Feasibility Study of an Interactive Medical Diagnostic Wikipedia

Jiří Grim

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic, Prague
Department of Pattern Recognition
http://www.utia.cas.cz/RO

SPMS 2016, June 20-24, 2016, Dobřichovice, Czech Republic

Outline

Medical Diagnostic System

Goal: platform to accumulate decision-making know-how

$\boldsymbol{x}_n \in \mathcal{X}_n$: discrete variables, $n \in \mathcal{N}$, $\mathcal{N} = \{1, \ldots, N\}$
$\mathcal{X}_n$: finite sets of values $(\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N)$

symptom variables (questions): $\bar{\boldsymbol{x}} = (x_1, \ldots, x_K) \in \bar{\mathcal{X}}$
diagnostic variables (diagnoses): $y_j = x_{K+j},\ j \in \mathcal{J},\ \mathcal{J} = \{1, \ldots, J\}$

$\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{X},\ \boldsymbol{y} = (y_1, y_2, \ldots, y_J) \in \mathcal{Y},\ N = K + J,\ (N \approx 10^3)$

Statistical model: mixture of product components (restrictive ?)

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m), \quad F(\boldsymbol{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m), \ \ \mathcal{M} = \{1, \ldots, M\}$$

$f_n(x_n|m), n \in \mathcal{N}$: univariate discrete distributions $\sum_{x_n \in \mathcal{X}_n} f_n(x_n|m) = 1$

Properties of product mixtures:

- simple evaluation of (conditional) marginals
- estimation from incomplete data vectors by EM algorithm
- "dimensionless" applicability (structural model)

Product Mixture Model as a Knowledge Base of PES

Estimated knowledge base: $P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|m), \quad \boldsymbol{x} \in \mathcal{X}$

INFERENCE MECHANISM: $(P_j(y_j) = P_{K+j}(x_{K+j}), j \in \mathcal{J})$

definite input: $\boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_k}), \quad \mathcal{C} = \{i_1, \ldots, i_k\} \subset \{1, 2, \ldots, K\}$

response: conditional distribution $P_{j|C}(y_j|\boldsymbol{x}_C)$ given $\boldsymbol{x}_C \in \mathcal{X}_C$

$$P_{j|C}(y_j|\boldsymbol{x}_C) = \frac{P_{jC}(y_j, \boldsymbol{x}_C)}{P_C(\boldsymbol{x}_C)} = \sum_{m\in\mathcal{M}} W_m(\boldsymbol{x}_C) f_j(y_j|m)$$

$$P_{jC}(y_j, \boldsymbol{x}_C) = \sum_{m\in\mathcal{M}} w_m f_j(y_j|m) F_C(\boldsymbol{x}_C|m), \quad F_C(\boldsymbol{x}_C|m) = \prod_{i\in\mathcal{C}} f_i(x_i|m)$$

$$W_m(\boldsymbol{x}_C) = \frac{w_m F_C(\boldsymbol{x}_C|m)}{\sum_{k\in\mathcal{M}} w_k F_C(\boldsymbol{x}_C|k)}, \quad \boldsymbol{x}_C \in \mathcal{X}_C, \ m \in \mathcal{M}$$

uncertain input: $P_C^*(\boldsymbol{x}_C), \boldsymbol{x}_C \in \mathcal{X}_C, \quad W_m^* = \sum_{\boldsymbol{x}_C\in\mathcal{X}_C} W_m(\boldsymbol{x}_C) P_C^*(\boldsymbol{x}_C)$

response:

$$P_n^*(y_j) = \sum_{\boldsymbol{x}_C\in\mathcal{X}_C} P_{j|C}(y_j|\boldsymbol{x}_C) P_C^*(\boldsymbol{x}_C) = \sum_{m\in\mathcal{M}} W_m^* f_j(y_j|m)$$

Ordered List of Relevant Diagnostic Variables

Ordering of diagnostic variables given the symptoms $\mathbf{x}_C$

set of symptoms $\mathbf{x}_C = (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{X}_C, \;\; \mathcal{C} = \{i_1, \ldots, i_k\}$

