Structural Poisson Mixtures for Classification of Documents

Jiří Grim, Jana Novovičová, Petr Somol

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

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

19th International Conference on Pattern Recognition
Tampa, Florida, December 8 - 11, 2008

Outline

Representation of Text Documents

PURPOSE: automatic sorting of text documents into predefined classes

text document: $\mathbf{d} = \langle w_{i_1}, \ldots, w_{i_k} \rangle \approx$ list of terms from a vocabulary $\mathcal{V}$

vocabulary of terms: $\mathcal{V} = \{t_1, \ldots, t_N\} \approx$ set of informative terms (obtained from training data by removing stop words and low-frequency words and by stemming, typically $N \approx 10^4$)

document as a "bag of words" (only frequency of terms is considered)

$\mathbf{x} = \mathbf{x}(\mathbf{d}) = (x_1, \ldots, x_N) \in \mathcal{X} = \Im^N \approx$ vector of integers

$x_n \approx$ the frequency of the term $t_n \in \mathcal{V}$

$|\mathbf{x}| = \sum_{n=1}^{N} x_n \approx$ the length of document $\mathbf{x}$

Remark: The "bag of words" representation disregards the position of words in the original documents.

Statistical Approach to Document Classification

probabilistic description:

$P(\mathbf{x}|c)p(c), \quad c \in \mathcal{C}$: conditional distributions of classes

$\mathcal{C} = \{c_1, \ldots, c_J\} \approx$ set of classes with *a priori* probabilities $p(c), c \in \mathcal{C}$

decision making based on Bayes formula:

$$p(c|\mathbf{x}) = \frac{P(\mathbf{x}|c)p(c)}{P(\mathbf{x})}, \qquad P(\mathbf{x}) = \sum_{c \in \mathcal{C}} p(c)P(\mathbf{x}|c)$$

the "naive" Bayes classifier: conditional independence of variables

$$P(\mathbf{x}|c) = \prod_{n \in \mathcal{N}} f_n(x_n|c), \; c \in \mathcal{C}, \;\; \mathcal{N} = \{1, \ldots, N\}$$

Remark: Naive Bayes classifier disregards statistical dependencies between vocabulary terms. Despite many attempts no essential improvement has been achieved by considering the dependencies in a way (cf. e.g. Lewis 1998).

ÚTIA

Application of Poisson Mixtures

Idea: approximation of $P(\mathbf{x}|c)$ by mixtures of Poisson distributions

$$P(\mathbf{x}|c) = \sum_{m \in \mathcal{M}_c} f(m) \prod_{n \in \mathcal{N}} f_n(x_n|\lambda_{mn}), \quad f(m) \geq 0, \sum_{m \in \mathcal{M}_c} f(m) = 1$$

$$f_n(x_n|\lambda_{mn}) = \frac{(\lambda_{mn})^{x_n}}{x_n!} \mathrm{e}^{-\lambda_{mn}}, \quad (|\mathbf{x}| = \sum_{n=1}^{N} x_n)$$

$f_n() \approx$ probability that $t_n \in \mathcal{V}$ occurs x_n-times in a document of length $|\mathbf{x}|$

$\lambda_{mn} \approx$ mean frequency of the term t_n in a document of a given length $|\mathbf{x}|$

the document length may be different $\rightarrow \lambda_{mn} = \theta_{mn}|\mathbf{x}|$

$$P(\mathbf{x}|c) = \sum_{m \in \mathcal{M}_c} F(\mathbf{x}|\boldsymbol{\theta}_m) f(m) = \prod_{n \in \mathcal{N}} f_n(x_n|\theta_{mn}|\mathbf{x}|) = \prod_{n \in \mathcal{N}} \frac{(\theta_{mn}|\mathbf{x}|)^{x_n}}{x_n!} \mathrm{e}^{-\theta_{mn}|\mathbf{x}|}$$

$F(\mathbf{x}|\boldsymbol{\theta}_m) \approx$ product Poisson distributions

Remark: Mixture of product Poisson distributions has $M(N+1)$ parameters $\Rightarrow$ very high number in case of documents.

Structural Mixture Model

"structural" multivariate Poisson mixtures:

$$P(\mathbf{x}|c) = \sum_{m \in \mathcal{M}_c} F(\mathbf{x}|\boldsymbol{\theta}_0) G(\mathbf{x}|\boldsymbol{\theta}_m, \phi_m) f(m), \quad c \in \mathcal{C}$$

$F(\mathbf{x}|\boldsymbol{\theta}_0) \approx$ "background" probability distribution common to all classes

$$F(\mathbf{x}|\boldsymbol{\theta}_0) = \prod_{n \in \mathcal{N}} f_n(x_n|\theta_{0n}|\mathbf{x}|) = \prod_{n \in \mathcal{N}} \frac{(\theta_{0n}|\mathbf{x}|)^{x_n}}{x_n!} \mathrm{e}^{-\theta_{0n}|\mathbf{x}|}$$

