Evidential Machine learning

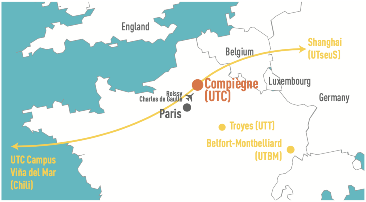
Thierry Denœux

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

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

University of Electronic Science and Technology of China
June 23, 2021

Compiègne University of Technology



- Founded in 1972, has played a pioneering role for bridging industry and education
- 4 500 engineering students, 300 PhD students
- 9 research laboratories (4 associated with CNRS)
- 120 start-ups

Sorbonne University Association

- Six academic institutions
- Four national research organisations

- 7,700 academic staff (2,900 tenured professors)
- 57,800 students including 23,000 undergraduates

International Journal of Approximate Reasoning

Home > Journals > International Journal of Approximate Reasoning

ISSN: 0888-613X

International Journal of Approximate Reasoning

Uncertainty in Intelligent Systems

Editor-in-Chief: Thierry Denoeux, PhD

> View Editorial Board

> **CiteScore: 7.1** ⓘ **Impact Factor: 2.678** ⓘ

- Submit Your Paper
- Supports Open Access
- View Articles
- Guide for Authors
- Abstracting/ Indexing
- Track Your Paper
- Order Journal
- Journal Metrics

The *International Journal of Approximate Reasoning* is intended to serve as a forum for the treatment of **imprecision** and **uncertainty** in **Artificial** and **Computational Intelligence**, covering both the foundations of uncertainty theories, and the design of intelligent systems for scientific and engineering applications. It publishes high-quality research papers describing theoretical developments or innovative applications, as well as review articles on topics of general interest.

Relevant topics include, but are not limited to, probabilistic **reasoning** and Bayesian networks, imprecise probabilities, random sets, belief functions (Dempster-Shafer theory), possibility theory, fuzzy sets, rough sets, decision theory, non-additive measures and integrals, qualitative reasoning about uncertainty, comparative probability orderings, game-theoretic probability, default reasoning, nonstandard logics, argumentation systems, inconsistency tolerant reasoning, elicitation techniques, philosophical foundations and psychological models of uncertain reasoning.

Domains of application for **uncertain reasoning systems** include risk analysis and assessment, information retrieval and database design, information fusion, machine learning, data and web mining, computer vision, image and signal

Theory of belief functions

- Also referred to as
  - Dempster-Shafer (DS) theory
  - Evidence theory
  - Transferable Belief Model
- Originates from Dempster's seminal work on statistical inference in the late 1960's. Formalized by Shafer in his seminal 1976 book. Further developed and popularized by Smets in the 1990's and early 2000's.
- DS theory has a level of generality that makes it applicable to a wide range problems involving uncertainty. It has been applied in may areas, including statistical inference, knowledge representation, information fusion, etc.
- This talk focuses on applications to machine learning (ML).

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

Everything than 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

Outline

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 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

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

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.

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

Special cases

A mass function is said to be

- Logical if it has only one focal set
- Vacuous if it has only one focal set and that focal set is Ω (represents total ignorance)
- Bayesian if all focal sets are singletons ($\sim$ probability distribution)
- Consonant if the focal sets are nested

Belief and plausibility functions

Definition

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

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

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

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 $\overline{A}$ (or consistency with A)

Two-dimensional representation

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

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

- 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):

$$(Bel(A), Pl(A)) = (p_A, p_A), \quad \text{with} \quad p_A \in [0, 1]$$

Outline

Dempster's rule

Definition (Degree of conflict)

Let m_1 and m_2 be two mass functions. Their degree of conflict is

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

Definition (Orthogonal sum)

Let m_1 and m_2 be two mass functions such that $\kappa < 1$. Their orthogonal sum is the mass function defined by

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

for all $A \neq \emptyset$ and $(m_1 \oplus m_2)(\emptyset) := 0$.

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:*

$$m \oplus m_0 = m.$$

3. *Let m_1 and m_2 be two mass functions with plausibility functions Pl_1 and Pl_2. We have*

$$(Pl_1 \oplus Pl_2)(\{\omega\}) \propto Pl_1(\{\omega\})Pl_2(\{\omega\}) \quad \forall \omega \in \Omega$$

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.

