Application du formalisme des Fonctions de Croyance en fusion d'informations et en 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

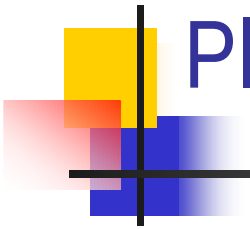
1. Problématique de la fusion d'informations
2. Approches « probabilistes »
   - Statistique classique
   - Cadre Bayésien
3. Théorie des possibilités
4. Théorie des fonctions de croyance
5. Applications en classification
6. Conclusions et perspectives

Problématique de la Fusion d'informations

- Combinaison de données issues de différentes sources, en vue de répondre à une certaine question (valeur d'un paramètre).
- Problématique récente (multiplication des capteurs, développement des ressources informatiques)
- Domaines d'applications :
  - Fusion multi-capteurs (application militaire, télédétection, etc.)
  - Systèmes d'interrogation de bases de données multiples
  - Combinaison d'avis d'experts
- Difficulté du problème : données incomplètes, incertaines, hétérogènes, issues de sources de fiabilité inconnue, éventuellement dépendantes.
- Cadres théoriques : probabilités, possibilités, croyances.

Formalisation

- Question = trouver la valeur d'un paramètre θ à valeur dans Θ (type quantitatif ou qualitatif, monodimensionnel ou vectoriel)
- Soit S un ensemble de sources, dont chacune apporte une information sur θ :
  - Information certaine et précise : une source suffit
  - Information incertaine ou imprécise : intérêt de la fusion
- Principales questions :
  - Comment représenter l'information sur θ apportée par chacune des sources ? Nécessité d'un cadre théorique pour la représentation des incertitudes
  - Comment combiner ces informations, en prenant en compte la fiabilité des sources ?



Plan

1. Problématique de la fusion d'informations
2. Approches « probabilistes »
   - Statistique classique
   - Cadre Bayésien
3. Théorie des possibilités
4. Théorie des fonctions de croyance
5. Applications en classification
6. Conclusions et perspectives

Approche statistique « classique »

- Source d'information = variable aléatoire X (associée à une expérience aléatoire $\mathcal{E}$) , dont la loi de probabilité $p_X(\cdot;\theta)$ dépend de θ

  $\int_A p_X(x;\theta)dx$ = fréquence limite de réalisation de n de l'événement $X \in A$, lorsque le nombre de ions de $\mathcal{E}$ tend vers l'infini

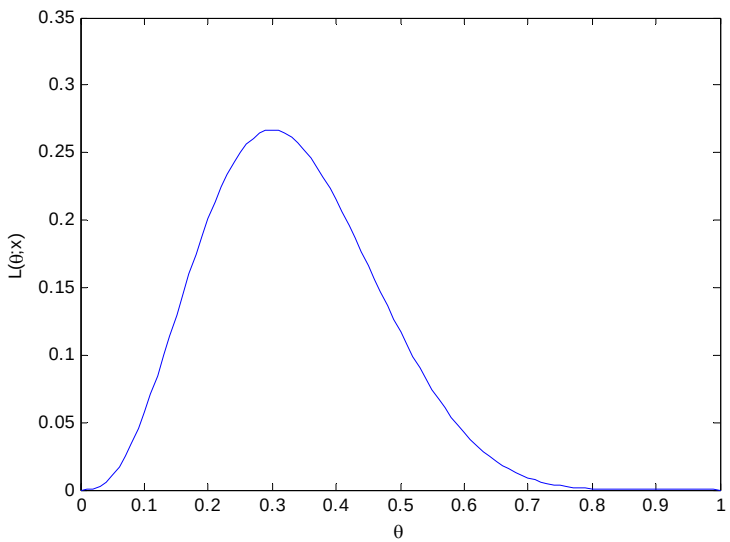
  rvé une réalisation x de X, l'information sur θ est exprimée par la fonction de vraisemblance $x;\theta)$.

  s une mesure de probabilité !

Exemple

- θ=proportion de boules blanches dans une urne.
- X = nombre de boules blanches obtenues au cours de n tirages avec remise
- $L(\theta;x)=p_X(x;\theta)=C_n^x\ \theta^x(1-\theta)^{n-x}$

n=10, x=3



Fusion de fonctions de vraisemblance

- Soient deux v.a. indépendantes X et Y, de lois $p(x;\theta)$ et $p(y;\theta)$.
- Ayant observé les réalisations x et y, l'information sur θ est représentée par la fonction de vraisemblance

  $L(\theta;x,y) = p_{X,Y}(x,y;\theta)=p_X(x;\theta)p_Y(y;\theta) = L(\theta;x)\ L(\theta;y)$

- Conclusion :
  - Cadre restrictif (sources = variables aléatoires dont les lois de probabilité sont gouvernées par le paramètre inconnu θ)
  - Information sur θ représentée de manière non probabiliste par la fonction de vraisemblance
  - Opérateur de combinaison = produit.

Le cadre Bayésien

- Postulat : les degrés de croyance d'un agent rationnel obéissent, comme les fréquences, à l'axiome d'additivité :
  - Soit P(A) = degrés de croyance dans la proposition $\theta \in A$, $\forall A \subseteq \Theta$
  - $\forall A,B \subseteq \Theta$, $P(A \cup B)=P(A)+P(B)-P(A \cap B)$
  - Par convention, $P(\Theta)=1$.
- P(A) = probabilité subjective (même formalisme mathématique que pour la représentation des fréquences, mais signification différente).
- Cas particulier : degrés de croyances $\equiv$ fréquences lorsque celle-ci sont connues (exemple de l'urne).

Fusion bayésienne

- Soient deux sources S_1 et S_2 fournissant des données x et y (par exemple : estimations de θ).
- On suppose connues :
  - $p(x,y|\theta)$, $\forall\ \theta \in \Theta$ (probabilité que les sources S_1 et S_2 fournissent les observations x et y, connaissant la vraie valeur de θ)
  - p_0 = probabilité a priori sur θ
- On en déduit :

$$p(\theta|x,y) = \frac{p(x,y|\theta)p_0(\theta)}{\sum_{\theta' \in \Theta} p(x,y|\theta')p_0(\theta')}$$

Fusion bayésienne (suite)

- Cas particulier : sources indépendantes $p(x,y|\theta)=p(x|\theta)p(y|\theta)$.

$$p(\theta|x,y) = \frac{p(x|\theta)p(y|\theta)p_0(\theta)}{\sum_{\theta' \in \Theta} p(x|\theta')p(y|\theta')p_0(\theta')}$$

$$p(\theta|x,y) = \frac{L(\theta;x)L(\theta;y)}{\sum_{\theta' \in \Theta} L(\theta';x)L(\theta';y)}$$

Fusion de lois de probabilités

- Combinaison d'avis d'experts : chaque source (expert) S_i fournir une distribution de probabilité p_i, traduisant sa connaissance sur θ
- Problème : construire une distribution de probabilité p agrégée reflétant l'opinion du groupe d'experts.
- Méthode du consensus :

$$p = \sum_i w_i \, p_i$$

- w_i = poids reflétant la fiabilité de l'expert i
- Des méthodologies ont été développées pour construire les p_i (quantiles) et les w_i

Critique de l'approche bayésienne

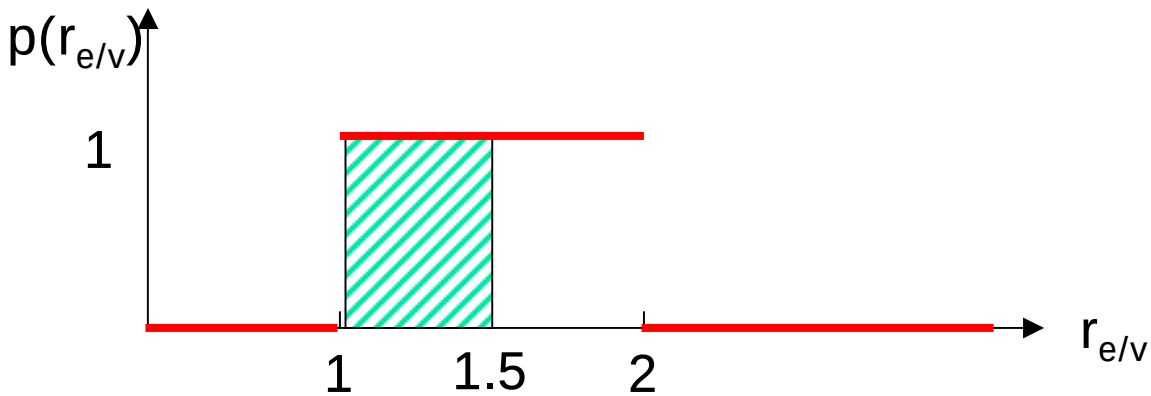
- Le postulat selon lequel tout état de connaissance peut être représenté par une distribution de probabilité semble critiquable dans le cas où l'information est très pauvre :
  - Quelle loi a priori sur θ lorsqu'on ne dispose d'aucune information initiale ?
  - Représentation de la quasi-ignorance d'un expert ?
- Principe d'indifférence ou « de raison insuffisante » (Bernouilli) :
  - l'ignorance totale est modélisée par une distribution de probabilité uniforme sur Θ, supposée non informative
  - Justification par le principe du maximum d'entropie.

