ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Computational Statistics & Data Analysis ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csd

Two-way Poisson mixture models for simultaneous document classification and word clustering

Jia Li[a,*], Hongyuan Zha[b]

[a] *Department of Statistics, Penn State, 417A Thomas Bldg, University Park, PA 16802, USA*
[b] *Department of Computer Science and Engineering, Penn State, University Park, PA 16802, USA*

Abstract

An approach to simultaneous document classification and word clustering is developed using a two-way mixture model of Poisson distributions. Each document is represented by a vector with each dimension specifying the number of occurrences of a particular word in the document in question. As a collection of documents across several classes usually makes use of a large number of words, the document vectors are of high dimension. On the other hand, the number of distinct words in any single document is usually substantially smaller than the size of the vocabulary, leading to sparse document vectors. A mixture of Poisson distributions is used to model the multivariate distribution of the word counts in the documents within each class. To address the issues of high dimensionality and sparsity, the parameters in the mixture model are regularized by imposing a clustering structure on the set of words. An EM-style algorithm for the two-way mixture model will be derived for parameter estimation with the clustering of words part of the estimation process. The connection of the two-way mixture model with dimension reduction will also be elucidated. Experiments on the newsgroup data have demonstrated promising results.

MSC: 62H30; 68T10

Keywords: Two-way mixture model; Mixture of Poisson distributions; Document classification; Word clustering; Simultaneous classification and clustering

* Corresponding author. Tel.: +1-814-863-3074; fax: +1-814-863-7114.
E-mail addresses: jiali@stat.psu.edu (J. Li), zha@cse.psu.edu (H. Zha).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.07.013

1. Introduction

The last few years saw an exponential growth of textual information available in the public World Wide Web, corporate intranets, news wires and elsewhere. While the amount of textual data is constantly increasing, our ability to process and utilize this information has remained largely unchanged. One of the great challenges for today's information science and technology is to develop algorithms and software for efficiently and effectively organizing, accessing and mining this vast amount of information. In this paper, we focus on the task of classifying natural language documents into a pre-defined set of topical categories, commonly referred to as document classification. Document classification is an enabling technology that is essential for many information processing applications. For example, it can be used as a building block to classify Web documents into a directory system such as Yahoo! or Open Directory. It can also be used to categorize the incoming emails in a company for spam detection, routing for automatic machine response or sending to the correct human recipient (Joachims, 2002; Yang, 1999; Zhang, 2001).

In this paper, we follow the general paradigm of representing text documents using the vector space model (Belew, 2000; Salton, 1989). Each document in a collection is represented by a p-dimensional vector, and each coordinate of the vector (variable) corresponds to a word in a vocabulary of size p. This formulation leads to the so-called term-document matrix $A = [a_{ij}]$ for the representation of the collection of documents, where a_{ij} is the so-called term frequency, i.e., the number of times word i occurs in document j. In this vector space model terms and documents are treated asymmetrically with terms considered as the covariates or attributes of documents.

For a collection of documents across several topical classes, it usually makes use of a large number of words leading to a large vocabulary size p, and all the document vectors live in a high-dimensional space. On the other hand, the number of distinct words in any single document is usually substantially smaller than the size of the vocabulary, leading to sparse document vectors, vectors with many zero components. High dimensionality and sparsity do pose a challenge to many classification algorithms. Several methods have been proposed to handle those problems, for example, there are methods for selecting a subset of words based on various heuristics such as document frequency cut-off and mutual information (see Yang and Pedersen, 1997 for a survey). However, it is also well-known that the reduction of the number of words based on feature selection cannot be too aggressive, otherwise the classification accuracy will suffer (Joachims, 2002). Notable among those that allow aggressive feature space reduction is the distributional clustering approach whereby words are clustered into groups based on the distribution of class labels associated with each word (Baker and McCallum, 1998). Distributional clustering is used as a pre-processing step to compress the size of the feature space used for document classification and the clustering is homogeneous across all the classes. As we will show later, inhomogeneous word clustering can improve classification accuracy. There are also several document classification methods that do not rely on aggressive feature reduction, among which we mention Naive Bayes method (McCallum and Nigam, 1998), support vector machines (Joachims, 2002) and regularized linear classifiers (Zhang, 2001). All these methods, unlike the two-way mixture model we will propose, do not produce word clustering as an integral part of the modeling and classification process. Another related line of research is the simultaneous clustering

approach (Hofmann, 2001) (also known as biclustering or co-clustering, Zha et al., 2001) whereby data instances and their attributes are simultaneously clustered to enhance clustering effectiveness and cluster interpretability. However, the simultaneous clustering approach focuses on a fixed set of documents, and therefore, in a strict sense, it does not provide a generative model for arbitrary documents. Ways to overcome this difficulty have been proposed in Blei et al. (2003) based on a model that involves a much complicated optimization problem.

In our approach, we characterize each class by a mixture model with a word clustering structure, and combine the class models in an overall classification framework. The distribution of the document vectors within each class is modeled by a mixture of multivariate probability mass functions (pmf) with independent variables each following a Poisson distribution. The variables are in general not conditionally independent given the class because a class may contain multiple mixture components. By combining several additive components, the mixture model is flexible for characterizing the distribution of the document vector. As we have mentioned, high dimensionality and sparsity pose difficulty for accurately estimating classification models. Therefore a clustering structure is imposed on the variables to tackle the issue, leading to a two-way mixture model for each class. Variables in the same cluster are assumed to have equal Poisson parameters within each mixture component. The searching for the optimal partition of words into clusters is an integrated part of fitting the mixture model.

The rest of the paper is organized as follows. In Section 2, the two-way mixture of Poisson distributions is described. Its dimension reduction property is discussed in Section 3. The algorithm to estimate the model and to form word clusters is described in Section 4 with some of the details relegated to the appendix. Experiments are presented in Section 5. We conclude in Section 6.

2. Two-way mixtures of Poisson distributions

We estimate the distribution of the document vector in each class using a parametric mixture model. In particular, every mixture component is a multivariate distribution that has independent variables each following a Poisson distribution. Denote the document vector by $X = (X_1, X_2, \ldots, X_p)^{\mathrm{T}}$, where p is the dimension, i.e., the size of the vocabulary. Let the class label be $Y \in \mathscr{K} = \{1, 2, \ldots, K\}$. Then

$$P(X = x \mid Y = k) = \sum_{r=1}^{R_k} \pi_{kr} \prod_{j=1}^{p} \phi(x_j \mid \lambda_{k,r,j}), \tag{1}$$

where ϕ denotes the pmf of a Poisson distribution: $\phi(x_j \mid \lambda_{k,r,j}) = \lambda_{k,r,j}^{x_j} e^{-\lambda_{k,r,j}} / x_j!$, and π_{kr} is the prior probability of component r. Suppose the marginal probability $P(Y = k) = a_k$, $\sum_{k=1}^{K} a_k = 1$. The total number of components in all the classes is $M = \sum_{k=1}^{K} R_k$. To avoid the notational complexity of specifying the distribution of X separately for each class,