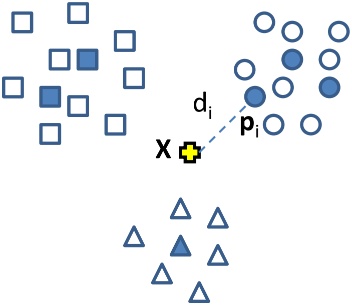
Outline

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.

Outline

Principle



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

Propagation equations

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

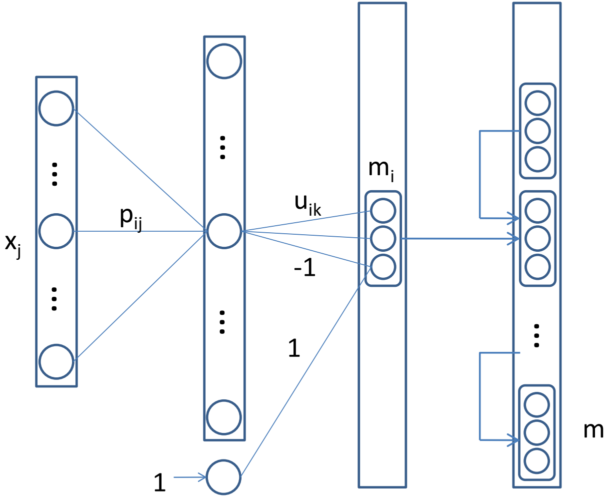
$$m_i(\{\omega_k\}) = \alpha_i u_{ik} \exp(-\gamma_i d_i^2), \quad k = 1, \ldots, c$$
$$m_i(\Omega) = 1 - \alpha_i \exp(-\gamma_i d_i^2)$$

- Combination:

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

- The combined mass function m has as focal sets the singletons $\{\omega_k\}$, $k = 1, \ldots, c$ and Ω.

Neural network implementation



Learning

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

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

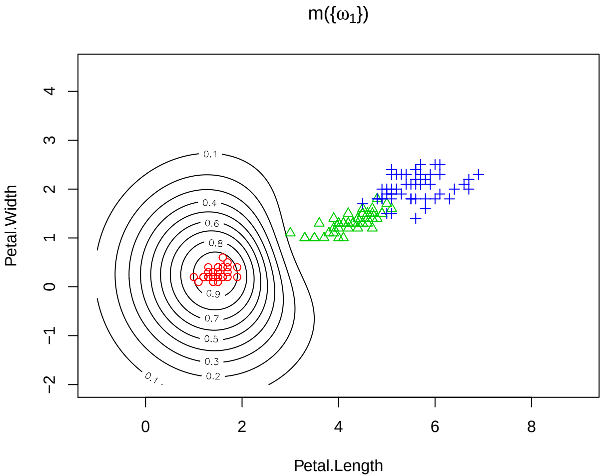
  where pl_{ik} is the output plausibility of class ω_k for instance i, $y_{ik} = I(y_i = \omega_k)$, and λ is a regularization coefficient (hyperparameter).
- The hyperparameter μ can be optimized by cross-validation.

