

Evidential Machine learning

Supervised and unsupervised learning using belief functions

Thierry Denœux

Université de technologie de Compiègne, Compiègne, France
and
Institut Universitaire de France, Paris, France

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

2nd International Conference on Digital Futures and
Transformative Technologies (ICoDT2)
May 25, 2022

Machine Learning

- **Machine Learning (ML)** started in the 1950's, but it has recently undergone important developments and immensely grown in popularity due to the advent of deep neural networks.
- Basically, deep networks make it possible to extract high-level (“semantic”) features from complex structured data (images, videos, texts, graphs, etc.).
- These features allow us to make **predictions** for classification or regression tasks, or to lay bare some **underlying structure** of the data (partition, hierarchy, etc.).
- One of the topical problems in ML is the **quantification of uncertainty**, including
 - Prediction uncertainty (supervised learning)
 - Cluster-membership uncertainty (unsupervised learning)

Uncertainty

- Main sources:
 - Randomness (aleatory uncertainty)
 - Lack of knowledge (epistemic uncertainty)
 - Conflict
- Theoretical frameworks:
 - Frequentist (confidence regions, p-values, etc.)
 - Bayesian (additive probabilities)
 - Imprecise probabilities (lower/upper previsions, etc.)
 - Fuzzy sets and possibility theory
 - Belief functions: Dempster-Shafer (DS) / Evidence theory
- Arguments for DS theory:
 - Extends both Bayesian and Possibility theories
 - Allows for the representation of aleatory and epistemic uncertainties
 - Well-suited for information fusion

Outline

- 1 Dempster-Shafer theory
 - Mass, belief and plausibility functions
 - Dempster's rule
- 2 Evidential machine learning
 - Evidential classification
 - Evidential clustering

Key features of DS theory

Generality: DS theory is based on the idea of **combining sets and probabilities**. It extends both

- Propositional logic, computing with sets (interval analysis)
- Probabilistic reasoning

All that can be done with sets or with probabilities alone can be done with belief functions, but DS theory can do much more!

Operationality: DS theory is easily put in practice by breaking down the available evidence into **elementary pieces of evidence**, and combining them by a suitable operator called **Dempster's rule of combination**.

Scalability: Contrary to a widespread misconception, evidential reasoning can be applied to **very large problems**.

Outline

- 1 Dempster-Shafer theory
 - Mass, belief and plausibility functions
 - Dempster's rule
- 2 Evidential machine learning
 - Evidential classification
 - Evidential clustering

Mass function

Definition (Mass function)

A *mass function* on a finite set Ω is a mapping $m : 2^\Omega \rightarrow [0, 1]$ such that

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

If $m(\emptyset) = 0$, m is said to be *normalized* (usually assumed).

Definition (Focal set)

Let m be a mass function on Ω . Every subset A of Ω such that $m(A) > 0$ is called *focal set* of m .

Interpretation

- Interpretation:
 - Ω is the set of possible answers to some question (called the **frame of discernment**)
 - Mass function m describes a **piece of evidence/information** pertaining to that question
 - Each mass $m(A)$ represents a share of a unit mass of belief allocated to focal set A , and which **cannot be allocated to any strict subset of A** .
- Example: consider an object recognition task, and

$$\Omega = \{\text{pedestrian, car, motorcycle, tree}\}$$

A sensor tells us that the object is a vehicle, and this information is 80% reliable. This information (evidence) can be represented by the following mass function:

$$m(\{\text{car, motorcycle}\}) = 0.8, \quad m(\Omega) = 0.2$$

Special cases

- If $m(A) = 1$ for some $A \subseteq \Omega$, m is said to be **logical**. It represents pure imprecision.
- The mass function m_0 such that $m_0(\Omega) = 1$ is said to be **vacuous**. It corresponds to complete ignorance.
- If $m(A) > 0 \Rightarrow |A| = 1$, m is said to be **Bayesian**. It can be used to represent aleatory uncertainty.

Belief and plausibility functions

Definition

Given a normalized mass function m on Ω , the *belief* and *plausibility* functions are defined, respectively, as

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

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

for all $A \subseteq \Omega$.

