Local Linear Matrix Factorization for Document Modeling

Conference Paper · November 2014

DOI: 10.1007/978-3-319-06028-6_33

4 authors, including:

Jiafeng Guo
Chinese Academy of Sciences
211 PUBLICATIONS **5,913** CITATIONS
SEE PROFILE

Yanyan Lan
Chinese Academy of Sciences
152 PUBLICATIONS **3,737** CITATIONS
SEE PROFILE

Local Linear Matrix Factorization for Document Modeling

Lu Bai, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng

Institute of Computing Technology, Chinese Academy of Sciences, BeiJing, China
bailu@software.ict.ac.cn, {guojiafeng,lanyanyan,cxq}@ict.ac.cn

Abstract. Mining low dimensional semantic representations of document is a key problem in many document analysis and information retrieval tasks. Previous studies show better representation mining results by incorporating geometric relationships among documents. However, existing methods model the geometric relationships between a document and its neighbors as independent pairwise relationship; while the pairwise relationship relies on some heuristic similarity/dissimilarity measures and predefined threshold. To address these problems, we propose a Local Linear Matrix Factorization (LLMF), for low dimensional representation learning. Specifically, LLMF exploits the geometric relationships between a document and its neighbors based on local linear combination assumption, which encodes richer geometric information among the documents. Moreover, the linear combination relationships can be learned from the data without any heuristic parameter definition. We present an iterative model fitting algorithm based on quasi-Newton method for the optimization of LLMF. In the experiments, we compare LLMF with the state-of-the-art semantic mining methods on two text data sets. The experimental results show that LLMF can produce better document representations and higher accuracy in document classification task.

Keywords: document modeling, local linear combination, matrix factorization.

1 Introduction

Extracting low dimensional semantic representations of documents has shown great success in wide applications[4][22]. Typically, by representing the corpus as a document-word matrix, matrix factorization can be applied to identify the semantic co-occurrence patterns of words (known as topics), and meanwhile extract low dimensional document representations over the topics [15][6]. Compared to the original document-word matrix, the low dimensional representations exhibit better semantics and achieve higher computation and storage efficiency.

Recent studies suggest that the documents are usually sampled from a nonlinear low dimensional subspace which is embedded in the high dimensional ambient space [7][8][14]. Thus, the local geometric structure is essential to reveal the hidden semantics in the corpora, and should be preserved when learning the low dimensional semantic representations. Based on this idea, some works (such

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 398–411, 2014.

as LapPLSA[7], LTM[8], DTM[14] and GNMF [6]) model the geometric relationship in a manifold way, which requires that the low dimensional representations of two documents to be close if they are neighbors in the original space. The empirical experiments demonstrated these methods can produce better results than the traditional factorization methods.

However, there are two clear drawbacks on only modeling the pairwise geometric relationship between documents. Firstly, the pairwise relationship usually relies on some heuristic similarity/dissimilarity measures, and the local neighborhood structure selected with predefined threshold as well. Unfortunately, it is unclear which measure can well capture the closeness of document pairs and the threshold is often hard to define in practice. Secondly, the pairwise geometric relationships between a document and its neighbors are assumed to be independent, making the rich geometric information among the local pairs lost. Moreover, these models may easily be affected by biased distribution of document pairs (especially when many redundant but less similar document pairs are included in the local neighborhood).

In this work, we propose Local Linear Matrix Factorization (LLMF), a novel low dimensional representation learning method by better exploiting the geometric relationship among documents. Specifically, inspired by the Local Linear Embedding (LLE [17]) method, we capture the geometric relationships among documents through representing a document by a linear combination of its neighbor documents. The linear combination coefficients demonstrate not only the geometric relationships between the document and its neighbors, but also the relationships among the neighbors. Unlike the LLE method, the linear combination coefficients are obtained by solving a regression problem with l_1 constraints [20], which simultaneously build the nearest neighborhood structure of the documents. Therefore, we do not resort to choosing any similarity/dissimilarity measure or predefining a threshold to select neighbors. With the learned combination coefficients, LLMF produces the low dimensional semantic representations by factorizing the document-word matrix with the local linear constraint. The learning process is straight forward, and can be summarized as an iterative process to fit the model in a quasi-Newton way.

