Pattern Classification using Belief Functions

Thierry Denœux

`tdenoeux@utc.fr`

University of Compiègne, France

Outline

1. Pattern classification
   - Definitions, applications
   - classical approaches, limitations
2. Learning evidential classifiers from data
   - Model-based approach
   - Case-based approach
   - Belief decision trees
3. Combination of unreliable sensors/experts
4. Conclusions

Pattern classification

- Classification = assignment of objects to predefined categories (classes)
- Applications:
  - character, speech recognition
  - diagnosis, fault identification, condition monitoring
  - target identification
  - face recognition, person identification
  - text categorization, context-based image retrieval, web mining, etc.

Formalization

- Population $\mathcal{P}$ of objects, each object described by two variables:
  - $\mathbf{x}$: vector of d attributes (features), quantitative, qualitative, mixed
  - c: class variable, qualitative, values in finite set $\Omega = \{\omega_1, \ldots, \omega_K\}$.
- Classifier: mapping $f : \mathbb{R}^d \to \Omega$ allowing to predict the class of any new object described by feature vector $\mathbf{x}$
- Building a classifier from data = supervised learning.

utc
Université de Technologie
Compiègne

Supervised Learning

- Learning set:

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

- Usual assumptions:
  1. $\mathcal{L}$ is a realization of an iid sample drawn from $F(\mathbf{x}, c)$,
  2. Future examples will be drawn from the same distribution.
  3. There exists a loss function $L : \Omega^2 \to \mathbb{R}_+$, $L(\omega_k, \omega_\ell) =$ loss incurred if one assigns to class ω_k an object belonging to class ω_ℓ.

utc Université de Technologie Compiègne

The Bayes classifier

- The optimal (Bayes) classifier $f^* : \mathbb{R}^d \to \Omega$ is defined by

$$f^* : \mathbf{x} \mapsto \omega_k \text{ such that } R(\omega_k|\mathbf{x}) \leq R(\omega_\ell|\mathbf{x}) \quad \forall \ell \neq k$$

with

$$R(\omega_k|\mathbf{x}) = \sum_{\ell=1}^{c} L(\omega_k, \omega_\ell) P(\omega_\ell|\mathbf{x})$$

- f^* minimizes the overall risk:

$$R(f) = \int R(f(\mathbf{x})|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

utc
Université de Technologie
Compiègne

Approximating the Bayes classifier

- Usual approach for approximating f^*: estimate the posterior probabilities $P(\omega_k|\mathbf{x})$.
- Different strategies
  - parametric (ML) or non parametric (k-NN, Parzen) estimation of the class-conditional densities $p(\mathbf{x}|\omega_\ell)$, combination with priors $P(\omega_\ell)$ ($\ell = 1, \ldots, K$)
  - direct estimation of $P(\omega_k|\mathbf{x})$:
    - logistic regression,
    - neural networks,
    - decision trees, etc.

utc
Université de Technologie
Compiègne

Applicability

The above framework is relevant in applications where the learning set is:

1. representative of the data expected in the operating environment (proportions $\approx$ prior probabilities)
2. large enough to provide reliable estimates of the class-conditional densities
3. composed of precise and certain observations.

This is not always the case in real-world applications !

utc Université de Technologie Compiègne

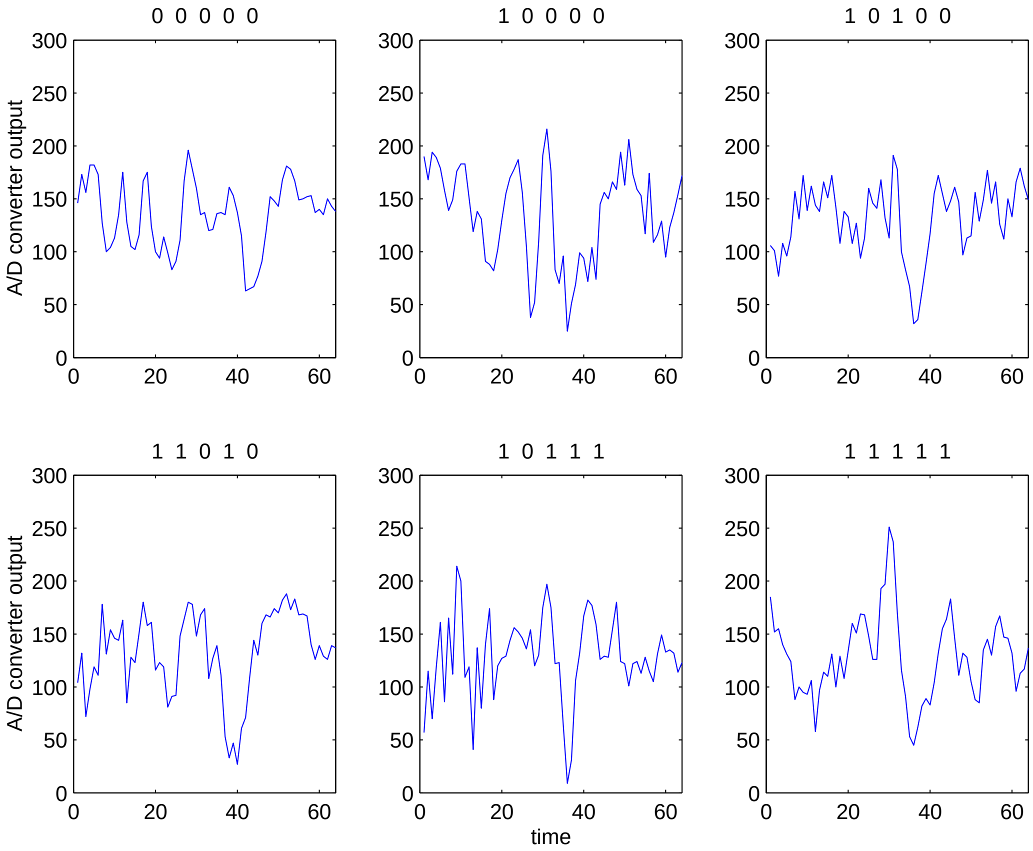
Analysis of sleep EEG

- Classification task: discriminate K-complexes from background activity in sleep EEG
- K-complexes = transient EEG patterns, play a major role in sleep stage assessment and diagnosis.
- Particular problems:
  - no “ground truth”: data has to be subjectively labeled by a panel of experts
  - the prior probability of a K-complex occurring in a given time window is unknown (depends on the patient)

utc
Université de Technologie
Compiègne

Data set

500 EEG signals encoded as 64-D patterns, 50 % negative (delta waves), 5 experts.



Sensor fusion

- Features are obtained from s sensors

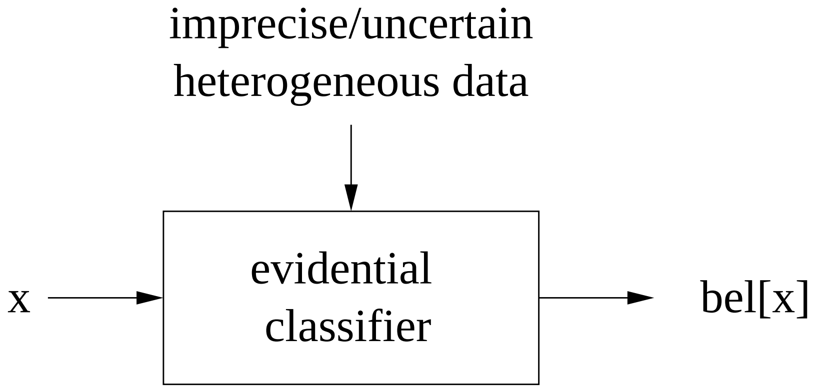
$$\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_s)$$