Critique du principe d'indifférence

- Non-invariance par rapport au choix des variables ou du domaine de référence :
  - Si X suit une loi uniforme, il n'en est pas de même de f(X) pour f non linéaire.
  - Si $\Theta=\{\theta_1,\theta_2\}$ est raffiné en $\Theta'=\{\theta_{1,1},\theta_{1,2},\theta_2\}$ avec $\theta_1=\{\theta_{1,1},\theta_{1,2}\}$, une loi uniforme sur Θ ne l'est plus sur Θ'.
- 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

- 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]



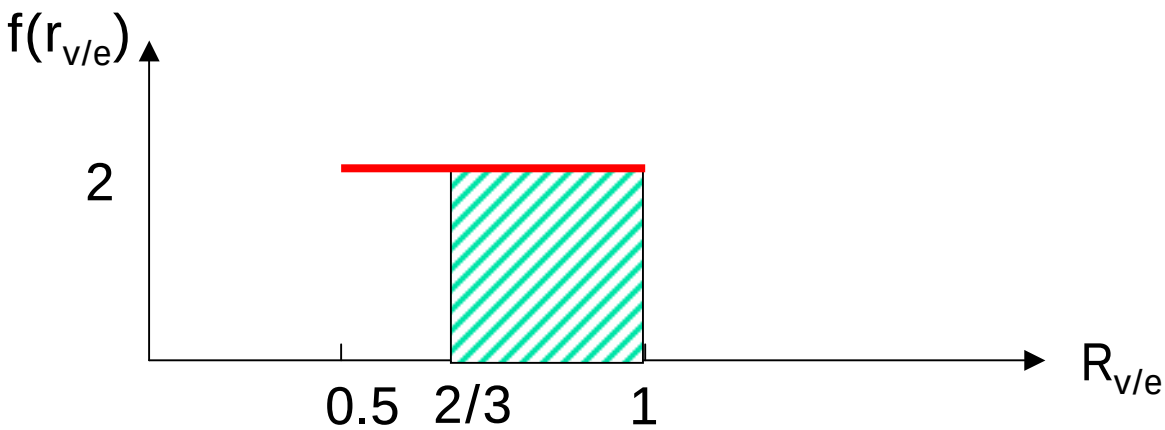
- $P(r_{e/v} \leq 1.5)=0.5$
- Probabilité que la bouteille contienne au plus 1.5 plus d'eau que de vin = 0.5

Paradoxe de Bertrand (suite)

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

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

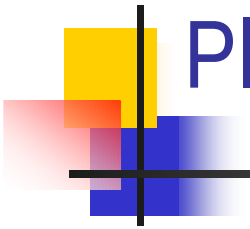
- PRI : loi de probabilité uniforme sur [0.5,1]



- $r_{e/v} \leq 1.5 \Leftrightarrow r_{v/e} \geq 2/3$. 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 !

Conclusion sur les approches probabilistes

- Assez peu de travaux en fusion dans un cadre strictement probabilistes car :
  - Théorie des probabilités objectives : domaine d'application limité (processus répétables)
  - Théorie des probabilités subjectives peu adaptée à la représentation d'états de connaissance partielle induits par des informations imprécises.
- Autres cadres théoriques proposés dans les années 1970 :
  - Théorie des possibilités (Zadeh, 1978)
  - Théorie des fonctions de croyance (Dempster, 1968 ; Shafer, 1976).



Plan

1. Problématique de la fusion d'informations
2. Approches « probabilistes »
   - Statistique classique
   - Cadre Bayésien
3. Théorie des possibilités
4. Théorie des fonctions de croyance
5. Applications en classification
6. Conclusions et perspectives

Théorie des possibilités (1)

- Information disponible : $\theta \in E$, $E \subseteq \Theta$:
  - Un événement $A \subseteq \Theta$ est possible ssi $A \cap E \neq \emptyset$:

$$\begin{aligned} \Pi(A) &= \begin{cases} 1 & \text{si } A \cap E \neq \emptyset \\ 0 & \text{sinon} \end{cases} \\ &= \sup_{u \in A} 1_E(u) \end{aligned}$$

  - Un événement $A \subseteq \Theta$ est nécessaire ssi $E \subseteq A$:

$$\begin{aligned} N(A) &= \begin{cases} 1 & \text{si } E \subseteq A \\ 0 & \text{sinon} \end{cases} \\ &= 1 - \Pi(\overline{A}) \end{aligned}$$

Théorie des possibilités : généralisation

- Soit E un sous-ensemble flou de Θ, de fonction d'appartenance $\pi : \Theta \rightarrow [0,1]$.
- Soient $\Pi, N : \ 2^{\Theta} \rightarrow [0,1]$ t.q.

$$\Pi(A) = \sup_{u \in A} \pi(u)$$

$$N(A) = 1 - \Pi(\overline{A})$$

- Π = mesure de possibilité

  N = mesure de nécessité duale

  π = distribution de possibilités

Propriétés

- Propriété d'additivité remplacée par :

$$\Pi(A \cup B) = \max(\Pi(A), \Pi(B))$$
$$N(A \cap B) = \min(N(A), N(B))$$

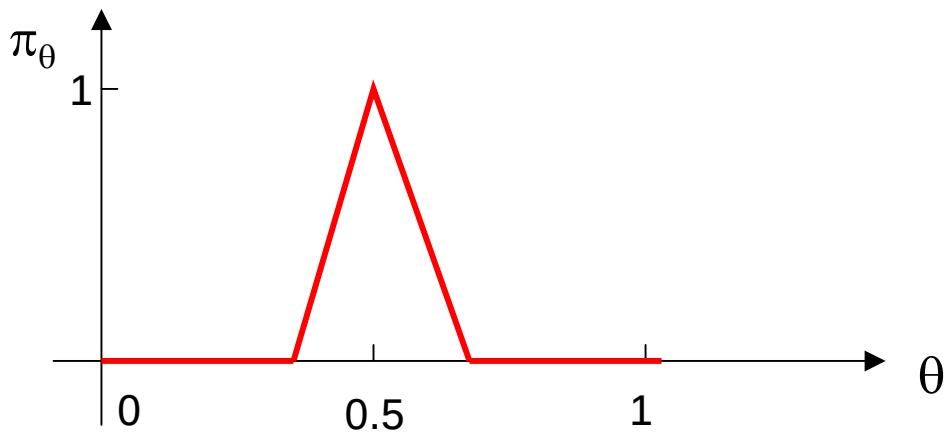
- $\Pi(\Theta)=N(\Theta)=1$, $\Pi(\emptyset)=N(\emptyset)=0$
- On a seulement

$$\max(\Pi(A), \Pi(\overline{A})) = 1$$

$\rightarrow$ incertitude relative à l'événement A caractérisée par deux valeurs Π(A) et N(A), avec N(A) $\leq \Pi$(A).

Interprétation 1 : contrainte flexible

- π = contrainte flexible définie, par exemple, de manière linguistique
  - Exemple : l'urne contient « environ 50 % » de boules blanches :



Interprétation 2 : famille de probabilités

- Soient $A_1 \subset A_2 \subset ... \subset A_n \subseteq \Theta$.
- Soit λ_i = borne inférieure sur la probabilité (définie mais inconnue) que A_i contienne la vraie valeur du paramètre θ
- Soit $\mathcal{P}=\{P \mid P(A_i) \geq \lambda_i, i=1,...,n\}$
- Alors Π défini par $\Pi(B)=\sup\{P(B) \mid P \in \mathcal{P}\}$ est une mesure de possibilité, et N défini par $N(B)=\inf\{P(B) \mid P \in \mathcal{P}\}$ est la mesure de nécessité duale.
- Réciproquement, toute mesure de possibilité (resp. nécessité) peut être vue comme l'enveloppe supérieure (resp. inférieure) d'une famille de lois de probabilités.

Interprétation 3 : vraisemblance

- Soit X = v.a. de loi $p_X(x;\theta)$.
- La fonction de vraisemblance normalisée

$$\pi(\theta) = \frac{L(\theta; x)}{\sup_{u \in \Theta} L(u; x)}$$

peut être interprétée comme une distribution de possibilité.

- Test du rapport de vraisemblance : hypothèse $\theta \in H$ rejetée si

$$\Pi(H) = \sup_{\theta \in H} \pi(\theta) < c$$

Fusion de distributions de possibilités

- Cas ensembliste :
  - Source S_1 : $\theta \in E_1$, $E_1 \subseteq \Theta$
  - Source S_2 : $\theta \in E_2$, $E_2 \subseteq \Theta$
  - Combinaison ?
    - Sources toutes deux fiables :

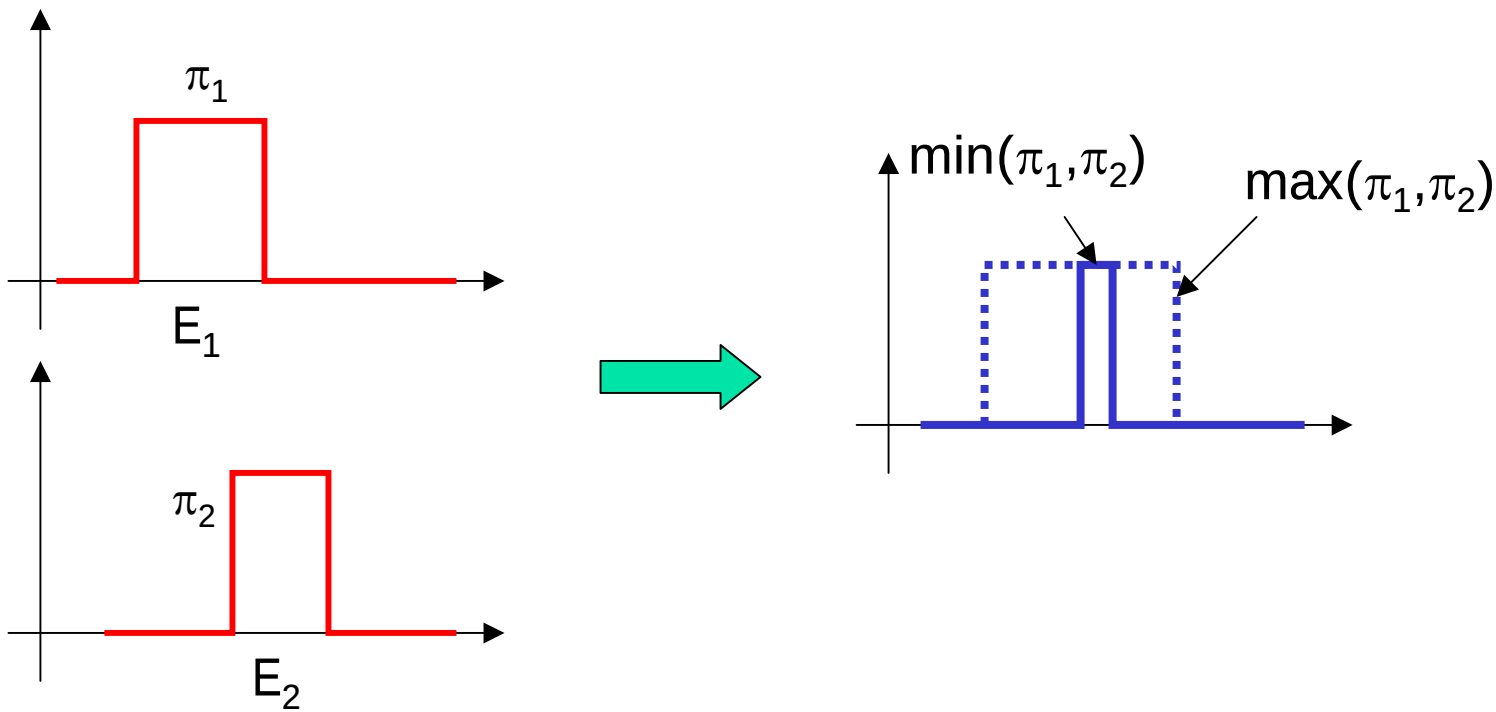
      $\theta \in E_1 \cap E_2$

      (fusion conjonctive)
    - L'une au moins des deux sources est fiable :

