# Théorie des fonctions de croyance: application en reconnaissance de formes et en fusion d'informations

Thierry Denœux

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

SCATI : Théorie et applications des fonctions de croyance pour les systèmes de vision
3 décembre 2010

# Qu'est-ce que la TFC ?
## Historique

- La théorie des fonctions de croyance (TFC) : un cadre formel pour le calcul et le raisonnement à partir d'informations partielles (incertaines, imprécises).
- Autres dénominations : théorie de Dempster-Shafer, théorie de l'évidence, Modèle des Croyances Transférables.
- Introduit par Dempster (1968) et Shafer (1976), développé par Smets dans les années 1980 et 1990.

# Qu'est-ce que la TFC ?

## Relation avec les approches classiques

- La théorie des fonctions de croyance étend à la fois les approches ensembliste and probabiliste de représentation des incertitudes.
- Une fonction de croyance peut être vue comme
  - un ensemble généralisé et
  - une mesure non additive ;
- Extension de notions
  - probabilistes (conditionnement, marginalisation) et
  - ensemblistes (intersection, union, inclusion, etc.).

# Qu’est-ce que la TFC ?

## Applications

- Inférence statistique (Dempster, 1968) ;
- Systèmes experts (1980-) ;
- Fusion d’informations (1990-) ;
- Classification, apprentissage, reconnaissance de formes (2000-).

# Plan

1. Théorie des fonctions de croyance
   - Concepts fondamentaux
   - Règle de Dempster
   - Règle prudente et extensions
2. Applications
   - Classification
   - Classification multi-label
   - Fusion adaptative de classifieurs

# Outline

# Fonction de masse
## Définition

- Soit $\Omega$ un ensemble fini de réponses possibles à une certaine question : cadre de discernement.
- Une fonction de masse (normalisée) sur $\Omega$ est une fonction $m : 2^{\Omega} \rightarrow [0, 1]$ telle que $m(\emptyset) = 0$ et

$$\sum_{A \subseteq \Omega} m(A) = 1.$$

- Les parties $A$ de $\Omega$ t.q. $m(A) > 0$ sont appelées éléments focaux de $m$.

# Fonction de masse
## Interprétation

- Une fonction de masse $m$ modélise un élément d'évidence sur une variable $X$ à valeur dans $\Omega$.
- $m(A)$ s'interprète comme une mesure de la croyance allouée exactement à l'hypothèse $X \in A$, et à aucune hypothèse plus spécifique.

# Fonction de masse

## Cas particuliers

- Fonction de masse catégorique :

$$m(A) = 1 \text{ for some } A \subseteq \Omega$$

  $\longrightarrow$ équivalente à un ensemble. La fonction de masse vide, correspondant à $A = \Omega$, représente l'ignorance totale.

- Fonction de masse bayésienne :

$$m(A) > 0 \Rightarrow |A| = 1$$

  $\longrightarrow$ équivalent à une distribution de probabilité.

- Dans le cas général, une fonction de masse peut donc être vue comme :
  - un ensemble généralisé ;
  - une distribution de probabilité généralisée.

# Exemple

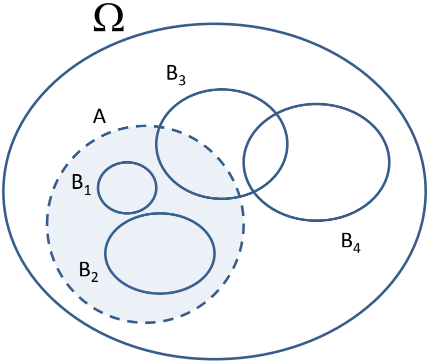
- Un meurtre a été commis. Il y a 3 suspects : $\Omega = \{Pierre, Jean, Marie\}$.
- Un témoin a vu le meurtrier s'enfuir, mais il est myope et peut seulement attester que c'était un homme. On sait que le témoin est ivre 20 % du temps.
- Représentation de l'élément d'évidence :

$$m(\{Pierre, Jean\}) = 0.8,$$

$$m(\Omega) = 0.2$$

- La masse 0.2 n'est pas allouée à $\{Marie\}$, car le témoignage n'accuse absolument pas Marie !

# Fonctions de croyance et de plausibilité
## Definitions




$$bel(A) = \sum_{\emptyset \neq B \subseteq A} m(B)$$

$$pl(A) = \sum_{B \cap A} m(B),$$

$$pl(A) \geq bel(A), \quad \forall A \subseteq \Omega.$$

# Fonctions de croyance et de plausibilité

## Interprétation et cas particuliers

- Interprétations :
  - $bel(A) =$ degré de croyance en $A$, justifié par l'élément d'évidence considéré.
  - $pl(A) =$ borne supérieure du degré de croyance susceptible d'être alloué à $A$ après prise ne compte de nouvelles informations.
- Cas particuliers :
  - Si $m$ est bayésienne, $bel = pl$ (mesure de probabilité).
  - Si les éléments focaux sont emboîtés, $pl$ est une mesure de possibilité, et $bel$ est la mesure de nécessité duale.

