Théorie des Fonctions de Croyance et Classification

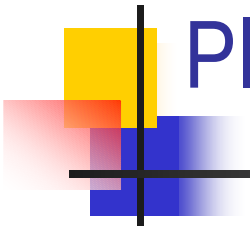
Thierry Denœux

Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)
http://www.hds.utc.fr/~tdenoeux

utc Université de Technologie Compiègne

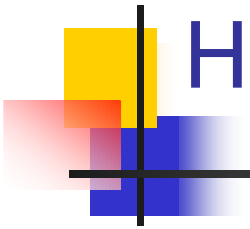
HEUDIASYC

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE



Plan

1. Théorie des Fonctions de Croyance (Dempster-Shafer, « théorie de l'évidence »)
   - Historique
   - Modèle des croyances transférables
     - partie statique (fonctions de croyance)
     - partie dynamique (combinaison, conditionnement, ...)
     - partie décisionnelle.
2. Applications en classification
   - Motivations
   - Construction de classifieurs crédibilistes
     - Théorème de Bayes généralisés
     - Classification à base de cas
     - Arbres de décision crédibilistes
   - Fusion de classifieurs
3. Conclusions et perspectives



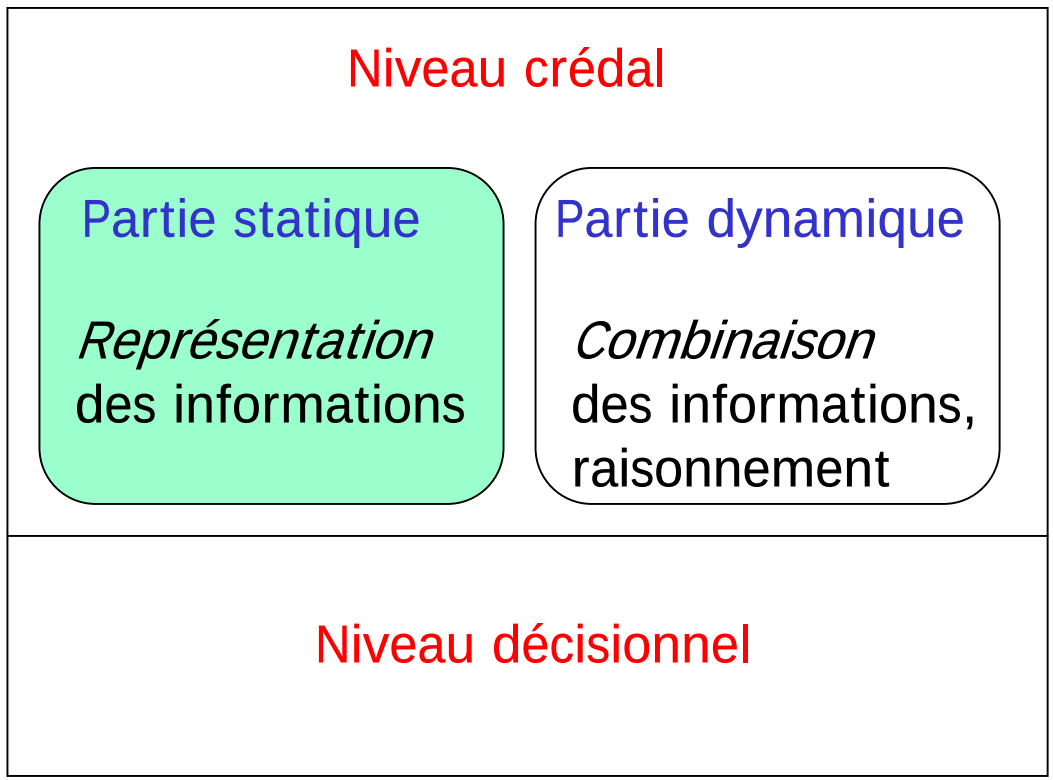
Historique

- Origines :
  - Dempster (1966-1968) : théorie de l'inférence statistique généralisant l'inférence Bayésienne (pas d'a priori sur les paramètres)
  - Shafer (1976) : proposition des fonctions de croyance comme cadre général de représentation des incertitudes, englobant la théorie des probabilités comme cas particulier.
- Applications :
  - Années 80 : IA, modélisation des incertitudes dans les systèmes experts.
  - Années 90 : fusion d'informations (télédétection, identification de cibles, imagerie médicale, ...).

Contexte actuel

- Fonctions de croyance utilisées dans deux modèles différents :
  - modèle de Dempster, théorie des Hints (Kohlas, Monney, Univ. de Fribourg) : fait appel à une mesure de probabilité sur un espace « sous-jacent » ;
  - modèle des croyances transférables (MCT), développé par P. Smets (ULB) depuis 1978 : interprétation subjectiviste, non probabiliste.
- Autres théories :
  - Théorie bayésienne des probabilités ($\subset$ MCT);
  - Possibilités (Zadeh, Dubois et Prade) ;
  - Probabilités imprécises (Walley, de Cooman, ...)

Le Modèle des Croyances Transférables



Fonction de Masse de Croyance

- Ω : ensemble fini (extension possible au cas infini)

$$m : 2^{\Omega} \rightarrow [0,1] \quad \text{t.q.} \quad \sum_{A \subseteq \Omega} m(A) = 1$$

- Interprétation :
  - traduit un état de connaissance partielle sur la valeur d'une variable ω à valeurs dans Ω.
  - m(A) = « part » de croyance allouée (par une source S) à l'hypothèse « $\omega \in$ A » et à aucune hypothèse plus restrictive, étant donnée une base de connaissances BC.
- Notation complète :

$$m_S^{\Omega}\{\omega\}[BC]$$

- A $\subseteq \Omega$ t.q. m(A) $\neq$ 0 : élément focal de m

Fonctions de masse particulières

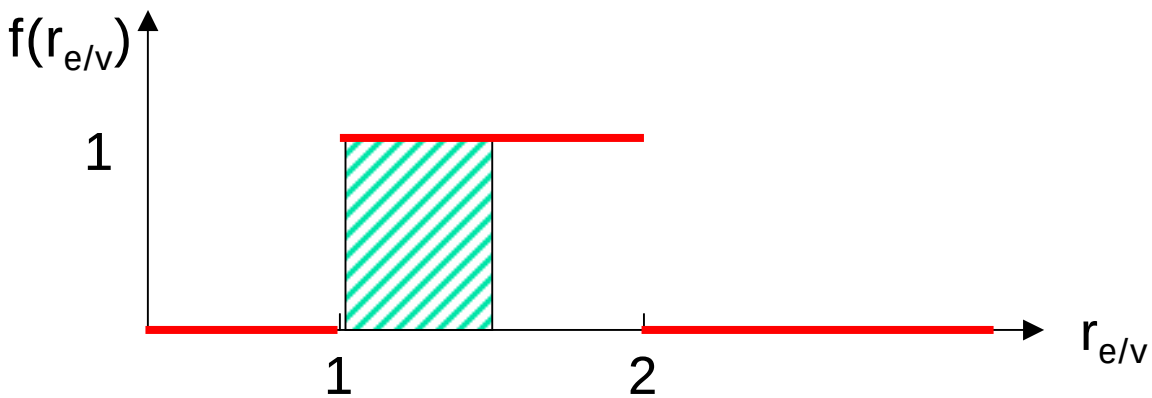
- Ignorance totale : $m(\Omega)=1$ (fonction de masse vide);
- Certitude : $m(\{\omega\})=1$ pour un $\omega \in \Omega$ (fonction de masse certaine) ;
- Connaissance imprécise et certaine : $m(A)=1$ pour un $A \subseteq \Omega$, $|A|>1$ (fonction de masse catégorique) ;
- Incertitude probabiliste : $m(A)=0$ pour tout $A \subseteq \Omega$ t.q. $|A| > 1$ (fonction de masse Bayésienne) ;
- Incertitude « possibiliste » : éléments focaux $A_1 \subseteq A_2 \subseteq ... \subseteq A_n$ (fonction de masse consonante).
- etc...

L'ignorance en Probabilités : principe d'indifférence

- Principe d'indifférence ou « de raison insuffisante » (Bernouilli) : l'ignorance totale est modélisée par une distribution de probabilité uniforme sur Ω (maximum d'entropie).
- Paradoxe de Bertrand :
  - Soit une bouteille contenant un mélange d'eau et de vin. La bouteille contient
    - au moins autant d'eau que de vin,
    - au plus deux fois plus d'eau que de vin.
  - Probabilité que la bouteille contienne au plus 1.5 plus d'eau que de vin ?

Paradoxe de Bertrand (1)

- Soit $r_{e/v} \in \mathbb{R}_+$ le rapport eau/vin : $1 \leq r_{e/v} \leq 2$.
- PRI : loi de probabilité uniforme sur [1,2]



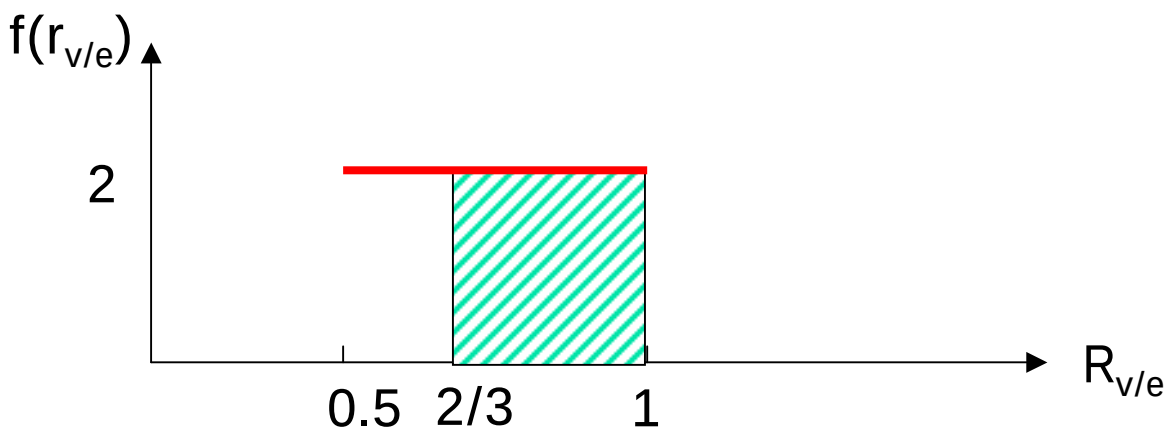
- $P(r_{e/v} \leq 1.5)=0.5$
- Probabilité que la bouteille contienne au plus 1.5 plus d'eau que de vin = 0.5

Paradoxe de Bertrand (2)

- Soit $r_{v/e} \in \mathbb{R}_+$ le rapport vin/eau :

$$1 \leq r_{e/v} \leq 2 \Leftrightarrow 0.5 \leq r_{v/e} \leq 1$$

- PRI : loi de probabilité uniforme sur [0.5,1]



- $r_{e/v} \leq 1.5 \Leftrightarrow r_{v/e} \geq 2/3$. Or, $P(r_{v/e} \geq 2/3)=2/3$
- Probabilité que la bouteille contienne au plus 1.5 plus d'eau que de vin = 2/3 !

