion selon le nombre de classes K, nous l'avons fait varier en considérant $K \in \{4,6,8,10\}$, i.e. en augmentant la précision requise pour l'estimation des temps de parcours.

La modélisation des connaissances, via les jeux de masses de croyance, est fondée sur l'apprentissage statistique. Plus précisément, elle repose sur l'exploitation des matrices de confusion associées aux deux sources disponibles. Ces matrices sont obtenues par apprentissage statistique et le critère de performance retenu est le taux de bien classés (TBC).

Lorsque l'on affecte des masses non nulles aux seules classes simples, ce qui revient à dire que l'on modélise l'incertitude mais non l'imprécision, la seule stratégie de décision consiste à choisir la classe de masse maximale. Les performances de la fusion sont synthétisées par le graphique suivant :

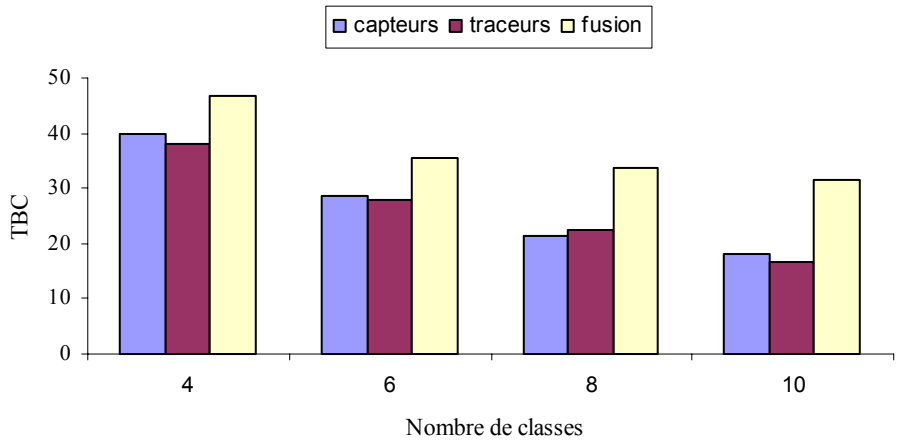

FIG. 3- *Performances de la fusion dans le cas de classes simples.*

On constate clairement que la méthode de fusion proposée supplante, dans chacun des cas, les méthodes d'estimation basées sur une seule source. L'amélioration de la qualité de l'estimation, en terme de taux de bien classés, varie de 6 % à 13,6 % et celle-ci est d'autant plus importante quand le nombre de classes augmente.

Afin de prendre en compte l'imprécision dans le processus de décision, les masses non nulles doivent être affectées, non seulement aux seuls éléments de $\mathfrak{X}$, mais aussi à toute jonction de ces derniers. L'imprécision dont il est question peut provenir soit d'un défaut d'information, soit éviter le choix arbitraire entre deux classes de masses très proches. Ainsi, comme les classes $\{\mathfrak{h}_1,\mathfrak{h}_2,\ldots,\mathfrak{h}_k\}$ sont ordonnées, les seules unions que l'on s'autorise sont celles formées entre classes contiguës.

Connaissant les masses des classes simples, la masse associée à une classe composée (union de deux classes simples) est calculée par :

$$m(\mathfrak{h}_i \cup \mathfrak{h}_j) = \begin{cases} \max\left(m(\mathfrak{h}_i), m(\mathfrak{h}_j)\right) + \frac{1}{2}\left|m(\mathfrak{h}_i) - m(\mathfrak{h}_j)\right| & \text{si} \quad \left|m(\mathfrak{h}_i) - m(\mathfrak{h}_j)\right| < \delta \\ 0 \quad \text{sinon} \end{cases} \qquad [15]$$

Ceci revient à dire que l'on affecte une masse à une classe composée lorsque les deux classes la composant possèdent des masses très similaires (à δ près). Ce qui sous-entend que, dans cette situation, on n'a aucune raison de choisir l'une ou l'autre du fait de la similitude des croyances que l'on a sur la véracité de chacune d'elles.

Nous donnons, ci-dessous, les résultats de la fusion uniquement dans le cas de six classes avec deux valeurs possibles du paramètre δ=0,1 et 0,2 et pour les trois stratégies de décision : maximum de crédibilité, de plausibilité et de probabilité pignistique (cf. §. 2.2.2.).

Rappelons que les deux premières stratégies peuvent retenir des classes composées alors que la règle pignistique fournie, quant à elle, des classes simples. Dans les résultats ci-dessous, nous considérons que le résultat est correct dès lors que l'intersection de la classe obtenue par fusion avec la classe de référence est non vide.