CRITERION: difference between the a priori distribution $P_j(y_j), j \in \mathcal{J}$ and the conditional distribution $P_{j|C}(y_j|\mathbf{x}_C)$

Kullback-Leibler information divergence:

$$I(P_j(\cdot)||P_{j|C}(\cdot|\mathbf{x}_C)) = \sum_{y_j \in \mathcal{Y}_j} P_j(y_j) \log \frac{P_j(y_j)}{P_{j|C}(y_j|\mathbf{x}_C)}, \; j \in \mathcal{J}$$

decrease of conditional entropy:

$$\triangle H = \frac{H(\mathcal{Y}_j) - H_{\mathbf{x}_c}(\mathcal{Y}_j)}{H(\mathcal{Y}_j)}, \; j \in \mathcal{J}$$

absolute difference: $\triangle P = \sum_{y_j \in \mathcal{Y}_j} |P_j(y_j) - P_{j|C}(y_j|\mathbf{x}_C)|$

REMARK: Medical aspects may be more important for the choice of y_j.

Optimal Sequential Decision-Making

Optimal choice of informative questions

given a set of symptoms $\boldsymbol{x}_C = (x_{i_1}, \ldots, x_{i_k}) \in \mathcal{X}_C$ and a diagnose y_j
$\Rightarrow$ the most informative question x_i, $(i \notin \mathcal{C})$ with respect to y_j

CRITERION: maximum conditional Shannon information given $\boldsymbol{x}_C$:

$$I_{\boldsymbol{x}_C}(\mathcal{X}_i, \mathcal{Y}_j) = H_{\boldsymbol{x}_C}(\mathcal{Y}_j) - H_{\boldsymbol{x}_C}(\mathcal{Y}_j|\mathcal{X}_i), \qquad i^* = \arg\max_{i \notin C}\{I_{\boldsymbol{x}_C}(\mathcal{X}_i, \mathcal{Y}_j)\}$$

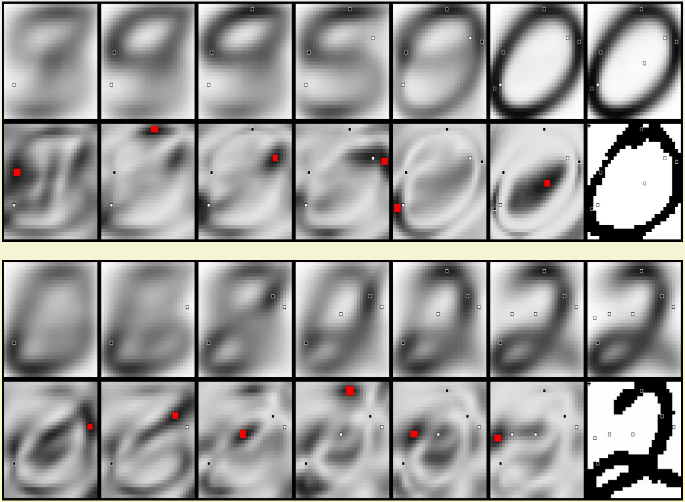
$$H_{\boldsymbol{x}_C}(\mathcal{Y}_j) = \sum_{y_j \in \mathcal{Y}_j} -P(y_j|\boldsymbol{x}_C) \log P(y_j|\boldsymbol{x}_C), \quad p(y_j|\boldsymbol{x}_C) = P(y_j, \boldsymbol{x}_C)/P(\boldsymbol{x}_C)$$

$$H_{\boldsymbol{x}_C}(\mathcal{Y}_j|\mathcal{X}_i) = \sum_{x_i \in \mathcal{X}_i} p(x_i|\boldsymbol{x}_C) \sum_{y_j \in \mathcal{Y}_j} -P(y_j|\boldsymbol{x}_C, x_i) \log P(y_j|\boldsymbol{x}_C, x_i)$$

$$P(y_j|\boldsymbol{x}_C, x_i) = P(y_j|\boldsymbol{x}_C, x_i)/P(\boldsymbol{x}_C, x_i), \; i \notin \mathcal{C}, \; y_j \in \mathcal{Y}_j$$

