# Structural Poisson Mixtures for Classification of Documents

Jiří Grim, Jana Novovičová, Petr Somol
*Institute of Information Theory and Automation*
*P.O.BOX 18, CZ-18208 Prague 8, Czech Republic*
*grim@utia.cas.cz, novovic@utia.cas.cz, somol@utia.cas.cz*

## Abstract

*Considering the statistical text classification problem we approximate class-conditional probability distributions by structurally modified Poisson mixtures. By introducing the structural model we can use different subsets of input variables to evaluate conditional probabilities of different classes in the Bayes formula. The method is applicable to document vectors of arbitrary dimension without any preprocessing. The structural optimization can be included into the EM algorithm in a statistically correct way.*

## 1. Introduction

Text classification as a problem of automatic sorting of documents into predefined classes is important in many information retrieval tasks. Various statistical and machine learning techniques have been explored to build automatically a classifier by learning from previously labeled documents. For a discussion of different approaches see e.g. Sebastiani [7].

We consider classification of the text documents in a Bayesian learning framework with a bag-of-words document representation. There are two common models in the representation of text documents (see e.g. [6], [5]). The multivariate Bernoulli model represents each document by a vector of binary feature variables indicating whether or not a certain word occurs in the document. Alternatively, in the multinomial model, the features are defined as frequencies of the related vocabulary terms in the document. In both cases the dimension of document vectors is very high because of a large number of vocabulary terms and therefore different feature selection methods have to be used as a rule [1]. Unfortunately, it is difficult to reduce the size of vocabulary since there are many different classes having different subsets of characteristic terms. An informative subset of features common to all classes often represents a difficult compromise possibly connected with a loss of classification accuracy.

In this paper we propose the use of a structural mixture of multivariate Poisson distributions to learn Bayesian text classifier. By introducing binary structural parameters we can reduce the evaluation of the Bayes formula only to subsets of informative variables which may be different for different classes and even for different mixture components. In this way we can reduce the number of parameters in the conditional distributions without reducing the number of vocabulary terms.

The paper is organized as follows. In Section 2 we describe the problem of statistical text classification, Section 3 introduces the structural Poisson mixture model and Section 4 describes the corresponding modified EM algorithm. In Section 5 we describe the computational experiments and finally we summarize the results in the Conclusion.

## 2. Statistical Document Classification

We assume that after standard preprocessing a text document $\boldsymbol{d}$ is reduced to a finite list of terms from a given vocabulary $\mathcal{V}$

$$\boldsymbol{d} = \langle w_{i_1}, \dots, w_{i_k} \rangle, \;\; w_{i_l} \in \mathcal{V} = \{t_1, \dots, t_N\}. \quad (1)$$

The vocabulary is chosen to characterize the semantic meaning of documents with a limited number of highly informative specific terms. For the sake of classification we ignore common short or rare words and disregard the position of words in the original document. A document is treated as a "bag of words", only the frequency of vocabulary terms is considered. In this sense, denoting $x_n$ the frequency of the term $t_n \in \mathcal{V}$, we describe a document (1) by $N$-dimensional vector of integers

$$\boldsymbol{x} = \boldsymbol{x}(\boldsymbol{d}) = (x_1, x_2, \dots, x_N) \in \mathcal{X} = \Im^N. \quad (2)$$

In the following we denote $|\boldsymbol{x}|$ the length of document $\boldsymbol{x}$ which may correspond to the total number of words in

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

the original non-reduced document or may be set equal to the sum of frequencies of the vocabulary terms.

Considering the statistical classification problem we assume random occurrence of documents from a finite set of classes $\mathcal{C} = \{c_1, \dots, c_J\}$ with *a priori* probabilities $p(c), c \in \mathcal{C}$ and according to some class-conditional probability distributions $P(\boldsymbol{x}|c)$ on $\mathcal{X}$. If the probabilistic description is known then the document classification is easily made by computing the conditional probabilities $p(c|\boldsymbol{x})$ by means of Bayes formula

$$p(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c)p(c)}{P(\boldsymbol{x})}, \quad P(\boldsymbol{x}) = \sum_{c \in \mathcal{C}} P(\boldsymbol{x}|c)p(c). \quad (3)$$

In this paper we propose a probabilistic model for text documents based on the mixtures of multivariate Poisson distributions. The Poisson distribution defines the probability that certain phenomenon occurs in a fixed period of time exactly $x-$times, given that the mean number of occurrences in this time interval is $\lambda$:

$$f(x|\lambda) = \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda}. \quad (4)$$

Formula (4) is usually applied in physics to describe the random number of breakdowns in a radioactive material in a unit of time or in telephony to characterize the number of calls in a given time interval.

