Knowledge-Based Systems 121 (2017) 41–57

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Discriminative locally document embedding: Learning a smooth affine map by approximation of the probabilistic generative structure of subspace

Chao Wei, Senlin Luo, Jia Guo, Zhouting Wu, Limin Pan*

Beijing Institute of Technology, Beijing 10081, China

ARTICLE INFO

Article history:
Received 19 July 2016
Revised 25 December 2016
Accepted 7 January 2017
Available online 11 January 2017

Keywords:
Document embedding
Smooth affine map
Generative probabilistic model
Multi-agent random walk
Regularized auto-encoders

ABSTRACT

Document embedding is a technology that captures informative representations from high-dimensional observations by some structure-preserving maps over corpus and has been intensively explored in machine learning. Recently, some manifold-inspired embedding methods become a hot topic, mainly due to their ability in capturing discriminative embedding. However, the existing methods capture the embeddings based on the geometrical information of nearest neighbors without considering the intrinsic documents-generating structure on a subspace, thus leads to a limitation to uncover intrinsic semantic information. In this paper, we propose a semi-supervised local-invariant method, called Discriminative Locally Document Embedding (Disc-LDE), aiming to build a smooth affine map for document embedding by preserving documents-generating structure on a subspace. Disc-LDE models the documents-generating structure as a pseudo-document by a generative probabilistic model of subspace, where the subspace is acquired by a transductive learning of multi-agent random walk on neighborhood graph, and regularizes the training of Auto-Encoders (AEs) to jointly recover the input document and its pseudo-document. Under a general regularized function learning framework, the regularized training can impact the parameterized encoder network become smooth to variations along the documents-generating structure of the local field on manifold. The experimental results on three widely-used corpora demonstrate Disc-LDE could efficient capture the intrinsic semantic structure to improve the clustering and classification performance to the state-of-the-arts methods.

1. Introduction

Many fields of document analysis and processing systems involving clustering [1,2] and documents retrieval [3] can be facilitated through certain structure-preserving embeddings in a low-dimensional space, since the semantic structure of documents may become easier to estimate. In the last decades, many unsupervised attempts to distil low-dimensional embeddings have been proposed, like Latent Semantic Indexing (LSI) [4], Nonnegative Matrix Factorization (NMF) [5], probabilistic Latent Semantic Analysis (pLSA) [6] and Latent Dirichlet Allocation (LDA) [7], etc. Alternatively, various architecture of neural networks, such as AEs [8], Long-Short Term Memory (LSTM) [9] and document neural autoregressive distribution estimators (DocNADE) [10], take the intermediate output of hidden layer as embeddings, which is called distributed representation, due to the feeding signals are distributed over each independently activated neuron. Afterward, to further improve the discriminability of document embeddings, some supervised or semi-supervised extensions based on the above methods incorporate some priori knowledge of labeled samples (i.e., category labels) as a side information into the learning process [11–16].

However, the aforementioned research neglect a fact that human-generated documents are probably sampled from a manifold of the ambient euclidean space [17,18], which means the local invariant structure[1] of the corpus needs to be incorporated into the document embedding. For this reason, Constrained Neighborhood Preserving Concept Factorization (CNPCF) [21] and Semi-Supervised Concept Factorization (SSCF) [22] are proposed to incorporate the pairwise constraints related to local invariant into concept factorization (CF) as the reward and penalty terms for better embedding. Locally discriminative Topic Model (LDTM) [23] and

* Corresponding author.
E-mail address: panlimin.bit@gmail.com (L. Pan).

[1] In mathematics, manifold is defined as a topological space that resembles Euclidean space near each point, which means low-dimensional embeddings depend crucially on modeling the local relation of samples. The structure is referred as Local Invariant.

http://dx.doi.org/10.1016/j.knosys.2017.01.012
0950-7051/

Locally-consistent Topic Model (LTM) [24] extend pLSA based on the proximity information of nearby documents. Furthermore, Discriminative Topic Model (DTM) [25] considers information of non-neighboring pairs for full discriminability. However, such methods are plagued by an out-of-sample (OOS) extension problem, since they cannot build an explicit projection of the new coming samples. As a result, they are successfully applied to improve the clustering rather classification. A straightforward remedy for classification is an inclusive approach that rebuilds proximity matrices with new coming documents, retraining the model based on these matrices [25].

The inclusive approach is obviously more computationally expensive. Another option is to build a parameterized mapping function to preserve the local invariant structure of the whole data distribution [26]. Recently, some manifold-inspired embedding methods are proposed based on this idea, such as [27], Graph regularized Auto-Encoders (GAEs) [28], Laplacian Auto-Encoders (LAEs) [29] and Flexible Constrained Sparsity Preserving Embedding (FCSPE) [30]. Among these methods, GAEs and LAEs both exploit the relation of nearby samples to regularize the training of AEs and can extract the embeddings of new coming samples with the encoder network directly. Study in [27] is designed for fitting a radial basis function interpolator using a progressive estimation for OOS extension. It is a cascade estimation for the training samples embeddings priorly obtained by some other manifold embedding algorithms, limiting the further enhancement its performance through a fusion of the explicit map and low-dimensional embeddings. FCSPE can build an approximate linear projection through a regression process while capturing the low-dimensional embeddings over the training samples [30], which will lead to more flexibility in practical applications. Although these methods take different approaches for OOS extension, the basic criterion for preserving the local invariant property is to encode the geometrical information by constructing a k-nearest neighbor (knn) graph. However, except for the geometrical information of knn, some salient statistical information for modeling local cluster of the corpus, like documents-generating structure on a subspace, is also worthy of attention for capturing the intrinsic semantic information and improving the application of document embedding.