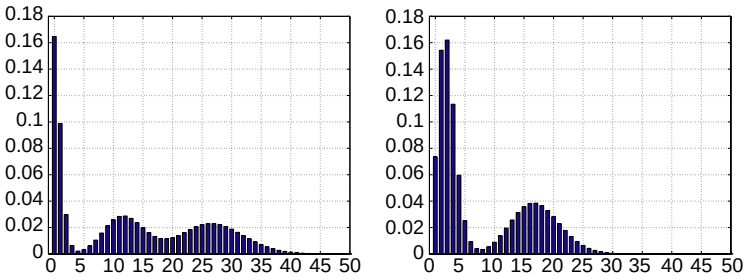

Fig. 1. The pmfs of two example mixtures of Poisson distributions.

we relabel the parameters of the model and let the joint distribution of X and Y be

$$P(X = x, Y = k) = \sum_{m=1}^{M} \pi_m p_m(k) \prod_{j=1}^{p} \phi(x_j \mid \lambda_{m,j}), \tag{2}$$

where $p_m(k)$ is a pmf for the class label Y when a mixture component is given. For a model equivalent to (1), we have $p_m(k) = 1$ if component m is from class k and 0 otherwise. Let $\bar{R}_k = \sum_{k'=1}^{k} R_{k'}$, $\bar{R}_0 = 0$. We have $M = \bar{R}_K$. Let the set $\mathscr{R}_k = \{\bar{R}_{k-1} + 1, \bar{R}_{k-1} + 2, \ldots, \bar{R}_k\}$. Without loss of generality, assume that $p_m(k) = 1$ if $m \in \mathscr{R}_k$ and 0 otherwise. That is, the $(\bar{R}_{k-1} + 1)$th to $\bar{R}_k$th components correspond to the R_k mixture components of class k. The associated class of component m is denoted by $b(m)$. Apparently, $b(m) = k$ if $p_m(k) = 1$. A one to one mapping between π_m and π_{kr}, a_k exists: $a_k = \sum_{m \in \mathscr{R}_k} \pi_m$, $\pi_{kr} = \pi_{\bar{R}_{k-1}+r}/a_k$.

The Poisson distribution is employed to model each component because of its ease of estimation and unbounded support. The additive combination of Poissons increases model flexibility and allows the true distribution to be better approximated. For instance, a mixture of Poissons can be multi-modal, while a Poisson distribution is always uni-modal. Fig. 1 shows the pmfs of two example mixtures of Poissons. According to (2), when the identity of a mixture component m is given, the variables are independent. This is, however, not true in general if only the class label is given because one class may contain multiple components.

We introduce clustering structures for the variables (i.e., words) next. In order not to confuse with the clustering structures for the documents indicated by the mixture components, we will obey the following convention: "cluster" is a simplified reference to "word cluster" (or variable cluster); and a "component" means a component in the mixture model, representing a "mode" of documents (or samples). The clustering structure of the variables is explored by constraining the parameters of the variables in the same cluster. For each class k, suppose the variables are clustered into L groups. The cluster identity of a variable j in class k is denoted by $c(k, j) \in \{1, 2, \ldots, L\}$, $k = 1, 2, \ldots, K$, $j = 1, 2, \ldots, p$, which is referred to as the *cluster assignment function*. For brevity, the same number of clusters is used for all the classes here. We refer to this model as the two-way mixture model due to its characteristic of simultaneous variable clustering. Suppose mixture component m is from class k, i.e., $b(m) = k$, and variable j_1 and j_2 are grouped into the same cluster in the class k, i.e., $c(k, j_1) = c(k, j_2)$, it is then assumed that $\lambda_{m,j_1} = \lambda_{m,j_2}$. There are two aspects implied by the assumption: (1) within each class, the clustering structures of the

variables are the same across all the components of the class; (2) within each component, variables in the same cluster are merged and have equal Poisson parameters while across different components of the same class, the Poisson parameters of any given variable can be different. Define the unique value of $\lambda_{m,j}$ for variable j's in the same cluster by $\theta_{m,l}$, where l is the cluster identity of j. Specifically, $l = c(k, j)$, where $k = b(m)$ is the class to which component m belongs to. Model (2) acquires the following parsimonious version with ML Poisson parameters:

$$P(X = x, Y = k) = \sum_{m=1}^{M} \pi_m p_m(k) \prod_{j=1}^{p} \phi(x_j \mid \theta_{m,c(b(m),j)}) \ . \tag{3}$$

To classify a sample $X = x$, the Bayes rule $\hat{y} = \arg\max_k P(Y = k \mid X = x) = \arg\max_k P(X = x, Y = k)$ is used.

3. Dimension reduction

The parameter constraint imposed on variables in the same cluster implies that these variables have the same marginal distribution within each mixture component, and consequently within each class as well. More importantly, Model (3) implies that for each cluster, only one statistic of the variables in this cluster is needed for the purpose of predicting the class label Y. The variable clustering leads to dimension reduction in a precise sense, as stated by the following proposition.

Proposition 3.1. Let $\bar{X}_{k,l} = \sum_{j:c(k,j)=l} X_j$, $k = 1, \ldots, K$, $l = 1, \ldots, L$. Given $\bar{X}_{k,l}$, $k = 1, \ldots, K$, $l = 1, \ldots, L$, the class Y is conditionally independent of $X_1, X_2, \ldots, X_p$.

According to the proposition, the sums of variables in all the clusters are sufficient for predicting the class label. The dimension of the sufficient statistics is KL. It will be shown in the experiment section that similar or even considerably better document classification can often be achieved with $KL \ll p$. For certain applications, if it is desirable to cluster the variables in the same way for all the classes, i.e., $c(k, l)$ is fixed over k, the dimension sufficient for predicting Y is L since $\bar{X}_{k,l}$'s are identical for different k's.

We now prove Proposition 3.1. Denote the number of variables in cluster l of class k by $\eta_{k,l}$, $\sum_{l=1}^{L} \eta_{k,l} = p$, for all k. Suppose variables in cluster l of class k are $\{j_1^{(k,l)}, j_2^{(k,l)}, \ldots, j_{\eta_{k,l}}^{(k,l)}\}$. Model (3) can be rewritten as

$$\begin{aligned} P(X = x, Y = k) &= \sum_{m=1}^{M} \pi_m p_m(k) \prod_{j=1}^{p} \phi(x_j \mid \theta_{m,c(b(m),j)}) \\ &= \sum_{m \in \mathscr{R}_k} \pi_m \prod_{l=1}^{L} \prod_{i=1}^{\eta_{k,l}} \phi(x_{j_i^{(k,l)}} \mid \theta_{m,l}). \end{aligned} \tag{4}$$