Fonction de Croyance

$$bel : A \rightarrow bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B), \quad \forall A \subseteq \Omega$$

- bel(A) = degré de croyance en l'hypothèse « $\omega \in$ A », compte tenu des masses de croyances affectées à toutes les hypothèses qui impliquent A. (bel(A) d'autant plus grand que l'ensemble des informations disponibles accréditent, directement ou non, l'hypothèse A).
- Propriété : capacité complètement monotone

$$bel\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_i bel(A_i) - \sum_{i>j} bel(A_i \cap A_j) \ldots$$
$$- (-1)^n bel\left(\bigcap_{i=1}^{n} A_i\right), \quad \forall A_1, \ldots, A_n \subseteq \Omega$$

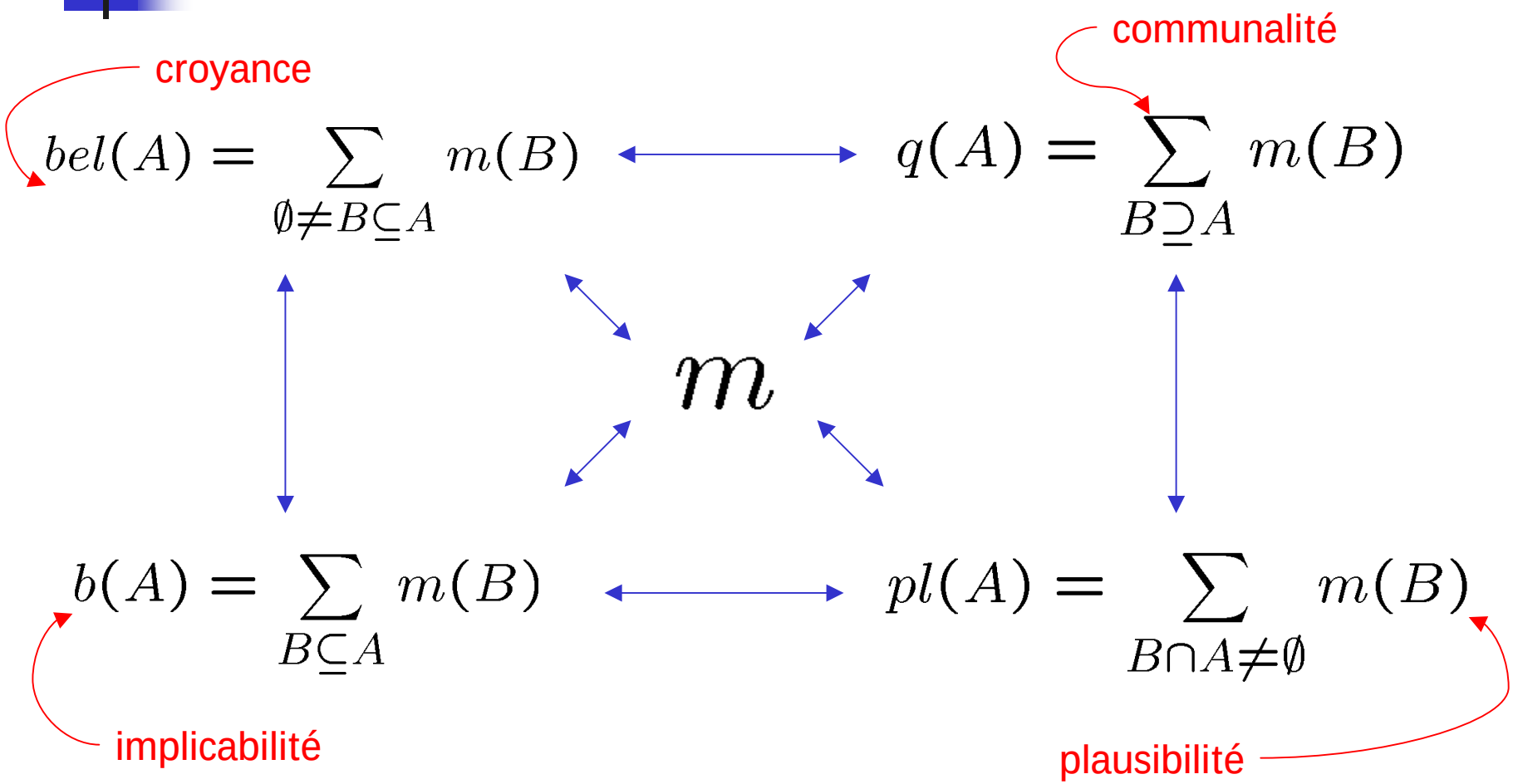
Fonction de plausibilité

$$pl : A \rightarrow pl(A) = \sum_{B \cap A \neq \emptyset} m(B)$$
$$= 1 - m(\emptyset) - bel(\overline{A}), \quad \forall A \subseteq \Omega$$

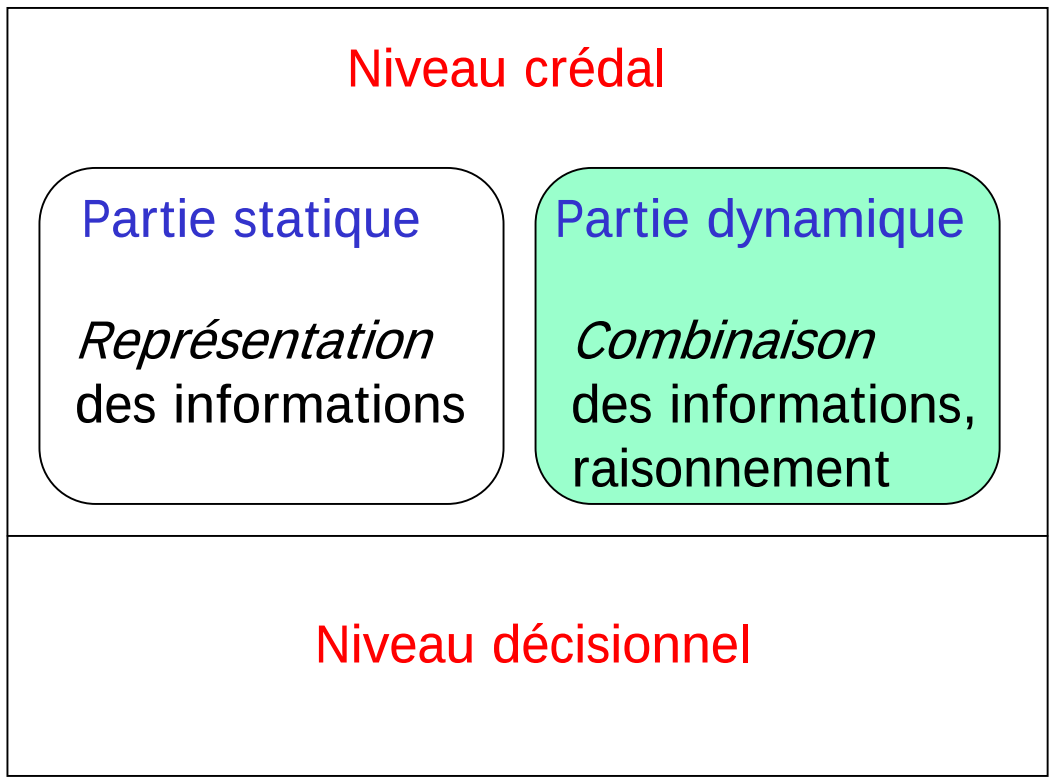
- pl(A) = degré maximal de croyance pouvant potentiellement être attribué à l'hypothèse « $\omega \in$ A » (conditionnellement à l'obtention de nouvelle informations).
- Propriétés:
  - bel(A) $\leq$ pl(A), $\forall$ A $\subseteq \Omega$
  - mesure sous-additive

$$pl\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_i pl(A_i) - \sum_{i>j} pl(A_i \cap A_j) \ldots$$
$$- (-1)^n pl\left(\bigcap_{i=1}^{n} A_i\right), \quad \forall A_1, \ldots, A_n \subseteq \Omega$$

Représentations équivalentes



Le Modèle des Croyances Transférables



Somme conjonctive

- Soient deux fonctions de masse m_1 et m_2 issues de deux sources d'informations distinctes.
- Somme conjonctive :

$$(m_1 \textcircled{\cap} m_2)(C) = \sum_{A \cap B = C} m_1(A) m_2(B), \quad \forall C \subseteq \Omega$$

$$q_{1\textcircled{\cap}2} = q_1 \cdot q_2$$

- Propriétés :
  - commutative,
  - associative,
  - élément neutre $m(\Omega)=1$,
  - non-idempotente.

Interprétation de m(∅)

- Degré de conflit :

$$(m_1 \textcircled{\cap} m_2)(\emptyset) = \sum_{A \cap B = \emptyset} m_1(A) m_2(B)$$

- Interprétation possible : m(∅) est la masse de croyance allouée à l'hypothèse $\omega \notin \Omega$ (hypothèse du monde ouvert).
- Si Ω exhaustif (hypothèse du monde clos) : on impose m(∅)=0.
- Règle de Dempster = somme conjonctive puis normalisation.

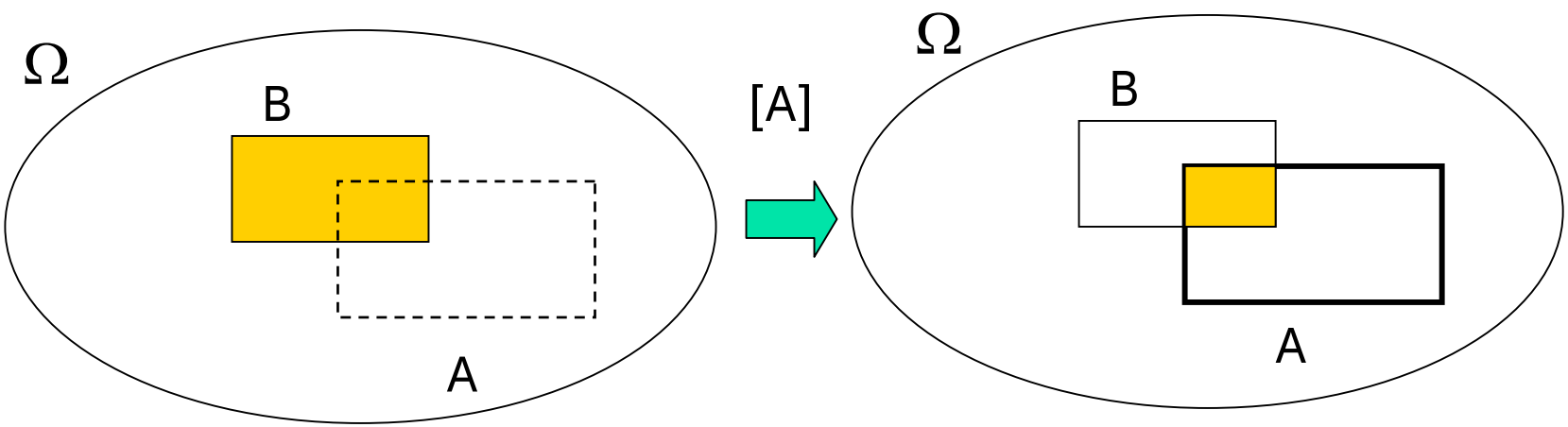
$$(m_1 \oplus m_2)(A) = \frac{(m_1 \textcircled{\cap} m_2)(A)}{1 - (m_1 \textcircled{\cap} m_2)(\emptyset)} \quad \forall A \neq \emptyset$$

Règle de conditionnement de Dempster

- Cas particulier de la somme conjonctive.

$$m[A] = m \textcircled{\cap} m_A \text{ avec } m_A(A) = 1, A \subseteq \Omega$$

- Chaque masse m(B) est transférée à B ∩ A (d'où l'appellation « Modèle des croyances Transférables »)
- Correspond au conditionnement probabiliste après renormalisation.