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`

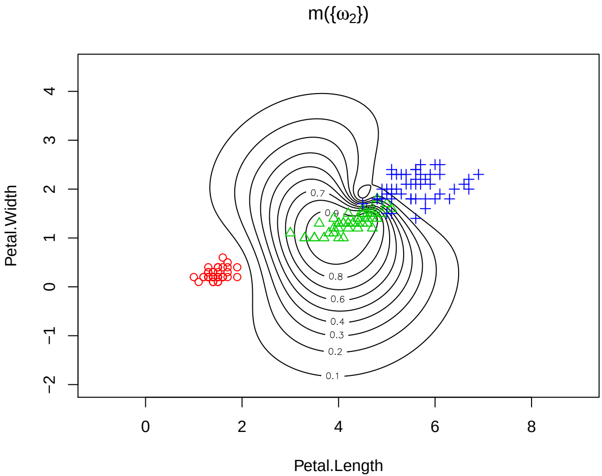
Results on the Iris data

Mass on $\{\omega_1\}$



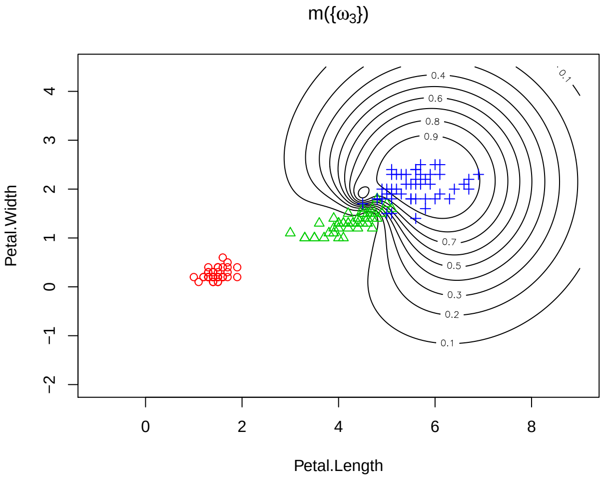
Results on the Iris data

Mass on $\{\omega_2\}$



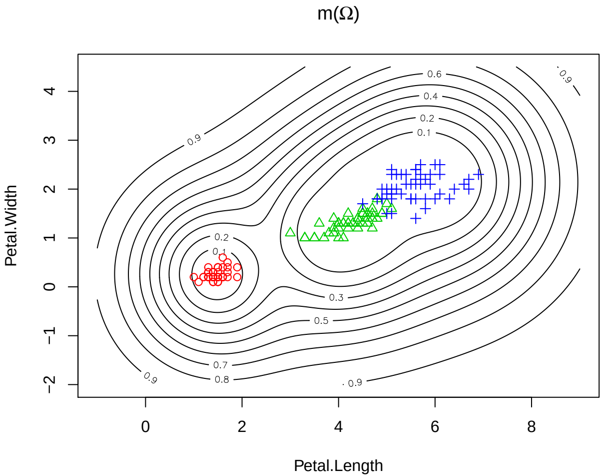
Results on the Iris data

Mass on $\{\omega_3\}$

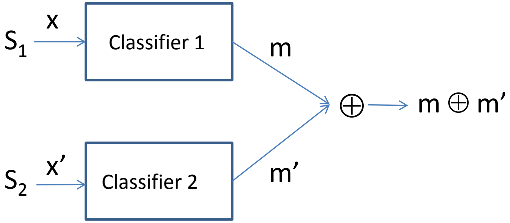


Results on the Iris data

Mass on Ω

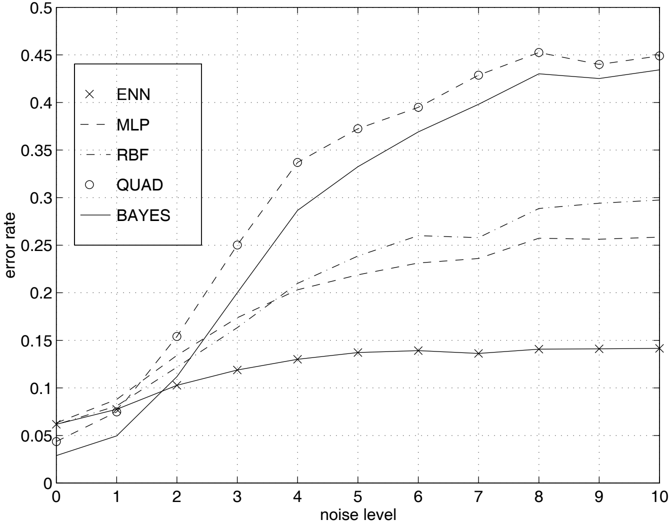


Data fusion example



- $c = 2$ classes
- Learning set ($n = 60$): $\boldsymbol{x} \in \mathbb{R}^5$, $\boldsymbol{x}' \in \mathbb{R}^3$, Gaussian distributions, conditionally independent
- Test set (real operating conditions): $\boldsymbol{x} \leftarrow \boldsymbol{x} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$

Results


ENN
MLP
RBF
QUAD
BAYES
error rate
noise level
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
1
2
3
4
5
6
7
8
9
10

Outline

Recent extensions

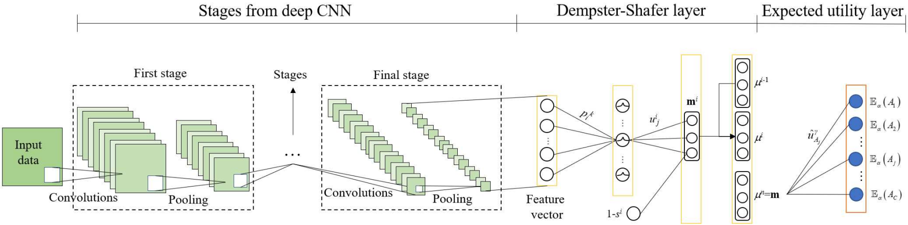
- Recent work on applying the previous ideas to modern convolutional networks (CNNs) for image classification and semantic segmentation:

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

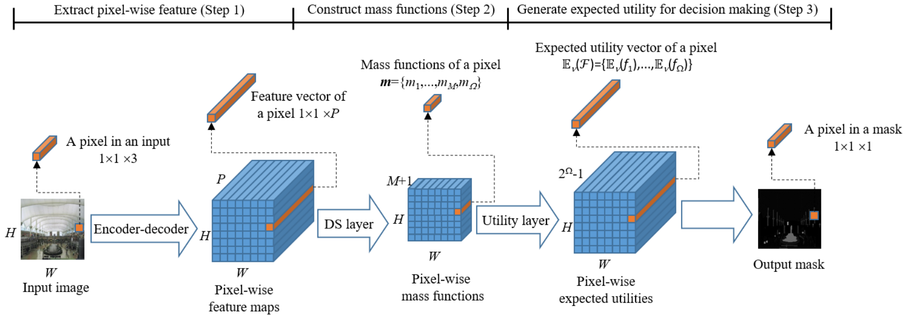
Z. Tong, Ph. Xu and T. Denœux.
Evidential fully convolutional network for semantic segmentation.
Applied Intelligence, 2021.
https://doi.org/10.1007/s10489-021-02327-0

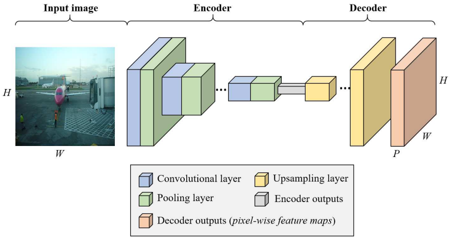
- Basic idea: plug in a “DS layer” at the output of a deep neural network architecture composed of convolutional layers.

Evidential CNN for image classification



Evidential FCN for semantic segmentation





Partial classification

- In partial classification, the input pattern is assigned to a set of classes, instead of being assigned to a single class.
- Partial classification is a refinement of classification with rejection: it makes it possible to decrease the error rate while still providing informative decisions most of the time.
- Decision and partial classification in the DS framework:

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

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

Partial classification in the DS framework

Definition of utilities

- As in standard classification, we start with a utility matrix $\boldsymbol{U} = (u_{ij})$, where u_{ij} is the utility of selecting class ω_i if the true class is ω_j.
- Matrix $\boldsymbol{U}$ is extended to set assignments by defining the utility of selecting set $K \subseteq \Omega$ if the true class is ω_j as

$$\widetilde{u}_{K,j} = \mathrm{OWA}_{\boldsymbol{w}}\left(\{u_{ij} : \omega_i \in K\}\right) = \sum_{k=1}^{|K|} w_k u^K_{(k)j}$$

  where $u^K_{(k)j}$ denotes the k-th largest element in the set $\{u_{ij},\ \omega_i \in K\}$ and $\sum_k w_k = 1$.
- The weight vector is determined to maximize the entropy subject to

$$\sum_{k=1}^{|K|} \frac{|K|-k}{|K|-1} w_k = \gamma,$$

- $\gamma = 0.5$ gives the average (minimum tolerance degree), $\gamma = 1$ gives the maximum (maximum tolerance degree).

Partial classification in the DS framework

Decision

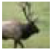
- Let $\mathcal{K} \subseteq 2^{\Omega}$ be a set of subsets of classes that can be selected (it includes singletons, and sets of "similar" classes).
- Let m be a mass function on Ω computed by the classifier.
- Hurwicz criterion: select the set of classes with maximum expected utility, defined as

$$\mathbb{E}_{\rho}(K) = \underbrace{\rho \sum_{A \subseteq \Omega} m(A) \min_{\omega_j \in A} \widetilde{u}_{K,j}}_{\text{lower expected utility}} + (1-\rho) \underbrace{\sum_{A \subseteq \Omega} m(A) \max_{\omega_j \in A} \widetilde{u}_{K,j}}_{\text{upper expected utility}},$$