Recall that $p_m(k) = 1$ if $m \in \mathscr{R}_k$ and 0 otherwise. By the definition in Proposition 3.1, $\sum_{i=1}^{\eta_{k,l}} X_{j_i^{(k,l)}} = \bar{X}_{k,l}$. For any given mixture component $m \in \mathscr{R}_k$, $\{X_{j_i^{(k,l)}} : i = 1, \ldots, \eta_{k,l}\}$

are i.i.d. Poisson random variables. By the property of independent Poisson random variables, the sum $\bar{X}_{k,l}$ follows a Poisson distribution with mean $\eta_{k,l}\theta_{m,l}$; and the condition distribution of $\{X_{j_i^{(k,l)}} : i = 1, \ldots, \eta_{k,l}\}$ given $\bar{X}_{k,l} = \bar{x}_{k,l}$ is a multinomial distribution: $M(\frac{1}{\eta_{k,l}}, \frac{1}{\eta_{k,l}}, \ldots, \frac{1}{\eta_{k,l}}; \bar{x}_{k,l})$. Consequently,

$$\prod_{i=1}^{\eta_{k,l}} \phi(x_{j_i^{(k,l)}} \mid \theta_{m,l}) = \phi(\bar{x}_{k,l} \mid \eta_{k,l}\theta_{m,l}) \frac{\bar{x}_{k,l}!}{\prod_{i=1}^{\eta_{k,l}} x_{j_i^{(k,l)}}!} \left(\frac{1}{\eta_{k,l}}\right)^{\bar{x}_{k,l}}. \tag{5}$$

Substitute (5) into (4)

$$\begin{aligned} P(X{=}x, Y{=}k) &= \sum_{m \in \mathscr{R}_k} \pi_m \left[\prod_{l=1}^{L} \phi(\bar{x}_{k,l} \mid \eta_{k,l}\theta_{m,l}) \bar{x}_{k,l}! \left(\frac{1}{\eta_{k,l}}\right)^{\bar{x}_{k,l}}\right] \left[\prod_{l=1}^{L} \prod_{i=1}^{\eta_{k,l}} \frac{1}{x_{j_i^{(k,l)}}!}\right] \\ &= \left[\sum_{m \in \mathscr{R}_k} \pi_m \prod_{l=1}^{L} \phi(\bar{x}_{k,l} \mid \eta_{k,l}\theta_{m,l}) \bar{x}_{k,l}! \left(\frac{1}{\eta_{k,l}}\right)^{\bar{x}_{k,l}}\right] \left[\prod_{j=1}^{p} \frac{1}{x_j!}\right]. \end{aligned}$$

We thus have

$$P(Y = k \mid X = x) \propto \sum_{m \in \mathscr{R}_k} \pi_m \prod_{l=1}^{L} \phi(\bar{x}_{k,l} \mid \eta_{k,l}\theta_{m,l}) \bar{x}_{k,l}! \left(\frac{1}{\eta_{k,l}}\right)^{\bar{x}_{k,l}},$$

subject to $\sum_{k=1}^{K} P(Y = k \mid X = x) = 1$. As the posterior probability of Y given $X = x$ only depends on $\bar{x}_{k,l}$, $k = 1, \ldots, K$, $l = 1, \ldots, L$, X and Y are conditionally independent given $\bar{X}_{k,l}$, $k = 1, \ldots, K$, $l = 1, \ldots, L$.

4. Model estimation

For Model (3), we need to estimate the prior probabilities of the mixture components π_m, the Poisson parameters $\theta_{m,l}$, $m{=}1, \ldots, M$, $l{=}1, \ldots, L$, and the cluster assignment function $c(k, j) \in \{1, 2, \ldots, L\}$, $k{=}1, \ldots, K$, $j{=}1, \ldots, p$, which determines how the variables are clustered in each class. Denote the collection of parameters, including the cluster assignment function c by ψ. The EM algorithm for the maximum-likelihood estimation performs in each iteration the following two steps. The estimation of ψ at iteration t is denoted by ψ_t: $\psi_t = \{\pi_m^{(t)}, \theta_{m,l}^{(t)}, c^{(t)}(k, j) : m = 1, \ldots, M, l = 1, \ldots, L, k = 1, \ldots, K, j = 1, \ldots, p\}$. Let the training data be $\{(x^{(i)}, y^{(i)}) : i = 1, \ldots, n\}$.

1. *E-step*: Compute the posterior probability, $q_{i,m}$, of each sample i belonging to component m

$$q_{i,m} \propto \pi_m^{(t)} p_m(y^{(i)}) \prod_{j=1}^{p} \phi(x_j^{(i)} \mid \theta^{(t)}_{m,c^{(t)}(b(m),j)}), \quad \text{subject to } \sum_{m=1}^{M} q_{i,m} = 1.$$

2. *M-step*: Update ψ_{t+1} by maximizing the objective function

$$Q(\psi_{t+1}|\psi_t)=\max_{\psi'}\sum_{i=1}^{n}\sum_{m=1}^{M} q_{i,m}\log\left(\pi'_m p_m(y^{(i)})\prod_{j=1}^{p}\phi(x_j^{(i)}\mid\theta'_{m,c'(b(m),j)})\right). \tag{6}$$

As described in Section 2, R_k components are used to model class k, where R_k is pre-selected. Components in $\mathscr{R}_k=\{\bar{R}_{k-1}+1,\ldots,\bar{R}_k\}$ belong to class k. The values of $p_m(k)$ as well as the associated class of a component m, denoted by $b(m)$, are fixed once R_k's are selected. Since $p_m(k)=0$ if $b(m)\neq k$, $q_{i,m}=0$ if $y^{(i)}\neq b(m)$. Eq. (6) and those below are written with the assumption $0\log(0)=0$.

Since (6) can be further written as

$$\begin{aligned}Q(\psi_{t+1}|\psi_t)=&\sum_{i=1}^{n}\sum_{m=1}^{M} q_{i,m}\log\pi_m^{(t+1)}+\sum_{i=1}^{n}\sum_{m=1}^{M} q_{i,m}\sum_{j=1}^{p}\log\phi(x_j^{(i)}\mid\theta^{(t+1)}_{m,c^{(t+1)}(b(m),j)})\\&+\sum_{i=1}^{n}\sum_{m=1}^{M} q_{i,m}\log\ p_m(y^{(i)})\end{aligned} \tag{7}$$

the optimal $\pi_m^{(t+1)}$ are analytically given by

$$\pi_m^{(t+1)}=\frac{\sum_{i=1}^{n} q_{i,m}}{\sum_{m'=1}^{M}\sum_{i=1}^{n} q_{i,m'}},\quad m=1,\ldots,M. \tag{8}$$

The optimization of $\theta_{m,l}^{(t+1)}$, $m=1,\ldots,M$, $l=1,\ldots,L$, and $c^{(t+1)}(k,j)$, $k=1,\ldots,K$, $j=1,\ldots,p$, requires a numerical procedure. One approach is to alternatively optimize them with one fixed in each turn. Note that $Q(\psi_{t+1}\mid\psi_t)$ depends on $c^{(t+1)}$ and $\theta_{\cdot,\cdot}^{(t+1)}$ only through the second term in (7).

For a given $c^{(t+1)}$, the following $\theta_{m,l}^{(t+1)}$ maximize $Q(\psi_{t+1}\mid\psi_t)$:

$$\theta_{m,l}^{(t+1)}=\frac{\sum_{i=1}^{n} q_{i,m}\sum_{j:c(b(m),j)=l} x_j^{(i)}}{\eta_{b(m),l}\sum_{i=1}^{n} q_{i,m}}, \tag{9}$$

where $\eta_{k,l}$ is the number of j's such that $c(k,j)=l$. The proof of the optimality of Equation (9) is in the appendix. To avoid computational difficulty caused by $\theta_{m,l}^{(t+1)}$ being zero, a constant can be added in the numerator in (9) to offset the value from zero. When n is large, the adjusted estimation approaches (9).

