Probabilistic Text Categorization using Sparse Topical Encoding[*]

Xipeng Qiu[†] Xuanjing Huang, Lide Wu
School of Computer Science, Fudan University

Abstract: In this paper, we propose a topic-based probabilistic text categorization model, which can be decomposed into two steps. Firstly, we present sparse non-negative matrix factorization (SNMF) algorithm, which can extract the sparse topical encoding for documents automatically and has an intuitive interpretation. Secondly, we calculate the similarity between documents by integrating the probabilistic similarities of their topics, we can assign the different weight to each topic by additive logistic regression. We compare our method with some traditional text categorization approaches, and experimental results show our method is better than the others.

Keywords: Probabilistic text categorization, Sparse Topical Encoding, Sparse Non-Negative Matrix Factorization.

1 Introduction

Text categorization[1] is the problem of automatically classifying the natural language text to predefined categories based on their content. Text categorization plays an importance role in many applications, such as information retrieval[2] and Anti-Spam Email Filtering[3], etc.

In recent years, research interest in text categorization has been growing in machine learning, as well as in information retrieval, computational linguistics, and other fields. Statistical machine learning methods have shown great benefit in text categorization[4]. The top-performing learning methods include Naive Bayes[5], kNN[4], Support Vector Machines(SVM) [6] and maximum entropy methods(ME) [7]. However, when these traditional machine learning methods are applied to large-scale text categorization, a major characteristic, or difficulty, is the high dimensionality of the feature space. Dimensionality reduction has attracted much attention recently since it make the text categorization task more efficient and save more storage space[8].

Generally, the dimensionality reduction approaches can be classified into feature extraction (FE) and feature selection (FS)[9].

The traditional FE methods reduce the dimensionality of data by linear algebra transformations. FE methods have been proved to be very effective for dimensionality reduction. However, they fail to explain how their projected subspace is homologous with the word space.

On the other hand, FS methods reduce the dimensionality of the data by select features from the original vectors directly. Thus FS methods are preferred for dimensionality reduction problems of text categorization. In the text domain, the most popular used FS algorithms are still the traditional ones such as Document Frequency (DF), χ^2 statistics (CHI), Information Gain (IG)

[*] The research described in this paper has been supported by NSF of China(60435020)

[†] Corresponding author. Email:xpqiu@fudan.edu.cn

and Mutual Information (MI) , etc[10]. However, most of FS methods select the features based the words, so it is inefficient than the FE methods.

Recently, Lee *et al.* [11] proposed non-negative matrix factorization (NMF) for extracting the basis documents that correspond with intuitive notions of the topics of documents. Non-negative matrix factorization imposes the non-negativity constraints instead of the orthogonality. As the consequence, the entries of basis functions and encodings are all non-negative, and hence only non-subtractive combinations are allowed. Ding et al. [12] have claimed the NMF have an equivalence with probabilistic latent semantic indexing[13].

However, NMF do not always result the sparse encoding. Moveover, since that the basis functions are not orthogonal and may include some redundancies, we cannot calculate the similarity between the encodings by traditional similarity measure directly.

In this paper, we propose a sparse non-negative matrix factorization(SNMF) to extract the sparse and non-negative topical encodings. SNMF is an improved version of non-negative matrix factorization (NMF) in two ways. Firstly, we impose a weight for each word. Secondly, we redefine the optimization function of NMF to obtain the sparse encoding. Once sparse encoding is obtained, we calculate the similarity using probabilistic measure. We propose a topic-based probabilistic similarity measure, which estimate the posterior probability of intra-category variations by the forward additive logistic regression.

The rest of the paper is organized as follows: In Section 2 we propose the sparse non-negative matrix factorization, whose convergence is also proved. Then we describe the probabilistic similarity measure in Section 3. Section 4 gives the experimental comparisons among our method and some state-of-art text categorization methods. Finally, we give the conclusions in Section 5.

2 Sparse Non-Negative Topical Encoding

Let a set of m documents (each document is formulated in n-dimensional vector with vector space model) be given as an $n \times m$ matrix $V = [V_{i\mu}]$, with each column consisting of the n non-negative word values of an document. Denote a set of $r \leq n$ basis documents by an $n \times r$ matrix W. Each document can be represented as a linear combination of the basis documents using the approximate factorization $V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$. The dimensions of the factor matrices W and H are $n \times r$ and $r \times m$, respectively. Usually r is chosen to be smaller than m and n. The r columns of W are called basis documents. Each row of H is called an encoding and is in one-to-one correspondence with a document in V. An encoding consists of the coefficients by which a document is represented with a linear combination of basis topic.