# Outline

1. Théorie des fonctions de croyance
   - Concepts fondamentaux
   - Règle de Dempster
   - Règle prudente et extensions
2. Applications
   - Classification
   - Classification multi-label
   - Fusion adaptative de classifieurs

# Définition

- Soient $m_1$ et $m_2$ deux fonctions de masse sur $\Omega$ modélisant deux éléments d'évidence indépendants. Comment les combiner ?
- Règle de Dempster :

$$(m_1 \oplus m_2)(A) = \begin{cases} \dfrac{1}{1-K} \displaystyle\sum_{B \cap C = A} m_1(B) m_2(C) & \text{si } A \neq \emptyset \\ 0 & \text{si } A = \emptyset \end{cases}$$

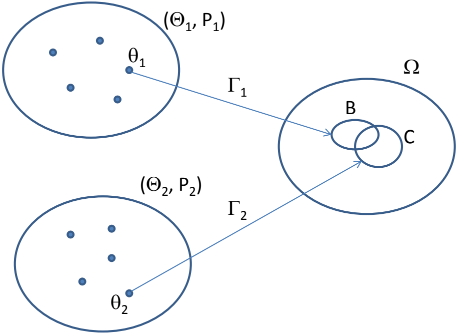
avec $K = \sum_{B \cap C = \emptyset} m_1(B) m_2(C)$ : degré de conflit.

# Exemple

- On a $m_1(\{Pierre, Jean\}) = 0.8$, $m_1(\Omega) = 0.2$.
- Nouvel élément d'évidence : le meutrier est blond, confiance=0.6 $\rightarrow$ $m_2(\{Jean, Marie\}) = 0.6$, $m_2(\Omega) = 0.4$.

| | $\{Pierre, Jean\}$ 0.8 | $\Omega$ 0.2 |
|---|---|---|
| $\{Jean, Marie\}$ 0.6 | $\{Jean\}$ 0.48 | $\{Jean, Marie\}$ 0.12 |
| $\Omega$ 0.4 | $\{Pierre, Jean\}$ 0.32 | $\Omega$ 0.08 |

# Justification




- Chaque $m_i$ induite par une mesure de probabilité $P_i$ sur un ensemble $\Theta_i$ (interprétations), et une application $\Gamma_i : \Theta_i \rightarrow 2^{\Omega}$.
- Si les interprétations $\theta_1$ et $\theta_2$ sont vraies, alors il faut croire $X \in \Gamma_1(\theta_1) \cap \Gamma_2(\theta_2)$.
- L'indépendance de $m_1$ et $m_2$ correspond à l'indépendance de $P_1$ et $P_2$.

# Propriétés

- Commutativité, associativité.
- Elément neutre : fonction de masse vide.
- Généralisation de l'intersection : si $m_A$ et $m_B$ sont des fonctions de masse catégoriques et $A \cap B \neq \emptyset$, alors

$$m_A \oplus m_B = m_{A \cap B}$$

- Généralisation du conditionnement probabiliste : si $m$ est bayésienne et $m_A$ catégorique, alors $m \oplus m_A$ est la fonction de masse bayésienne correspondant au conditionnement de $m$ par $A$.

# Décomposition canonique
## Fonction de masse simple

- Le plus souvent, un élément d'évidence se modélise sous forme d'une fonction de masse simple de la forme :

$$\begin{aligned} m(A) &= 1 - w \\ m(\Omega) &= w, \end{aligned}$$

avec $A \neq \emptyset$, $A \subset \Omega$ et $w \in [0, 1]$.

- Notation : $A^w$.
- Propriété : $A^{w_1} \oplus A^{w_2} = A^{w_1 w_2}$.

# Décomposition canonique

## Fonction de masse séparable

- Une fonction de masse (normalisée) est dite séparable si elle s'obtient comme résultat de la combinaison par la règle de Dempster de fonctions de masse simples :

$$m = \bigoplus_{A \subseteq \Omega} A^{w(A)}$$

- La fonction $w : A \rightarrow w(A)$ est une représentation alternative de $m$.
- Généralisation possible dans le cas d'une fonction de masse quelconque ($w$ est alors à valeurs dans $\mathbb{R}_+$).

# Outline

1. Théorie des fonctions de croyance
   - Concepts fondamentaux
   - Règle de Dempster
   - Règle prudente et extensions
2. Applications
   - Classification
   - Classification multi-label
   - Fusion adaptative de classifieurs

# Motivations

- La règle de Dempster suppose l'indépendance des sources d'information.
- Dans beaucoup de situations, les sources ne peuvent être considérées comme indépendantes :
  - Experts partageant certaines informations ;
  - Classifieurs entraînés sur des bases d'apprentissage identiques ou non disjointes ;
  - Informations échangées dans un réseau mobile (fusion distribuée), etc.
- Nécessité de disposer de règles de combinaison tolérant la dépendance et la redondance des informations combinées.

