EM Cluster Analysis for Categorical Data

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

Conference S+SSPR 2006, Hong Kong 2006

Outline

Product Mixture: Conditional Independence Model

discrete random variables: $\xi_n \in \mathcal{X}_n, \quad n \in \mathcal{N}, \quad \mathcal{N} = \{1, \ldots, N\}$
$\mathcal{X}_n$: finite sets of categorical values (no ordering)

random vector: $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N) \in \mathcal{X}, \quad \mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$

discrete random (latent) variable: $\mu \in \mathcal{M}, \quad \mathcal{M} = \{1, \ldots, M\}$

$$P\{\mu = m\} = w_m, \quad m \in \mathcal{M}, \quad \sum_{m \in \mathcal{M}} w_m = 1$$

ASSUMPTION: variables ξ_n are conditionally independent given μ

$$P\{\boldsymbol{\xi} = \mathbf{x} \mid \mu = m\} = F(\mathbf{x}|m) = \prod_{n \in \mathcal{N}} f_n(x_n|m)$$

model of conditional independence (product mixture):

$$P(\mathbf{x}) = \sum_{m \in \mathcal{M}} w_m F(\mathbf{x}|m) = \sum_{m \in \mathcal{M}} w_m \prod_{n \in \mathcal{N}} f_n(x_n|m)$$

EM Algorithm For Discrete Product Mixtures

independent observations of the random vector $\boldsymbol{\xi}$:

$$\mathcal{S} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(J)}\}, \quad \mathbf{x}^{(j)} \in \mathcal{X}$$

log-likelihood function:

$$L = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} \log \left(\sum_{m \in \mathcal{M}} w_m \prod_{n \in \mathcal{N}} f_n(x_n|m) \right) \ \rightarrow \ max$$

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

$$q(m|\mathbf{x}) = \frac{w_m \prod_{n \in \mathcal{N}} f_n(x_n|m)}{\sum_{j \in \mathcal{M}} w_j \prod_{n \in \mathcal{N}} f_n(x_n|j)}, \quad w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x})$$

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

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

$\Rightarrow$ convergence to local/global maximum of the log-likelihood function

$\Rightarrow$ starting-point dependent estimates

Application of Product Mixtures to Cluster Analysis

CATEGORICAL DATA: the discrete space $\mathcal{X}$ has no structure in itself, conditional independence assumption is the only source of information about possible clusters ("latent classes"):

$$P(\mathbf{x}) = \sum_{m\in\mathcal{M}} w_m F(\mathbf{x}|m) = \sum_{m\in\mathcal{M}} w_m \prod_{n\in\mathcal{N}} f_n(x_n|m)$$

the values of "latent variable" $m \in \mathcal{M}$ correspond to "hidden causes" (remove statistical dependences between $\xi_1, \dots, \xi_N$)

$$q(m|\mathbf{x}) = \frac{w_m F(\mathbf{x}|m)}{\sum_{j\in\mathcal{M}} w_j F(\mathbf{x}|j)}, \qquad d(\mathbf{x}) = \arg\max_{m\in\mathcal{M}}\{q(m|\mathbf{x})\}$$

$q(m|\mathbf{x})$: **membership function of the m-th cluster**

$$\Re = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_M\}, \qquad \mathcal{S}_m = \{\mathbf{x} \in \mathcal{S} : d(\mathbf{x}) = m\}, \quad \mathcal{S} = \cup_{m\in\mathcal{M}} \mathcal{S}_m,$$

Remark. Clusters $\mathcal{S}_m$ are defined by the mixture components $w_m F(\mathbf{x}|m)$. If the mixture $P(\mathbf{x})$ is not defined uniquely, then the result of cluster analysis becomes questionable.

Discrete Product Mixtures Are Non-Identifiable

Definition of Identifiability

A class of distribution mixtures $\mathcal{F}$ is identifiable if the equality of any two mixtures $P, P^{'}$ from $\mathcal{F}$

$$P(\mathbf{x}) = \sum_{m \in \mathcal{M}} w_m F(\mathbf{x}|m) = \sum_{m \in \mathcal{M}'} w_m^{'} F^{'}(\mathbf{x}|m) = P^{'}(\mathbf{x}), \ \forall \ \mathbf{x} \in \mathcal{X}$$

implies that the parameters of the two mixtures $P, P^{'}$ are identical, except for order of components.