We conduct empirical experiments on two benchmark text data sets. The results demonstrate that LLMF can produce better document representations and achieve higher accuracy in document classification task compared to several state-of-the-art semantic representation learning methods.

2 Preliminary Studies

In this section, we briefly introduce some previous studies on matrix factorization and the incorporation of geometric information.

2.1 A Brief View of Matrix Factorization

Matrix factorization or matrix decomposition is a technique that factorizes a source matrix into the production of several matrices based on the discipline of

linear algebra. Given a data matrix $W^T = [W_1, W_2, \ldots, W_N]^T \in \mathbb{R}^{N \times M}$, where $W_i, i \in [1, n]$ denotes a M-dimension data vector. A simple factorization of W^T is to produce two related matrices $U \in \mathbb{R}^{N \times K}$, $V \in \mathbb{R}^{K \times M}$ that approximate W^T with the production of U and V:

$$W^T \approxeq UV \tag{1}$$

The dimension K is usually set as a small value, e.g. $K \ll M$, thus U is a low dimensional and compact representation of W^T with the new basis depicted by V. In some domains, such as document modeling and image processing, U or V are required to be non-negative. Components thus are additive-only to construct data that makes the factorization more interpretable and favorable in practice.

Square Euclidean distance between W^T and UV is a typical objective function for matrix factorization (e.g. in NMF).

$$\min \sum_{i,j} (W_{ij}^T - U_{i\cdot} V_{\cdot j})^2 \tag{2}$$

The problem in (2) is bi-convex, which means the problem is convex on $U(V)$ when $V(U)$ is fixed. We can apply an iterative algorithm for optimization. As suggested in [2], it is convenient to optimize U or V as a linear regression problem alternatively, then stop the iteration when the loss is small enough.

Alternatively, KL-divergence is also a popular measure for the approximation of W^T with U and V, especially when U and V fall into the probabilistic perspective (e.g. PLSA). Usually, the statistic inference methods, e.g. EM or MCMC, are employed to solve the optimization problem.

2.2 Document Modeling with Geometric Constraints

Matrix factorization has been applied in document modeling to find compact representations by minimizing the reconstruction error. Recent studies show that by incorporating geometric information among documents, one can learn better low dimensional representations. A variety of document modeling methods employ the idea of Laplacian Eigenmap(LE) to enhance the geometric properties of learned topics, such as GNMF [6], LapPLSA [7] and LTM [8]. Specifically, a document manifold is first constructed by selecting the neighbors of each document with some similarity measure and threshold. The geometric constraints of LE can then be formulated as the following optimization problem.

$$\min \sum_{i,j} S_{ij} \parallel x_i - x_j \parallel_2^2 \tag{3}$$

where S_{ij} denotes the similarity between i and j. Intuitively, the optimization of (3) is equivalent to make the low dimensional representation x_i, x_j close when i and j are close in the original space.

Obviously, the existing document modeling methods based on LE preserve the geometric information among documents by only modeling the geometric

relationships between independent document pairs. Thus, the rich geometric information among the neighbor pairs are lost in this case. Moreover, the pairwise relationship relies on some heuristic similarity measure as well as the predefined threshold for selecting the neighbors. However, it is unclear which measure can well capture the closeness of document pairs and the threshold is often hard to define in practice.

3 Local Linear Matrix Factorization

In this section, we introduce a novel low dimensional representation learning method better exploiting the geometric information among documents, namely Local Linear Matrix Factorization(LLMF). We also provide an effective algorithm for optimization. In addition, we give some detailed discussions about the differences between LLMF and other document modeling methods.

3.1 Model Formalization

Suppose we have N documents over the vocabulary of size M. Let $D^T = [D_1, D_2, \ldots, D_N]^T \in \mathbb{R}_+^{N\times M}$ denote a document-word matrix, where D_{ij}^T is the occurrence number of word j in document i. $\theta \in \mathbb{R}_+^{N\times K}$ is the low dimensional representation of D^T with $K \ll N$. $\beta \in \mathbb{R}_+^{K\times M}$ is the corresponding basis of the latent semantic space. From the perspective of matrix factorization, D^T can be expressed as

$$D^T \approxeq \theta\beta \tag{4}$$

By using the square error to measure the approximation in formula (4), we obtain the following non-negative matrix factorization problem for document modeling

$$\begin{gathered}\mathcal{L}_\theta = \sum_{i=1}^{M}\sum_{j=1}^{N}(D_{ij}^T - \theta_{i\cdot}\beta_{\cdot j})^2 + \lambda_\theta \parallel \theta \parallel_2^2 + \lambda_\beta \parallel \beta \parallel_2^2 \\ s.t. \quad \theta \geq 0, \beta \geq 0\end{gathered} \tag{5}$$

where λ_θ and λ_β are the weights of l_2 regularizers used to reduce the over-fitting, and the non-negative constraints over θ and β make the learned components interpretable.