      $\theta \in E_1 \cup E_2$

      (fusion disjonctive)

Fusion de distributions de possibilité



Généralisation

- Soient π_i, i=1,...,n des distributions de possibilité fournies par n sources.
- On peut définir :

$$\pi_\wedge(u) = \top_{i=1,n}\pi_i(u) \text{ [intersection floue]}$$
$$\pi_\vee(u) = \perp_{i=1,n}\pi_i(u) \text{ [union floue]}$$

où $\top$ est une norme triangulaire, et $\perp$ la co-norme duale $\alpha \perp \beta = 1-(1-\alpha)\top(1-\beta)$

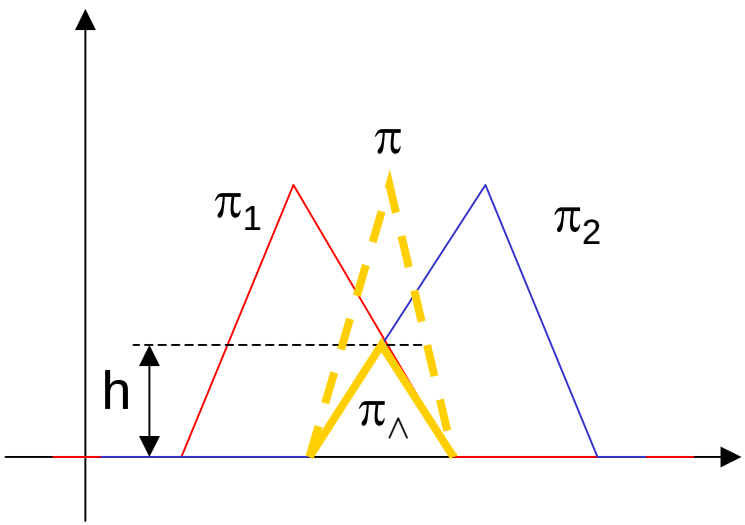
- Exemples de solutions :

$$\alpha \top \beta = \min(\alpha,\beta) \leftrightarrow \alpha \perp \beta = \max(\alpha,\beta)$$
$$\alpha \top \beta = \alpha\beta \leftrightarrow \alpha \perp \beta = \alpha + \beta - \alpha\beta$$
$$\alpha \top \beta = \max(0,\alpha+\beta-1) \leftrightarrow \alpha \perp \beta = \min(1,\alpha + \beta)$$

Critères de choix

- Fiabilité des sources (combinaison disjonctive plus prudent lorsque les sources ne sont pas toutes fiables)
- Dépendance entre les sources (opérateur min idempotent : ne nécessite pas l'hypothèse d'indépendance entre les sources, contrairement à l'opérateur produit)
- Conflit entre les sources : si deux sources ont un conflit important, il est vraisemblable que l'une au moins est erronée $\rightarrow$ opérateur disjonctif.

Fusion conjonctive normalisée



Degré de recouvrement entre π_1 et π_2 :

$h(\pi_1,\pi_2) = \sup_{u\in\Theta} \pi_1(u) \top \pi_2(u)$

Fusion conjonctive normalisée :

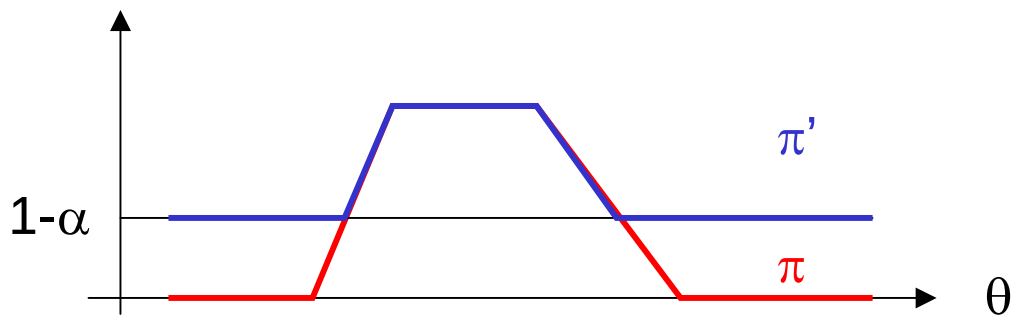
$$\pi(u) = \frac{\pi_\wedge(u)}{h(\pi_1,\pi_2)}$$

Prise en compte de la fiabilité des sources

- Soit π une distr. de possibilité fournie par une source, et α le degré de certitude que la source soit fiable.
- Affaiblissement de π :

$$\pi' = \max(\pi, 1-\alpha)$$

  - $\alpha=1$: $\pi'=\pi$
  - $\alpha=0$: $\pi'(u)=1, \forall$ u (ignorance totale)



Combinaison de sources de fiabilités variables

- n sources fournissent π_i, i=1,...,n
- Coefficients de fiabilités α_i, i=1,...,n

$$\pi_{\wedge\alpha}(u) = \min_{i=1,n} \max(\pi_i(u), 1-\alpha_i)$$
$$\pi_{\vee\alpha}(u) = \max_{i=1,n} \min(\alpha_i, \pi_i(u))$$

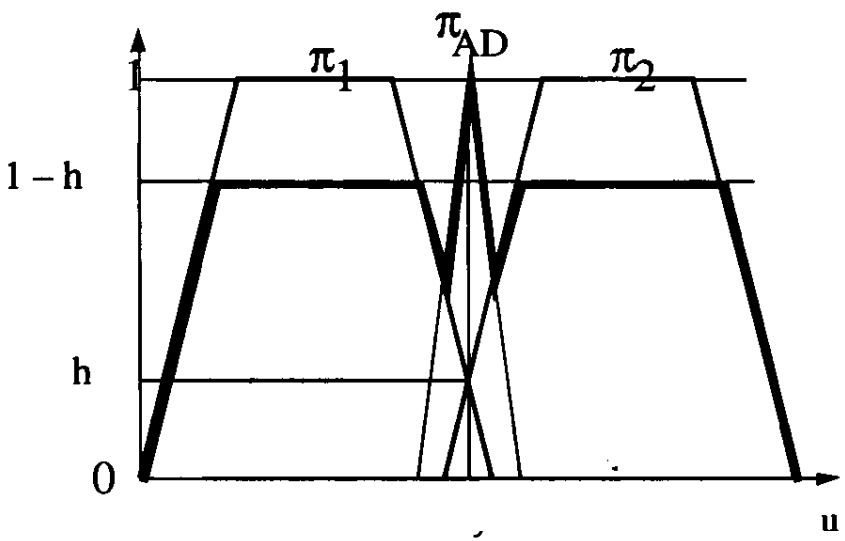
- Généralisation
  - Affaiblissement : $\pi' = \alpha \top \pi + 1-\alpha$
  - Combinaison :

$$\pi_{\wedge\alpha}(u) = \min_{i=1,n} \alpha_i \top \pi_i(u) + 1 - \alpha_i$$
$$\pi_{\vee\alpha}(u) = \max_{i=1,n} \alpha_i - \alpha_i \top (1-\pi_i(u))$$

Fusion adaptative

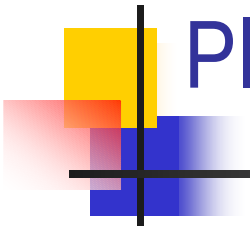
- Idée : adapter le caractère conjonctif ou disjonctif de la combinaison en fonction du conflit

$$\pi_{AD}(u) = \max\left(\frac{\min(\pi_1(u), \pi_2(u))}{h(\pi_1, \pi_2)}, \min(\max(\pi_1(u), \pi_2(u)), 1 - h(\pi_1, \pi_2))\right)$$



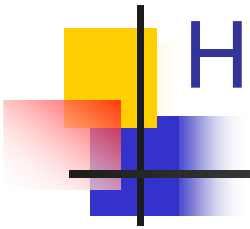
Conclusion sur l'approche possibiliste

- Cadre plus souple que le cadre probabiliste, mieux adapté à la représentation d'informations imprécises.
- Nombreux travaux en fusion d'information :
  - multiplicité d'opérateurs de fusion,
  - choix empirique lié aux caractéristiques de l'application.
- Capacités d'expression limitées : impossibilité de représenter une incertitude probabiliste (ex : couleur d'une boule devant être extraite d'une urne de composition connue).
- Intérêt d'un formalisme plus général englobant les probabilités et les possibilités : fonctions de croyance.