In the last years the Poisson distribution has been used repeatedly to characterize the frequency of terms in text documents. In this case the Poisson mean $\lambda_n$ analogously denotes the mean frequency of a vocabulary term $t_n \in \mathcal{V}$ in a document of a given length $|\boldsymbol{x}|$. However, the length of documents may be different. For this reason the document vectors have to be normalized since otherwise formula (4) would not be applicable. The term frequencies $x_n \in \Im$ are usually recomputed to a huge document length in order to avoid large rounding errors. Simultaneously, Laplace smoothing parameters are often applied to damp down excessive influence of very small documents (cf. [4]).

In this paper we prefer a simple statistically correct solution of this problem. Making substitution $\lambda_n = \theta_n|\boldsymbol{x}|$ we replace the parameter $\lambda_n$ by the respective relative frequency $\theta_n$

$$f_n(x_n|\theta_n|\boldsymbol{x}|) = \frac{(\theta_n|\boldsymbol{x}|)^{x_n}}{x_n!}\mathrm{e}^{-\theta_n|\boldsymbol{x}|}. \quad (5)$$

Note that formula (5) naturally includes the length of documents $|\boldsymbol{x}|$ into decision making.

## 3. Structural Poisson Mixture

Assuming statistically independent frequencies $x_n$ of different vocabulary terms we could write:

$$P(\boldsymbol{x}|c) = \prod_{n \in \mathcal{N}} f_n(x_n|c), \; c \in \mathcal{C}, \; \mathcal{N} = \{1, \dots, N\}. \quad (6)$$

In spite of the unrealistic independence assumption (6) the resulting "naive" Bayes classifier (3) has been shown to give good results in practical experiments.

There have been many attempts to improve the naive Bayes classifiers by considering the statistical dependence of vocabulary terms in documents [5]. One possibility is to approximate the class-conditional distributions $P(\boldsymbol{x}|c)$ by finite Poisson mixtures of the form

$$P(\boldsymbol{x}|c) = \sum_{m \in \mathcal{M}_c} F(\boldsymbol{x}|\boldsymbol{\theta}_m)f(m), \; \boldsymbol{x} \in \mathcal{X}, \; c \in \mathcal{C} \quad (7)$$

where $f(m) \geq 0, \; \sum_{m \in \mathcal{M}_c} f(m) = 1$ are some conditional probabilistic weights, $F(\boldsymbol{x}|\boldsymbol{\theta}_m)$ are the component Poisson distributions

$$F(\boldsymbol{x}|\boldsymbol{\theta}_m) = \prod_{n \in \mathcal{N}} f_n(x_n|\theta_{mn}|\boldsymbol{x}|) =$$

$$= \prod_{n \in \mathcal{N}} \frac{(\theta_{mn}|\boldsymbol{x}|)^{x_n}}{x_n!}\mathrm{e}^{-\theta_{mn}|\boldsymbol{x}|} \quad (8)$$

and $\mathcal{M}_c$ is the index set for the class $c \in \mathcal{C}$. In order to keep the notation simple we assume a consecutive indexing of components in the conditional mixtures (7). In this way the component index $m \in \mathcal{M}_c$ uniquely identifies the class $c \in \mathcal{C}$ and therefore the parameter $c$ can be partly omitted whenever tolerable.

In the present paper we apply the multivariate Poisson mixture (7) in a structural modification originally proposed for pattern recognition [2]. Recently the structural mixtures have been applied to texture modeling [3]. By introducing binary structural parameters in the components (8) we can reduce the evaluation of Bayes formula only to subsets of informative variables. In particular, making substitution

$$F(\boldsymbol{x}|\boldsymbol{\theta}_m) = F(\boldsymbol{x}|\boldsymbol{\theta}_0)G(\boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m), \quad (9)$$

we introduce a "structural" distribution mixture

$$P(\boldsymbol{x}|c) = \sum_{m \in \mathcal{M}_c} F(\boldsymbol{x}|\boldsymbol{\theta}_0)G(\boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m)f(m) \quad (10)$$

where $F(\boldsymbol{x}|\boldsymbol{\theta}_0)$ is a nonzero "background" probability distribution common to all classes

$$F(\boldsymbol{x}|\boldsymbol{\theta}_0) = \prod_{n \in \mathcal{N}} f_n(x_n|\theta_{0n}|\boldsymbol{x}|) =$$

$$= \prod_{n\in\mathcal{N}} \frac{(\theta_{0n}|\boldsymbol{x}|)^{x_n}}{x_n!} \mathrm{e}^{-\theta_{0n}|\boldsymbol{x}|} \tag{11}$$