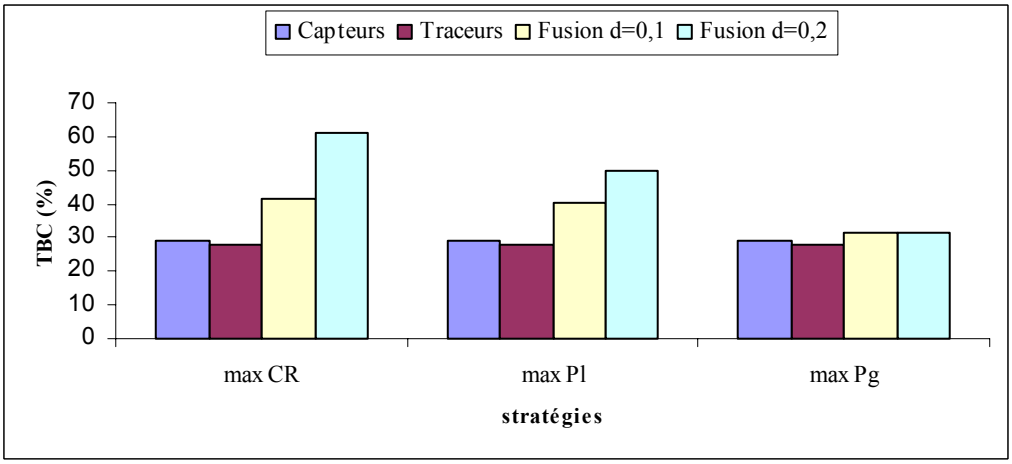

FIG. 4 - *Performances de la fusion dans le cas de six classes contiguës*.

On remarque une fois encore que la fusion permet d'obtenir de meilleurs taux de bien classés. En comparaison avec le cas de classes simples, les taux de bien classés sont meilleurs. De plus, les taux les meilleurs sont obtenus avec la crédibilité et la plausibilité.

Il est à noter que la contrepartie de ces bons résultats c'est une perte de précision. En effet, retenir une classe composée revient à augmenter l'amplitude des classes finales ce qui revient à privilégier la certitude par rapport à la précision. C'est ainsi que les résultats de la figure 4 sont obtenus en retenant le plus souvent des hypothèses composées : à 68 % pour la crédibilité et à 88 % pour la plausibilité. Ceci souligne que l'imprécision modélisée dans le cadre de la théorie des crédibilités se propage jusqu'à l'étape de décision.

Notons par ailleurs les résultats décevants obtenus avec la règle pignistique. Bien qu'elle ne retient que des clases simples dans sa phase de décision, les taux de bien classés obtenus sont inférieurs à ceux obtenus par la fusion sur les classes simples. Ce fait peut trouver son explication dans le principe même de cette règle qui procède à une réallocation uniforme des masses associées aux classes composées au bénéfice de classes simples les composant. Ceci induit un accroissement artificiel des masses associées aux classes simples aboutissant à des classifications erronées.

4 Conclusions

Dans ce travail, nous avons abordé successivement le problème complexe de l'agrégation d'estimateurs et de données issus de sources multiples. Bien que cette étude ait porté sur une classe très restreinte de schéma d'agrégation d'estimateurs (schémas linéaires), elle a permis de donner quelques éléments de réponse au problème d'extraction de connaissances dans un contexte de données multiformes, issues de sources hétérogènes. Ces résultats indiquent la propension de ces schémas à améliorer la qualité de l'estimation. Intuitivement, il est facile de voir que l'agrégation d'estimateurs peut améliorer la qualité de l'estimation lorsque les estimateurs individuels sont uniformément distribués autour de la vraie valeur. Hormis cette situation, il est bien difficile d'avoir une réponse définitive quant à l'apport de l'agrégation. De plus, cette approche n'exploite pas les structures locales des estimateurs. En effet, certains estimateurs, bien que possédant des erreurs d'estimation importante, peuvent fournir de faibles erreurs dans certaines régions de l'espace. Dans ce cas, il convient alors d'adopter des schémas projectifs au lieu de schémas agrégatifs que nous avons présenté ici.

Les techniques de fusion de données forment une alternative à la fusion d'estimateurs et plus particulièrement la théorie des crédibilités. L'utilisation de cette théorie exige une bonne connaissance du problème traité, car elle ne propose aucune méthodologie, ni pour constituer les hypothèses, ni pour modéliser les jeux de masses. Or ces deux étapes s'avèrent déterminantes pour le bon déroulement du processus de fusion. Ceci vient de son caractère très souple qui, paradoxalement, constitue en même temps un avantage non négligeable : tout problème de gestion d'informations issues de sources hétérogènes peut être abordé par cette théorie. Or, ce type de problème est aujourd'hui rencontré dans de nombreux domaines, et il est sans doute intéressant d'élaborer un cadre méthodologique général de modélisation des jeux de masses. Le développement de ce cadre ainsi que l'élaboration de schémas projectifs pour la combinaison d'estimateurs (estimation distribuée) font l'objet de nos travaux actuels.