Interpretation:

- $Bel(A)$ is a measure of **total support** in A
- $Pl(A)$ is a measure of the **lack of support** in \bar{A} (or **consistency** with A)

Two-dimensional representation

- The uncertainty about a set of possibilities $A \subseteq \Omega$ is thus described by two numbers

$$(Bel(A), Pl(A)) \quad \text{with} \quad Bel(A) \leq Pl(A)$$

- Total ignorance (vacuous mass function):

$$(Bel(A), Pl(A)) = (0, 1), \quad \forall A \in 2^\Omega \setminus \{\Omega, \emptyset\}$$

- Infinitely precise information (Bayesian mass function):

$$Bel(A) = Pl(A)$$

Outline

- 1 Dempster-Shafer theory
 - Mass, belief and plausibility functions
 - Dempster's rule
- 2 Evidential machine learning
 - Evidential classification
 - Evidential clustering

Dempster's rule

In DS theory, Dempster's rule is the fundamental mechanism for combining belief functions representing **independent items of evidence**.

Definition (Orthogonal sum, degree of conflict)

Let m_1 and m_2 be two mass functions on the same frame Ω . Their **orthogonal sum** is the mass function defined by

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

for all $A \neq \emptyset$ and $(m_1 \oplus m_2)(\emptyset) := 0$. In (1), κ is the **degree of conflict** defined as

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

Properties

Proposition

- ① If several pieces of evidence are combined, *the order does not matter*:

$$m_1 \oplus m_2 = m_2 \oplus m_1$$

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

- ② A mass function m is *not changed if combined with the vacuous mass function m_0* :

$$m \oplus m_0 = m.$$

Misconception about Dempster's rule

- Following a 1979 report by Zadeh, it is repeated that “Dempster's rule yields counterintuitive results” (which is usually used as a justification to introduce new combination rules)
- Zadeh's example: $\Omega = \{a, b, c\}$, two experts

$$m_1(\{a\}) = 0.99, \quad m_1(\{b\}) = 0.01 \quad m_1(\{c\}) = 0$$

$$m_2(\{a\}) = 0, \quad m_2(\{b\}) = 0.01 \quad m_2(\{c\}) = 0.99$$

We get $(m_1 \oplus m_2)(\{b\}) = 1$, which is claimed to be “counterintuitive” because both experts considered b as very unlikely.

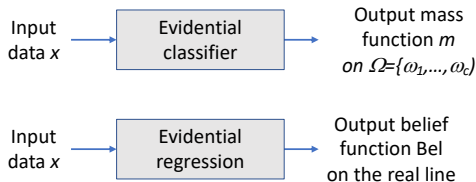
- But Expert 1 claims that c is absolutely impossible, and Expert 2 claims that a is absolutely impossible, so b is the only remaining possibility!
- Dempster's rule does produce sound results when used and interpreted correctly.

Outline

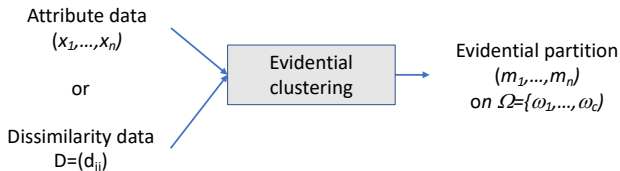
- 1 Dempster-Shafer theory
 - Mass, belief and plausibility functions
 - Dempster's rule
- 2 Evidential machine learning
 - Evidential classification
 - Evidential clustering

Main learning tasks

Supervised learning



Unsupervised learning



Outline

- 1 Dempster-Shafer theory
 - Mass, belief and plausibility functions
 - Dempster's rule
- 2 Evidential machine learning
 - Evidential classification
 - Evidential clustering

Application of DS theory to classification

- Two of the first papers applying DS theory to classification:



T. Denœux.

A k-nearest neighbor classification rule based on Dempster-Shafer theory.

IEEE Transactions on SMC, 25(05):804–813, 1995.



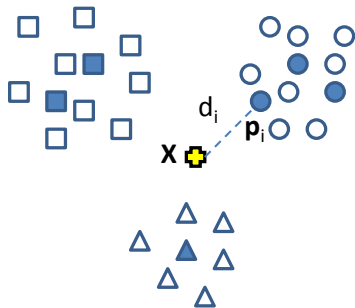
T. Denœux.

A neural network classifier based on Dempster-Shafer theory.

IEEE transactions on SMC A, 30(2):131–150, 2000.

- I will briefly recall the evidential neural network and describe some recent developments.

