Recognition of Properties by 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'09, Limassol, September 14-17, 2009

Outline

Bayes Decision Scheme

$\mathbf{x} = (x_1, \ldots, x_N) \in \mathcal{X}$: N-dimensional binary data vectors, $\mathcal{N} = \{1, \ldots, N\}$

$\Omega = \{\omega_1, \omega_2, \ldots, \omega_J\}$: finite number of classes, probability $p(\omega), \omega \in \Omega$

$P(\mathbf{x}|\omega)p(\omega), \quad \omega \in \Omega$: conditional distributions of classes

full statistical decision information: *a posteriori* probabilities

$$p(\omega|\mathbf{x}) = \frac{P(\mathbf{x}|\omega)p(\omega)}{P(\mathbf{x})}, \quad \mathbf{x} \in \mathcal{X}, \quad \omega \in \Omega,$$

Bayes decision function: minimum-error classification

$$d : \mathcal{X} \to \Omega, \qquad d(\mathbf{x}) = \arg\max_{\omega \in \Omega}\{p(\omega|\mathbf{x})\} = \arg\max_{\omega \in \Omega}\{P(\mathbf{x}|\omega)p(\omega)\}, \qquad \mathbf{x} \in \mathcal{X}$$

formula of complete probability implies mutually exclusive classes

$$P(\mathbf{x}) = \sum_{\omega \in \Omega} P(\mathbf{x}|\omega)p(\omega), \quad \mathbf{x} \in \mathcal{X}$$

real life categories are usually non-exclusive

$\Rightarrow$ **we propose: identification of non-exclusive properties by using statistical one-class classifiers**

Binary Recognition of Non-exclusive Properties

finite set of non-exclusive properties: $\Omega = \{\omega_1, \ldots, \omega_K\}$

two alternatives for each property $\omega \in \Omega$:

ω: the property has been identified (positive decision)
$\bar{\omega}$: the property has not been identified (negative decision)

$\Rightarrow$ **finite number of binary classification problems:** $\{\omega, \bar{\omega}\}, \quad (\omega \in \Omega)$

two training data sets: $\mathcal{S}_\omega = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K_\omega)}\}, \ \mathcal{S}_{\bar{\omega}} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K_{\bar{\omega}})}\}$

probabilistic description: $P(\mathbf{x}|\omega)p(\omega), \quad P(\mathbf{x}|\bar{\omega})p(\bar{\omega})$

$\Rightarrow$ **Bayesian decision-making:**

$$p(\omega|\mathbf{x}) = \frac{P(\mathbf{x}|\omega)p(\omega)}{P(\mathbf{x})}, \qquad p(\bar{\omega}|\mathbf{x}) = \frac{P(\mathbf{x}|\bar{\omega})p(\bar{\omega})}{P(\mathbf{x})}$$

$$d_\omega : \mathcal{X} \to \{\omega, \bar{\omega}\}, \quad d_\omega(\mathbf{x}) = \begin{cases} \omega, & p(\omega|\mathbf{x}) \geq p(\bar{\omega}|\mathbf{x}), \\ \bar{\omega}, & p(\omega|\mathbf{x}) < p(\bar{\omega}|\mathbf{x}), \end{cases} \qquad \mathbf{x} \in \mathcal{X}$$

PROBLEM: "negative" training data sets $\mathcal{S}_{\bar{\omega}}$ are rarely available

ÚTIA

Identification of Properties by One-Class-Classifiers

one-class classifier: only one training set for each property $\omega \in \Omega$:
$\mathcal{S}_\omega = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(K_\omega)}\}$, (*a priori* probability $p(\omega)$ unknown)

$\Rightarrow$ class-conditional distribution $P(\mathbf{x}|\omega)$ for the "target" class only
$\Rightarrow$ Bayes rule not applicable
$\Rightarrow$ **classification of properties by thresholding** $P(\mathbf{x}|\omega)$

we propose: thresholding of the log-likelihood ratio:

$$\Delta_\omega(\mathbf{x}) = \log \frac{P(\mathbf{x}|\omega)}{\prod_{n\in\mathcal{N}} f_n(x_n|0)} \geq \epsilon_\omega, \quad \mathbf{x} \in \mathcal{X}$$

$f_n(x_n|0) = \theta_{0n}^{x_n}(1-\theta_{0n})^{1-x_n}, \; n \in \mathcal{N} \;\approx$ unconditional marginal probabilities

the threshold ϵ_ω can be related to the mean value of $\Delta_\omega(\mathbf{x})$:

$$\epsilon_\omega = \frac{1}{|\mathcal{S}_\omega|} \sum_{x\in\mathcal{S}_\omega} \log \frac{P(\mathbf{x}|\omega)}{\prod_{n\in\mathcal{N}} f_n(x_n|0)} \;\rightarrow\; \sum_{x\in\mathcal{X}} P^*(\mathbf{x}|\omega) \log \frac{P^*(\mathbf{x}|\omega)}{\prod_{n\in\mathcal{N}} f_n(x_n|0)}$$

