Extraction of Binary Features by Probabilistic Neural Networks

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 ICANN'08, Prague, September 3-6, 2008

Outline

Feature Extraction for Classification

statistical classification methods:

- **purpose of feature extraction:** to reduce dimensionality in order to simplify decision making
- **goal:** small number of highly informative features

biological neural networks:

- **output of neuron** $\approx$ feature extracted from input layer neurons
- **purpose:** to extract simple features rather than to reduce dimensionality
- **goal:** large number of output neurons (features) which respond to highly specific input patterns
- $\Rightarrow$ complex input signals are coded by labels of output neurons
- $\Rightarrow$ no decision making is necessary

Statistical Pattern Recognition Based on Mixtures

$\mathbf{x} = (x_1, \dots, x_N) \in \mathcal{X}$: N-dimensional data vectors

$\Omega = \{\omega_1, \omega_2, \dots, \omega_J\}$: 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:

$$P(\mathbf{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} f(m) F(\mathbf{x}|m), \quad \sum_{m \in \mathcal{M}_\omega} f(m) = 1, \quad \mathcal{M} = \bigcup_{\omega \in \Omega} \mathcal{M}_\omega.$$

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

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)

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

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

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

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

structural model can be optimized in full generality by ▸ EM algorithm

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

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

$G(\mathbf{x}|m,\phi_m) \approx$ **may depend on different subsets of variables**

Properties of Features in Probabilistic Neural Networks

one output neuron for each class $\omega \in \Omega$:

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

(statistically correct subspace approach to Bayesian decision making)

hidden layer neuron:

$$y_m(\mathbf{x}) = \mathrm{T}_m(\mathbf{x}) = \log\left[q(m|\mathbf{x})\right] = \log\left[\frac{G(\mathbf{x}|m, \phi_m) w_m}{\sum_{j \in \mathcal{M}} G(\mathbf{x}|j, \phi_j) w_j}\right]$$

nearly binary properties of $q(m|\mathbf{x})$:

$$q_{max}(\mathbf{x}) = \max_{m \in \mathcal{M}} \{q(m|\mathbf{x})\}, \qquad \bar{q}_{max} = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q_{max}(\mathbf{x}) \quad \rightarrow \; 1$$

Remark: for $N \approx 10^2 \; : \; \bar{q}_{max} \; \approx \; 0.99$

Information Preserving Property

Theorem (Grim et al. 1996, 1998):

The mixture based transform $\mathbf{T} : \mathcal{X} \to \mathcal{Y},\ \mathcal{Y} \subset R^M$ defined by

$$y_m = \mathbf{T}^*_m(\mathbf{x}) = \log(q^*(m|\mathbf{x})), \quad \mathbf{x} \in \mathcal{X}, \quad m \in \mathcal{M}$$

preserves the statistical decision information by Shannon

$$I(\mathcal{X}, \mathcal{M}) = I(\mathcal{Y}, \mathcal{M})$$

given the true conditional probabilities $q^*(m|\mathbf{x})$.
Simultaneously the entropy of the transformed distribution is minimized:

$$H(\mathcal{Y}) = \sum_{\mathbf{y} \in \mathcal{Y}} -Q(\mathbf{y}) \log Q(\mathbf{y}) \ \to \ \min, \qquad Q(\mathbf{y}) = P(\mathbf{T}^{-1}(\mathbf{y}))$$

Idea of the proof: The transform **T** "unifies" the points $\mathbf{x} \in \mathcal{X}$ with the same posterior distributions $q^*(.|\mathbf{x})$ $\Rightarrow$ no information loss.

Extraction of Binary Features by PNN

regularized binary features:

$$y_m = \mathbf{T}_m(\mathbf{x}) = \log[q(m|\mathbf{x}) + \delta w_m], \quad \delta > 0, \quad \mathbf{x} \in \mathcal{X}, \quad m \in \mathcal{M}$$

Theorem (Grim 1998):

If the transform $\mathbf{T}$ satisfies for some $\delta, \epsilon > 0$ the inequality

$$|\mathbf{T}_m(\mathbf{x}) - \ln[q^*(m|\mathbf{x}) + w_m^*\delta]| < \epsilon, \quad \epsilon > 0, \quad m \in \mathcal{M}, \quad \mathbf{x} \in \mathcal{X}$$

then the arising information loss is bounded by the inequality

$$I(\mathcal{X}, \mathcal{M}) - I(\mathcal{Y}, \mathcal{M}) < \delta + 2\epsilon$$

binary features obtained by simple thresholding:

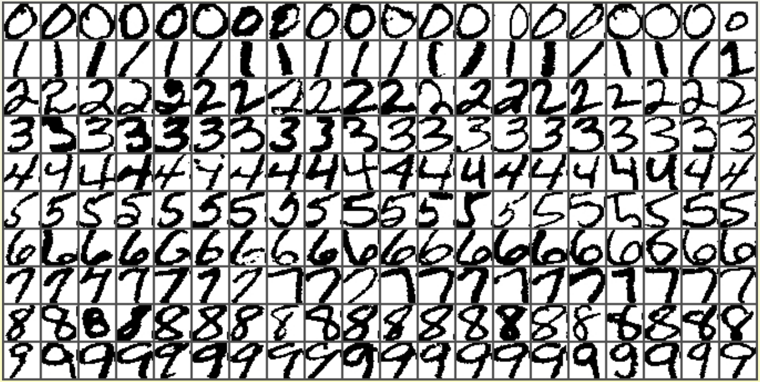
$$y_m = \mathbf{T}_m(\mathbf{x}) = \begin{cases} 1, & q(m|\mathbf{x}) \geq \theta, \\ 0, & q(m|\mathbf{x}) < \theta, \end{cases} \qquad 0 < \theta \ll 1$$

$$y_m = \mathbf{T}_m(\mathbf{x}) = \begin{cases} 1, & \log[G(\mathbf{x}|m, \phi_m) w_m] \geq \log\theta + \log P(\mathbf{x}) \\ 0, & \log[G(\mathbf{x}|m, \phi_m) w_m] < \log\theta + \log P(\mathbf{x}) \end{cases}$$

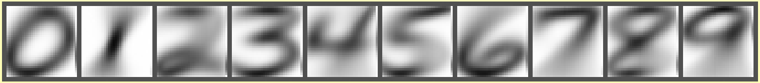
NIST SD19 Database of Hand-Written Numerals

NIST Special Database SD19: about 400000 **handwritten numerals**

examples of numerals normalized to 32x32 binary raster



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



Numerical Experiment: Recognition of the NIST Numerals

data split: odd data vectors for training, even data vectors for testing (200000 **training numerals and** 200000 **testing numerals)**

- all numerals normalized to 32x32 binary raster
- three differently rotated variants of each digit pattern included
- initial number of components chosen identically in all classes
- randomly initialized mixture parameters
- stopping rule: relative increment threshold

goals of the computational experiments:

- to compare recognition accuracy in the input space and feature space
- to test the influence of model complexity
- to illustrate the decrease of entropy in the feature space:

$$H(\mathcal{Y}) \approx \lim_{|\mathcal{S}|\to\infty} \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} -\log Q(\mathbf{y}(\mathbf{x})) \approx \sum_{y\in\mathcal{Y}} -\tilde{Q}(\mathbf{y}) \log Q(\mathbf{y})$$

Recognition of Numerals From the NIST SD19 Database

Classification of numerals from the NIST SD19 database by differently complex mixtures. ▸ Component Means
Comparison of the accuracy in the input space and in the feature space.

Experiment No.:	**1**	**2**	**3**	**4**	**5**
Number of Components	100	357	695	1119	1382
Number of Parameters	96046	243293	533628	574159	1027691
No. of Parameters in %	93.8	66.5	75.0	50.1	72.6
Mean No. of Units in $\mathbf{y}$	1.22	1.32	1.44	1.39	1.50
Log-Likelihood for $P(\mathbf{x})$	-295.8	-265.8	-242.0	-239.8	-235.3
Log-Likelihood for $P(\mathbf{y})$	-6.21	-7.95	-9.19	-9.48	-10.09
Recognition Accuracy					
Error in % (Input space)	5.46	3.24	2.52	2.21	2.12
Error in % (Feature space)	5.21	3.17	2.46	2.10	2.08

Concluding Remarks

Properties of Probabilistic Binary Features

- probabilistic features simplify decision making by reducing the feature complexity rather than dimensionality of the problem
- in the input space $\mathcal{X}$ the recognition accuracy increases with the model complexity (number of components)
- the recognition accuracy based on the proposed binary features is slightly better in all experiments
- in the binary feature space $\mathcal{Y}$ the recognition accuracy does not increase with the model complexity, only one component has been used to estimate the class-conditional distributions $Q(\mathbf{y}|\omega)$
- $\Rightarrow$ the resulting binary features appear to be almost conditionally independent with respect to classes
- in the feature space $\mathcal{Y}$ the entropy of the transformed distribution is much less than in the input space $\mathcal{X}$