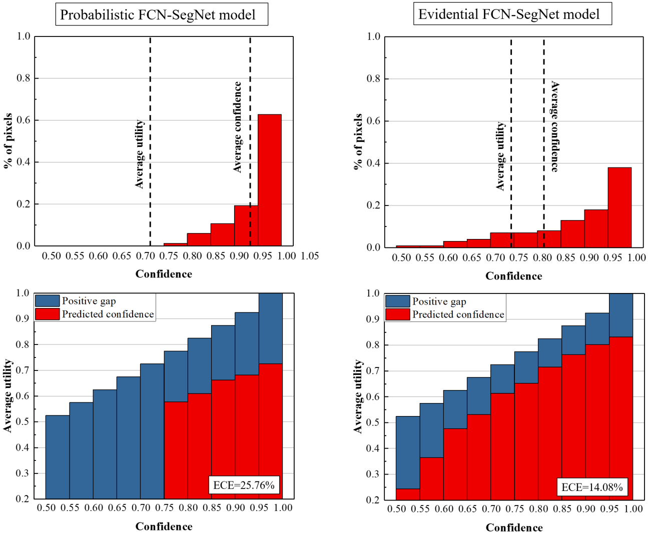
where ρ is a coefficient called the pessimism index.

Example

Label classification/utilities with different γ.

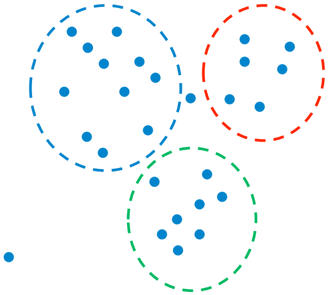
	#1(ω^*=cat)	#2(ω^*=dog)	#3(ω^*=deer)
γ=0.5	{dog}/0	{dog}/1	{deer}/1
γ=0.6	{cat,dog}/0.6	{cat,dog}/0.6	{deer}/1
γ=0.7	{cat,dog}/0.7	{cat,dog}/0.7	{deer,horse}/0.7
γ=0.8	{cat,dog}/0.8	{cat,dog}/0.8	{deer,horse}/0.8
γ=0.9	{cat,dog}/0.9	{cat,dog}/0.9	{cat,deer,dog,horse}/0.7104
γ=1.0	Ω/1.0	Ω/1.0	Ω/1.0
			

Calibration results (MIT-scene Parsing database)



Outline

Clustering



- n objects described by
  - Attribute vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ (attribute data) or
  - Dissimilarities (proximity data)
- Goals:
  1. Discover groups in the data
  2. Assess the uncertainty in group membership

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: $(\underline{u}_{ik}, \overline{u}_{ik}) \in \{0,1\}^2$, with $\underline{u}_{ik} \leq \overline{u}_{ik}$, $\sum_{k=1}^{c} \underline{u}_{ik} \leq 1$ and $\sum_{k=1}^{c} \overline{u}_{ik} \geq 1$
- 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.

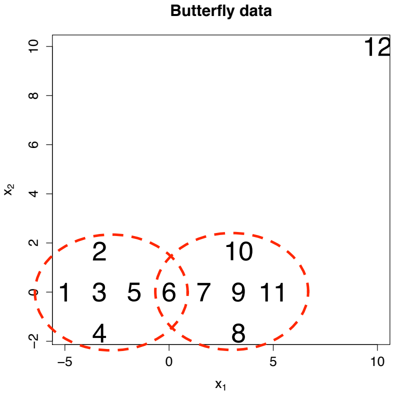
Credal partition

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

Definition

An credal partition is an n-tuple $M := (m_1, \ldots, m_n)$, where each m_i is a (not necessarily normalized) mass function on Ω representing evidence about the cluster membership of object o_i .