Plan

1. Problématique de la fusion d'informations
2. Approches « probabilistes »
   - Statistique classique
   - Cadre Bayésien
3. Théorie des possibilités
4. Théorie des fonctions de croyance
5. Applications en classification
6. Conclusions et perspectives



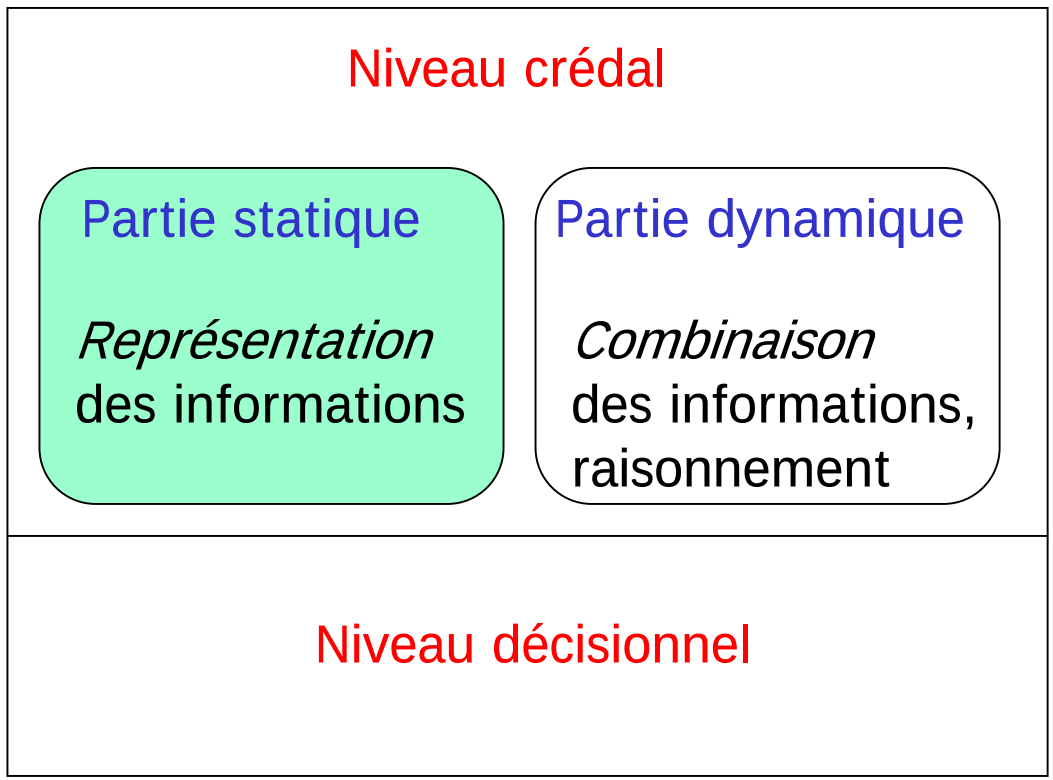
Historique

- 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

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

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

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

$$m_S^{\Theta}\{\theta\}[BC]$$

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

Exemple

- Morceau de journal annonçant une tempête pour le lendemain.
  - Je suis sûr à 75% que le journal est d'aujourd'hui.
  - Si le journal est d'aujourd'hui, l'information disponible accrédite l'hypothèse qu'il y aura une tempête.
  - Si le journal n'est pas d'aujourd'hui, l'élément d'évidence n'accrédite aucune hypothèse particulière

$$\Theta=\{T,NT\}$$

$$m(\{T\})=0.75 \qquad m(\Theta)=0.25$$

- Ignorance totale : $m(\Theta)=1$ (fonction de masse vide)

Fonction de Croyance

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

- bel(A) = degré de croyance en l'hypothèse « $\theta \in$ 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

$$bel\left(\bigcup_{i=1}^{n} A_i\right) \geq \sum_i bel(A_i) - \sum_{i>j} bel(A_i \cap A_j) \ldots$$

$$- (-1)^n bel\left(\bigcap_{i=1}^{n} A_i\right), \quad \forall A_1, \ldots, A_n \subseteq \Theta$$

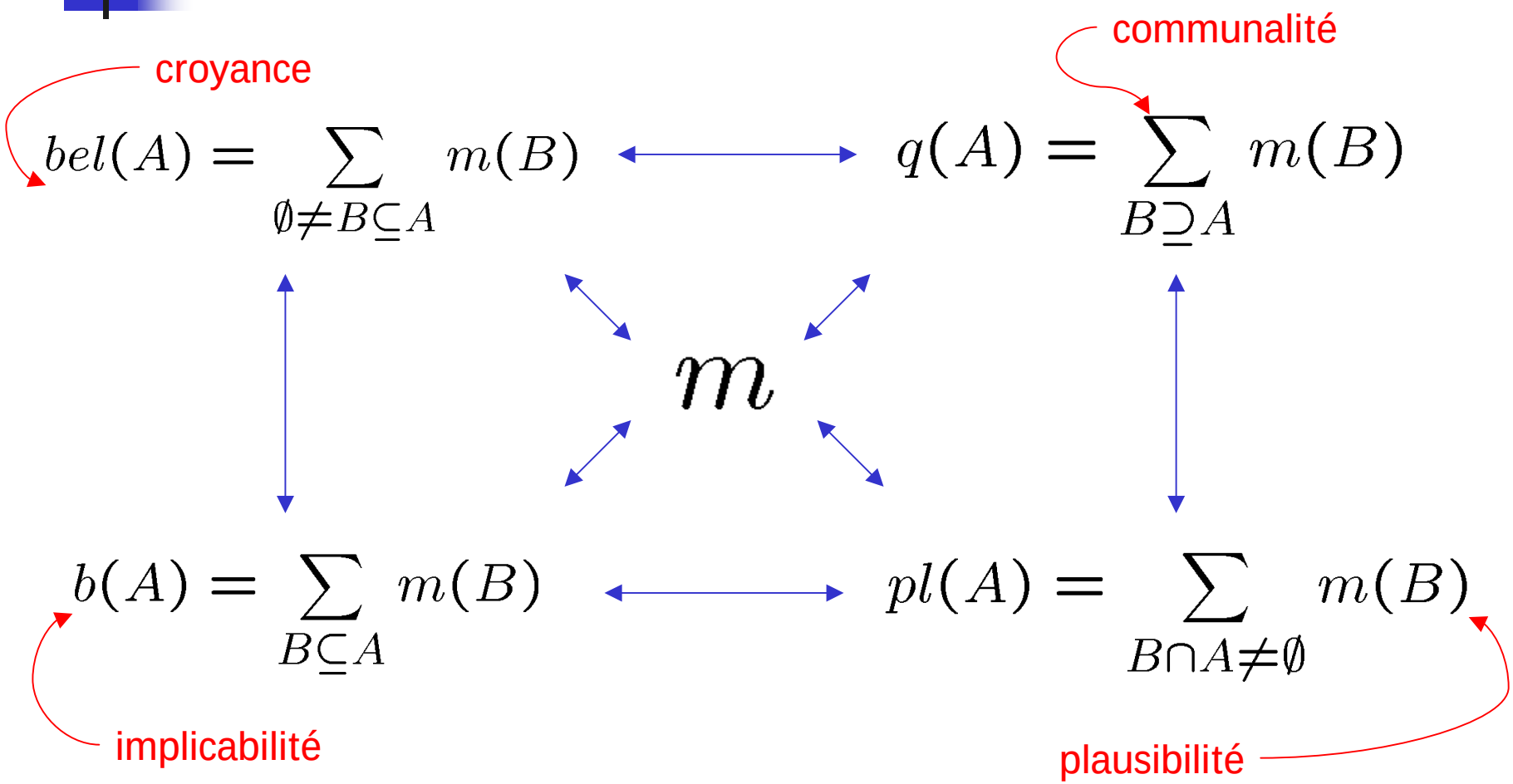
Fonction de plausibilité

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

- pl(A) = degré maximal de croyance pouvant potentiellement être attribué à l'hypothèse « θ ∈ A » (conditionnellement à l'obtention de nouvelle informations).
- Propriétés:
  - bel(A) ≤ pl(A), ∀ A ⊆ Θ
  - mesure sous-additive