Evidential neural network classifier



- The learning set is summarized by r **prototypes**.
- Each prototype p_j has **membership degree** u_{ik} to each class ω_k , with $\sum_{k=1}^c u_{ik} = 1$.
- Each prototype p_j is a **piece of evidence** about the class of x ; its **reliability decreases with the distance d_j** between x and p_j .

Propagation equations

- Mass function induced by prototype \mathbf{p}_i :

$$m_i(\{\omega_k\}) = \alpha_i u_{ik} \exp(-\gamma_i d_i^2), \quad k = 1, \dots, c$$
$$m_i(\Omega) = 1 - \alpha_i \exp(-\gamma_i d_i^2)$$

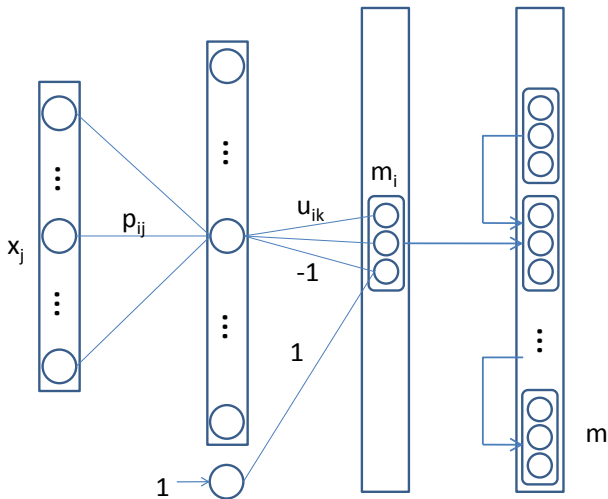
Remark: when $d_i \rightarrow +\infty$, $m_i \rightarrow m_0$.

- Combination:

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

- The focal sets of the combined mass function m are the singletons $\{\omega_k\}$, $k = 1, \dots, c$ and Ω .

Neural network implementation



Learning

- The parameters are the
 - The prototypes $\mathbf{p}_i, i = 1, \dots, r$ (rp parameters)
 - The membership degrees $u_{ik}, i = 1, \dots, r, k = 1 \dots, c$ (rc parameters)
 - The α_i and $\gamma_i, i = 1 \dots, r$ ($2r$ parameters).
- Let θ denote the vector of all parameters. It can be estimated by minimizing a **loss function** such as

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

where p_{ik} is the output plausibility of class ω_k for instance i , $y_{ik} = I(y_i = \omega_k)$, and λ is a regularization coefficient (hyperparameter).

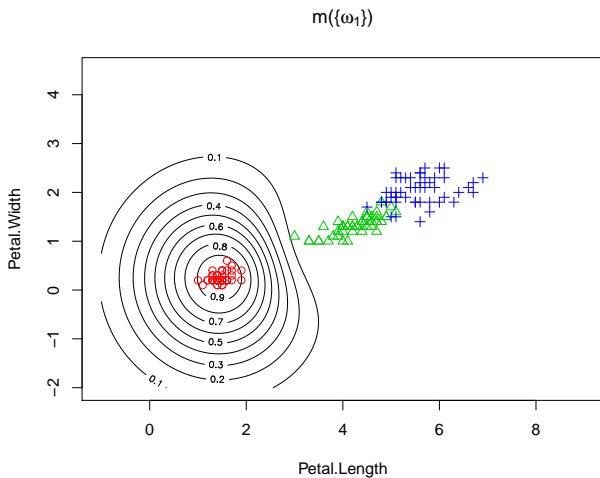
- The hyperparameter λ can be optimized by cross-validation.

Implementations

- **Matlab:** http://www.hds.utc.fr/~tdenoeux/software/belief_NN/belief_NN.zip
- **R package** `evclass`, **available at** <https://cran.r-project.org/web/packages/evclass/index.html>

