Knowledge discovery from data sets under tenable assumptions

Marco Zaffalon

IDSIA

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale
SWITZERLAND
http://www.idsia.ch/~zaffalon
zaffalon@idsia.ch

School on Imprecise Probabilities, Lugano (CH), 27–31 July 2004

Introduction to Knowledge discovery from data sets (KDD)

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

KDD

- Learning models from data (alone)
- Using models to make "predictions" about new data
- Discipline related to AI and Statistics
- Emphasis on computer-intensive methods

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Prototypical framework

- Data (***d***) come in a table

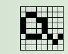
C	A_1	$\cdots$	A_j	$\cdots$	A_m
c_1	a_{11}	$\cdots$	a_{1j}	$\cdots$	a_{1m}
			$\vdots$		
c_i	a_{i1}	$\cdots$	a_{ij}	$\cdots$	a_{im}
			$\vdots$		
c_N	a_{N1}	$\cdots$	a_{Nj}	$\cdots$	a_{Nm}

- $c_i \in \mathcal{C} \quad (i = 1, \ldots, N)$
  - $\mathcal{C}$ is a finite set of *classes*
- $a_{ij} \in \mathcal{A}_j \; (i = 1, \ldots, N; j = 1, \ldots, m)$
  - $\mathcal{A}_j{}'$s are finite sets of *attribute values*
- A row of the table is called *observation*
- The data in the table are also called *sample*
- N, the number of units, is also called *sample size*

(Pattern) Classification

- Use data to learn a function $f : \mathcal{A}_1 \times \cdots \times \mathcal{A}_m \to \mathcal{C}$
  - *f* is "learned by examples"
  - *f* is called *classifier*
  - Learning = inferential approach
- Use the classifier to predict the *unknown* class of a *new* observation
  - $(a_1, \ldots, a_m) \xrightarrow{f} c$
- Why being interested in classification?

Some applications

- Hand-written character recognition  ⇨ a
  - Observation = Image
  - Classes = {a,b,…,z}
- Face recognition  ⇨ John
- Medical diagnosis List of Symptoms ⇨ Disease
- … Fraud detection, network intrusion, gene expression profiling, … … many others: very general paradigm

Common assumption

- The process underlying the data is *multinomial*
  - Observations are generated in independent and identically distributed way
  - C is called *class variable*, A_j $(j = 1,\ldots,m)$ *attribute variable*
- Notation
  - $\mathcal{X} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$, with generic element $x = (a_1,\ldots,a_m)$
  - $\mathcal{D} = \mathcal{C} \times \mathcal{X}$, with generic element $d = (c,x)$
- The *physical probability* (or *chance*) of d is θ_d $\left(\sum_{d \in \mathcal{D}} \theta_d = 1\right)$
  - Let θ be the vector of chances
  - θ is usually unknown
- The probability $p(\boldsymbol{d}|\theta)$ of the observed data is then $\prod_{d \in \mathcal{D}} \theta_d^{n_d}$
  - n_d is the number of times the instance d appears in the data set $\boldsymbol{d}$
  - As a function of θ, $p(\boldsymbol{d}|\theta)$ is called the *likelihood function*

Bayesian approach to classification

- Model your state of knowledge about θ by a *prior* density $p(\theta)$
- Compute the *posterior* density $p(\theta|\boldsymbol{d})$ by Bayes' rule

$$p(\theta|\boldsymbol{d}) = \frac{p(\theta)p(\boldsymbol{d}|\theta)}{\int p(\theta)p(\boldsymbol{d}|\theta)d\theta}$$

- Focus on $p(c,x|\boldsymbol{d}) = E[\theta_{c,x}|\boldsymbol{d}]$
  - E denotes expectation w.r.t $p(\theta|\boldsymbol{d})$
- Select $\quad c^* = \operatorname*{argmax}_{c \in \mathcal{C}} p(c|x, \boldsymbol{d}) = \operatorname*{argmax}_{c \in \mathcal{C}} p(c, x|\boldsymbol{d})$

The multinomial case

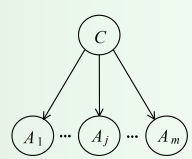
- Traditional choice for $p(\theta)$ is Dirichlet(s,$\boldsymbol{t}$): $p(\theta) \propto \prod_{d \in \mathcal{D}} \theta_d^{st_d - 1}$
  - s and $\boldsymbol{t} = (t_d)_{d \in \mathcal{D}}$ are hyperparameters
  - t_d = expectation of θ_d w.r.t. the prior
  - *Informal* view (prior is conjugate)
    - $s > 0$ as the number of additional (fractional) *virtual* observations
    - t_d as the proportion of observations equal to d in the virtual sample
    - $0 < t_d < 1, \sum_{d \in \mathcal{D}} t_d = 1$
  - e.g., uniform prior: $s = |\mathcal{D}|, t_d = 1/|\mathcal{D}|$; Perks prior: $s = 1, t_d = 1/|\mathcal{D}|$
- Posterior is Dirichlet(N+s,$\boldsymbol{t}^*$): $p(\theta|\boldsymbol{d}) \propto \prod_{d \in \mathcal{D}} \theta_d^{n_d + st_d - 1}$
  - $\boldsymbol{t}^* = (t_d^*)_{d \in \mathcal{D}}, t_d^* = (n_d + st_d)/(N+s)$
  - t_d^* = expectation of θ_d w.r.t. the posterior
- The needed probability is then

$$p(c,x|\boldsymbol{d}) = E[\theta_{c,x}|\boldsymbol{d}] = \frac{n_{c,x} + st_{c,x}}{N+s}$$

Some popular classifiers

The naive Bayes classifier (NBC)

- Unstructured models do not work well when $\mathcal{D}$ large/small sample
  - Problem of *overfitting*
    - The model memorizes the data rather than learning from them ⇒ bad predictions
- Structural assumption
  - Attribute variables are mutually independent given class variable, i.e.:
  - $\theta_{c,x} = \theta_c \prod_{j=1}^m \theta_{a_j|c}$ for all (c,x)
    - θ_c is the chance of $C = c$
    - $\theta_{a_j|c}$ is the chance of $A_j = a_j$ given $C = c$



- Simple but effective classifier, and very popular
  - Independence assumption not critical for classification tasks

Bayesian approach to NBC classification

- Independence assumption makes the likelihood factorize

$$\prod_{(c,x) \in \mathcal{D}} \theta_{c,x}^{n_{c,x}} = \prod_{c \in \mathcal{C}} \theta_c^{n_c} \prod_{j=1}^{m} \prod_{a_j \in \mathcal{A}_j} \theta_{a_j|c}^{n_{c,a_j}}$$