In this paper, we propose a semi-supervised local-invariant document embedding method, called Discriminative Locally Document Embedding (Disc-LDE), which can build a smooth affine map for OOS extension of document embedding by preserving documents-generating structure on a subspace, so that improve clustering and classification. Disc-LDE models the documents-generating structure as a pseudo-document by a generative probabilistic model of a discriminative subspace, where the subspace is acquired by a transductive learning of multi-agent random walk on improved knn graph, and regularizes the training of AEs to recover the input document and its pseudo-document under the given document. Several contributions of Disc-LDE are summarized as follows:

(1) Under a general regularized function learning framework, the recovering of pseudo-document acts as a regularized term to guide the parameterized encoder network to capture the intrinsic semantic information for a high-precision estimation of documents-generating structure on a subspace, thus when handling some nearby documents around the subspace, it ensures the local smooth property of learned encoder network.
(2) When building a knn graph, Disc-LDE adopts a normalized maximum matching distance based on word embeddings to weight every documents pair, called normalized matching distance (NMD), providing a robust similarity measurement even for the documents pair that has many synonyms but does not have the same vocabulary.
(3) Disc-LDE can adaptively find a discriminative subspace by taking a semi-supervised view of multi-agent Markov random walks on a knn graph, thus ensure the subspace achieve stronger local invariant. Meanwhile, the found subspace has a large overlap, facilitating efficient propagation of the discriminative information along with knn graph. Besides, some related results proved the basic idea is available to other manifold learning algorithms.

We offer empirical evidence on three different scales real-world text corpora (20 newsgroups, Amazon reviews and RCV1) and demonstrate the superiority of Disc-LDE to some state-of-the-art methods in clustering and classification tasks.

2. Related work

The pioneering manifold learning works include Locally Linear Embedding (LLE) [18] and Isometric Feature Mapping (Isomap) [19]. Since then, a graph-based manifold learning approach, Laplacian Eigenmaps (LEs) [20], shows that instinct geometrical structure can be encoded as a discrete approximation by a knn graph of scattered observation points, defined as follows:

$$\min \sum_{i,j} W_{ij}\mathrm{Dist}\left(\boldsymbol{y}_i, \boldsymbol{y}_j\right) \tag{1}$$

where $\boldsymbol{y_i}$ indicates the embeddings of sample $\boldsymbol{x_i}$ and $\boldsymbol{W_{ij}}$ indicates the edge weight connecting samples $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ in knn graph, **Dist**(.) indicates a distance metric function. To keep the instinct geometrical structure, the defined discrete approximation is used to regularize many traditional embedding methods, like pLSA, NMF, CF and AEs.

More precisely, LTM [24] and DTM [25] incorporate the discrete approximation to the optimization objective of pLSA. DTM takes the Euclidean distance to measure Dist(.), while LTM employs the Kullback-Leibler (KL) divergence as the distance metric. Moreover, to respect the full discriminability of the manifold, DTM goes further to explore negative relationships of documents [25]. Due to LTM and DTM need the explicit assignment of the edge weights, LDTM [23] introduces a local linear regression model for learning the graph Laplacian automatically, thus greatly improving the robustness of the method. Besides, the incorporation of the discrete approximation with NMF and CF also provide two improved techniques, called Graph Regularized Nonnegative Matrix Factorization (GNMF) [31] and Locally Consistent Concept Factorization (LCCF) [32]. The main benefit of both techniques is that their resulting embeddings can efficiently uncover the geometric structure of the dataset, but they are purely unsupervised methods, failing to incorporate other prior knowledge (e.g., category pairwise constraints). Recently, both CNPCF and SSCF are proposed to incorporate the category pairwise constraints related to local invariant for better embeddings, and achieve state-of-the-art performance in clustering. However, such methods are plagued by an OOS extension problem, mainly due to their inability of building an explicit map of the new samples. Therefore, in practice, they are more suitable for clustering rather than classification.

The OOS problem is also referred to as the generalization of the embeddings to new coming samples, which is one open problem within manifold learning. Some related studies focus on an idea that maintains the learned model on the training samples and extrapolate the embeddings for new coming samples by performing a cascade estimation. Similar previous works include [33,34], where [33] use the Nyström formula to speed up the computations of eigenvectors of a data kernel matrix and [34] performs an extrapolation by fitting a radial basis function interpolator. These two methods present a generalized solution for the existing methods but fail to further enhance the performance through a joint