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

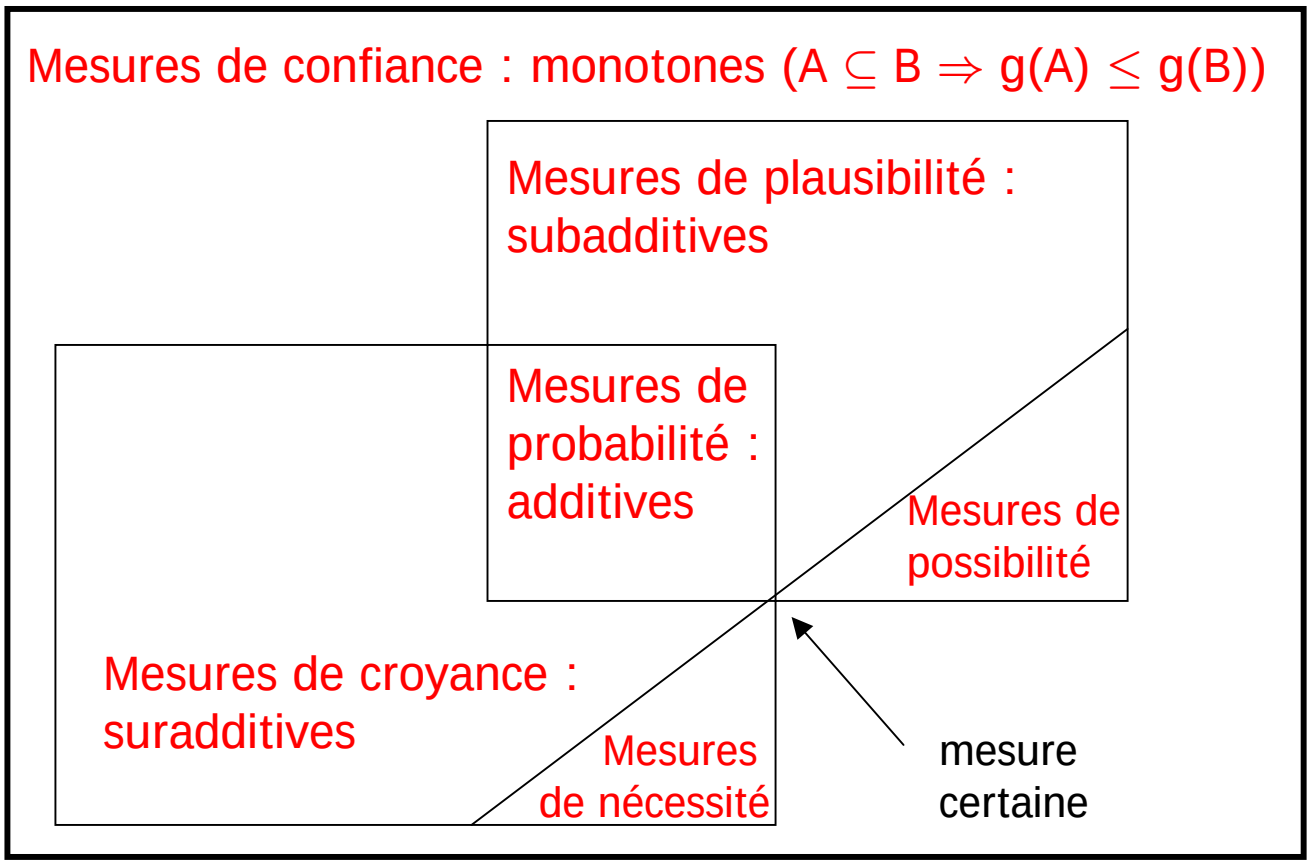
Représentations équivalentes



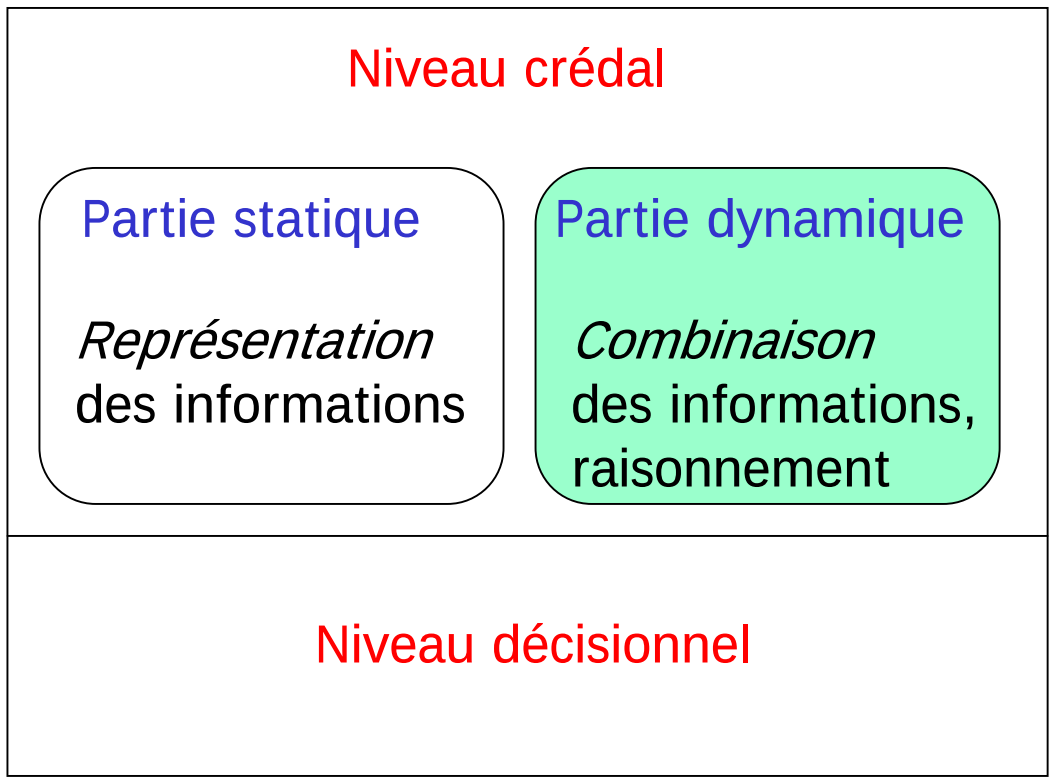
Cas particuliers

- Ensembles focaux singletons : $m(A)>0 \Rightarrow |A|=1$
  - Bel=pl, mesure de probabilité
  - Probabilité = fonction de croyance maximalement précise.
- Ensembles focaux emboîtés $A_1 \subset ... \subset A_n$
  - $pl(A \cup B)=\max(pl(A),pl(B))$: pl est une mesure de possibilité
  - Bel est la mesure de nécessité duale
  - Mesure de possibilité = fonction de plausibilité consonante (absence de « conflit interne »)
- Le formalisme des fonctions de croyance englobe donc la théorie des probabilités et la théorie des possibilités comme cas particuliers.

Relations d'inclusion entre types de mesures de confiance



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 :

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

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

- Propriétés :
  - commutative,
  - associative,
  - élément neutre $m(\Theta)=1$,
  - 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 $\theta \notin \Omega$ (hypothèse du monde ouvert).
- Si Θ exhaustif (hypothèse du monde clos) : on impose m(∅)=0.
- Règle de Dempster = somme conjonctive puis normalisation.

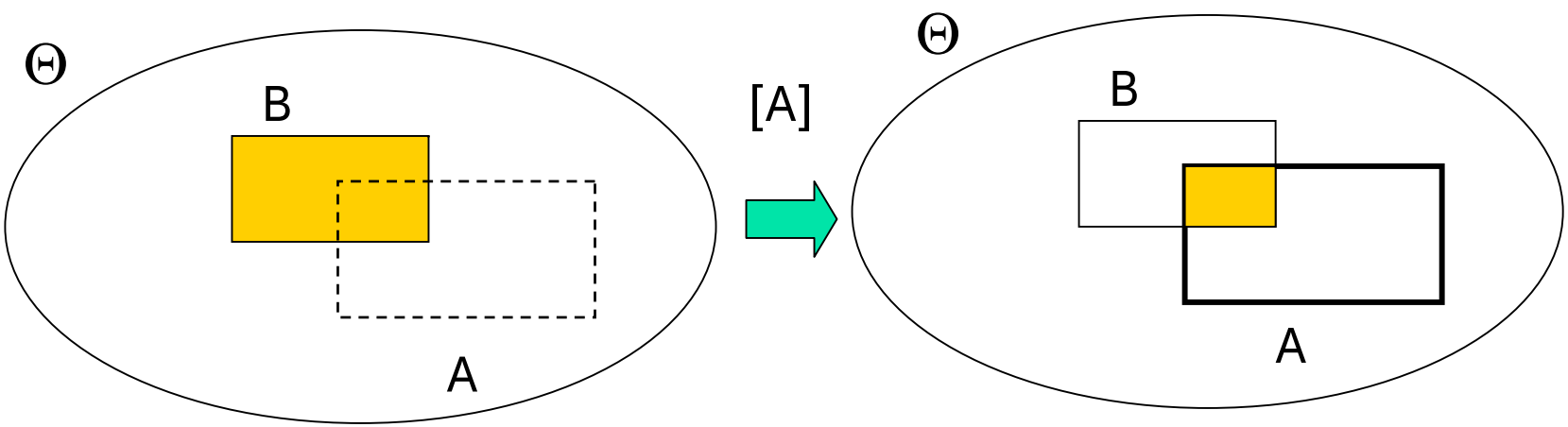
$$(m_1 \oplus m_2)(A) = \frac{(m_1 \textcircled{\cap} m_2)(A)}{1 - (m_1 \textcircled{\cap} m_2)(\emptyset)} \quad \forall A \neq \emptyset$$

Règle de conditionnement de Dempster

- Cas particulier de la somme conjonctive.

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

- 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 :

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

Autres règles de combinaison

- Règle de Yager

$$(m_1 Y m_2)(C) = \sum_{A \cap B = C} m_1(A) m_2(B), \quad \forall C \in 2^{\Theta} \setminus \{\emptyset, \Theta\}$$

$$(m_1 Y m_2)(\Theta) = \sum_{A \cap B = \Theta} m_1(A) m_2(B) + \sum_{A \cap B = \emptyset} m_1(A) m_2(B)$$

- Règle de Dubois et Prade

$$(m_1 D m_2)(C) = \sum_{A \cap B = C} m_1(A) m_2(B) + \sum_{A \cup B = C, A \cap B = \emptyset} m_1(A) m_2(B), \quad \forall C \in 2^{\Theta} \setminus \{\emptyset\}$$

- Règles non associatives

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

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

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

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

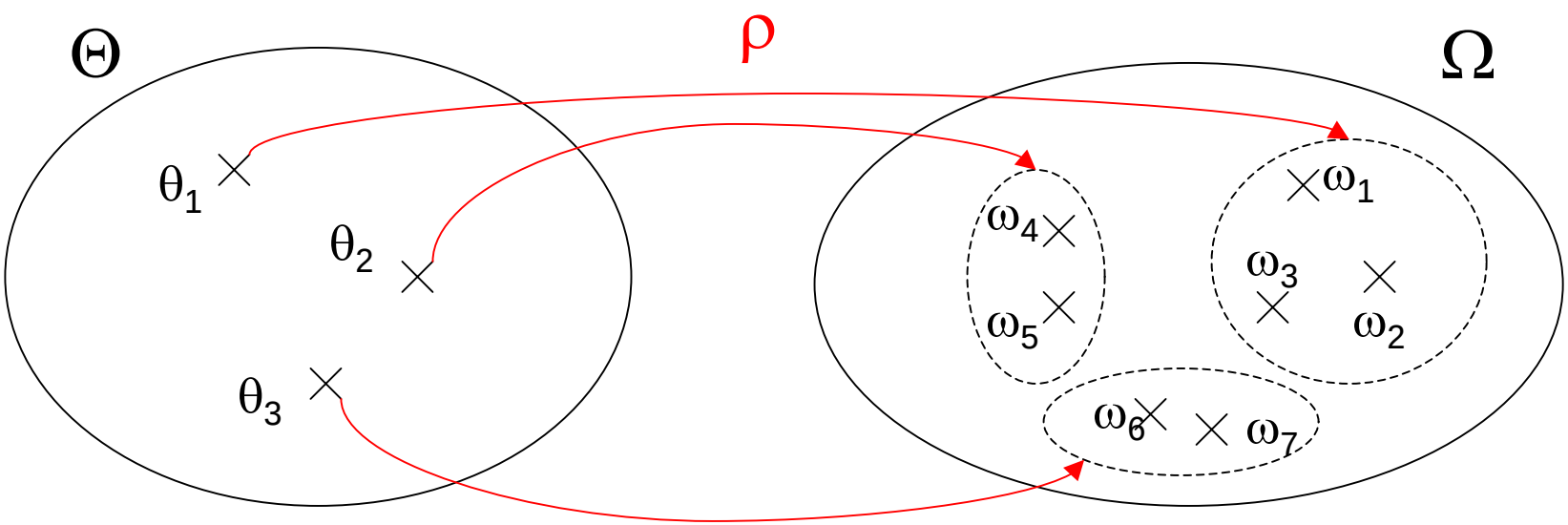
Principe d'information minimum

- 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.