# Notions préliminaires
## Principe d'engagement minimal

- Principe d'engagement minimal :
  Lorsque plusieurs fonctions de croyance sont compatibles avec un ensemble de contraintes, la moins informative doit être choisie.
- Que signifie « plus ou moins informative » ?
- Plusieurs définitions :

$$m_1 \sqsubseteq_{pl} m_2 \Leftrightarrow pl_1 \leq pl_2$$

$$m_1 \sqsubseteq_{w} m_2 \Leftrightarrow w_1 \leq w_2$$

# Notions préliminaires

## Opérateurs conjonctifs et disjonctifs

- Soient $m_1$ et $m_2$ deux fonctions de masse sur $\Omega$ issues de 2 sources.
- Un opérateur de combinaison fusionne $m_1$ et $m_2$ en une nouvelle fonction de masse $m_{1*2} = m_1 * m_2$ (nouvel état de connaissance).
- L'opérateur $*$ est
  - conjonctif si $m_{1*2}$ est plus informative que $m_1$ et $m_2$.
  - disjonctif si $m_{1*2}$ est moins informative que $m_1$ et $m_2$.
- Le choix d'un opérateur conjonctif suppose que les sources soient toutes deux fiables.
- Combinaison disjonctive : stratégie prudente (l'une au moins des deux sources est fiable).

# Règle conjonctive prudente
## Principe

- Un agent reçoit deux fonctions de masse $m_1$ et $m_2$ issues de sources fiables.
- Soit $m_{12}$ la fct de masse représentant son état de connaissance après réception de $m_1$ et $m_2$.
- $m_{12}$ doit être plus informative que $m_1$ et $m_2$. Formellement : $m_{12} \in \mathcal{S}(m_1) \cap \mathcal{S}(m_2)$, avec $\mathcal{S}(m)$= ens. des fcts de masse plus informatives que $m$ (au sens d'un certain ordre).
- Principe d'engagement minimal : on choisit dans $\mathcal{S}(m_1) \cap \mathcal{S}(m_2)$ la fonction de masse la moins informative (si elle existe).

# Règle conjonctive prudente
## Définition

- Il est possible de définir une règle conjonctive prudente en prenant comme relation d'inclusion la relation $\sqsubseteq_w$.
- On obtient alors l'opérateur suivant :

$$m_1 \textcircled{\wedge} m_2 = \bigoplus_{A \subset \Omega} A^{w_1(A) \wedge w_2(A)},$$

avec $\wedge$ = minimum.

# Règle prudente
## Propriétés

- Commutative, associative
- Idempotente : $\forall m,\ m \owedge m = m$
- Distributivité de $\oplus$ par rapport à $\owedge$ : $\forall m_1, m_2, m_3$

$$(m_1 \oplus m_2) \owedge (m_1 \oplus m_3) = m_1 \oplus (m_2 \owedge m_3).$$

L'élément d'évidence $m_1$ n'est compté qu'une fois !

- Absence d'élément neutre, mais $m_\Omega \owedge m = m$ ssi $m$ est séparable.

# Extensions

- La règle de Dempster et la règle prudente s'écrivent sous la forme :

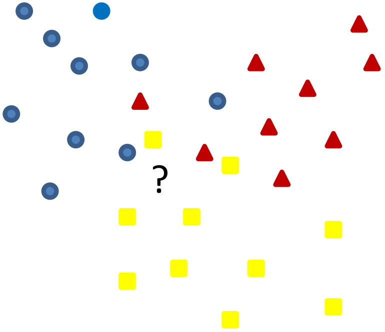
$$m_1 \circledast m_2 = \bigoplus_{A \subset \Omega} A^{w_1(A) \star w_2(A)},$$

avec $\star =$ produit (Dempster) ou minimum (règle prudente).

- En se restreignant aux fonctions de masse séparables, on peut définir une infinité de règles aux comportements intermédiaires entre ceux de $\oplus$ et $\owedge$, en prenant comme opérateur $\star$ une norme triangulaire (famille des t-normes de Frank par exemple).
- Généralisation possible au cas non séparable (Pichon et Denœux, 2009).

# Outline

# Problème

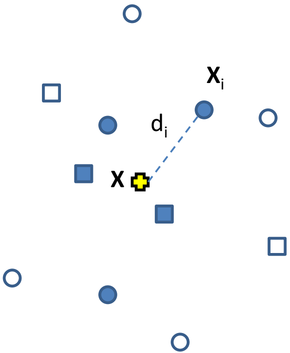



- Soit une population partitionnée en $c$ groupes ou classes.
- Soit $\Omega = \{\omega_1, \ldots, \omega_c\}$ l'ensemble des classes.
- Chaque individu est décrit par :
  - Un vecteur d'attributs $\mathbf{x} \in \mathbb{R}^p$ ;
  - Une étiquette de classe $y \in \Omega$.
- Problème : étant donné un ensemble d'apprentissage $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, prédire la classe d'un nouvel individu décrit par $\mathbf{x}$.