- Some sensors may be unreliable in certain operating conditions (not all represented in the training set)
- Incomplete information, different granularity levels:
  - sensor S_1: n_1 training patterns labeled as $\{\omega_1, \omega_2\}$ or ω_3
  - sensor S_2: n_2 training patterns labeled as ω_1 or ω_3, etc...

utc Université de Technologie Compiègne

The TBM framework

- A rich and flexible framework for representing various levels of uncertainties (from total ignorance to full knowledge),
- Requires fewer assumptions and less information than Probability theory
- Application to classification problems: evidential classifier.



Three approaches

1. Model-based (GBT):
   - Smets (1978)
   - Appriou (1991)
2. Cased-Based:
   - Denœux (1995)
3. Belief decision tree:
   - Elouedi and Smets (2000),
   - Denœux and Skarstein-Bjanger (2000)

utc
Université de Technologie
Compiègne

The model-based approach

- Based on the Generalized Bayesian Theorem ($\mathbf{x}$ discrete):

$$\boxed{\mathrm{pl}^{\Omega}[\mathbf{x}](A) = 1 - \prod_{\omega_k \in A} (1 - \mathrm{pl}^{X}[\omega_k](\mathbf{x})) \quad \forall A \subseteq \Omega}$$

- Problems:
  - How to determine $\mathrm{pl}^{X}[\omega_k]$?
  - Extension to continuous $\mathbf{x}$

Determination of pl$^X[\omega_k]$

- Let $\mathcal{L}_k$ be a learning set of n_k patterns of class ω_k.
- Assuming $\mathcal{L}_k$ to an iid sample from $p(\mathbf{x}|\omega_k)$, this conditional distribution can be estimated: $\widehat{p}(\mathbf{x}|\omega_k)$.
- pl$^X[\omega_k]$ can then be defined by discounting the estimated probability function $\widehat{p}(\mathbf{x}|\omega_k)$:

$$\text{pl}^X[\omega_k](\mathbf{x}) = 1 - \alpha_k + \alpha_k \widehat{p}(\mathbf{x}|\omega_k)$$

- We then have

$$\text{pl}^\Omega[\mathbf{x}](A) = 1 - \prod_{\omega_k \in A} \alpha_k (1 - \widehat{p}(\mathbf{x}|\omega_k)) \quad \forall A \subseteq \Omega$$

utc
Université de Technologie
Compiègne

The GBT in the continuous case

- Generalization to continuous $\mathbf{x}$:

$$\boxed{\mathrm{pl}^{\Omega}[\mathbf{x}](A) = 1 - \prod_{\omega_k \in A} \alpha_k (1 - \rho \cdot \widehat{p}(\mathbf{x}|\omega_k)) \quad \forall A \subseteq \Omega}$$

with

$$\rho = (\max_k \sup_{\mathbf{x}} \widehat{p}(\mathbf{x}|\omega_k))^{-1}.$$

- The reliability coefficients α_k can be fixed a priori or learnt from the data by minimizing an error function.

utc
Université de Technologie
Compiègne

1. Consistency with the Bayesian approach in the case where the class-conditional distributions $p(\mathbf{x}|\omega_k)$ and the prior probabilities $P(\omega_k)$ are known.
2. Separability of hypothesis evaluation: $m^\Omega[\mathbf{x}]$ can be decomposed as the conjunctive combination of K bba's $m_k^\Omega[\mathbf{x}]$ defined by

$$
\begin{aligned}
m_k^\Omega[\mathbf{x}](\overline{\{\omega_k\}}) &= \alpha_k(1-\rho\cdot\widehat{p}(\mathbf{x}|\omega_k)) \\
m_k^\Omega[\mathbf{x}](\Omega) &= 1-\alpha_k(1-\rho\cdot\widehat{p}(\mathbf{x}|\omega_k))
\end{aligned}
$$

3. Equivalence of aleatory and epistemic combination of observations: $m^\Omega[\mathbf{x},\mathbf{y}] = m^\Omega[\mathbf{x}] \bigcirc\!\!\!\!\cap\, m^\Omega[\mathbf{y}]$

Experiment 1 (1)

- Target classification problem with two classes $\Omega = \{\omega_1, \omega_2\}$ (e.g., aircraft and missile) and two sensors S_1 and S_2 (e.g. radar and infrared).
- Each sensor S_j allows to compute one feature x_j.
- Distributions of x_1 and x_2 in each class learnt in controlled experimental conditions:

$$p(x_1|\omega_1) = \mathcal{N}(0,1) \qquad p(x_1|\omega_2) = \mathcal{N}(6,1)$$

$$p(x_2|\omega_1) = \mathcal{N}(0,1) \qquad p(x_2|\omega_2) = \mathcal{N}(2,1)$$

- Equiprobability assumption: $P(\omega_1) = P(\omega_2)$.