In order to extract the sparse topical encoding for the documents, we propose a sparse non-negative matrix factorization (SNMF) algorithm, basis documents can correspond with intuitive notions of the topics of documents (non-negative). We first give a description to the original NMF algorithm.

2.1 Original NMF

To quantify the quality of the approximation $V \approx WH$, a cost function needs to be constructed with some measure of distance between two non-negative matrices A and B. NMF uses the divergence defined as

$$D(A||B) = \sum_{i,\mu} d_{i\mu} = \sum_{i,\mu} (A_{i\mu} \log \frac{A_{i\mu}}{B_{i\mu}} - A_{i\mu} + B_{i\mu}). \tag{1}$$

A NMF factorization is defined as

$$\begin{aligned} \min_{W,H} \quad & D(V||WH), \\ \text{s.t} \quad & W,H \geq 0, \textstyle\sum_i W_{ia} = 1. \end{aligned} \tag{2}$$

The objective function in Eq.(2) can be related to the likelihood of generating the documents in V from the basis W and encodings H. An iterative approach to reach a local maximum of this objective function is given by the multiplicative update rules [11]:

$$W_{ia} = W_{ia} \sum_{\mu} \frac{V_{i\mu}}{(WH)_{i\mu}} H_{a\mu}, \tag{3}$$

$$W_{ia} = \frac{W_{ia}}{\sum_j W_{ja}}, \tag{4}$$

$$H_{a\mu} = H_{a\mu} \sum_{i} W_{ia} \frac{V_{i\mu}}{(WH)_{i\mu}}. \tag{5}$$

Initialization is performed using positive random initial conditions for matrices W and H. The convergence of the process is also ensured[14].

2.2 Sparse Non-Negative Matrix Factorization

In general sense, a topic can be regarded as a basis document whose word values are positive in some localized region and are zeros elsewhere. However, NMF does not guarantee its basis documents are localized, which conflict with our nature concept of "topics". In order to obtain localized representation, Li *et al.* [15] proposed local NMF, but there are too much approximate in their proof of convergence and LNMF has a high reconstruction error. Hoyer [16] also proposed local NMF with constraints to the specific sparseness. They all give some localized constraint on the basis documents.

Here we improve NMF to obtain the localized basis documents from the different point of view. Without adding constraint on basis documents directly, we seek the sparse topical encoding for document, which is different with the methods in [15] and [16].

Olshausen [17] have shown that when one seeks a sparse, linear decomposition of objects, the basis functions that emerge are spatially localized, oriented, and bandpass (selective to structure at different spatial scales), comparable to the basis functions of wavelet transform.

The search for a sparse encoding can be formulated as an optimization problem by constructing the following cost function to be minimized:

$$E = [\text{reconstruct error}] - \lambda[\text{sparseness of encoding}], \tag{6}$$

where λ is a positive constant that determines the importance of the second term relative to the first. The first term measures how well the code describes the document, and we evaluate it by Eq. (1). The second term assesses the sparseness of the encoding for a given document by assigning a cost depending on how activity is distributed among the coefficients. Olshausen [17] constructed a series criterion by a function which takes the sum of each coefficient's activity passed through a nonlinear function $S(x)$:

$$[\text{sparseness of } x] = -\sum_i S(\frac{x_i}{\delta}), \tag{7}$$

where δ is scaling constant. $S(x)$ can be chosen as $-e^{-x^2}$, $log(1+x^2)$ or $|x|$, and they all favor among activity states with equal variance those with the fewest number of non-zero coefficients.

For extracting more meaningful sparse topical encoding, we improve the NMF by two observations:

1. Since different word values have different variances for a given set of documents, it's reasonable that the different words should have different contributions in evaluating the divergence between the document and its reconstruction. So we impose a weight to each word $w_i = \sigma_i^2$, where σ_i is the standard deviation of the observed values corresponding to the word i. We normalize $\sum_i w_i = 1$, and the weighted divergence $D_w(A||B) = \sum_{i,\mu} w_i d_{i\mu}$. Thus, reconstructed documents put emphases on the more important words.

2. For computing efficiency, we use $|x|^2$ as $S(x)$ to evaluate the sparseness of encoding. Due to each column of H represents the encoding coefficients. Based on Eq.(7), sparseness of encoding is imposed by $-\sum_a H_{a\mu}^2$.