$G(\mathbf{x}|\boldsymbol{\theta}_m, \phi_m) \approx$ component functions
$\phi_{mn} \in \{0, 1\} \approx$ structural parameters

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

$F(\mathbf{x}|\boldsymbol{\theta}_0)$ **can be canceled in the Bayes formula:**

$$p(c|\mathbf{x}) = \frac{p(c) \sum_{m \in \mathcal{M}_c} G(\mathbf{x}|\boldsymbol{\theta}_m, \phi_m) f(m)}{\sum_{c \in \mathcal{C}} p(c) \sum_{j \in \mathcal{M}_c} G(\mathbf{x}|\boldsymbol{\theta}_j, \phi_j) f(j)}.$$

Structural Mixture Model Estimation

log-likelihood function:

$$L = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} \log[\sum_{m \in \mathcal{M}_c} G(\mathbf{x}|\boldsymbol{\theta}_m, \phi_m) f(m)], \qquad \mathcal{S}_c = \{\mathbf{x}_1, \ldots, \mathbf{x}_{K_c}\}$$

EM algorithm:

$$q(m|\mathbf{x}) = \frac{G(\mathbf{x}|\boldsymbol{\theta}_m, \phi_m) f(m)}{\sum_{j \in \mathcal{M}_c} G(\mathbf{x}|\boldsymbol{\theta}_j, \phi_j) f(j)}, \qquad m \in \mathcal{M}_c,\ n \in \mathcal{N},\ \mathbf{x} \in \mathcal{S}_c$$

$$\tilde{x}_n^{(m)} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} x_n q(m|\mathbf{x}), \qquad |\bar{\mathbf{x}}|^{(m)} = \frac{1}{|\mathcal{S}_c|} \sum_{x \in \mathcal{S}_c} |\mathbf{x}| q(m|\mathbf{x})$$

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

$$\phi^{'}_{mn} = \begin{cases} 1, & \gamma^{'}_{mn} \in \Gamma^{'}_r, \\ 0, & \gamma^{'}_{mn} \notin \Gamma^{'}_r, \end{cases}, \qquad \gamma^{'}_{mn} = \tilde{x}_n^{(m)} \log \frac{\theta^{'}_{mn}}{\theta_{0n}} + |\bar{\mathbf{x}}|^{(m)} (\theta^{'}_{0n} - \theta_{mn})$$

$\Gamma^{'}_r$ is the set of r highest quantities $\gamma^{'}_{mn}$ ▸ Proof

Example 1: Classification of REUTERS text documents

REUTERS text documents:

(small classes and multiply labeled documents removed)
8941 documents partitioned into 33 different classes
10105 vocabulary terms (by removing stop words and after stemming)
6431 training documents, 2510 test documents ($\approx$ APTE split)

Experiment No.	**1**	**2**	**3**	**4**	**5**
Components M	33	33	35	35	43
Parameters $\sum \phi_{mn}$	333465	208366	285220	327184	201417
Parameters [in %]	100.0	62.5	80.6	92.5	46.4
Classification Errors	155	156	162	152	147
Classif. Error [in %]	6.17	6.21	6.45	6.07	5.86

Remark. The best classification result (experiment 5) is only slightly better than the "naive" Bayes classification accuracy (experiment 1).

Example 2: Classification of 20 NEWSGROUPS documents

20 NEWSGROUPS text documents:

19956 documents partitioned nearly evenly into 20 different classes
31826 vocabulary terms (by removing stop words and after stemming)
13314 training documents, 6632 test documents
(random partition, no multiple labels)

Experiment No.	**1**	**2**	**3**	**4**	**5**
Components M	20	40	40	60	80
Parameters $\sum \phi_{mn}$	636520	1204262	1102073	1276602	1024782
Parameters [in %]	100.0	94.6	86.6	66.8	40.2
Classification Errors	1406	1379	1370	1362	1412
Classif. Error [in %]	21.20	20.79	20.66	20.54	21.29

Remark. The results differ only by several tens of erroneously classified documents, the “naive” Bayes classification is only slightly worse.

Concluding Remarks

Properties of Structural Poisson Mixtures

- enable statistically correct subspace approach to Bayes classification of documents
- the class-conditional distributions and even individual components may be defined on different subspaces
- $\Rightarrow$ the number of parameters in the conditional distributions can be reduced without restricting the set of vocabulary terms

Classification Performance

- the recognition error slightly decreases with increasing model complexity and simultaneously decreasing number of parameters
- **probable reason:** the number of documents in the training data sets is not sufficient to utilize more complex statistical properties

Literatura 1/2

G. Forman.
An experimental study of feature selection metrics for text categorization.
Journal of Machine Learning Research, 3:1289–1305, 2003.