# Approches crédibilistes

1. Approche 1 : exprimer les sorties de classifieurs standards sous forme de fonctions de croyance et les combiner par la règle de Dempster ou une autre règle ;
2. Approche 2 : construction de classifieurs crédibilistes basés sur
   - le théorème de Bayes généralisé (extension du théorème de Bayes permettant d'exprimer une connaissance partielle sur les distributions des classes et les probabilités a priori) (Denœux et Smets, *IEEE SMC*, 2008) ;
   - les distances à des voisins ou à des prototypes : règle des *k*-ppv crédibiliste (Denœux, *IEEE SMC*, 1995), réseaux de neurones crédibilistes (Denœux, *IEEE SMC*, 2000).

# Règle des $k$-ppv crédibiliste (1/2)




- Soit $\mathcal{N}_k(\mathbf{x}) \subset \mathcal{L}$ l'ensemble des $k$ plus proches voisins de $\mathbf{x}$ dans $\mathcal{L}$, pour une certaine distance.
- Chaque $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$ peut être considéré comme un élément d'évidence concernant la classe de $\mathbf{x}$.
- La fiabilité de cet élément d'évidence décroît avec la distance $d_i$ entre $\mathbf{x}$ et $\mathbf{x}_i$.

# Règle des *k*-ppv crédibiliste (2/2)

- L'élément d'évidence $(\mathbf{x}_i, y_i)$ peut être représenté par

$$m_i(\{y_i\}) = \varphi(d_i)$$

$$m_i(\Omega) = 1 - \varphi(d_i),$$

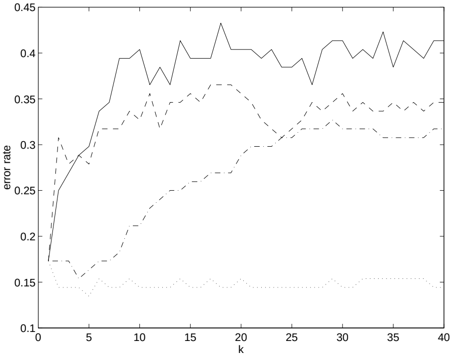
  où $\varphi$ est une fonction décroissante de $[0, +\infty)$ dans $[0, 1]$ telle que $\lim_{d \to +\infty} \varphi(d) = 0$.
- Les éléments d'évidence correspondant aux $k$ ppv de $\mathbf{x}$ sont combinés par la règle de Dempster :

$$m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} m_i.$$

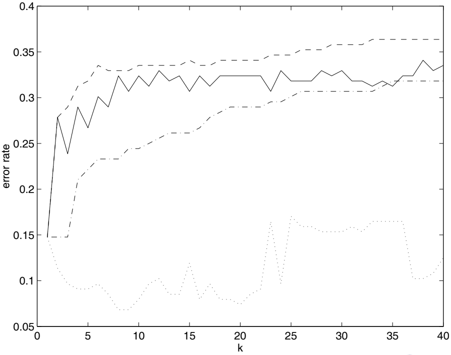
- La fonction $\varphi$ peut être déterminée de manière heuristique ou sélectionnée parmi une famille $\{\varphi_\theta | \theta \in \Theta\}$ par validation croisée.

# Résultats (bases UCI)

Données Sonar




Donées Ionosphere




Taux d'erreur de test en fonction de *k* pour les règles des *k* ppv avec vote (-), crédibiliste ( :), floue (–) et pondérée (-.).

# Données partiellement supervisées

- Soit un ensemble d'apprentissage de la forme

$$\mathcal{L} = \{(\mathbf{x}_i, m_i), i = 1, \ldots, n\}$$

avec
  - $\mathbf{x}_i$ le vecteur d'attributs pour l'exemple $i$, et
  - $m_i$ une fonction de masse représentant une connaissance partielle sur la classe $y_i$ de l'exemple $i$.
- Cas particuliers :
  - $m_i(\{\omega_k\}) = 1$ pour tout $i$ : apprentissage supervisé ;
  - $m_i(\Omega) = 1$ pour tout $i$ : apprentissage non supervisé ;