- Similarly for the prior, and the posterior is then proportional to

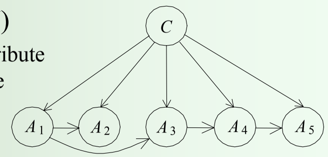
$$\prod_{c \in \mathcal{C}} \theta_c^{n_c + st_c - 1} \prod_{j=1}^{m} \prod_{a_j \in \mathcal{A}_j} \theta_{a_j|c}^{n_{c,a_j} + st_{c,a_j}}$$

which is a product of Dirichlet densities, from which

$$p(c,x|\boldsymbol{d}) = E[\theta_{c,x}|\boldsymbol{d}] = \frac{n_c + st_c}{N+s} \prod_{j=1}^{m} \frac{n_{c,a_j} + st_{c,a_j}}{n_c + st_c} = p(c|\boldsymbol{d}) \prod_{j=1}^{m} p(a_j|c,\boldsymbol{d})$$

Relaxing the independence assumption: the TAN model

- *Tree augmented naive Bayes* (TAN)
  - The dependence structure between attribute variables, given class variable, is a tree



- Learning problem
  - Learning the tree structure + probabilities
  - Mix of *maximum likelihood* and Bayesian techniques
- Classification

$$p(c, x|\boldsymbol{d}) = p(c|\boldsymbol{d}) \prod_{j=1}^{m} p(a_j|\pi_{A_j}, \boldsymbol{d})$$

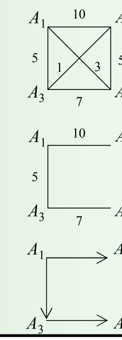
Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Learning TAN structure

- Consider the *conditional mutual information* (MI)

$$\mathrm{MI}(A_j, A_k|C) = \sum_{c\in\mathcal{C}} \theta_c \sum_{(a_j,a_k)\in\mathcal{A}_j\times\mathcal{A}_k} \theta_{a_j,a_k|c} \log \frac{\theta_{a_j,a_k|c}}{\theta_{a_j|c}\theta_{a_k|c}}$$

- Procedure
  - Set up a fully connected undirected graph
    - Node ⇔ attribute variable
    - Weight each edge (A_j, A_k) by the *empirical* $\mathrm{MI}(A_j, A_k|C)$
      - Empirical MI = replace chances with relative frequencies
  - Compute the *maximum weight spanning tree*
  - Arbitrarily choose root node
- The TAN T constructed this way maximizes the likelihood $p(\boldsymbol{d}|T)$



Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Learning TAN probabilities

- Focus on mass function $p(A_j|c,a_k,\boldsymbol{d})$
- Consider the sub-sample related to (c,a_k)
- Apply usual Bayesian learning to the sub-sample

$$p(a_j|c, a_k, \boldsymbol{d}) = \frac{n_{a_j,c,a_k} + st_{a_j,c,a_k}}{n_{c,a_k} + s}$$

  - e.g, $s = 1$ and $t_{a_j,c,a_k} = 1/|\mathcal{A}_j|$ for all a_j
  - (similarly for the class variable)

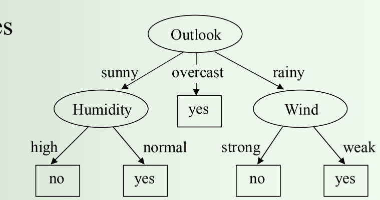
Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

The ID3 classification tree

- Example: the weather problem

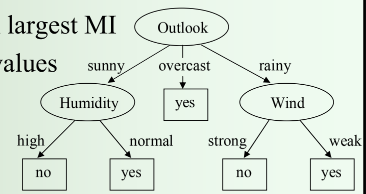
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
d1	sunny	hot	high	weak	no
d2	sunny	hot	high	strong	no
d3	overcast	hot	high	weak	yes
d4	rainy	mild	high	weak	yes
d5	rainy	cool	normal	weak	yes
d6	rainy	cool	normal	strong	no
d7	overcast	cool	normal	strong	yes
d8	sunny	mild	high	weak	no
d9	sunny	cool	normal	weak	yes
d10	rainy	mild	normal	weak	yes
d11	sunny	mild	normal	strong	yes
d12	overcast	mild	high	strong	yes
d13	overcast	hot	normal	weak	yes
d14	rainy	mild	high	strong	no

- Inner nodes = attribute variables
- Edges = attribute values
- Leaves = classes
- Path = sub-sample
- Top-down classification



ID3 learning

- Compute the empirical MI between class and attribute variables
- Root node = attribute variable with largest MI
- Branch on the variable's attribute values
- Recursion on sub-samples

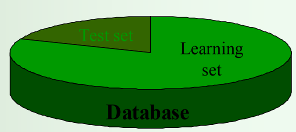


- Stop criteria
  - No more attribute variables
  - All observations in the same class
  - …others (avoid overfitting)
- Leaf: class with maximum (empirical) probability

Further considerations about KDD

Empirical evalutation: one pillar of KDD

- Empirical evaluation of classifiers
  - Split the data into learning and test sets
  - Learn a classifier from the learning set
  - Test it on the test set (hide classes)



- *Prediction accuracy*
  - Relative number of correct predictions
  - e.g., 99% means that 99% of times the predicted and the actual class coincide
- Empirically driven field to some extent
  - Classifier "works" in practice ⇒ OK

Cross validation

- A more sophisticated way to exploit data for testing
- *Tenfold cross validation*
  - Randomly partition the data $\boldsymbol{d}$ in the subsets $\boldsymbol{d_1,\ldots,d_{10}}$ of approx. equal size
  - For i = 1…10:
    - Learn the classifier from $\boldsymbol{d} \setminus \boldsymbol{d_i}$ and test it on $\boldsymbol{d_i}$
  - Average the prediction accuracies obtained over the 10 trials
- Repeated cross validation
  - Repeat the above procedure n (e.g., 10) times and average the results

Serious problems of KDD or … The necessity of imprecise probability

- The dream of KDD
  - Learning good models from data with minimal, or no, human intervention
- Ignorance matters
  - The learning process is started in conditions of ignorance about the domain
    - This is called *prior ignorance*
  - We may not know some of the values in the data set
    - This is a problem of *missing data*
- In Bayesian terms
  - Prior ignorance: there is no domain knowledge to choose a prior reliably
  - Missing data are a form of partial ignorance about the likelihood

The Bayesian way

- Prior ignorance
  - Model ignorance by so-called *non-informative priors*
  - *e.g., Bayes-Laplace's (uniform) prior, Perks'prior, …*
  - Controversial, long-debated, problem
  - Credible conclusions?
    - e.g., inferred probabilities are precise for any sample size, even equal to zero
    - Problem severe especially for *small* data sets
- Missing data
  - Assume *MAR,* data *missing at random*
    - The mechanism that turns complete into incomplete values is not systematic
    - Missing data can be discarded
  - Problem 1: MAR is often strong and unrealistic
  - Problem 2: MAR cannot be tested statistically
  - Serious problem for any size of data set

Credibility of conclusions

- Does empirical evaluation help?
  - Prior ignorance
    - Empirical evaluations are unreliable when the sample size is small
  - Missing data
    - Boolean attribute variables, A_1, A_2. Class variable represents the XOR function, i.e. $C = 1$ iff exactly one of the attribute variables is equal to 1.
    - In the available sample, A_2 is missing iff $A_2 = 0$. The prediction accuracy of the classifier is 100% on the pattern (A_1=1, A_2=*), i.e., the classifier learns to predict 1 all the times.
    - When put to work in practice, the classifier only faces cases s.t. A_2 is missing iff $A_2 = 1$. The accuracy of the classifier drops to 0%. (This is a case of selectively missing data.)

