Thierry Denœux

1Université de Technologie de Compiègne
HEUDIASYC (UMR CNRS 6599)
http://www.hds.utc.fr/~tdenoeux

Spring School BFTA 2011
Autrans, April 4-8, 2011
Outline

1. Classification
   - Evidential $k$-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential $c$-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
The classification problem

- A population is assumed to be partitioned in $c$ groups or classes.
- Let $\Omega = \{\omega_1, \ldots, \omega_c\}$ denote the set of classes.
- Each instance is described by
  - A feature vector $\mathbf{x} \in \mathbb{R}^p$;
  - A class label $y \in \Omega$.
- Problem: given a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, predict the class label of a new instance described by $\mathbf{x}$. 

Thierry Denœux
What can we expect from belief functions?

- Problems with “weak” information:
  - Non exhaustive learning sets;
  - Learning and test data drawn from different distributions;
  - Partially labeled data (imperfect class information for training data), etc.

- Information fusion: combination of classifiers or clusterers trained using different data sets or different learning algorithms (ensemble methods).
Main belief function approaches

1. Approach 1: Convert the outputs from standard classifiers into belief functions and combine them using Dempster’s rule or any other alternative rule (e.g., Quost al., *IJAR*, 2011);

2. Approach 2: Develop evidence-theoretic classifiers directly providing belief functions as outputs:
   - **Generalized Bayes theorem**, extends the Bayesian classifier when class densities and priors are ill-known (Appriou, 1991; Denœux and Smets, *IEEE SMC*, 2008);
Outline

1. Classification
   - Evidential $k$-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential $c$-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
Evidential $k$-NN rule

**Principle**

- Let $\mathcal{N}_k(x) \subset \mathcal{L}$ denote the set of the $k$ nearest neighbors of $x$ in $\mathcal{L}$, based on some distance measure.
- Each $x_i \in \mathcal{N}_k(x)$ can be considered as a piece of evidence regarding the class of $x$.
- The strength of this evidence decreases with the distance $d_i$ between $x$ and $x_i$. 
The evidence of \((x_i, y_i)\) can be represented by the simple mass function:

\[ m_i(\{y_i\}) = \varphi(d_i) \]

\[ m_i(\Omega) = 1 - \varphi(d_i), \]

where \(\varphi\) is a decreasing function from \([0, +\infty)\) to \([0, 1]\) such that \(\lim_{d \to +\infty} \varphi(d) = 0\).

The evidence of the \(k\) nearest neighbors of \(x\) is pooled using Dempster’s rule of combination:

\[ m = \bigoplus_{x_i \in N_k(x)} m_i = \bigoplus_{x_i \in N_k(x)} \{y_i\}^{1-\varphi(d_i)} \]
Evidential $k$-NN rule

Decision-making

- Let $a_k$ be the act of assigning the object to $\omega_k$ and $A = \{a_1, \ldots, a_c\}$ the set of acts.
- Let $L(a_k, \omega_\ell)$ be the loss incurred if an object from class $\omega_\ell$ is assigned to $\omega_k$.
- **Decision rule:** assign the object to the class $\omega_{k^*}$ such that $\underline{R}(a_{k^*})$, $\overline{R}(a_{k^*})$ or $R_{bet}(a_{k^*})$ is minimized.
- The three rules are equivalent when $L(a_k, \omega_\ell) = 1$ if $k \neq \ell$ or 0 otherwise.
Evidential \(k\)-NN rule

Learning

- Example of function \(\varphi\):
  \[ \varphi(d) = \alpha \exp\left(-\gamma d^2\right). \]
  with \(\alpha \in (0, 1)\) and \(\gamma > 0\).

- Parameters \(\alpha\) and \(\gamma\) can be fixed heuristically or learned by minimizing an error function, e.g.:
  \[ E(\alpha, \gamma) = \sum_{i=1}^{n} \sum_{k=1}^{c} (p_k^{(-i)} - u_{ik})^2 \]
  where \(u_{ik} = 1\) if \(y_i = \omega_k\) and 0 otherwise, and \((p_k^{(-i)})\) is the pignistic probability of class \(\omega_k\) computed using the leave-one-out method.
Performance comparison (UCI database)

Sonar data

Ionosphere data

Test error rates as a function of $k$ for the voting (-), evidential (:), fuzzy (−) and distance-weighted (−.) $k$-NN rules.
We now consider a learning set of the form

$$\mathcal{L} = \{(x_i, m_i), i = 1, \ldots, n\}$$

where

- $x_i$ is the attribute vector for instance $i$, and
- $m_i$ is a mass function representing uncertain expert knowledge about the class $y_i$ of instance $i$.