Inspired by LLE [17], we consider that local geometric information can be captured by local linear combination relationship (i.e. document can be reconstructed by linear combination of its neighbors), rather than independent pairwise relationships. Specifically, document d can be approximate as

$$D_d^T \approxeq \phi_d^T \hat{D}^T \tag{6}$$

where ϕ_d denotes the combination weight vector for document d, and $\hat{D}^T$ denotes the normalized document-word matrix obtained by $\hat{D}_{ij}^T = \frac{D_{ij}^T}{L_i}$, where L_i is the length of document i. We use normalized document-word matrix in local linear combination to avoid the bias of long documents. Note that for the combination

weight vector of document d, documents not belonging to the neighbors of d have the value 0 at the corresponding entries of ϕ_d.

The local linear combination constraints can then be expressed by the following objective function

$$\mathcal{L}_\phi = \sum_{i,j}^{N} (D_{ij}^T - \phi_{i\cdot}^T \hat{D}_{\cdot j}^T)^2 + \gamma \parallel \phi \parallel + \lambda_\phi \parallel \phi \parallel_2^2 \tag{7}$$

where γ and λ_ϕ denotes the weights for l_1 and l_2 norm over ϕ, respectively. Here we put the l_1 norm over ϕ due to the assumption that the number of neighbors of each document is small. It is worth noting that minimizing the formula (7) actually conducts neighbor selection and combination weight learning simultaneously. In this way, we can avoid using heuristic similarity measures and threshold to select neighbors as previous methods.

Unlike previous document modeling methods with geometric constraints (e.g. GNMF, LapPLSA and LTM), it is not straightforward to combine the local linear constraints $\mathcal{L}_\phi$ with the matrix factorization objective $\mathcal{L}_\theta$, since the θ and ϕ is not shared by both optimizations. However, we can bridge θ and ϕ by the normalized basis $\hat{D}$. Based on formula (4), we have that

$$\hat{D}^T = [\hat{D}_1^T, \ldots, \hat{D}_N^T] = [\hat{\theta}_1\beta, \ldots, \hat{\theta}_N\beta] = \Big([\hat{\theta}_1, \ldots, \hat{\theta}_N]\Big)\beta = \hat{\theta}\beta \tag{8}$$

where the entries of $\hat{\theta}_i$ can be derived as $\hat{\theta}_{ik} = \frac{\theta_{ik}}{L_i}$, $k \in [1, \cdots, K]$. In this way, formula (6) can be rewritten as

$$D_d^T \approxeq \phi_d \hat{D}^T \approxeq \phi_d \hat{\theta}\beta \tag{9}$$

Compared formula (9) with formula (4), we obtain that

$$\theta_d \approxeq \phi_d[\hat{\theta}_1, \ldots, \hat{\theta}_N] \tag{10}$$

Then we can integrate the local linear constraints into the matrix factorization, and the objective function can be expressed as

$$\begin{aligned}\mathcal{L}(\theta, \beta) = &\sum_{i=1}^{N}\sum_{j=1}^{M}(\sum_{k=1}^{K}\theta_{ik}\beta_{kj} - D_{ij})^2 + \lambda_\theta \sum_{i=1}^{N}\sum_{k=1}^{K}\theta_{ik}^2 + \lambda_\beta \sum_{k=1}^{K}\sum_{j=1}^{M}\beta_{kj}^2 \\ &+\eta \sum_{i=1}^{N}\sum_{k=1}^{K}(\theta_{ik} - \sum_{n=1}^{N}\frac{1}{L_n}\phi_{in}\theta_{nk})^2\end{aligned} \tag{11}$$

where the coefficient η controls the trade-off between the matrix factorization objective and the local linear constraints.