Based on the above two modifications, the objective function of sparse NMF is as follow:

$$\begin{aligned}\bar{D}(V||WH) &= \sum_{i,\mu} w_i d_{i\mu} + \alpha \sum_{\mu}\sum_{a} H_{a\mu}^2 \\ &= \sum_{i,\mu} w_i (V_{i\mu} \log \frac{V_{i\mu}}{WH_{i\mu}} - V_{i\mu} + WH_{i\mu}) \\ &+ \alpha \sum_{\mu}\sum_{a} H_{a\mu}^2, \qquad (8)\end{aligned}$$

A local minimum solution to $\bar{D}(V||WH)$ can be obtained by using the following three update rules:

$$H_{a\mu} = \frac{-\sum_i w_i W_{ia} + \sqrt{(\sum_i w_i W_{ia})^2 + \beta}}{4\alpha}, \qquad (9)$$

$$W_{ia} = W_{ia} \frac{\sum_\mu V_{i\mu} H_{a\mu}/(WH)_{i\mu}}{\sum_\nu H_{a\nu}}, \qquad (10)$$

$$W_{ia} = \frac{W_{ia}}{\sum_k W_{ka}}. \qquad (11)$$

where $\beta = 8\alpha H_{a\mu} \sum_i w_i W_{ia} V_{i\mu}/(WH)_{i\mu}$.

2.3 Proof of Convergence

To minimize Eq.(8), we will make use of an auxiliary function similar to that used in the Expectation-Maximization algorithm [18]. $G(x, x')$ is an auxiliary function of $F(x)$ if the conditions $G(x, x') \geq F(x)$ and $G(x, x) = F(x)$ are satisfied. Thus $F(x)$ is non-increasing when x is updated using $x^{t+1} = \arg\min_x G(x, x^t)$. This is because $F(x^{t+1}) \leq G(x^{t+1}, x^t) \leq G(x^t, x^t) = F(x^t)$.

Updating H: H is updated by minimizing $F(H) = D(V||WH)$ with W fixed. We need construct an auxiliary function $G(H, H^t)$ and H is updated by $H^{t+1} = \arg\min_H G(H, H^t)$.

$G(H, H^t)$ is constructed as

$$\begin{aligned} G(H, H^t) &= \sum_{i,\mu} w_i (V_{i\mu} \log V_{i\mu} - V_{i\mu}) + \sum_{i,\mu,a} w_i W_{ia} H_{a\mu} \\ - & \sum_{i,\mu,a} w_i \Big(V_{i\mu} \frac{W_{ia} H^t_{a\mu}}{\sum_b W_{ib} H^t_{b\mu}} (\log W_{ia} H_{a\mu} - \log \frac{W_{ia} H^t_{a\mu}}{\sum_b W_{ib} H^t_{b\mu}}) \Big) \\ + & \alpha \sum_\mu \sum_a H^2_{a\mu}. \end{aligned} \tag{12}$$

It's straightforward to verify that $G(H, H) = F(H)$. To show that $G(H, H^t) \leq F(H)$, we use convexity of the log function to derive the inequality

$$-\log \sum_a W_{ia} H_{a\mu} \leq - \sum_a \alpha_a \log \frac{W_{ia} H_{a\mu}}{\alpha_a}, \tag{13}$$

which holds for all nonnegative α_a that sum to unity. Setting $\alpha_a = \frac{W_{ia} H^t_{a\mu}}{\sum_b W_{ib} H^t_{b\mu}}$, we can obtain $G(H, H^t) \geq F(H)$.

To minimize $F(H)$, we can update H using $H^{t+1} = \arg\min_H G(H, H^t)$. Such an H can be found by letting $\frac{\partial G(H,H^t)}{\partial H_{a\mu}} = 0$ for all a, μ.

$$-\sum_i w_i V_{i\mu} \frac{W_{ia} H^t_{a\mu}}{\sum_b W_{ib} H^t_{b\mu}} \frac{1}{H_{a\mu}} + \sum_i w_i W_{ia} + 2\alpha H_{a\mu} = 0. \tag{14}$$

Therefore, the update rule of H takes the form:

$$H^{t+1}_{a\mu} = \frac{-\sum_i w_i W_{ia} + \sqrt{(\sum_i w_i W_{ia})^2 + \beta}}{4\alpha}, \tag{15}$$

where $\beta = 8\alpha H^t_{a\mu} \sum_i w_i W_{ia} V_{i\mu} / (WH^t)_{i\mu}$.