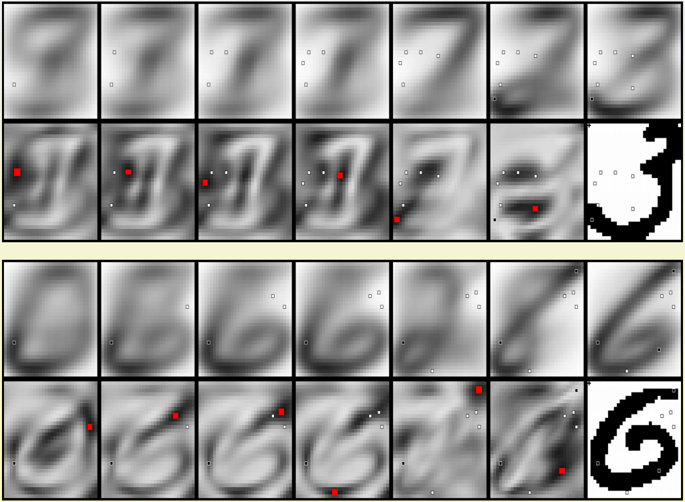
REMARK: A unique possibility to compute conditional Shannon information.

Example: Optimal Sequential Recognition of Numerals



First row: What the classifier expects given some raster fields
Second row: Conditional informativity of the raster fields.

Example: Optimal Sequential Recognition of Numerals



First row: What the classifier expects given some raster fields
Second row: Conditional informativity of the raster fields

Controlled Dialog Scheme

Input information and inference mechanism

- choice of input questions and specification of symptoms: $\boldsymbol{x}_C \in \mathcal{X}_C$
- $\Rightarrow$ conditional distributions of diagnostic variables: $P_{j|C}(y_j|\boldsymbol{x}_C), j \in \mathcal{J}$

Evaluation of diagnostic variables

- ordered list of relevant diagnostic variables $y_j, j \in \mathcal{J}$ given $\boldsymbol{x}_C$
- $\Rightarrow$ choice of the most relevant diagnostic variable y_j

Choice of the most informative question with respect to diagnose y_j

- conditional informativity $I_{\boldsymbol{x}_C}(x_i, y_j)$ of questions $x_i, i \notin \mathcal{C}$
- $\Rightarrow$ ordered list of informative questions x_i given y_j and $\boldsymbol{x}_C \in \mathcal{X}_C$

Knowledge base update

- inclusion of new data vectors and re-estimation of the mixture model
- dialog scheme produces incomplete (sparse) data vectors

EM Algorithm for Discrete Product Mixtures

independent identically distributed observations:

$$\mathcal{S} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(I)}\}, \quad \boldsymbol{x}^{(i)} \in \mathcal{X}$$

log-likelihood function:

$$L = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} \log[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m)] = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} \log \left(\sum_{m \in \mathcal{M}} w_m \prod_{n \in \mathcal{N}} f_n(x_n|m) \right)$$

EM iteration equations: $(m \in \mathcal{M}, \quad \boldsymbol{x} \in \mathcal{S})$

$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|m)}{\sum_{j \in \mathcal{M}} w_j F(\boldsymbol{x}|j)}, \quad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} q(m|\boldsymbol{x})$$

$$f_n^{'}(\xi|m) = \frac{1}{\sum_{\boldsymbol{x} \in \mathcal{S}} q(m|\boldsymbol{x})} \sum_{\boldsymbol{x} \in \mathcal{S}} \delta(\xi, x_n) q(m|\boldsymbol{x}), \quad \xi \in \mathcal{X}_n, \quad n \in \mathcal{N}$$

MONOTONIC PROPERTY: $L^{(t+1)} - L^{(t)} \geq 0, \quad t = 0, 1, 2, \ldots$

REMARK: Risk of multiple latent underflow in evaluation of $q(m|\boldsymbol{x})$!