ϵ_ω **converges to the Kullback-Leibler I-divergence**

ÚTIA

Structural Mixture Model

$\phi_{mn} \in \{0,1\}$ $\approx$ **binary structural parameters** (Grim et al. 1999, 2002)

$$P(\mathbf{x}|\omega) = \sum_{m \in \mathcal{M}_\omega} F(\mathbf{x}|m) f(m) = \sum_{m \in \mathcal{M}_\omega} f(m) \prod_{n \in \mathcal{N}} f_n(x_n|m)^{\phi_{mn}} f_n(x_n|0)^{1-\phi_{mn}}$$

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

$\mathcal{M}_\omega \approx$ component index set of the property $\omega \in \Omega$

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

univariate distributions: $f_n(x_n|m) = \theta_{mn}^{x_n}(1-\theta_{mn})^{1-x_n}, \quad \theta_{mn} \in \langle 0,1 \rangle$

$$G(\mathbf{x}|m, \phi_m) = \prod_{n \in \mathcal{N}} \left[\frac{f_n(x_n|m)}{f_n(x_n|0)} \right]^{\phi_{mn}} = \prod_{n \in \mathcal{N}} \left[\left(\frac{\theta_{mn}}{\theta_{0n}} \right)^{x_n} \left(\frac{1-\theta_{mn}}{1-\theta_{0n}} \right)^{1-x_n} \right]^{\phi_{mn}}$$

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

structural model can be optimized in full generality ▸ EM algorithm

Identification of Non-exclusive Properties by PNN

mixture-based one-class classifier:

$$\Delta_\omega(\mathbf{x}) = \log \frac{P(\mathbf{x}|\omega)}{\prod_{n\in\mathcal{N}} f_n(x_n|0)} = \log \frac{P(\mathbf{x}|\omega)}{F(\mathbf{x}|0)} = \log \sum_{m\in\mathcal{M}_\omega} G(\mathbf{x}|m,\phi_m)f(m) \geq \epsilon_\omega$$

decision threshold ϵ_ω : related to the mean value of $\Delta_\omega(\mathbf{x})$

$$\epsilon_\omega = \frac{c_0}{|\mathcal{S}_\omega|}\sum_{x\in\mathcal{S}_\omega}\Delta_\omega(\mathbf{x}) = \frac{c_0}{|\mathcal{S}_\omega|}\sum_{x\in\mathcal{S}_\omega}\log\sum_{m\in\mathcal{M}_\omega} G(\mathbf{x}|m,\phi_m)f(m)$$

c_0 $\approx$ a constant to balance false positive and false negative decisions (to be optimized by validation set)

mean value of $\Delta_\omega(\mathbf{x})$ $\approx$ maximum-likelihood criterion

$$L_\omega = \frac{1}{|\mathcal{S}_\omega|}\sum_{x\in\mathcal{S}_\omega}\log P(\mathbf{x}|\omega)$$

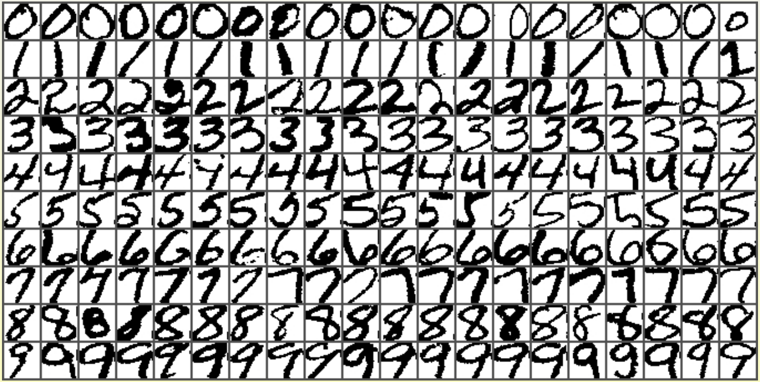
$$L_\omega = \frac{1}{|\mathcal{S}_\omega|}\sum_{x\in\mathcal{S}_\omega}\log F(\mathbf{x}|0) + \frac{1}{|\mathcal{S}_\omega|}\sum_{x\in\mathcal{S}_\omega}\log\left[\sum_{m\in\mathcal{M}_\omega} G(\mathbf{x}|m,\phi_m)f(m)\right]$$