Autres opérateurs de combinaison

- Somme disjonctive

$$(m_1 \textcircled{\cup} m_2)(C) = \sum_{A \cup B = C} m_1(A) m_2(B)$$

$$b_{1\textcircled{\cup}2} = b_1 \cdot b_2$$

  - Propriétés : commutative, associative, élément neutre m(∅)=1.
  - Interprétation : l'une au moins des sources est fiable (opérateur plus « prudent »)

- Moyenne :

$$m = \frac{1}{n} \sum_{i=1}^{n} m_i$$

Notion d'information

- Comment définir le « contenu informationnel », ou le « degré d'incertitude » d'une fonctions de croyance ?
- Approche ordinale : définition d'un ordre partiel sur l'ensemble des fonctions de croyance.
  - Soient bel_1 et bel_2 deux fonctions de croyance normalisées ($m_1(\emptyset)=m_2(\emptyset)=0$).
  - bel_1 est moins informative que bel_2 ssi

$$bel_1(A) \leq bel_2(A), \forall A \subseteq \Omega \Leftrightarrow$$
$$pl_1(A) \geq pl_2(A), \forall A \subseteq \Omega$$

  - élément minimum : fonction de croyance vide.
- Approche quantitative : définition de « mesures d'incertitude », par exemple la nonspécificité :

$$N(m) = \sum_{\emptyset \neq A \subseteq \Omega} m(A) \log_2(|A|)$$

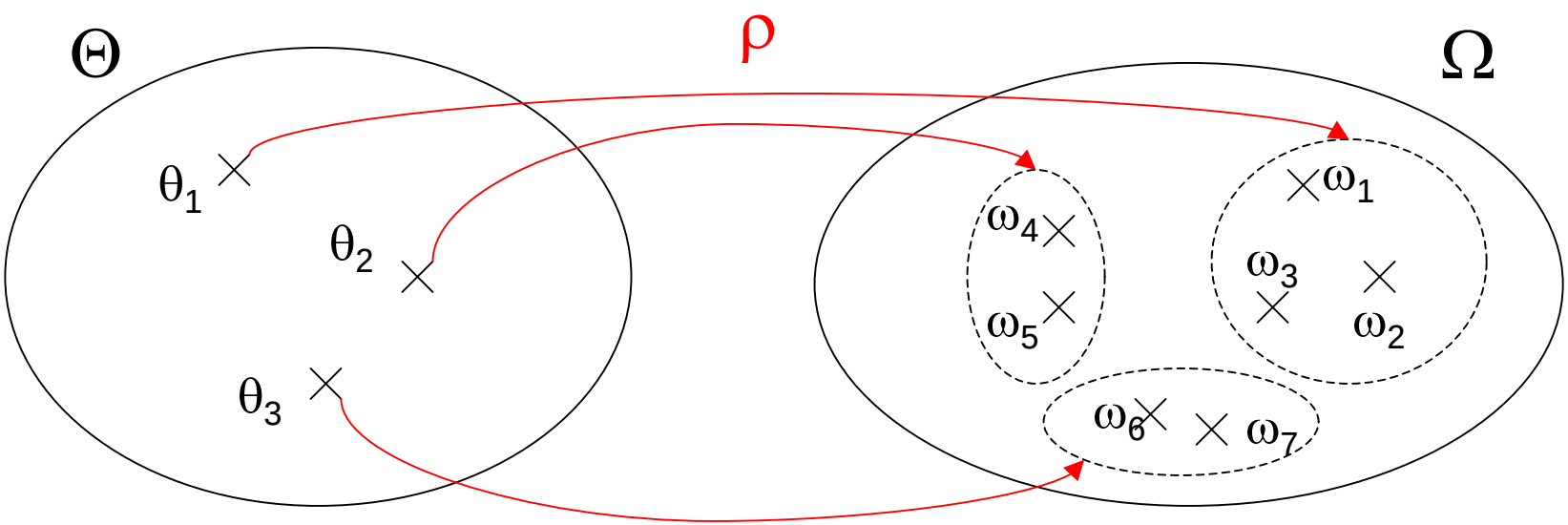
Principe d'information minimum

- Principe d'information minimum :

> Choisir la fonction de croyance
> la moins informative (lorsqu'elle existe)
> parmi l'ensemble des fonctions de croyance
> compatibles avec les informations disponibles.

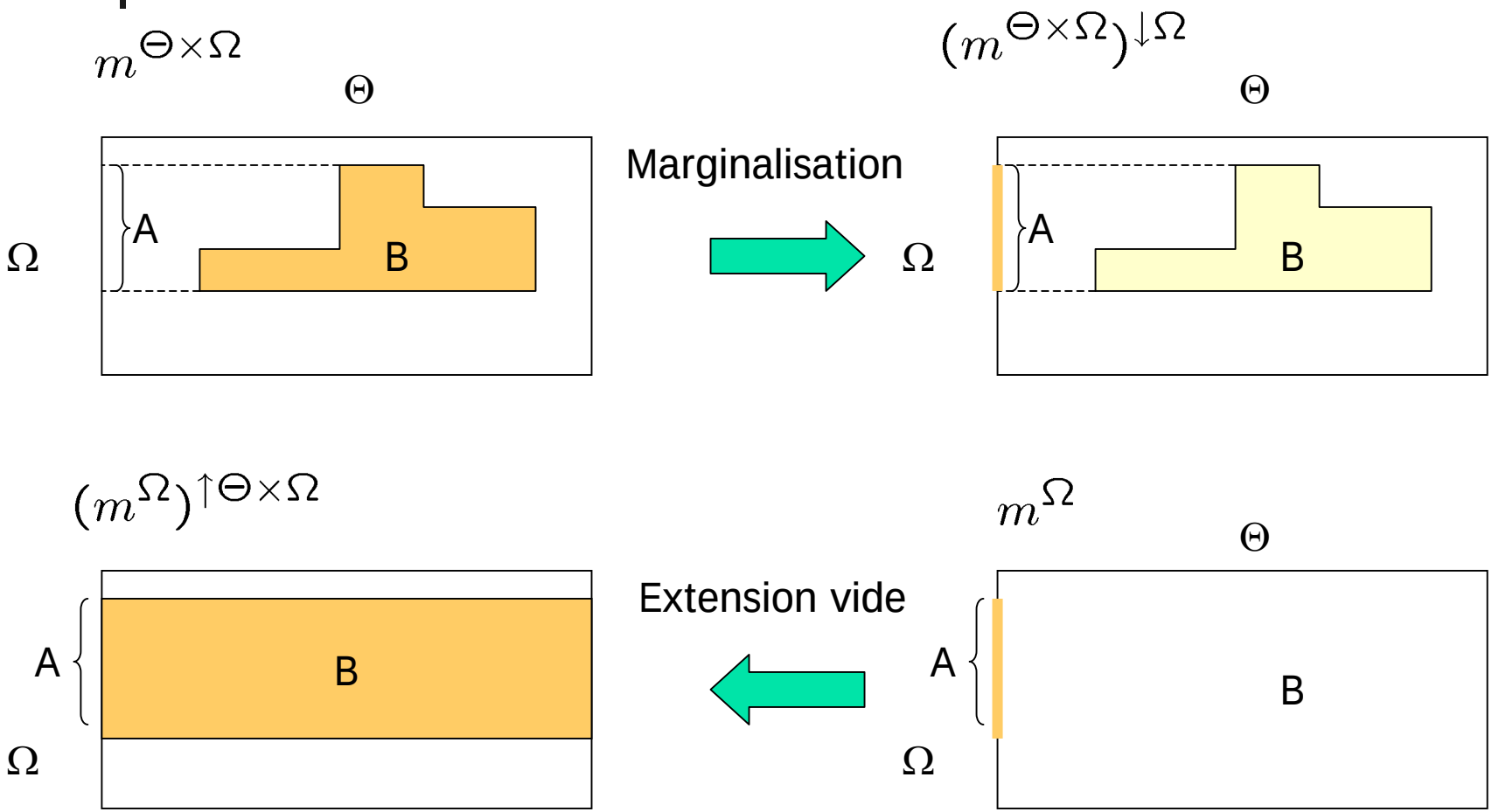
- Joue le même rôle que le *principe du maximum d'entropie* en probabilités.
- Applications :
  - extension vide
  - déconditionnement.

Application 1 : extension vide



- Soit m^{Θ} une fonction de masse sur Θ, traduisant un certain état de connaissance
- Problème : comment exprimer cet état de connaissance dans un référentiel Ω plus fin ? (transporter m^{Θ} dans Ω)
- Solution la moins informative : $m^{\Omega}(\rho(A)) = m^{\Theta}(A), \forall A \subseteq \Theta$

Cas d'un espace produit



Application 2 : déconditionnement

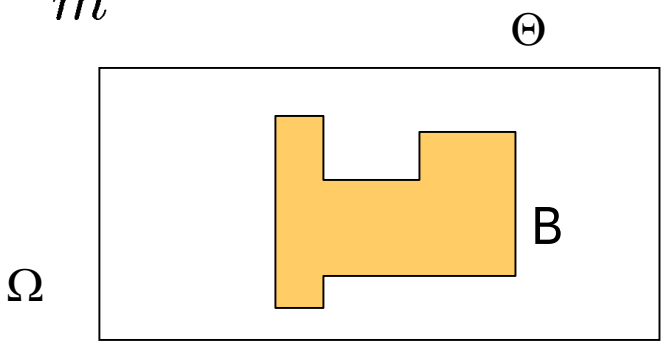
- Supposons que l'on connaisse $m^{\Omega}[A]$ pour $A \subseteq \Omega$ (état de connaissance sur $\omega \in \Omega$, dans un contexte où l'on sait que $\omega \in A$).
- Comment en déduire une fonction de masse non conditionnelle sur Ω ?
- Approche : recherche de la fonction de masse la moins informative, dont le conditionnement par rapport à A redonne $m^{\Omega}[A]$.
- Solution la moins informative :

$$m^{\Omega}(B \cup \overline{A}) = m^{\Omega}[A](B), \quad \forall B \subseteq A$$

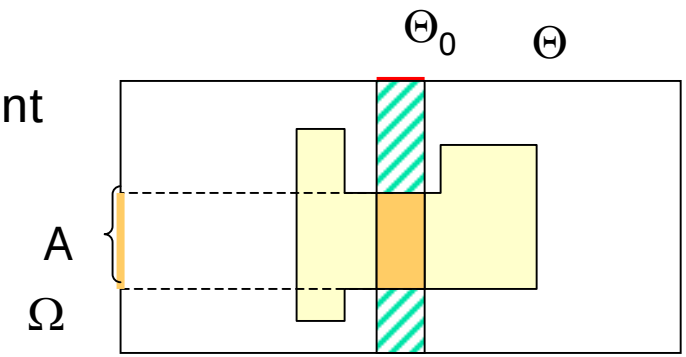
Cas d'un espace produit

$$m^{\Omega}[\theta \in \Theta_0] = \left(m^{\Theta\times\Omega} \textcircled{\cap} m_{\Theta_0}^{\Theta\uparrow\Theta\times\Omega}\right)^{\downarrow\Omega}$$