  $\Longrightarrow$ Empirical evaluation can be unreliable with missing data!
- The *law of decreasing credibility* (C. F. Manski):
  - "the credibility of inference decreases with the strength of the assumptions maintained"
- Assumptions should be weak enough to be tenable

An imprecise probability approach to prior ignorance

The imprecise Dirichlet model (IDM)

- An imprecise probability method for multinomial sampling
  - Generalization of Bayesian approach
    - Model prior ignorance by the set of all the Dirichlet(s,$\boldsymbol{t}$) priors with s fixed
    - s is a *degree of caution*
  - From this, a set of posterior densities

$$p(\theta) \propto \prod_{d \in \mathcal{D}} \theta_d^{st_d - 1}$$
$$\sum_{d \in \mathcal{D}} t_d = 1$$
$$0 < t_d < 1 \; \forall d \in \mathcal{D}$$

  - Lower and upper probabilities

$$p(c, x|\boldsymbol{d}, s, \boldsymbol{t}) = \frac{n_{c,x} + st_{c,x}}{N + s}$$
$$\Downarrow$$
$$p(c, x|\boldsymbol{d}, s) \in \left[\frac{n_{c,x}}{N+s}, \frac{n_{c,x}+s}{N+s}\right]$$

  - Imprecision (upper – lower) is $\frac{s}{N+s}$

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Credal classification

- There is generally no single *optimal* class with imprecise models
  - e.g., $p(c'|x) \in [0.5, 0.8], p(c''|x) \in [0.3, 0.6]$
    - No class *dominates* the other (they are *incomparable*); both are plausible options
  - Opposite approach: rather than looking for the optimal class, the focus in on **discarding the dominated classes**
- *Credal dominance* (*strict preference*)
  - $c' > c''$ iff $p(c'|\boldsymbol{d}, x, s, \boldsymbol{t}) > p(c''|\boldsymbol{d}, x, s, \boldsymbol{t})$ for all $\boldsymbol{t}$ in the IDM
    iff $\inf_{\boldsymbol{t}} p(c', x|\boldsymbol{d}, s, \boldsymbol{t}) / p(c'', x|\boldsymbol{d}, s, \boldsymbol{t}) > 1$
- Classification
  - Compare all the classes and output the undominated ones
    - The output is a set of classes! The set shrinks with sample size; **reliability**
- *Credal classification*

A *credal classifier* is a function $f : \mathcal{A}_1 \times \cdots \times \mathcal{A}_m \to \wp(\mathcal{C})$

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Some credal classifiers

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

The naive credal classifier (NCC)

- Recall that $p(c, x|\boldsymbol{d}, s, \boldsymbol{t}) = \frac{n_c + st_c}{N+s} \prod_{j=1}^{m} \frac{n_{c,a_j} + st_{c,a_j}}{n_c + st_c}$
- The optimization problem for testing credal dominance is

$$\begin{aligned}
\inf_{\boldsymbol{t}} \frac{p(c', x|\boldsymbol{d}, s, \boldsymbol{t})}{p(c'', x|\boldsymbol{d}, s, \boldsymbol{t})} &= \inf \left[\left(\frac{n_{c''} + st_{c''}}{n_{c'} + st_{c'}}\right)^{m-1} \prod_{j=1}^{m} \frac{n_{c',a_j} + st_{c',a_j}}{n_{c'',a_j} + st_{c'',a_j}}\right] \\
&\quad \text{s.t. } \sum_{c \in \mathcal{C}} t_c = 1 \\
&\quad\quad 0 < t_c < 1 \quad \forall c \in \mathcal{C} \\
&\quad\quad 0 < t_{c,a_j} < t_c \quad \forall c \in \mathcal{C}, j \in \{1, \ldots, m\} \\
&= \inf \left[\left(\frac{n_{c''} + st_{c''}}{n_{c'} + st_{c'}}\right)^{m-1} \prod_{j=1}^{m} \frac{n_{c',a_j}}{n_{c'',a_j} + st_{c''}}\right] \\
&\quad \text{s.t. } t_{c'} + t_{c''} = 1 \\
&\quad\quad t_{c'}, t_{c''} > 0
\end{aligned}$$

- The latter is a convex optimization problem in a single variable!
  - *Global* optimum found rapidly via numerical method, e.g., Newton-Raphson

The "grass grub" application

- Agricultural problem
  - Grass grubs are one of the major insect pests of pasture in Canterbury, New Zealand
  - Objective is to qualitatively predict the grass grub quantity based on characteristics of the paddock and on farming practice
- Data set
  - 155 observations
  - 9 attribute variables
    - Damage ranking, dry or irrigated paddock, position of the paddock, …
  - Class variable
    - Amount of grass grubs per square meter: low (l), average (a), high (h), very high (v)
    - Relative frequency of the majority class (low) is 0.316

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

NBC performs well

- Tenfold cross-validated performances of several classifiers

Classifier	Accuracy (%)
Decision Table	40.00
IB5	45.16
J48	42.58
Naive Bayes	**49.03**
OneR	45.16
PART	36.77
SMO	40.64

- The prediction accuracy tells only part of the story …

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

NBC vs NCC

- Tenfold cross validation
- NCC
  - In 60% of cases, NCC outputs a single class, with accuracy 52%
  - In the rest, NCC outputs ~2.3 classes on average (out of 4)
    - Actual class is in this set 82% of times
    - Robust way to deal with scarce knowledge
- NBC
  - A row, a prior
  - N is the accuracy of the NBC
  - N is the accuracy of the NBC, restricted to the observations where NCC produces indeterminate classifications
  - Rs is the accuracy of a uniformly random predictor
- NBC is guessing at random when NCC suspends judgment

	N	Ns	Rs
Perks	48.21	42.74	44.47
Uniform	48.83	44.24	44.47
Jeffreys	48.58	43.65	44.47