utc
Université de Technologie
Compiègne

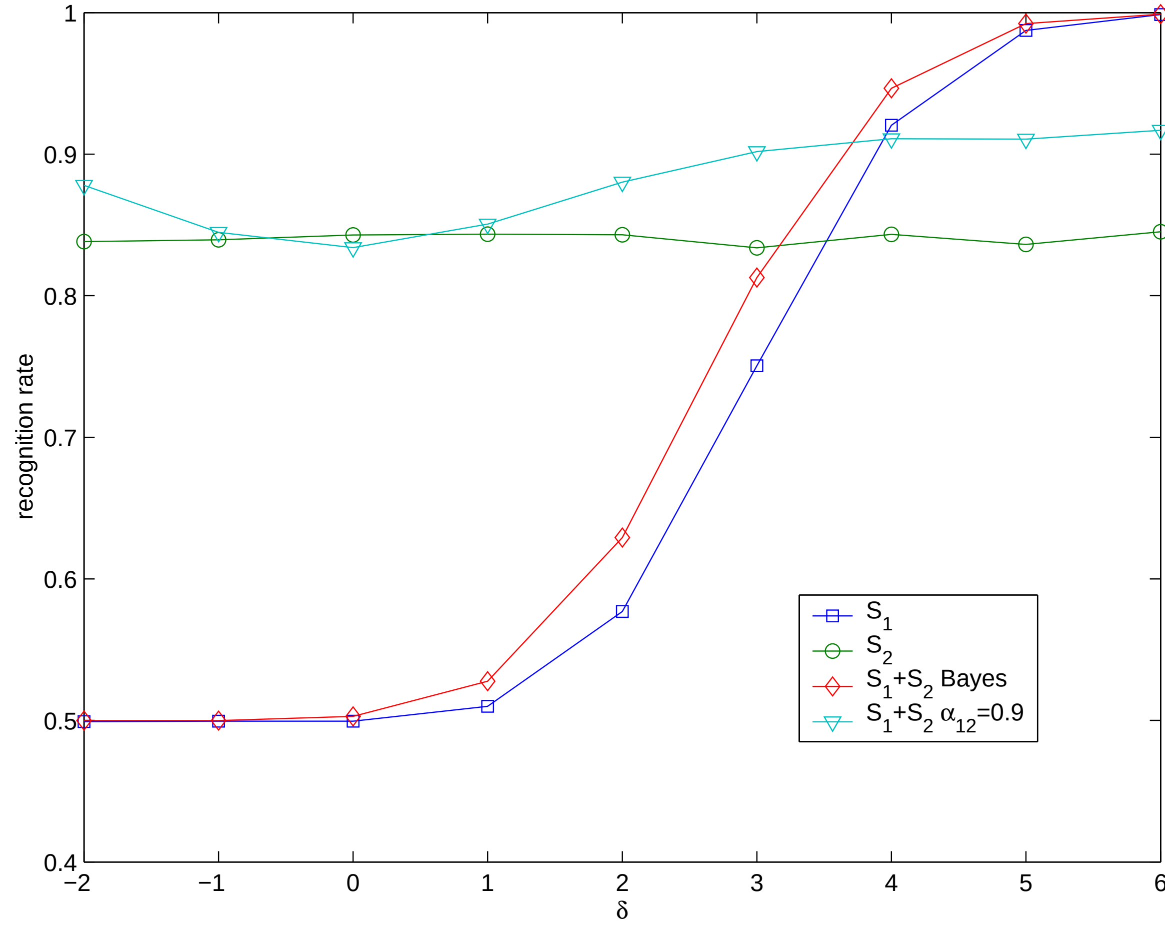
Experiment 1 (2)

- If the distributions of x_1 and x_2 were the same in the operational context, the best performances would be achieved by the Bayes classifier, BUT
- it is known that the distribution of x_1 for class ω_2 objects is altered due to environmental conditions.
- This is can be taken into account by discounting $p(x_1|\omega_2)$ with rate $1 - \alpha_{1,2} > 0$.
- $pl^\Omega[x_1] \bigcirc\!\!\!\!\cap\, pl^\Omega[x_2]$ are then computed using the GBT and combined:

$$pl^\Omega[x_1, x_2] = pl^\Omega[x_1] \bigcirc\!\!\!\!\cap\, pl^\Omega[x_2]$$

Experiment 1: result

$$p(x_1|\omega_2) = \mathcal{N}(\delta, 1)$$



Experiment 2

- Two sensors S_1 and S_2, three classes $\Omega = \{\omega_1, \omega_2, \omega_3\}$.
- We know
  - Sensor S_1: $p(x_1|\omega_1)$, $p(x_1|\omega_2)$
  - Sensor S_2: $p(x_2|\omega_1) = p(x_2|\omega_2)$, $p(x_2|\omega_3)$
- We do not know:
  - distribution of x_1 in class 3: $p(x_1|\omega_3)$
  - prior probabilities $P(\omega_1), P(\omega_2), P(\omega_3)$
- Two solutions:
  - TBM solution
  - Bayesian solution

utc
Université de Technologie
Compiègne

The TBM solution

- Frame for sensor S_1: $\Omega_{12} = \{\omega_1, \omega_2\}$
- Frame for sensor S_2: $\Omega_{\{12\}3} = \{\omega_{12}, \omega_3\}$ with $\omega_{12} = \{\omega_1, \omega_2\}$.

$$
\begin{tikzcd}
p(x_1|\omega_1), p(x_1|\omega_2) \arrow[r, "\text{GBT}"] & m^{\Omega_1{}^2}[x_1] \arrow[r, "\text{vac. ext.}"] & m^{\Omega}[x_1] \arrow[d] \\
& \boxed{m^{\Omega}[x_1, x_2]} & \cap \arrow[l] \\
p(x_2|\omega_{12}), p(x_2|\omega_3) \arrow[r, "\text{GBT}"] & m^{\Omega_{\{12\}3}}[x_2] \arrow[r, "\text{refinement}"] & m^{\Omega}[x_2] \arrow[u]
\end{tikzcd}
$$

The Bayesian solution

- A prior distribution on Ω and a conditional probability density $p(x_1|\omega_3)$ must be defined.
- Natural choice: "non-informative" priors

$$P(\omega_1) = P(\omega_2) = P(\omega_3) = 1/3$$

$$p(x_1|\omega_3) = \mathcal{U}_{[-1,5]}$$

- Computation of posterior probabilities:

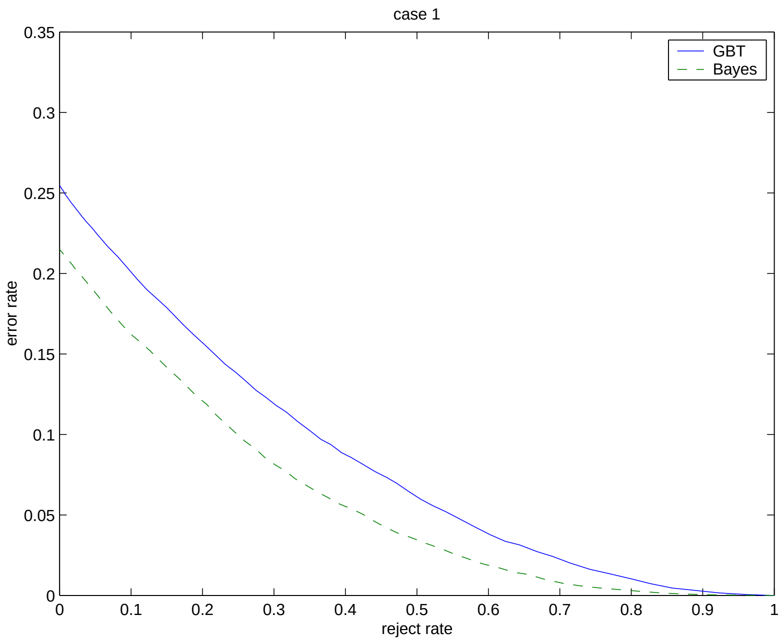
$$P(\omega_k|x_1, x_2) = \frac{p(x_1|\omega_k)p(x_2|\omega_k)P(\omega_k)}{p(x_1, x_2)}$$

