

Théorie des fonctions de croyance et apprentissage automatique

Thierry Denœux

Université de technologie de Compiègne
Heudiasyc (UMR CNRS 7253)
et
Institut universitaire de France

<https://www.hds.utc.fr/~tdenoeux>

GdR ISIS, Journée « Apprentissage automatique multimodal et fusion d'informations »(2ième édition)
19 janvier 2022

Motivation

- Ces dernières années, l'apprentissage automatique a connu des développements importants, avec notamment la montée en puissance des réseaux de neurones profonds.
- Ces derniers offrent des mécanismes très efficaces pour extraire des **caractéristiques non-linéaires** à partir de masses de données complexes (images, vidéos, textes, etc.).
- Ces caractéristiques sont utilisées pour faire des **prédictions**, généralement par régression logistique (couche de sortie softmax) ou par régression linéaire (couche de sortie linéaire).
- Un des problèmes d'actualité en apprentissage est celui de la **quantification des incertitudes** de prédiction.

Importance de la quantification des incertitudes

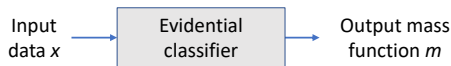
- Sources d'incertitude :
 - Variabilité statistique (incertitude **aléatoire**)
 - Méconnaissance de la distribution des données (incertitude **épistémique**)
 - Exemples de test issus d'une autre distribution
- La quantification des incertitudes est importante pour :
 - Permettre l'application de règles de décision **prudentes** (rejet, affectation à plusieurs classes, etc.)
 - Dans le cas d'un système interactif d'aide à la décision, rendre la main à l'utilisateur lorsque l'incertitude est trop grande (aide à la conduite automobile, aide au diagnostic médical, etc.)
 - Permettre la **fusion d'informations** (apprentissage multi-modal, systèmes multi-capteurs, etc.)

Approches de la quantification des incertitudes

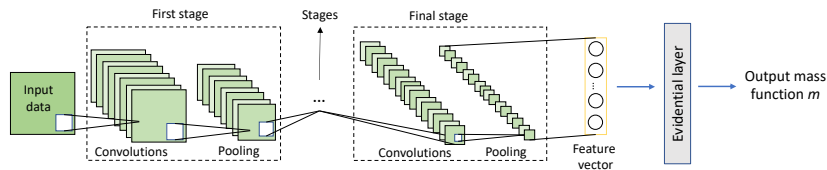
- Deux problèmes :
 - 1 Choix d'un cadre théorique de représentation des incertitudes
 - 2 Définition de méthodes pratiques et rationnelles pour la description des incertitudes de prédiction dans le formalisme choisi.
- Principaux cadres théoriques :
 - Bayésien (probabilités)
 - Fréquentiste (régions de confiance)
 - Probabilités imprécises (probabilités inférieures/supérieures, etc.)
 - Théorie des possibilités
 - Dempster-Shafer (fonctions de croyance)
- La théorie de DS est plus générale que la théorie bayésienne (qui en est un cas particulier). Elle est donc potentiellement plus riche et peut permettre de distinguer différentes sources d'incertitude.
- La théorie de DS est particulièrement bien adaptée à la fusion d'information (qui joue un rôle central dans ce formalisme).

Classifieur évidentiel

De la notion de classifieur évidentiel¹...



... à celle de couche évidentielle² :



1. T. Denœux. A neural network classifier based on Dempster-Shafer theory. *IEEE transactions on Systems, Man and Cybernetics A* 30(2) :131-150, 2000.

2. Z. Tong, Ph. Xu and T. Denœux. An evidential classifier based on Dempster-Shafer theory and deep learning. *Neurocomputing* 450 :275-293, 2021.

Plan

- 1 Théorie des fonctions de croyance
 - Représentation de l'information
 - Combinaison de l'information
 - Décision
- 2 Classifieurs évidentiels
 - Approche basée sur des prototypes
 - Approche basée sur les poids d'évidence
- 3 Application à la fusion de classifieurs

Plan

- 1 **Théorie des fonctions de croyance**
 - Représentation de l'information
 - Combinaison de l'information
 - Décision
- 2 **Classifieurs évidentiels**
 - Approche basée sur des prototypes
 - Approche basée sur les poids d'évidence
- 3 **Application à la fusion de classifieurs**

Théorie des fonctions de croyance

- Aussi appelée **théorie de Dempster-Shafer**, d'après les travaux fondateurs d'Arthur Dempster³ et de Glenn Shafer⁴
- Un modèle de l'incertain basé sur deux idées principales :
 - 1 La représentation d'éléments d'évidence élémentaires par des **fonctions de croyance**
 - 2 Leur combinaison par un opérateur appelé **règle de Dempster**
- Le modèle généralise les approches logique/ensembliste et probabiliste.

3. A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics* 38 :325–339, 1967.

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

Plan

- 1 Théorie des fonctions de croyance
 - Représentation de l'information
 - Combinaison de l'information
 - Décision
- 2 Classifieurs évidentiels
 - Approche basée sur des prototypes
 - Approche basée sur les poids d'évidence
- 3 Application à la fusion de classifieurs

Fonction de masse

Définition

Définition

Une *fonction de masse* sur un ensemble fini Ω (appelé cadre de discernement) est une application $m : 2^\Omega \rightarrow [0, 1]$ vérifiant

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

et $m(\emptyset) = 0$. Tout sous-ensemble A de Ω tel que $m(A) > 0$ est appelé *ensemble focal*.