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

From another angle

- Sequential learning
- Evaluation of NBC performance based on probabilities
  - Loss = logarithmic score $= -\log_2 p(c|\boldsymbol{d}, x)$
  - Low probability, high loss

#	c	NBC	$p(c\|\boldsymbol{d},x)$	loss	NCC	$\underline{p}(c\|\boldsymbol{d},x)$, $\overline{p}\ (c\|\boldsymbol{d},x)$
1	l	lahv	0.25	2.00	lahv	0.00,1.00
2	h	l	0.04	4.64	lahv	0.00,1.00
3	h	l	0.31	1.67	lahv	0.00,1.00
4	h	h	0.75	0.41	lahv	0.06,0.94
5	l	h	0.05	4.32	lahv	0.00,0.63
6	l	h	0.20	2.33	lahv	0.00,0.68
7	h	lh	0.49	1.02	lahv	0.00,1.00
8	l	h	0.30	1.76	lahv	0.14,0.67
9	a	h	0.02	5.51	lahv	0.00,1.00
10	a	a	0.53	0.92	lahv	0.00,1.00
11	l	a	0.30	1.75	lahv	0.00,1.00
12	h	l	0.32	1.64	lahv	0.00,1.00
13	h	h	0.40	1.33	lahv	0.00,1.00
14	h	h	0.75	0.42	lahv	0.00,1.00
15	v	h	0.00	8.38	ahv	0.00,0.96

- NBC's total loss = 38.09
- Total loss of uniformly random predictor = 30
  - Probability = ¼ every time

NCC's sequential learning

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Concluding remarks on the NCC

- The NCC is reliable and fast
  - Computational complexity
    - $O(N)$ for learning
    - $O(m|\mathcal{C}^2|)$ for classification
    - As fast as NBC
- Open issues
  - e.g., NCC sometimes too cautious
  - Not clear yet why
    - NCC sensitive to irrelevant attribute variables
    - Feature selection probably a necessary step

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Tree-augmented naive credal classifier (TANC)

- Straightforward extension of TAN to credal classification
- Structure learning as in the precise case
- IDM-based learning of probabilities
  - Replace $p(a_j|c, a_k, \boldsymbol{d}, s, \boldsymbol{t}) = \dfrac{n_{a_j,c,a_k} + st_{a_j,c,a_k}}{n_{c,a_k} + s}$ with

$$\boxed{p(a_j|c, a_k, \boldsymbol{d}, s) \in \left[\frac{n_{a_j,c,a_k}}{n_{c,a_k} + s}, \frac{n_{a_j,c,a_k} + s}{n_{c,a_k} + s}\right]}$$

that is, $p(a_j|\pi_{A_j}) \in \left[\underline{p}(a_j|\pi_{A_j}), \overline{p}(a_j|\pi_{A_j})\right]$

  - This way makes different credal sets be *separately specified*
    - (Conditional) Credal sets in the same node and in different nodes

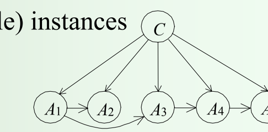
Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

TANC classification

- $\underline{p}(c, a_1, \ldots, a_m) = \underline{p}(c) \prod_{j=1}^{m} \underline{p}(a_j|\pi_{A_j})$
- Credal dominance test becomes (complete instance)

$$\frac{\underline{p}(c', a_1, \ldots, a_m)}{\overline{p}(c'', a_1, \ldots, a_m)} = \frac{\underline{p}(c')}{\overline{p}(c'')} \prod_{j=1}^{m} \frac{\underline{p}(a_j|\pi'_{A_j})}{\overline{p}(a_j|\pi''_{A_j})} > 1$$

  - Linear time
- More involved with incomplete (ignorable) instances
  - $\underline{p}(c, a_5)$?
  - $A_1, \ldots, A_4$ must be "marginalized out"
  - The expression to minimize involves a sum



- Exact minimization by propagating intervals over the tree
  - Linear time again

Some numbers

- Empirical evaluation of TANCs

Data set name (and size)		M%	C_1 %	C_s%	T%	T_s%	R_s%	S%	P_s
Iris	(150)	33.3	97.0	100.0	92.0	52.9	30.4	11.3	2.6/3
Splice	(3170)	51.9	97.7	97.8	94.7	70.4	47.1	11.1	2.1/3
Vehicle	(946)	25.4	85.1	85.8	73.6	63.2	36.8	52.4	2.1/4

  - M = relative frequency of the majority class
  - C_1 = accuracy of the TANC when precise
  - C_s = set-based accuracy of the TANC
  - T = accuracy of the TAN
  - T_s = accuracy of the TAN in the area of imprecision
  - R_s = accuracy of the uniformly random predictor in the area of imprecision
  - S = size of the area of imprecision
  - P_s = average number of classes produced by the TANC in the area of imprecision
- Good, but sometimes too cautious results (?)

Concluding remarks on the TANC

- Encouraging results
- Sometimes too cautious (?)
  - A single IDM?
    - More difficult solving of the optimization problems
  - Irrelevant feature variables
    - Problem probably more severe here than with the NCC ⇒ feature selection
- Extension to non-ignorable missing data
- Extension of structure learning to imprecise probabilities

A general framework for incomplete data based on imprecise probability

The problem of incomplete data: an introductory example

- C and A are Boolean random variables
  - $C = 1$ is the presence of a disease
  - $A = 1$ is the positive result of a medical test
- Let us do diagnosis
- Good point: you know that
  - $p(C = 0, A = 0) = 0.99$
  - $p(C = 1, A = 1) = 0.01$
  - Whence $p(C = 0 \mid A = \text{a})$ allows you to make a sure diagnosis
- Bad point: the test result can be missing
  - This is an incomplete, or *set-valued*, observation $\{0,1\}$ for A

What is $p(C = 0 \mid A \text{ is missing})$?

Example ctd

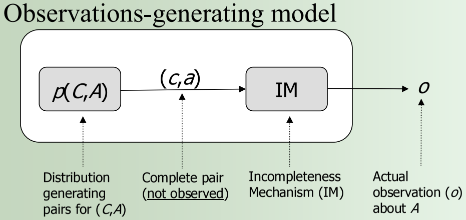
- Kolmogorov's *definition* of conditional probability *seems* to says
  - $p(C = 0 \mid A \in \{0,1\}) = p(C = 0) = 0.99$
  - i.e., with high probability the patient is healthy
- Is this right?
- In general, <u>it is not</u>
- Why?

Why?