Example 2 - Results

Case 1:

$$P = (1/3, 1/3, 1/3)$$

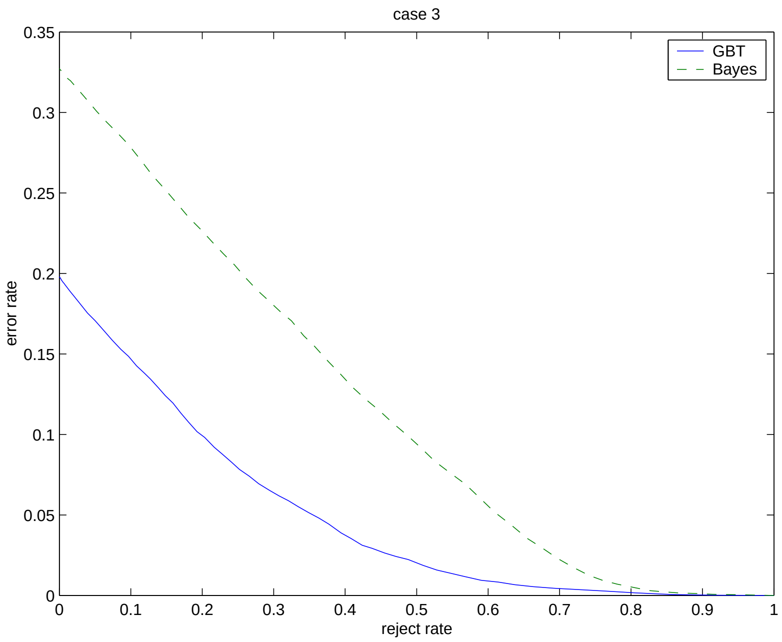
$$p(x_1|\omega_3) = \mathcal{U}_{[-1,5]}$$



Case 2

$$P = (0.1, 0.2, 0.7)$$

$$p(x_1|\omega_3) = \mathcal{N}(2, 1)$$



The case-based approach

- Does not use any probabilistic model of the distribution of attributes in each class;
- Treats each example $(\mathbf{x}_i, c_i)$ in the learning set as a piece of evidence, whose relevance depends on the dissimilarity beween the current vector $\mathbf{x}$ and $\mathbf{x}_i$;
- The n items of evidence are combined using the Dempster's rule of combination.
- Allows to use training data with imprecise and/or uncertain class labels (semi-supervised learning).

The learning set

- A more general form of the learning set:
  $\mathcal{L} = \{e_i = (\mathbf{x}_i, m_i), i = 1, \ldots, n\}$
- m_i: a bba representing Your partial knowledge regarding the class of object i.
- Special cases:
  - $m_i(\{\omega_k\}) = 1$: precise (standard) labelling
  - $m_i(A) = 1$ for $A \subseteq \Omega$: imprecise labelling
  - m_i is a probability function: probabilistic labeling (opinions of N experts)
  - m_i has nested focal elements: possibilistic labeling (“object i is big”), etc...

utc
Université de Technologie
Compiègne

Impact of 1 example

- The relevance of e_i as an item of evidence regarding the class of $\mathbf{x}$ is related to the dissimilarity between the 2 vectors:
  - If $\mathbf{x} = \mathbf{x}_i$, e_i is totally relevant, $m[\mathbf{x}, e_i] \approx m_i$.
  - If $\mathbf{x}$ and $\mathbf{x}_i$ are very dissimilar, e_i is irrelevant and $m[\mathbf{x}, e_i](\Omega) = 1$.
  - If $\mathbf{x}$ and $\mathbf{x}_i$ are somewhat disimilar, e_i is partially relevant. m_i must be discounted:

$$\boxed{m[\mathbf{x}, e_i] = m_i^{\alpha(\mathbf{x},\mathbf{x}_i)}}$$

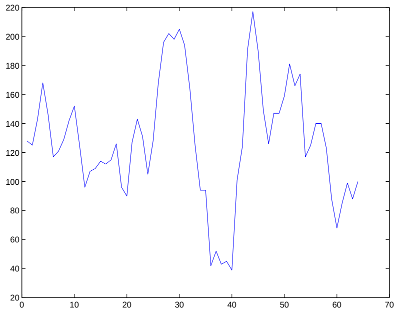
where $\alpha(\mathbf{x}, \mathbf{x}_i) \in [0, 1]$ is a dissimilarity measure.

utc Université de Technologie Compiègne

Example

$$\Omega = \{K\text{-complex}, \delta\text{-wave}\}$$

$\mathbf{x}_i =$

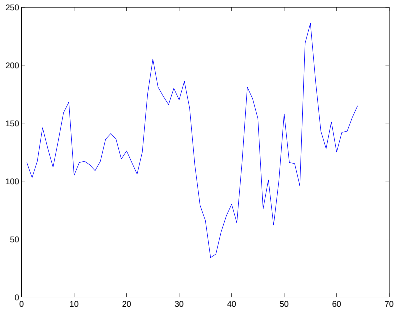


A	$\emptyset$	$\{K\}$	$\{\delta\}$	Ω
$m_i(A)$	0	0.8	0.2	0

$\updownarrow \alpha(\mathbf{x}, \mathbf{x}_i) = 0.5$

$\downarrow$ discounting

$\mathbf{x} =$



A	$\emptyset$	$\{K\}$	$\{\delta\}$	Ω
$m[\mathbf{x}, e_i](A)$	0	0.4	0.1	0.5

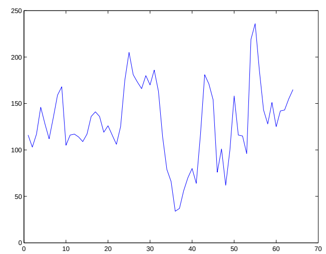
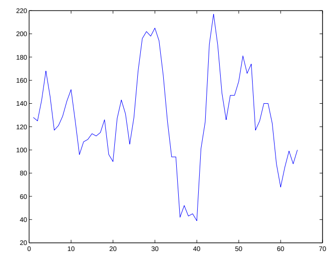
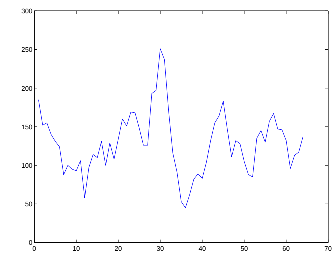
Impact of n examples

- Each learning example induces a bba $m[\mathbf{x}, e_i]$.
- Assuming the n learning examples to be n distinct items of evidence, the evidence of the n examples is pooled using Dempster's rule of combination:

$$\boxed{m[\mathbf{x}, \mathcal{L}] = m[\mathbf{x}, e_1] \bigcap\!\!\!\!\bigcirc \ldots \bigcap\!\!\!\!\bigcirc m[\mathbf{x}, e_n]}$$

utc Université de Technologie Compiègne

Example

$\mathbf{x}=$  $\mathbf{x}_i=$  $\mathbf{x}_j=$ 

$\alpha(\mathbf{x}, \mathbf{x}_i) = 0.5$ $\alpha(\mathbf{x}, \mathbf{x}_i) = 0.3$

A	$\emptyset$	$\{K\}$	$\{\delta\}$	Ω
$m_i(A)$	0	0.8	0.2	0
$m_j(A)$	0	0.6	0.4	0
$m[\mathbf{x}, e_i](A)$	0	0.4	0.1	0.5
$m[\mathbf{x}, e_j](A)$	0	0.42	0.28	0.3
$m[\mathbf{x}, e_i, e_j](A)$	0.196	0.282	0.372	0.15

