Computational Properties of Probabilistic Neural Networks

Jiří Grim and Jan Hora

Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic, Prague

Department of Pattern Recognition
http://www.utia.cas.cz/RO

ICANN'10, Thessaloniki, September 15-18, 2010

Outline

Overfitting in Neural Networks

problem of overfitting

- small multidimensional training data sets ($\approx$ insufficiently representative)
- $\Rightarrow$ "overfitting" of parameters to training data
- $\Rightarrow$ bad "generalizing property" caused by overfitting
- a general analysis of overfitting is difficult

to reduce the risk of overfitting:

- dimensionality reduction and/or large data sets
- cross-validation techniques
- "under-computing": stopping rule for training
- optimal complexity of classifiers

Probabilistic Neural Networks:

structural mixtures $\Rightarrow$ reduced complexity $\Rightarrow$ less prone to overfitting

Statistical Pattern Recognition Based on Mixtures

$\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}$:	N-dimensional data vectors
$\Omega = \{\omega_1, \omega_2, \ldots, \omega_J\}$:	finite number of classes
$P(\mathbf{x}\|\omega)p(\omega), \quad \omega \in \Omega$:	conditional distributions of classes

approximation of $P(\mathbf{x}|\omega)$ by mixtures of product components:

$$P(\mathbf{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} f(m) F(\mathbf{x}|m), \quad P(\mathbf{x}) = \sum_{\omega \in \Omega} p(\omega) P(\mathbf{x}|\omega)$$

$$F(\mathbf{x}|m) = \prod_{n \in \mathcal{N}} \theta_{mn}^{x_n} (1 - \theta_{mn})^{1 - x_n}, \quad 0 \le \theta_{mn} \le 1, \quad m \in \mathcal{M}_\omega, \quad \mathcal{M} = \sum_{\omega \in \Omega} \mathcal{M}_\omega$$

decision making based on Bayes formula:

$$p(\omega|\mathbf{x}) = \frac{P(\mathbf{x}|\omega)p(\omega)}{P(\mathbf{x})} = \sum_{m \in \mathcal{M}_\omega} q(m|\mathbf{x}), \qquad q(m|\mathbf{x}) = \frac{w_m F(\mathbf{x}|m)}{\sum_{j \in \mathcal{M}} w_j F(\mathbf{x}|j)}$$

probabilistic neuron $\approx$ mixture component

Structural Mixture Model (Grim et al. 1986, 1999, 2002)

$$F(\mathbf{x}|m) = \prod_{n\in\mathcal{N}} f_n(x_n|m)^{\phi_{mn}} f_n(x_n|0)^{1-\phi_{mn}}, \qquad \phi_{mn} \in \{0,1\} \approx \text{structure}$$

$$f_n(x_n|m) = \theta_{mn}^{x_n}(1-\theta_{mn})^{1-x_n}, \quad n \in \mathcal{N}, \quad \mathcal{N} = \{1,\ldots,N\}$$

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

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

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

$G(\mathbf{x}|m,\phi_m) \quad \approx$ **defined on different subspaces** OPTIMIZATION: EM algorithm

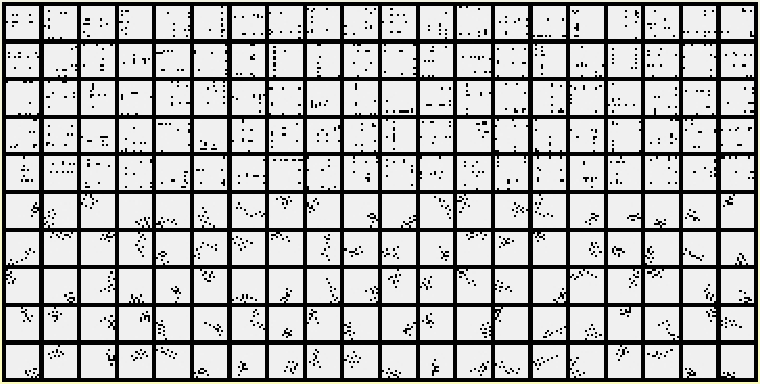
$$p(\omega|\mathbf{x}) = \frac{P(\mathbf{x}|\omega)p(\omega)}{P(\mathbf{x})} = \frac{\sum_{m\in\mathcal{M}_\omega} w_m G(\mathbf{x}|m,\phi_m)}{\sum_{j\in\mathcal{M}} w_j G(\mathbf{x}|j,\phi_j)}$$

$\Rightarrow$ **"background distribution"** $F(\mathbf{x}|0)$ **cancels in the Bayes formula**