For fixed $\theta_{\cdot,\cdot}^{(t+1)}$, $Q(\psi_{t+1}\mid\psi_t)$ is maximized by optimizing the cluster assignment function $c^{(t+1)}(k,j)$ separately for each class k and each variable j because the second term in Eq. (7) equals $\sum_{k=1}^{K}\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{m\in\mathscr{R}_k} q_{i,m}\ \log\ \phi(x_j^{(i)}\mid\theta^{(t+1)}_{m,c^{(t+1)}(k,j)})$. A straightforward

search is used to optimize $c^{(t+1)}(k, j)$

$$c^{(t+1)}(k, j) = \arg\max_{l} \sum_{i=1}^{n} \sum_{m \in \mathscr{R}_k} q_{i,m} \log \phi(x_j^{(i)} \mid \theta_{m,l}^{(t+1)})$$
$$= \arg\max_{l} \sum_{i=1}^{n} \sum_{m \in \mathscr{R}_k} q_{i,m} (x_j^{(i)} \log \theta_{m,l}^{(t+1)} - \theta_{m,l}^{(t+1)}). \quad (10)$$

The second step comes from the substitution of the Poisson pmf for ϕ.

It is computationally intensive to embed an iterative procedure within each M-step of EM. When it is numerically difficult to perform the M-step, the generalized EM (GEM) algorithm is suggested (Dempster et al., 1977). GEM computes a ψ_{t+1} that satisfies $Q(\psi_{t+1} \mid \psi_t) \geqslant Q(\psi_t \mid \psi_t)$, but ψ_{t+1} does not necessarily maximize $Q(\psi' \mid \psi_t)$ over ψ'. It is shown in (Dempster et al., 1977) that for any such ψ_{t+1}, the log likelihood under ψ_{t+1} is greater than or equal to that under ψ_t (the algorithm is thus ascending). Equality cannot occur if $Q(\psi_{t+1} \mid \psi_t) > Q(\psi_t \mid \psi_t)$. The GEM approach is taken in our estimation algorithm. In particular, we initialize $c^{(t+1)}$ by $c^{(t)}$ and then alternate the optimization of $\theta_{\cdot,\cdot}^{(t+1)}$ and $c^{(t+1)}$ only once using Eqs. (9) and (10). Let $\tilde{\psi} = (\pi_{\cdot}^{(t+1)}, \theta_{\cdot,\cdot}^{(t+1)}, c^{(t)})$.Note $\psi_t = (\pi_{\cdot}^{(t)}, \theta_{\cdot,\cdot}^{(t)}, c^{(t)})$. According to the optimality of Eqs. (8)–(10), we have $Q(\psi_{t+1} \mid \psi_t) \geqslant Q(\tilde{\psi} \mid \psi_t) \geqslant Q(\psi_t \mid \psi_t)$. The amount of computation in each iteration of the GEM is linearly proportional to *npML*.

To initialize the estimation algorithm, we randomly assign each sample to a mixture component m that belongs to the given class of the sample. The posterior probability $q_{i,m}$ is set to 1 if sample i is assigned to component m and 0 otherwise. Each value of the cluster assignment function $c(k, j)$, $k = 1, \ldots, K$, $j = 1, \ldots, p$, is randomly set to a number in $\{1, \ldots, L\}$. In our current implementation, we start with the same variable partition for all the classes, i.e., $c(k, j)$'s are initially identical over k. With the initial posterior probabilities and the cluster assignment function, an M-step is performed to obtain the initial parameters. The EM iterations then follow. The pseudo-code for the estimation algorithm is provided in the appendix.

5. Experiments

5.1. Data set

We perform experiments on the newsgroup data (Lang, 1995). In this data set, there are 20 topics, each containing about 1000 documents (email messages). Fourteen topics listed in Table 1 are used in our experiments. We used the *bow* toolkit to process this dataset. Specifically, we used the tokenization option so that the UseNet headers were stripped; and we also applied stemming (McCallum, 1996). Some of the newsgroups have large overlaps, for example, the five newsgroups *comp.** about computers. In fact several articles are posted to multiple newsgroups. Roughly half of the documents on each topic are randomly selected as training samples and the rest test samples.

Table 1
The 14 topics used in the experiments

1	*comp.sys.ibm.pc.hardware*	*comp.sys.mac.hardware*
2	*comp.os.ms-windows.misc*	*comp.windows.x*
3	*alt.atheism*	*soc.religion.christian*
4	*sci.med*	*sci.space*
5	*talk.politics.guns*	*talk.politics.mideast*
6	*rec.sport.baseball*	*rec.sport.hockey*
7	*rec.autos*	*rec.motorcycles*

5.2. Pre-selection of words

The total number of stemmed words in the newsgroup of 20 topics is 77, 952. In our experiments, to classify a set of topics, we pre-selected words to include in the document word count vector. One reason is to reduce the amount of computation. A second reason is that many words are related only to certain topics and are barely useful for the topics to be classified. Nevertheless, the number of words we used in classification is always substantially larger than the number of training documents. Several feature selection methods are discussed in Yang and Pedersen (1997) based on various heuristics such as deleting words that occur less than a certain number of times in the dataset; deleting words that occur in less than a certain number of documents in the dataset; and several mutual information based selection schemes.

In this paper, we use the following approach for feature selection. Let the training data set in an experiment be $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(n)}, y^{(n)})\}$. The vector $x^{(i)}$ is originally of dimension $\bar{p}=77,952$, including the counts of all the words. The topic class of document i is $y^{(i)} \in \{1, \ldots, K\}$. For each word j, $j = 1, \ldots, \bar{p}$, compute its total counts within each class k, $k = 1, \ldots, K$. Denote the total count of word j in class k by $\delta_{j,k}$. Then $\delta_{j,k} = \sum_{i=1}^{n} x^{(j)} I(y^{(i)} = k)$. As usual, $I(\cdot)$ is the indicator function that equals 1 when the argument is true and 0 otherwise. The variance of $\delta_{j,k}$ over class label k for each word j is computed. Specifically, the variance $\sigma_j^2 = \sum_{k=1}^{K} (\delta_{j,k} - \bar{\delta}_j)^2 / K$, where $\bar{\delta}_j = \sum_{k=1}^{K} \delta_{j,k} / K$ is the mean of $\delta_{j,k}$. A large value of the variance indicates that the overall counts of a word in different classes vary in a large range. The word is thus of high potential for distinguishing the classes. For each set of topics under classification, we choose a given number of words that have the highest σ_j^2 computed from the training data.

5.3. Classification results

The two-way mixture model approach is applied to both binary and multi-class classification. We first present results on binary classification of seven pairs of topics, each listed in one row of Table 1. The two topics selected in every pair are content-wise most close to each other among all the topics in the collection, as suggested by the topic names in Table 1. For each of the seven classification data sets, a document is represented by a vector containing the counts of 5000 words selected using the method described previously. The number of samples in each training set is around 1000, so is that in each testing set.