Interprétation :

- Ω est le domaine d'une variable inconnue Y
- m modélise un élément d'évidence (une information) sur Y
- $m(A)$ est la probabilité que l'élément d'évidence nous permette d'affirmer que $Y \in A$, et rien de plus.

Fonctions de masse particulières

- Si m a un seul élément focal A , on dit que c'est une fonction de masse **catégorique** et on la note m_A .
- En particulier, la **fonction de masse vide** a pour seul élément focal Ω , on la note m_\emptyset . La fonction de masse vide représente l'absence totale d'information (ignorance totale).
- Une fonction de masse dont tous les ensembles focaux sont des singletons est dite **Bayésienne**. Elle correspond à une distribution de probabilité.

Affaiblissement

- Soit m une fonction de masse sur Ω et $\alpha \in [0, 1]$.
- On appelle **affaiblissement** l'opération qui associe à m la fonction de masse

$${}^{\alpha}m = (1 - \alpha)m + \alpha m_{\gamma}$$

En particulier, ${}^0m = m$ et ${}^1m = m_{\gamma}$. Le coefficient α est appelé coefficient d'affaiblissement.

- Interprétation : ${}^{\alpha}m$ représente l'information apportée par une source S qui fournit une fonction de masse m , supposée fiable avec une probabilité $1 - \alpha$.

Fonctions de croyance et de plausibilité

Définition

À une fonction de masse m sur Ω sont associées des *fonctions de croyance et de plausibilité* définies par

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) = 1 - Bel(\bar{A}),$$

pour tout $A \subseteq \Omega$

- Interprétation :
 - $Bel(A)$ est la probabilité qu'on puisse déduire que $Y \in A$
 - $Pl(A)$ est la probabilité que l'on ne puisse pas déduire que $Y \notin A$
- Cas de l'ignorance totale : $Bel(A) = 0$ pour tout $A \neq \Omega$ et $Pl(A) = 1$ pour tout $A \neq \emptyset$.

Caractérisation des fonctions de croyance

Théorème

Une application $Bel : 2^\Omega \rightarrow [0, 1]$ est une fonction de croyance pour une certaine fonction de masse m ssi elle vérifie $Bel(\emptyset) = 0$, $Bel(\Omega) = 1$ et elle est **complètement monotone** : pour tout $k \geq 2$ et toute famille A_1, \dots, A_k in 2^Ω ,

$$Bel\left(\bigcup_{i=1}^k A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1, \dots, k\}} (-1)^{|I|+1} Bel\left(\bigcap_{i \in I} A_i\right) \quad (1)$$

La fonction de masse m est alors obtenue par

$$m(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B)$$

Plan

- 1 Théorie des fonctions de croyance
 - Représentation de l'information
 - **Combinaison de l'information**
 - Décision
- 2 Classifieurs évidentiels
 - Approche basée sur des prototypes
 - Approche basée sur les poids d'évidence
- 3 Application à la fusion de classifieurs

Règle de Dempster

Définition (Règle de Dempster)

Soient deux fonctions de masse m_1 et m_2 définies sur le même ensemble Ω . leur **somme orthogonale** est la fonction de masse définie par

$$(m_1 \oplus m_2)(A) := \frac{1}{1 - \kappa} \sum_{B \cap C = A} m_1(B)m_2(C), \quad \forall A \neq \emptyset$$

et $(m_1 \oplus m_2)(\emptyset) = 0$, où κ est le **degré of conflit** défini par

$$\kappa := \sum_{B \cap C = \emptyset} m_1(B)m_2(C).$$

Interprétation : \oplus est l'opérateur permettant de combiner des éléments d'évidence indépendants.

Propriétés de la règle de Dempster

- Commutativité : $\forall m_1, m_2, m_1 \oplus m_2 = m_2 \oplus m_1$
- Associativité : $\forall m_1, m_2, m_3, (m_1 \oplus m_2) \oplus m_3 = m_1 \oplus (m_2 \oplus m_3)$
- Élément neutre : $\forall m, m \oplus m_\tau = m_\tau \oplus m = m$
- Généralisation de l'intersection : pour toutes fonctions de masses catégoriques m_A et m_B telles que $A \cap B \neq \emptyset$, $m_A \oplus m_B = m_{A \cap B}$
- **Généralisation du conditionnement bayésien** : si m est une fonction de masse bayésienne de fonction de croyance (additive) P et m_A est une fonction de masse catégorique d'ensemble focal A , alors $m \oplus m_A$ est une fonction de masse bayésienne, et la fonction de croyance (additive) associée est $P(\cdot | A)$

Poids d'évidence

La règle de Dempster peut être calculée en additionnant des **poids d'évidence**.

Définition (Poids d'évidence)

Soit m une **fonction de masse simple** d'ensembles focaux A et Ω :

$$m(A) = s$$

$$m(\Omega) = 1 - s.$$

La quantité $w = -\ln(1 - s)$ est appelée **poids d'évidence** pour A . La fonction de masse m est notée A^w .

Proposition

Étant données deux fonctions de masse simples A^{w_1} et A^{w_2} de même ensemble focal $A \subset \Omega$, leur somme orthogonale s'écrit

$$A^{w_1} \oplus A^{w_2} = A^{w_1 + w_2}$$

Fonctions de masse séparables

Définition (Fonction de masse séparable)

Une fonction de masse m est dite *séparable* si elle est la somme orthogonale de fonctions de masse simples. Elle peut alors s'écrire

$$m = \bigoplus_{\emptyset \neq A \subset \Omega} A^{w(A)},$$

où $w(A) \in [0, +\infty)$ est le *poids d'évidence* pour A .