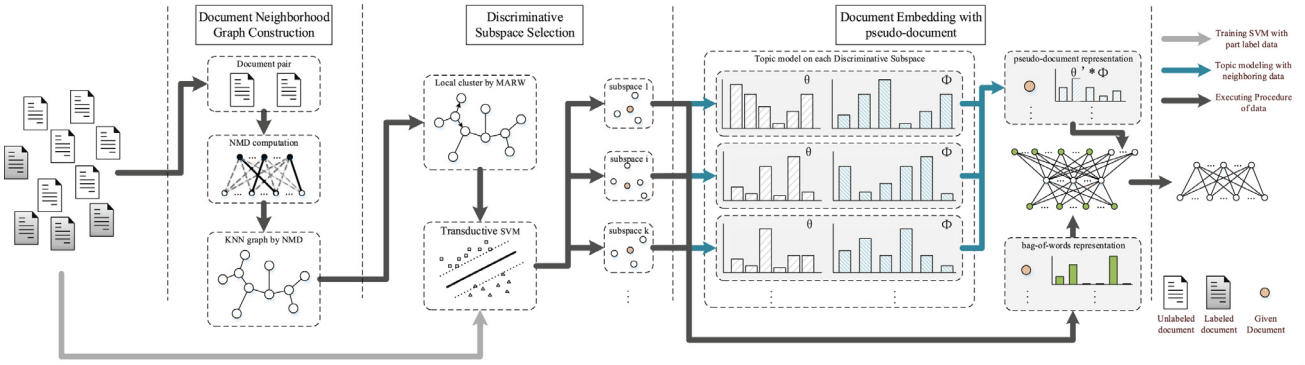

Fig. 1. The schematic diagram of Disc-LDE.

learning of the low-dimensional embeddings and the extrapolation. Therefore, another solution is to develop a unified framework of joint learning a parameterized mapping function and low-dimensional embeddings for preserving the local invariant structure of the whole data distribution. FCSPE can build an approximate linear projection through a regression process while capturing the low-dimensional embeddings over the training samples. Recently, some variations of AEs have been proposed to build the unified framework for preserving the local invariant structure, such as Denoising Auto-Encoders (DAEs) [35], Contractive Auto-Encoders (CAEs) [36], GAEs [28] and LAEs [29]. These variations can be summarized as a regularized training of AEs. Specifically, the denoising training is to recover the clean observations under a corrupted version, thus ensuring DAEs focus on the essential factors for observed distribution. CAEs incorporates the Frobenius norm of the encoder's Jacobian to the optimization objective, causing the encoding relationship become strongly contracting in the neighborhood space. However, their main purpose is to recover the input sample, which limits their ability to model the local invariance structure. For this reason, GAEs and LAEs both explicitly exploit the instinct geometrical structure of the nearest neighbor samples to regularize the training of AEs, so that the learned encoder network can directly extract the embeddings of new coming samples. Although the geometric information has been successfully used for local invariance preservation, some salient statistical information for modeling local cluster of the corpus, like documents-generating structure on a subspace, is also worthy of attention for capturing the intrinsic semantic information and improving the application of document embedding.

3. Methods

We present Disc-LDE to build a smooth affine map for document embedding by preserving documents-generating structure on a subspace. For this purpose, we incorporate the documents-generating structure as a pseudo-document by a generative probabilistic model of subspace, where the subspace is acquired by a transductive learning of multi-agent random walk on neighborhood graph, and regularize the training of AEs to recover the input document and its pseudo-document as precision as possible. The behind intuition is that if a document embedding incorporates the essential information behind the input document and its pseudo-document, then the embedding is likely to provide a good recovery. The details of Disc-LDE are introduced in three stages, neighborhood graph construction, discriminative subspace selection, document embedding on discriminative subspace, shown in Fig. 1. Supposed there are sufficient documents (so that the document manifold is well-sampled), during the first step, a knn graph is built to depict the manifold structure of the entire corpus, in which every connecting edge is weighted based on NMD of word embeddings. In the next step, Disc-LDE adaptively finds a discriminative subspace for each given document by a transductive learning of multi-agent random walk on the knn graph. At last, a pseudo-document related to given document is inferred by topic modeling on the subspace and the training of AEs is regularized through the additional recovery of pseudo-documents.

3.1. Document neighborhood graph construction

In the literature review, the neighborhood graph is commonly used to depict the manifold structure of the entire corpus, mainly due to a neighborhood graph can be seen as a discrete approximation with respect to a smooth manifold [20]. The common neighborhood graph is either the ε-neighborhood graph connecting neighbors within a radius of ε, or the knn graph connecting k-nearest neighbors. In our work, we take knn graph as our neighborhood graph, since the ε-neighborhood graph provides consistently weaker performance [37]. Let $\mathbf{G} = (\mathbf{X}, \mathbf{A})$ denotes a knn graph, where $\mathbf{X}$ is the set of vertices with each corresponding to a document $\boldsymbol{x}_i$ and $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{\mathbf{m}\times\mathbf{m}}$ is the adjacency matrix with each element corresponding to a similarity of document pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Firstly, a full adjacency matrix $\mathbf{A}$ is acquired by a similarity metric of every document pair. Then, the matrix $\mathbf{A}$ is sparsified by connecting $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with an undirected edge if $\boldsymbol{x}_i$ is among the k-nearest neighbors of $\boldsymbol{x}_j$ or if $\boldsymbol{x}_j$ is among the k-nearest neighbors of $\boldsymbol{x}_i$. A common similarity metric for document pair is the cosine distance of the bag-of-words vector. However, due to the bag-of-words representation views words independently and cannot capture the semantic relationship of individual words, leading a limitation that the cosine distance fails to give a reliable similarity measure when document pair has many synonyms but do not have the same word.