Special cases:

- $m_i(\{\omega_k\}) = 1$ for all $i$: supervised learning;
- $m_i(\Omega) = 1$ for all $i$: unsupervised learning;
Each instance \((x_i, m_i)\) in \(\mathcal{L}\) is an item of evidence regarding \(y\), whose reliability decreases with the distance \(d_i\) between \(x\) and \(x_i\).

Each mass function \(m_i\) is discounted to produce a “weaker” mass function \(m'_i\):

\[
m'_i(A) = \phi(d_i) m_i(A), \quad \forall A \subset \Omega.
\]

\[
m'_i(\Omega) = 1 - \sum_{A \subset \Omega} m'_i(A).
\]

The \(k\) mass functions are combined using Dempster’s rule:

\[
m = \bigoplus_{x_i \in \mathcal{N}_k(x)} m'_i.
\]
Example: EEG data

EEG signals encoded as 64-D patterns, 50 % positive (K-complexes), 50 % negative (delta waves), 5 experts.
Results on EEG data
(Denoeux and Zouhal, 2001)

- $c = 2$ classes, $p = 64$
- For each learning instance $x_i$, the expert opinions were modeled as a mass function $m_i$.
- $n = 200$ learning patterns, $300$ test patterns

<table>
<thead>
<tr>
<th>$k$</th>
<th>$k$-NN</th>
<th>$w$</th>
<th>$k$-NN</th>
<th>Ev. $k$-NN (crisp labels)</th>
<th>Ev. $k$-NN (uncert. labels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0.30</td>
<td>0.30</td>
<td>0.31</td>
<td>0.31</td>
<td>0.27</td>
</tr>
<tr>
<td>11</td>
<td>0.29</td>
<td>0.30</td>
<td>0.29</td>
<td>0.29</td>
<td>0.26</td>
</tr>
<tr>
<td>13</td>
<td>0.31</td>
<td>0.30</td>
<td>0.31</td>
<td>0.31</td>
<td>0.26</td>
</tr>
</tbody>
</table>
Outline

1. Classification
   - Evidential $k$-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential $c$-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
The learning set is summarized by $r$ prototypes.

Each prototype $p_i$ has membership degree $u_{ik}$ to each class $\omega_k$, with $\sum_{k=1}^{c} u_{ik} = 1$.

Each prototype $p_i$ induces a piece of evidence regarding the class of $x$, whose reliability decreases with the distance $d_i$ between $x$ and $p_i$. 
Mass function induced by $p_i$:

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

Combination:

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

All parameters are learnt from data by minimizing an error function.
## Results on classical data

<table>
<thead>
<tr>
<th>Classifier</th>
<th>test error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-layer perceptron (88 units)</td>
<td>0.49</td>
</tr>
<tr>
<td>Radial Basis Function (528 units)</td>
<td>0.47</td>
</tr>
<tr>
<td>Gaussian node network (528 units)</td>
<td>0.45</td>
</tr>
<tr>
<td>Nearest neighbor</td>
<td>0.44</td>
</tr>
<tr>
<td>Linear Discriminant Analysis</td>
<td>0.56</td>
</tr>
<tr>
<td>Quadratic Discriminant Analysis</td>
<td>0.53</td>
</tr>
<tr>
<td>CART</td>
<td>0.56</td>
</tr>
<tr>
<td>BRUTO</td>
<td>0.44</td>
</tr>
<tr>
<td>MARS (degree=2)</td>
<td>0.42</td>
</tr>
<tr>
<td>Evidential NN (33 prototypes)</td>
<td>0.38</td>
</tr>
<tr>
<td>Evidential NN (44 prototypes)</td>
<td>0.37</td>
</tr>
<tr>
<td>Evidential NN (55 prototypes)</td>
<td>0.37</td>
</tr>
</tbody>
</table>

**Vowel data**

\[ c = 11, \quad p = 10 \]

\[ n = 568 \]