3.2 Model Fitting

In this section we would show how to infer the latent factor ϕ, θ and β respectively. The inference process can be divided into two optimization problems.

ϕ can be firstly evaluated by optimizing the objective function (7) as a regression problem. Since the basis $\hat{D}^T$ can be pre-computed by normalizing the words count for every documents, the optimization leaves ϕ unknown. It is easy to prove that minimizing the objective function (7) is a convex problem, where the global optimal solution can be found. However, the l_1 regularizer makes the optimization problem in formular (7) not differentiable when some dimension of ϕ is 0. To address this problem, we adopt the OWL-QN [3] algorithm for optimization.

OWL-QN algorithm is based on the famous Quasi-Newton algorithm that leverages the second order information to accelerate the optimization. The algorithm would check the state for each iteration step and revise the value if some dimension cross the orthant boundary. The OWL-QN algorithm requires to calculate the gradient of the function without the l_1 norm part, which is shown as follows

$$\frac{\partial \mathcal{L}'_{\phi}}{\partial \phi_{in}} \propto \sum_{w=1}^{M} \big(\sum_{n=1}^{N} \phi_{in} \hat{D}^T_{nw} - D^T_{iw} \big) \hat{D}^T_{nw} + \lambda_{\phi} \phi_{in} \tag{12}$$

The iteration stops when the change of objective function (7) is small enough.

After evaluating ϕ, θ and β can be obtained by objective function (5). Since both θ and β are unknown, the optimization problem is bi-convex and we can use the alternative strategy for optimization. Here we solve the problem with gradient method. The gradient of θ and β can be calculate as following:

Algorithm 1. Learning Procedure of LLMF

Require: $D^T, K \in Z_+, \epsilon > 0, \lambda_\theta, \lambda_\phi, \lambda_\beta, \eta$
  $f \leftarrow$ the function to calculate the loss of ϕ as function 7
  $g \leftarrow$ the function to calculate the gradient of ϕ as function 12
  initialize $\phi \in \mathbb{R}_+^{N \times N}$ randomly
  Learn $\phi \leftarrow$ OWL-QN$(D^T, \phi, \eta, \epsilon, f, g)$
  initialize $\theta^{(0)} \in \mathbb{R}_+^{N \times K}$ randomly
  initialize $\beta^{(0)} \in \mathbb{R}_+^{K \times M}$ randomly
  $f \leftarrow$ the function to calculate the loss of ϕ as function 5
  $g_\theta \leftarrow$ the function to calculate the gradient of ϕ as function 14
  $g_\beta \leftarrow$ the function to calculate the gradient of β as function 13
  $l^{old} \leftarrow 0$
  for $t = 1 : T$ **do**
    Learn $\theta^{(t)} \leftarrow$ OWL-QN$(D^T, \theta^{(t-1)}, 0, \epsilon, f, g_\theta)$
    Learn $\beta^{(t)} \leftarrow$ OWL-QN$(D^T, \beta^{(t-1)}, 0, \epsilon, f, g_\beta)$
    $l \leftarrow$ calculate loss function value as function 5
    if $\| l - l^{old} \| < \epsilon$ **then**
      break
    end if
    $l^{old} \leftarrow l$
  end for
  return θ, β, ϕ

$$\frac{\partial \mathcal{L}_\theta}{\partial \beta_{kw}} = \sum_{d=1}^{N} \Big(\sum_{k=1}^{K} \theta_{dk}\beta_{kw} - D_{dw}^{T} \Big)\theta_{dk} + \lambda_\beta \beta_{kw} \tag{13}$$

$$\begin{aligned} \frac{\partial \mathcal{L}_\theta}{\partial \theta_{dk}} = & \sum_{w=1}^{M} \Big(\sum_{k=1}^{K} \theta_{dk}\beta_{kw} - D_{dw}^{T} \Big)\beta_{kw} + \lambda_\theta \theta_{dk} + \eta \sum_{b=1}^{N} \Big(\sum_{n=1}^{N} \frac{1}{L_n} \phi_{bn}\theta_{nk} - \theta_{bk} \Big) \Big(\frac{1}{L_b} \phi_{bd} \Big) \\ & - \eta \Big(\sum_{n=1}^{N} \frac{1}{L_n} \phi_{dn}\theta_{nk} - \theta_{dk} \Big) \end{aligned} \tag{14}$$

The algorithm to infer ϕ, θ and β are described in the algorithm 1.