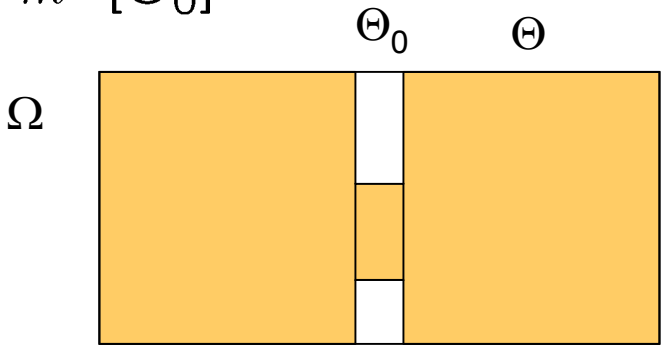
$m^{\Theta\times\Omega}$



Conditionnement

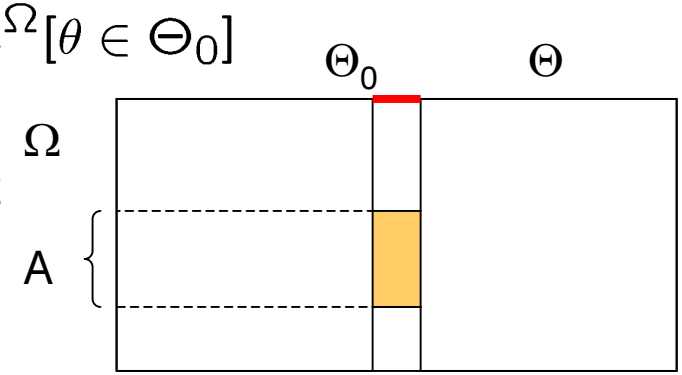


$m^{\Omega}[\Theta_0]^{\Uparrow\Theta\times\Omega}$



Déconditionnement

$m^{\Omega}[\theta \in \Theta_0]$



Affaiblissement

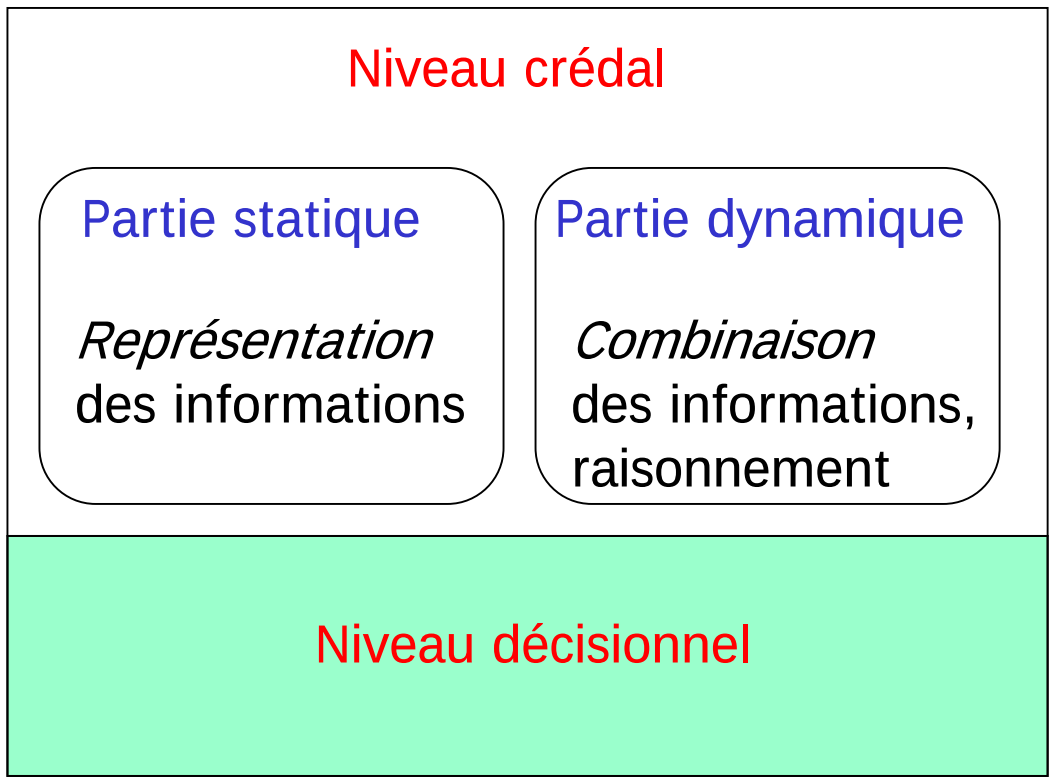
- m^{Ω}_{S} fonction de masse traduisant l'information sur une variable ω produite par une source S.
- Fiabilité de S inconnue : $\mathcal{R}$={F,NF}.
- On suppose :
  - $m^{\Omega}[F] = m^{\Omega}_{S}$
  - $m^{\Omega}[NF](\Omega)=1$.
  - croyances sur $\mathcal{R}$: $m^{\mathcal{R}}(NF) = \alpha$, $m^{\mathcal{R}}(F)=1-\alpha$
- Combinaison :

$$m^{\Omega}[m^{\Omega}_{S}, m^{\mathcal{R}}] = \left(m^{\Omega}[F]^{\Uparrow \Omega \times \mathcal{R}} \cap m^{\Omega}[NF]^{\Uparrow \Omega \times \mathcal{R}} \cap m^{\mathcal{R} \uparrow \Omega \times \mathcal{R}} \right)^{\downarrow \Omega}$$

- Résultat : Affaiblissement de m

$$\begin{aligned} {}^{\alpha}m(A) &= (1-\alpha)m(A) \quad \forall A \in 2^{\Omega} \setminus \Omega \\ {}^{\alpha}m(\Omega) &= m(\Omega) + \alpha(1-m(\Omega)) \end{aligned}$$

Le Modèle des Croyances Transférables



Justification des probabilités dans un cadre décisionnel

- « Un comportement rationnel face au risque suppose que les décisions soient prises sur la base d'une mesure de probabilité ».
- Soit Ω un ensemble d'états de la nature, $A \subseteq \Omega$
- Jeu : le joueur donne $\pi(A)$ au banquier. En échange, il reçoit 1 euro si A se réalise, 0 sinon.
- On demande de fixer $\pi(A)$, sans savoir si on jouera le rôle du joueur ou du banquier (on est obligé de participer au jeu).
- Si, π ne vérifie pas les axiomes des probabilités, on s'expose à « pari hollandais » (*Dutch book*) : suite de paris entraînant une perte à coup sûr.

Exemple

- Supposons que je fixe $\pi(A)=0.3$ et $\pi(\text{non } A)=0.6$.
- On me donne le rôle du banquier dans les deux jeux. Je reçois 0.3+0.6=0.9 euros, et je reverse 1 euro au joueur quoiqu'il arrive. Je perds à coup sûr 0.1 euro.
- Supposons maintenant que je fixe $\pi(A)=0.5$ et $\pi(\text{non } A)=0.6$.
- On me donne le rôle du joueur dans les deux jeux. Je verse 0.5+0.6=1.1 euros, et je gagne 1 euro quoiqu'il arrive. Je perds à coup sûr 0.1 euro.
- Pour éviter cette situation, il faut donc

$$\pi(A)+\pi(\text{non } A)=1$$

Probabilité pignistique

- Le raisonnement précédent justifie seulement l'existence d'une mesure de probabilité.
- MCT : construction de la fonction de probabilité pignistique

$$BetP(\omega) = \sum_{\{A \subseteq \Omega, \omega \in A\}} \frac{m(A)}{(1 - m(\emptyset))|A|}$$

Décision dans un cadre général

- Soient Ω un ensemble d'états, $\mathcal{A}$ un ensemble d'actions, C : $\mathcal{A} \times \Omega \rightarrow \mathbb{R}$ une fonction de coûts, et m une fonction de masse sur Ω. Choix d'une action ?
- MCT : Choix de l'action $\alpha \in \mathcal{A}$ qui minimise le minimise le « risque pignistique »

gnistique »

$$R_{BetP}(\alpha) = \sum_{\omega \in \Omega} C(\alpha, \omega) BetP(\omega)$$

nis

istique »

stique »

maximum de probabilité pignistique ».

nistique ».

Décision : maximum de plausibilité

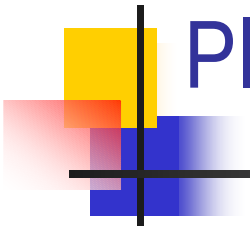
- Autre règle : « maximum de plausibilité ».
- Dans le contexte précédent, revient à choisir l'action qui minimise l'espérance inférieure du coût :

$$R_*(\alpha) = \sum_{A \subseteq \Omega} m(A) \min_{\omega \in A} C(\alpha, \omega)$$
$$= \min_{bel \leq P \leq pl} R_P(\alpha)$$

- Les deux règles peuvent conduire à des résultats très différents.
- Exemple : $\Omega=\{\omega_1,\ldots,\omega_{70}\}$, $m(\{\omega_1\})=0.3$, $m(\{\omega_2\})=0.01$, $m(\{\omega_2,\ldots,\omega_{70}\})=0.69$

  $pl(\{\omega_1\})=0.3$, **$pl(\{\omega_2\})=0.7$**, $pl(\{\omega_i\})=0.69$, $i=3,\ldots,70$.

  $BetP(\omega_1)=0.3$, $BetP(\omega_2)=0.02$, $BetP(\omega_i)=0.01$, $i=3,\ldots,70$



Plan

1. Théorie des Fonctions de Croyance (Dempster-Shafer, « théorie de l'évidence »)
   - Historique
   - Modèle des croyances transférables
     - partie statique (fonctions de croyance)
     - partie dynamique (combinaison, conditionnement, ...)
     - partie décisionnelle.
2. Applications en classification
   - Motivations
   - Construction de classifieurs crédibilistes
     - Théorème de Bayes généralisés
     - Classification à base de cas
     - Arbres de décision crédibilistes
   - Fusion de classifieurs
3. Conclusions et perspectives

Application en classification : Motivations

- « Points forts » de la théorie des fonctions de croyance :
  - représentation de connaissances partielles, depuis l'ignorance totale jusque la connaissance parfaite ;
  - combinaison d'informations issues de différentes sources.
- Application en classification :
  - problèmes pour lesquels l'information disponible est relativement « pauvre » :
    - données imprécises, incertaines, partiellement étiquetées ;
    - données d'apprentissage non totalement représentatives de l'environnement opérationnel ;
    - ensembles d'apprentissages hétérogènes, non exhaustifs.
  - problème de fusion de classifieurs.

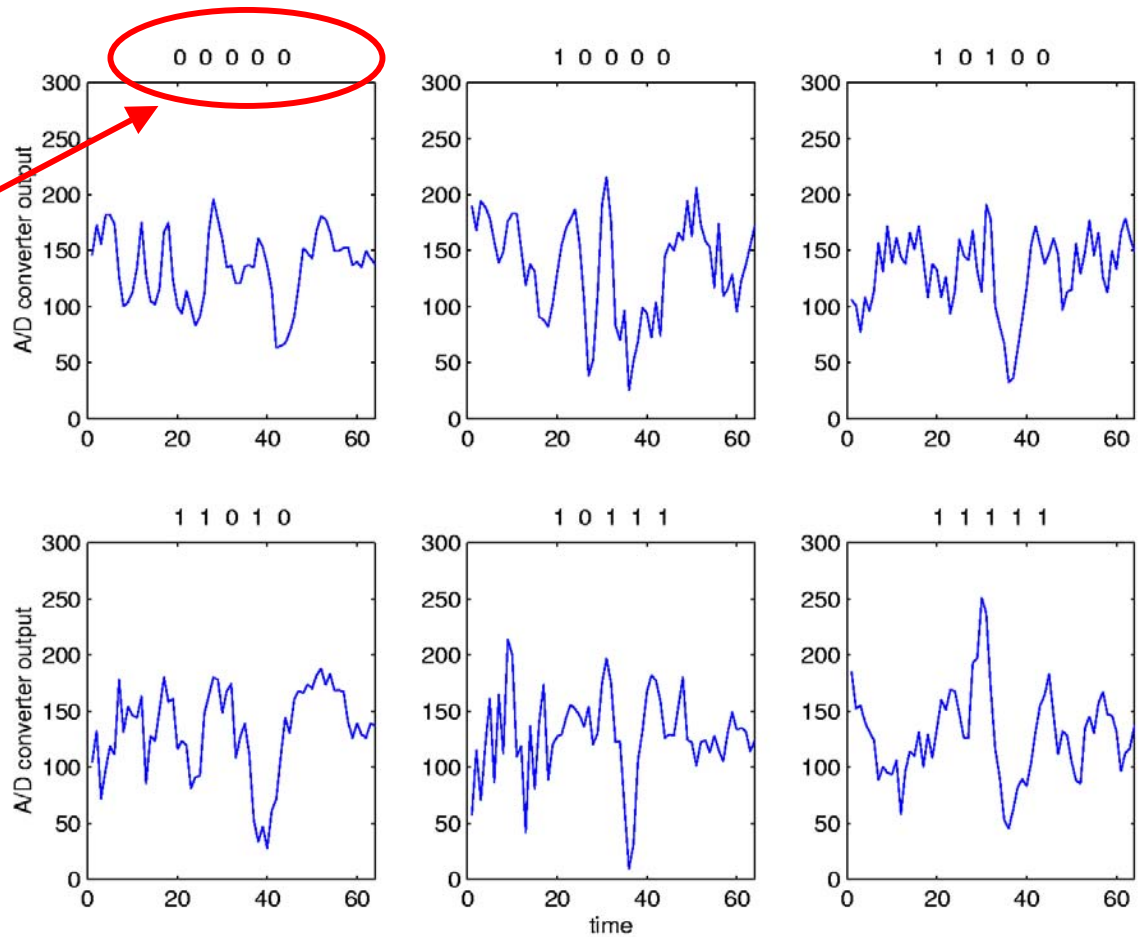
Exemple : classification de signaux EEG

- Pb : discriminer les complexes K du signal de fond dans des signaux EEG enregistrés pendant le sommeil (Richard, 1998).
- Complexe K = forme transitoire, utile pour l'étiquetage des stades de sommeil et le diagnostic en psychiatrie.
- Problèmes :
  - absence de critères objectifs : étiquetage des données par un panel d'experts ;
  - probabilité a priori d'apparition d'un complexe K dans une fenêtre temporelle inconnue (dépend du patient).