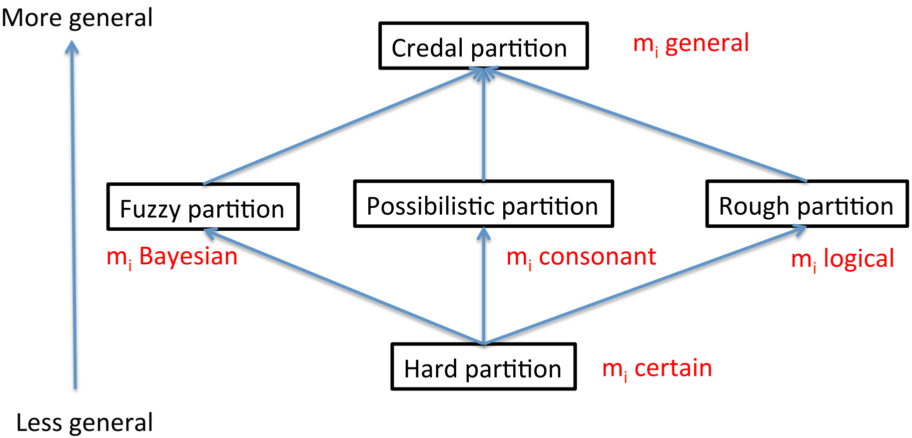
Example



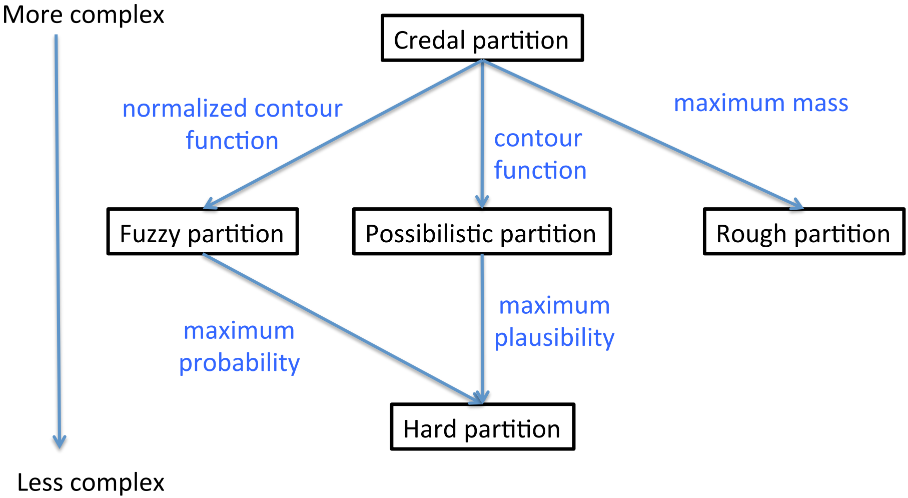
Credal partition

	$\emptyset$	$\{\omega_1\}$	$\{\omega_2\}$	$\{\omega_1, \omega_2\}$
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



Summarization of a credal partition



Evidential clustering algorithms

1. Evidential *c*-means (ECM)[1]:
   - Attribute data
   - HCM, FCM family
2. EVCLUS[2]:
   - Attribute or proximity (possibly non metric) data
   - Multidimensional scaling approach
3. Bootstrapping approach[3]
   - 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`

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

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

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

EVCLUS and variants

- First published evidential clustering algorithm:
  - 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.
- Improved and accelerated version:
  - T. Denœux, S. Sriboonchitta and O. Kanjanatarakul.
    Evidential clustering of large dissimilarity data.
    Knowledge-Based Systems, 106:179–195, 2016.
- Neural-network implementation:
  - T. Denœux
    NN-EVCLUS: Neural Network-based Evidential Clustering.
    Information Sciences, 572:297–330, 2021.

Outline

1. Dempster-Shafer theory
   - Mass, belief and plausibility functions
   - Dempster's rule
2. Evidential classification
   - Evidential neural network classifier
   - Deep evidential neural network classifiers
3. Evidential clustering
   - EVCLUS
   - NN-EVCLUS

Learning a credal partition from proximity data

- Problem: given the dissimilarity matrix $D = (d_{ij})$, how to build a relevant credal partition ?
- We need a model that relates cluster membership to dissimilarities.
- Basic idea: "The more similar two objects, the more plausible it is that they belong to the same group".
- How to formalize this idea?

Formalization

- Let m_i and m_j be mass functions regarding the group membership of objects o_i and o_j.
- We can show that the plausibility that objects o_i and o_j belong to the same group is

$$pl_{ij}(S_{ij}) = \sum_{A \cap B \neq \emptyset} m_i(A) m_j(B) = 1 - \kappa_{ij}$$

  where κ_{ij} = degree of conflict between m_i and m_j.
- Problem: find a credal partition $M = (m_1, \ldots, m_n)$ such that larger degrees of conflict κ_{ij} correspond to larger dissimilarities d_{ij}.