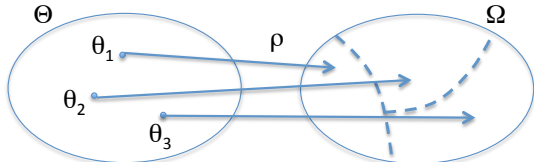
Proposition

Pour combiner des fonctions de masse séparables, il suffit d'ajouter leurs poids :

$$\left(\bigoplus_{\emptyset \neq A \subset \Omega} A^{w_1(A)} \right) \oplus \left(\bigoplus_{\emptyset \neq A \subset \Omega} A^{w_2(A)} \right) = \bigoplus_{\emptyset \neq A \subset \Omega} A^{w_1(A) + w_2(A)}$$

Combinaison d'informations hétérogènes

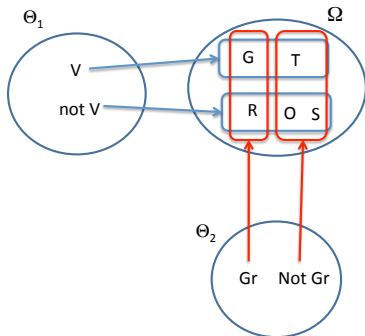
- Il arrive souvent que différents éléments d'information soient exprimés avec différents degrés de granularité.
- Exemple : $\Theta = \{\text{véhicule}, \neg\text{véhicule}\}$, $\Omega = \{\text{voiture}, \text{vélo}, \text{camion}, \dots\}$.
- On dit qu'un cadre de discernement Ω est un **raffinement** d'un cadre Θ si on peut faire correspondre à chaque élément θ de Θ un élément $\rho(\theta)$ d'une partition de Ω .



- Étant donné une fonction de masse m^Θ sur Θ et un raffinement Ω de Θ , on appelle **extension** de m^Θ dans Ω la fonction de masse $m^{\Theta \uparrow \Omega}$ obtenue en transférant chaque masse $m^\Theta(A)$ à $\bigcup_{\theta \in A} \rho(\theta)$.

Combinaison d'informations hétérogènes (suite)

- Deux cadres de discernement sont dits **compatibles** s'ils admettent un raffinement commun :



- Etant données deux fonctions de masse m^{Θ_1} et m^{Θ_2} définies sur deux cadres de discernement compatibles Θ_1 et Θ_2 ayant pour raffinement commun Ω , on définit leur somme orthogonale par

$$m^\Omega = m^{\Theta_1 \uparrow \Omega} \oplus m^{\Theta_2 \uparrow \Omega}$$

Plan

- 1 Théorie des fonctions de croyance
 - Représentation de l'information
 - Combinaison de l'information
 - **Décision**
- 2 Classifieurs évidentiels
 - Approche basée sur des prototypes
 - Approche basée sur les poids d'évidence
- 3 Application à la fusion de classifieurs

Nécessité d'une théorie de la décision

- Pour appliquer le modèle de DS en classification, il faut une **théorie de la décision** adaptée au cas où l'incertitude est modélisée par une fonction de croyance sur l'ensemble des classes.
- On trouve dans la littérature de nombreuses approches plus ou moins bien justifiées⁵.
- Une théorie de la décision justifiée axiomatiquement a récemment été proposée⁶. La méthode décrite ici est un cas particulier de cette théorie.

5. T. Denoeux. Decision-Making with Belief Functions : a Review. *International Journal of Approximate Reasoning* 109 :87–110, 2019.

6. T. Denoeux and P. P. Shenoy. An Interval-Valued Utility Theory for Decision Making with Dempster-Shafer Belief Functions. *International Journal of Approximate Reasoning* 124 :194–216, 2020.

Cadre général

- Soit Ω un ensemble d'états, C un ensemble de **conséquences**, \mathcal{F} un ensemble d'applications $f : \Omega \rightarrow C$ appelées **actes**, et u une application de C dans \mathbb{R} appelée **fonction d'utilité**.
- Soit m une fonction de masse sur Ω . Pour tout acte f , on définit des **utilités espérées inférieure et supérieure** par

$$U_*(f) = \sum_{A \subseteq \Omega} m(A) \min_{\omega \in A} u[f(\omega)]$$

$$U^*(f) = \sum_{A \subseteq \Omega} m(A) \max_{\omega \in A} u[f(\omega)]$$

- Relation de préférence : f est préféré (au sens large) à f' ssi

$$\rho U_*(f) + (1 - \rho)U^*(f) \geq \rho U_*(f') + (1 - \rho)U^*(f'),$$

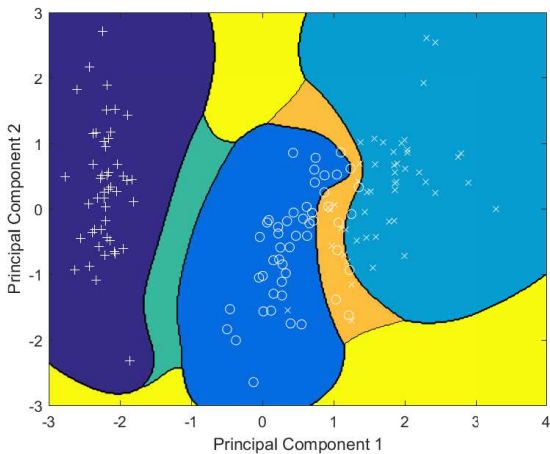
où $\rho \in [0, 1]$ est un indice de pessimisme.