Table 2
The binary classification error rates in percent achieved by the two-way mixture model and SVM for the seven pairs of topics

%	$M=10$ $L=10$	$M=10$ $L=20$	$M=10$ $L=30$	$M=10$ n.v.c.	$M=20$ $L=10$	$M=20$ $L=20$	$M=20$ $L=30$	$M=20$ n.v.c.	SVM
1	11.09	11.39	9.70	12.87	11.58	10.59	10.59	**8.02**	10.40
2	9.28	10.19	9.17	**8.87**	9.38	9.28	10.30	13.25	11.93
3	6.08	4.99	5.48	6.88	5.78	5.18	**4.39**	6.08	5.08
4	4.20	3.40	**2.90**	4.60	3.50	3.40	3.50	3.70	3.80
5	**1.96**	2.64	2.35	6.07	2.06	**1.96**	2.55	4.02	2.74
6	3.21	2.80	4.25	3.83	3.21	**2.70**	3.11	7.98	2.80
7	5.79	5.89	5.69	4.90	5.59	**4.50**	5.09	7.99	4.70

For the two-way mixture model, the number of mixture components $M=10$, 20. For each value of M, different numbers of variable clusters are tested. The number of variable clusters $L=10$–30. Classification is also performed using the mixture model without variable clustering, denoted by "n.v.c.". The minimum error rate achieved for each pair of topics is in bold font.

Classification results have been obtained by estimating two-way mixture models with different numbers of components M and different numbers of variable clusters L. Table 2 lists the percentages of mis-classified test samples obtained with $M=10$, 20 and $L=10$, 20, 30. The number of mixture components M is evenly divided to each class. For instance, when $M=10$, both classes have $R_k=5$ components. Classification error rates obtained by SVM are listed in the table for comparison. For SVM classification, we used the SVM-Light program (Joachims, 2002) and the linear kernel with different values of the penalty parameter C. In Table 2, only the results obtained from the default selection of C are reported because different values of C lead only to slight changes and the default selection yields the best result for some data sets. We have also tested SVM-Light applying to the normalized document vectors, but the results are similar and therefore are not presented here. To explore the effect of variable clustering on classification, error rates based on mixtures of Poisson distributions without variable clustering are also computed.

For all the seven data sets, the lowest error rates listed in Table 2 are achieved by a mixture model. Except for the first two data sets which contain topics related to computer, variable clustering results in better classification. According to Proposition 3.1, data sets 3–7 can be better classified using no more than 60 ($KL \leqslant 60$) dimensions, significantly smaller than the original dimension 5000. For the two data sets that are best classified without variable clustering, a lack of word clusters is indicated. It is difficult to know a priori whether variable clustering yields better or worse classification. However, a data driven method such as cross validation can be applied to decide which is preferred. If homogeneous word clustering is enforced across different classes, classification accuracy is usually worse than that achieved by inhomogeneous clustering, as shown by the error rates listed in Table 3.

As indicated by the results in Table 2, classification accuracy varies with different M and L although only marginally across many values. It is also observed that variation in performance is more prominent in the low value range of M and L. Hence, a set of candidate M and L can be formed by sampling smaller values more densely, e.g., a grid on a log

Table 3
The binary classification error rates in percent achieved by the two-way mixture model with identical word clustering for the two classes

%	$M = 10$ $L = 10$	$M = 10$ $L = 20$	$M = 10$ $L = 30$	$M = 20$ $L = 10$	$M = 20$ $L = 20$	$M = 20$ $L = 30$
1	18.12	14.55	14.26	17.03	16.44	**14.16**
2	13.35	11.21	**9.79**	14.58	12.03	12.95
3	6.18	**5.58**	**5.58**	6.78	6.58	5.68
4	6.21	4.70	5.31	**3.50**	4.50	3.90
5	4.51	3.82	3.43	2.84	3.62	**2.15**
6	5.60	5.08	5.29	4.46	4.56	**4.25**
7	8.40	**7.19**	7.69	7.89	8.59	8.59

The number of mixture components $M = 10, 20$. The number of variable clusters $L = 10$–30. The minimum error rate achieved for each pair of topics is in bold font.

scale. Cross validation can then be applied to select the values of M and L. Unlike many situations where physical constraints dictate the existence of a certain unique *intrinsic* dimension and/or number of clusters/classes, for modeling collections of documents, the number of clusters is far less well-determined: within a certain range, the topical structures can be equally well captured by clustering with several different cluster numbers. Choosing the number of clusters from a purely statistical perspective is a difficult problem and is out of the scope of this paper. Effort in this direction has been made by Tibshirani et al. (2002).

If a single component is used to model each class and variable clustering is not performed, classification based on the mixture model is essentially the naive Bayes algorithm with each dimension modeled by a Poisson distribution. For all the seven data sets we tested above, classification accuracy is improved by having multiple components for each class. Fig. 2 shows the error rates obtained for four data sets using mixture models with different numbers of components. Variables are not clustered in these models. Results obtained at $M=2$ correspond to modeling each class by a single Poisson distribution in every dimension. Except for the data set containing class *rec.autos* and *rec.motorcycles*, there is a marked drop of classification error rate when M begins to increase. This indicates that the diversity of documents in one class usually demands several component Poisson distributions to capture. We expect that a collection of documents not possessing multiple modes content-wise, i.e., focusing exclusively on one topic, tends to exhibit a similar pattern to that of the *rec.autos* vs. *rec.motorcycles* data set.

To test multi-class classification performance, we form a data set using the following 8 topics: *comp.os.ms-windows.misc*, *comp.windows.x*, *alt.atheism*, *soc.religion.christian*, *sci.med*, *sci.space*, *talk.politics.guns*, *talk.politics.mideast*. There are 3960 randomly selected documents in the training set. Every class covers roughly 500 samples. The test data have 4004 samples. By the word pre-selection method, 10000 words are chosen to represent these documents. When the number of components $M = 100$, the classification error rates achieved by having $L = 5, 10, 20$ variable clusters are 8.92%, 9.02%, and 7.84% respectively. If variable clustering is not used, the mixture model with 100 components

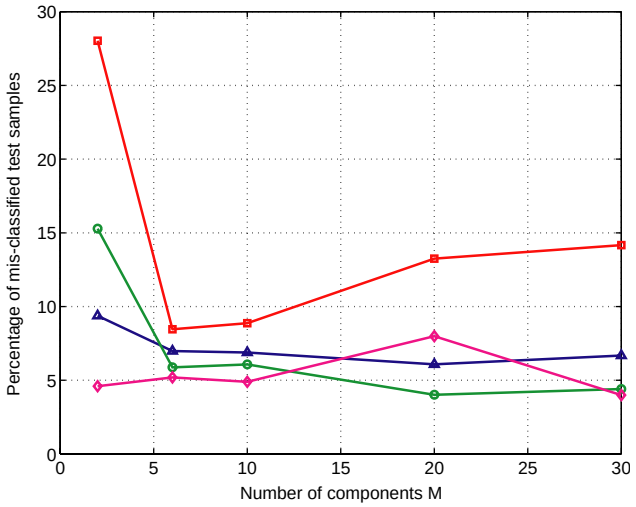

Fig. 2. The percentages of mis-classified test samples obtained for four data sets using mixture models with different numbers of components. The number of mixture components ranges from 2 to 30. Square: data set 2, *comp.os.ms-windows.misc* vs. *comp.windows.x*; Triangle: data set 3, *alt.atheism* vs. *soc.religion.christian*; Circle: data set 5, *talk.politics.guns* vs. *talk.politics.mideast*; Diamond: data set 7, *rec.autos* vs. *rec.motorcycles*.