Structural Mixture Model

structural variables: $\phi_{mn} \in \{0,1\}, \; m \in \mathcal{M}, \; n \in \mathcal{N}$

$\phi_{mn} = 0 \Rightarrow$ distribution $f_n(x_n|m)$ is replaced by fixed "background" $f_n(x_n|0)$

$$F(\boldsymbol{x}|m) = \prod_{n\in\mathcal{N}} f_n(x_n|m)^{\phi_{mn}} f_n(x_n|0)^{1-\phi_{mn}}, \; m \in \mathcal{M}, \;\; (f_n(x_n|0) = P_n(x_n))$$

$$P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} F(\boldsymbol{x}|m) f(m) = F(\boldsymbol{x}|0) \sum_{m\in\mathcal{M}} w_m G(\boldsymbol{x}|m, \phi_m)$$

$$F(\boldsymbol{x}|0) = \prod_{n\in\mathcal{N}} f_n(x_n|0), \qquad G(\boldsymbol{x}|m, \phi_m) = \prod_{n\in\mathcal{N}} \left[\frac{f_n(x_n|m)}{f_n(x_n|0)}\right]^{\phi_{mn}}$$

"background distribution" $F(\boldsymbol{x}|0)$ **cancels in the conditional weights:**

$$W_m(\boldsymbol{x}_C) = \frac{w_m G_C(\boldsymbol{x}_C|m, \phi_m)}{\sum_{j=1}^{M} w_j G_C(\boldsymbol{x}_C|j, \phi_j)}, \qquad G_C(\boldsymbol{x}_C|m, \phi_m) = \prod_{i\in\mathcal{C}} \left[\frac{f_i(x_i|m)}{f_i(x_i|0)}\right]^{\phi_{mi}}$$

Consequence: **structural mixture can be treated as dimensionless**

ÚTIA

Structural Modification of EM Algorithm

structural variables can be optimized in full generality:

$$L = \frac{1}{|\mathcal{S}|} \sum_{x \in \mathcal{S}} \log \left[F(\boldsymbol{x}|0) \sum_{m \in \mathcal{M}} w_m G(\boldsymbol{x}|m, \phi_m) \right], \quad F(\boldsymbol{x}|0) = \prod_{n \in \mathcal{N}} f_n(x_n|0)$$

EM Algorithm: ($F(\boldsymbol{x}|0)$ cancels in the conditional weights $q(m|\boldsymbol{x})$)

$$q(m|\boldsymbol{x}) = \frac{G(\boldsymbol{x}|m, \phi_m) w_m}{\sum_{j \in \mathcal{M}} G(\boldsymbol{x}|j, \phi_j) w_j}, \quad w_m' = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} q(m|\boldsymbol{x}), \quad m \in \mathcal{M}, \boldsymbol{x} \in \mathcal{S}$$

$$f_n'(\xi|m) = \frac{1}{\sum_{\boldsymbol{x} \in \mathcal{S}} q(m|\boldsymbol{x})} \sum_{\boldsymbol{x} \in \mathcal{S}} \delta(\xi, x_n) q(m|\boldsymbol{x}), \quad n \in \mathcal{N}$$

structural criterion: Kullback-Leibler information divergence

$$\gamma_{mn}' = \sum_{\boldsymbol{x} \in \mathcal{S}} \frac{q(m|\boldsymbol{x})}{|\mathcal{S}|} \log \left[\frac{f_n'(x_n|m)}{f_n(x_n|0)} \right] = w_m' \sum_{x_n \in \mathcal{X}_n} f_n'(x_n|m) \log \frac{f_n'(x_n|m)}{f_n(x_n|0)}$$

optimization: $\phi_{mn}' = 1$ for a fixed number of the highest values γ_{mn}'