J. Grim.
Multivariate statistical pattern recognition with nonreduced dimensionality.
Kybernetika, 22(2):142–157, 1986.

J. Grim, J. Kittler, P. Pudil, and P. Somol.
Multiple classifier fusion in probabilistic neural networks.
Pattern Analysis & Appl., 7(5):221–233, 2002.

J. Grim, M. Haindl, P. Somol, and P. Pudil.
A subspace approach to texture modelling by using gaussian mixtures.
In B. Haralick and T. K. Ho, eds., *Proc. of the 18th Conference ICPR 2006*, pages 235–238, Hong Kong, 2006.

Literatura 2/2

S. Kim, K. Han, H. Rim, and S. Myaeng.
Some effective techniques for naive bayes text classification.
IEEE Trans. on Knowl. and Data Engineering, 18(11):1457–1466, 2006.

D. Lewis.
Naive (bayes) at forty: The independence assumption in information retrieval.
In *10-th European Conf. on Machine Learning ECML-98*, pages 4–15, 1998.

A. McCallum and K. Nigam.
A comparison of event models for naive Bayes text classification.
In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, 1998.

F. Sebastiani.
Machine learning in automated text categorization.
ACM Comp. Surveys, 34(1):1–47, March 2002.

Proof of the Monotonic Property

Kullback-Leibler information divergence is nonnegative:

$$\frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}}[\sum_{m\in\mathcal{M}} q(m|\mathbf{x}) \log \frac{q(m|\mathbf{x})}{q^{'}(m|\mathbf{x})}] \geq 0$$

by making substitution we can write

$$\frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} \log \frac{P^{'}(\mathbf{x})}{P(\mathbf{x})} - \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}}\sum_{m\in\mathcal{M}} q(m|\mathbf{x}) \log \left[\frac{f^{'}(m) G(\mathbf{x}|\boldsymbol{\theta}^{'}_m, \phi^{'}_m)}{f(m) G(\mathbf{x}|\boldsymbol{\theta}_m, \phi_m)}\right] \geq 0.$$

$$L^{'} - L \geq \sum_{m\in\mathcal{M}} f^{'}(m) \log \frac{f^{'}(m)}{f(m)} + \sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}) \log \left[\frac{G(\mathbf{x}|\boldsymbol{\theta}^{'}_m, \phi^{'}_m)}{G(\mathbf{x}|\boldsymbol{\theta}_m, \phi_m)}\right]$$

the first sum on the right is nonnegative and therefore

$$L^{'} - L \geq \sum_{m\in\mathcal{M}} \frac{1}{|\mathcal{S}|}\sum_{\mathbf{x}\in\mathcal{S}} q(m|\mathbf{x}) \log \left[\frac{G(\mathbf{x}|\boldsymbol{\theta}^{'}_m, \phi^{'}_m)}{G(\mathbf{x}|\boldsymbol{\theta}_m, \phi_m)}\right]$$

Proof of the Monotonic Property

further, making substitution, we obtain

$$L^{'} - L \geq \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \frac{1}{|\mathcal{S}|} \sum_{\mathbf{x} \in \mathcal{S}} q(m|\mathbf{x}) \log \frac{\left[(\frac{\theta^{'}_{mn}}{\theta_{0n}})^{x_n} \exp\left(\theta_{0n} - \theta^{'}_{mn}\right)|\mathbf{x}| \right]^{\phi^{'}_{mn}}}{\left[(\frac{\theta_{mn}}{\theta_{0n}})^{x_n} \exp\left(\theta_{0n} - \theta_{mn}\right)|\mathbf{x}| \right]^{\phi_{mn}}}$$

the last inequality can be rewritten in the form

$$L^{'} - L \geq \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \left[\phi^{'}_{mn} \gamma_{mn}(\theta^{'}_{mn}) - \phi_{mn} \gamma_{mn}(\theta_{mn}) \right]$$

where

$$\gamma_{mn}(\theta_{mn}) = \tilde{x}_n^{(m)} \log \frac{\theta_{mn}}{\theta_{0n}} + |\bar{\mathbf{x}}|^{(m)} (\theta_{0n} - \theta_{mn})$$

in view of the definition of $\theta^{'}_{mn}$ and $\phi^{'}_{mn}$ we can write Return

$$\theta^{'}_{mn} = \frac{\tilde{x}_n^{(m)}}{|\bar{\mathbf{x}}|^{(m)}} = \arg\max_{\theta_{mn}} \{\gamma_{mn}(\theta_{mn})\} \quad \Rightarrow \quad \gamma_{mn}(\theta^{'}_{mn}) \geq \gamma_{mn}(\theta_{mn})$$

$$\Rightarrow \quad L^{'} - L \geq \sum_{m \in \mathcal{M}} \sum_{n \in \mathcal{N}} \left[\phi^{'}_{mn} - \phi_{mn} \right] \gamma_{mn}(\theta^{'}_{mn}) \geq 0$$