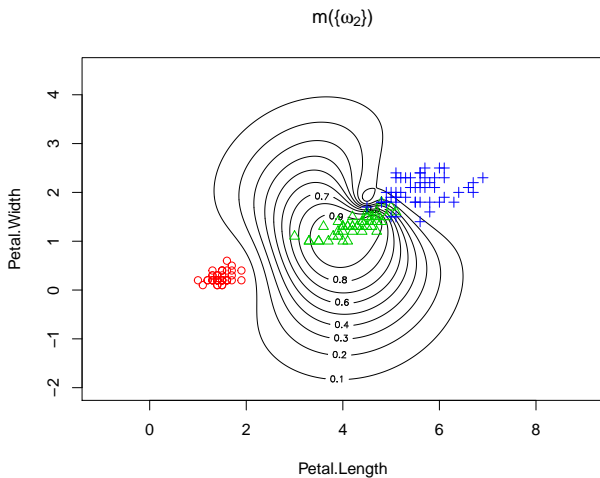
Results on the Iris data

Mass on $\{\omega_1\}$



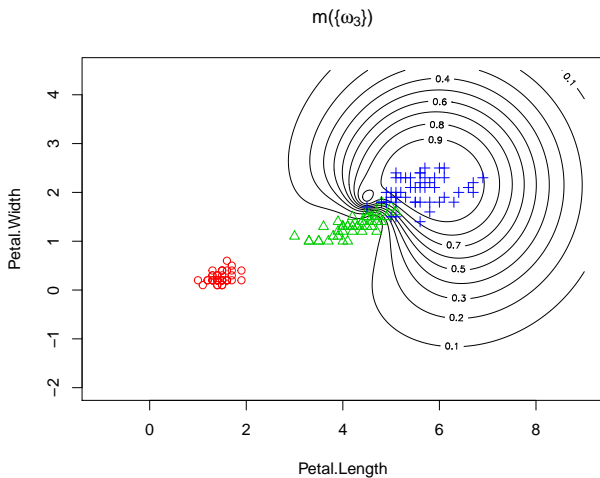
Results on the Iris data

Mass on $\{\omega_2\}$



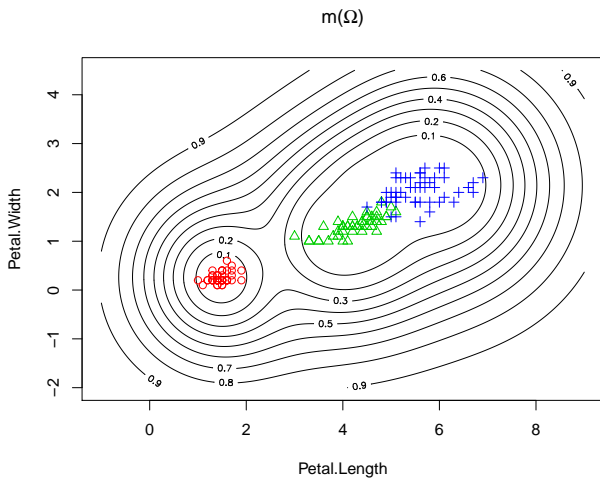
Results on the Iris data

Mass on $\{\omega_3\}$

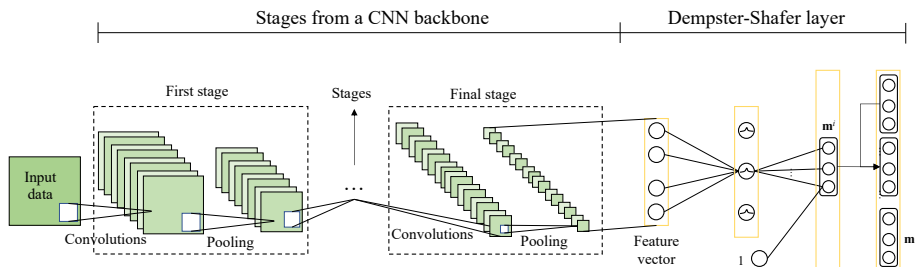


Results on the Iris data

Mass on Ω



Deep evidential classifier



Z. Tong, Ph. Xu and T. Denœux

An evidential classifier based on Dempster-Shafer theory and deep learning.

Neurocomputing 450:275–293, 2021.

Application to classifier fusion

- Training deep networks on very large datasets requires considerable computational resources
- For a given problem, there usually exist many smaller **heterogenous datasets** with different sets of classes and levels of granularity.
- Proposed approach: combine classifiers trained on different, heterogenous learning sets.