Test: 462 ex. (different speakers)
Data fusion example

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

\[
\begin{align*}
S_1 \rightarrow & \quad x \\
& \quad \text{Classifier 1} \quad m \\
\oplus & \quad m \oplus m' \\
S_2 \rightarrow & \quad x' \\
& \quad \text{Classifier 2} \quad m'
\end{align*}
\]
Results

Test error rates: $x + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2 I)$

![Graph showing error rates for different classifiers as a function of noise level. The graph includes lines for ENN, MLP, RBF, QUAD, and BAYES, with error rates plotted on the y-axis and noise levels on the x-axis.](image)
Outline

1. Classification
   - Evidential $k$-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential $c$-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
The clustering problem

- \( n \) objects described by
  - Attribute vectors \( \mathbf{x}_1, \ldots, \mathbf{x}_n \) (attribute data) or
  - Dissimilarities (proximity data).

- Goal: find a meaningful structure in the data set, usually a partition into \( c \) crisp or fuzzy subsets.

- The language of belief functions may allow us to extract richer information from the data using a more general data structure.
Different partition concepts

- **Hard partition**: each object belongs to one and only one group. Group membership is expressed by binary variables $u_{ik}$ such that $u_{ik} = 1$ if object $i$ belongs to group $k$ and $u_{ik} = 0$ otherwise.

- **Fuzzy partition**: each object has a degree of membership $u_{ik} \in [0, 1]$ to each group, with $\sum_{k=1}^{C} u_{ik} = 1$. The membership degrees $(u_{i1}, \ldots, u_{ic})$ define a probability distribution over the set $\Omega$ of groups.

- **Credal partition**: the group membership of each object is described by a *mass function* $m_i$ over $\Omega$. 
### Credal partition

#### Example

<table>
<thead>
<tr>
<th>A</th>
<th>$m_1(A)$</th>
<th>$m_2(A)$</th>
<th>$m_3(A)$</th>
<th>$m_4(A)$</th>
<th>$m_5(A)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\emptyset$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>${\omega_1}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td>${\omega_2}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>${\omega_1, \omega_2}$</td>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>${\omega_3}$</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
<td>0.4</td>
<td>0</td>
</tr>
<tr>
<td>${\omega_1, \omega_3}$</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>${\omega_2, \omega_3}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>0.3</td>
<td>0</td>
<td>0.3</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Hard and fuzzy partitions are recovered as special cases when all mass functions are certain or Bayesian, respectively.
Algorithms for computing a credal partition

- **EVCLUS** (Denoeux and Masson, 2004):
  - Proximity (possibly non metric) data,
  - Multidimensional scaling approach.

- **Evidential c-means (ECM)**: (Masson and Denoeux, 2008):
  - Attribute data,
  - HCM, FCM family (alternate optimization of a cost function).

- **Relational Evidential c-means (RECM)**: (Masson and Denoeux, 2009): ECM for proximity data.

- **Constrained Evidential c-means (CECM)** (Antoine et al., 2011): ECM with pairwise constraints.
Outline

1. Classification
   - Evidential $k$-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential $c$-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
Learning a Credal Partition from proximity data

Problem: given the dissimilarity matrix $D = (d_{ij})$, how to build a “reasonable” credal partition?

We need a model that relates class membership to dissimilarities.

Basic idea: “The more similar two objects, the more plausible it is that they belong to the same class”.

How to formalize this idea?
Let $m_i$ and $m_j$ be mass functions regarding the class membership of objects $o_i$ and $o_j$. The plausibility of the proposition $S_{ij}$: “objects $o_i$ and $o_j$ belong to the same class” can be shown to be equal to:

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

where $K_{ij} =$ degree of conflict between $m_i$ and $m_j$.

Problem: find $M = (m_1, \ldots, m_n)$ such that larger degrees of conflict $K_{ij}$ correspond to larger dissimilarities $d_{ij}$. 
**EVCLUS algorithm**

**Cost function**

- Approach: **minimize the discrepancy** between the dissimilarities $d_{ij}$ and the degrees of conflict $K_{ij}$.

- Example of a **cost function**:

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

  (assuming the $d_{ij}$ have been scaled to $[0, 1]$).

- $M$ can be determined by minimizing $J$ using an alternate directions method, solving a QP problem at each step.

- To reduce the complexity, focal sets are reduced to $\{\omega_k\}^c_{k=1}$, $\emptyset$, and $\Omega$. 
Butterfly example

one additional object (#1) similar to all other objects ("inlier")
Proximity matrix from the structural comparison of 213 protein sequences.