# Règle des $k$-ppv crédibiliste
## Cas de données partiellement supervisées

- Chaque exemple $(\mathbf{x}_i, m_i)$ dans $\mathcal{L}$ représente un élément d'évidence sur $y$, dont la fiabilité décroît avec la distance $d_i$ entre $\mathbf{x}$ et $\mathbf{x}_i$.
- Chaque fonction de masse $m_i$ est transformée (affaiblie) en ue fonction de masse $m_i'$ moins informative :

$$m_i'(A) = \varphi(d_i)\, m_i(A), \quad \forall A \subset \Omega.$$

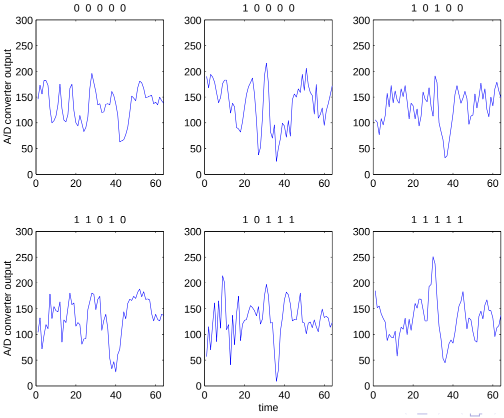
$$m_i'(\Omega) = 1 - \sum_{A \subset \Omega} m_i'(A).$$

- Les $k$ fonctions de masse sont combinées par la règle de Dempster :

$$m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} m_i'.$$

# Exemple : données EEG

Signaux EEG codés comme des vecteurs de dimension $p = 64$ ; $c = 2$ classes : 50 % positifs (complexes K), 50 % négatifs (ondes delta) ; étiquetage par 5 experts.



# Résultats
(Denoeux and Zouhal, 2001)

- Pour chaque exemple $\mathbf{x}_i$, les opinions des experts sont modélisées par une fonction de masse $m_i$.
- $n = 200$ exemples d'apprentissage, 300 exemples de test

| $k$ | $k$-ppv | $k$-ppv pond. | $k$-ppv créd. | $k$-ppv créd. (étiq. incertain) |
|---|---|---|---|---|
| 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 |

# Réseau de neurones crédibiliste

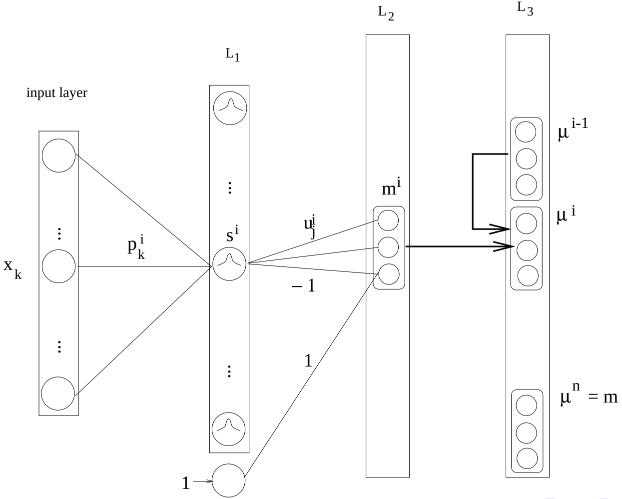
- Implémentation de la règle des $k$-ppv crédibilistes dans un réseau de neurones de type RBF avec $r$ prototypes : $\mathbf{p}_1, \ldots, \mathbf{p}_r$.
- Chaque prototype $\mathbf{p}_i$ est caractérisé par des degrés d'appartenance $u_{ik}$ à chaque classe $\omega_k$ avec $\sum_{k=1}^{c} u_{ik} = 1$
- La distance entre $\mathbf{x}$ et $\mathbf{p}_i$ induit une fonction de masse :

$$\begin{aligned} m_i(\{\omega_k\}) &= \alpha_i u_{ik} \exp(-\gamma_i \|\mathbf{x} - \mathbf{p}_i\|^2) \quad \forall k \\ m_i(\Omega) &= 1 - \alpha_i \exp(-\gamma_i \|\mathbf{x} - \mathbf{p}_i\|^2) \end{aligned}$$

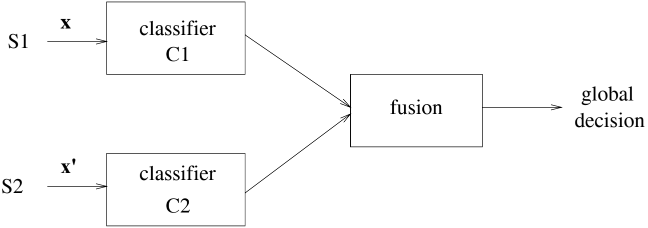
$$m = \bigoplus_{i=1}^{r} m_i$$

- Apprentissage des paramètres $\mathbf{p}_i$, $u_{ik}$, $\gamma_i$, $\alpha_i$ à partir des données par minimisation d'une fonction d'erreur.