and the component functions $G(\boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m)$ include additional binary structural parameters $\phi_{mn} \in \{0,1\}$:

$$G(\boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m) = \prod_{n\in\mathcal{N}} \left[\frac{f_n(x_n|\theta_{mn}|\boldsymbol{x}|)}{f_n(x_n||\boldsymbol{x}|\theta_{0n})}\right]^{\phi_{mn}} = \tag{12}$$

$$= \prod_{n\in\mathcal{N}} \left[\left(\frac{\theta_{mn}}{\theta_{0n}}\right)^{x_n} \mathrm{e}^{(\theta_{0n}-\theta_{mn})|\boldsymbol{x}|}\right]^{\phi_{mn}}.$$

The background distribution (11) is usually defined as a product of fixed marginals estimated from the union of the available data sets ($\mathcal{S} = \bigcup_{c\in\mathcal{C}} \mathcal{S}_c$):

$$\theta_{0n} = \frac{\bar{x}_n}{|\bar{\boldsymbol{x}}|}, \;\; \bar{x}_n = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} x_n, \;\; |\bar{\boldsymbol{x}}| = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}} |\boldsymbol{x}|.$$

Note that in the Bayes formula the background distribution $F(\boldsymbol{x}|\boldsymbol{\theta}_0)$ can be canceled and we can write

$$p(c|\boldsymbol{x}) = \frac{p(c)\sum_{m\in\mathcal{M}_c} G(\boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m)f(m)}{\sum_{c\in\mathcal{C}} p(c)\sum_{j\in\mathcal{M}_c} G(\boldsymbol{x}|\boldsymbol{\theta}_j, \boldsymbol{\phi}_j)f(j)}. \tag{13}$$

Consequently, the *a posteriori* probability $p(c|\boldsymbol{x})$ is proportional to the weighted sum of the component functions $G(\boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m)$ which can be defined on different subspaces. Obviously, the subspaces may be different for different classes and even for different components. In this way we can reduce the number of parameters in the conditional distributions without restricting the number of vocabulary terms.

## 4. Structural Model Estimation

An advantage of the structural mixture model (10) is the computationally feasible optimization which can be included into EM algorithm. In particular, we assume that for each class $c \in \mathcal{C}$ there is a corresponding sample $\mathcal{S}_c$ of typical documents

$$\mathcal{S}_c = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{K_c}\}, \quad \boldsymbol{x}_k \in \mathcal{X}. \tag{14}$$

In order to compute maximum-likelihood estimates of the mixture parameters $f(m), \theta_{mn}, \phi_{mn}$ we have to maximize the corresponding log-likelihood function. As the background distribution $F(\boldsymbol{x}|\boldsymbol{\theta}_0)$ is fixed we can maximize the following criterion

$$L = \frac{1}{|\mathcal{S}_c|}\sum_{x\in\mathcal{S}_c} \log\left[\sum_{m\in\mathcal{M}_c} G(\boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m)f(m)\right]. \tag{15}$$

We can derive the following modified EM iteration equations ($m \in \mathcal{M}_c, n \in \mathcal{N}, \boldsymbol{x} \in \mathcal{S}_c$):

$$q(m|\boldsymbol{x}) = \frac{G(\boldsymbol{x}|\boldsymbol{\theta}_m, \boldsymbol{\phi}_m)f(m)}{\sum_{j\in\mathcal{M}_c} G(\boldsymbol{x}|\boldsymbol{\theta}_j, \boldsymbol{\phi}_j)f(j)}, \tag{16}$$

$$\tilde{x}_n^{(m)} = \frac{1}{|\mathcal{S}_c|}\sum_{x\in\mathcal{S}_c} x_n q(m|\boldsymbol{x}), \tag{17}$$

$$|\bar{\boldsymbol{x}}|^{(m)} = \frac{1}{|\mathcal{S}|}\sum_{x\in\mathcal{S}_c} |\boldsymbol{x}| q(m|\boldsymbol{x}), \tag{18}$$

$$f^{'}(m) = \frac{1}{|\mathcal{S}_c|}\sum_{x\in\mathcal{S}_c} q(m|\boldsymbol{x}), \;\; \theta^{'}_{mn} = \frac{\tilde{x}_n^{(m)}}{|\bar{\boldsymbol{x}}|^{(m)}}, \tag{19}$$

$$\gamma^{'}_{mn} = \tilde{x}_n^{(m)} \log\frac{\theta^{'}_{mn}}{\theta_{0n}} + |\bar{\boldsymbol{x}}|^{(m)}(\theta^{'}_{0n} - \theta_{mn}), \tag{20}$$

$$\phi^{'}_{mn} = \begin{cases} 1, & \gamma^{'}_{mn} \in \Gamma^{'}_r, \\ 0, & \gamma^{'}_{mn} \notin \Gamma^{'}_r, \end{cases}, \tag{21}$$

where the apostrophe denotes the new parameter values and $\Gamma^{'}_r$ denotes the set of $r$ highest quantities $\gamma^{'}_{mn}$.