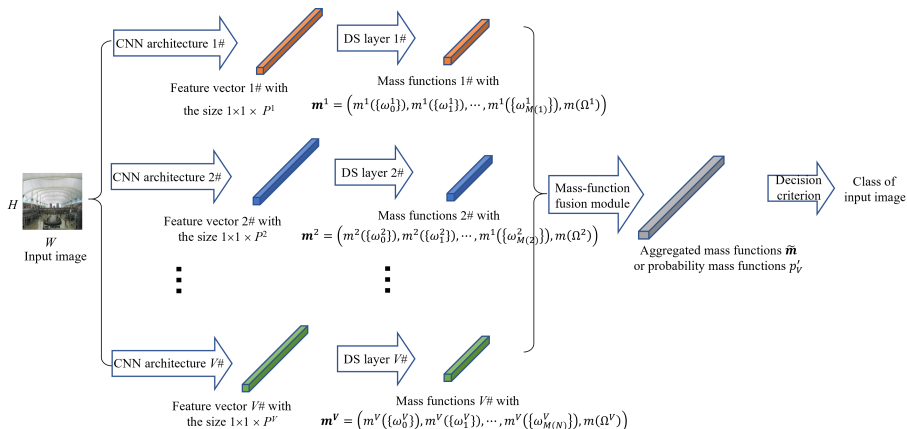


Z. Tong, Ph. Xu and T. Denœux

Fusion of Evidential CNN Classifiers for Image Classification.

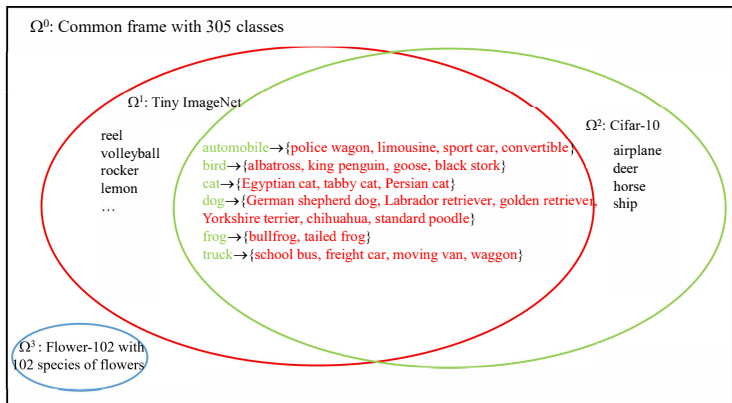
In T. Denœux et al. (Eds), *Belief Functions: Theory and Applications*, Springer International Publishing, Cham, pp 168–176, 2021.

Classifier fusion for image classification



Data sets

Dataset	# classes	# training instances	# test instances
Tiny ImageNet	200	10^5	10^4
Flower-102	102	4080	4129
CIFAR-10	10	5×10^4	10^4



Method

- An evidential classifier is trained on each of the learning sets.
- Each classifier has its own frame of discernment, containing the classes represented in its training set, and a class ω_0 representing everything else.
- The classifier outputs are expressed in a common refined frame Ω^0 and combined by Dempster's rule.
- Optionally, the classifiers can be fine-tuned together using the whole data set (partially supervised learning).




Alternative fusion methods

Probability-to-mass fusion (PMF)¹ probabilistic classifiers (softmax outputs), combination of probabilities (extended in Ω^0) by Dempster's rule

Bayesian fusion (BF): probabilistic classifiers (softmax outputs), probabilities computed in Ω^0 using Laplace's indifference principle, combination by Dempster's rule

Probabilistic feature combination (PFC): concatenation of the feature vectors + softmax layer

Evidential feature combination (EFC): concatenation of the feature vectors + DS layer


¹Ph. Xu, F. Davoine, J.-B. Bordes, H. Zhao and Th. Dencœur. Multimodal Information Fusion for Urban Scene Understanding. *Machine Vision and Applications* 27(3):331–349, 2016.   




Results

Classifier	Tiny ImageNet	Flower-102	CIFAR-10	Overall
E-ResNet-101	18.66	4.68	4.61	-
P-ResNet-101 ²	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

²Y. Luo et al. Direction concentration learning: Enhancing congruency in machine learning. *IEEE Trans. PAMI* 43(6), 2021.