- Because A can be **selectively** reported
- e.g., the medical test machine is broken;
  it produces an output $\Leftrightarrow$ the test is negative ($A = 0$)
  - In this case $p(C = 0 \mid A \text{ is missing}) = p(C = 0 \mid A = 1) = 0$
  - The patient is definitely ill!
  - Compare this with the naive application of Kolmogorov's updating

Modeling it the right way

- Observations-generating model

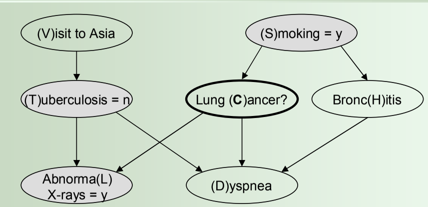


  - o is a generic value for O, another random variable
  - o can be 0, 1, or * (i.e., missing value for A)
- IM = $p(O \mid C,A)$ should not be neglected!

The correct *overall* model we need is $p(C,A)p(O \mid C,A)$

What about Bayesian nets?

- Asia net



- Let us predict C on the basis of the observation $(L,S,T) = (y,y,n)$
- Bayesian network *updating* instructs us to use $p(C \mid L = y, S = y, T = n)$ to predict C

Asia ctd

- Should we really use $p(C \mid L = y, S = y, T = n)$ to predict C?

(V,H,D) is missing

⇩

$(L,S,T,V,H,D) = (y,y,n,*,*,*)$ **is an incomplete observation**

- $p(C \mid L = y, S = y, T = n)$ is just the "naive" updating
- By using the naive updating, we are neglecting the IM!

⇩

Wrong inference in general

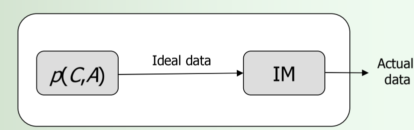
What's the problem with the IM?

- Actually, it can be neglected, if it does not act systematically
  - i.e., if CAR/MAR holds: $p(o|c,a) = \alpha$ for each (c,a)
  - Mainstream assumption in literature
- CAR/MAR is very strong and cannot be tested statistically
- Why not modeling the IM explicitly?
  - Often very difficult/costly
  - Partly, because we are not really dealing with "mechanisms" but with humans!
- In many real cases we are left with ignorance about the IM
  - No way but making ignorance become part of our models

Statistical framework

Statistical treatment of incomplete data

- Pervasive problem in statistical practice, important theoretical issue
  - Subtleties
- Fundamental distinction
  - Ideal data
    - Produced by a certain process
    - Directly unobservable
  - Actual data
    - Set-based view of ideal data, produced by the incompleteness mechanism
      - e.g., Space of possibilities = {1,2,3}; ideal data = 2; actual data = {2,3}; missing data = {1,2,3}
    - Observable



- Learning and classification with incomplete data
  - Learning set and/or observation to classify incomplete
- Focus on very general framework
  - e.g., no i.i.d. process

Learning – ideal variables

- Ideal learning data are regarded as an instance of the random matrix

$$\begin{bmatrix} C_1 & A_{11} & \cdots & A_{1j} & \cdots & A_{1k} \\ \vdots & \vdots & & \vdots & & \vdots \\ C_i & A_{i1} & \cdots & A_{ij} & \cdots & A_{ik} \\ \vdots & \vdots & & \vdots & & \vdots \\ C_N & A_{N1} & \cdots & A_{Nj} & \cdots & A_{Nk} \end{bmatrix} = \begin{bmatrix} C_1 & X_1 \\ \vdots & \vdots \\ C_i & X_i \\ \vdots & \vdots \\ C_N & X_N \end{bmatrix} = \begin{bmatrix} D_1 \\ \vdots \\ D_i \\ \vdots \\ D_N \end{bmatrix} = \boldsymbol{D}$$

  - C_i's are *class variables*, with values in $\mathcal{C}$
  - A_{ij}'s are *attribute variables*, with values in $\mathcal{A}_j$ for each (i,j)
  - Each row D_i in the matrix is called *unit*, with values in $\mathcal{D} = \mathcal{C} \times \mathcal{X} = \mathcal{C} \times \mathcal{A}_1 \times \cdots \times \mathcal{A}_m$

Learning – actual variables

- Consider one more random variable

- $\boldsymbol{O}$ is the *actual observation* of $\boldsymbol{D}$
  - $\boldsymbol{D}$ takes values $\boldsymbol{d}$ from $\mathcal{D}^N$
  - $\boldsymbol{O}$ takes values $\boldsymbol{o}$ from $\wp(\mathcal{D}^N)$
    - Rather than observing $\boldsymbol{d}$, you observe a set $\boldsymbol{o}$ that contains it
    - This is called *coarsening*, i.e., looking at $\boldsymbol{d}$ with different levels of detail
    - Note that $\boldsymbol{o}$ is simply a symbol, not a set, when regarded as the value of $\boldsymbol{O}$

Learning – assumptions

- Overall model = joint density $p(T, \boldsymbol{D}, \boldsymbol{O})$
  - T is a random parameter with values $\theta \in \Theta$
- Factorization
  - $p(T, \boldsymbol{D}, \boldsymbol{O}) = p(T)p(\boldsymbol{D}|T)p(\boldsymbol{O}|\boldsymbol{D})$
    - $p(T)$ is an imprecise prior for T, i.e., it belongs to a certain non-empty set
    - The IM depends only on $\boldsymbol{D}$
- Accuracy (of mechanism)
  - $p(\boldsymbol{o}|\boldsymbol{d}) = 0$ if $\boldsymbol{d}$ $\boldsymbol{o}$
  - Connection between complete and incomplete observations
- Positivity
  - Ideal data: $p(\boldsymbol{D}) > 0$
  - Actual observation: $p(\boldsymbol{o}) > 0$
    - There exists $\boldsymbol{d}$ s.t. $p(\boldsymbol{o}|\boldsymbol{d}) > 0$

Learning – problem

- Formulation of the learning problem
  - Using observed data $\boldsymbol{o}$ to update beliefs about a function $f : \Theta \to \mathbb{R}$
    - e.g., θ real number, $f(\theta) = \theta$
  - In the precise framework, one would compute $E(f|\boldsymbol{o})$

- $$\begin{aligned} E(f|\boldsymbol{o}) &= \frac{\int f(\theta)p(\theta,\boldsymbol{o})d\theta}{p(\boldsymbol{o})} \\ &= \frac{\int f(\theta)\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\theta)p(\boldsymbol{d}|\theta)p(\boldsymbol{o}|\boldsymbol{d})d\theta}{\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d})p(\boldsymbol{d})} \end{aligned}$$
  - Using Factorization, Accuracy and Positivity

- What about $p(\boldsymbol{o}|\boldsymbol{d})$ and $p(\theta)$?
  - Imprecise knowledge