It can be verified that the iterative equations (16) - (21) generate a nondecreasing sequence of values $\{L\}_0^\infty$ converging to a possibly local maximum of the criterion (15).

## 5. Numerical Examples

The structural Poisson mixture model has been applied to classify the standard document collections Reuters-21578 and 20 Newsgroups. The results are given in Table 1 and Table 2 respectively.

For the sake of classification the Reuters data have been reduced to 8941 documents and 33 classes by discarding documents having multiple or no class labels and by removing classes with less than twenty documents. By removing stop words and low frequency words (occurring less than 3 times) we have obtained a vocabulary of 10105 words after stemming. Finally, according to the Apte split, we have obtained a training set of 6431 documents and test set of 2510 documents.

The 20 Newsgroups data set is a collection of 19956 documents partitioned nearly evenly into 20 different classes. This data set is exclusively single-labeled. By removing stop words and stemming we have obtained 31826 vocabulary terms. We have selected randomly two-thirds of documents for training and the rest has been used for testing.

In both examples we have estimated in several experiments the class-conditional mixtures of different complexity by means of EM algorithm of Sec. 4. Unlike

**Table 1. Classification of Reuters text documents (APTE split) by structural Poisson mixtures (33 classes, 10105 vocabulary terms, 6431 training documents, 2510 test documents).**

| **Experiment No.:** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| Number of Components | 33 | 33 | 35 | 35 | 35 | 35 | 43 |
| Number of Parameters | 333465 | 208366 | 209366 | 285220 | 334781 | 327184 | 201417 |
| Number of Parameters [in %] | 100.0 | 62.5 | 59.2 | 80.6 | 94.6 | 92.5 | 46.4 |
| Number of Errors | 155 | 156 | 163 | 162 | 156 | 152 | 147 |
| Classification Error [in %] | 6.17 | 6.21 | 6.49 | 6.45 | 6.21 | 6.07 | 5.86 |

**Table 2. Classification of 20 NEWSGROUPS text documents by structural Poisson mixtures (20 classes, 31826 vocabulary terms, 13314 training documents, 6632 test documents).**

| **Experiment No.:** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
|---|---|---|---|---|---|---|---|
| Number of Components | 20 | 20 | 40 | 40 | 60 | 60 | 80 |
| Number of Parameters | 668346 | 580733 | 1204262 | 1102073 | 1407277 | 1276602 | 1024782 |
| Number of Parameters [in %] | 100.0 | 91.2 | 94.6 | 86.6 | 73.7 | 66.8 | 40.2 |
| Number of Errors | 1406 | 1404 | 1379 | 1370 | 1390 | 1362 | 1412 |
| Classification Error [in %] | 21.20 | 21.17 | 20.79 | 20.66 | 20.95 | 20.54 | 21.29 |

Eq. (21) the structural parameters $\phi_{mn}$ have been chosen in each iteration by simple thresholding. The EM algorithm has been initialized randomly and stopped by relative increment threshold after several iterations. For each experiment the total number of mixture components is given in the second row of the tables. The resulting total number of component specific parameters is given in the third row of both tables. In the fourth row the same number is expressed relatively in % of the corresponding "full" mixture model. The last two rows compare the classification performance. Roughly speaking, the recognition error slightly decreases with increasing model complexity and simultaneously decreasing number of parameters.

## 6. Concluding Remarks

The proposed method of structural Poisson mixtures represents statistically correct subspace approach to Bayesian decision-making. It is directly applicable to a non-reduced input space of arbitrary dimension and allows for different feature subsets for different classes. We succeeded to improve the classifier performance by relaxing the popular feature independence assumption and using structural mixtures with reduced number of parameters. The advantage of different class-specific feature sets should be more relevant in case of multiclass problems.

**Acknowledgment:** This work was supported by the project GAČR No. 102/07/1594 of the Czech Grant Agency and partially by the MŠMT projects 2C06019 ZIMOLEZ and 1M0572 DAR.

## References

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

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

[3] 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.

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

[5] 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.

[6] 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.

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