To incorporates the semantic relationships of individual words, we consider the document as an independent set of word embeddings, thus any document pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is a bipartite graph and their similarity metric can be defined as a normalized maximum matching distance of word embeddings. Specifically, let $\mathbf{D} = \{\boldsymbol{w}_i\} \in \mathbb{R}^{\mathbf{d}\times\mathbf{v}}$ denotes the set of word embeddings,[2] where $\mathbf{d}$ is the dimension of each word embedding and $\mathbf{v}$ is the number of vocabulary. Let $\Omega_i \in \mathbb{R}^{\mathbf{d}\times|\boldsymbol{x}_i|}$ denotes the set of word embedding related to document $\boldsymbol{x}_i$, where $|\boldsymbol{x}_i|$ is the number of unique words in $\boldsymbol{x}_i$. Given a pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$, the edge connecting word $\boldsymbol{w}_s \in \Omega_i$ and word $\boldsymbol{w}_t \in \Omega_j$ is weighted by a maximized softmax score, denote as δ_{st}. The softmax function takes a K-dimensional vector $\mathbf{z}$ of arbitrary real

[2] Word embeddings could be achieved by the tool of word2vec [38], furthermore, some pre-trained word embeddings are provided by GloVe [39] in here http://nlp.stanford.edu/projects/glove/.

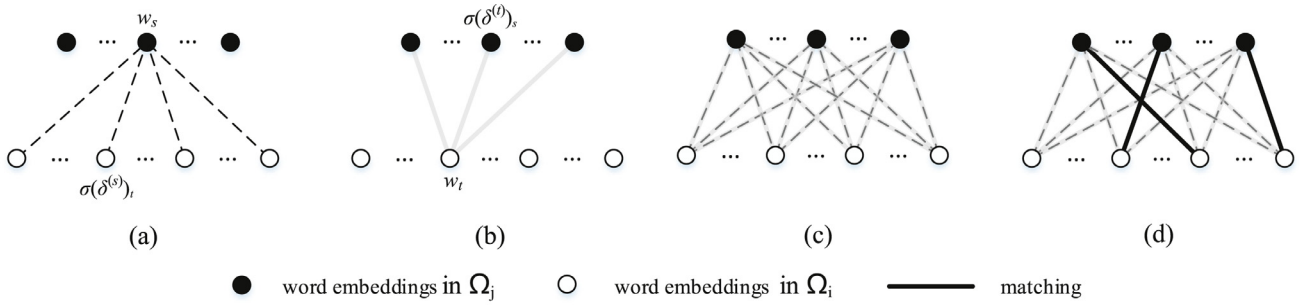

Fig. 2. (a) and (b) is an illustration of two forms of real values vector, (c) is the produced weighted bipartite graph of documents $\boldsymbol{\Omega_i}$ and $\boldsymbol{\Omega_j}$ and (d) is an example of maximum weighted bipartite matching.

values and squashes it to a K-dimensional real values vector $\sigma(\mathbf{z})$ between zero and one that sum to 1, where the jth value is given as follows,

$$\sigma(z)_j = e^{z_j} / \sum_K e^{z_k}, \tag{2}$$

As we can see, the softmax function gives a normalized probability to the values of the input vector **z,** and acts as a smoothed version of the 'max' function because, if $z_j \gg z_k$ for all $j \neq k$, then $\sigma(\mathbf{z})_j \approx 1$ and $\sigma(\mathbf{z})_k \approx 0$. For a document pair ($\boldsymbol{x}_i$, $\boldsymbol{x}_j$), there are two forms of real values vector $\mathbf{z}$, shown as Fig. 2(a) and (b), one is taken by inner product between given word embeddings $w_s \in \Omega_i$ and every word embeddings in Ω_j, denote $\delta^{(s)}$; the other one is taken by inner product between given word embeddings $w_t \in \Omega_j$ and every word embeddings in Ω_i, denote $\delta^{(t)}$. Finally, the weight δ_{st} is taken by the maximum of the corresponding softmax score,

$$\delta_{st} = \max\left(\sigma\left(\delta^{(s)}\right)_t, \sigma\left(\delta^{(t)}\right)_s\right), \tag{3}$$

Defined in this way, δ_{st} reflects the most likelihood of similarity between corresponding words in two document. After connecting all word, it produce a weighted bipartite graph related to pair ($\boldsymbol{x}_i$, $\boldsymbol{x}_j$), like Fig. 2(c). Therefore, the similarity metric between document pairs can be easily solved by finding a maximum weighted bipartite matching between the bipartite graph[3] (Fig. 2(d)). The maximum weighted bipartite matching is viewed as the assignment problem, which is find a matching of maximal sum of the connecting weight, and it can be solved by the famous Hungarian algorithm.[4] Formally, NMD is real values between zero and one, which is the average of maximum weighted bipartite matching.

$$NMD\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \frac{1}{m} \sum_{s \in \Omega_i, t \in \Omega_j} \delta_{st} \tag{4}$$

where m is the numbers of edges in matching. As we can see that NMD takes the normalization of the maximum matching, which helps to eliminate the unfairness case when the words number of document pairs does not match. Additionally, note that NMD is a metric about similarity, which means the similarity of document pairs increases as result of increasing of NMD. Algorithm I give the procedure to build the neighborhood graph in our model.