Learning – express ignorance about the IM

- Focus on $p(\boldsymbol{o}|\boldsymbol{d})$, $\boldsymbol{d} \in \boldsymbol{o}$
  - Call $p(\boldsymbol{o}|\boldsymbol{D})$ the vector with elements $p(\boldsymbol{o}|\boldsymbol{d})$, $\boldsymbol{d} \in \boldsymbol{o}$
- We can only constrain the set of possible vectors $p(\boldsymbol{o}|\boldsymbol{D})$

$$\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) > 0$$
$$0 \le p(\boldsymbol{o}|\boldsymbol{d}) \le 1,\ \boldsymbol{d} \in \boldsymbol{o}$$

  - The restriction on the sum is due to Accuracy
  - The inequality following the sum is due to Positivity
- The inequalities define an open linear set called $\mathcal{P}(\boldsymbol{o}|\boldsymbol{D})$
- Call $\mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})$ the approximating closed set

$$\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \ge \varepsilon$$
$$0 \le p(\boldsymbol{o}|\boldsymbol{d}) \le 1,\ \boldsymbol{d} \in \boldsymbol{o}$$

Learning – goal

- Goal: $\underline{E}(f|\boldsymbol{o}) = \inf_{p(T) \in \mathcal{P}(T)} \inf_{p(\boldsymbol{o}|\boldsymbol{D}) \in \mathcal{P}(\boldsymbol{o}|\boldsymbol{D})} E(f|\boldsymbol{o})$
- Focus on $\min_{p(\boldsymbol{o}|\boldsymbol{D}) \in \mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} E(f|\boldsymbol{o})$
  - i.e., $\min_{p(\boldsymbol{o}|\boldsymbol{D}) \in \mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \frac{\sum_{\mathrm{d} \in \mathrm{o}} p(\boldsymbol{o}|\boldsymbol{d}) \int f(\theta) p(\theta) p(\boldsymbol{d}|\theta) d\theta}{\sum_{\mathrm{d} \in \mathrm{o}} p(\boldsymbol{o}|\boldsymbol{d}) p(\boldsymbol{d})}$
  - Objective function is ratio of linear functions
- Fractional programming theorem
  - Consider $\min_{x \in S} \frac{q(x)}{r(x)}$ where S = compact subset of $\mathbb{R}^v$
    - q,r continuous and r positive on S
  - Define $h: \mathbb{R} \to \mathbb{R}$, $h(\mu) = \min_{x \in S}[q(x) - \mu r(x)]$
  - Then $\mu^* = \mathrm{argmin}_{x \in S} \frac{q(x)}{r(x)} \iff h(\mu^*) = 0$
    - h has a single zero

Learning – solution: conservative learning rule

- $$\begin{aligned} h(\mu) &= \min_{p(\boldsymbol{o}|\boldsymbol{D}) \in \mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \left[\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \int f(\theta) p(\theta) p(\boldsymbol{d}|\theta) d\theta - \mu \sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) p(\boldsymbol{d}) \right] \\ &= \min_{p(\boldsymbol{o}|\boldsymbol{D}) \in \mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) p(\boldsymbol{d}) \left[\int f(\theta) p(\theta|\boldsymbol{d}) d\theta - \mu \right] \\ &= \min_{p(\boldsymbol{o}|\boldsymbol{D}) \in \mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) p(\boldsymbol{d}) \left[E(f|\boldsymbol{d}) - \mu \right] \end{aligned}$$
- $\mu^* = \min_{\boldsymbol{d} \in \boldsymbol{o}} E(f|\boldsymbol{d}) \implies h(\mu^*) = 0$
- $\min_{p(\boldsymbol{o}|\boldsymbol{D}) \in \mathcal{P}_\varepsilon(\boldsymbol{o}|\boldsymbol{D})} E(f|\boldsymbol{o}) = \min_{\boldsymbol{d} \in \boldsymbol{o}} E(f|\boldsymbol{d})$

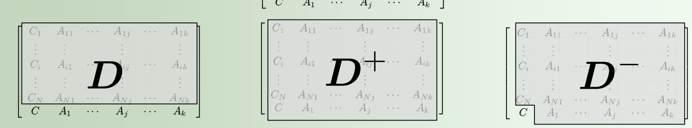
$$\Downarrow$$

$$\boxed{\underline{E}(f|\boldsymbol{o}) = \inf_{p(T) \in \mathcal{P}(T)} \min_{\boldsymbol{d} \in \boldsymbol{o}} E(f|\boldsymbol{d})}$$

Classification – variables

- Learning variable $\boldsymbol{D}$ as before
- The ideal observation to classify is regarded as a (partial) instance of the further unit $(C, A_1, \ldots, A_j, \ldots, A_N) = (C, X) = D$ with values in $\mathcal{D}$
- Summary of ideal variables

$$\begin{bmatrix} C_1 & A_{11} & \cdots & A_{1j} & \cdots & A_{1k} \\ \vdots & \vdots & & \vdots & & \vdots \\ C_i & A_{i1} & \cdots & A_{ij} & \cdots & A_{ik} \\ \vdots & \vdots & & \vdots & & \vdots \\ C_N & A_{N1} & \cdots & A_{Nj} & \cdots & A_{Nk} \\ C & A_1 & \cdots & A_j & \cdots & A_k \end{bmatrix}$$



- Actual variables: $\boldsymbol{O}$, $\boldsymbol{O}^+$, $\boldsymbol{O}^-$
  - Taking values in the related power sets

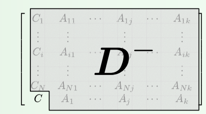
Classification – assumptions

- Factorization, Accuracy, Positivity, as before
- One more: Independence
  - $p(\boldsymbol{O}^+|\boldsymbol{D}^+) = p(\boldsymbol{O}^+|\boldsymbol{D}^-)$
  - The mechanism does not depend on what we want to predict, i.e., C
  - Without it, we could never exclude that $p(c|\boldsymbol{o}^+) = 0$ or $p(c|\boldsymbol{o}^+) = 1$
    - Independence avoids problem of vacuous conclusions
- On the meaning of Independence
  - Equivalent to $p(C|\boldsymbol{O}^+, \boldsymbol{D}^-) = p(C|\boldsymbol{D}^-)$
  - Independence characterizes problems of incomplete data, in the sense that
    - Once you know $\boldsymbol{D}^-$, there is no more a problem of incomplete (or missing) data
    - Consider the opposite case: you have not included some factor
  - Independence not very restrictive in practice
    - If $\boldsymbol{D}^-$ represents the set of all "factors" that you deem important to predict C, Independence follows automatically

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Classification – goal and solution: conservative updating rule

- Focus on $E_C(g|\boldsymbol{o}^+) = \sum_{c\in\mathcal{C}} g(c)p(c|\boldsymbol{o}^+)$
  - Generic g
  - e.g., with credal dominance: $g(c) = \begin{cases} 1 & c = c' \\ -1 & c = c'' \\ 0 & \text{otherwise} \end{cases}$
- Goal: $\underline{E}_C(g|\boldsymbol{o}^+)$
- Theorem: $\boxed{\underline{E}_C(g|\boldsymbol{o}^+) = \inf_{p(T)\in\mathcal{P}(T)} \min_{\boldsymbol{d}^-\in\boldsymbol{o}^-} E_C(g|\boldsymbol{d}^-)}$



- Incomplete data, via the *conservative updating rule*, naturally produce credal classifiers!