▸ probabilistic neuron

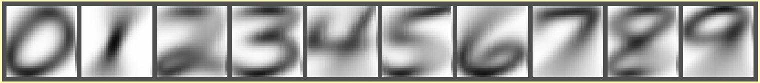
NIST SD19 Database of Hand-Written Numerals

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

examples of numerals normalized to 32x32 binary raster



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



Computational Experiment

DATA

- **odd data vectors for training ($\sum |\mathcal{S}_\omega| = 201485$ numerals)**
- **even data vectors for testing ($\sum |\mathcal{S}_\omega^T| = 201479$ numerals)**
- all numerals normalized to 32x32 binary raster
- database extend by three differently rotated variants of each pattern (by -10,-5,+5 degrees)

estimation of class-conditional structural mixtures:

- conditional distributions $P(\mathbf{x}|\omega)$ estimated from $\mathcal{S}_\omega$ for all $\omega \in \Omega$
- initial number of components chosen identically in all classes: $M_\omega = 200$
- randomly initialized mixture parameters
- stopping rule: relative increment threshold

GOAL: to compare one-class classifiers with the mutually exclusive Bayesian decision-making - using the same distributions $P(\mathbf{x}|\omega)$

Bayesian Recognition of Mutually Exclusive Numerals

in rows: frequencies of decisions for test data from the respective classes
last column: percentage of false negative decisions in the class
last row: total false positive frequencies (in % of all test patterns)

CLASS	0	1	2	3	4	5	6	7	8	9	False n.
♯ Test s.	20182	22352	20038	20556	19577	18303	19969	20947	19790	19767	
0	19950	8	43	19	39	32	36	0	38	17	1.1%
1	2	22162	30	4	35	7	18	56	32	6	0.9%
2	32	37	19742	43	30	9	8	29	90	16	1.5%
3	20	17	62	20021	4	137	2	28	210	55	2.6%
4	11	6	19	1	19170	11	31	51	30	247	2.1%
5	25	11	9	154	4	17925	39	6	96	34	2.1%
6	63	10	17	6	23	140	19652	1	54	3	1.6%
7	7	12	73	10	73	4	0	20497	22	249	2.1%
8	22	25	53	97	30	100	11	11	19369	72	2.1%
9	15	13	25	62	114	22	3	146	93	19274	2.5%
♯ False p.	197	139	537	396	352	462	148	328	665	699	
False p.	0.09%	0.07%	0.27%	0.20%	0.17%	0.23%	0.07%	0.16%	0.33%	0.35%	**1.84%**

Remark: **one decision for each data vector**
mean false negative: **1.84%**

Recognition of Numerals by Using One-Class Classifiers

in columns: frequencies of positive decisions for the respective classifiers

CLASS	0	1	2	3	4	5	6	7	8	9
♯ Test s.	20182	22352	20038	20556	19577	18303	19969	20947	19790	19767
0	18815	2	954	23	30	292	76	0	406	103
1	6	21857	55	46	2756	111	52	4436	5039	410
2	4	9	18660	105	5	2	6	6	207	3
3	6	2	43	18971	1	1733	0	12	3177	373
4	1	0	6	1	18494	5	5	83	265	3229
5	7	2	4	918	0	17211	35	0	1246	282
6	50	10	30	0	60	888	18833	0	360	1
7	1	5	601	324	209	4	0	19817	242	6735
8	9	13	22	620	19	289	6	5	18201	154
9	3	4	6	70	1722	90	2	1060	1266	18667
♯ False neg.	1367	495	1378	1585	1083	1092	1136	1130	1599	1100
False neg.	6.8%	2.2%	6.9%	7.7%	5.5%	6.0%	5.7%	5.4%	8.0%	5.6%
♯ False pos.	87	47	1721	2107	4802	3414	182	5602	12208	11290
False pos.	0.00%	0.00%	0.01%	0.01%	0.02%	0.02%	0.00%	0.03%	0.06%	0.06%

Remark: **ten(!) independent decisions for each data vector**
mean false negative: **5.98%**

Concluding Remarks

Bayes Decision Function

- minimizes the classification error in case of mutually exclusive classes
- fails completely in case of non-exclusive properties

One-Class Classifier

- less precise in case of mutually exclusive classes
- clearly preferable in case of non-exclusive properties
- applicable both to the non-exclusive and exclusive properties
- can be transformed to Bayes decision function by norming in case of equiprobable mutually exclusive classes
- provides a unified approach to recognition of properties and feature extraction

$\Rightarrow$ **assumption of mutually exclusive classes is not suitable to model biological neural networks**