Each protein belongs to one of 4 classes of globins: hemoglobin-\(\alpha\) (HA), hemoglobin-\(\beta\) (HB), myoglobin (M) and heterogeneous globins (G).

Non-metric dissimilarities: most relational fuzzy clustering algorithms converge to a trivial solution.

EVCLUS recovers the true partition with only one error.
Advantages and drawbacks

- **Advantages**
  - Applicable to **proximity data** (not necessarily Euclidean, or even numeric).
  - **Robust** against atypical observations (similar or dissimilar to all other objects).
  - **Usually performs better** than relational fuzzy clustering procedures.

- **Drawback**: computational complexity (iterative optimization, limited to datasets of a few thousands of objects and less than 20 classes).

- **A more efficient procedure**: the Evidential c-means algorithm (Masson and Denoeux, 2008).
Outline

1. Classification
   - Evidential $k$-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential $c$-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
Problem: generate a credal partition $M = (m_1, \ldots, m_n)$ from attribute data $X = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^p$.

Generalization of hard and fuzzy $c$-means algorithms:
- Each class represented by a prototype;
- Alternate optimization of a cost function with respect to the prototypes and to the credal partition.
Fuzzy c-means (FCM)

- Minimize

\[ J_{\text{FCM}}(U, V) = \sum_{i=1}^{n} \sum_{k=1}^{c} u_{ik}^{\beta} d_{ik}^2 \]

where \( d_{ik} = ||x_i - v_k|| \) under the constraints \( \sum_k u_{ik} = 1, \forall i \).

- Alternate optimization algorithm:

\[ v_k = \frac{\sum_{i=1}^{n} u_{ik}^{\beta} x_i}{\sum_{i=1}^{n} u_{ik}^{\beta}} \quad \forall k = 1, \ldots, c, \]

\[ u_{ik} = \frac{d_{ik}^{-2/(\beta-1)}}{\sum_{\ell=1}^{c} d_{i\ell}^{-2/(\beta-1)}}. \]
ECM algorithm

Principle

- Each class $\omega_k$ represented by a prototype $v_k$.
- Each non empty set of classes $A_j$ represented by a prototype $\bar{v}_j$ defined as the center of mass of the $v_k$ for all $\omega_k \in A_j$.

Basic ideas:
- For each non empty $A_j \in \Omega$, $m_{ij} = m_i(A_j)$ should be high if $x_i$ is close to $\bar{v}_j$.
- The distance to the empty set is defined as a fixed value $\delta$. 
ECM algorithm
Objective criterion

- Criterion to be minimized:

\[ J_{\text{ECM}}(M, V) = \sum_{i=1}^{n} \sum_{\{j/A_j \neq \emptyset, A_j \subseteq \Omega\}} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^{2} + \sum_{i=1}^{n} \delta^{2} m_{i\emptyset}^{\beta}, \]

- Parameters:
  - \(\alpha\) controls the specificity of mass functions;
  - \(\beta\) controls the hardness of the evidential partition;
  - \(\delta\) controls the amount of data considered as outliers.

- \(J_{\text{ECM}}(M, V)\) can be iteratively minimized with respect to \(M\) and \(V\) using an alternate optimization scheme.
Butterfly dataset
4-class data set
4-class data set
Hard credal partition
4-class data set

Lower approximation

Thierry Denœux
4-class data set
Upper approximation
Brain data

Problem

- Magnetic resonance imaging of pathological brain, 2 sets of parameters.
- Three regions: normal tissue (Norm), ventricles + cerebrospinal fluid (CSF/V) and pathology (Path).
- Image 1 highlights CSF/V (dark), image 2 highlights pathology (bright).
Brain data
Results in grey level space

Diagram showing the results in grey level space with clusters represented by different symbols.
Brain data
Image segmentation

Pathology (left); CSF and ventricles (center); normal brain tissues (right). The lower approximations of the clusters are represented by light grey areas, the upper approximations by the union of light and dark grey areas.
Determining the number of groups

Validity index

- If a proper number of classes is chosen, the prototypes will be close to the cluster centers and most of the mass will be allocated to singletons of $\Omega$.
- On the contrary, if $c$ is too small or too high, the mass will be distributed to subsets with higher cardinality or to $\emptyset$.