Chess-Board Patterns Made by Rook and Knight

16x16 chess-board patterns: 256-dimensional binary vectors

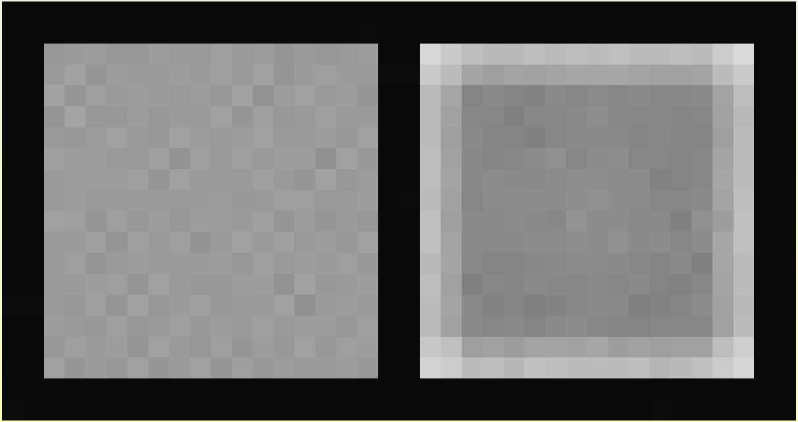
generating patterns: random moves until 10 different positions
chess-piece position coded by $x_n = 1$



training and testing data: 100000 binary vectors for each class

Marginal Probabilities of the Chess-Board Patterns

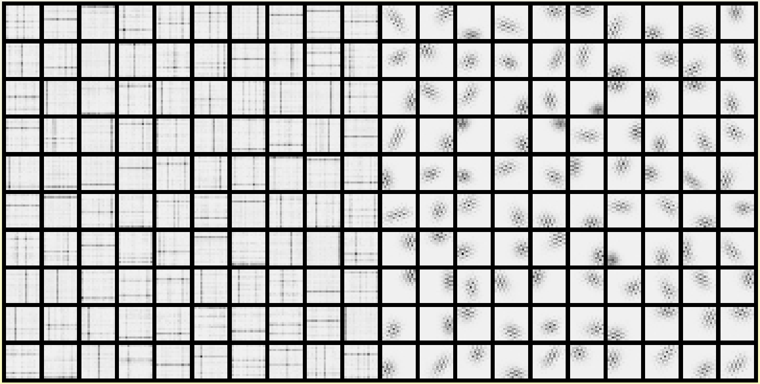
class-means of training numerals ("mean images")



left: rook-made patterns **right:** knight-made patterns

Component Parameters of the Estimated Mixtures $P(\mathbf{x}|\omega)$

component parameters θ_{mn} in chess-board arrangement



left: rook-made patterns **right:** knight-made patterns

Recognition of Chess-board patterns: full model

Recognition error of the full multivariate Bernoulli mixture model in %

M	**1 000**	200 000	**10 000**	200 000	**100 000**	200 000
2	**34.70**	41.56	**39.59**	40.38	**39.90**	40.02
4	**13.10**	15.83	**16.54**	16.65	**16.42**	16.48
10	**1.65**	7.72	**6.60**	7.00	**6.49**	6.60
20	**0.95**	9.21	**5.40**	5.90	**4.04**	4.34
40	**0.15**	8.76	**3.91**	4.90	**2.73**	2.89
100	**0.00**	9.35	**2.01**	4.54	**1.37**	1.90
200	**0.00**	11.02	**1.22**	5.57	**0.84**	1.68
400	**0.00**	15.40	**0.69**	8.35	**0.45**	1.92
1000	**0.00**	17.77	**0.20**	14.66	**0.14**	3.76

Recognition of Chess-board patterns: subspace model

Recognition error of the structural Bernoulli mixture model in %

M	**1 000**	200 000	**10 000**	200 000	**100 000**	200 000
4	**13.80**	16.48	**25.53**	26.53	**16.91**	16.92
8	**6.45**	9.97	**10.96**	11.32	**7.28**	7.29
20	**5.70**	10.97	**5.14**	5.77	**4.70**	4.72
40	**8.65**	13.63	**4.20**	4.73	**3.29**	3.32
80	**4.20**	12.91	**6.12**	6.88	**1.91**	1.92
200	**0.25**	11.46	**3.36**	4.76	**1.83**	1.85
400	**0.00**	18.11	**3.54**	4.70	**3.10**	3.20
800	**0.00**	18.50	**3.88**	4.82	**5.42**	5.45
2000	**0.00**	18.75	**2.84**	6.39	**2.71**	2.73