Lemma

Any discrete distribution mixture

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

can be equivalently described by infinitely many different parameter sets if at least one of the univariate conditional distributions $f_n(x_n|m)$ is non-degenerate in the sense that $f_n(x_n|m) < 1, \quad \text{for all} \quad x_n \in \mathcal{X}_n$.

Discrete Product Mixtures Are Non-Identifiable

Proof. If $f_n(x_n|m)$ is a non-degenerate distribution then we can write

$$f_n(\cdot|m) = \alpha f_n^{(\alpha)}(\cdot|m) + \beta f_n^{(\beta)}(\cdot|m), \quad f_n^{(\alpha)}(\cdot|m) \neq f_n^{(\beta)}(\cdot|m)$$

where $0 < \alpha < 1, \ \ \beta = 1 - \alpha$. By substitution we obtain

$$w_m F(\mathbf{x}|m) = w_m^{(\alpha)} F^{(\alpha)}(\mathbf{x}|m) + w_m^{(\beta)} F^{(\beta)}(\mathbf{x}|m), \quad \mathbf{x} \in \mathcal{X}$$

where $F^{(\alpha)}(\mathbf{x}|m), \ F^{(\beta)}(\mathbf{x}|m)$ are different components:

$$w_m^{(\alpha)} = \alpha w_m, \quad F^{(\alpha)}(\mathbf{x}|m) = f_n^{(\alpha)}(x_n|m) \prod_{i \in \mathcal{N}, i \neq n} f_i(x_i|m),$$

$$w_m^{(\beta)} = \beta w_m, \quad F^{(\beta)}(\mathbf{x}|m) = f_n^{(\beta)}(x_n|m) \prod_{i \in \mathcal{N}, i \neq n} f_i(x_i|m)$$

$\Rightarrow$ **The original mixture is described equivalently by non-trivially different parameters.**

Unique Mixture Parameters by Additional Constraints

EM Algorithm & Sequential Adding of Components

- starting with one component: $M = 1$, uniform distributions $f_n(x_n|1)$
- adding new component after sufficient convergence ($\frac{L'-L}{L} < \epsilon$): $M \to M+1$, uniform distributions $f_n(x_n|M+1)$, $w_{M+1} = 0.5$
- repeat adding of components until the new weight is "suppressed"

Properties:

- $\oplus$ the method avoids random influences of initial values
- $\oplus$ the resulting mixture is defined (almost) uniquely
- $\oplus$ newly added component fits to currently "outlying" data
- $\oplus$ reasonable choice of a proper number of components
- $\ominus$ the method is based on heuristical idea, no theoretical arguments
- $\ominus$ adding new components disturbs preceding convergence phase

Artificial Problem: Re-Identification of Bernoulli Mixture

mixture of multivariate Bernoulli distributions

$$P^*(\mathbf{x}) = \sum_{m\in\mathcal{M}} w_m \prod_{n\in\mathcal{N}} \theta_{nm}^{x_n}(1-\theta_{nm})^{1-x_n}, \quad \mathbf{x}\in\{0,1\}^N, \quad 0<\theta_{nm}<1$$

SOLUTION: re-estimation of parameters w_m, θ_{nm} by using weighted modification of EM algorithm:

$$L^* = \lim_{|\mathcal{S}|\to\infty} \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \log\left[\sum_{m\in\mathcal{M}} w_m F(\mathbf{x}|m)\right] = \sum_{x\in\mathcal{X}} P^*(\mathbf{x}) \log\left[\sum_{m\in\mathcal{M}} w_m F(\mathbf{x}|m)\right]$$

modified EM iteration equations: $(m\in\mathcal{M},\ \mathbf{x}\in\mathcal{X})$

$$q(m|\mathbf{x}) = \frac{w_m F(\mathbf{x}|m)}{\sum_{j\in\mathcal{M}} w_j F(\mathbf{x}|j)}, \quad w_m^{'} = \sum_{\mathbf{x}\in\mathcal{X}} P^*(\mathbf{x}) q(m|\mathbf{x})$$

$$\theta_{nm}^{'} = \frac{1}{\sum_{\mathbf{x}\in\mathcal{X}} P^*(\mathbf{x}) q(m|\mathbf{x})} \sum_{\mathbf{x}\in\mathcal{X}} x_n P^*(\mathbf{x}) q(m|\mathbf{x}), \quad \xi\in\mathcal{X}_n$$

Remark. Computation is equivalent to infinite data set $\mathcal{S}$ (avoids random small-sample fluctuations).