**Nonspecificity** of a mass function:

$$N(m) \triangleq \sum_{A \in 2^\Omega \setminus \emptyset} m(A) \log_2 |A| + m(\emptyset) \log_2 |\Omega|,$$

- Proposed **validity index** of a credal partition:

$$N^*(c) \triangleq \frac{1}{n \log_2(c)} \sum_{i=1}^{n} N(m_i).$$

Thierry Denœux
Determining the number of groups
Result with the 4-class dataset

![Graph showing the number of clusters against different values of \( \alpha \)]
Outline

1. Classification
   - Evidential k-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential c-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
In the worst case, representing beliefs on a finite frame of discernment of size $K$ requires the storage of $2^K - 1$ numbers, and operations on belief functions have exponential complexity.

In classification and clustering, the frame of discernment (set of classes) is usually of moderate size (less than 100). Can we address more complex problems in machine learning, involving considerably larger frames of discernment?

Examples of such problems:
- Multi-label classification (Denœux, *Art. Intell.*, 2010);
Outline of the approach:

1. Consider a partial ordering \( \leq \) of the frame \( \Omega \) such that \((\Omega, \leq)\) is a lattice.
2. Define the set of propositions as the set \( \mathcal{I} \subset 2^\Omega \) of intervals of that lattice.
3. Define \( m \), \( bel \) and \( pl \) as functions from \( \mathcal{I} \) to \([0, 1] \) (this is possible because \((\mathcal{I}, \subseteq)\) has a lattice structure).

As the cardinality of \( \mathcal{I} \) is at most proportional to \(|\Omega|^2\), all the operations of Dempster-Shafer theory can be performed in polynomial time (instead of exponential when working in \((2^\Omega, \subseteq)\)).
Outline

1. Classification
   - Evidential $k$-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential $c$-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
Multi-label classification

- In some problems, learning instances may belong to several classes at the same time.
- For instance, in image retrieval, an image may belong to several semantic classes such as “beach”, “urban”, “mountain”, etc.
- If $\Theta = \{\theta_1, \ldots, \theta_c\}$ denotes the set of classes, the class label of an instance may be represented by a variable $y$ taking values in $\Omega = 2^\Theta$.
- Expressing partial knowledge of $y$ in the Dempster-Shafer framework may imply storing $2^{2^c}$ numbers.

<table>
<thead>
<tr>
<th>$c$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^{2^c}$</td>
<td>16</td>
<td>256</td>
<td>65536</td>
<td>4.3e9</td>
<td>1.8e19</td>
<td>3.4e38</td>
<td>1.2e77</td>
</tr>
</tbody>
</table>
The frame of discernment is $\Omega = 2^\Theta$, where $\Theta$ is the set of classes. The natural ordering in $2^\Theta$ is $\subseteq$, and $(2^\Theta, \subseteq)$ is a (Boolean) lattice.

The intervals of $(2^\Theta, \subseteq)$ are sets of subsets of $\Theta$ of the form:

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

for $A \subseteq B \subseteq \Theta$. 
Example (diagnosis)

- Let $\Theta = \{a, b, c, d\}$ be a set of faults.
- Item of evidence 1 → $a$ is surely present and $\{b, c\}$ may also be present, with confidence 0.7:

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

- Item of evidence 2 → $c$ is surely present and either faults $\{a, b\}$ (with confidence 0.8) or faults $\{a, d\}$ (with confidence 0.2) may also be present:

$$m_2([\{c\}, \{a, b, c\}]) = 0.8, \quad m_2([\{c\}, \{a, c, d\}]) = 0.2$$
Example
Combination by Dempster’s rule

<table>
<thead>
<tr>
<th>Evidence 1</th>
<th>Evidence 2</th>
<th>Evidence 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>{a}, {a, b, c}</td>
<td>{\emptyset_\Theta, \emptyset}</td>
<td>{\emptyset_\Theta, \emptyset}</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>{c}, {a, b, c}</td>
<td>{a, c}, {a, b, c}</td>
<td>{c}, {a, b, c}</td>
</tr>
<tr>
<td>0.8</td>
<td>0.56</td>
<td>0.24</td>
</tr>
<tr>
<td>{c}, {a, c, d}</td>
<td>{a, c}, {a, c}</td>
<td>{c}, {a, c, d}</td>
</tr>
<tr>
<td>0.2</td>
<td>0.14</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Based on this evidence, what is our belief that