Table 4
The confusion table for classifying the 8 topics: *comp.os.ms-windows.misc*, *comp.windows.x*, *alt.atheism*, *soc.religion.christian sci.med*, *sci.space*, *talk.politics.guns*, *talk.politics.mideast*

Rate (%)	*ms-win*	*win.x*	*atheism*	*christian*	*sci.med*	*sci.space*	*guns*	*mideast*
ms-win	84.27	13.15	0.22	0.22	0.65	1.51	0.00	0.00
win.x	4.84	91.10	0.00	0.00	1.16	2.32	0.19	0.39
atheism	0.40	0.20	93.37	3.61	1.20	0.40	0.20	0.60
christian	0.59	0.40	4.36	89.70	1.98	0.20	0.79	1.98
sci.med	0.99	0.99	0.60	0.00	93.44	2.39	0.99	0.60
sci.space	0.81	0.60	1.01	0.00	2.62	93.55	1.01	0.40
guns	0.20	0.00	0.59	0.40	0.79	0.20	95.65	2.17
mideast	0.58	0.58	0.78	0.58	0.78	0.19	0.97	95.53

A two-way mixture model with $M = 100$ components and $L = 20$ variable clusters is used. For each topic, the percentages of its test samples that are classified to the 8 topics, respectively, are listed in one row. Values on the diagonal indicate the classification accuracy.

yields an error rate of 13.64%, considerably higher than those obtained by employing clustering. For $M = 100$, $L = 20$, the classification result is provided in detail in Table 4. The confusion table shows that pairs of topics listed in one row of Table 1 are most likely to be confused with each other, e.g., *comp.os.ms-windows.misc* vs. *comp.windows.x*,

and *sci.med* vs. *sci.space*. This is anticipated since these pairs of topics are about similar subjects.

5.4. Word clustering

The parameters $\theta_{m,l}$, $m = 1, \ldots, M$, $l = 1, \ldots, L$ in the two-way mixture model summarize the high-dimensional document vectors X_i, $i = 1, \ldots, n$ across samples as well as variables. Each mixture component m can be regarded as one "mode" of X in a certain class k. Within one "mode" m, each dimension j of X follows a uni-modal distribution, specifically, a Poisson parameterized by its mean value $\theta_{m,l}$, where $l = c(k, j)$. The vector $(\theta_{m,c(k,1)}, \theta_{m,c(k,2)}, \ldots, \theta_{m,c(k,p)})^{\mathrm{T}}$ is the mean (centroid) of X generated in "mode" m. The representation of this centroid can be further simplified to $(\theta_{m,1}, \ldots, \theta_{m,L})^{\mathrm{T}}$ due to the clustering structure of variables embedded in the two-way mixture model. We can thus regard $(\theta_{m,1}, \ldots, \theta_{m,L})^{\mathrm{T}}$ as characteristic vectors for class k, where $m \in \mathscr{R}_k$ indexes mixture components of class k. As shown in the experiments, high classification accuracy can often be achieved with $L \ll p$. The characteristic vectors are hence of a much lower dimension than X. As there is no definition for "correct" word clusters from the perspective of document analysis, an intrinsically right number of word clusters does not exist. We will use a particular value of L to illustrate the summarization of data provided by the two-way mixture model.

We now use the data set containing topics *comp.os.ms-windows.misc* and *comp.windows.x* and the one containing *sci.med* and *sci.space* as examples to examine the characteristic vectors and word clusters. The two-way mixture models investigated have $M = 10$ components and $L = 30$ word clusters. Fig. 3(a) shows the number of words in each of the 30 word clusters for the class *comp.os.ms-windows.misc* and *comp.windows.x*, respectively. These word clusters are indexed in an order of descending sizes. The sizes of the word clusters are highly uneven although they are roughly equal by initialization. The largest word cluster accounts for about 30% of all the words. Fig. 3(b) shows two characteristic vectors for each class, which correspond to the two components with highest prior probabilities. Every curve in the plot shows how $\theta_{m,l}$ varies with the cluster index l for a fixed component m. Fig. 3(c) and (d) show the cluster sizes and the dominant characteristic vectors for *sci.med* and *sci.space*. Similarly, the largest cluster for each class accounts for a substantial percentage of words. For all the four topics, the average word counts $\theta_{m,l}$ tend to be small for words in large clusters. This indicates that for every topic, there are a large number of words that on average occur very infrequently.

As shown in Fig. 3, the 29th word cluster of *comp.os.ms-windows.misc* has the largest average counts for both dominant "modes". This cluster contains 11 words: *articl*, *card*, *driver*, *file*, *program*, *problem*, *run*, *system*, *window*, *work*, *write*. These words are generic ones a document related to computer is likely to contain. In fact, the last 8 words in this cluster all appear in the 30th cluster of the other class *comp.windows.x*, which also has the largest average counts for both dominant "modes". The word *microsoft* is contained in the 23rd cluster of *comp.os.ms-windows.misc*, which has the second largest average counts for both "modes". For class *comp.windows.x*, the word *microsoft* belongs to its 23rd cluster (by coincidence), which has rather low average counts for both "modes", as shown by the right panel of Fig. 3(b). It is not surprising that a document on ms-windows tends to