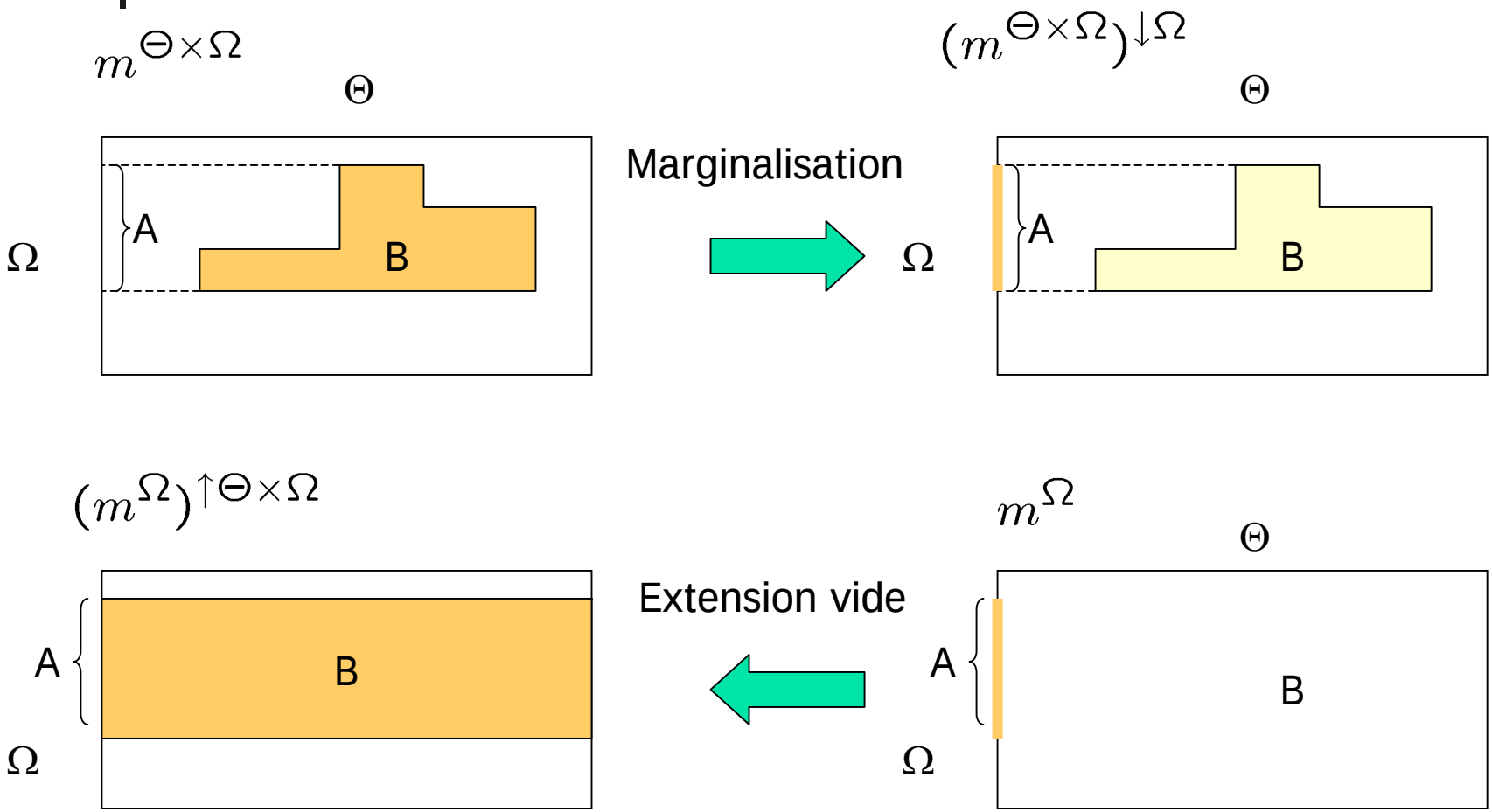
- Joue le même rôle que le *principe du maximum d'entropie* en probabilités.
- Applications :
  - extension vide
  - déconditionnement.

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

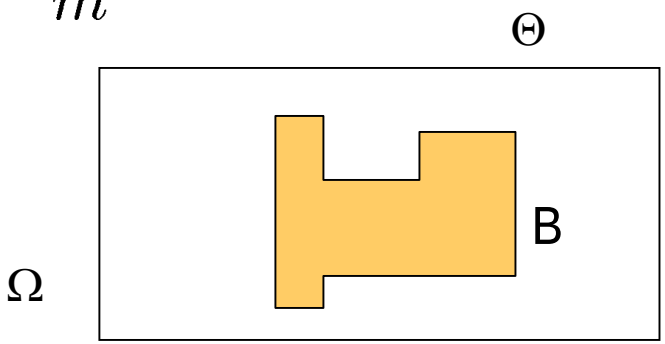
- Supposons que l'on connaisse $m^{\Theta}[A]$ pour $A \subseteq \Theta$ (état de connaissance sur $\theta \in \Theta$, dans un contexte où l'on sait que $\theta \in A$).
- Comment en déduire une fonction de masse non conditionnelle sur Θ ?
- Approche : recherche de la fonction de masse la moins informative,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

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

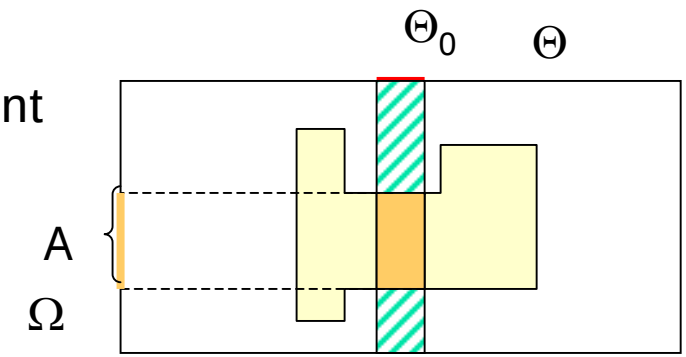
Cas d'un espace produit

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

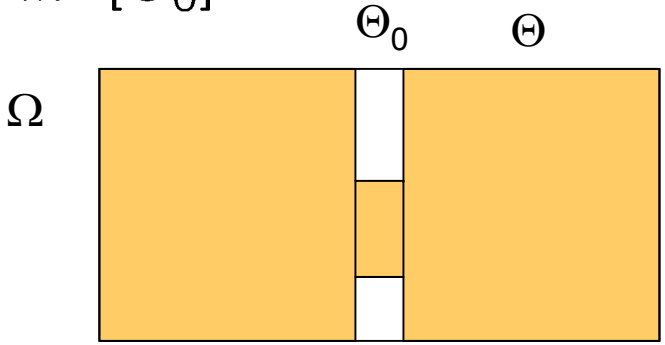
$m^{\Theta\times\Omega}$



Conditionnement

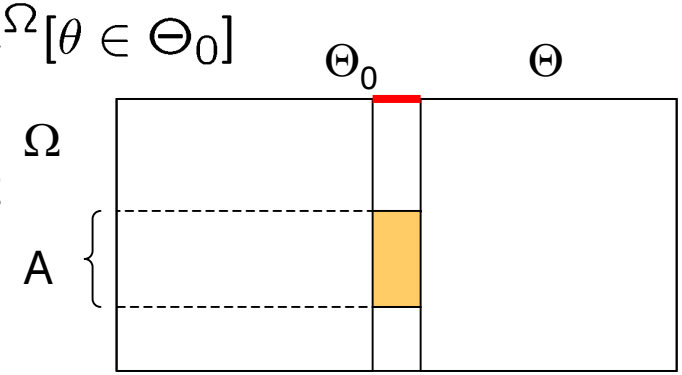


$m^{\Omega}[\Theta_0]^{\Uparrow\Theta\times\Omega}$



Déconditionnement

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



Affaiblissement

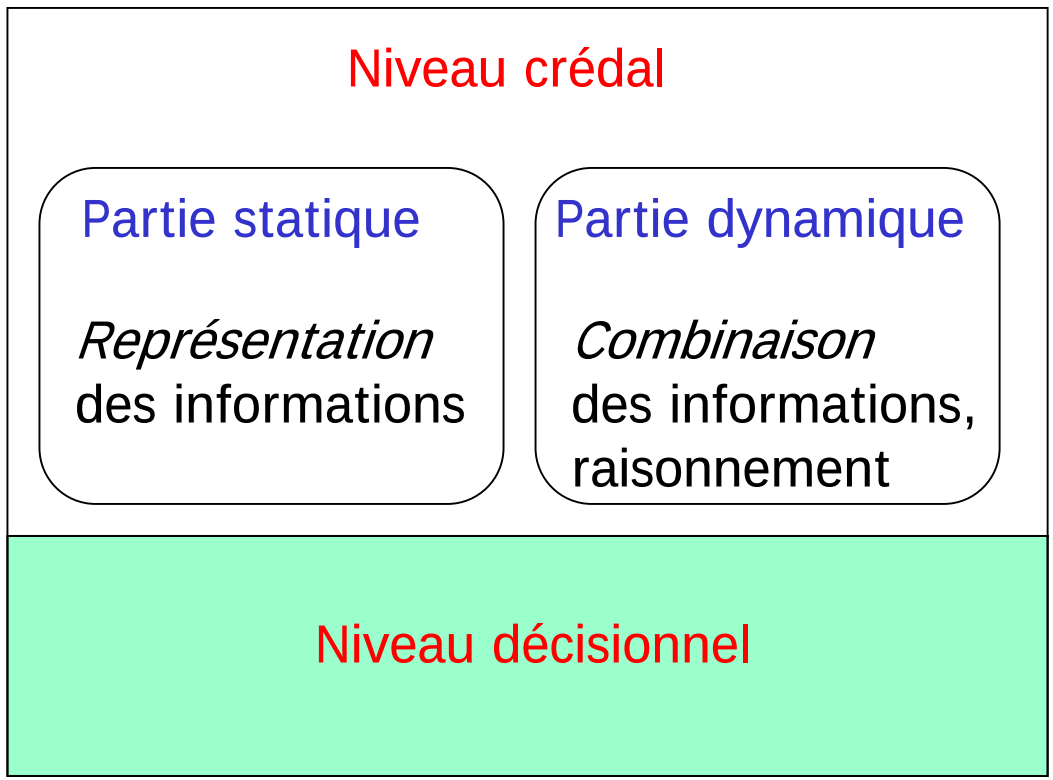
- m^{Θ}_{S} fonction de masse traduisant l'information sur une variable θ produite par une source S.
- Fiabilité de S inconnue : $\mathcal{R}$={F,NF}.
- On suppose :
  - $m^{\Theta}[F] = m^{\Theta}_{S}$
  - $m^{\Theta}[NF](\Theta)=1$.
  - croyances sur $\mathcal{R}$: $m^{\mathcal{R}}(NF) = \alpha$, $m^{\mathcal{R}}(F)=1-\alpha$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

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

- Résultat : Affaiblissement de m

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

Le Modèle des Croyances Transférables



Décision

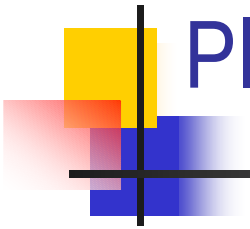
- Soient $\mathcal{A}$ un ensemble d'actions, C : $\mathcal{A} \times \Theta \rightarrow \mathbb{R}$ une fonction de coûts, et m une fonction de masse sur Θ. Choix d'une action ?
- MCT :
  - Calcul de la probabilité pignistique

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

  - Choix de l'action $\alpha \in \mathcal{A}$ qui minimise le « risque pignistique »

$$R_{BetP}(\alpha) = \sum_{\theta \in \Theta} C(\alpha, \theta) BetP(\theta)$$

- $\mathcal{A} = \Theta$ et $C(\theta_i,\theta_j)=1-\delta_{ij}$: règle du « maximum de probabilité pignistique ».