Détection de Complexes K

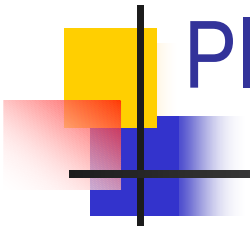
étiquetage subjectif par 5 experts

0= onde delta
1 = complexe K



Fusion multi-capteurs

- Attributs issus de s capteurs : $\mathbf{x}=(x_1,\ldots,x_s)$
- Exemples : télédétection (fusion d'images radar, visibles, infra-rouge), imagerie médicale (fusion d'images IRM multi-écho), applications militaires (identification de cible, etc.)
- Problèmes :
  - capteurs peu fiables dans certaines conditions opérationnelles, non nécessairement représentées dans l'ensemble d'apprentissage ;
  - données d'apprentissage incomplètes, hétérogènes, par ex :
    - capteur S_1 : n_1 exemples étiquetés $\{\omega_1,\omega_2\}$ ou ω_3
    - capteur S_2 : n_2 exemples étiquetés ω_1 ou ω_3 , etc ...

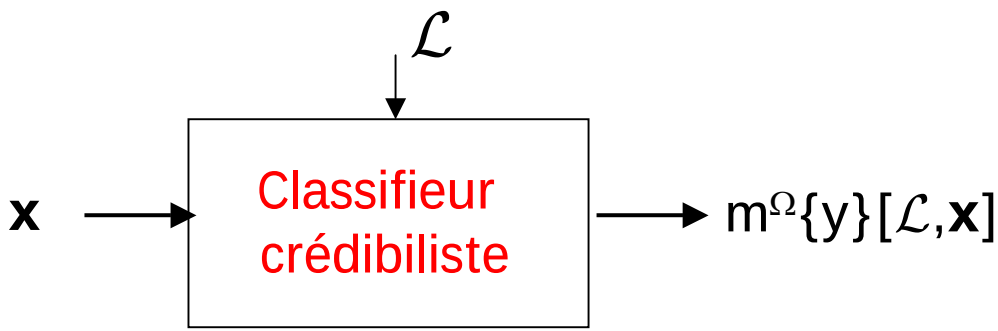


Plan

1. Théorie des Fonctions de Croyance (Dempster-Shafer, « théorie de l'évidence »)
   - Historique
   - Modèle des croyances transférables
     - partie statique (fonctions de croyance)
     - partie dynamique (combinaison, conditionnement, ...)
     - partie décisionnelle.
2. Applications en classification
   - Motivations
   - Construction de classifieurs crédibilistes
     - Théorème de Bayes généralisés
     - Classification à base de cas
     - Arbres de décision crédibilistes
   - Fusion de classifieurs
3. Conclusions et perspectives

Construction de classifieurs crédibilistes

- Problème de classification (discrimination) :
  - objets décrits par un vecteur d'attributs $\mathbf{x} \in \mathbb{R}^d$ et une variable de classe $y \in \Omega=\{\omega_1,...,\omega_K\}$.
  - ensemble d'apprentissage $\mathcal{L}$ (observations, eventuellement partielles, des variables **x** et y pour n objets).
  - Problème : prédire y sachant **x**, pour un nouvel objet.



- Trois approches :
  - Théorème de Bayes Généralisé (Smets 1978, Appriou 1991);
  - approche à base de cas (Denoeux, 1995);
  - arbres de décision crédibilistes (Elouedi 2000, Denoeux 2000).

Approche basée sur le TBG

- Théorème de Bayes Généralisé (Smets, 1978) :
  - Modèle $\{m^X[\omega_k], \forall\, \omega_k \in \Omega\}$ connu (X discret)
  - Pas de connaissance a priori sur Ω.
  - On observe $\mathbf{x} \in X_0 \subset X$. Croyances sur Ω ?
- Solution :

$$m^{\Omega}[X_0] = \left(\textcircled{\cap}_{k=1}^{K} m^X[\omega_k]^{\Uparrow X\times\Omega}\right)[X_0]$$

$$pl^{\Omega}[X_0](A) = 1 - \prod_{\omega_k \in A} (1 - pl^X[\omega_k](X_0)) \quad \forall A \subseteq \Omega$$

- Si connaissance a priori m_0^{Ω} : combinaison conjonctive
- Généralise le théorème de Bayes
- Application en classification (Appriou, 1991) : détermination des $pl^X[\omega_k]$, k=1,...,K (cas discret et continu)

Détermination de $pl^X[\omega_k]$ (X discret)

- Soit $\mathcal{L}_k$ un ensemble de n_k exemples issus de la classe ω_k.
- Hypothèse : $\mathcal{L}_k$ est un échantillon iid de $p(\mathbf{x}|\omega_k)$.
- $pl^X[\omega_k]$ peut être définie par un affaiblissement de la fonction de probabilité conditionnelle estimée :

$$pl^X[\omega_k](\mathbf{x}) = 1 - \alpha_k + \alpha_k \widehat{p}(\mathbf{x}|\omega_k)$$

- On a alors :

$$pl^\Omega[\mathbf{x}](A) = 1 - \prod_{\omega_k \in A} \alpha_k (1 - \widehat{p}(\mathbf{x}|\omega_k)) \quad \forall A \subseteq \Omega$$

Cas X continu

- On pose (Appriou, 1991) :

$$pl^{\Omega}[\mathbf{x}](A) = 1 - \prod_{\omega_k \in A} \alpha_k (1 - \rho \cdot \hat{p}(\mathbf{x}|\omega_k)) \quad \forall A \subseteq \Omega$$

avec

$$\rho = (\max_k \sup_{\mathbf{x}} \hat{p}(\mathbf{x}|\omega_k))^{-1}.$$

- Les coefficients α_k peuvent être fixés a priori ou appris à partir des données par minimisation d'une fonction d'erreur.
- Si une information a priori est disponible, elle est modélisée par une fonction de masse m^{Ω}_0 est combinée conjonctivement avec $pl^{\Omega}[\mathbf{x}]$

Propriétés

- Équivalence avec l'approche bayésienne dans le cas où les densités conditionnelles p(**x**|ω_k) et les probabilités a priori P(ω_k) sont connus.
- Séparabilité de l'évaluation des hypothèses : m^{Ω}[**x**] peut être décomposée comme la somme conjonctive de K fonctions m^{Ω}_k[**x**] définies par :

$$\begin{aligned} m_k^{\Omega}[\mathbf{x}](\overline{\{\omega_k\}}) &= \alpha_k(1-\rho\cdot\widehat{p}(\mathbf{x}|\omega_k)) \\ m_k^{\Omega}[\mathbf{x}](\Omega) &= 1-\alpha_k(1-\rho\cdot\widehat{p}(\mathbf{x}|\omega_k)) \end{aligned}$$

- Dans le cas de deux vecteurs d'attributs **x** et **x**' indépendants :

$$m^{\Omega}[\mathbf{x},\mathbf{x}'] = m^{\Omega}[\mathbf{x}] \textcircled{\cap} m^{\Omega}[\mathbf{x}']$$

Expérience 1 (Appriou, 1991)

- Problème de reconnaissance de cibles : 2 classes

  classes $\Omega=\{\omega_1,\omega_2\}$ (avion, missile) et 2 capteurs S_1

  $_1$ et S_2 (radar et infrarouge).

  rouge).

  ibut x_j

  e, dans des conditions expérimentales contrôlées :

  $$p(x_1|\omega_1) = \mathcal{N}(0,1) \qquad p(x_1|\omega_2) = \mathcal{N}(6,1)$$

  $$p(x_2|\omega_1) = \mathcal{N}(0,1) \qquad p(x_2|\omega_2) = \mathcal{N}(2,1)$$

  es :

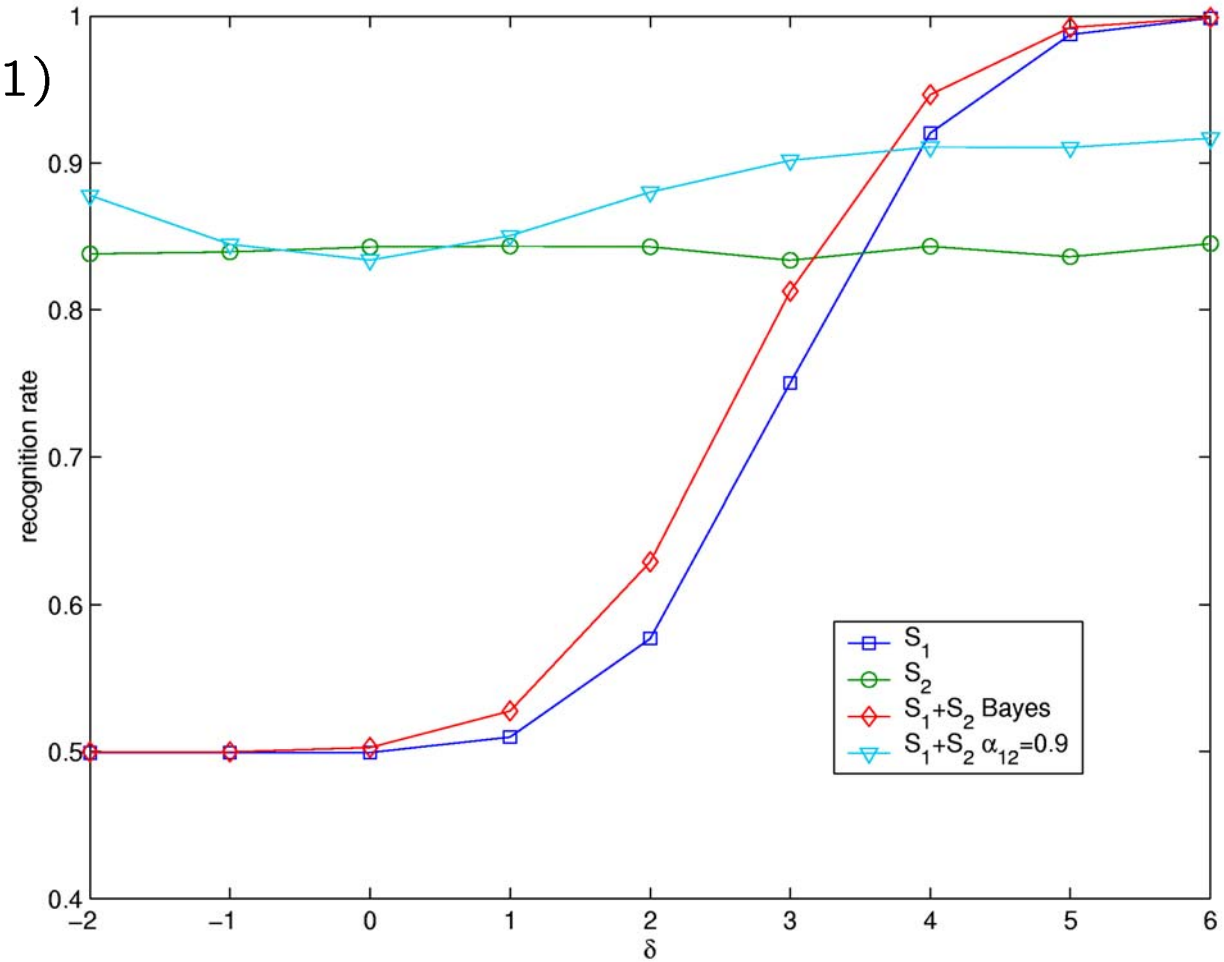
Expérience 1 (suite)