# Architecture



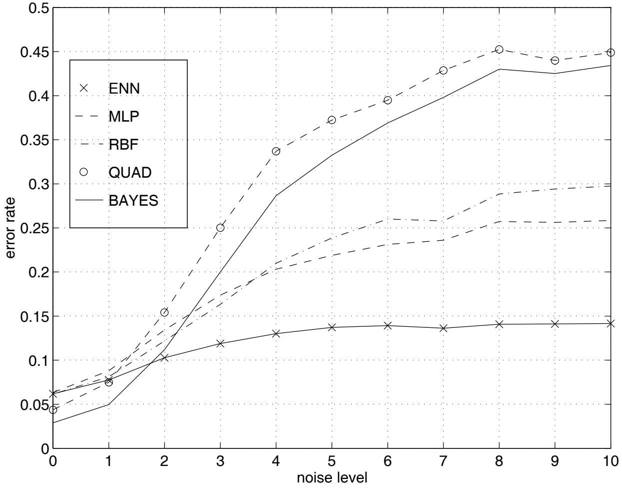
# Exemple (fusion de décisions)




- $c = 2$ classes
- Ensemble d'apprentissage ($n = 60$) : $\mathbf{x} \in \mathbb{R}^5, \mathbf{x}' \in \mathbb{R}^3$, distributions normales, conditionnellement indépendantes
- Ensemble de test ( conditions opératoires réelles) : $\mathbf{x} \leftarrow \mathbf{x} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$.

# Resultats

## Taux d'erreur de test : $\mathbf{x} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$



# Outline

# Classification multi-label

- Dans certaines applications, les individus peuvent appartenir simultanément à plusieurs classes.
- Par exemple, en indexation d'images, le contenu d'une image peut être décrit par plusieurs termes tels que : "montagne", "mer", "ville", etc.
- Soit $\Theta = \{\theta_1, \ldots, \theta_c\}$ l'ensemble des classes, l'étiquette d'un exemple peut être représentée par une variable $y$ à valeurs dans $\Omega = 2^\Theta$.
- L'expression d'une connaissance partielle sur $y$ par une fonction de croyance peut nécessiter la stockage de $2^{2^c}$ nombres.

| $c$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| $2^{2^c}$ | 16 | 256 | 65536 | 4.3e9 | 1.8e19 | 3.4e38 | 1.2e77 |

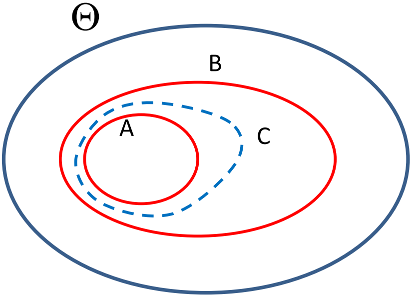
# Grands cadres de discernement

## Approche générale

- Principe :
  1. Considérer un ordre partiel $\leq$ sur le cadre $\Omega$ tel que $(\Omega, \leq)$ soit un treillis.
  2. Définir l'ensemble des propositions comme l'ensemble $\mathcal{I} \subset 2^\Omega$ des intervalles du treillis.
  3. Définir *m*, *bel*, *pl*, *w*, etc., comme des fonctions de $\mathcal{I}$ dans $[0, 1]$ (possible car $(\mathcal{I}, \subseteq)$ est un treillis).
- La cardinalité de $\mathcal{I}$ étant au plus proportionnelle à $|\Omega|^2$, les opérations de la théorie des fonctions de croyance peuvent être effectuées en temps polynomial (et non exponentiel comme dans $(2^\Omega, \subseteq)$).

# Classification multi-label

- Le cadre de discernement est $\Omega = 2^\Theta$, où $\Theta$ est l'ensemble des classes.
- L'ordre naturel dans $2^\Theta$ est $\subseteq$, est $(2^\Theta, \subseteq)$ est le treillis booléen des parties de $\Theta$.




Les intervalles de $(2^\Theta, \subseteq)$ sont les ensembles de parties de $\Theta$ de la forme :

$$[A, B] = \{C \subseteq \Theta | A \subseteq C \subseteq B\}$$

pour $A \subseteq B \subseteq \Theta$.

# Exemple (diagnostic)

- Soit $\Theta = \{a, b, c, d\}$ un ensemble de défauts.
- Elément d'évidence 1 $\rightarrow$ $a$ est certainement présent et $\{b, c\}$ sont peut-être aussi présents, avec un degré de confiance de 0.7 :

$$m_1([\{a\}, \{a, b, c\}]) = 0.7, \quad m_1([\emptyset_\Theta, \Theta]) = 0.3$$

- Elément d'évidence 2 $\rightarrow$ $c$ est certainement présent et soit $\{a, b\}$ (with confiance = 0.8), soit $\{a, d\}$ (confiance = 0.2) sont peut-être aussi présents :

$$m_2([\{c\}, \{a, b, c\}]) = 0.8, \quad m_2([\{c\}, \{a, c, d\}]) = 0.2$$

# Exemple

## Combinaison par la règle de Dempster

| | $[\{a\},\{a,b,c\}]$ 0.7 | $[\emptyset_\Theta,\Theta]$ 0.3 |
|---|---|---|
| $[\{c\},\{a,b,c\}]$ 0.8 | $[\{a,c\},\{a,b,c\}]$ 0.56 | $[\{c\},\{a,b,c\}]$ 0.24 |
| $[\{c\},\{a,c,d\}]$ 0.2 | $[\{a,c\},\{a,c\}]$ 0.14 | $[\{c\},\{a,c,d\}]$ 0.06 |

Degrés de croyance dans l'hypothèse que

- Le défaut $a$ est présent : $bel([\{a\},\Theta]) = 0.56 + 0.14 = 0.70$ ;
- Le défaut $d$ n'est pas présent : $bel([\emptyset_\Theta,\overline{\{d\}}]) = bel([\emptyset_\Theta,\{a,b,c\}]) = 0.56 + 0.14 + 0.24 = 0.94$.