- Fault $a$ is present: $bel(\{a\}, \emptyset) = 0.56 + 0.14 = 0.70$;
- Fault $d$ is not present: $bel(\emptyset, \{d\}) = bel(\emptyset, \emptyset) + bel(\{a, b, c\}) = 0.56 + 0.14 + 0.24 = 0.94$. 
Let us consider a learning set of the form:

\[ \mathcal{L} = \{(x_1, [A_1, B_1]), \ldots, (x_n, [A_n, B_n])\} \]

where

- \( x_i \in \mathbb{R}^p \) is a feature vector for instance \( i \)
- \( A_i \) is the set of classes that certainly apply to instance \( i \);
- \( B_i \) is the set of classes that possibly apply to that instance.

In a multi-expert context, \( A_i \) may be the set of classes assigned to instance \( i \) by all experts, and \( B_i \) the set of classes assigned by some experts.
Multi-label evidential \( k \)-NN rule

**Construction of mass functions**

- Let \( \mathcal{N}_k(\mathbf{x}) \) be the set of \( k \) nearest neighbors of a new instance \( \mathbf{x} \), according to some distance measure \( d \).
- Let \( \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}) \) with label \([A_i, B_i]\). This item of evidence can be described by the following mass function in \((\mathcal{I}, \subseteq)\):

\[
\begin{align*}
    m_i([A_i, B_i]) &= \varphi(d_i), \\
    m_i([\emptyset, \Theta]) &= 1 - \varphi(d_i),
\end{align*}
\]

where \( \varphi \) is a decreasing function from \([0, +\infty)\) to \([0, 1]\) such that \( \lim_{d \to +\infty} \varphi(d) = 0 \).

- The \( k \) mass functions are combined using Dempster’s rule:

\[
m = \bigoplus_{\mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})} m_i
\]
Multi-label evidential $k$-NN rule

Let $\hat{Y}$ be the predicted label set for instance $x$.

To decide whether to include in $\hat{Y}$ each class $\theta \in \Theta$ or not, we compute

- the degree of belief $\text{bel}(\{\theta\}, \Theta)$ that the true label set $Y$ contains $\theta$, and
- the degree of belief $\text{bel}(\emptyset, \{\theta\})$ that it does not contain $\theta$.

We then define $\hat{Y}$ as

$$\hat{Y} = \{ \theta \in \Theta \mid \text{bel}(\{\theta\}, \Theta) \geq \text{bel}(\emptyset, \{\theta\}) \}.$$ 

Other method: find the set of labels $\hat{Y}$ with the largest plausibility (linear programming problem).
Example: emotions data (Trohidis et al. 2008)

- Problem: Predict the emotions generated by a song.
- 593 songs were annotated by experts according to the emotions they generate.
- The emotions were: amazed-surprise, happy-pleased, relaxing-calm, quiet-still, sad-lonely and angry-fearful.
- Each song was described by 72 features and labeled with one or several emotions (classes).
- The dataset was split in a training set of 391 instances and a test set of 202 instances.
- Evaluation of results:

\[
Acc = \frac{1}{n} \sum_{i=1}^{n} \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}
\]
Results

Emotions

Accuracy

k

EML–kNN imprecise labels
EML–kNN noisy labels
ML–kNN noisy labels
Outline

1. Classification
   - Evidential $k$-NN rule
   - Evidential neural network classifier

2. Clustering
   - Credal partition
   - EVCLUS
   - Evidential $c$-means

3. Working in very large frames
   - Motivation and general approach
   - Multi-label classification
   - Ensemble clustering
Problem statement

- Clustering may be defined as the search for a partition of a set $E$ of $n$ objects.
- The natural frame of discernment for this problem is the set $\mathcal{P}(E)$ of partitions of $E$, with size $s_n$.
- Expressing such evidence in the Dempster-Shafer framework implies working with sets of partitions.

<table>
<thead>
<tr>
<th>$n$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_n$</td>
<td>5</td>
<td>15</td>
<td>52</td>
<td>203</td>
<td>876</td>
</tr>
<tr>
<td>$2^{s_n}$</td>
<td>23</td>
<td>32768</td>
<td>4.5e15</td>
<td>1.3e61</td>
<td>5.0e263</td>
</tr>
</tbody>
</table>
A partition $p$ is said to be finer than a partition $p'$ (or, equivalently $p'$ is coarser than $p$) if the clusters of $p$ can be obtained by splitting those of $p'$; we write $p \preceq p'$.