3.2. Discriminative subspace selection

Given a knn graph $\mathbf{G} = (\mathbf{X}, \mathbf{A})$, let $\mathbf{X}^c = \{(\boldsymbol{x}_i, \boldsymbol{c}_i)\}_1^n, n \ll |\mathbf{X}|$ denote a labeled subset and $L = \{1, 2, \ldots, l, \ldots\}$ denote the set of category indicator, where $\boldsymbol{c}_i \in L$ is the label of $\boldsymbol{x}_i \in \mathbf{X}$. A straightforward approach to define a discriminative neighborhood is to select the k-nearest neighbors of the same class. As shown in [41], large enough overlaps among a neighborhood may produce better global embedding, since it facilitates efficient propagation of the discriminative information along with the knn graph. Unfortunately, it may not be easy to determine the suitable k-nearest neighbors for an efficient propagation of the discriminative information. In this section, we take a local cluster as neighborhood for modeling documents-generating structure in third step, where the local cluster is defined as a subset of nearby documents that belong to the same category of given document, called discriminative subspace. Besides, we adopt the transductive support vector machine (TSVM) to improve the multi-agent random walk (MARW) [42] on knn graph that enforce all agents to pick neighboring documents of the same class as given document, denoted as Transductive-MARW.

3.2.1. Transductive-MARW

Consider a knn graph $\mathbf{G} = (\mathbf{X}, \mathbf{A})$, the moving probability of one agent random walk from vertex $\boldsymbol{u}$ to $\boldsymbol{v}$ is proportional to the edge weight $a_{u,v}$, defined as follows,

$$m_{u,v} \overset{\text{def}}{=\!=} \frac{a_{u,v}}{\sum_w a_{u,w}} \tag{5}$$

where $a_{u,v}$ indicates the above mentioned NMD score between vertexes $\boldsymbol{u}$ and $\boldsymbol{v}$. For convenient, the probabilities matrix can be represented as $\boldsymbol{M} \overset{\text{def}}{=\!=} \boldsymbol{D}^{-1}\mathbf{A}$, where $\boldsymbol{D}$ denotes the diagonal matrix of degree of all vertexes and $\boldsymbol{D}^{-1}$ is the corresponding inverse matrix. Given a start vertex $\boldsymbol{u}_0$, the initial probability is commonly formed as a column vector $\boldsymbol{P}^0 = [0, \ldots, p(\boldsymbol{u}_0), \ldots 0]$ of size $|\boldsymbol{X}|$, where $p(\boldsymbol{u}_0) = 1$. So, the probability distribution $\boldsymbol{P}^t$ at timestep t can be iteratively expressed as follows,

$$\boldsymbol{P}^t \overset{\text{def}}{=\!=} \boldsymbol{M}\boldsymbol{P}^{t-1} = \boldsymbol{M}\left(\boldsymbol{M}\boldsymbol{P}^{t-2}\right) = \cdots = \boldsymbol{M}^t\boldsymbol{P}^0 \tag{6}$$

Based on the above analysis, the transition probability for k agents defined as follows.

$$m_{U,V} \overset{\text{def}}{=\!=} f(x) = \begin{cases} 0, & \text{otherwise} \\ \sum\limits_{i \in U_k} \prod\limits_{j \in V_k} \boldsymbol{M}\left(U_i, V_j\right), & \text{if } \boldsymbol{L_U} = \boldsymbol{L_V} = \boldsymbol{L_{P^0}} \end{cases} \tag{7}$$

$\boldsymbol{U}$ and $\boldsymbol{V}$ denote two K-dimensional vectors, which represents the permutations of current vertexes and next vertexes with respect to the k agents, respectively. $\boldsymbol{L}_U$ and $\boldsymbol{L}_V$ denote the class label vector of current vertexes and next vertexes. This behavior is equivalent to perform k agents' random walks independently, but all agents are permitted to move a new step simultaneously when satisfying the constraint that current vertexes and next vertexes belong to the same category.

To improve the prediction accuracy of the above constraint, we adopt TSVM to determine whether all agents satisfy above constraint, mainly due to it successfully takes into account the distribution information behind the unlabeled examples. TSVM makes

[3] Alternatively, Word Mover's Distance consider this issue as Transportation problem which can be solved by the Earth Mover's Distance with time complexity $O(V^3 \log\{V\})$ [40].

[4] The running time of the Hungarian algorithm is up to $O(V^2E)$ using the Bellman–Ford algorithm for shortest path search, or achieve $O(V^2 \log\{V\} + VE)$ with the Dijkstra algorithm and Fibonacci heap.

Algorithm I. neighborhood graph construction.