In Eq.(15), β is non-negative due that W and H are initialized as random positive numbers and $\alpha > 0$. So H^{t+1} keeps the non-negative property.

Updating W: W is updated by minimizing $F(W) = D(V||WH)$ with H fixed. We can calculate the update rule for W similarly by an auxiliary function $G(W, W^t)$. W is updated by $W^{t+1} = \arg\min_W G(W, W^t)$.

$G(W, W^t)$ is constructed as

$$\begin{aligned} G(W, W^t) &= \sum_{i,\mu} w_i (V_{i\mu} \log V_{i\mu} - V_{i\mu}) + \sum_{i,\mu,a} w_i W_{ia} H_{a\mu} \\ &- \sum_{i,\mu,a} w_i \Big(V_{i\mu} \frac{W^t_{ia} H_{a\mu}}{\sum_b W^t_{ib} H_{b\mu}} (\log W_{ia} H_{a\mu} - \log \frac{W^t_{ia} H_{a\mu}}{\sum_b W^t_{ib} H_{b\mu}}) \Big) \\ &+ \alpha \sum_\mu \sum_a H^2_{a\mu}. \end{aligned} \tag{16}$$

Based on Eq. (13) and setting $\alpha_a = \frac{W^t_{ia} H_{a\mu}}{\sum_b W^t_{ib} H_{b\mu}}$. It's verified that $G(W, W^t) \leq F(W)$ and $G(W, W) = F(W)$.

To minimize $F(W)$, we can update H using $W^{t+1} = \arg\min_W G(W, W^t)$. Such an W can be found by letting $\frac{\partial G(W,W^t)}{\partial W_{ia}} = 0$ for all i, a.

$$-w_i \sum_\mu V_{i\mu} \frac{W_{ia}^t H_{a\mu}}{\sum_b W_{ib}^t H_{b\mu}} \frac{1}{W_{ia}} + w_i \sum_\mu H_{a\mu} = 0. \tag{17}$$

Therefore, the update rule of W takes the form:

$$W_{ia}^{t+1} = W_{ia}^t \frac{\sum_\mu V_{i\mu} H_{a\mu}/(W^t H)_{i\mu}}{\sum_\mu H_{a\mu}}. \tag{18}$$

In each update, W is normalized to meet $\sum_i W_{ia} = 1$.

3 Topic-based Probabilistic Similarity Measure

Most of the text categorization systems rely on similarity metrics, such as Cosine or Euclidean distance. We use $S(D_1, D_2)$ to represent the similarity measure between two documents D_1 and D_2. Such a simple metrics suffers from a major drawback: it does not exploit knowledge of which types of variation are critical in expressing similarity.

Moghaddam [19] presented a probabilistic similarity measure based on the Bayesian belief that the image intensity differences, denoted by $\Delta = D_1 - D_2$ are characteristic of typical variations in appearance of an individual. Similar to their ideas, we define two classes of document variations: *intra-category variations* Ω_I and *extra-category variations* Ω_E. The similarity measure is then expressed in terms of a posterior probability

$$\begin{aligned} S(D_1, D_2) &= P(\Omega_I|\Delta) \\ &= \frac{P(\Delta|\Omega_I)P(\Omega_I)}{P(\Delta|\Omega_I)P(\Omega_I) + P(\Delta|\Omega_E)P(\Omega_E)}. \end{aligned} \tag{19}$$

Because Δ is high dimensional variable, it is difficult to estimate its probability. Moghaddam assume that Δ is the gaussian distribution and computed the likelihood using the PCA-based method [20]. Unlike the encoding coefficients of PCA, those of SNMF aren't pairwise independent, so we don't use the method in [19, 20].

Assuming that I' is the sparse encoding for document I, we define

$$x = I'_i - I'_j, \tag{20}$$

where $x = [x_1, \ldots, x_m, \ldots, x_r]^T$ is r-dim vector and represents the topical encoding variations between document i and j.

For the r-dim topical encoding variant x, if we estimate it using gaussian density, we need calculate the full covariance matrix because the encoding coefficients of SNMF aren't pairwise independent, which results in $r(r+1)/2$ independent covariance parameters to be estimated. Clearly, the samples are usually insufficient in the text categorization system, the estimated parameters aren't reliable.