# Classification multi-label
## Etiquettes imprécises

- Soit un ensemble d'apprentissage de la forme :

$$\mathcal{L} = \{(\mathbf{x}_1, [A_1, B_1]), \ldots, (\mathbf{x}_n, [A_n, B_n])\}$$

  où
  - $\mathbf{x}_i \in \mathbb{R}^p$ est le vecteur de caractéristiques pour l'exemple $i$
  - $A_i$ est l'ensemble des classes auxquelles appartient certainement l'exemple $i$ ;
  - $B_i$ est l'ensemble des classes auxquelles appartient possiblement l'exemple $i$.
- Dans un contexte multi-expert, $A_i$ peut être l'ensemble des classes affectées à l'exemple $i$ par tous les experts, et $B_i$ l'ensemble des classes affectées par au moins un expert.

# Règle des $k$-ppv crédibiliste multi-label

## Construction des fonctions de masse

- Soit $\mathcal{N}_k(\mathbf{x})$ l'ensemble des $k$ ppv d'un nouvel exemple $\mathbf{x}$, pour une distance $d$.
- Soit $\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})$ avec l'étiquette $[A_i, B_i]$. Cet élément d'évidence peut être décrit par une fonction de masse dans $(\mathcal{I}, \subseteq)$ :

$$\begin{aligned} m_i([A_i, B_i]) &= \varphi(d_i), \\ m_i([\emptyset_\Theta, \Theta]) &= 1 - \varphi(d_i), \end{aligned}$$

  où $\varphi$ est une fonction décroissante de $[0, +\infty)$ ans $[0, 1]$ telle que $\lim_{d \to +\infty} \varphi(d) = 0$.
- Les $k$ fonctions de masse sont combinées par la règle de Dempster :

$$m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} m_i$$

# Règle des *k*-ppv crédibiliste multi-label

## Décision

- Soit $\widehat{Y}$ be the l'ensemble des classes prédites pour l'exemple **x**.
- Pour décider d'inclure ou non chaque classe $\theta \in \Theta$ dans $\widehat{Y}$, on calcule
  - le degré de croyance $bel([\{\theta\}, \Theta])$ dans l'hypothèse que l'ensemble de labels $Y$ contient $\theta$, et
  - le degré de croyance $bel([\emptyset, \overline{\{\theta\}}])$ dans l'hypothèse que l'ensemble de labels $Y$ ne contient pas $\theta$.
- On définit alors $\widehat{Y}$ comme

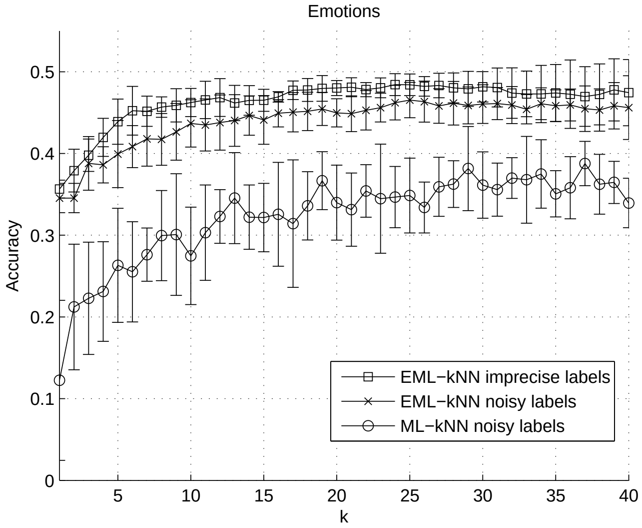
$$\widehat{Y} = \{\theta \in \Theta \mid bel([\{\theta\}, \Theta]) \geq bel([\emptyset, \overline{\{\theta\}}])\}.$$

# Exemple : données Emotions
Trohidis et al., 2008

- Problème : prédire les émotions engendrées par un morceau de musique.
- 593 morceaux annotés par des juges.
- Six émotions (surprise, joie, calme, tristesse, etc.).
- Chaque morceau décrit par 72 attributs et étiqueté par une ou plusieurs émotions (classes).
- Données partitionnées en un ensemble d'apprentissage de 391 exemples et un ensemble de test de 202 exemples.
- Evaluation des résultats :

$$Acc = \frac{1}{n}\sum_{i=1}^{n}\frac{|Y_i \cap \widehat{Y}_i|}{|Y_i \cup \widehat{Y}_i|}$$

# Résultats


Emotions
Accuracy
k
0
0.1
0.2
0.3
0.4
0.5
5
10
15
20
25
30
35
40
EML−kNN imprecise labels
EML−kNN noisy labels
ML−kNN noisy labels

# Outline

# Problème

- Soient $q$ sources (classifieurs) fournissant des fonctions de masse $m_{(1)}, \ldots, m_{(q)}$ sur $\Omega$.
- Différents cas :
  - Attributs différents (capteurs différents, etc.) ;
  - Ensembles d'apprentissages différents (disjoints ou non) ;
  - Algorithmes d'apprentissages différents, etc.
- Problème : comment combiner les sorties des classifieurs ?