1: **Input**: k is the nearest neighbors numbers and $\mathbf{X} = \{\boldsymbol{x}_i\}_1^m$, where $\boldsymbol{x}_i$ is bag-of-words vector
2: **For** each document pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$
3: Calculate the similarity weight δ by expression (3);
4: Find maximum weighted bipartite matching of document pair $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ by Hungarian algorithm;
5: Calculate the NMD$(\boldsymbol{x}_i, \boldsymbol{x}_j)$ by expression (4);
6: **End for**
7: Connect the k-nearest neighbors and weighted the edge according to NMD$(\boldsymbol{x}_i, \boldsymbol{x}_j)$;
8: **Return:** knn graph $\mathbf{G} = (\mathbf{X}, \mathbf{A})$

the decision by constructing a non-linear separating hyperplane $f = \mathbf{K}w^{\mathrm{T}} + b$, where $\mathbf{K}$ is the kernel function. It can be achieved by solving the following minimization problem.

$$\min_{f, y_u} J(f) + C\sum_{i\in X^l} L(y_i o_i) + C^* \sum_{j\in X^u} L\left(y_j o_j\right) \tag{8}$$

where $L(\cdot)$ is the hinge loss like $L(y_i o_i) = \max(0, 1 - y_i o_i)$. y_i is the class label and o_i is output of decision function f. $J(f)$ indicates the inverse of the geometric separation margin. In our work, we take radial basis function (RBF) as the kernel function to construct a non-linear separating hyperplane and $J(f) = 1/2w^{\mathrm{T}}\mathbf{K}w$. Our improvement derives from two considerations:

(1) To guide the current separating hyperplane could move to the optimal orientation gradually, the transduction should adhere to the semantic relation behind knn graph that neighboring documents are likely to share similar category information.
(2) The inference of every timestep should simultaneously consider the semantic information conveyed by each category, preventing the separating hyperplane to move in the direction of large scale category.

Specifically, given a knn graph $\mathbf{G} = (\mathbf{X}, \mathbf{A})$, the scheme of Discriminative Subspace Selection is summarized as follows.

Step 0: Consider the labeled subset $\mathbf{X}^c$ including $|L|$ categories, an initial separating hyperplane $H_0 : \mathbf{K}w^{\mathrm{T}} + b = 0$ is founded by inductive learning with fixed parameter C.

Step 1: For every category $\Omega^{l\in L} = \{(\boldsymbol{x}_i, \boldsymbol{c}_i)\}_{c_i=l}$, stochastically pick a start document $\boldsymbol{U} \in \Omega^l$ with initial probability $\boldsymbol{P}^0(\boldsymbol{U})$ and start a k agents random walk at $\boldsymbol{U}$. Therefore, it could produce $|L|$ multi-agent random walk synchronously, each one is denoted as MARW^l.

Step 2: For each MARW^l, at timestep $t = \{i\}_1^{last}$, pick all permutations of neighboring documents $\boldsymbol{V}$ around current document $\boldsymbol{U}$ and execute a labeling by H_{t-1}, and retain the permutations that the labeling is consistency to current document, denote as $\aleph^l(t)$.

Step 3: For each MARW^l, at timestep $t = \{i\}_1^{last}$, execute a transductive inference by optimizing expression (8) with respect to the $\mathbf{X}^c$ and each labeling permutation in $\aleph^*(t)$ with a same temporary penalty parameter C^*, denote as C_t^*, then find one labeling permutation that provide the biggest decreasing as start vertex at next timestep, denote this permutation as $\aleph^*(t)$ and the corresponding separating hyperplane as H_t.

Step 4: repeat steps 2 and 3 at new start vertex, but execute a new transduction with a decreasing penalty parameter C_t^*. The iterative process is finished when $t = last$ or $C_t^* \ll C^*$, and for every MARW^l, the subspace is achieved as the union of permutation at every timestep but excluding start documents $\boldsymbol{U}$, denote as $\aleph_i^l = \{\boldsymbol{x}_s | \boldsymbol{x}_s \neq \boldsymbol{U}$ and $\boldsymbol{x}_s \in \bigcup_{t=1}^{last} \aleph^*(t)\}$.

Step 5: according to the achieved result at step 4, the labeled subset $\mathbf{X}^c$ is expanded by some reasonable labeling examples, like non-support vector examples, and repeat the whole scheme using the expanded labeled subset for another start document.

3.2.2. More discussions

In the scheme, an initial separating hyperplane based on the labeled subset $\mathbf{X}^c$ is used to execute a labeling at next step. Hence, it is very important to provide a high-accuracy labeling by H_0 for improving the effect of transduction of hyperplane at successive step. In practice, one could increase the penalty parameter C to pursue the hyperplane with lower tolerance of misclassification. However, this will lead to a more complex boundary and may increase the risk of over-fitting [43]. Alternatively, one could pick those non-support vector examples as start vertexes since they have higher confidence in labeling than support vector examples.

Note that, at step 2, the permutations include a case that some agents may pick the same neighboring document. Therefore, this behavior is essentially an oversampling process and could provide balance labeling examples for transduction at next timestep, avoiding the degradation of labeling because of data imbalance. Furthermore, at step 1, if we cannot find any permutations of neighboring documents that satisfied the constraint, we will pick another one from that category. In addition, the scheme is suitable for an implement of parallelization, which is helpful for improving the transduction efficiency.