Implementation

- Dissimilarity measure defined as a function of a distance measure (e.g. Euclidean if $\mathbf{x} \in \mathbb{R}^d$). For instance:

$$\alpha(\mathbf{x}, \mathbf{x}_i) = 1 - \alpha_0 \exp(-\gamma ||\mathbf{x} - \mathbf{x}_i||^2)$$

- parameters α_0 and γ can be learnt by minimization of an error function
- For faster computation:
  - use only the k nearest neighbors of $\mathbf{x}$ (evidential k-NN rule)
  - summarize $\mathcal{L}$ using p prototypes (learnt in unsupervised or supervised mode)

utc Université de Technologie Compiègne

Results on 'classical data'

Vowel data

$K = 11,$

$d = 10$

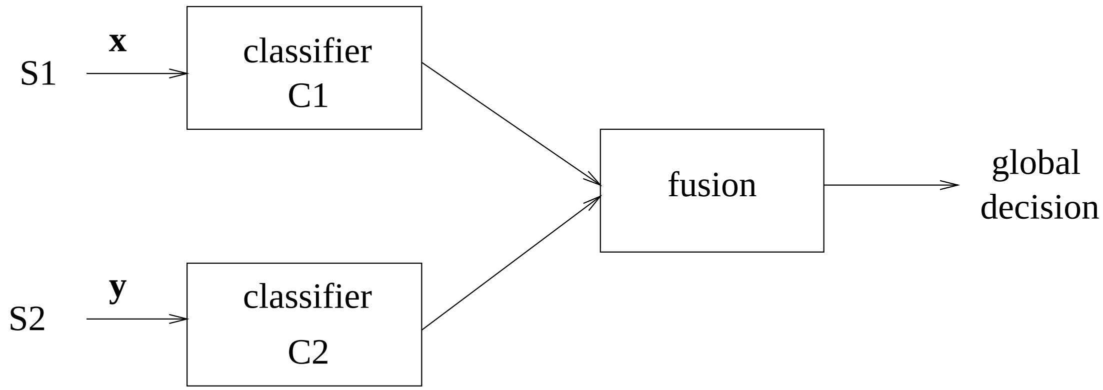
$n = 568$

test : 462 ex. (different speakers)

Classifier	test error rate
Multi-layer perceptron (88 hidden units)	0.49
Radial Basis Function (528 hidden units)	0.47
Gaussian node network (528 hidden units)	0.45
Nearest neighbor	0.44
Linear Discriminant Analysis	0.56
Quadratic Discriminant Analysis	0.53
CART	0.56
BRUTO	0.44
MARS (degree=2)	0.42
Case-based classifier (33 prototypes)	0.38
Case-based classifier (44 prototypes)	0.37
Case-based classifier (55 prototypes)	0.37

utc
Université de Technologie
Compiègne

Data fusion example



- $K = 2$ classes
- $\mathbf{x} \in \mathbb{R}^5, \mathbf{y} \in \mathbb{R}^3$, Gaussian distribution, conditionally independent
- Learning set: $n = 60$, cross-validation: $n_{cv} = 100$
- test: 5000 vectors

Data fusion: results (1)

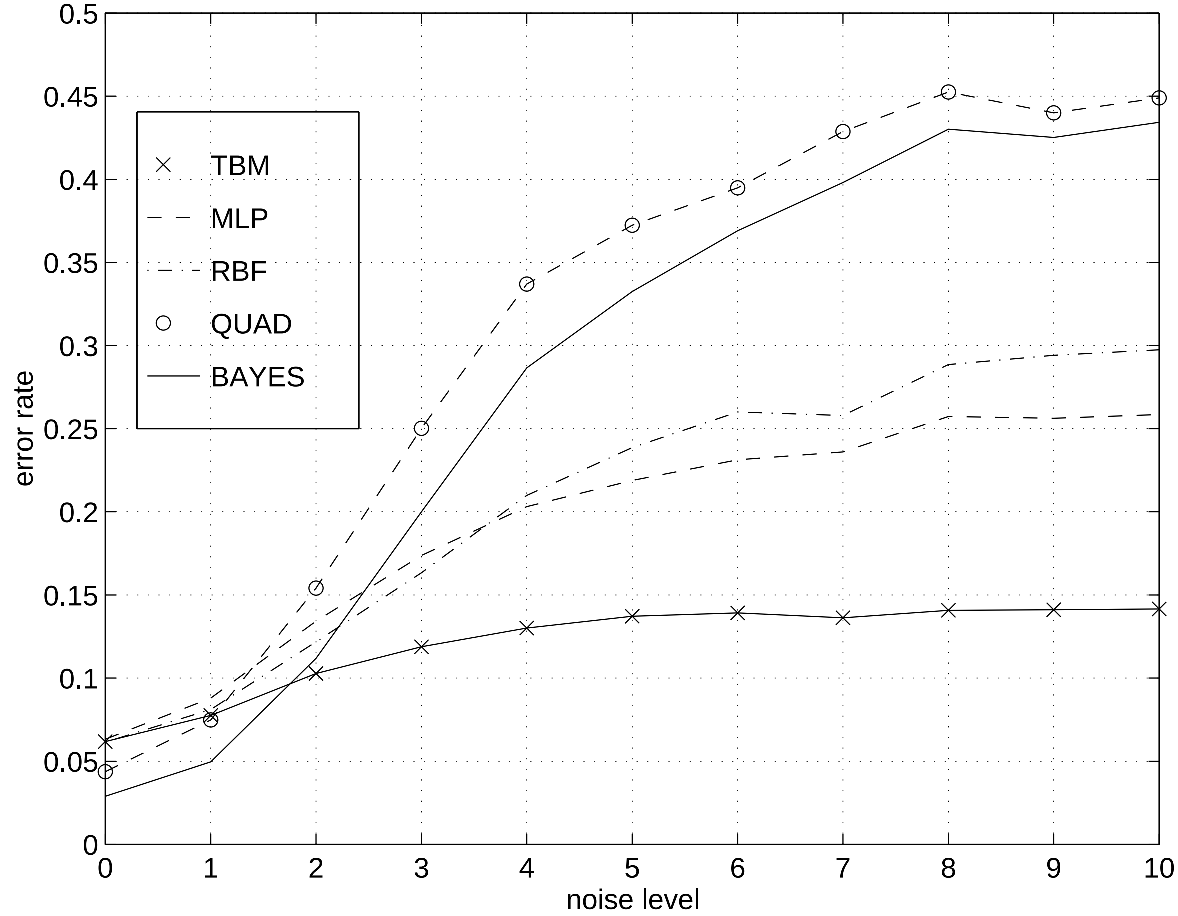
Test error rates: uncorrupted data

Method	$\mathbf{x}$ alone	$\mathbf{y}$ alone	$\mathbf{x}$ and $\mathbf{y}$
TBM	0.106	0.148	0.061
MLP	0.113	0.142	0.063
RBF	0.133	0.159	0.083
QUAD	0.101	0.141	0.049
BAYES	0.071	0.121	0.028

utc Université de Technologie Compiègne

Data fusion: results (2)