We propose a topic-based probabilistic similarity measure, which has a more intuitive interpretation. We call the basis of SNMF as the topics of document. For the m-th topic, the corresponding topic-based variations is x_m. Thus, the log-likelihood ratios of topic-based variations are calculated on individual topic. Then we use an additive logistic regression model [21] to combine the similarities of different topics. Because the number of topics are greatly less than the dimension of Δ, our method can avoid the curse of dimensionality problem.

Given the topic-based variations x_m on topic m, we need calculate the corresponding similarity measure $f(x_m, \theta_m)$ We define the topic-based similarity measure $f(x, \theta_m)$ to be log-likelihood ratio between $P(x_m|\Omega_I, \theta_m)$ and $P(x_m|\Omega_E, \theta_m)$.

Here we use $y = \{+1, -1\}$ to represent the *intra-category variations* Ω_I and *extra-category variations* Ω_E.

$$f(x_m, \theta_m) = \log \frac{P(x_m|y = +1, \theta_m^+)}{P(x_m|y = -1, \theta_m^-)}, \tag{21}$$

where $\theta_m = (\theta_m^+, \theta_m^-)$.

Assuming that we estimate the likelihood of topic-based variations using gaussian density and θ_m be the parameters (*mean* and *variance*) of gaussian distribution on the topic m. For each $\Delta = D_1 - D_2$, there exits a $\Delta = D_2 - D_1$. Thus, densities of topic-based variations x are zero-mean. So we just estimate the parameter of variance of x_m. Here, parameters θ_m^+ and θ_m^- represent the variances of intra-category and extra-category topic-based variations on topic m, respectively.

Assuming $f(x_m, \theta_m)$ to be the similarity calculated on topic m, we need combine these topic-based similarities $f(x_m, \theta_m)$ $(m = 1, \ldots, r)$ to approximate the posterior probability $P(\Omega_I|\Delta)$. From $P(\Omega_I|\Delta) \in [0, 1]$, we can formulate it to additive logistic regression model [22] [23], which has the form:

$$\log \frac{P(y = +1|\Delta)}{P(y = -1|\Delta)} = \sum_{m=1}^{r} \alpha_m f(x_m, \theta_m). \tag{22}$$

At last, the posterior probability of facial variation can be calculated by

$$P(y = +1|\Delta) = \frac{e^{F(x)}}{1 + e^{F(x)}}, \tag{23}$$

where $F(x) = \sum_{m=1}^{r} \alpha_m f(x_m, \theta_m)$.

3.1 Parameter Learning

To find the optimal parameters α and θ in Eq. 22, the additive logistic model is usually fitted by maximizing the binomial log-likelihood, and enjoy all the associated asymptotic optimality features of maximizing likelihood estimation. Friedman showed that fitting an additive logistic regression can be replaced by minimizing the exponential criterion $J(F) = E(e^{-yF(x)})$ [24] [21], where E represents *expectation*.

We adopt a greedy forward algorithm to find the optimal parameters. At the mth stage we fix $F_{m-1}(x) = \sum_{i=1}^{m-1} \alpha_i f(x_i, \theta_i)$ and find α_m, θ_m minimize the exponential criterion,

$$\begin{aligned}
\{\alpha_m, \theta_m\} &= \arg\min_{\alpha,\theta} J\big(F_{m-1} + \alpha f(x_m, \theta)\big) \\
&= \arg\min_{\alpha,\theta} E\big(e^{-yF_{m-1}} e^{-y\alpha f(x_m,\theta)}\big) \\
&= \arg\min_{\alpha,\theta} E_w\big(e^{-y\alpha f(x_m,\theta)}\big),
\end{aligned} \tag{24}$$

where $E_w(\cdot)$ refers to a *weighted conditional expectation*, and $w = w(x, y) = e^{-yF_{m-1}(x))}$.

For $\alpha > 0$, minimizing Eq.(24) is equivalent to maximizing

$$E_w\left(e^{yf(x_m,\theta)}\right), \tag{25}$$

where $f(x_m, \theta) = \log \frac{P(x_m|y=+1,\theta^+)}{P(x_m|y=-1,\theta^-)}$ which is the log-ratio of likelihood between topic-based intra-category and extra-category variation x_m. For maximizing Eq.(25) we need estimate variance θ_m of Gaussian density (mean is zero) on weighted samples.