At step 4, the decreasing of penalty parameter C_t^* will produce a contraction of neighborhood, reducing the probability of escaping from current subspace. Since the decreasing of penalty parameter enforce a large width margin, which push the labeling example far away the separating hyperplane. Consider the minimization problem shown as expression (8), the hinge loss reflects a misclassification that the loss increase linearly when output $o_i < 1$. Stated differently, the last term plays a role of slack variable to allow the separating hyperplane have a soft margin so that it could handle the overlap of different categories. In practice, this influence is reflected by the corresponding regularization parameters C^*. For example, if they are fixed a large value, a low tolerance of misclassification is allowed for the model, which means the minimization of (8) concerns more about the distance minimization of misclassified examples from the hyperplane than the margin maximization. In particular, in the case of $C^* \to inf$, the model are equivalent to the hard-margin that almost all of the data points are classified correctly, but the hyperplane may be sensitive to outlier data; on the contrary, if they are small, the model will focus on the margin maximization, allowing higher deviations from the margin. As we all known, the distance from support vectors (SVs) to the separating hyperplane determines the margin width. Fig. 3 is an example of binary classification by soft-margin SVM with a changing regularization parameter. As we can see, when increasing the regularization parameter, the margin width become thin, such as, when $C = 1000$ the hyperplane become sensitive to outliers.

Intuitively, Transductive-MARW could find the local cluster around given vertex more accurately compared to the simple random walk, since it is less probable that all agents simultaneously jump over the bottleneck of a graph than only one agent. This intuition could be formally proved by showing the spectral gap of Transductive-MARW is smaller than the simple random walk. Based on the discrete version of Cheeger inequality, a clustering quality metric known as conductance is related to the spectral gap, revealing that a small spectral gap implies that a graph contains a

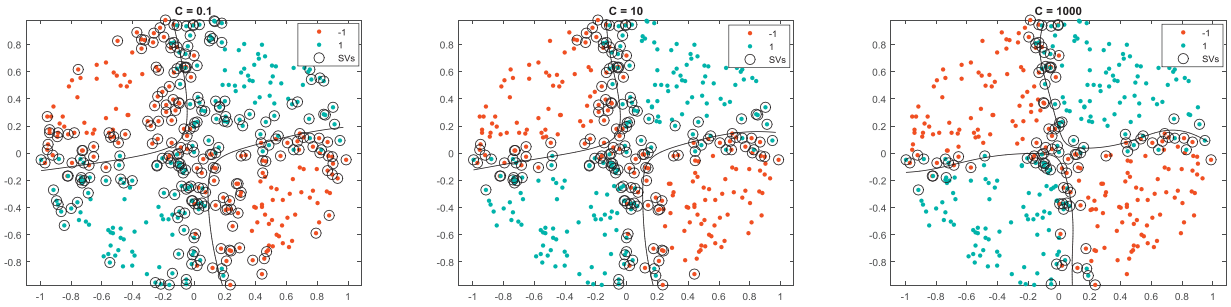

Fig. 3. example of influence of changing penalty parameter for overlap case of two categories. SVs are marked with circle and the solid line indicates the separating hyperplane.

set of vertices with low conductance.[5] Let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_n$ are the eigenvalues of $\boldsymbol{M}$, and the spectral gap is defined as the difference between the moduli of the two largest eigenvalues $\lambda_1 - \lambda_2$, where $\lambda_1 = 1$, because $\boldsymbol{M}$ is positive semi-definite. According to the corollary in [44], a lower bounds of spectral gap for weighted graph is represented as follows,

$$1 - \lambda_2 \geq \frac{vol(G)w_{min}}{d_{\max}^2 |\gamma_{max}| b} \tag{9}$$

where $vol(G) = \sum_i d_i$ denote the volume of graph and d_i is the degree of vertex. w_{min} denote the minimal edge weight and d_{max} is the maximal degree. $|\gamma_{max}|$ indicates the maximal number of edges in any canonical paths connecting two vertices in graph and b is defined as the maximal load over all edges,

$$\mathrm{b} \sim= \max_{\{edge\ e\}} |\{\gamma_{uv} | e \in \gamma_{uv}\}| \tag{10}$$

The behave difference between Transductive-MARW and simple random walk is the edge weights, as the weight $w_{u,v}$ is replaced by $w_{u,v}^k$. Hence, the difference of spectral gap depends on the minimal and maximal weights and degrees. Formally, assume the spectral gap is β and the corresponding symbols of simple random walk are denoted with a tilde, then we have:

$$\tilde{w}_{min} = w_{min}^k \tag{11}$$

$$\tilde{d}_{max} = d_{max}^k \tag{12}$$

$$\mathrm{vol}(G)w_{min} = w_{min} \sum_e d_e \leq w_{min}|E|w_{max} \leq |E| d_{max}^2 \tag{13}$$

$$\mathrm{vol}\left(\tilde{G}\right) = \sum_e \tilde{w}_e = \sum_e w_e^k \geq \frac{\left(\sum_e w_e\right)^k}{|E|^{k-1}} = \frac{vol(G)^k}{|E|^{k-1}} \tag{14}$$

$$\tilde{\beta} \geq \frac{vol\left(\tilde{G}\right)\tilde{w}_{min}}{\tilde{d}_{max}^2 |\gamma_{max}| b} \geq \frac{vol(G)^k w_{min}^k}{|E|^{k-1} d_{max}^{2k} |\gamma_{max}| b} = \beta \left(\frac{vol(G) w_{min}}{|E| d_{max}^2} \right)^{k-1} := \beta (G)^{k-1} \tag{15}$$