References

EM Algorithm for Structural Mixture Model

structural optimization can be included into EM algorithm

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

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

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

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

structural criterion: Kullback-Leibler I-divergence

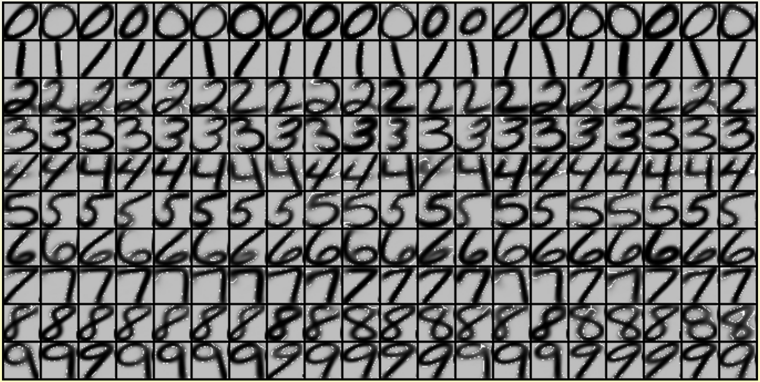
$$\gamma^{'}_{mn} = f^{'}(m) \left[\theta^{'}_{mn} \log \frac{\theta^{'}_{mn}}{\theta_{0n}} + (1 - \theta^{'}_{mn}) \log \frac{(1 - \theta^{'}_{mn})}{(1 - \theta_{0n})} \right]$$

structural parameter optimization: $\phi^{'}_{mn} = 1$ for the r highest values $\gamma^{'}_{mn}$

Remark. The “structural” EM algorithm converges monotonically. Back

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

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



Back

Probabilistic Neuron for Identification of Properties

output layer neuron for the property ω **:**

$$\Delta_\omega(\mathbf{x}) = \sum_{m \in \mathcal{M}_\omega} q(m|\mathbf{x}) \log \left[G(\mathbf{x}|m, \phi_m) f(m) \right] + \sum_{m \in \mathcal{M}_\omega} -q(m|\mathbf{x}) \log q(m|\mathbf{x})$$

$$q(m|\mathbf{x}) = \frac{G(\mathbf{x}|m, \phi_m) f(m)}{\sum_{j \in \mathcal{M}_\omega} G(\mathbf{x}|j, \phi_j) f(j)}, \qquad H(q(\cdot|\mathbf{x})) = \sum_{m \in \mathcal{M}_\omega} -q(m|\mathbf{x}) \log q(m|\mathbf{x})$$

$$\Delta_\omega(\mathbf{x}) = H(q(\cdot|\mathbf{x})) + \sum_{m \in \mathcal{M}_\omega} q(m|\mathbf{x}) \left[\log f(m) + \sum_{n \in \mathcal{N}} \phi_{mn} \log \left(\frac{f_n(x_n|m)}{f_n(x_n|0)} \right) \right]$$

hidden layer neuron: $(m \in \mathcal{M}_\omega)$

$$y_m(\mathbf{x}) = \left[\log f(m) + \sum_{n \in \mathcal{N}} \phi_{mn} \log \left(\frac{1 - \theta_{mn}}{1 - \theta_{0n}} \right) + x_n \log \left(\frac{\theta_{mn}(1 - \theta_{0n})}{\theta_{0n}(1 - \theta_{mn})} \right) \right]$$

$$\Delta_\omega(\mathbf{x}) = H(q(\cdot|\mathbf{x})) + \sum_{m \in \mathcal{M}_\omega} q(m|\mathbf{x}) y_m(\mathbf{x})$$

References 1/4

- Grim J., Hora, J.: Iterative principles of recognition in probabilistic neural networks. *Neural Networks*. Special Issue, Vol. 21, No. 6, pp. 838-846 (2008)
- Grim J.: Extraction of Binary Features by Probabilistic Neural Networks. In: *Artificial Neural Networks - ICANN 2008 Part II*, Springer: Berlin, LNCS **5164** (2008) 52–61
- 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)
- Grim, J., Kittler, J., Pudil, P., Somol, P.: Multiple classifier fusion in probabilistic neural networks. *Pattern Analysis & Applications*, **5** (2002) 221-233

Back

References 2/4

- 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.: Neuromorphic features of probabilistic neural networks. *Kybernetika*, Vol. 43, No. 5, pp. 697-712, (2007)
- Grim J.: Self-organizing maps and probabilistic neural networks. *Neural Network World*, 10, 3, 407–415 (2000)
- 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)

References 3/4

- 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., 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.: Multivariate statistical pattern recognition with non-reduced dimensionality. *Kybernetika*, **22** 6, 142–157 (1986)
- Vajda, I., Grim, J.: About the maximum information and maximum likelihood principles in neural networks. *Kybernetika*, **34** (1998) 485-494
- Haykin, S.: *Neural Networks: a comprehensive foundation*, Morgan Kaufman: San Mateo CA (1993)

Back

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. IEEE International 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)