Application à la classification

- En classification, Ω est l'ensemble des classes, un acte f_A est le choix d'un sous-ensemble non vide de classes A , parmi un ensemble $\mathcal{A} \subseteq 2^\Omega \setminus \{\emptyset\}$.
- Par exemple, si $\mathcal{A} = \{\{\omega\} : \omega \in \Omega\}$ on n'autorise que le choix d'une seule classe, si $\mathcal{A} = \{\{\omega\} : \omega \in \Omega\} \cup \{\Omega\}$ on autorise en plus le rejet, si $\mathcal{A} = \{A \subset \Omega : 0 < |A| \leq 2\}$ on autorise le choix d'un singleton ou d'une paire de classes, etc.
- En définissant les utilités $u[f_A(\omega)]$ de choisir l'ensemble de classes A si la vraie classe est ω , pour tout $A \in \mathcal{A}$ et tout $\omega \in \Omega$, on peut appliquer la théorie précédente pour choisir l'acte de plus grande utilité espérée⁷.

7. L. Ma and T. Denoeux. Partial Classification in the Belief Function Framework. *Knowledge-Based Systems* 214 :106742, 2021.

Exemple



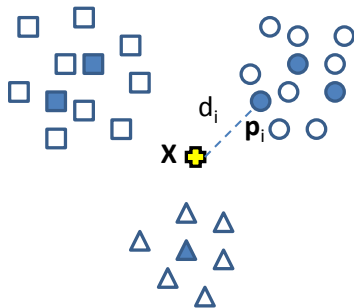
Plan

- 1 Théorie des fonctions de croyance
 - Représentation de l'information
 - Combinaison de l'information
 - Décision
- 2 **Classifieurs évidentiels**
 - Approche basée sur des prototypes
 - Approche basée sur les poids d'évidence
- 3 Application à la fusion de classifieurs

Plan

- 1 Théorie des fonctions de croyance
 - Représentation de l'information
 - Combinaison de l'information
 - Décision
- 2 **Classifieurs évidentiels**
 - **Approche basée sur des prototypes**
 - Approche basée sur les poids d'évidence
- 3 Application à la fusion de classifieurs

Hypothèses



- L'ensemble d'apprentissage est résumé par r **prototypes**.
- Chaque prototype p_i a un **degré d'appartenance** $u_{ik} \in [0, 1]$ à chaque classe ω_k , avec $\sum_{k=1}^C u_{ik} = 1$.
- Chaque prototype p_i est un **élément d'évidence** sur la classe de x , dont la **fiabilité décroît avec la distance** $d_i = \|x - p_i\|$ entre x et p_i .

Formalisation

- Soit $s_i = \exp(-\gamma_i d_i^2)$ le degré de similarité entre \mathbf{x} et \mathbf{p}_i , avec $\gamma_i > 0$.
- L'information apportée par le prototype \mathbf{p}_i peut être représentée par la fonction de masse⁸ :

$$m_i(\{\omega_k\}) = \alpha_i u_{ik} s_i, \quad k = 1, \dots, c$$

$$m_i(\Omega) = 1 - \alpha_i s_i,$$

avec $\alpha_i \in [0, 1]$.

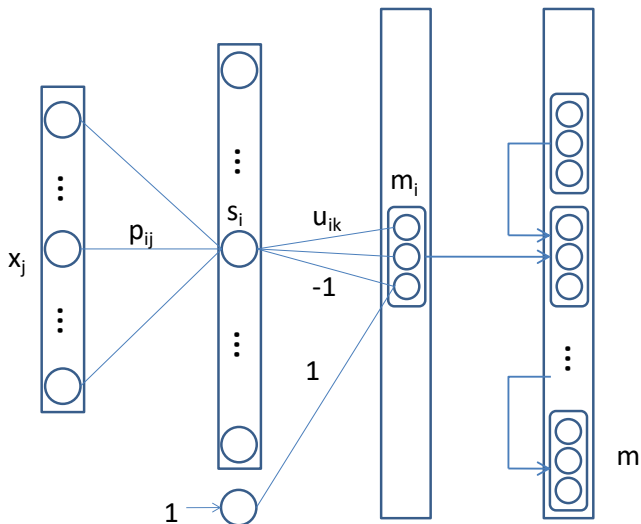
- Combinaison :

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

- La fonction de masse combinée m a pour ensemble focaux les singletons $\{\omega_k\}$, $k = 1, \dots, c$ and Ω .

8. T. Dencœux. A neural network classifier based on Dempster-Shafer theory. *IEEE transactions on Systems, Man and Cybernetics A* 30(2) :131–150, 2000.

Implémentation connexionniste



Apprentissage

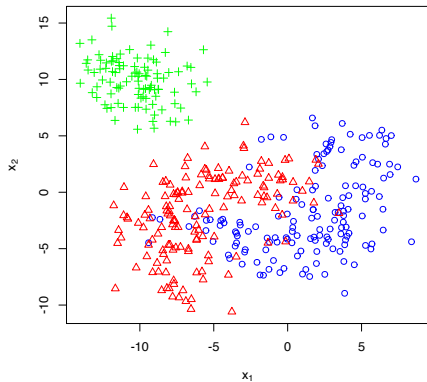
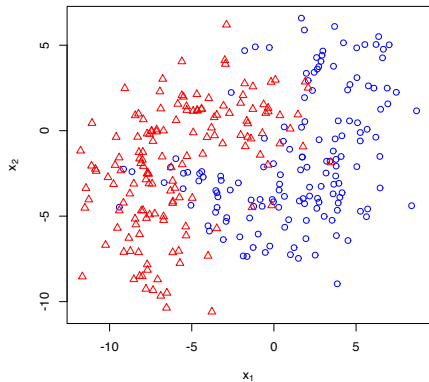
- Paramètres du modèle :
 - Prototypes $\mathbf{p}_j, j = 1, \dots, r$ (rp paramètres)
 - Degrés d'appartenance $u_{ik}, i = 1, \dots, r, k = 1 \dots, c$ ($r(c - 1)$ paramètres)
 - α_j et $\gamma_j, j = 1 \dots, r$ ($2r$ paramètres).
- Fonction de coût :

$$J(\theta) = \underbrace{\sum_{i=1}^n \sum_{k=1}^c (p_{ik} - y_{ik})^2}_{\text{erreur}} + \lambda \underbrace{\sum_{i=1}^r \alpha_i}_{\text{régularisation}}$$