Interpretation

Instance/label	Before fusion			p' on Ω^0 after fusion
	p' from Tiny ImageNet	p' from CIFAR-10	p' from Flower102	
 Egyptian cat	$p'(\text{Egyptian cat}) = 0.472$	$p'(\text{cat}) = 0.873$	$p'(\text{buttercup}) = 0.001$	$p'(\text{Egyptian cat}) = 0.860$
	$p'(\text{chihuahua}) = 0.511$	$p'(\text{dog}) = 0.116$	$p'(\text{camellia}) = 0$	$p'(\text{chihuahua}) = 0.125$

 king penguin	$p'(\omega_0^1) = 0.001$	$p'(\omega_0^2) = 0.001$	$p'(\omega_0^3) = 0.998$	$p'(\omega_0^0) = 0.001$
	$p'(\text{king penguin}) = 0.453$	$p'(\text{bird}) = 0.732$	$p'(\text{buttercup}) = 0$	$p'(\text{king penguin}) = 0.988$
	$p'(\text{academic gown}) = 0.532$	$p'(\{\text{frog}\}) = 0.102$	$p'(\text{camellia}) = 0.001$	$p'(\text{academic gown}) = 0.006$
 bull frog	$p'(\omega_0^1) = 0.001$	$p'(\omega_0^2) = 0.004$	$p'(\omega_0^3) = 0.993$	$p'(\omega_0^0) = 0.001$
	$p'(\text{bull frog}) = 0.382$	$p'(\text{frog}) = 0.972$	$p'(\text{buttercup}) = 0.001$	$p'(\text{bull frog}) = 0.388$
	$p'(\text{tailed frog}) = 0.602$	$p'(\text{cat}) = 0.010$	$p'(\text{camellia}) = 0$	$p'(\text{tailed frog}) = 0.611$
 bull frog
	$p'(\omega_0^1) = 0$	$p'(\omega_0^2) = 0$	$p'(\omega_0^3) = 0.999$	$p'(\omega_0^0) = 0$

Outline

- 1 Dempster-Shafer theory
 - Mass, belief and plausibility functions
 - Dempster's rule
- 2 Evidential machine learning
 - Evidential classification
 - Evidential clustering

Evidential clustering

- Several **soft clustering** methodologies to have been proposed over the years:
 - **Fuzzy clustering**: $u_{ik} \in [0, 1]$, $\sum_{k=1}^c u_{ik} = 1$
 - **Possibilistic clustering**: $u_{ik} \in [0, 1]$
 - **Rough clustering**: $(\underline{u}_{ik}, \bar{u}_{ik}) \in \{0, 1\}^2$, with $\underline{u}_{ik} \leq \bar{u}_{ik}$, $\sum_{k=1}^c \underline{u}_{ik} \leq 1$ and $\sum_{k=1}^c \bar{u}_{ik} \geq 1$
- **Evidential clustering** generalizes and unifies these approaches. First references:



T. Denœux and M.-H. Masson.

EVCLUS: Evidential Clustering of Proximity Data.

IEEE Transactions on Systems, Man and Cybernetics B
34(1):95-109, 2004.



M.-H. Masson and T. Denœux.

ECM: An evidential version of the fuzzy c-means algorithm.

Pattern Recognition 41(4):1384–1397, 2008.

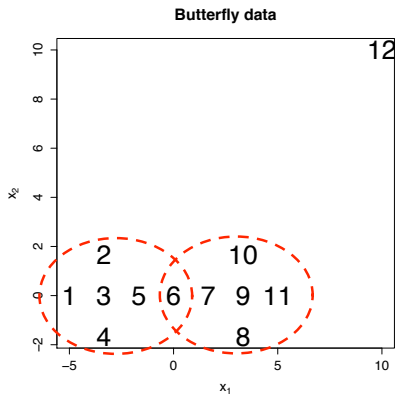
Credal partition

- Let $O = \{o_1, \dots, o_n\}$ be a set of n objects and $\Omega = \{\omega_1, \dots, \omega_c\}$ be a set of c groups (clusters).
- Assumption: each object o_i belongs to **at most one group**.

Definition

A *credal/evidential partition* is an n -tuple $M := (m_1, \dots, m_n)$, where each m_i is a mass function on Ω representing *uncertain knowledge* about the cluster membership of object o_i .