3.3 Discussion

In this section, we further provide some discussions on the differences between the proposed LLMF method and existing state-of-the-art document modeling methods.

Compared to NMF, PLSA and LDA, LLMF smoothed the low dimensional representations of documents with its neighbors. The weights for each neighbors are evaluated by solving a least square regression problem that captures the geometric information among documents. Both LDA and NMF smooth the latent representations with a unimodal prior distribution (Dirichlet distribution for LDA and Gaussian distribution for NMF), but the posterior distribution would prefer to fit the most intensive areas globally. In all the three methods (i.e. NMF, PLSA and LDA), the local geometric information among documents is not considered.

When compared with document modeling methods with geometric constraints, like LapPLSA and LTM, the way of capturing the geometric information in these

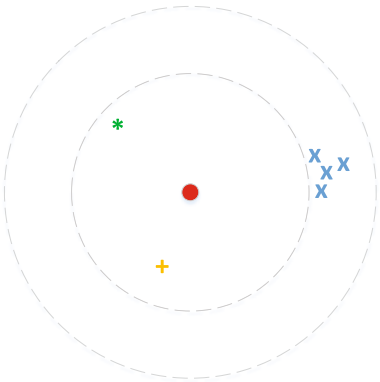

Fig. 1. Biased latent representations when data distribution is unbalanced

methods is quite different from that of our proposed LLMF. Previous methods smooth the latent representations with their neighbors in a pairwise way. It is illustrated in [5] that in these methods the similarity measure and neighborhood threshold should be carefully defined when constructing the local manifold patch, otherwise the improper neighborhood would dramatically bias the latent representations and affect the performance. We explain this phenomenon here and show the advantage of local linear constraints over pairwise constraints. As shown in Fig. 1, one aims to learn the latent representation of red circle point, where the green snow point ($*$) and yellow plus point ($+$) are more close than the blue cross points (x). When the blue points are included as neighbors of the target red point, the learned representation of the red circle point would be biased to the blue points due to the pairwise smoothing regularization and unbalanced data distribution. This problem would become more severe when more neighbors are used for pairwise smoothing. However, in our LLMF , the local geometric information is preserved by linear combination of its neighbors, and thus the neighbors are competitive in representing the data. Therefore, the importance of the green and red points in representing the red point would not be affected much even when the blue points are involved, and the weight of each blue point would be reduced due to the redundancy. As a result, LLMF can better preserve the rich geometric relationships among data and learn better low dimensional representations.

4 Experiments

In this section, we demonstrate the results in the task of document classification, with experiments conducted on two widely used benchmark text corpora, i.e. 20newsgroup and la1. Firstly, we introduce the experimental settings. Then we qualitatively evaluate the latent topics learned by LLMF. At last, we evaluate the effectiveness of the proposed LLMF in document classification by comparing with the state-of-the-art semantic learning methods.

4.1 Data Sets and Baseline Methods

Our empirical studies on text semantics learning were conducted on two real-word text corpora, i.e. 20newsgroup dataset and la1 data set.

- The 20newsgroup[1] is a benchmark text collection for topic modeling, which contains almost 18,744 postings to Usenet newsgroups in 20 different categories almost evenly. The vocabulary is pruned by stemming each term to its root and removing the stop words for noise concern.
- la1 is a public dataset in Weka[2] containing 2,850 documents in 5 categories. The vocabulary consists of 13,195 unique words, and is also preprocessed by

[1] `http://people.csail.mit.edu/jrennie/20Newsgroups/`
[2] `http://www.cs.waikato.ac.nz/ml/weka/`

stemming and stop words removing. Unlike 20newsgroup, la1 is an unbalanced dataset where different categories contain quite different number of documents.

To evaluate the performance of LLMF, we provide comparisons of our methods against several state-of-the-art topic learning methods, including PLSA [13], LDA [4], NMF [15] and LapPLSA [13]. Here we briefly introduce the experimental settings about these methods beside our LLMF.