Plan

1. Problématique de la fusion d'informations
2. Approches « probabilistes »
   - Statistique classique
   - Cadre Bayésien
3. Théorie des possibilités
4. Théorie des fonctions de croyance
5. Applications en classification
6. 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 complète ;
  - combinaison d'informations issues de différentes sources.
- Application en classification :
  - problèmes pour lesquels l'information disponible est incomplète, parcellaire :
    - 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.

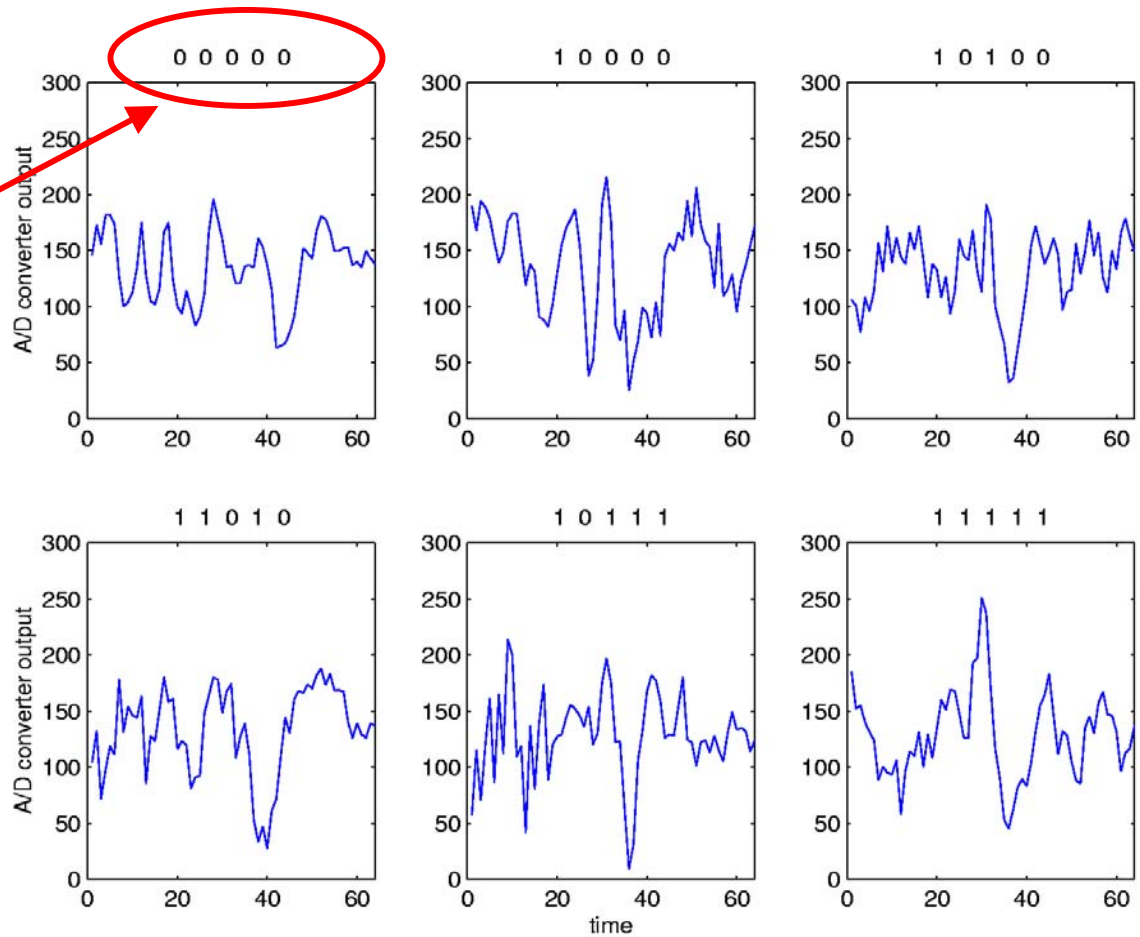
Exemple : classification de signaux EEG

- 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

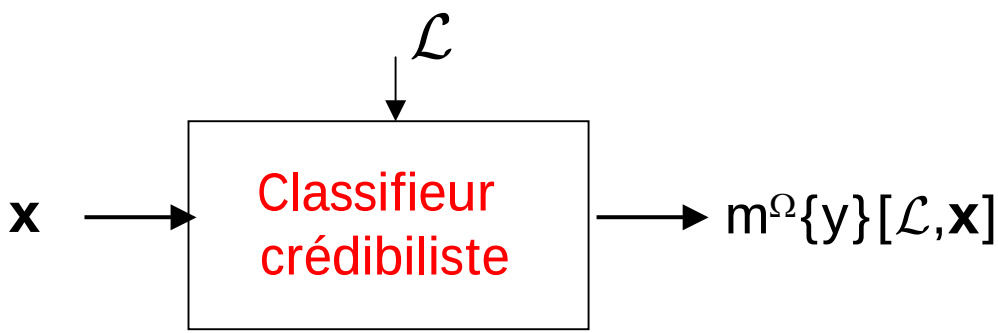


Fusion multi-capteurs

- 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 ...

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 $y \in \Omega=\{\omega_1,\ldots,\omega_K\}$.
  - ensemble d'apprentissage $\mathcal{L}$ (observations, éventuellement partielles, des variables **x** et y pour n objets).
  - Problème : prédire y sachant **x**, pour un nouvel objet.



- Deux approches :
  - Théorème de Bayes Généralisé;
  - Approche à base de cas

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 discrimination (Appriou, 1991)

- Pb : détermination des $pl^X[\omega_k]$, k=1,...,K :
  - Cas 1 : densités des classes connues

$$pl^X[\omega_k](\mathbf{x}) = \rho \cdot L(\omega_k; \mathbf{x}) = \rho \cdot p(\mathbf{x}|\omega_k)$$

  - Cas 2 : densités estimées

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

- On a alors :

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

Application en discrimination (suite)

- Remarques :
  - Les coefficients α_k peuvent être fixés a priori ou appris à partir des données par minimisation d'une fonction d'erreur.
  - La formule précédente ne nécessite pas d'information a priori sur les classes. Si une telle information 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 $p(\mathbf{x}|\omega_k)$ et les probabilités a priori $P(\omega_k)$ sont connus.
  - Dans le cas de deux vecteurs d'attributs **x** et **x**' indépendants :

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

Exemple (Appriou, 1991)

- Problème de reconnaissance de cibles : 2 classes $\Omega=\{\omega_1,\omega_2\}$ (avion, missile) et 2 capteurs S_1 et S_2 (radar et infrarouge).
- Chaque capteur S_j fournit un attribut x_j
- Distributions de x_1 et x_2 dans chaque classe, dans des conditions expérimentales contrôlées :

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

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

- Hypothèse d'équiprobabilité : $P(\omega_1)=P(\omega_2)$

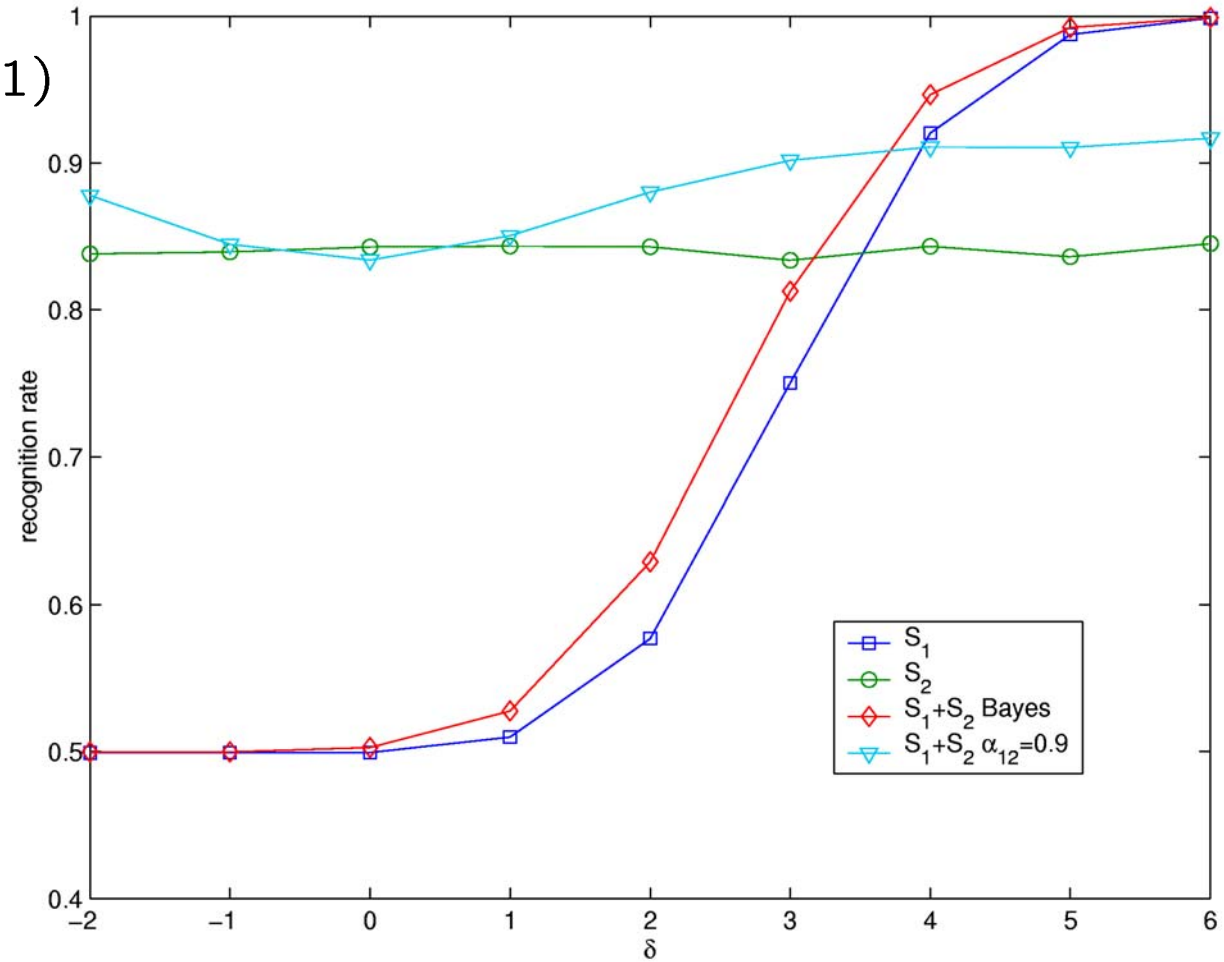
Exemple (suite)