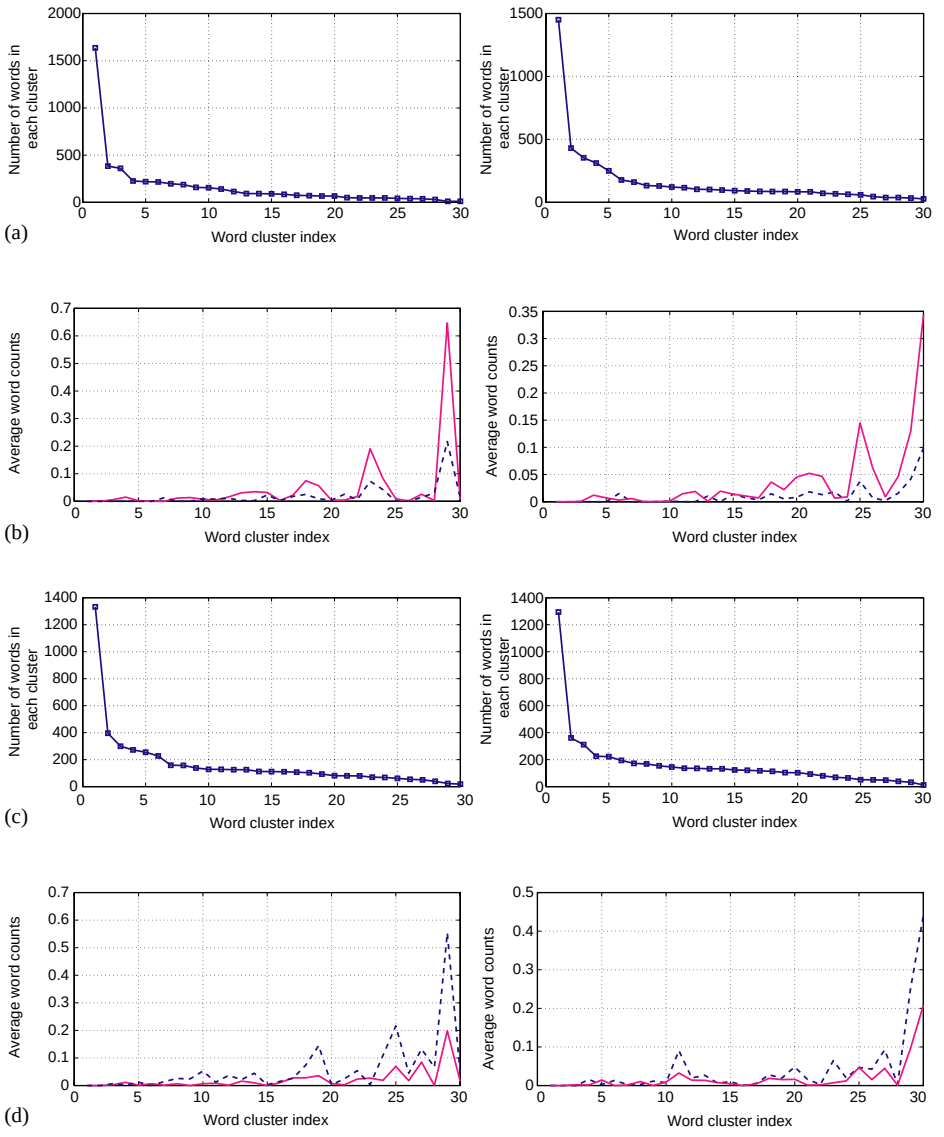

Fig. 3. The sizes of the word clusters and the characteristic vectors. (a): the sizes of the word clusters in the class *comp.os.ms-windows.misc* (left) and *comp.windows.x* (right); (b): the two dominant characteristic vectors for *comp.os.ms-windows.misc* and *comp.windows.x*, respectively; (c): the sizes of the word clusters in the class *sci.med* (left) and *sci.space* (right); (d): the two dominant characteristic vectors for *sci.med* and *sci.space*.

include *microsoft* much more frequently than one on x-windows. Similarly, the following words particularly related to x-windows: *openwindow*, *xdm*, *xterm*, *xview* are in a high count cluster (30) of *comp.windows.x*. For *comp.os.ms-windows.misc*, however, these words are in clusters (1 and 7) with very low average counts.

For either *comp.os.ms-windows.misc* or *comp.windows.x*, words in the largest cluster (cluster 1) have the lowest average counts in both of the corresponding dominant "modes". The largest cluster of *comp.os.ms-windows.misc* includes words hardly related to computer, e.g., *angel*, *christian*, *lake*, *sooner*, *wind*, *wood*, and words specific to x-windows but not ms-windows, e.g., *ximage*, *xloadimag*, *xfig*, *xlib*, *xstorecolor*, *xtiff*, *xmdrawingarea*, *xcopyarea*. Similarly, the largest cluster of *comp.windows.x* contains words unlikely to be associated with this topic, e.g., *gift*, *giant*, *mbyte*, *mpeg*, *microsofti*, *foreign*, *canon*, *sharp*, *warehous*.

As we have mentioned before that the word clustering structures for different classes can be different, and in our numerical experiments this approach was shown to perform better than using a homogeneous clustering structure across all the classes. Examination of some of the word clusters seems to indicate that this inhomogeneity of word clustering tends to disambiguate word senses to a certain extent. For example, the cluster containing the word *window* in the class *comp.os.ms-windows.misc* includes another 10 generic words related to computer systems, the cluster in the class *comp.windows.x* that contains *window* also includes words *font*, *motif*, *widget*, *xterm*, which are clearly x-window specific.

6. Conclusions

In this paper, a two-way mixture model of Poisson distributions is proposed for simultaneous document classification and word clustering. By employing multiple components for each document class, the distribution of document vectors can be better approximated. The issues of high dimensionality and sparsity of the document vectors are addressed by assuming a variable clustering structure in the mixture model. The two-way mixture model possesses a dimension reduction property. In particular, given the sum of counts for words in each cluster, the class label of a sample is conditionally independent of the document vector. A GEM algorithm is derived to estimate the model. Experiments have been performed with the newsgroup data set. Comparisons made with SVM demonstrate competitive performance. Word clustering is shown to improve classification in many cases. As the two-way mixture model leads to clustering across both documents and words, a large number of documents represented by high-dimensional vectors are summarized by a small number of low-dimensional vectors, which can be better visualized.

Acknowledgements

We are deeply grateful for the reviewer's insightful suggestions and detailed comments.

Appendix

We prove $\theta_{m,l}^{(t+1)}$, $m = 1, \ldots, M$, $l = 1, \ldots, L$, in Eq. (9) maximize $Q(\psi_{t+1} \mid \psi_t)$ when $c^{(t+1)}$ is given. According to (7), $Q(\psi_{t+1} \mid \psi_t)$ depends on $\theta_{\cdot,\cdot}^{(t+1)}$ only through the second

term. Hence, we only need to maximize the second term, which can be written as

$$\sum_{i=1}^{n}\sum_{m=1}^{M} q_{i,m}\sum_{j=1}^{p}\log\,\phi(x_j^{(i)}\mid\theta^{(t+1)}_{m,c^{(t+1)}(b(m),j)})$$
$$=\sum_{m=1}^{M}\sum_{l=1}^{L}\left(\sum_{i=1}^{n} q_{i,m}\sum_{j:c^{(t+1)}(b(m),j)=l}\log\,\phi(x_j^{(i)}\mid\theta^{(t+1)}_{m,l})\right). \tag{11}$$

To maximize (11), we can maximize the summand in the parenthesis by optimizing each $\theta_{m,l}$ separately, where $m=1,\ldots,M$, $l=1,\ldots,L$. Substitute in the pmf of the Poisson distribution:

$$\sum_{i=1}^{n} q_{i,m}\sum_{j:c^{(t+1)}(b(m),j)=l}\log\,\phi(x_j^{(i)}\mid\theta^{(t+1)}_{m,l})$$
$$=\sum_{i=1}^{n} q_{i,m}\sum_{j:c^{(t+1)}(b(m),j)=l}\left(x_j^{(i)}\log\,\theta^{(t+1)}_{m,l}-\theta^{(t+1)}_{m,l}-\log(x_j^{(i)}!)\right)$$
$$=-\left(\eta_{b(m),l}\sum_{i=1}^{n} q_{i,m}\right)\theta^{(t+1)}_{m,l}+\left(\sum_{i=1}^{n} q_{i,m}\sum_{j:c^{(t+1)}(b(m),j)=l} x_j^{(i)}\right)\log\,\theta^{(t+1)}_{m,l}$$
$$-\sum_{i=1}^{n} q_{i,m}\sum_{j:c^{(t+1)}(b(m),j)=l}\log(x_j^{(i)}!),$$

where $\eta_{k,l}$ is the number of j's such that $c(k,j)=l$. The above function is concave in $\theta^{(t+1)}_{m,l}$ and is maximized by setting the first derivative to zero

$$-\eta_{b(m),l}\sum_{i=1}^{n} q_{i,m}+\frac{1}{\theta^{(t+1)}_{m,l}}\sum_{i=1}^{n} q_{i,m}\sum_{j:c^{(t+1)}(b(m),j)=l} x_j^{(i)}=0.$$

Hence $\theta^{(t+1)}_{m,l}=\sum_{i=1}^{n} q_{i,m}\sum_{j:c(b(m),j)=l} x_j^{(i)}/\eta_{b(m),l}\sum_{i=1}^{n} q_{i,m}$. The optimality of Eq. (9) is proved.