When $f(x_m, \theta_m)$ is obtained, the α that minimizes Eq.(24) satisfies

$$\frac{\partial E_w(e^{-y\alpha f(x_m,\theta_m)})}{\partial \alpha} = E_w(yf(x_m, \theta_m)e^{-y\alpha f(x_m,\theta_m)}) = 0. \tag{26}$$

This equation has no closed-form solution and requires an iterative solution by Newton-Raphson. The initial value of α is set zero.

Fig. 1 shows the flowchart of parameter learning of topic-based probabilistic similarity measure.

- Sample the intra-category variations and a random subset of extra-category variations from a given document database.
- Project the variations to topic-based representation, we get $(x_1, y_1), \ldots, (x_N, y_N)$, where $x_i = [x_{i1}, \ldots, x_{ir}]^T$ when there are r topics, and $y = \{+1, -1\}$ is class label representing the intra-category and extra-category variations.
- Assume the numbers of positive and negative samples are respectively N^+ and N^-, $N^+ + N^- = N$. Initialize sample weights: $w_1(x, y = +1) = \frac{1}{N^+}$,$w_1(x, y = -1) = \frac{1}{N^-}$.
- For $m = 1, \ldots, r$:
  1. Normalize the weights of samples so that $\sum_i w_i = 1$.
  2. Learn the intra-category and extra-category variances θ_m^+, θ_m^- on the topic m.
  3. Estimate $f(x_{im}, \theta_m)$ using Eq.(21)
  4. Calculate the optimal α by finding the solution of Eq.(26) using Newton-Raphson method.
  5. Set $w_i = w_i \exp[-y_i \alpha f(x_{im}, \theta_m)]$, $i = 1, \ldots, N$.
- The final topic-based probabilistic similarity measure is given in Eq.(23)

Figure 1: The flowchart of parameter learning of topic-based probabilistic similarity measure

3.2 Link to Naïve Bayes

Assuming the topics are independent of one another, the posterior probabilities can be calculated by

$$P(y = c|x) \quad = \quad P(y = c|x_1, \ldots, x_r)$$

$$
\begin{aligned}
&= \frac{P(x_1,\ldots,x_r|y=c)P(y=c)}{P(x)} \\
&= \Big(\prod_i (P(x_i|y=c))\Big)/P(x). \qquad (27)
\end{aligned}
$$

Leting $p = P(y=+1|x)$ and taking logarithms of $P(y=+1|x)/P(y=-1|x)$, Naïve Bayes is equivalent to the following formula:

$$\log\frac{p}{1-p} = \sum_i log(\frac{P(x_i|y=+1)}{P(x_i|y=-1)}) + \log(\frac{P(y=+1)}{P(y=-1)}). \qquad (28)$$

From Eq.(28), the link between Naïve Bayes and our topic-based probability model is shown that our model can be regarded as a weighted version of Naïve Bayes. Another difference from Naïve Bayes is that our model use the reweighting procedure in estimating the likelihoods of variations. We give larger weights to the samples which cannot be correctly predicted by the previous likelihood estimation. In fact, all the topics are not guaranteed to be independent, so the Naïve Bayes cannot reach a better performance. An improved version, semi-Naïve Bayes classifier [25], was proposed, which decomposes the input variables into subsets, representing statistical dependency within each subset, while treating the subsets as statistically independent. However, it is challenging to learn its structure because the search space is enormous.

4 Experiments

In this section, we conduct our experiments on two real large scale text data sets.

4.1 Datasets

We performed experiments on two data sets: Reuters21587[5] and 20 Newsgroups[26].

4.1.1 Reuters21587

The ApteMod version of Reuters21587[5] which was obtained by eliminating unlabeled documents and selecting the categories which have at least one document in the training set and the test set. This process resulted in 90 categories in both the training and test sets. After eliminating documents which do not belong to any of these 90 categories, we obtained a training set of 7768 documents, a test set of 3019 documents, and a vocabulary of 23793 unique words after stemming and stop word removal. The number of categories per document is 1.23 on average. The category distribution is skewed; the most common category has a training set frequency of 2877 but 82% of the categories have less than 100 instances, and 33% of the categories have less than 10 instances.

4.1.2 20 Newsgroups

The 20 Newsgroups data consists of Usenet articles Lang collected from 20 different newsgroups[26]. Over a period of time 1000 articles were taken from each of the newsgroups, which make up of an overall number of 20000 documents in this collection. Except for a small fraction of the articles, each document belongs to exactly one newsgroup. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hardware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). We use the "bydate" version of data whose training and testing data are split previously by the data provider[1]. There are 18941 documents and 70776 unique words after stemming and stop word removal.