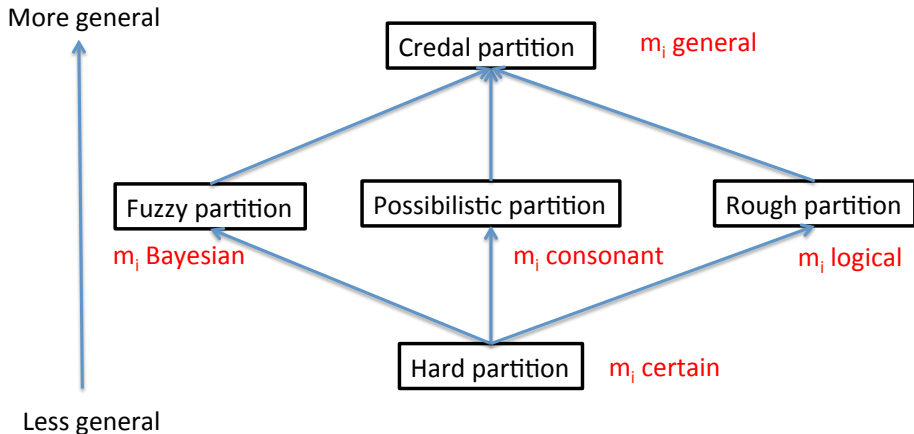
Example



Credal partition

	\emptyset	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_1, \omega_2\}$
m_3	0	1	0	0
m_5	0	0.5	0	0.5
m_6	0	0	0	1
m_{12}	0.9	0	0.1	0

Relationship with other clustering structures



Evidential clustering algorithms

- 1 **Evidential *c*-means (ECM)³:**
 - Attribute data
 - HCM, FCM family
- 2 **EVCLUS⁴:**
 - Attribute or proximity (possibly non metric) data
 - Multidimensional scaling approach
- 3 **Bootstrapping approach⁵**
 - Based on a mixture models and bootstrap confidence intervals
 - The resulting credal partition has frequentist properties

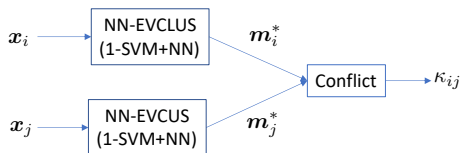
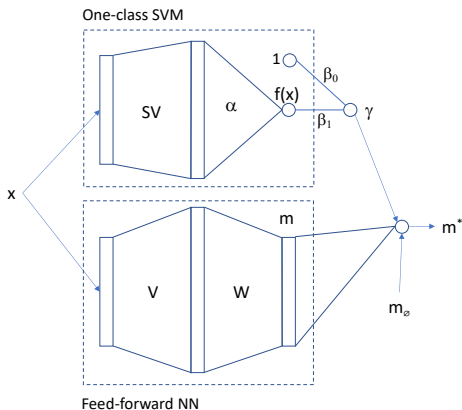
All these algorithms are implemented in the R package `evclust`, see <https://cran.r-project.org/web/packages/evclust/index.html>

³M.-H. Masson and T. Denœux. ECM: An evidential version of the fuzzy *c*-means algorithm. *Pattern Recognition* 41(4):1384–1397, 2008.

⁴T. Denœux *et al.* Evidential clustering of large dissimilarity data. *KBS* 106:179–195, 2016.

⁵T. Denœux. Calibrated model-based evidential clustering using bootstrapping. *Information Sciences* 528:17–45, 2020.

NN-EVCLUS



Loss function:

$$J(M) = \sum_{i < j} (\kappa_{ij} - d_{ij})^2$$

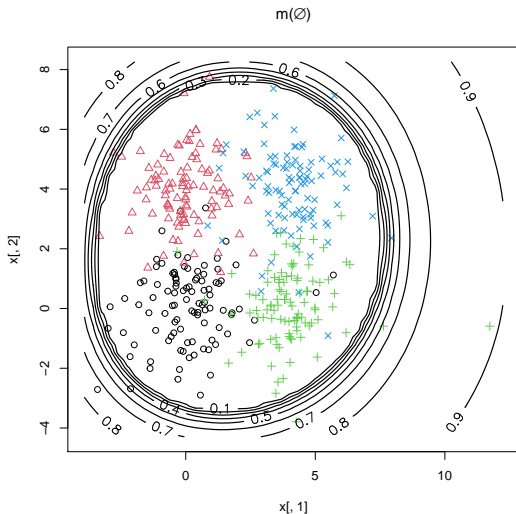


T. Denœux

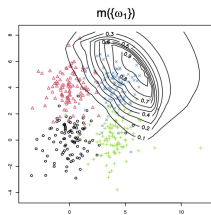
NN-EVCLUS: Neural Network-based Evidential Clustering.

Information Sciences, 572:297–330, 2021.