- **PLSA:** Probabilistic Latent Semantic Analysis(PLSA) is introduced by Hoffman[13] as a probabilistic version of LSI [9]. We use the code from Peter's homepage[3] for experiment.
- **LDA:** Latent Dirichlet Allocation(LDA) is a full Bayes version of PLSA. We user the code from the author's homepage of LDA[4], which is implemented in c-language by Blei.
- **NMF:** Non-negative Matrix Factorization(NMF) is a traditional dimension reduction methods. We adopt the alternating constrained least squares (ACLS) [2] to factorize the document-word matrix. To avoid over-fitting, l_2 norm is added as a constraint over the factorized matrices.
- **LapPLSA:** LapPLSA smooths the topic representations of nearby document pairs. In our experiments, we applied different number of neighbors and weights of the geometric regularizer, and select the best performance to report.
- **LLMF:** In our proposed LLMF, we set the parameters $\lambda_\theta, \lambda_\beta, \lambda_\phi$ as $0.01, 0.1, 1$, respectively. The l_1 norm in the model is weighted by $\gamma \in \{0.001, 0.01, 0.1, 1\}$.

All the above methods are conducted on both dataset several times with random initialization by setting the dimension $K \in \{30, 50, 70, 100\}$. We compare the best results of different methods and demonstrate the results in the following sections.

4.2 Topic Learning

In this section, we qualitatively evaluate the learned semantic information by LLMF. To illustrate the meaning of the inferred semantic factors, we randomly select several column from the matrix β, and re-range the words according to the corresponding weights in that column in descending order.

In Table 1, we list the top 10 important words from the randomly selected 5 learned components over the two datasets, respectively. For better understanding, we manually label each topic according to the meaning of the selected words. It is interesting to see that the selected words are closely related, and show similar semantic meanings expressed by the labels. Therefore, the results show that our LLMF can effectively learn the latent semantic information from the corpus.

[3] `http://people.kyb.tuebingen.mpg.de/pgehler/code/index.html`
[4] `http://www.cs.princeton.edu/blei/lda-c`

Table 1. Topics Learned by LLMF over the Two Datasets

20news					la1				
image	**hardware**	**hockey**	**ibm**	**motor-race**	**sports**	**financial**	**national**	**market**	**computer**
jpeg	tape	team	ibm	motorcyc	game	fund	west	bank	stor
gif	driver	hockei	hardwar	ride	plai	stock	soviet	compani	disk
compress	adaptec	leag	ram	rec	team	market	chemic	million	electron
viewer	sys	nhl	machin	bmw	season	price	plant	loan	comput
jfif	backup	season	memori	bike	player	invest	weapon	amp	data
convert	memori	game	card	club	coach	trad	german	card	ibm
format	run	player	monitor	time	basketbal	bond	govern	sav	machin
quantis	cdrom	championship	dos	rider	goal	investor	libya	credit	softwar
imag	floppi	wing	cpu	moto	time	exchang	israel	billion	pc
displai	hardwar	vs	bus	biker	win	trade	american	market	user

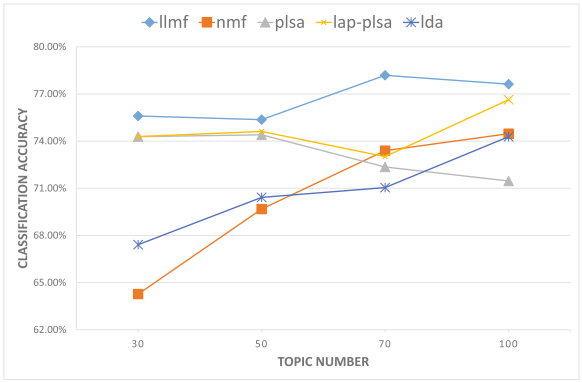

Fig. 2. Classification performance on 20newsgroups

4.3 Document Classification

In this section, we quantitatively evaluate the effectiveness of the learned low dimensional representations. Since the learned low dimensional representation is usually taken as feature engineering in real application tasks such as classification and clustering, we propose to evaluate the representations by comparing the performances of different methods in these tasks.

Here we conduct the task of document classification on the two data sets mentioned above, and compare the classification accuracies of different models. Specifically, we use all the documents of each data set to learn the parameters of different models. Then we randomly select 60% documents with their inferred representations as features to build the multi-class SVM classifier, and the rest 40% documents for test. We adopted the LIBSVM toolbox[5] as our implementation for SVM. Cross validation is conducted to select the parameter C in SVM.