4.2 Classifiers

In order to evaluate our method, we compare it with the traditional categorization methods: kNN with RCut[10][27] and SVM[28].

4.2.1 kNN with RCut

The RCut is a thresholding strategies[27]. In RCut, we sort categories by score for each document, and assign this document to each of t top-ranking categories. RCut is parameterized by t which can be either specified manually or automatically tuned using a validation set. In this paper, we set $t = 1$ since that the number of categories per document is closer to 1 in our experiments.

4.2.2 SVM

Besides kNN classification, we also test the performance of kNNFS with SVM to see whether kNNFS is helpful for the other classifiers. For SVM classifier, we choose SVMlib[28] with multi-class classification. We use the radial basis kernel in our experiments.

4.3 Performance Measurement

We evaluate their performances with precision, recall and $F1$ scores.

Precision, Recall and $F1$ are the most widely used performance measurements for text categorization problems nowadays[4]. Precision is the ratio of correct assignments by the system divided by the total number of the system's assignments. Recall is defined to be the ratio of correct assignments by the system divided by the total number of correct assignments. The $F1$ is a common measure in text categorization that combines recall and precision into a single score with an equal weight in the following formula:

$$F1 = \frac{2PR}{P + R}, \tag{29}$$

where P is the precision and R is the recall.

These scores (P, R, $F1$) can be computed in two ways: macro-averaging and micro-averaging. The macro-averaging is to calculate these scores for the binary decisions on each individual category and obtain the averaged scores over categories. The micro-averaging is that there scores are computed globally over all the $n \times C$ binary decisions where n is the number of total test documents and C is the number of categories. The micro-averaged F1 have been widely used to measure text categorization methods. The micro-averaged scores (recall, precision and F1) tend to depend on the classifier's performance on common categories, and the macro-averaged scores are more influenced by the performance on rare categories.

In this section, we use micro-averaging $F1$ to measure the performance of each result.

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/

Table 1: Micro-averaging $F1$ on Reuters21587 dataset.

Classifier	FS	Number of Features									
		2	4	8	16	32	64	128	256	512	1024
RCut kNN	DF	21.3	38.2	40.4	52.9	56.4	62.0	67.3	70.4	73.0	73.9
	CHI	21.3	21.3	21.4	26.7	30.2	31.1	64.9	73.7	76.4	**79.6**
	OCFS	23.8	24.3	26.2	41.0	51.6	68.9	73.0	72.8	76.1	76.7
	SNMF	**50.3**	**54.6**	**64.4**	**68.5**	**70.0**	**72.0**	**74.2**	**75.6**	**76.5**	76.9
SVM	DF	38.4	38.2	43.2	53.7	59.3	64.8	70.4	72.4	73.9	74.2
	CHI	32.1	32.1	32.2	37.6	55.3	55.9	67.8	74.1	77.5	78.9
	OCFS	34.9	40.0	44.1	53.5	58.6	71.6	76.1	77.1	77.6	76.6
	SNMF	**50.9**	**55.4**	**63.2**	**68.7**	**72.5**	**74.7**	**76.9**	**77.6**	**78.8**	**79.7**
Additive LR	SNMF	50.2	**55.6**	**64.4**	68.5	**73.0**	**74.9**	**77.1**	**78.0**	**78.9**	**79.9**

Table 2: Micro-averaging $F1$ on 20 Newsgroups dataset.

Classifier	FS	Number of Features									
		2	4	8	16	32	64	128	256	512	1024
RCut kNN	DF	4.9	6.3	9.8	12.6	16.1	20.7	29.6	40.1	50.8	58.4
	CHI	8.6	11.5	20.8	27.3	32.9	45.9	54.5	59.8	65.3	69.6
	OCFS	4.3	4.7	5.0	7.5	12.4	29.4	46.6	56.1	61.0	64.5
	SNMF	**10.3**	**25.0**	**38.5**	**49.0**	**55.4**	**56.9**	**57.4**	**61.1**	**70.5**	**71.4**
SVM	DF	5.9	7.2	12.1	13.6	19.5	27.2	39.3	49.7	62.3	69.6
	CHI	9.6	12.6	22.0	28.3	41.8	51.1	60.2	65.0	70.5	74.6
	OCFS	5.4	5.9	9.4	10.2	15.4	33.2	52.0	61.4	68.3	73.0
	SNMF	**10.8**	**25.7**	**40.1**	**51.3**	**62.2**	**65.3**	**68.7**	**70.9**	**73.4**	**75.1**
Additive LR	SNMF	10.1	**27.5**	**42.6**	50.2	**65.2**	**69.4**	**71.6**	**73.2**	**74.9**	**76.7**