- Si les distributions de x_1 et x_2 sont inchangées dans un contexte opérationnel, alors le classifieur de Bayes est optimal.
- On considère le cas où la distribution de x_1 dans la classe ω_2 est modifiées par l'environnement du capteur.
- Ceci peut être modélisé en affaiblissant $p(x_1|\omega_2)$ par un facteur $1-\alpha_{1,2} > 0$
- On calcule ensuite $pl^{\Omega}[x_1]$, $pl^{\Omega}[x_2]$, puis

$$pl^{\Omega}[x_1, x_2] = pl^{\Omega}[x_1] \textcircled{\cap} pl^{\Omega}[x_2]$$

Expérience 1 : résultat

$$p(x_1|\omega_2) = \mathcal{N}(\delta, 1)$$

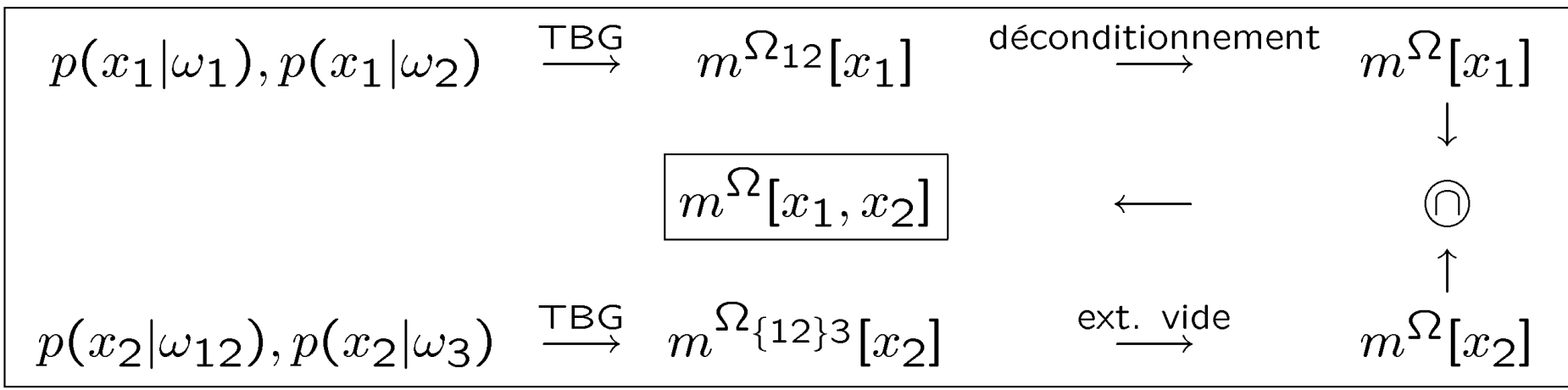


Expérience 2

- Deux capteurs S_1 et S_2, 3 classes $\Omega=\{\omega_1,\omega_2,\omega_3\}$.
- On connaît :
  - capteur S_1 : $p(x_1|\omega_1)$, $p(x_1|\omega_2)$
  - capteur S_2 : $p(x_2|\omega_1)=p(x_2|\omega_2)$, $p(x_2|\omega_3)$
- On ne connaît pas :
  - loi de x_1 dans la classe ω_3 : $p(x_1|\omega_3)$
  - probabilités a priori $P(\omega_1)$, $P(\omega_2)$, $P(\omega_3)$
- Deux solutions :
  - MCT
  - Bayes

Solution par le MCT

- Cadre de discernement pour le capteur S_1 : $\Omega_{12}=\{\omega_1,\omega_2\}$
- Cadre de discernement pour le capteur S_2 : $\Omega_{\{12\}3}=\{\omega_{12},\omega_3\}$ avec $\omega_{12}=\{\omega_1,\omega_2\}$



Solution bayésienne

- Nécessité de définir une distribution a priori sur Ω, et une densité conditionnelle p($x_1|\omega_3$)
- Solutions « non informatives »

$$P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$$

$$p(x_1|\omega_3) = \mathcal{U}_{[-1,5]}$$

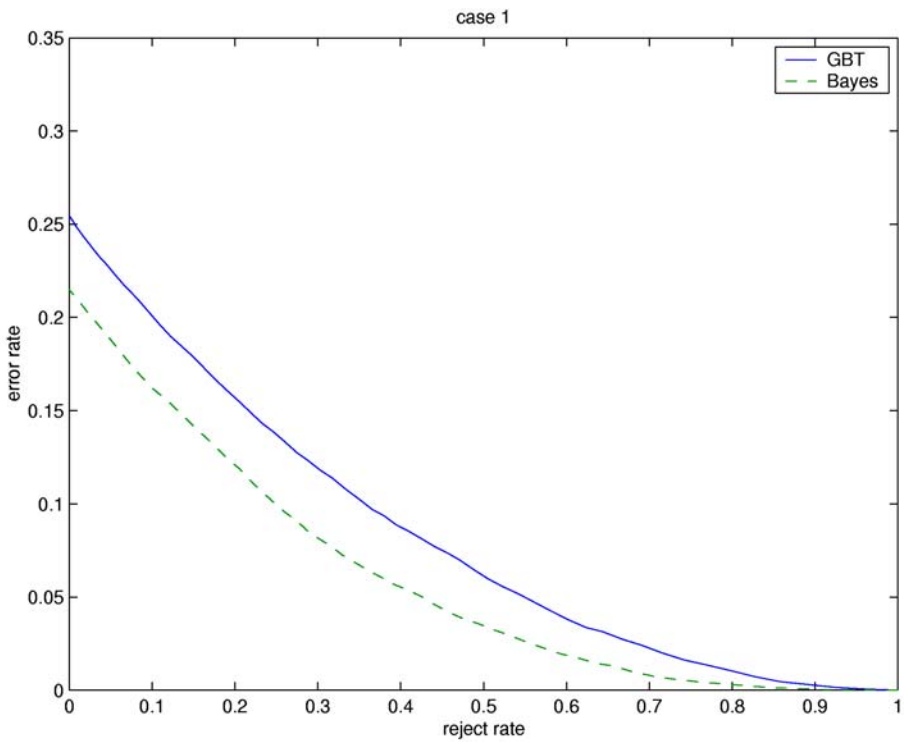
- Calcul des probabilités a posteriori :

$$P(\omega_k|x_1, x_2) = \frac{p(x_1|\omega_k)p(x_2|\omega_k)P(\omega_k)}{p(x_1, x_2)}$$

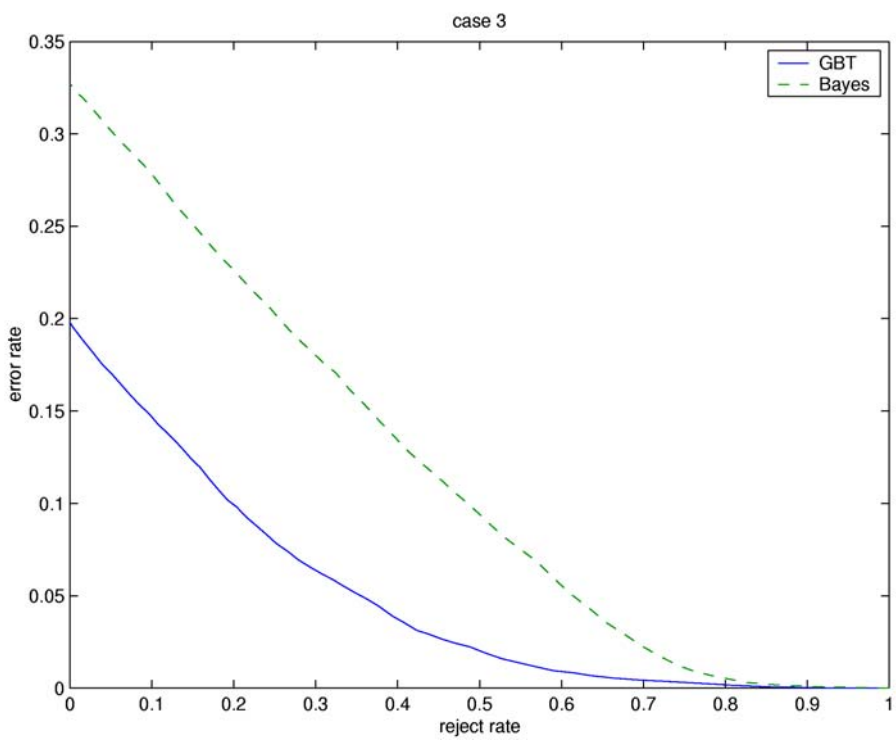
Exemple 2 : résultats

Cas 1 :
$P = (1/3, 1/3, 1/3)$

$p(x_1|\omega_3) = \mathcal{U}_{[-1,5]}$



Cas 2 :
$P = (0.1, 0.2, 0.7)$
$p(x_1|\omega_3) = \mathcal{N}(2, 1)$



Approche à base de cas (1)

- Ensemble d'apprentissage $\mathcal{L}=\{e_1,...,e_n\}$ avec $e_i=(\mathbf{x}_i,m_i)$ où $m_i=m^{\Omega}\{y_i\}$.
- Cas particuliers :
  - $m_i(\{\omega_k\})=1$: étiquetage précis, certain ;
  - $m_i(A)=1$, $A \subseteq \Omega$, $|A|>1$: étiquetage certain, imprécis ;
  - m_i est une fct de masse bayésienne : étiquetage incertain (probabiliste) ;
  - m_i est une fct de masse consonante : étiquetage possibiliste,
  - etc...
- On suppose définie une mesure de dissimilarité $\delta : X^2 \rightarrow \mathbb{R}_+$. (Par exemple, distance euclidienne si $X=\mathbb{R}^d$).

Approche à base de cas (2)

- Problème : construction d'une fonction de croyance $m^{\Omega}\{y\}$ concernant la classe y d'un nouvel exemple décrit par un vecteur d'attributs **x**.
- Principe : chaque exemple d'apprentissage e_i est une source d'information distincte sur y, d'autant plus pertinente que la dissimilarité $\delta(\mathbf{x},\mathbf{x}_i)$ est faible.
- Modélisation : affaiblissement de m_i

$$m^{\Omega}\{y\}[\mathbf{x}, e_i] = {}^{\alpha_i} m_i$$

avec $\alpha_i = \phi(\delta(\mathbf{x},\mathbf{x}_i)) \in [0,1]$, ϕ fonction croissante (le facteur d'affaiblissement est d'autant plus proche de 1 que la dissimilarité entre **x** et $\mathbf{x}_i$ est plus grande).

Approche à base de cas (3)

- Impact des n exemples :

$$m^{\Omega}\{y\}[\mathbf{x}, \mathcal{L}] = {}^{\alpha_1}m_1 \textcircled{\cap} \ldots \textcircled{\cap}^{\alpha_n}m_n$$

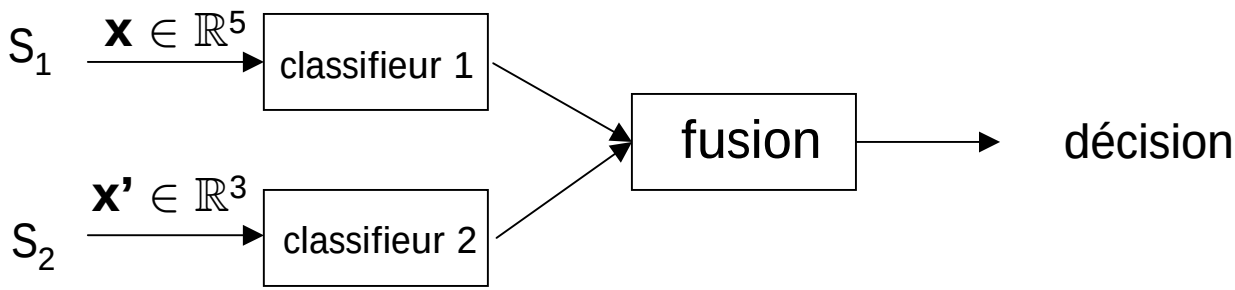
- Mise en œuvre de la méthode :
  - optimisation de la fonction d'affaiblissement ϕ par apprentissage ;
  - prise en compte uniquement des k plus proches voisins de **x** dans $\mathcal{L}$ (ou des vecteurs $\mathbf{x}_i$ t.q. $\delta(\mathbf{x},\mathbf{x}_i) \leq \delta_{min}$) ;
  - synthèse de $\mathcal{L}$ sous forme de p prototypes, déterminés par apprentissage supervisé ou non supervisé.
- Justification possible par le TBG (α_i = 1-plausibilité d'observer une distance $\delta(\mathbf{x},\mathbf{x}_i)$ pour 2 exemples de la même classe).