The classification results on the two data sets are reported in Figure 2 and 3. We can see that LLMF consistently achieves better accuracy than all the baseline methods in both data sets. The results indicate that we can learn better

[5] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

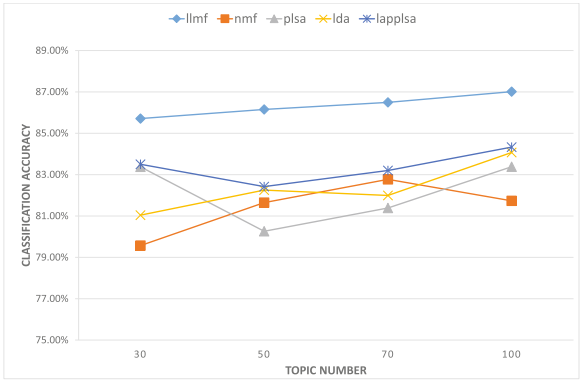

Fig. 3. Classification performance on la1

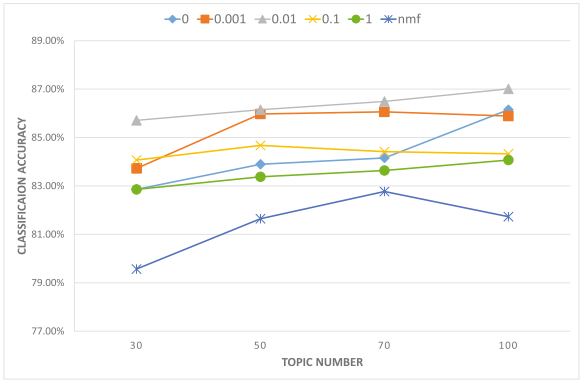

Fig. 4. Classification performance on la1 with the variation of the parameter γ

semantic representation by LLMF. It demonstrates that it is valuable to preserve more detailed local geometric information using local linear combination for learning semantic representation. Moreover, we can see that the difference between between LLMF and LapPLSA in Figure 3 is even larger than that in Figure 2. The reason may lie in the imbalance of la1, since the pairwise relation will be highly biased by imbalanced data, as shown in our discussion.

We also study the robustness of our methods under different sizes of neighborhood. In our proposed LLMF, different γs represent different sparseness of neighbors. Intuitively, the bigger γ is, the fewer neighbors wil be used in the linear combination, and vise versa. Therefore, we show the variation of classification accuracy on the unbalanced dataset la1 as the weight γ changes in LLMF in Figure 4. Since LLMF is essentially a generalized non-negative matrix factorization method, we take NMF as the baseline for comparison. From the results we can clearly see that with different γs, LLMF are consistently better than the

basic NMF method in terms of document classification accuracy. It demonstrates the robustness of our proposed LLMF method.

5 Related Work

Learning an effective semantic representation of data can greatly improve the effectiveness and efficiency of many real applications, such as information retrieval [22], network analysis [1], recommend systems [21] and so on.

For document modeling, matrix factorization is a typical way to learn the low dimensional representations of documents, such as LSI [9] and NMF [15]. These methods directly factorize the document-word matrix into low rank matrices according to different criteria. As an alternative, topic models, such as PLSA [13] and LDA [4], provide a probabilistic view on document modeling. Specifically, each document is taken as a distribution over topics, where each topic is a distribution over words. Through the posterior optimization, [18] provides a probabilistic generative view in interpreting the matrix factorization. [12] and [10] demonstrate the close connections between PLSA and NMF.

Integrating the geometric relationship among documents into the document modeling methods has been proved reasonable and effective. Some studies [11][19][16] leverage the explicit relationship, such as links and citations, in topic learning. Alternatively, several researches employ the geometric relationship, e.g. manifold assumption, to improve the document modeling. For example, LapPLSA [7] increases the proximity between the topics of document pairs in neighborhood using Laplacian eigenmap constraints based on PLSA. LTM [8] takes the same assumption as LapPLSA, but leverages the KL-divergence to evaluate the difference of topics instead. GNMF [6] and GraphSC [23] leverage the graph embedding to regularize the latent factor learning. DTM [14] not only enhances the topical proximity between nearby document pairs, but also increases the topical separability between the unfavorable pairs. As far as we known, all these above methods need firstly select the documents' neighbors with heuristic similarity measure and thresholds, and then preserve the geometric relationship by enhancing the pairwise proximity.