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Obtaining stronger conclusions: the mixed rules

- CAR might hold in some cases
  - This should not be neglected in order to strengthen conclusions
- Some attribute variables are CAR, the others non-ignorable
- Mixed rules
  - Learning: $\underline{E}(f|\boldsymbol{o},\hat{\boldsymbol{o}}) = \inf_{p(T)\in\mathcal{P}(T)} \min_{\boldsymbol{d}\in\boldsymbol{o}} E(f|\boldsymbol{d},\hat{\boldsymbol{d}}\in\hat{\boldsymbol{o}})$
  - Classification: $\underline{E}_C(g|\boldsymbol{o}^+,\hat{\boldsymbol{o}}^+) = \inf_{p(T)\in\mathcal{P}(T)} \min_{\boldsymbol{d}^-\in\boldsymbol{o}^-} E_C(g|\boldsymbol{d}^-,\hat{\boldsymbol{d}}^-\in\hat{\boldsymbol{o}}^-)$

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

The conservative rules in practice

Unstructured case

Conservative learning in the unstructured case

- Table with missing data
- Focus on the empirical mass function $p(X)$
  - and on derived descriptive indexes
- Conservative learning:
  - Consider all the complete tables (completions) consistent with the incomplete one
- <u>Problem of complexity</u>
  - n missing values $\Rightarrow 3^n$ completions
    - 3^5 in the example

Obs.	O_{A1}	O_{A2}
o_1	1	1
o_2	1	*
o_3	1	*
o_4	*	2
o_5	3	2
o_6	*	*

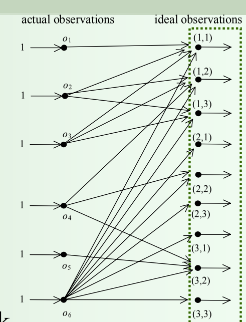
$(A_1, A_2) = X$ in $\{1,2,3\}^2$
$(O_{A_1}, O_{A_2}) = O$ in $\{1,2,3,*\}^2$

Alternative representation

- Data set $\Rightarrow$ network of flow

Obs.	O_{A1}	O_{A2}
o_1	1	1
o_2	1	*
o_3	1	*
o_4	*	2
o_5	3	2
o_6	*	*

$\Longrightarrow$



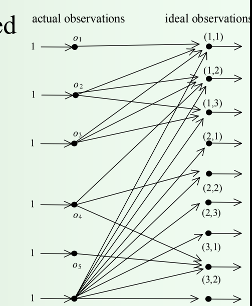
- Let the vector $\boldsymbol{f}$ be a given *flow* in the network
  - With elements $f(a,b)$: flow on the arc $a \rightarrow b$
  - Denote by $\boldsymbol{f}_X$ the sub-vector for these arcs
  - $p(X) = (1/N)\, \boldsymbol{f}_X$

Flow network = polytope

- The possible flows make up a polytope Ψ
- Let $R(o_i)$ = set of nodes to which o_i is mapped
- $\boldsymbol{f} \in \Psi$ where Ψ is the defined by

$$\begin{array}{ll} \sum_{x \in R(o_i)} f(o_i, x) = 1 & i = 1, \ldots, N \\ \sum_{j: x \in R(o_j)} f(o_j, x) = f(x, \cdot) & \forall x \in \mathcal{X} \\ \boldsymbol{f} \geq 0 & \end{array}$$

$\Longleftarrow$



- Linear constraints
- Note that $\boldsymbol{f}$ is not required to be integer

Properties of the flow model

- $\tilde{\mathcal{P}}$ = finite set of joint mass functions from all the completions
- $\tilde{\mathcal{P}}_f$ = finite set of joint mass functions from all the integer flows
- $\mathcal{P}$ and $\mathcal{P}_f$ are their respective convex hulls

- Lemma 1: $\tilde{\mathcal{P}} = \tilde{\mathcal{P}}_f$
- Lemma 2: The extreme points of Ψ are integer flows

- Theorem: $\mathcal{P} = \mathcal{P}_f = \left\{ \frac{1}{N} \boldsymbol{f}_X \middle| \boldsymbol{f} \in \Psi \right\}$

- The constraints for Ψ provide us with an implicit description of $\mathcal{P}$
- No need to consider the extreme distributions explicitly

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Linear-programming computations

- $\mathcal{X}', \mathcal{X}'' \subseteq \mathcal{X}$
- Optimizing linear functions
  - e.g., $\underline{p}(\mathcal{X}'), \overline{p}(\mathcal{X}')$
  - Exact computation, polynomial time

- Optimizing special types of fractional linear functions
  - e.g., $\underline{p}(\mathcal{X}'|\mathcal{X}''), \overline{p}(\mathcal{X}'|\mathcal{X}'')$
  - Exact computation, polynomial time

- Optimizing general fractional linear functions
  - e.g., $\underline{E}(\mathcal{X}'|\mathcal{X}''), \overline{E}(\mathcal{X}'|\mathcal{X}'')$
  - Approximate computation (any precision), polynomial time

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Further properties

- $\underline{p}$ is a belief function
  - Special algorithms on the network lead to linear-time computations
- Modeling more general patterns of missing data
  - Partially missing type 1
    - Not all values are possible replacements
  - Partially missing type 2
    - Intra-observation dependencies
- Easy extension to an inferential approach

  imprecise Dirichlet model + treatment of missing data

  $\Updownarrow$ $\Updownarrow$

  prior ignorance + partial ignorance about the likelihood

  - e.g., $\left[\underline{p}(\mathcal{X}'|\boldsymbol{o},s), \overline{p}(\mathcal{X}'|\boldsymbol{o},s)\right] = \left[\frac{\underline{n}(\mathcal{X}')}{N+s}, \frac{\overline{n}(\mathcal{X}')+s}{N+s}\right]$