Résultats sur des données « classiques »

Vowel data $K = 11$, $d = 10$ $n = 568$ test : 462 ex. (different speak-ers)

Classifier	test error rate
Multi-layer perceptron (88 hidden units)	0.49
Radial Basis Function (528 hidden units)	0.47
Gaussian node network (528 hidden units)	0.45
Nearest neighbor	0.44
Linear Discriminant Analysis	0.56
Quadratic Discriminant Analysis	0.53
CART	0.56
BRUTO	0.44
MARS (degree=2)	0.42
Case-based classifier (33 prototypes)	0.38
Case-based classifier (44 prototypes)	0.37
Case-based classifier (55 prototypes)	0.37

Exemple : fusion de classifieurs

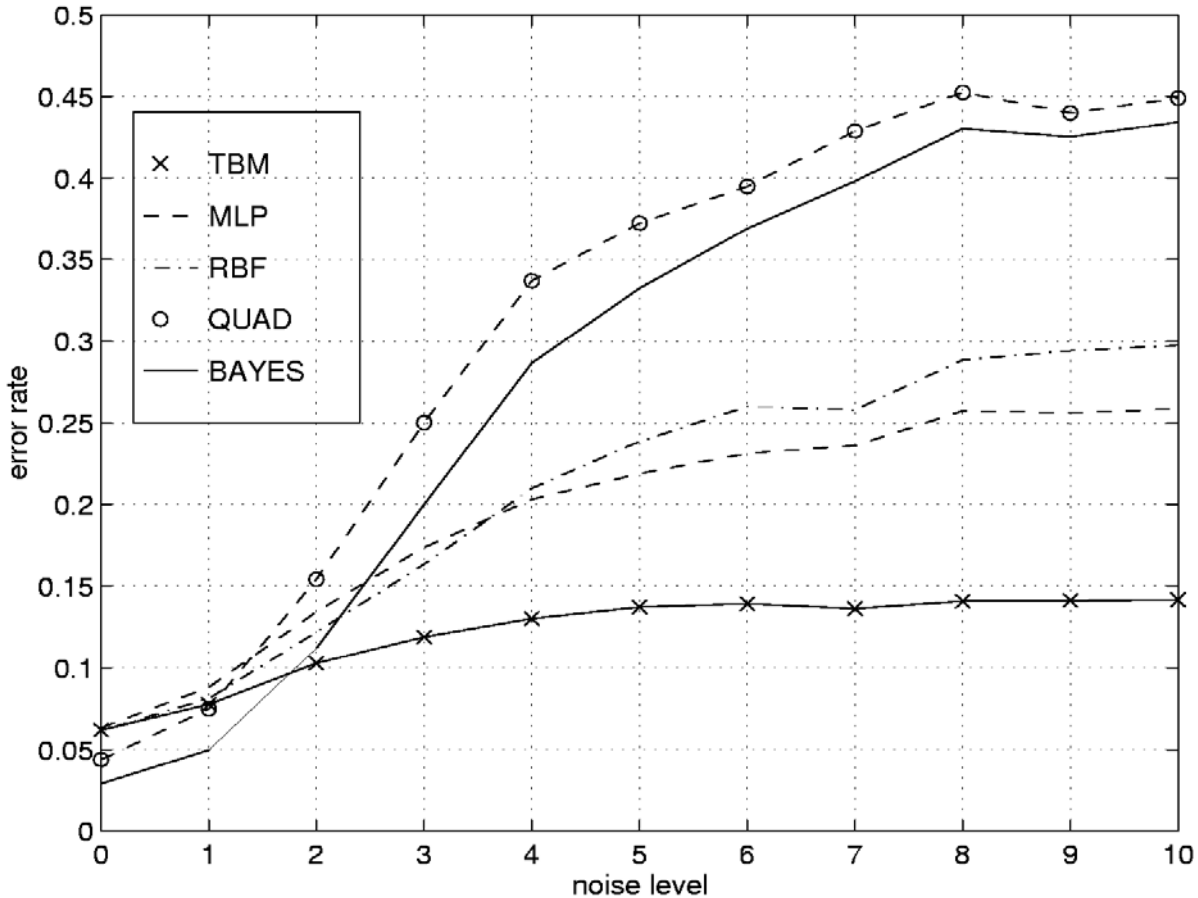


- 2 classes
- **x** et **x**' gaussiens, conditionnellement indépendants
- apprentissage : n=60, validation croisée : n_{cv}=100
- test : n_t=5000

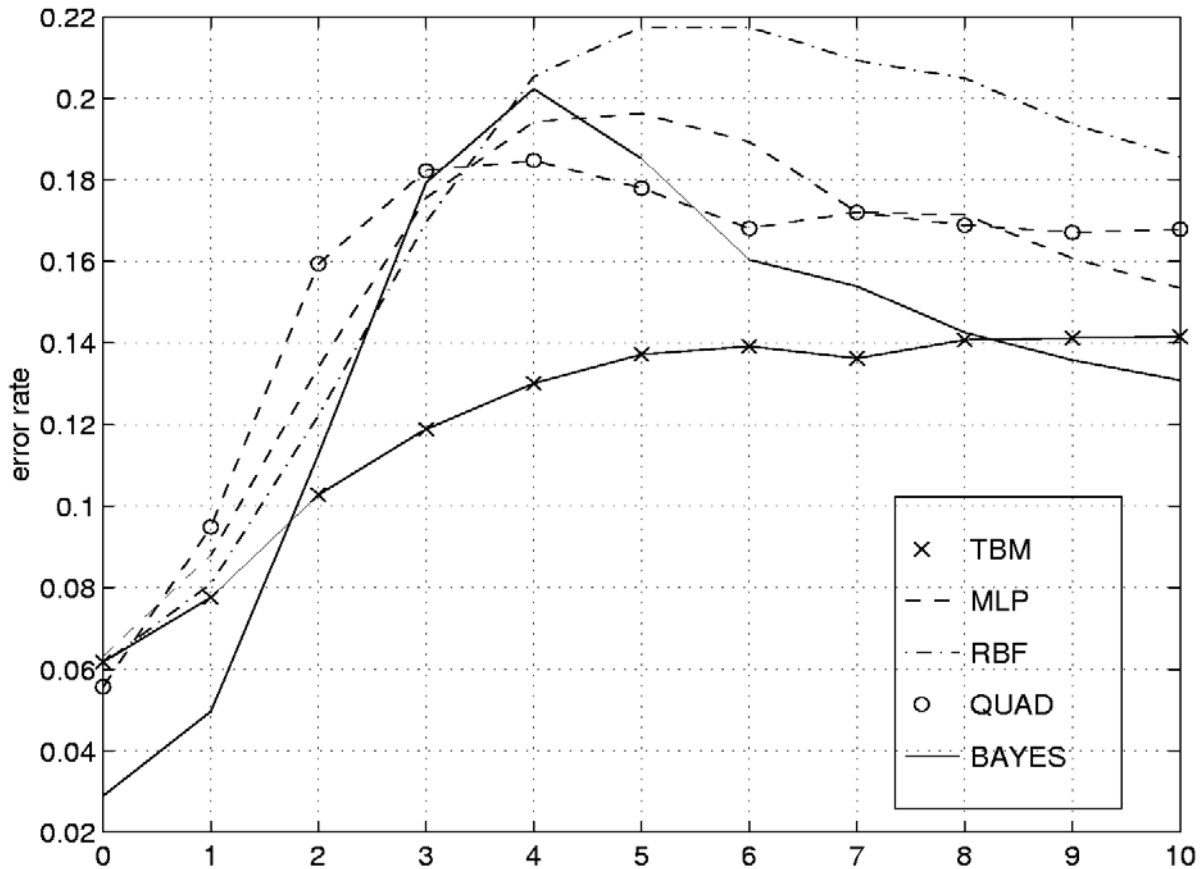
Résultats : test $\sim$ apprentissage

Méthode	$\mathbf{x}$ seul	$\mathbf{x}'$ seul	$\mathbf{x}$ et $\mathbf{x}'$
MCT	0.106	0.148	0.061
MLP	0.113	0.142	0.063
RBF	0.133	0.159	0.083
QUAD	0.101	0.141	0.049
BAYES	0.071	0.121	0.028

Test : $\mathbf{x} \leftarrow \mathbf{x} + \mathcal{N}(0,\sigma^2)$



Test : $\mathbf{x} \leftarrow \mathbf{x} + \mathcal{N}(0,\sigma^2)$ et rejet



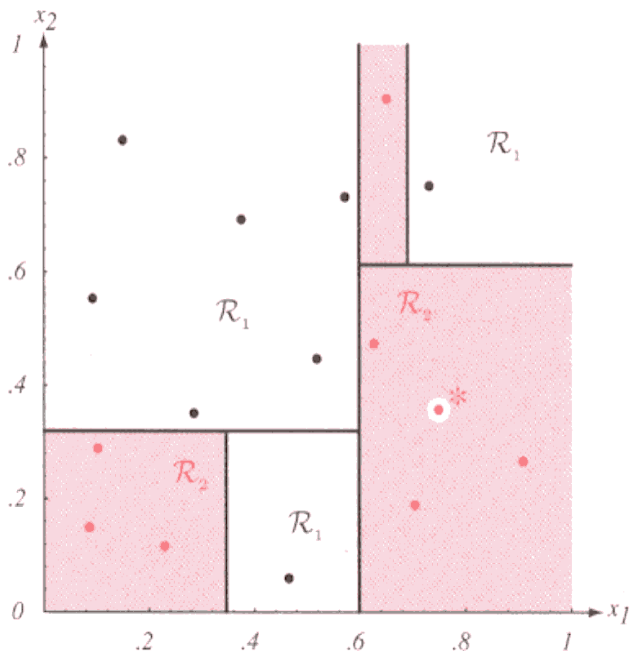
Données EEG

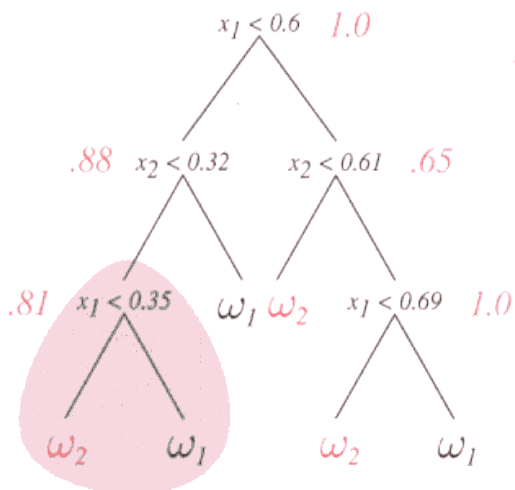
- K=2 classes, d=64
- données étiquetées par 5 experts
- n=200 exemples d'apprentissage, 300 exemples de test.

k	k-ppv	k-ppv p.	MCT (crisp labels)	MCT (uncert. labels)
9	0.30	0.30	0.31	0.27
11	0.29	0.30	0.29	0.26
13	0.31	0.30	0.31	0.26

Arbres de décision crédibilistes (1)

- Génération d'arbres de décision : partitionnement récursif de l'espace de représentation par des hyperplans correspondant à des tests effectués sur une ou plusieurs variables





Arbres de décision crédibilistes (2)

- Principe : partitionner les données en sous-ensembles les plus purs possible en une seule classe.
- Critères d'impureté classiquement utilisés : basés sur la distribution empirique des classes $p_k(t)=n_k(t)/n(t)$, k=1,…,K. Exemple : entropie

$$i(t) = -\sum_{k=1}^{K} p_k(t) \log_2 p_k(t)$$

- Pour un nouvel exemple **x** appartenant à la feuille t, l'incertitude sur y est décrite par les $p_k(t)$.
- Principe des AD crédibilistes : remplacer la loi de probabilité empirique $p_k(t)$ par une fonction de croyance traduisant mieux l'incertitude sur la classe d'un nouvel exemple $\in$ t.

Arbres de décision crédibilistes (3)

- Analogie :
  - suite de tirages dans une urne avec remise (cas K=2).
  - Ayant obtenu $n_1(t)$ boules blanches et $n_2(t)$ boules noires sur $n(t)=n_1(t)+n_2(t)$ tirages, quelle est ma croyance concernant la couleur de la prochaine boule tirée ?
- Solution proposée par Smets (1994) dans le cadre du MCT :

$$m^{\Omega}\{y\}[n_1(t), n_2(t)](\{\omega_k\}) = \frac{n_k(t)}{n(t)+1}, \quad k = 1, 2$$

$$m^{\Omega}\{y\}[n_1(t), n_2(t)](\Omega) = \frac{1}{n(t)+1}$$

- Définition d'un critère d'impureté : extension de la notion d'entropie aux fonctions de croyance.