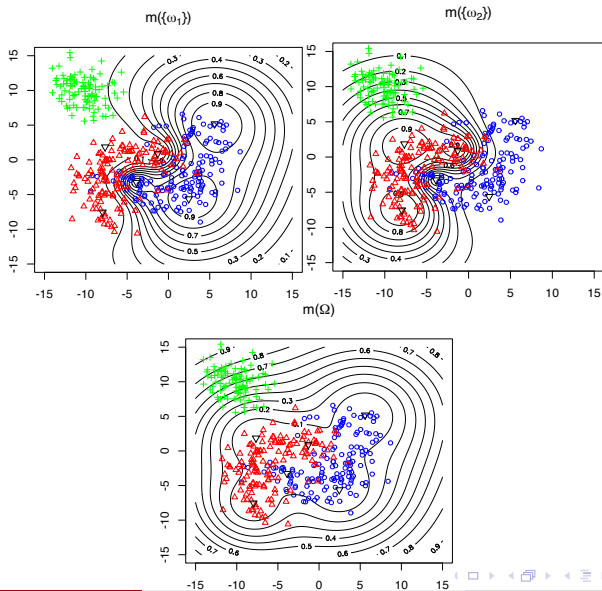
où p_{ik} est la plausibilité de la classe ω_k pour l'exemple i , $y_{ik} = I(y_i = \omega_k)$, et λ est un hyperparamètre (déterminé par validation croisée)⁹.

9. T. Denœux. evclass : Evidential Distance-Based Classification, R package version 1.1.1, <https://CRAN.R-project.org/package=evclass>.

Exemple



Exemple



Plan

- 1 Théorie des fonctions de croyance
 - Représentation de l'information
 - Combinaison de l'information
 - Décision
- 2 Classifieurs évidentiels
 - Approche basée sur des prototypes
 - Approche basée sur les poids d'évidence
- 3 Application à la fusion de classifieurs

Formalisation (cas $c = 2$)

- Comme précédemment, on considère r prototypes $\mathbf{p}_1, \dots, \mathbf{p}_r$, et on note $s_i = \exp(-\gamma_i d_i^2)$ le degré de similarité entre \mathbf{x} et \mathbf{p}_i (avec $\gamma_i > 0$).
- On associe à chaque prototype \mathbf{p}_i un paramètre supplémentaire $v_i \in \mathbb{R}$ tel que la proximité de \mathbf{x} avec le prototype \mathbf{p}_i accrédite
 - La classe ω_1 avec un poids d'évidence $w_i = s_i v_i$ si $v_i \geq 0$;
 - La classe ω_2 avec un poids d'évidence $w_i = -s_i v_i$ si $v_i < 0$.
- La fonction de masse induite par le prototype \mathbf{p}_i s'écrit donc ¹⁰

$$m_i = \{\omega_1\}^{w_i^+} \oplus \{\omega_2\}^{w_i^-},$$

où $w_i^+ = \max(0, w_i)$ et $w_i^- = \max(0, -w_i)$ sont les parties positive et négative de w_i .

10. T. Denœux. Logistic Regression, Neural Networks and Dempster-Shafer Theory : a New Perspective. *Knowledge-Based Systems* 176 :54–67, 2019.

Formalisation (cas $c = 2$)

En combinant l'information apportée par les r prototypes, on obtient

$$\begin{aligned}
 m &= \bigoplus_{i=1}^r \left(\{\omega_1\}^{w_i^+} \oplus \{\omega_2\}^{w_i^-} \right) \\
 &= \left(\bigoplus_{i=1}^r \{\omega_1\}^{w_i^+} \right) \oplus \left(\bigoplus_{i=1}^r \{\omega_2\}^{w_i^-} \right) \\
 &= \{\omega_1\}^{w^+} \oplus \{\omega_2\}^{w^-},
 \end{aligned}$$

où

- $w^+ = \sum_{i=1}^r w_i^+$ est le poids total de l'évidence en faveur de ω_1
- $w^- = \sum_{i=1}^r w_i^-$ est le poids total de l'évidence en faveur de ω_2 .

Formalisation (cas $c = 2$)

Expression de m :

$$m(\{\omega_1\}) = \frac{[1 - \exp(-w^+)] \exp(-w^-)}{1 - \kappa}$$

$$m(\{\omega_2\}) = \frac{[1 - \exp(-w^-)] \exp(-w^+)}{1 - \kappa}$$

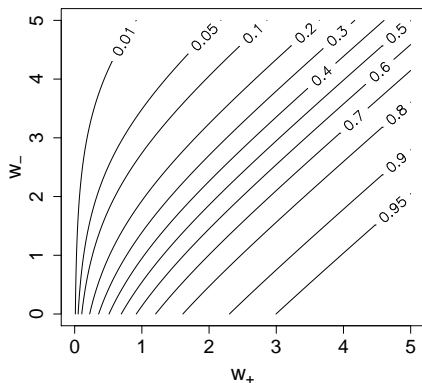
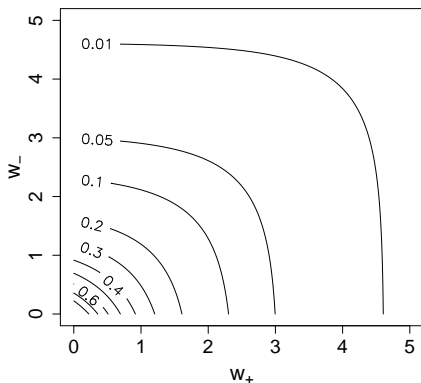
$$m(\Omega) = \frac{\exp(-w^+ - w^-)}{1 - \kappa} = \frac{\exp(-\sum_{i=1}^I |w_i|)}{1 - \kappa},$$