Modification of EM Algorithm for Incomplete Data

"sparse" data vectors, e.g. $\bar{\boldsymbol{x}} = (x_1, -, x_3, x_4, -, -, x_7, \ldots, x_N), \quad x_n \in \mathcal{X}_n$

$\mathcal{N}(\bar{\boldsymbol{x}}) = \{n \in \mathcal{N} : \text{for which variable } x_n \text{ is defined in } \bar{\boldsymbol{x}}\}, \quad \bar{\boldsymbol{x}} \in \mathcal{S}$

$\mathcal{S}_n = \{\bar{\boldsymbol{x}} \in \mathcal{S} : n \in \mathcal{N}(\bar{\boldsymbol{x}})\}, \quad \approx$ vectors $\bar{\boldsymbol{x}} \in \mathcal{S}$ with defined variable x_n

m.-l. estimation from incomplete data:

$$L = \frac{1}{|\mathcal{S}|} \sum_{\bar{\boldsymbol{x}} \in \mathcal{S}} \log[\sum_{m=1}^{M} w_m F(\bar{\boldsymbol{x}}|m)], \qquad F(\bar{\boldsymbol{x}}|m) = \prod_{n \in \mathcal{N}(\bar{\boldsymbol{x}})} f_n(x_n|m)$$

iteration equations: $(m \in \mathcal{M}, n \in \mathcal{N}, \bar{\boldsymbol{x}} \in \mathcal{S})$

$$q(m|\bar{\boldsymbol{x}}) = \frac{w_m F(\bar{\boldsymbol{x}}|m)}{\sum_{j=1}^{M} w_j F(\bar{\boldsymbol{x}}|j)}, \qquad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{\bar{\boldsymbol{x}} \in \mathcal{S}} q(m|\bar{\boldsymbol{x}})$$

$$f_n^{'}(\xi|m) = \frac{1}{\sum_{\bar{\boldsymbol{x}} \in \mathcal{S}_n} q(m|\bar{\boldsymbol{x}})} \sum_{\bar{\boldsymbol{x}} \in \mathcal{S}_n} \delta(\xi, x_n) q(m|\bar{\boldsymbol{x}}), \quad \xi \in \mathcal{X}_n$$

Remark: Only available data are used without estimating missing values

Modification of EM Algorithm for Weighted Data

NOTATION: $\gamma(\boldsymbol{x})$: relative frequency of $\boldsymbol{x}$ in $\mathcal{S}$

$$\boldsymbol{x} \notin \mathcal{S} \Rightarrow \gamma(\boldsymbol{x}) = 0, \quad \Rightarrow \sum_{\boldsymbol{x} \in \mathcal{X}} \gamma(\boldsymbol{x}) = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} 1 = 1$$

$\Rightarrow$ **equivalent expression:**

$$L = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x} \in \mathcal{S}} \log\Big[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m)\Big] = \sum_{\boldsymbol{x} \in \mathcal{X}} \gamma(\boldsymbol{x}) \log\Big[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x}|m)\Big]$$

$\Rightarrow$ **weighted iteration equations:** $(m \in \mathcal{M},\ n \in \mathcal{N},\ \boldsymbol{x} \in \mathcal{X})$

$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|m)}{\sum_{j \in \mathcal{M}} w_j F(\boldsymbol{x}|j)}, \quad w'_m = \sum_{\boldsymbol{x} \in \mathcal{X}} \gamma(\boldsymbol{x}) q(m|\boldsymbol{x}),$$

$$f'_n(\xi|m) = \sum_{\boldsymbol{x} \in \mathcal{X}} \delta(\xi, x_n) \gamma(\boldsymbol{x}) \frac{q(m|\boldsymbol{x})}{w'_m}, \quad \xi \in \mathcal{X}_n, \quad m \in \mathcal{M}$$