Cost function

- Approach: minimize the discrepancy between the dissimilarities d_{ij} and the degrees of conflict κ_{ij}.
- Example of a cost (stress) function:

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

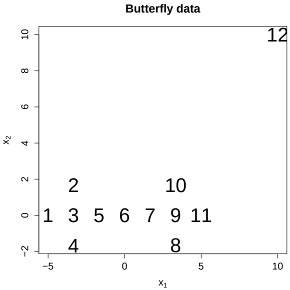
where φ is an increasing function from $[0, +\infty)$ to $[0, 1]$, for instance

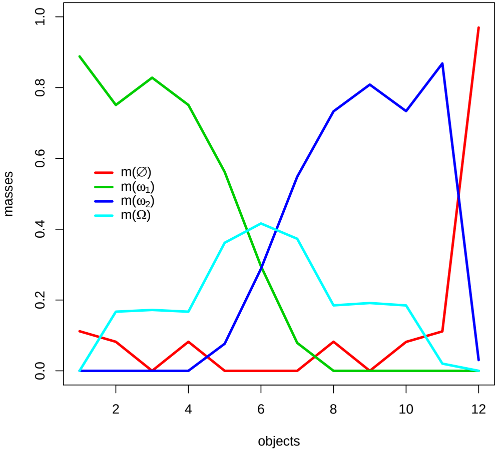
$$\varphi(d) = 1 - \exp(-\gamma d^2),$$

where γ is a scaling coefficient.

Butterfly example

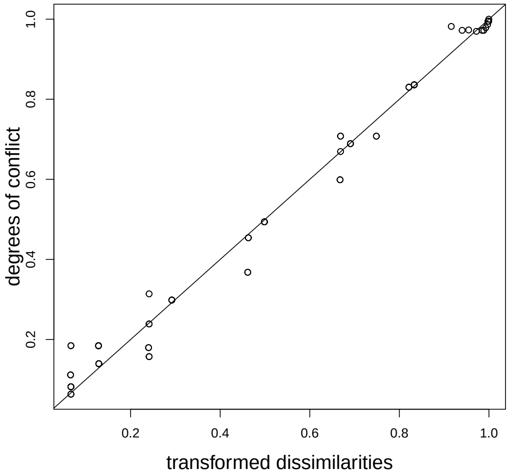
Credal partition





Butterfly example

Shepard diagram

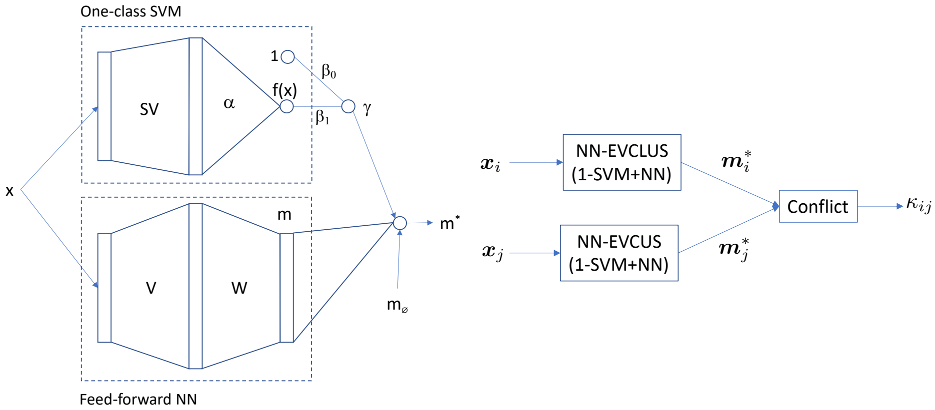


Advantages of EVCLUS

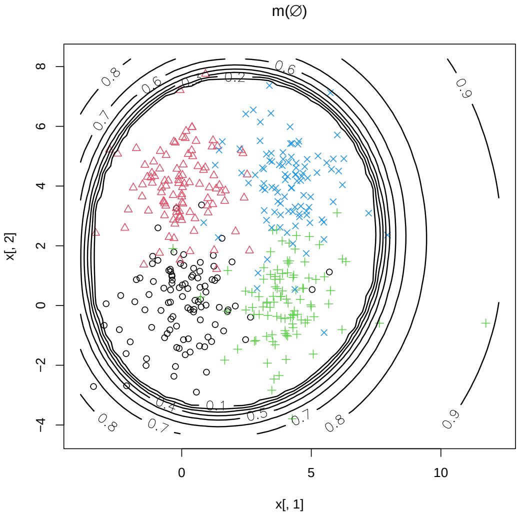
- Conceptually simple, clear interpretation.
- EVCLUS can handle nonmetric dissimilarity data (even expressed on an ordinal scale).
- It was also shown to outperform some of the state-of-the-art clustering techniques on proximity datasets.