Concluding Remarks

Overfitting of Structural Mixture Model

- probabilistic

Structural Modification of EM Algorithm

STRUCTURAL OPTIMIZATION: can be included into EM algorithm

$$L = \frac{1}{|\mathcal{S}_\omega|} \sum_{x \in \mathcal{S}_\omega} \log \Big[\sum_{m \in \mathcal{M}_\omega} F(\mathbf{x}|0) G(\mathbf{x}|m, \phi_m) w_m \Big], \qquad F(\mathbf{x}|0) = \prod_{n \in \mathcal{N}} f_n(x_n|0)$$

EM Algorithm: $(m \in \mathcal{M}_\omega, n \in \mathcal{N}, \mathbf{x} \in \mathcal{S}_\omega)$

$$q(m|\mathbf{x}) = q(m|\mathbf{x}, \omega) = \frac{G(\mathbf{x}|m, \phi_m) w_m}{\sum_{j \in \mathcal{M}_\omega} G(\mathbf{x}|j, \phi_j) w_j},$$

$$w_m^{'} = \frac{1}{|\mathcal{S}_\omega|} \sum_{\mathbf{x} \in \mathcal{S}_\omega} q(m|\mathbf{x}), \qquad \theta_{mn}^{'} = \frac{1}{\sum_{\mathbf{x} \in \mathcal{S}_\omega} q(m|\mathbf{x})} \sum_{\mathbf{x} \in \mathcal{S}_\omega} x_n q(m|\mathbf{x})$$

structural criterion: (Kullback-Leibler $\mathbb{I}$-divergence)

$$\gamma_{mn}^{'} = w_m^{'} \left[\theta_{mn}^{'} \log \frac{\theta_{mn}^{'}}{\theta_{0n}} + (1 - \theta_{mn}^{'}) \log \frac{(1 - \theta_{mn}^{'})}{(1 - \theta_{0n})} \right]$$

structural parameter optimization: $\phi_{mn}^{'} = 1$ for the r highest values $\gamma_{mn}^{'}$

Remark. The "structural" EM algorithm converges monotonically.

Back

References 1/4

- Dempster, A.P., Laird, N.M., & Rubin, D.B., (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. of the Royal Stat. Soc.*, **B 39**, pp. 1-38.
- Dietterich, T.: Overfitting and undercomputing in machine learning. *ACM Computing Surveys*, 3 **27** (1995) 326-327
- Grim J.: Multivariate statistical pattern recognition with nonreduced dimensionality. *Kybernetika*, **22** (1986) 142-157.
- Grim J.: Information approach to structural optimization of probabilistic neural networks. In proceedings of: *4th System Science European Congress*, Ferrer, L. et al. (Eds.), (pp. 527-540), Valencia: Soc. Espanola de Sistemas Generales, (1999)
- Grim, J.: Neuromorphic features of probabilistic neural networks. *Kybernetika.*, 5 **43** (2007) 697–712

References 2/4

- Grim, J., Pudil, P., Somol, P.: Recognition of handwritten numerals by structural probabilistic neural networks. In: *Proc. Second ICSC Symposium on Neural Computation.* (Bothe, H., Rojas, R. eds.). ICSC, Wetaskiwin (2000) 528-534
- Grim, J., Just, P., Pudil, P.: Strictly modular probabilistic neural networks for pattern recognition. Neural Network World, **13** (2003) 599-615
- Grim J., Hora, J.: "Iterative principles of recognition in probabilistic neural networks." *Neural Networks, Special Issue, 6* **21** *(2008) 838–846*
- *McLachlan, G.J., Peel, D.: Finite Mixture Models, John Wiley and Sons, New York, Toronto (2000)*
- *Vajda, I., (1992): Theory of Statistical Inference and Information. Boston: Kluwer.*

References 3/4

- Sarle, W.S.: Stopped training and other remedies for overfitting. In: Proceedings of the 27th Symposium on the Interface. (1995) Available via ftp://ftp.sas.com/pub/neural/inter95.ps.Z.
- Schaffer C.: Overfitting avoidance as Bias. *Machine Learning*, 2 **10** (1993) 153–178.
- Schlesinger, M.I.: Relation between learning and self-learning in pattern recognition. (in Russian), *Kibernetika*, (Kiev), No. 2 (1968) 81-88.
- Specht, D.F.: Probabilistic neural networks for classification, mapping or associative memory. In: *Proc. IEEE Int. Conf. on Neural Networks*, **I** (1988) 525–532
- Yinyin L., Starzyk J.A., Zhen Zhu: Optimized Approximation Algorithm in Neural Networks Without Overfitting. *IEEE Tran. Neural Networks*, **19** (2008) 983–995