The poset $(\mathcal{P}(E), \preceq)$ is a lattice.
Lattices of partition intervals \((n = 3)\)

13 partition intervals < \(2^5 = 32\) sets of partitions.
Ensemble clustering aims at combining the outputs of several clustering algorithms ("clusterers") to form a single clustering structure (crisp or fuzzy partition, hierarchy).

This problem can be addressed using evidential reasoning by assuming that:

- There exists a "true" partition $p^*$;
- Each clusterer provides evidence about $p^*$;
- The evidence from multiple clusterers can be combined to draw plausible conclusions about $p^*$.

To implement this scheme, we need to manipulate Dempster-Shafer mass functions, the focal elements of which are sets of partitions.

This is feasible by restricting ourselves to intervals of the lattice $(\mathcal{P}(E), \preceq)$. 
Method
Mass construction and combination

- Compute $r$ partitions $p_1, \ldots, p_r$ with large numbers of clusters using, e.g., the FCM algorithm.
- For each partition $p_k$, compute a validity index $\alpha_k$.
- The evidence from clusterer $k$ can be represented as a mass function

$$
\begin{align*}
  m_k([p_k, p_E]) &= \alpha_k \\
  m_k([p_0, p_E]) &= 1 - \alpha_k,
\end{align*}
$$

where $p_E$ is the coarsest partition.
- The $r$ mass functions are combined using Dempster’s rule:

$$
m = m_1 \oplus \ldots \oplus m_r
$$
Method

Exploitation of the results

Let $p_{ij}$ denote the partition with $(n − 1)$ clusters, in which objects $i$ and $j$ are clustered together.

The interval $[p_{ij}, p_E]$ is the set of all partitions in which objects $i$ and $j$ are clustered together.

The degree of belief in the hypothesis that $i$ and $j$ belong to the same cluster is then:

$$Bel_{ij} = bel([p_{ij}, p_E]) = \sum_{[p_k, \bar{p}_k] \subseteq [p_{ij}, p_E]} m([p_k, \bar{p}_k])$$

Matrix $Bel = (Bel_{ij})$ can be considered as a new similarity matrix and can be processed by, e.g., a hierarchical clustering algorithm.
Results

Individual partitions

6 clusters ($\alpha=0.228$)

7 clusters ($\alpha=0.276$)

8 clusters ($\alpha=0.258$)

9 clusters ($\alpha=0.282$)

10 clusters ($\alpha=0.291$)

11 clusters ($\alpha=0.320$)
Results

Synthesis

Consensus
Gaussian data, 8 features, 5 clusters

(\(\alpha=0.0095\))
Distributed clustering
8D5K data (Strehl and Gosh, 2002)
Distributed clustering
8D5K data (Strehl and Gosh, 2002)

\[(\alpha = 0.0062)\]
Distributed clustering

Here, each clusterer provides a partition $p_k$ that tends to be coarser than the true partition $p_k$.

The output from clusterer $k$ can be represented as a mass function

$$\begin{align*}
m_k([p_0, p_k]) &= \alpha_k \\
m_k([p_0, p_E]) &= 1 - \alpha_k.
\end{align*}$$

As before, the mass functions are combined and synthesized in the form of a similarity matrix.
Distributed clustering
Consensus

Thierry Denœux
Classification and clustering using belief functions 75/78
The theory of belief function has great potential for solving challenging machine learning problems:

- Classification (supervised learning);
- Clustering (unsupervised learning).

Belief functions allow us to:

- Learn from weak information (partially supervised learning, imprecise and uncertain data);
- Express uncertainty on the outputs of a learning system (e.g., credal partition);
- Combine the outputs from several learning systems (ensemble classification and clustering).

Recent developments make it possible to address problems in very large frames (multilabel classification, clustering, preference learning, etc.).
T. Denœux.  
A k-nearest neighbor classification rule based on Dempster-Shafer theory.  

T. Denœux.  
A neural network classifier based on Dempster-Shafer theory.  

EVCLUS: Evidential Clustering of Proximity Data.  
ECM: An evidential version of the fuzzy c-means algorithm.  

T. Denœux, Z. Younes and F. Abdallah.  
Representing uncertainty on set-valued variables using belief functions.  

Ensemble clustering in the belief functions framework.  