Outline

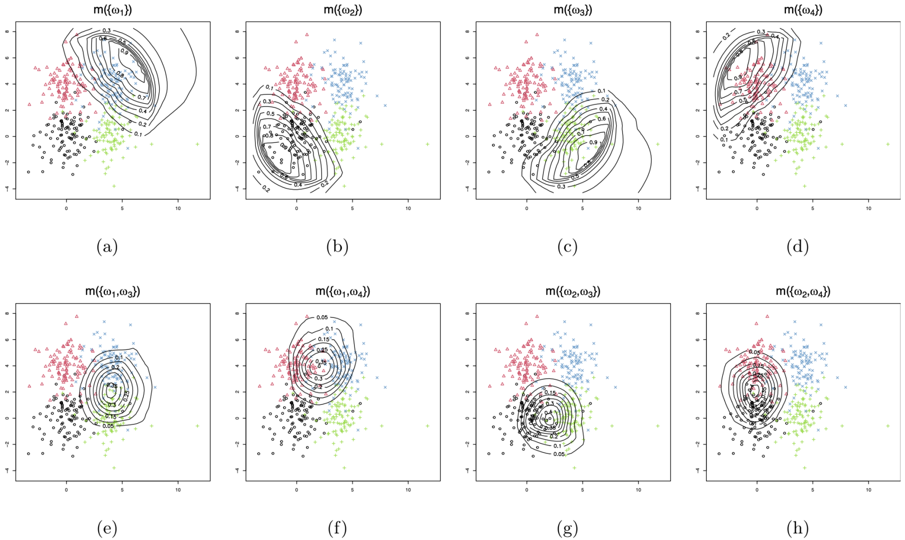
Principle of NN-EVCLUS



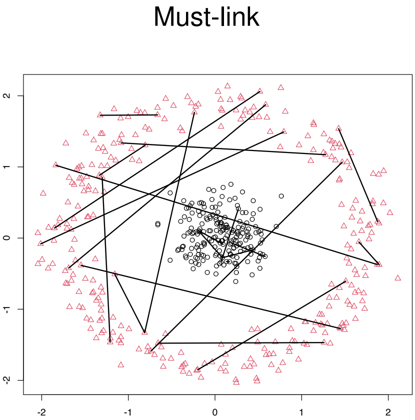
Example: mass on the emptyset

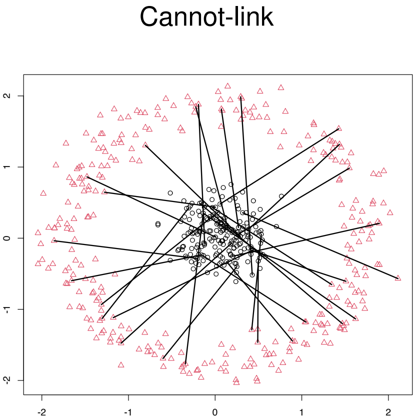


Example: masses on singletons and pairs



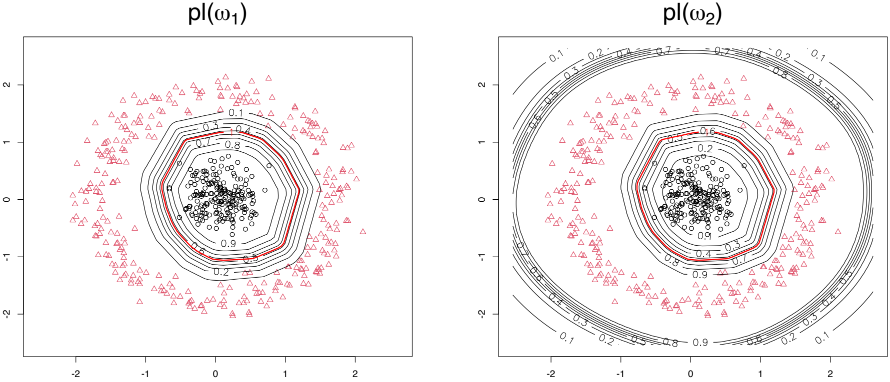
Constrained clustering





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
  - Preference learning, etc.
- Many classical ML techniques can be revisited from a DS perspective, with possible implications in terms of
  - Interpretation
  - Decision strategies
  - Model combination, etc.

References

T. Denœux, D. Dubois and H. Prade.
Representations of Uncertainty in Artificial Intelligence: Beyond Probability and Possibility
In P. Marquis, O. Papini and H. Prade (Eds), A Guided Tour of Artificial Intelligence Research, Volume 1, Chapter 4, Springer Verlag, pages 119-150, 2020.

Full text of all my papers and other resources at:

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

THANK YOU !