6 Conclusions

In this paper, we present a novel method for learning low dimensional representations of document, namely Local Linear Matrix Factorization(LLMF). LLMF exploits the geometric relationships between a document and its neighbors based on local linear combination assumption in document modeling. In this way, LLMF can better capture the rich geometric information among documents than those based on independent pairwise relationships. The experimental results on document classification show LLMF can learn better low dimensional semantic representations than the state-of-the-art baseline methods.

In the future, we would like to extend LLMF to the paralleled and distributed settings for computation efficiency. Moreover, it would also be interesting to

apply LLMF to other scenarios, e.g. recommender systems, where dimension reduction has shown benefits for the application.

Acknowledgments. This work was funded by the 973 Program of China under Grants No. 2012CB316303 and No. 2014CB340401, and National Natural Science Foundation of China under Grant No. 61232010, No. 61003166, No. 61203298, No. 61173064 and No. 61272536.

References

1. Airoldi, E.M., Blei, D.M., Fienberg, S.E., Xing, E.P.: Mixed membership stochastic blockmodels. J. Mach. Learn. Res. 9, 1981–2014 (2008)
2. Albright, R., Cox, J., Duling, D., Langville, A.N., Meyer, C.D.: Algorithms, initializations, and convergence for the nonnegative matrix factorization. Matrix (919), 1–18 (2006)
3. Andrew, G., Gao, J.: Scalable training of l1-regularized log-linear models. In: ICML 2007, pp. 33–40. ACM, New York (2007)
4. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent dirichlet allocation. Journal of Machine Learning Research 3 (2003)
5. Cai, D., He, X., Han, J.: Locally consistent concept factorization for document clustering. IEEE Transactions on Knowledge and Data Engineering 23(6), 902–913 (2011)
6. Cai, D., He, X., Han, J., Huang, T.S.: Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. Pattern Anal. Mach. Intell. 33(8), 1548–1560 (2011)
7. Cai, D., Mei, Q., Han, J., Zhai, C.: Modeling hidden topics on document manifold. In: CIKM 2008, pp. 911–920. ACM, New York (2008)
8. Cai, D., Wang, X., He, X.: Probabilistic dyadic data analysis with local and global consistency. In: ICML 2009, pp. 105–112. ACM, New York (2009)
9. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JASIS 41(6), 391–407 (1990)
10. Ding, C., Li, T., Peng, W.: On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. Comput. Stat. Data Anal. 52(8), 3913–3927 (2008)
11. Erosheva, E., Fienberg, S., Lafferty, J.: Mixed-membership models of scientific publications 101(suppl. 1), 5220–5227 (2004)
12. Gaussier, E., Goutte, C.: Relation between plsa and nmf and implications. In: SIGIR 2005, pp. 601–602. ACM, New York (2005)
13. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. In: Machine Learning (2001)
14. Huh, S., Fienberg, S.E.: Discriminative topic modeling based on manifold learning. In: KDD 2010, pp. 653–662. ACM, New York (2010)
15. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–791 (1999)
16. Nallapati, R.M., Ahmed, A., Xing, E.P., Cohen, W.W.: Joint latent topic models for text and citations. In: KDD 2008, pp. 542–550. ACM, New York (2008)
17. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. SCIENCE 290, 2323–2326 (2000)
18. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: NIPS (2007)

19. Sun, C., Gao, B., Cao, Z., Li, H.: Htm: a topic model for hypertexts. In: EMNLP 2008, Stroudsburg, PA, USA, pp. 514–522. Association for Computational Linguistics (2008)
20. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)
21. Wang, C., Blei, D.M.: Collaborative topic modeling for recommending scientific articles. In: KDD 2011, pp. 448–456. ACM, New York (2011)
22. Wei, X., Croft, W.B.: Lda-based document models for ad-hoc retrieval. In: SIGIR 2006, pp. 178–185. ACM, New York (2006)
23. Zheng, M., Bu, J., Chen, C., Wang, C., Zhang, L., Qiu, G., Cai, D.: Graph regularized sparse coding for image representation. Trans. Img. Proc. 20(5), 1327–1336 (2011)