Next, the pseudo-code for estimating the two-way mixture model is provided. The number of mixture components for each class R_k, $k=1,\ldots,K$, is pre-specified.

(1) *Initialization*
(a) *Set* $t=0$.
(b) *Let* $\bar{R}_0=0$, $\bar{R}_k=\sum_{k'=1}^{k} R_{k'}$, $\mathscr{R}_k=\{\bar{R}_{k-1}+1,\ldots,\bar{R}_k\}$, $k=1,\ldots,K$; $M=\sum_{k'=1}^{K} R_{k'}$.
(c) *For* $m=1,\ldots,M$, *find* k *such that* $m\in\mathscr{R}_k$. *Set* $b(m)=k$; $p_m(k)=1$ *and* $p_m(k')=0$ *for all* $k'\neq k$, $k'\in\{1,\ldots,K\}$.
(d) *Set the initial log likelihood of the data set* $\Gamma=0.0$.
(e) *For each training sample* $i=1,\ldots,n$, *let* $k=y^{(i)}$. *Randomly select a mixture component* m *from* $\mathscr{R}_k$. *Let* $q^{(t)}_{i,m}=1$ *and* $q^{(t)}_{i,m'}=0$ *for all* $m'\neq m$, $m'\in\{1,2,\ldots,M\}$.
(f) *For each variable* $j=1,\ldots,p$, *randomly select a number* l *from* $\{1,\ldots,L\}$, *and set* $c^{(t)}(k,j)=l$ *for all* $k\in\{1,\ldots,K\}$.

(2) *M-step*

(a) *M-stepUpdate* $\pi_m^{(t+1)}$, $m = 1, \ldots, M$: $\pi_m^{(t+1)} = \sum_{i=1}^n q_{i,m}^{(t)} / \sum_{m'=1}^M \sum_{i=1}^n q_{i,m'}^{(t)}$.

(b) *Update* $\theta_{m,l}^{(t+1)}$, $m = 1, \ldots, M$, $l = 1, \ldots, L$:

$$\theta_{m,l}^{(t+1)} = \frac{\sum_{i=1}^n q_{i,m}^{(t)} \sum_{j=1}^p x_j^{(i)} I(c^{(t)}(b(m), j) = l)}{\sum_{i=1}^n q_{i,m}^{(t)} \cdot \sum_{j=1}^p I(c^{(t)}(b(m), j) = l)}.$$

(c) *Update* $c^{(t+1)}(k, j)$, $k = 1, \ldots, K$, $j = 1, \ldots, p$,

$$c^{(t+1)}(k, j) = \arg\max_l \sum_{i=1}^n \sum_{m \in \mathscr{R}_k} q_{i,m}^{(t)} (x_j^{(i)} \log \theta_{m,l}^{(t+1)} - \theta_{m,l}^{(t+1)}).$$

(3) *E-step*: *update the posterior probabilities* $q_{i,m}^{(t+1)}$, $i = 1, \ldots, n$, $m = 1, \ldots, M$:

$$q_{i,m}^{(t+1)} = \frac{\pi_m^{(t+1)} p_m(y^{(i)}) \prod_{j=1}^p \phi(x_j^{(i)} \mid \theta_{m,c^{(t+1)}(b(m),j)}^{(t+1)})}{\sum_{m'=1}^M \pi_{m'}^{(t+1)} p_{m'}(y^{(i)}) \prod_{j=1}^p \phi(x_j^{(i)} \mid \theta_{m',c^{(t+1)}(b(m'),j)}^{(t+1)})}.$$

(4) *Let* $\Gamma' = \Gamma$. *Compute the new log likelihood of the data set under the updated parameters*:

$$\Gamma = \sum_{i=1}^n \log \left[\sum_{m=1}^M \pi_m^{(t+1)} p_m(y^{(i)}) \prod_{j=1}^p \phi(x_j^{(i)} \mid \theta_{m,c^{(t+1)}(b(m),j)}^{(t+1)}) \right].$$

(5) *Set* $t + 1 \to t$. *If* $t > 1$ *and the increase in the log likelihood is very small, specifically,* $\frac{\Gamma - \Gamma'}{|\Gamma|} < \varepsilon$, *go back to Step* 2. *Otherwise, stop.*

Essentially, the above algorithm estimates the mixture model of each class individually. However, if homogeneous word clustering across all the classes is enforced, i.e., $c(k, j)$ is fixed for different k's, the update of $c^{(t+1)}(k, j)$ in Step 2(c) should be modified to

$$c^{(t+1)}(k, j) = \arg\max_l \sum_{i=1}^n \sum_{m=1}^M q_{i,m}^{(t)} \left(x_j^{(i)} \log \theta_{m,l}^{(t+1)} - \theta_{m,l}^{(t+1)} \right),$$

and the estimation of the mixture models for the K classes cannot be separated.

References

Baker, L., McCallum, A., 1998. Distributional clustering of words for text classification. In: Proceedings of SIGIR-98.pp. 96–103.

Belew, R., 2000. Finding Out About: A Cognitive Perspective on Search Engine Technology and the WWW. Cambridge University Press, Cambridge.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. J. Mach. Learning Res. 3, 993–1022.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. vol. 39, no. 1, pp. 1–21.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learning J. vol. 42, pp. 177–196.

Joachims, T., 2002. Learning to Classify Text Using Support Vector Machines, Kluwer Academic Publishers, Boston.

Lang, K., 1995. Learning to filter net news. In: Proceeding of the International Conference Machine Learning. pp. 331–339.

McCallum, A., 1996. Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering, http://www.cs.cmu.edu/ mccallum/bow.

McCallum, A., Nigam, K., 1998. A comparison of event models for naive Bayes text classification. AAAI-98 Workshop on Learning for Text Categorization.

Salton, G., 1989. Automatic Text Processing, Addison-Wesley, New York.

Tibshirani, R., Walther, G., Botstein, D., Brown, P., 2002. Cluster validation by prediction strength. Manuscript, Stanford University.

Yang, Y., 1999. An evaluation of statistical approaches to text categorization. Inform. Retrieval 1, 69–90.

Yang, Y., Pedersen, J., 1997. A comparative study on feature selection in text categorization. In: Proceedings of ICML-97. pp. 412–420.

Zhang, T., 2001. Text classification based regularized linear classification methods. Inform. Retrieval 4, 5–31.

Zha, H., He, X., Ding, C., Gu, M., Simon, H., 2001. Bipartite graph partitioning and data clustering. In: Proceedings of ACM CIKM. pp. 25–32.