- 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

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

Résultats

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



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 :

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

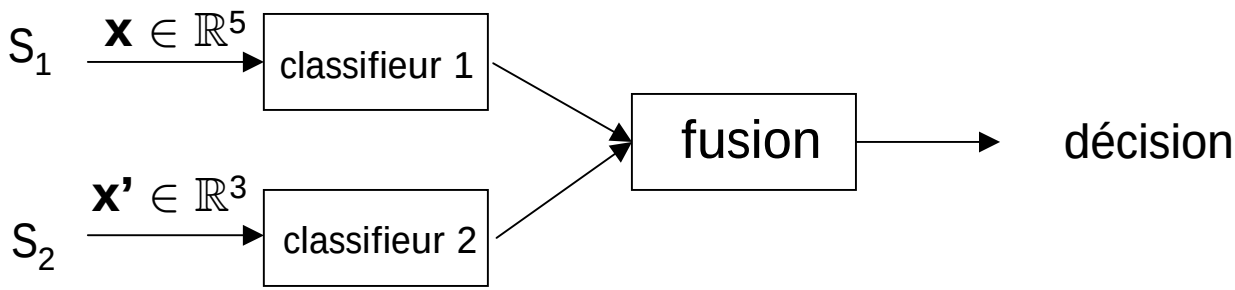
- Mise en œuvre de la méthode :
  - optimisation de la fonction d'affaiblissement ϕ par apprentissage ;
  - prise en compte uniquement des k plus proches voisins de **x** dans $\mathcal{L}$ (ou des vecteurs $\mathbf{x}_i$ t.q. $\delta(\mathbf{x},\mathbf{x}_i) \leq \delta_{min}$) ;
  - synthèse de $\mathcal{L}$ 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

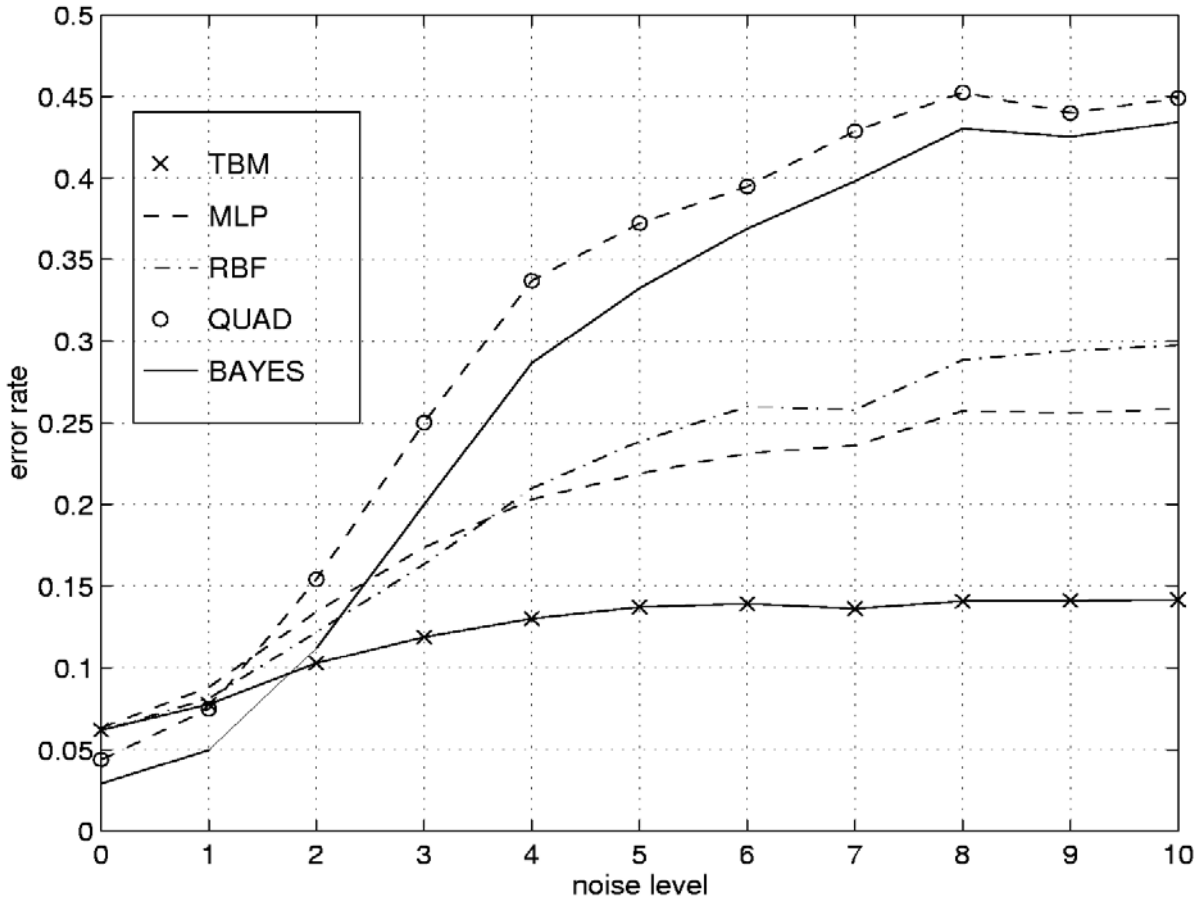


- 2 classes
- **x** et **x**' gaussiens, conditionnellement indépendants
- apprentissage : n=60, validation croisée : n_{cv}=100
- test : n_t=5000

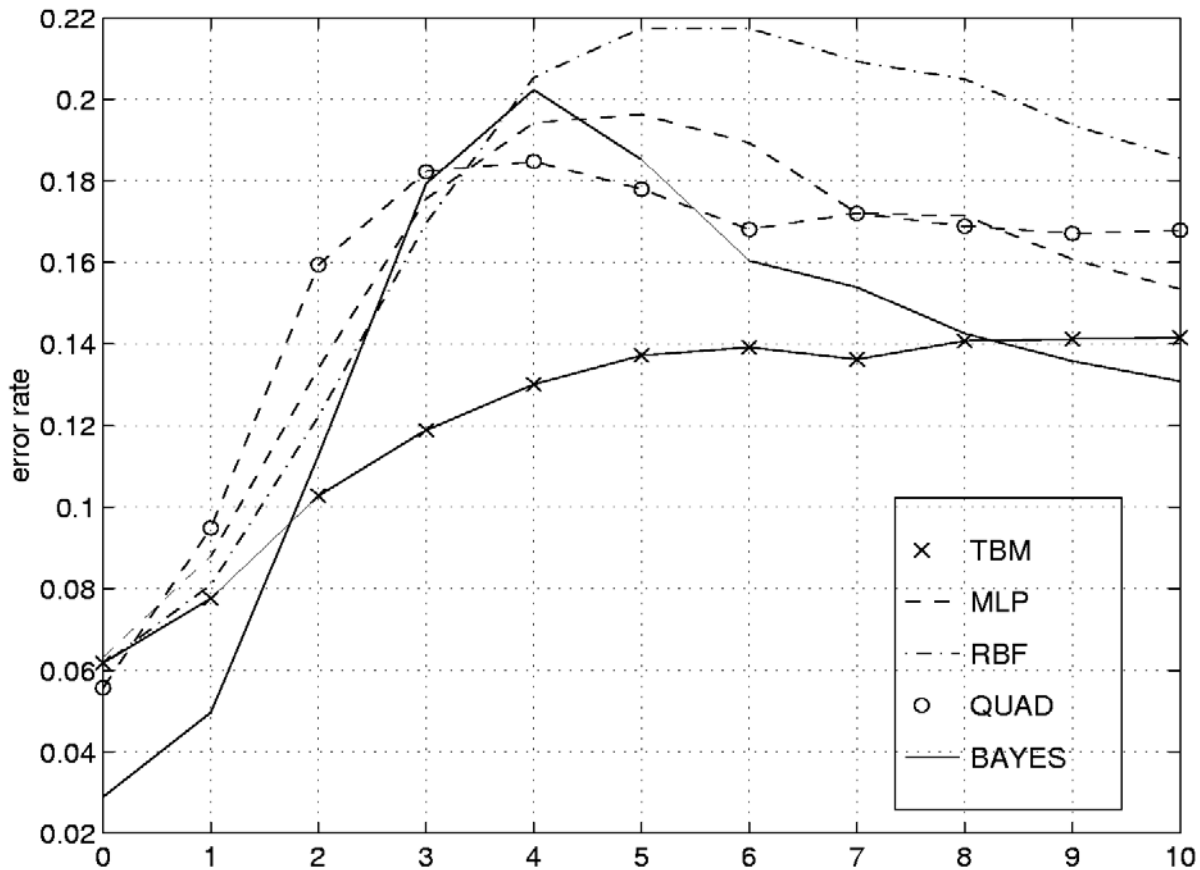
Résultats : test $\sim$ apprentissage

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



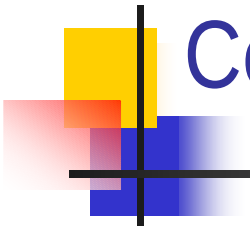
Test : $\mathbf{x} \leftarrow \mathbf{x} + \mathcal{N}(0,\sigma^2)$ et rejet



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



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.