Références

Bates M. et Granger W. J., (1969), The combination of forecasts, Operational Research Quarterly, vol. 20, 1969, pp 451-468.

Bloch I. et Maître H. (2002) Fusion d'informations en traitement d'images : spécificités, modélisation et combinaison, Technique de l'ingénieur, Vol. 5, n° 230, pp 1-26.

Bonvalet F. et Robin-Prevallée Y., (1987), Mise au point d'un indicateur permanent des conditions de circulation en Île-de-France, T.E.C., n° 84-85, Septembre-décembre 1987.

Breiman L. (1996), Stacked regressions, Machine Learning, vol. 24, n° 3, pp 49-64.
Clemen T. et Guerard J. (1989), Econometric GNP forecasts: Incremental Information Relative to Naïve Extrapolation, International Journal of Forecasting, vol. 5, pp 417-426.
Caven P. et Wahba G. (1979), Smoothing noisy data with spline function, Numerische Mathmatik, vol. 31, pp 277-403
Dasarathy B. V. (2003), Information fusion, data mining, and knowledge discovery: Editorial, Information fusion, 4(1), pp 1.
Dempster A. P. (1967), Upper and lower probabilities induced by multivalued mapping, Annals of Mathematical Statistics, vol. 38, pp 325-339.
Dempster A. P. (1968) A generalization of Bayesian inference, J.R.S.S., 30(B), pp 205-247.
EL Faouzi N.-E. (1997), Fusion linéaire d'estimateurs multiples, rapport technique n° 265, septembre 1997, LICIT, INRETS-ENTPE.
EL Faouzi N.-E. (1999), Combining Predictive Schemes in Short-Term Traffic Forecasting, Proceedings of the 14th International Symposium on Transportation and Traffic theory (ISTTT), Jerusalem, 26-28 July 1999, A. Ceder (Ed.), Pergamon, Elsevier, pp 471-487.
EL Faouzi N.-E. (2000a), Fusion de données : Concepts et méthodes, rapport technique n° 265, septembre 2000, LICIT, INRETS-ENTPE.
EL Faouzi N.-E. (2000b), Fusion de données pour l'estimation des temps de parcours via la théorie de l'évidence, Revue Transport-Sécurité, n° 68, 2000, pp 15-30.
EL Faouzi N.-E. Eds. (2003), Recueil multiforme et fusion de données en circulation routière, Collection Actes-INRETS (A paraître).
Fraedrich K. et Leslie M. (1988), Real-times short-term forecasting of precipitation at an Australian tropical station , Weather and Forecasting, vol. 3, 1988, pp 104-114.
Han J. et Kamber M. (2001), Data Mining Concepts and Techniques., Academic Press.
Klosgen W. et Zytkow J. M. Eds. (2002), Handbook of Data mining and Knowledge Discovery. Oxford University Press.
Lawson, C. et Hanson, R. (1974). Solving Least-Squares Problems. Prentice-Hall, 1974.
Saporta G. (2002), Data fusion and data grafting, Computational Statistics and Data Analysis, vol. 38, pp 465-473.
Shafer G. (1976), Mathematical theory of Evidence, Princeton University Press, 1976.
Smets Ph. (1989), Constructing the pignistic probability function in a context of uncertainty, Proceedings of the fifth Workshop on Uncertainty in AI, Windsor, Canada, pp 319-326.
Smith D. (1989), Combination of forecasts in electricity demand prediction, Int. Journal of Forecasting, vol. 8, pp 349-356.
Waltz E. et Llinas J. (1990), Multisensor Data fusion, Artech House, Boston, 1990.
Wolpert D. H. (1992), Stacked Generalization. Neural Networks, n° 5, pp 241-259.

Summary

The data fusion permits to enrich a set of fragmentary data, often collected via multiple devices, by combining the information that they contain in order to achieve improvement of the extracted knowledge quality. So, knowledge Discovery process (KDD) is clearly one of the many purposes of a multisource data and information fusion.
In this paper, several schemes of aggregation and data fusion are proposed, exploiting complementarities and the redundancies of mutisources data and information. These schemes are then illustrated on the traffic application, namely travel time estimation in a given path of a road network based on conventional traffic data and probe vehicle reports.