Arbres de décision crédibilistes (4)

- Mesures d'incertitude d'une fonction de croyance :
  - non-spécificité :
  - conflit :

$$N(m) = \sum_{\emptyset \neq A \subseteq \Omega} m(A) \log_2(|A|)$$

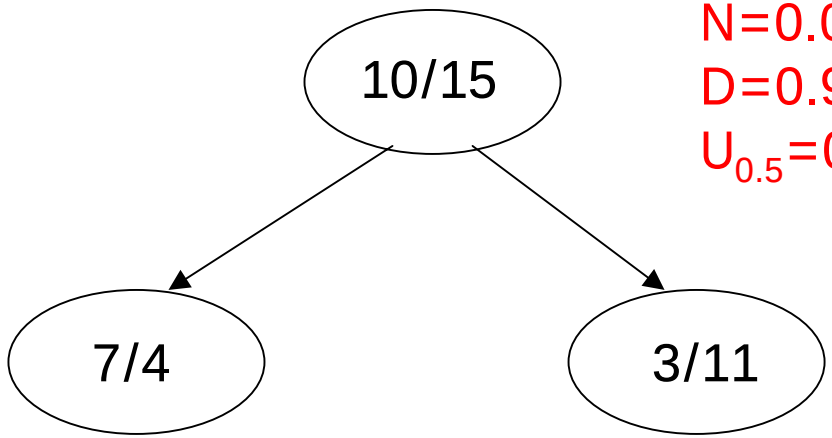
$$D(m) = -\sum_{A \subseteq \Omega} m(A) \log_2 BetP(A)$$

- Mesure d'incertitude totale :

$$U_\gamma(m) = (1-\gamma)\, N(m) + \gamma\, D(m), \quad 0 < \gamma < 1$$

- Propriété : lorsque l'on descend dans l'arbre, N(m)↗ et D(m)↘
- Possibilité d'utiliser l'augmentation de $U_\gamma(m)$ comme critère d'arrêt.
- γ peut être déterminé par validation croisée.

Exemple



N=0.039
D=0.93
$U_{0.5}$=0.49, $U_{0.1}$=0.128

N=0.083
D=0.87
$U_{0.5}$=0.48, $U_{0.1}$=0.16

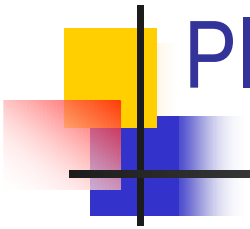
N=0.067
D=0.70
$U_{0.5}$=0.38, $U_{0.1}$=0.130

$\Delta U_{0.5}$=(11/25 * 0.48 + 14/25 *0.38)-0.49 = - 0.062 < 0
$\rightarrow$ partition acceptée

$\Delta U_{0.1}$= (11/25 * 0.16 + 14/25 *0.130) - 0.128 = 0.016 > 0
$\rightarrow$ partition rejetée

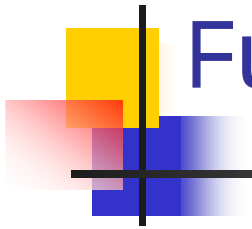
Arbres de décision crédibilistes

- Performances comparables à celles des méthodes classiques (CART, ...) sur des données « classiques »
- La méthode peut être étendue de manière rigoureuse au cas de données avec étiquetage incertain $\mathcal{L}=\{(\mathbf{x}_i,m_i),\ i=1,...,n\}$.
- L'arbre obtenu permet de décrire l'incertitude sur la classe y d'un nouvel exemple sous la forme d'une fonction de croyance $m^{\Omega}\{y\}[\mathbf{x},\mathcal{L}]$ qui peut être combinée avec d'autres sources d'information.



Plan

1. Théorie des Fonctions de Croyance (Dempster-Shafer, « théorie de l'évidence »)
   - Historique
   - Modèle des croyances transférables
     - partie statique (fonctions de croyance)
     - partie dynamique (combinaison, conditionnement, ...)
     - partie décisionnelle.
2. Applications en classification
   - Motivations
   - Construction de classifieurs crédibilistes
     - Théorème de Bayes généralisés
     - Classification à base de cas
     - Arbres de décision crédibilistes
   - Fusion de classifieurs
3. Conclusions et perspectives



Fusion de classifieurs

- Problématique générale : améliorer les performances d'un système de décision en combinant plusieurs classifieurs.
- Diversité des situations :
  - attributs identiques ou non ;
  - ensembles d'apprentissage identiques ou non ;
  - domaines de décision (ensemble des classes) identiques ou non ...
- Les outils de la théorie des fonctions de croyance doivent être choisis en fonction de chaque situation.
- Deux exemples de problèmes :
  - combinaison de classifieurs dans des espaces d'attributs différents ;
  - combinaison de classifieurs dans des espaces de décision différents.

Combinaison de classifieurs opérant dans des espaces d'attributs différents

- s classifieurs (a priori quelconques) opérant dans des espaces de représentation indépendants : sources d'information distinctes.
- Deux méthodes différent par la nature des sorties des classifieurs :
  - méthode de Xu et al. (1992) : chaque classifieur produit une décision d'affectation à une classe ou de rejet.
  - méthode « expert tuning » (Elouedi et Smets, 2004) : chaque classifieur produit une fonction de croyance (méthode plus générale que la précédente).

Méthode de Xu et al. (1992)

- Principe : transformer la décision de chaque classifieur en une fonction de masse, en utilisant la matrice de confusion.
- Soient e_r et c_r les taux d'erreur et de bonne classification du classifieur r ($e_r + c_r < 1$ car rejet possible) ;
- Soit $D_r \in \{0,1,...,K\}$ la décision du classifieur r.
- Traduction sous forme d'une fonction de masse :
  - si $D_r=0$, $m_r(\Omega)=1$;
  - si $D_r=k \in \{1,...,K\}$,
    $m_r(\{\omega_k\})=c_r$,
    $m_r(\Omega \setminus \{\omega_k\})=e_r$
    $m_r(\Omega)=1-e_r-c_r$.
- Combinaison des m_r (r=1,...,s) par la règle de Dempster.
- Bons résultats sur des problèmes de reconnaissance de caractères.

Expert tuning (Elouedi et Smets, 2004)

- Chaque classifieur r produit une fonction de masse

$$m_r = m^{\Omega}_{S_r}\{y\}$$

- Les classifieurs sont considérés comme des sources d'information indépendantes, avec des degrés de fiabilité différents.
- Modèle : chaque classifieur r est « fiable » avec une probabilité 1-α_r. Chaque masse m_r doit être affaiblie par un facteur α_r.
- Les s classifieurs sont ensuite combinés conjonctivement :

$$m = {}^{\alpha_1}m_1 \textcircled{\cap} \ldots \textcircled{\cap}^{\alpha_s}m_s$$

- Les coefficients α_r sont déterminés automatiquement par minimisation d'une fonction d'erreur.

Expert tuning : exemple

$$\Omega = \{\text{Airplane}, \text{Helicopter}, \text{Rocket}\}$$

	A	H	R	$\{A,H\}$	$\{A,R\}$	$\{H,R\}$	Ω	y_i
$m_1\{y_1\}$	0	0	0.5	0	0	0.3	0.2	A
$m_1\{y_2\}$	0	0.5	0.2	0	0	0	0.3	H
$m_1\{y_3\}$	0	0.4	0	0	0.6	0	0	A
$m_1\{y_4\}$	0	0	0	0	0.6	0.4	0	R
$m_2\{y_1\}$	0	0	0	0.7	0	0	0.3	A
$m_2\{y_2\}$	0.3	0	0	0.4	0	0	0.3	H
$m_2\{y_3\}$	0.2	0	0	0	0	0.6	0.2	A
$m_2\{y_4\}$	0	0	0	0	0	1	0	R

$$\alpha_1^* = 0.66 \qquad \alpha_2^* = 0.52$$

Combinaison de classifieurs opérant dans des espaces de décision différents (1)

- Certains algorithmes d'apprentissage sont adaptés aux problèmes à 2 classes (classifieurs binaires), ex : SVM.
- Extension à K classes : combinaison de plusieurs classifieurs binaires.
- Plusieurs stratégies, par exemple :
  - « un contre tous » : K classifieurs, espaces de décision

$$\Theta_k = \{\theta_{k,1}, \theta_{k,2}\}, \quad \theta_{k,1} = \{\omega_k\}, \theta_{k,2} = \overline{\{\omega_k\}}$$

  - « deux à deux » : K(K-1)/2 classifieurs, espaces de décision

$$\Omega_{k\ell} = \{\omega_k, \omega_\ell\}$$

- Combinaison de ces classifieurs : méthodes le plus souvent empiriques, mal justifiées dans un cadre probabiliste.
- Formalisation dans le cadre de la théorie des fonctions de croyance ?

Combinaison de classifieurs opérant dans des espaces de décision différents (2)

- Principe général : expression des sorties de chaque classifieur binaire dans l'espace d'origine Ω, puis combinaison.
- Cas 1 : stratégie « un-contre-tous »
  - La sortie $m_k^{\Theta k}$ du classifieur k s'exprime dans l'espace Θ_k , défini comme une partition (grossissement, *coarsening*) de Ω
  - Expression dans Ω : extension vide

$$\begin{aligned} m_k^{\Omega}(\{\omega_k\}) &= m_k^{\Theta_k}(\{\theta_{k,1}\}) \\ m_k^{\Omega}(\overline{\{\omega_k\}}) &= m_k^{\Theta_k}(\{\theta_{k,2}\}) \\ m_k^{\Omega}(\Omega) &= m_k^{\Theta_k}(\Theta_k) \end{aligned}$$

- Combinaison :

$$m^{\Omega} = \frac{1}{K} \sum_{k=1}^{K} m_k^{\Omega}$$

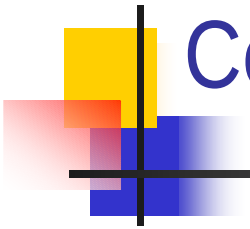
Combinaison de classifieurs opérant dans des espaces de décision différents (3)

- Cas 2 : stratégie « deux-à-deux »
  - La sortie du classifieur (k,l) s'exprime dans l'espace $\Omega_{k,l} \subset \Omega$
  - $\rightarrow$ peut être vue comme une fonction de masse conditionnelle
  - Expression dans Ω : déconditionnement

$$
\begin{aligned}
m_{k,\ell}^{\Omega_{k,\ell}\Uparrow\Omega}(\{\omega_k\} \cup \overline{\Omega_{k,\ell}}) &= m_{k,\ell}^{\Omega_{k,\ell}}(\{\omega_k\}) \\
m_{k,\ell}^{\Omega_{k,\ell}\Uparrow\Omega}(\{\omega_\ell\} \cup \overline{\Omega_{k,\ell}}) &= m_{k,\ell}^{\Omega_{k,\ell}}(\{\omega_\ell\}) \\
m_{k,\ell}^{\Omega_{k,\ell}\Uparrow\Omega}(\Omega) &= m_{k,\ell}^{\Omega_{k,\ell}}(\Omega_{k,\ell})
\end{aligned}
$$

- Combinaison :

$$
m^{\Omega} = \frac{2}{K(K-1)} \sum_{k<\ell} m_{k,\ell}^{\Omega_{k,\ell}\Uparrow\Omega}
$$



Conclusion

- Théorie des fonctions de croyance : outil de modélisation riche et flexible permettant la représentation et la gestion de différentes formes d'incertitudes.
- Potentialités d'application en classification :
  - problèmes dans lesquels l'information disponible est trop parcellaire pour être modélisée dans un cadre probabiliste sans hypothèses arbitraires ;
  - problèmes dans lesquels la combinaison d'informations hétérogènes joue un rôle important (fusion multi-capteurs, intégration de connaissances expertes, systèmes interactifs d'aide à la décision).
- Perspectives :
  - développement d'outils d'inférence statistique permettant la construction de fonctions de croyance à partir d'observations ;
  - extension à des référentiels continus.