4.4 Experimental Results

The performances of the different feature selection algorithms on Reuters21587 and 20 Newsgroups data are reported in Table 1. From the tables, we can see that the features selected by SNMF have better performances than the other methods(Document Frequency (DF), χ^2 statistics (CHI)[10] and and Orthogonal Centroid Feature Selection(OCFS)[29]) with different classifiers, especially when the number of features is small. The reason is that the features selected by SNMF are based on topics and not words, so they cover more information, especially in low dimensionality. Besides, our proposed probabilistic model for text categorization has competitive performence with the traditional methods, such as SVM, KNN.

5 Conclusions

We have presented a probabilistic model using sparse topical encoding for text categorization. There are two contributions in this paper. Firstly, we propose a sparse non-negative matrix factorization method, which can extract the sparse topical encoding and the localized and intuitive topics automatically. Secondly, we propose topic-based probabilistic similarity measure, whose parameters are learned through a forward stage-wise additive logistic regression algorithm. In the further work, we will investigate our method and extract more intuitive and effective topics, which have an obvious corresponding to our “notional topics”.

References

[1] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.

[2] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval.* Addison-Wesley Harlow, England, 1999.

[3] H. Drucker, D. Wu, and V.N. Vapnik. Support vector machines for spam categorization. *IEEE Transactions On Neural Networks*, 10(5):1048–1054, 1999.

[4] Y. Yang. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1):69–90, 1999.

[5] D.D. Lewis and M. Ringuette. A comparison of two learning algorithms for text categorization. *Third Annual Symposium on Document Analysis and Information Retrieval*, pages 81–93, 1994.

[6] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. *Proc. of Euro. Conf. on Mach. Learn. (ECML)*, pages 137–142, 1998.

[7] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[8] H. Kim, P. Howland, and H. Park. Dimension reduction in text classification with support vector machines. *J. Mach. Learn. Res.*, 6:37–53, 2005.

[9] K. Fukunaga. *Introduction to statistical pattern recognition.* Academic Press, Boston, 2nd edition, 1990.

[10] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In *Proc. of Int. Conf. on Mach. Learn. (ICML)*, volume 97, 1997.

[11] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[12] Chris H. Q. Ding, Tao Li, and Wei Peng. NMF and PLSI: equivalence and a hybrid algorithm. In *Proceedings of the annual international ACM SIGIR conference on Research and development in information retrieval*, pages 641–642, 2006.

[13] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM Press New York, NY, USA, 1999.

[14] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *Proc. of Neural Inform. Processing Syst. (NIPS)*, 2000.

[15] S. Li, X. Hou, H. Zhang, and Q. Cheng. Learning spatially localized, parts-based representation. In *Proc. of IEEE Conf. on Comput. Vision (ICCV)*, 2001.

[16] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, 5:1457–1469, 2004.

[17] B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[18] A. Dempster, N. Laird, and D. Bubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, 39:1–38, 1977.

[19] B. Moghaddam, T. Jebara, and A. Pentland. Bayesian face recognition. *Pattern Recogn.*, 33:1771–1782, 2000.

[20] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(7):696–710, 1997.

[21] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. Technical report, Stanford University, 1998.

[22] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[23] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[24] M. Collins, R. Schapire, and Y. Singer. Logistic regression, adaboost and bregman distances. In *Proc. of Annual Conf. on Comput. Learn. Theory (COLT)*, pages 158–169, 2000.

[25] I. Kononenko. Semi-Naïe bayesian classifier. In *Sixth European Working Session on Learning*, pages 206–219, 1991.

[26] Ken Lang. Newsweeder: Learning to filter netnews. In *Proc. of Int. Conf. on Mach. Learn. (ICML)*, pages 331–339, 1995.

[27] Y. Yang. A study of thresholding strategies for text categorization. In *Proc. of Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval(SIGIR)*, pages 137–145. ACM Press New York, NY, USA, 2001.

[28] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[29] J. Yan, N. Liu, B. Zhang, S. Yan, Z. Chen, Q. Cheng, W. Fan, and W.Y. Ma. Ocfs: optimal orthogonal centroid feature selection for text categorization. In *Proc. of Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval(SIGIR)*, pages 122–129. ACM Press New York, NY, USA, 2005.