où

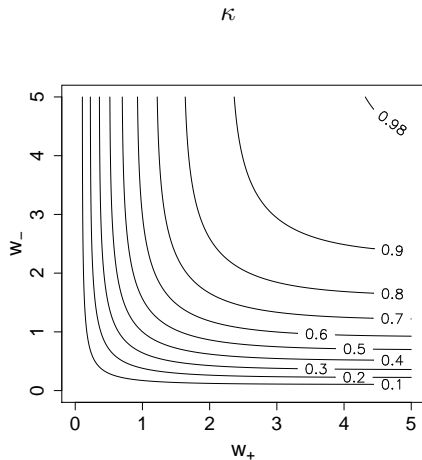
$$\kappa = [1 - \exp(-w^+)][1 - \exp(-w^-)]$$

est le degré de conflit entre les fonctions de masse $\{\omega_1\}^{w^+}$ et $\{\omega_2\}^{w^-}$.

$m(\{\omega_1\})$ et $m(\Omega)$ vs. poids d'évidence

 $m(\{\omega_1\})$

 $m(\Omega)$


Degré de conflit vs. poids d'évidence



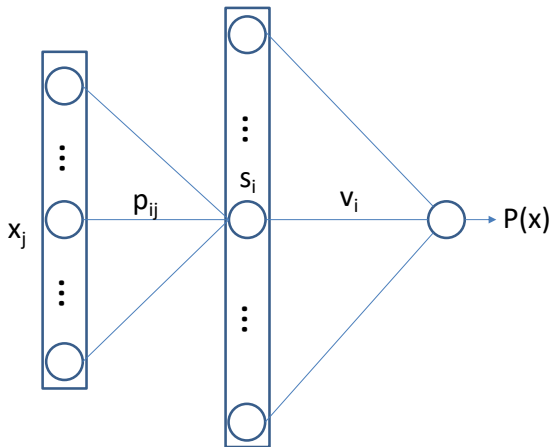
Plausibilités normalisées

- La plausibilité normalisée de la classe ω_1 est

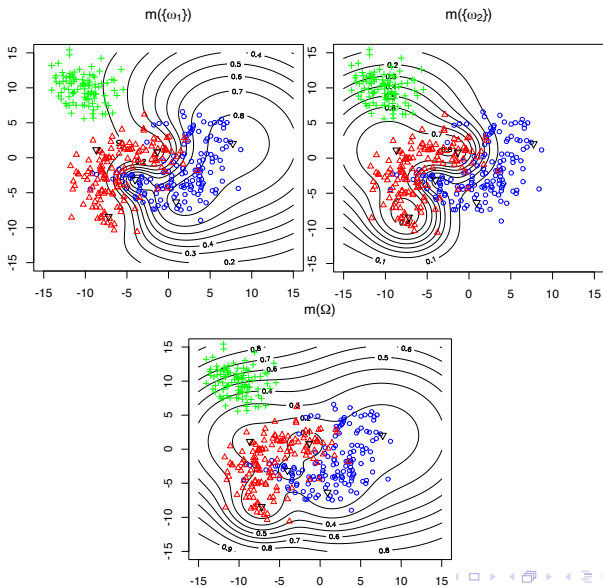
$$\begin{aligned}
 P(\mathbf{x}) &= \frac{PI(\{\omega_1\})}{PI(\{\omega_1\}) + PI(\{\omega_2\})} = \frac{m(\{\omega_1\}) + m(\Omega)}{m(\{\omega_1\}) + m(\{\omega_2\}) + 2m(\Omega)} \\
 &= \frac{1}{1 + \exp[-\sum_{i=1}^r v_i s_i]}
 \end{aligned}$$

- C'est la sortie d'un réseau de neurones à **fonctions de base radiales (FBR)** avec une neurone de sortie à fonction d'activation logistique.
- Pour un tel réseau, la probabilité calculée en sortie est une plausibilité normalisée pour une **fonction de masse latente**.

Implémentation connexionniste



Exemple



Extension à plus de 2 classes

- Soit $\Omega = \{\omega_1, \dots, \omega_c\}$ avec $c > 2$. L'information apportée par chaque prototype \mathbf{p}_i est représentée par c fonctions de masse simples m_{i1}, \dots, m_{ic} .
- La fonction de masse m_{ik} s'écrit

$$m_{jk} = \{\omega_k\}^{w_{jk}^+} \oplus \overline{\{\omega_k\}}^{w_{jk}^-}$$

où $w_{ik} = s_i v_{ik}$ et v_{ik} est un paramètre associé au prototype i et à la classe k .