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)

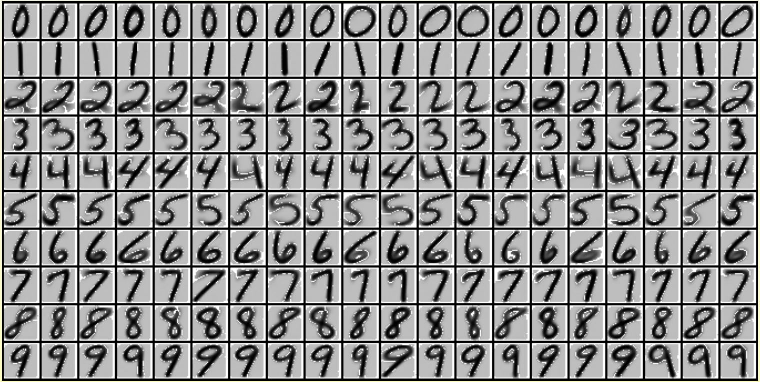
$$\gamma_{mn}^{'} = w_m^{'} \left[\theta_{mn}^{'} \log \frac{\theta_{mn}^{'}}{\theta_{0n}} + (1 - \theta_{mn}^{'}) \log \frac{(1 - \theta_{mn}^{'})}{(1 - \theta_{0n})} \right]$$

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

Remark. The "structural" EM algorithm converges monotonically.

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

component parameters $\theta_{mn} \in \langle 0, 1 \rangle$ displayed as grey levels in raster arrangement (the white fields denote unused variables with $\phi_{mn} = 0$)



Back

References 1/4

- Grim J.: Multivariate statistical pattern recognition with non-reduced dimensionality. Kybernetika, **22** 6, 142–157 (1986)
- Grim J.: Maximum-likelihood design of layered neural networks. In: International Conference on Pattern Recognition, (Proceedings), pp. 85–89, IEEE Computer Society Press, Los Alamitos (1996)
- Grim J.: Design of multilayer neural networks by information preserving transforms. In: Third European Congress on Systems Science, pp. 977–982, Pessa E., Penna M. P., Montesanto A. eds., Edizioni Kappa, Roma (1996)
- Grim J.: Discretization of probabilistic neural networks with bounded information loss. In: Computer-Intensive Methods in Control and Data Processing, pp. 205–210, J. Rojicek et al., eds., UTIA Prague (1998)

References 2/4

- Grim, J., Kittler, J., Pudil, P., Somol, P.: Multiple classifier fusion in probabilistic neural networks. Pattern Analysis & Applications **5** (2002) 221-233
- 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.: Self-organizing maps and probabilistic neural networks. Neural Network World, 10, 3, 407–415 (2000)
- Vajda, I., Grim, J.: About the maximum information and maximum likelihood principles in neural networks. Kybernetika, **34** (1998) 485-494

References 3/4

- Grim J., Somol P., Novovičová J., Pudil P., Ferri F., (1998b): Initializing normal mixture of densities. In *Proc. 14th Int. Conf. ICPR'98*, A.K. Jain et al. (Eds.), pp. 886-890, IEEE Computer Society: Los Alamitos, California, 1998
- Grim J.: Information approach to structural optimization of probabilistic neural networks. In: Fourth European Congress on Systems Science, pp. 527–539, Ferrer L., Caselles A. eds., SESGE, Valencia (1999)
- Grim J., Hora J.: Recurrent Bayesian Reasoning in Probabilistic Neural Networks. *Artificial Neural Networks – ICANN 2007*, Ed. Marques de Sá et al., LNCS 4669, pp. 129–138, Berlin: Springer (2007)
- Haykin, S.: Neural Networks: a comprehensive foundation, Morgan Kaufman: San Mateo CA (1993)

References 4/4

- Kohonen, T. (1997). The Self-Organizing Maps. New York, Berlin: Springer Verlag, (1997)
- McLachlan, G.J., Peel, D.: Finite Mixture Models, John Wiley and Sons, New York, Toronto (2000)
- Specht, D.F.: Probabilistic neural networks for classification, mapping or associative memory. In: Proc. of the IEEE Itternational Conference on Neural Networks, I, 525–532 (1988)
- Streit, L.R., Luginbuhl, T.E.: Maximum-likelihood training of probabilistic neural networks, IEEE Trans. on Neural Networks 5, 764–783 (1994)
- Watanabe S. and Fukumizu K.: Probabilistic design of layered neural networks based on their unified framework. *IEEE Trans. on Neural Networks*, 6, 3, 691–702 (1995)