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

The naively structured case

Inference of the NCC from incomplete data

- Test of credal dominance with an incomplete learning sample

$$\min_{\boldsymbol{d}\in\boldsymbol{o}} \inf_{\boldsymbol{t}} \frac{p(c',x|\boldsymbol{d},s,\boldsymbol{t})}{p(c'',x|\boldsymbol{d},s,\boldsymbol{t})} = \quad \inf \left[\left(\frac{n_{c''}+st_{c''}}{n_{c'}+st_{c'}} \right)^{m-1} \prod_{j=1}^{m} \frac{\underline{n}_{c',a_j}}{\overline{n}_{c'',a_j}+st_{c''}} \right]$$
$$\text{s.t. } t_{c'}+t_{c''}=1, \quad t_{c'},t_{c''}>0$$

- Same complexity as with complete data

The dementia application

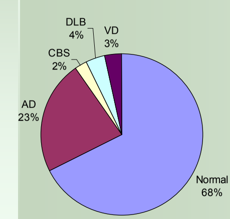
- Real problem of diagnosing dementias
- Application of the NCC
  - Inference from incomplete data
  - Conservative learning
- Aim
  - To show that developed methods are useful
  - To compare them with more traditional methods
  - To show that developed methods can be used in practice right now

The Cognitive Drug Research (CDR) System

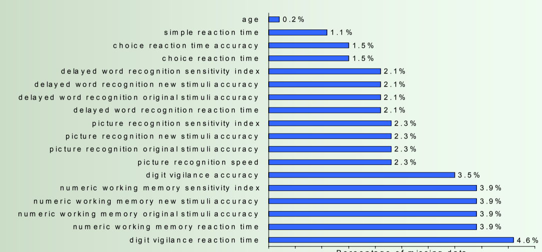
- Assessment of cognitive functions
- Most widely used system in clinical research
- Computer-based automated system
- CDR tasks and functions assessed
  - Attention, concentration and vigilance
    - Simple reaction time, choice reaction time, digit vigilance
  - Working memory and executive function
  - Episodic secondary memory
    - Picture recognition

The CDR database

- 3400 records of patients (observations)
  - Test results + disease
- 5 categories of patients (classes)
  - normal (NORM), to undergo Coronary Bypass Surgery (CBS), Dementia with Lewy Bodies (DLB), Alzheimer Disease (AD), Vascular Dementia (VD)



Distribution of the classes



Differentiating dementias

- Placing a demented patient in the right class
- Training and test data
  - ~50% - 50%
  - Results

C_1%	Cs%	N%	Ns%	Rs%	S%	Ps%
94.05	98.42	89.76	75.59	45.6	23.22	2.3/4

- C_1 = accuracy of the NCC when precise
- C_s = set-based accuracy of the NCC
- N = accuracy of the NCC
- N_s = accuracy of the NCC in the area of imprecision
- R_s = accuracy of the uniformly random predictor in the area of imprecision
- S = size of the area of imprecision
- P_s = average number of classes produced by the NCC in the area of imprecision

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Other cases

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

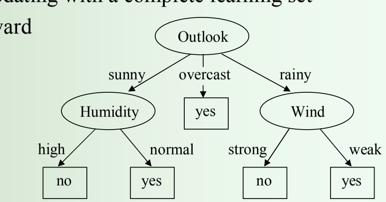
Robust classification with TANCs

- TANCs can be extended to classifying non-ignorable missing data
  - Case of ignorable missing data already there
  - New extension = conservative updating with a complete learning set
  - As the Markov blanket of the class node is singly connected (a tree)
    - Work in progress
- Extension to conservative learning
  - This is more difficult
  - Missing data create non-separately specified credal sets

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

The case of ID3

- ID3 can be extended to classifying non-ignorable missing data
  - New extension = conservative updating with a complete learning set
  - How: follow all the paths downward
- Learning is more difficult
  - Many trees



Concluding remarks

- Credible conclusions need tenable assumptions
  - Even if empirical validations are possible!
- Imprecise probabilities permit working with weak assumptions
  - Prior ignorance, incomplete data
- Methods exist that already work in practice
  - Useful
  - Efficient
- More work is (always) needed
  - Creating more methods
  - Developing efficient implementations

Annotated bibliography

- Introduction to pattern classification
  - R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification.* Wiley, 2001. 2nd edition.
- Bayesian data analysis
  - A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman, 1995.
- First definition of the NBC
  - R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley, New York, 1973.
- TAN classifier
  - N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian networks classifiers. *Machine Learning*, 29(2/3):131–163, 1997.

Annotated bibliography

- Classification trees
  - J. R. Quinlan. *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Mateo, 1993.
- The classical book on (ignorable) missing data
  - R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, 1987.
- Imprecise probability
  - P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York, 1991.
- The imprecise Dirichlet model
  - P. Walley. Inferences from multinomial data: learning about a bag of marbles. *J. R. Statist. Soc. B*, 58(1):3–57, 1996.

Annotated bibliography

- Definition of credal classification
  - M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.
- Inference of the NCC from (possibly incomplete non-ignorable) data
  - M. Zaffalon. Statistical inference of the naive credal classifier. In G. de Cooman, T. L. Fine, and T. Seidenfeld, editors, *ISIPTA '01: Proceedings of the Second International Symposium on Imprecise Probabilities and Their Applications*, pages 384–393, The Netherlands, 2001. Shaker.
- Grass grub application
  - M. Zaffalon. Credible classification for environmental problems. *Environmental modelling and software.* To appear.
- TANC classifier
  - M. Zaffalon and E. Fagiuoli. Tree-based credal networks for classsification. *Reliable Computing*, 9(6):487–509, 2003.

Annotated bibliography

- Wide view of incomplete non-ignorable data in identification prob.
  - C. F. Manski. *Partial Identification of Probability Distributions*. Springer-Verlag, New York, 2003.
- Non-ignorable missing data in artificial intelligence
  - M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2):147–170, 2001.
- Problems of naive updating raised for the first time
  - G. Shafer. Conditional probability. *International Statistical Review*, 53:261–277, 1985.
- Naive updating works iff CAR/MAR; CAR/MAR is strong
  - P. Grünwald and J. Halpern. Updating probabilities. *Journal of Artificial Intelligence Research*, 19:243–278, 2003.

Marco Zaffalon: Knowledge discovery from data sets under tenable assumptions. School on Imprecise Probabilities, Lugano (CH), 27-31 July 2004.

Annotated bibliography

- First definition of the conservative updating rule (expert systems)
  - G. de Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*. Forthcoming.
- Statistical framework for non-ignorable incomplete data
  - M. Zaffalon. Statistical classification with incomplete observations. Technical Report IDSIA-10-04, IDSIA, 2004. In progress.
- Dealing with non-ignorable missing data in the unstructured case
  - M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105–122, 2002.
- Dementia application
  - M. Zaffalon, K. Wesnes, and O. Petrini. Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. *Artificial Intelligence in Medicine*, 29(1–2):61–79, 2003.