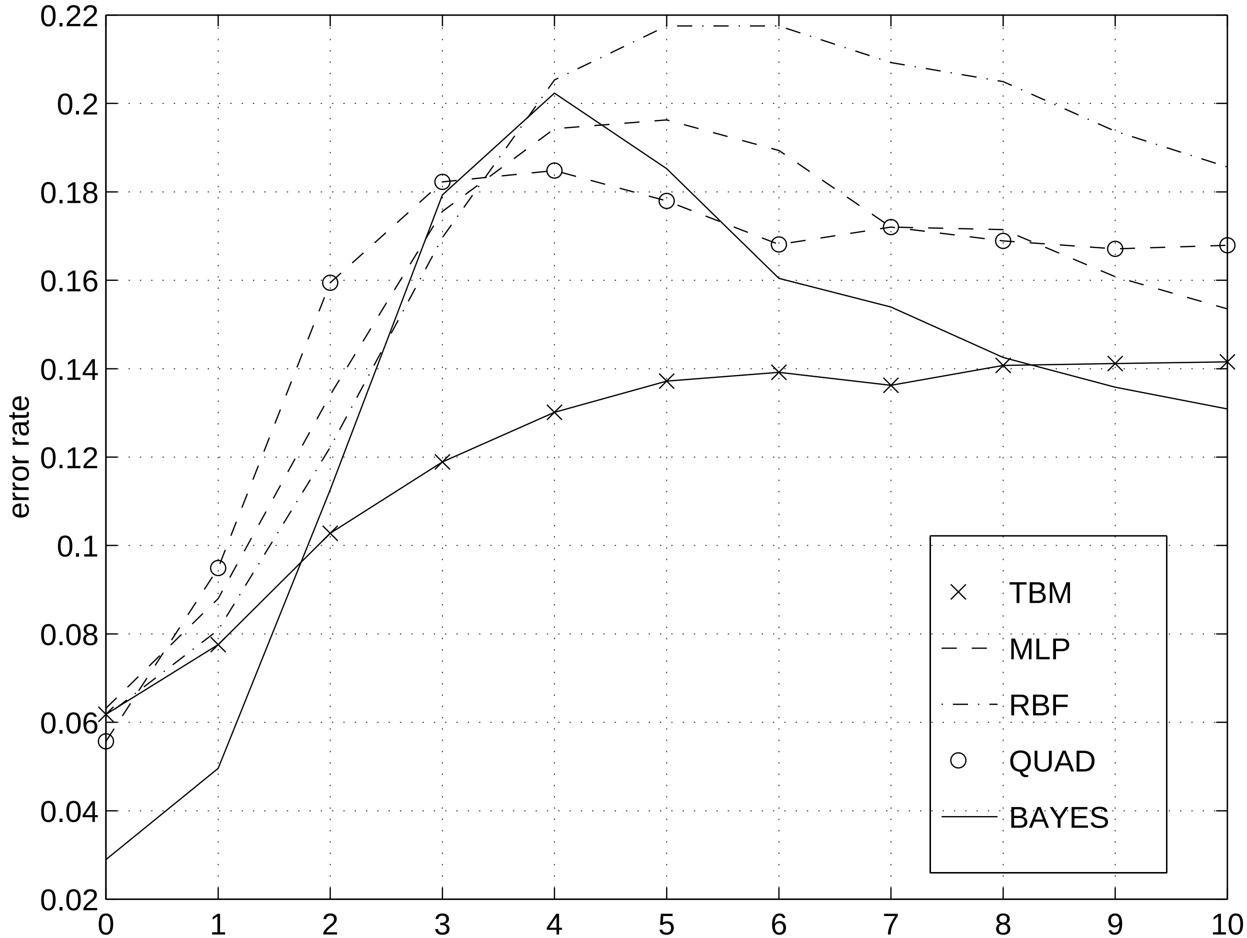
Test error rates: $\mathbf{x} + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$



utc Université de Technologie Compiègne

Data fusion: results (3)

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



Results on EEG data

- $K = 2$ classes, $d = 64$
- data labeled by 5 experts
- $n = 200$ learning patterns, 300 test patterns

k	k-NN	w K-NN	TBM (crisp labels)	TBM (uncert. labels)
9	0.30	0.30	0.31	0.27
11	0.29	0.30	0.29	0.26
13	0.31	0.30	0.31	0.26