- En combinant les $r \times k$ fonctions de masse simples, on obtient

$$m = \bigoplus_{i=1}^r \bigoplus_{k=1}^c \left(\{\omega_k\}^{w_{ik}^+} \oplus \overline{\{\omega_k\}}^{w_{ik}^-} \right) = \bigoplus_{k=1}^c \left(\{\omega_k\}^{w_k^+} \oplus \overline{\{\omega_k\}}^{w_k^-} \right),$$

où

- $w_k^+ = \sum_{i=1}^r w_{ik}^+$ est le poids total de l'évidence en faveur de la classe ω_k
- $w_k^- = \sum_{i=1}^r w_{ik}^-$ est le poids total de l'évidence contre la classe ω_k

Plausibilité normalisée

- La plausibilité normalisée de la classe ω_k est

$$P_k(\mathbf{x}) = \frac{PI(\{\omega_k\})}{\sum_{l=1}^c PI(\{\omega_l\})} = \frac{\exp(\sum_{i=1}^r v_{ik} s_i)}{\underbrace{\sum_{l=1}^c \exp(\sum_{i=1}^r v_{il} s_i)}_{\text{transformation softmax}}},$$

- Ce sont les sorties d'un réseau de neurones à FBR avec une **couche de sortie softmax** de c neurones.

Plan

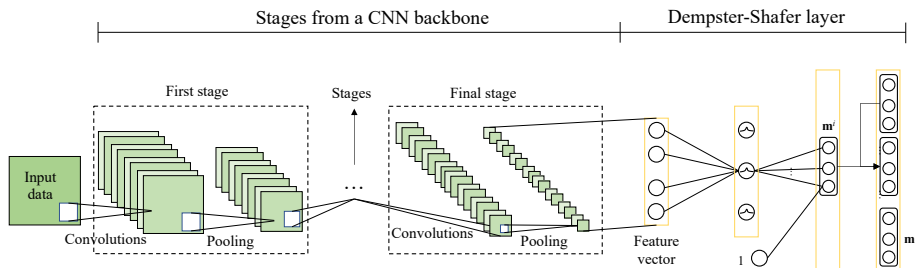
- 1 Théorie des fonctions de croyance
 - Représentation de l'information
 - Combinaison de l'information
 - Décision
- 2 Classifieurs évidentiels
 - Approche basée sur des prototypes
 - Approche basée sur les poids d'évidence
- 3 Application à la fusion de classifieurs

Motivations

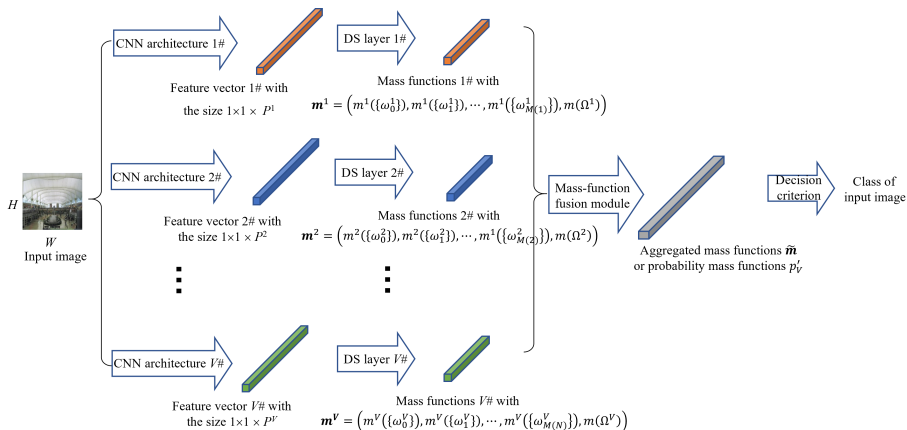
- L'apprentissage sur de très grands jeux de données nécessite des capacités de calcul considérables.
- D'autre part, pour un problème donné, il existe en général de nombreuses **bases de données hétérogènes** (avec des ensembles de classes différents, et différents niveau de granularité).
- Une solution aux problèmes précédents consiste à fusionner des classifieurs entraînés sur des ensembles d'apprentissage différents ¹¹.

11. Zh. Tong, Ph. Xu and T. Denœux. *Fusion of Evidential CNN Classifiers for Image Classification*. In T. Denœux, E. Lefèvre, Zh. Liu and F. Pichon (Eds), *Belief Functions : Theory and Applications*, Springer International Publishing, Cham, pp 168–176, 2021.

Classifieur évidentiel « profond »

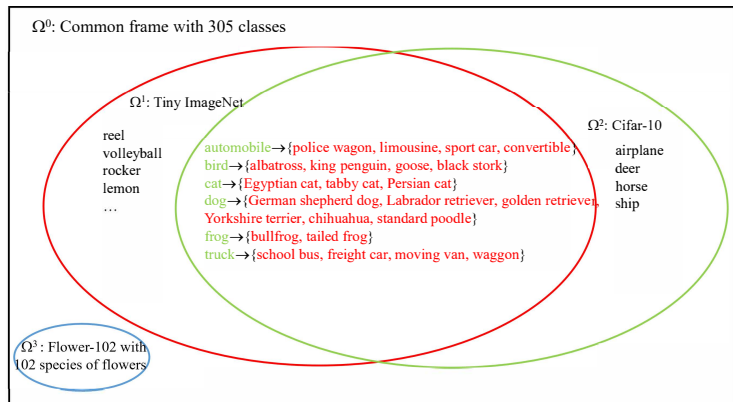


Fusion de classifieurs pour la classification d'images



Bases de données

Nom	# classes	# exemples d'app.	# exemples de test
Tiny ImageNet	200	10^5	10^4
Flower-102	102	4080	4129
CIFAR-10	10	5×10^4	10^4



Approche

- Un classifieur évidentiel est entraîné sur chacun des trois ensembles d'apprentissage.
- Chaque classifieur a son propre cadre de discernement, contenant l'ensemble des classes représentées dans l'ensemble d'apprentissage correspondant auquel on adjoint **une classe ω_0 correspondant à tout le reste** (les trois cadres de discernement sont alors compatibles).
- Les sorties des classifieurs sont étendues au plus petit raffinement commun Ω^0 et combinées par la règle de Dempster.
- Optionnellement, les trois classifieurs peuvent être réentraînés simultanément sur l'ensemble des données (on a alors un problème d'apprentissage partiellement supervisé).