Application: relevance of data in EM algorithm

Self-correcting Mechanisms

Self-correcting Mechanisms for Data

- unreliable, suspect or incorrect data can be removed or suppressed by weighting, e.g $\gamma(\boldsymbol{x}) = \log P(\boldsymbol{x})/\bar{L}, \quad \bar{L} = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} \log P(\boldsymbol{x})$
- sparsely used variable x_n can be removed, e.g. if $\sum_{m\in\mathcal{M}} \phi_{mn}$ is small
- a new variable x_n can be included at any level of design process by specifying the corresponding background distribution $f_n(x_n|0)$

Self-correcting Mechanisms for Knowledge Base

- initial parameters: expert design of elementary diagnoses (components)
- intuitively designed components will be modified by EM algorithm
- components having small "support" in the data can be identified by small weights w_m and removed
- a new component can be included at any level of design process

Concluding Remarks

Final output for a user:

printed recommendation for a physician including input symptoms, relevant diagnoses, suggested medication and comments

Concluding Remarks and Comments

- administration background (admin.): support of medical experts
- hierarchical lists of diagnoses and questions (admin.)
- sequential inclusion of new questions (open access and admin.)
- sequential inclusion of new components (open access and admin.)
- sequential inclusion of diagnostic areas (admin.)
- supervised accumulation of initial data (physicians, students)
- detailed comments to diagnoses (open access and admin.)
- supervised self-correcting mechanisms (admin.)
- suggested medication and treatment (admin.) $\Rightarrow$ financial support

References 1/3

- Ben-Bassat, M. (1978): Myopic policies in sequential classification, IEEE Trans. Computers, C-27, pp. 170-174.
- Ben-Bassat, M. (1980): Pattern-Based Interactive Diagnosis of Multiple Disorders: The MEDAS System, IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-2, 148-160.
- Fu, K.S. (1968): Sequential Methods in Pattern Recognition and Machine Learning, New York: Academic. *http:// dx.doi.org/10.1016/j.patrec.2014.02.024*
- Hussain, A.B.S., Donaldson, R.W. (1974): Suboptimal sequential decision schemes with on-line feature ordering, IEEE Trans. Computers, C-23, pp. 582-590.

References 2/3

- J. Grim, On numerical evaluation of maximum - likelihood estimates for finite mixtures of distributions, *Kybernetika*, Vol. l8, No.3, pp.173-190, 1982. *http://dml.cz/dmlcz/124132*
- Grim J.: Multivariate statistical pattern recognition with non-reduced dimensionality. Kybernetika, **22** 6, 142–157 (1986)
- Grim J. (1992): A dialog presentation of census results by means of the probabilistic expert system PES, in *Proceedings of the Eleventh Europeccn Meeting on Cybernetics and Systems Research*, (Ed. R.Trappl), Vienna, April 1992, World Scientific, Singapore 1992, pp. 997-1005.
- Grim J.: Knowledge representation and uncertainty processing in the probabilistic expert system PES, *International Journal of General Systems*, Vol. 22, No. 2, pp. 103 - 111, 1994.

References 3/3

- Grim J., Hora J., Somol P., Boček P., Pudil, P. (2010): Statistical Model of the 2001 Czech Census for Interactive Presentation. *Journal of Official Statistics*. Vol. 26, No. 4, pp. 673–694.
- Grim J. DEMO: Interactive Statistical Model of the 2001 Czech Census: *http://ro.utia.cas.cz/dem.html*
- Grim, J., Pudil, P., Somol, P.: Recognition of handwritten numerals by structural probabilistic neural networks. In: Proceedings of the Second ICSC Symposium on Neural Computation. (Bothe, H., Rojas, R. eds.). ICSC, Wetaskiwin (2000) 528-534
- Grim J., Somol P., Haindl M., Danes J. (2009): Computer-Aided Evaluation of Screening Mammograms Based on Local Texture Models. *IEEE Transactions on Image Processing 18(4): 765-773* Paper Award
- Grim, J. (2014). Sequential pattern recognition by maximum conditional informativity. *Pattern Recognition Letters*, Vol. 45C, pp. 39-45.
- Grim J. Product Mixtures I., Product Mixtures II., (Presentation), *http://ro.utia.cas.cz/dem.html*

ÚTIA