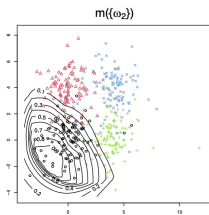
Example: mass on the emptyset



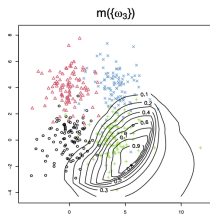
Example: masses on singletons and pairs



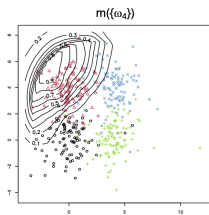
(a)



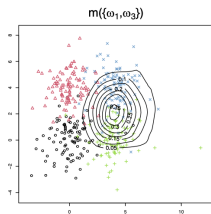
(b)



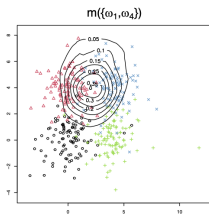
(c)



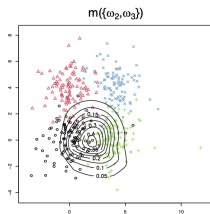
(d)



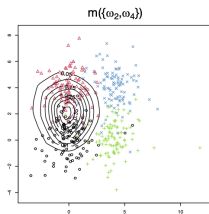
(e)



(f)



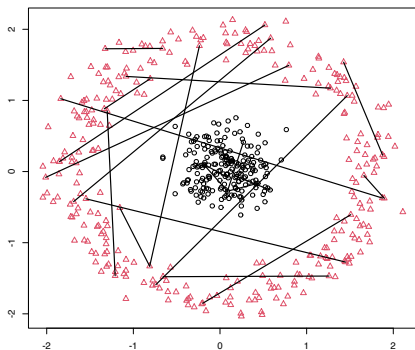
(g)



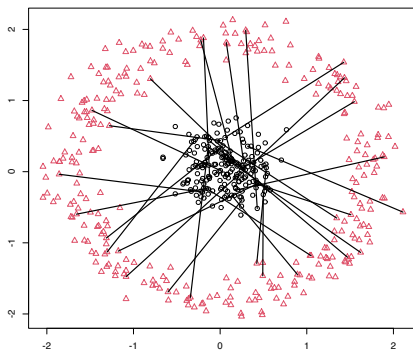
(h)

Constrained clustering

Must-link

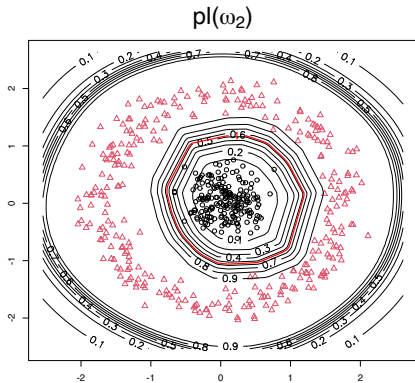
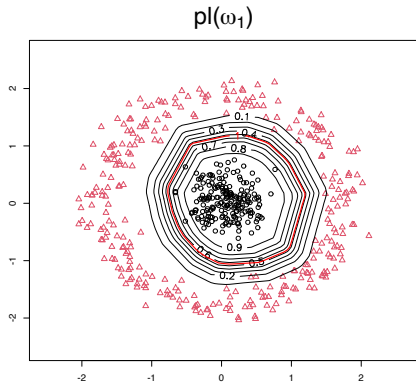


Cannot-link



Constrained clustering

Result



Summary

- Until recently, ML has been mostly based on probability theory. As a more general model, DS theory offers a **radically new and promising approach to uncertainty quantification in ML**.
- Other applications of belief functions in ML include
 - Classifier/clusterer ensembles
 - Partially labeled data
 - Regression
 - Multilabel classification
 - Partial classification
 - Transfer learning
 - Preference learning, etc.
- Many classical ML techniques can be **revisited from a DS perspective**, with important implications in terms of
 - Interpretation
 - Decision strategies
 - Model combination, etc.

References I

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.



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

References II

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

-  Z.-G. Liu, L.-Q. Huang, K. Zhou, and T. Denœux.
Combination of Transferable Classification with Multisource Domain Adaptation Based on Evidential Reasoning
IEEE Transactions on Neural Networks and Learning Systems 32:5: 2015–2029, 2021.
-  L. Ma and T. Denœux.
Partial Classification in the Belief Function Framework
Knowledge-Based Systems 214:106742, 2021.
-  T. Denœux.
Calibrated model-based evidential clustering using bootstrapping
Information Sciences 528:17–45, 2020.
-  T. Denœux
NN-EVCLUS: Neural Network-based Evidential Clustering
Information Sciences 572:297–330, 2021.