Comparison of the Original and Re-Estimated Parameters

original parameters: $M = 8, N = 16$, Carreira-Perpignan et.al. (2000)
(the weights $P^*(\mathbf{x})$ computed for all the 65536 binary vectors $\mathbf{x} \in \mathcal{X}$)

comparison of original and re-estimated parameters (upper$\times$lower row):

w_m	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{16}
$.222\dot{2}$	.80	.80	.80	.80	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20
.2220	.80	.80	.80	.80	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20
$.194\dot{4}$	.20	.20	.20	.20	.80	.80	.80	.80	.20	.20	.20	.20	.20	.20	.20	.20
.1943	.20	.20	.20	.20	.80	.80	.80	.80	.20	.20	.20	.20	.20	.20	.20	.20
$.166\dot{6}$	.20	.20	.20	.20	.20	.20	.20	.20	.80	.80	.80	.80	.20	.20	.20	.20
.1666	.20	.20	.20	.20	.20	.20	.20	.20	.80	.80	.80	.80	.20	.20	.20	.20
$.138\dot{8}$	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.80	.80	.80	.80
.1388	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.20	.80	.80	.80	.80
$.111\dot{1}$	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20
.1109	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20
$.083\dot{3}$	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20
.0832	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20
$.055\dot{5}$	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20
.0555	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20
$.027\dot{7}$	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80
.0277	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80	.20	.20	.20	.80
.0008	.44	.39	.37	.35	.39	.34	.32	.31	.37	.32	.30	.29	.35	.31	.29	.28

Remark. EM algorithm has been stopped after adding 9-th component. The weight w_9 of the last added component is by two orders less than w_8.

Concluding Remarks

Conditional Independence Model as a Tool of Cluster Analysis

- **goal:** identification of unknown mixture parameters
- applicable to multivariate categorical data
- drawback: discrete product mixtures are non-identifiable
- unique solution: additional constraints
  (sequential adding of components)

Application of Conditional Independence Model for Approximation

- **goal:** approximation of unknown probability distribution
- statistical pattern recognition
- statistical modelling of large databases
- texture modelling and evaluation
- non-identifiability is useful (increased flexibility)

Literatura 1/2

- Carreira-Perpignan M.A., Renals S. (2000): Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, Vol. 12, pp. 141-152
- Dempster A.P., Laird N.M. and Rubin D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc., B*, Vol. 39, pp. 1-38
- Grim J. (1982): On numerical evaluation of maximum - likelihood estimates for finite mixtures of distributions. *Kybernetika*, Vol.18, No.3, pp. 173-190
- Gyllenberg M., Koski T., Reilink E., Verlaan M. (1994): Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, Vol. 31, pp. 542-548
- Lazarsfeld P.F., Henry N. (1968): *Latent structure analysis*. Houghton Mifflin: Boston

Literatura 2/2

- McLachlan G.J. and Peel D. (2000): *Finite Mixture Models*, John Wiley & Sons, New York, Toronto: 2000
- Schlesinger, M.I. (1968): Relation between learning and self-learning in pattern recognition." (in Russian), *Kibernetika*, (Kiev), No. 2, pp. 81-88
- Teicher, H. (1968): Identifiability of mixtures of product measures. *Ann. Math. Statist.*, Vol. 39, pp. 1300-1302
- Titterington, D.M., Smith, A.F.M., & Makov, U.E. (1985): *Statistical analysis of finite mixture distributions.* John Wiley & Sons, New York: 1985
- Vermunt J.K., Magidson J. (2002): Latent Class Cluster Analysis. In: *Advances in Latent Class Analysis*, (Eds. Hagenaars J.A. et al.), Cambridge University Press

ÚTIA