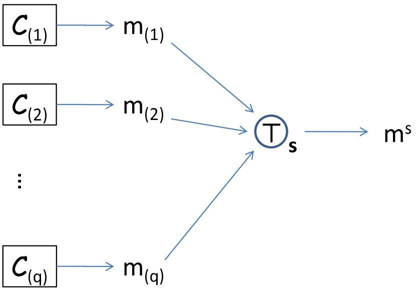
# Choix d'une règle de combinaison

- Hypothèses sur l'indépendance des sources :
  - Classifieurs indépendants :

$$m_{\oplus} = m_{(1)} \oplus \ldots \oplus m_{(q)}$$

  - Classifieurs dépendants :
    - Règle prudente : $m_{\owedge} = m_{(1)} \owedge \ldots \owedge m_{(q)}$
    - Autre règle ?
- Comment apprendre une règle optimale ?
- B. Quost, M.-H. Masson, T. Denœux. Classifier fusion in the Dempster-Shafer framework using optimized t-norm based combination rules. *International Journal of Approximate Reasoning* (à paraître en 2011).

# Approche 1 : optimisation d'une règle unique



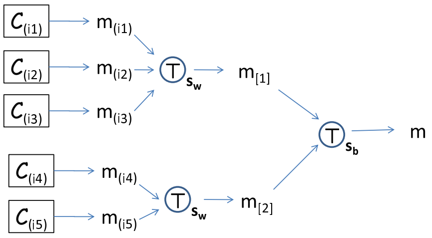

- Combinaison par une règle de la forme

$$m_1 \mathbin{\textcircled{$\top$}}_s m_2 = \bigoplus_{A \subset \Omega} A^{w_1(A) \top_s w_2(A)}$$

où $\top_s$ est une t-norme dépendant d'un paramètre $s$.

- Détermination de $s$ par minimisation d'un critère d'erreur.

# Approche 2 : optimisation de deux règles




- Regroupement des classifieurs selon un critère de distance entre fonctions de masse (distance de Jousselme).
- Combinaison par deux règles : intra-groupe et inter-groupes.
- Optimisation simultanée de ces deux règles par apprentissage (validation croisée).

# Résultats

| données | Satimage | Segment | Vowel | Waveform |
|---|---|---|---|---|
| Dempster | 24.45 | 17.91 | **<u>57.36</u>** | 16.90 |
| OPT1 $(\widehat{s})$ | **21.80** | **<u>15.16</u>** | 59.52 | **15.22** |
| OPT2 $(\widehat{s_w}, \widehat{s_b})$ | **<u>21.69</u>** | **<u>15.16</u>** | 59.52 | **<u>15.19</u>** |
| règle prudente | **21.80** | **<u>15.16</u>** | 59.52 | 16.59 |
| moyenne | 28.92 | 23.74 | 59.96 | 20.52 |

# Conclusion

- La théorie des fonctions de croyance constitue un cadre très général (englobant les formalismes ensembliste et probabiliste) pour la représentation et la manipulation d'informations imparfaites.
- Ce cadre est adapté à la résolution de problèmes complexes en reconnaissance de formes, particulièrement ceux impliquant :
  - Des informations imparfaites (données partiellement supervisées, capteurs peu fiables, etc.) ;
  - Une combinaison d'informations objectives (données) et subjectives (opinions d'experts, perceptions) : intégration d'informations a priori en classification, appréciation du contexte par des experts, etc. ;
  - Des sources d'informations multiples (fusion multi-capteurs, combination multi-experts, méthodes d'ensemble en classification supervisée ou non).

# Problèmes ouverts

1. Développer des stratégies de fusion sophistiquées incluant
   - de nouvelles règles de combinaison permettant la fusion d'informations dépendantes et/ou partiellement inconsistantes ;
   - des méta-connaissances sur la qualité (fiabilité) des sources d'information.
2. Nouveaux problèmes en RdF et en apprentissage :
   - Apprentissage de préférences (définition de fonctions de croyance sur des ensembles de relations de préférence) ;
   - Apprentissage à partir de données incertaines (attributs ou étiquettes de classes).

# Références

Articles et sources Matlab disponibles à l'adresse :

`http ://www.hds.utc.fr/~tdenoeux`

MERCI !