Méthodes de fusion alternatives

Probability-to-mass fusion (PMF) ¹² réseaux probabilistes (sortie softmax), combinaison des probabilités (après extension à Ω^0) par la règle de Dempster.

Bayesian-fusion (BF) : réseaux probabilistes (sortie softmax), probabilités calculées sur Ω^0 en utilisant le principe d'indifférence, combinaison par Dempster)

Probabilistic feature-combination (PFC) : concaténation des vecteurs de caractéristiques extraits par les trois réseaux + couche softmax

Evidential feature-combination (EFC) : concaténation des vecteurs de caractéristiques extraits par les trois réseaux + couche DS




12. Ph. Xu, F. Davoine, J.-B. Bordes, H. Zhao and Th. Denœux. Multimodal Information Fusion for Urban Scene Understanding. *Machine Vision and Applications* 27(3), 331–349, 2016.   

Résultats

Classifieur	Tiny ImageNet	Flower-102	CIFAR-10	Overall
E-ResNet-101	18.66	4.68	4.61	-
P-ResNet-101 ¹³	18.70	4.69	4.66	-
MFE-ResNet-101	18.52	4.68	<u>3.94</u>	<u>10.31</u>
PMF-ResNet-101	18.54	4.69	4.42	10.40
BF-ResNet-101	19.18	5.07	6.04	11.10
E2E MFE-ResNet-101	<u>18.50</u>	4.67	3.82	10.27
E2E PMF-ResNet-101	18.49	<u>4.68</u>	4.28	10.35
E2E BF-ResNet-101	18.87	4.99	5.74	10.89
E2E PFC-ResNet-101	18.59	5.74	4.89	10.94
E2E EFC-ResNet-101	21.68	5.46	7.57	12.56

13. Y. Luo et al. Direction concentration learning : Enhancing congruency in machine learning. *IEEE Trans. PAMI* 43(6), 2021.

Interprétation





Instance/label	Before fusion			p' on Ω^0 after fusion
	p' from Tiny ImageNet	p' from CIFAR-10	p' from Flower102	
	p' (Egyptian cat) = 0.472 p' (chihuahua) = 0.511 ...	p' (cat) = 0.873 p' (dog) = 0.116 ...	p' (buttercup) = 0.001 p' (camellia) = 0 ...	p' (Egyptian cat) = 0.860 p' (chihuahua) = 0.125 ...
Egyptian cat	$p'(\omega_0^1) = 0.001$	$p'(\omega_0^2) = 0.001$	$p'(\omega_0^3) = 0.998$	$p'(\omega_0^0) = 0.001$
	p' (king penguin) = 0.453 p' (academic gown) = 0.532 ...	p' (bird) = 0.732 p' ({frog}) = 0.102 ...	p' (buttercup) = 0 p' (camellia) = 0.001 ...	p' (king penguin) = 0.988 p' (academic gown) = 0.006 ...
king penguin	$p'(\omega_0^1) = 0.001$	$p'(\omega_0^2) = 0.004$	$p'(\omega_0^3) = 0.993$	$p'(\omega_0^0) = 0.001$
	p' (bull frog) = 0.382 p' (tailed frog) = 0.602 ...	p' (frog) = 0.972 p' (cat) = 0.010 ...	p' (buttercup) = 0.001 p' (camellia) = 0 ...	p' (bull frog) = 0.388 p' (tailed frog) = 0.611 ...
bull frog	$p'(\omega_0^1) = 0$	$p'(\omega_0^2) = 0$	$p'(\omega_0^3) = 0.999$	$p'(\omega_0^0) = 0$

Conclusions

- La théorie des fonctions de croyance fournit un cadre formel, plus général que le cadre probabiliste, pour la quantification des incertitudes.
- Cette théorie peut être utilisée pour la quantification des incertitudes de prédiction en apprentissage, en ajoutant une **couche évidentielle** à des réseaux de neurones profonds.
- Cette approche permet, en particulier, de combiner des classifieurs dont les sorties sont exprimées dans des cadres de discernement différents.
- Pistes de recherche :
 - Proposer de nouvelles fonctions de coût adaptées au cas où les sorties du classifieur sont des fonctions de croyance
 - Application à la régression (suppose de manipuler des fonctions de croyance sur \mathbb{R}^p)

References

cf. <https://www.hds.utc.fr/~tdenoeux>

-  **T. Denœux**
Logistic Regression, Neural Networks and Dempster-Shafer Theory : a New Perspective
Knowledge-Based Systems 176 :54–67, 2019.
-  **L. Ma and T. Denœux**
Partial Classification in the Belief Function Framework
Knowledge-Based Systems 214 :106742, 2021.
-  **Z. Tong, Ph. Xu and T. Denœux**
An evidential classifier based on Dempster-Shafer theory and deep learning
Neurocomputing 450 :275–293, 2021.
-  **Z. Tong, Ph. Xu and T. Denœux**
Evidential fully convolutional network for semantic segmentation
Applied Intelligence 51 :6376–6399, 2021.