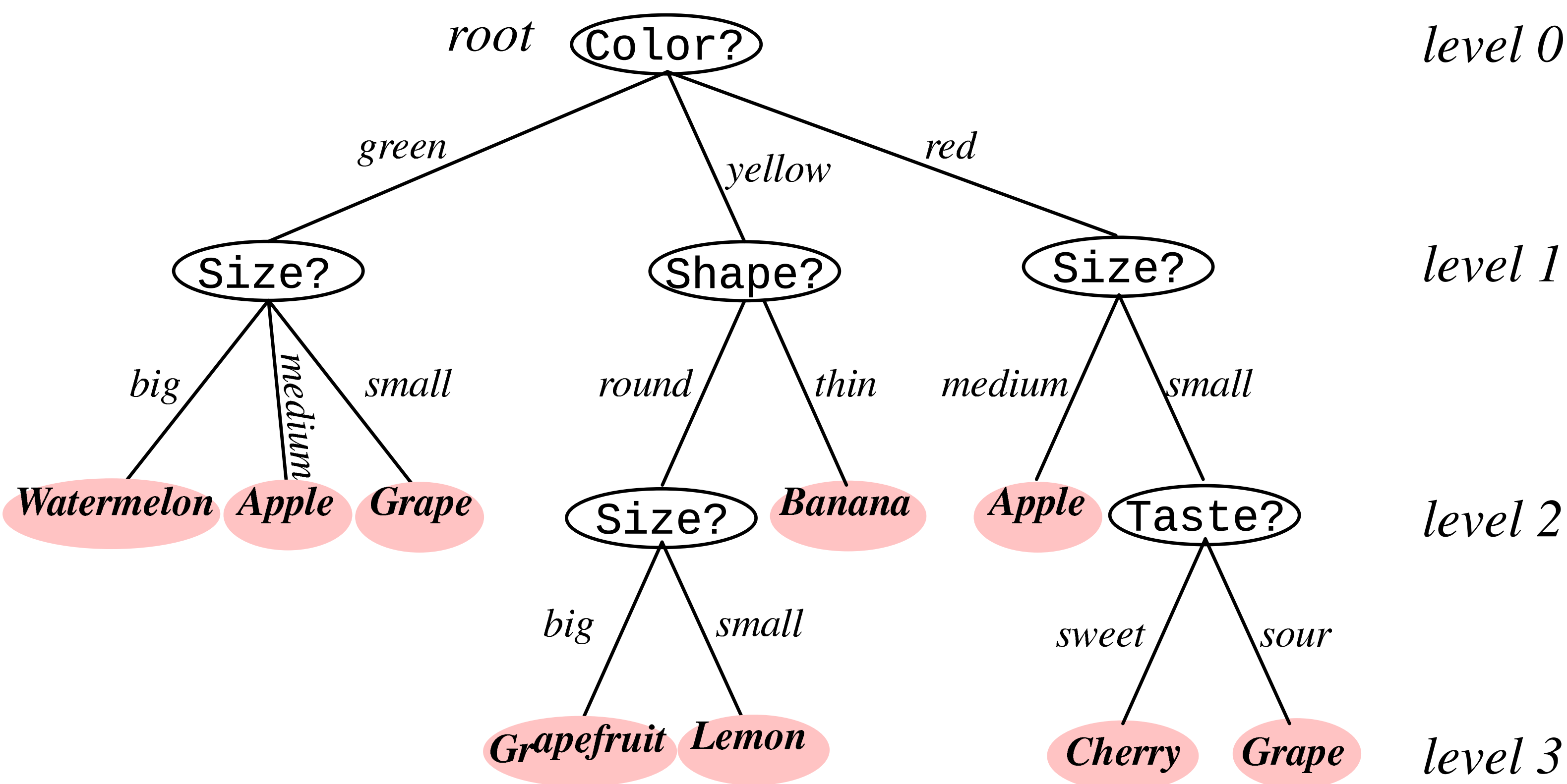
utc
Université de Technologie
Compiègne

Belief decision trees

- Recently introduced by Elouedi and Smets (2000), Denœux and Skarstein-Bjanger (2000);
- Goals:
  - extend the DT induction methodology to learning data with imprecise or uncertain class labels
  - allow for imprecise or uncertain attribute values in the testing phase
- Several algorithms, e.g. averaging approach.

A decision tree

Decision tree = representation of a sequential decision procedure



DT induction

- Basic principle: recursively partition the training set using one attribute at a time.
- At each step, try to split a node (=subset of patterns) in such a way that the child nodes are, on average, 'purer' in one class than their parents.
- Classical impurity criterion: $I(\mathcal{L}) = -\sum_{k=1}^{c} \widehat{p}_k \log_2 \widehat{p}_k$ where $\widehat{p}_k = n_k/n$ is the proportion of class ω_k in $\mathcal{L}$.
- Information gain of a categorical attribute a

$$\Delta I(a, \mathcal{L}) = I(\mathcal{L}) - \sum_{v=1}^{n_a} \frac{|\mathcal{L}_v|}{|\mathcal{L}|} I(\mathcal{L}_v)$$

utc
Université de Technologie
Compiègne

Extension to uncertain labels (1)

Interpretation of $I(\mathcal{L})$ in the classical case:

- C=class of the case selected at random from $\mathcal{L}$ with equiprobability.

$$
\begin{aligned}
P(C=\omega_k) &= \sum_{i=1}^{n} P(\text{selected case is } i) P(c_i=\omega_k) \\
&= \frac{1}{n}\sum_{i=1}^{n} P(c_i=\omega_k) = \frac{n_k}{n}
\end{aligned}
$$

- $I(\mathcal{L})$ is the entropy of the distribution of r.v. C.

Extension to uncertain labels (2)

- Let $\mathcal{L} = \{(\mathbf{x}_i, m_i), i = 1, \ldots, n\}$, where m_i is the bba about the class of case i.
- Select a case at as random from $\mathcal{L}$. For all $A \subseteq \Omega$,

$$
\begin{aligned}
m(C \in A) &= \sum_{i=1}^{n} P(\text{selected case is } i) m(c_i \in A) \\
&= \frac{1}{n} \sum_{i=1}^{n} m_i(A) = \overline{m}(A)
\end{aligned}
$$

- Hence, $\overline{m}$ generalizes the empirical class distribution $n_k/n, k = 1, \ldots, K$.

Extension to uncertain labels (3)

- The impurity of $\mathcal{L}$ can be defined as the entropy of the corresponding pignistic probability distribution:

$$I(\mathcal{L}) = -\sum_{k=1}^{c} \overline{\mathsf{BetP}}(\omega_k) \log_2 \overline{\mathsf{BetP}}(\omega_k)$$

  with $\overline{\mathsf{BetP}} = \frac{1}{n}\sum_{i=1}^{n} \mathsf{BetP}_i$

- Prepruning: discount $\overline{m}$ with reliability factor

$$1 - \alpha = \frac{|\mathcal{L}|}{|\mathcal{L}| + \eta}$$

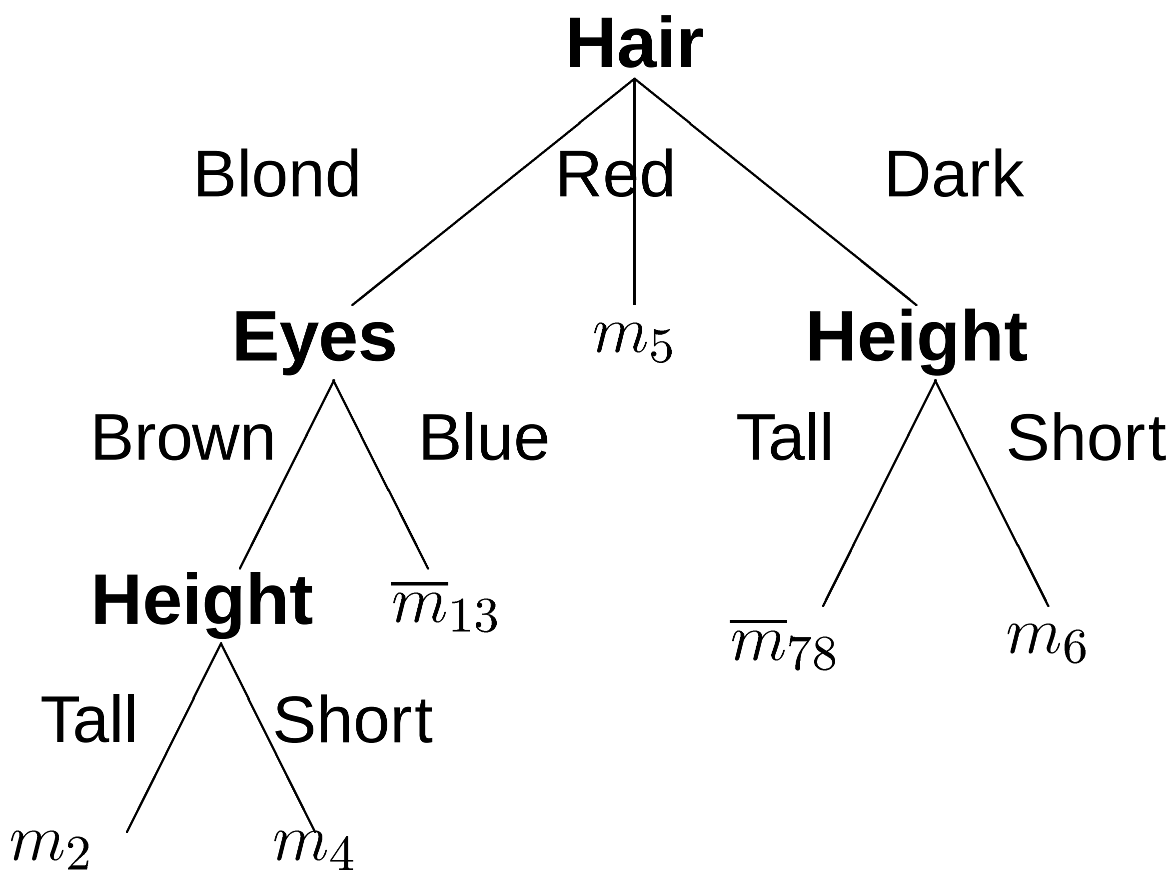
utc
Université de Technologie
Compiègne

Example: data

Hair	Eyes	Height	m_i
Blond	Blue	Tall	$\omega_1, .8; \Omega, .2$
Blond	Brown	Tall	$\omega_2, .4; \omega_1 \cup \omega_2, .4; \Omega, .2$
Blond	Blue	Tall	$\omega_1, .9; \Omega, .1$
Blond	Brown	Short	$\omega_2, .6; \omega_3, .2; \Omega, .2$
Red	Blue	Tall	$\omega_2, .8; \Omega, .2$
Dark	Brown	Short	$\omega_3, .6; \Omega, .4$
Dark	Brown	Tall	$\omega_3, .9; \Omega, .1$
Dark	Brown	Tall	$\omega_3, .5; \omega_1 \cup \omega_3, .2; \Omega, .3$

utc
Université de Technologie
Compiègne

Example: tree

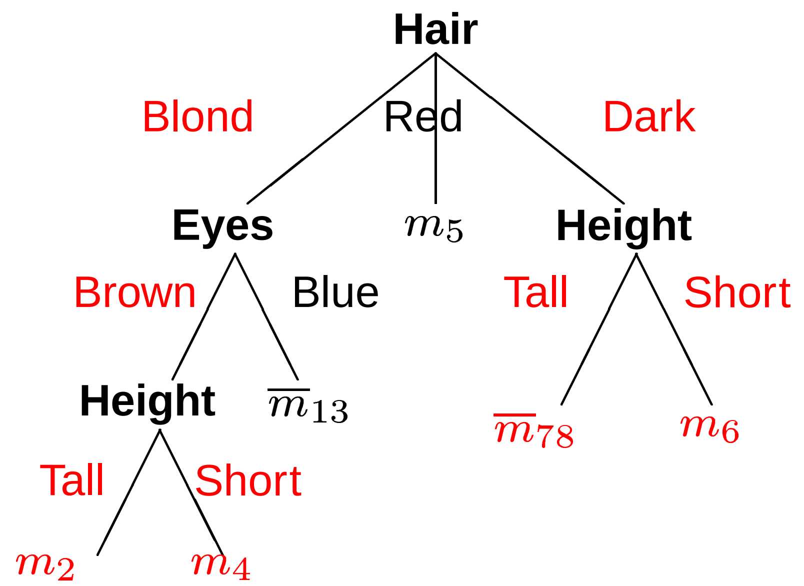


Imprecise/uncertain attribute values

- Disjunctive case: the values of some attributes are only known to belong to subset of values.

Ex: hair $\neq$ red,
eyes=brown
height $\in \{\text{tall}, \text{short}\}$

$$\boxed{m = m_2 \circledcup m_4 \circledcup \overline{m}_{78} \circledcup m_6}$$



- General case: knowledge about each attribute a_j described by a bba m^{a_j}.

Sensor Tuning

- Pb: s sensors (experts, classifiers) express beliefs regarding the class c of an object.
- The s sensors are assumed to be distinct sources, but they may have different degrees of reliability.
- Let $\alpha_j = P(S_j$ is not reliable$)$. Then a discounting rate α_j should be applied to the bba m_{S_j} before combining the s sensor reports:

$$\boxed{m = m_{S_1}^{\alpha_1} \bigcirc\!\!\!\!\cap \ldots \bigcirc\!\!\!\!\cap m_{S_s}^{\alpha_s}}$$

- How to learn the discounting rates α_j from a set of data with known classification ?

utc
Université de Technologie
Compiègne

Learning the α_j

- Let $\{o_1, \ldots, o_n\}$ denote a set of n objects, with known class c_i, $i = 1, \ldots, n$.
- The discounted bba provided by sensor S_j regarding the class of object o_i is $m_{S_j}^{\alpha_j}\{o_i\}$.
- The result of the combination for object o_i is

$$\boxed{m\{o_i\} = m_{S_1}^{\alpha_1}\{o_i\} \bigcirc\!\!\!\!\cap \ldots \bigcirc\!\!\!\!\cap m_{S_s}^{\alpha_s}\{o_i\}}$$

utc Université de Technologie Compiègne

Learning the α_j

- The error for object o_i may be measured as:

$$\boxed{\text{err}_i(\alpha_1, \ldots, \alpha_s) = \sum_{k=1}^{c} (\text{BetP}\{o_i\}(\omega_k) - t_{ik})^2}$$

  where $t_{ik} = 1$ if $c_i = \omega_k$, 0 otherwise.

- The optimal discounting rates may be determined by minimizing the overall error

$$(\alpha_1^*, \ldots, \alpha_s^*) = \arg \min_{\alpha_1, \ldots, \alpha_s} \sum_{i=1}^{n} \text{err}_i(\alpha_1, \ldots, \alpha_s)$$

Example

$$\Omega = \{\text{Airplane}, \text{Helicopter}, \text{Rocket}\}$$

	A	H	R	$\{A,H\}$	$\{A,R\}$	$\{H,R\}$	Ω	c_i
$m_{S_1}\{o_1\}$	0	0	0.5	0	0	0.3	0.2	A
$m_{S_1}\{o_2\}$	0	0.5	0.2	0	0	0	0.3	H
$m_{S_1}\{o_3\}$	0	0.4	0	0	0.6	0	0	A
$m_{S_1}\{o_4\}$	0	0	0	0	0.6	0.4	0	R
$m_{S_2}\{o_1\}$	0	0	0.5	0	0	0.3	0.2	A
$m_{S_2}\{o_2\}$	0	0.5	0.2	0	0	0	0.3	H
$m_{S_2}\{o_3\}$	0	0.4	0	0	0.6	0	0	A
$m_{S_2}\{o_4\}$	0	0	0	0	0.6	0.4	0	R

$$\alpha_1^* = 0.28 \qquad \alpha_2^* = 0.12$$

utc Université de Technologie Compiègne

Conclusions

- The main approaches to pattern classification (parametric, distance-based, tree-structured classifiers) can be transposed in the TBM framework, resulting in
  - greater flexibility to handle various sources of uncertainty (e.g. imprecise or bad quality data)
  - reduced need for unjustified assumption in situations of weak available information,
  - more robust decision procedures (unreliable sensor data)
- BF- based techniques also available for related problems such as regression and clustering.

utc
Université de Technologie
Compiègne