According to the expression (13), one can show that $G \leq 1$, so $\beta \leq \tilde{\beta}$. Hence, we conclude that the spectral gap is smaller than simple random walk and the more discriminative local cluster could be found with better theoretical guarantee. An illustration of MARW and Transductive-MARW on two moon dataset is presented in Fig. 4. As we can see that some ambiguous samples connect two moons, causing a degradation of MARW, but Transductive-MARW achieves a better performance because of the use of the neighboring Transductive learning strategy.

[5] The conductance is a ratio of the number of edges with one end in S (subsets of vertices of graph) and one end out of S. A low conductance indicates good cluster quality because it means that few edges connect the cluster to the rest of the graph, and many edges connect vertices inside the cluster to each-other.

3.3. Document embedding with discriminative subspace

In order to incorporate salient statistical information of the discriminative subspace into the building of a smooth affine map, we encode the probabilistic generative information of documents as a pseudo-document by LDA, and regularize the training of AEs to recover the input document and its pseudo-document as precision as possible.

3.3.1. Pseudo-document inference by LDA

The pseudo-document inference derives from probabilistic generative process of topic modeling on discriminative subspace, which computes the word distribution as the combination of topic-word distribution and document-topic distribution $p(w|d) = \sum_K p(w|z_k)p(z_k|d)$. It consists of a partial vocabulary rather than the entire vocabulary of corpus, because some words are very likely to appear in documents that not in current subspace, like different categories documents. This reduces the scale of the characterization of semantic structure and may relieve the entanglement and redundancy of variation around the subspace. Consider the corpus $\mathbf{X} = \{\boldsymbol{x}_i\}_1^m$ with a vocabulary $\boldsymbol{V}$, given a document $\boldsymbol{x}_i$ of a category $\Omega^l = \{\boldsymbol{x}_i\}_{c_i=l}$, let $\aleph_i^l = \{\boldsymbol{x}_n\}_{c_n=l, n\neq i}$ denote the subset of discriminative subspace around $\boldsymbol{x}_i$ but excluding $\boldsymbol{x}_i$, where $\boldsymbol{x}_i \in \mathbb{R}^{\boldsymbol{V}\times 1}$ indicates the bag-of-word vector, let $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{\boldsymbol{V}_i \times 1}$ indicates the bag-of-word approximation of pseudo-document, where $\boldsymbol{V}_i \subseteq \boldsymbol{V}$ denote the vocabulary consisting of $\aleph_i^l$ and $\boldsymbol{x}_i$. Let $\boldsymbol{K}_i$ denote the number of topics.

Note that document $\boldsymbol{x}_i$ is excluded from the subspace $\aleph_i^l$, thus we adopt LDA for topic modeling on discriminative subspace, since it provides well-defined inference procedures for previously unseen documents and achieved a better performance for document modeling [7]. From the view of smoothed LDA, the pseudo-document $\tilde{\boldsymbol{x}}_i$ related to $\boldsymbol{x}_i$ can be inferred as follows:

1. Draw topic proportions $\theta_i|\alpha \sim \mathrm{Dir}(\alpha)$, where $\mathrm{i} \in \{1, \ldots, m\}$, and $\mathrm{Dir}(\alpha)$ indicates a Dirichlet distribution on $\boldsymbol{K}_i$-vector parameter $\alpha = (\alpha_k)_1^{\boldsymbol{K}_i}$.
2. Draw word proportions $\varphi_k|\beta \sim \mathrm{Dir}(\beta)$, where $\mathrm{k} \in \{1, \ldots, k_i\}$, and $\mathrm{Dir}(\beta)$ indicates a Dirichlet distribution on $\boldsymbol{V}_i$-vector parameter $\beta = (\beta_w)_1^{\boldsymbol{V}_i}$.
3. For each of the word positions $\mathrm{i} \in \{1, \ldots, m\}$ and $\mathrm{j} \in \{1, \ldots, |\boldsymbol{V}_i|\}$
   a. Pick a topic $z_{i,j}|\theta_i \sim \mathrm{Multinomial}(\theta_i)$.
   b. Pick a word $w_{i,j}|\varphi_k \sim \mathrm{Multinomial}(\varphi_k|k = z_{i,j})$.

where θ_i indicates the topic distribution for pseudo-document $\tilde{\boldsymbol{x}}_i$ with a Dirichlet prior parameter α, and φ_k indicates the word distribution for topic k, and $z_{i,j}$ indicates the topic for the jth word in $\tilde{\boldsymbol{x}}_i$ and $w_{i,j}$ indicates the specific observed word in $\tilde{\boldsymbol{x}}_i$.

Based on the above process, the pseudo-document inference can be formalized as $p(w|d) = \theta_d \times \varphi$, where $\varphi = [\varphi_1, \ldots, \varphi_{k_i}]^{\mathrm{T}}$ indicates the word distribution of specific topics and θ_d indicates a

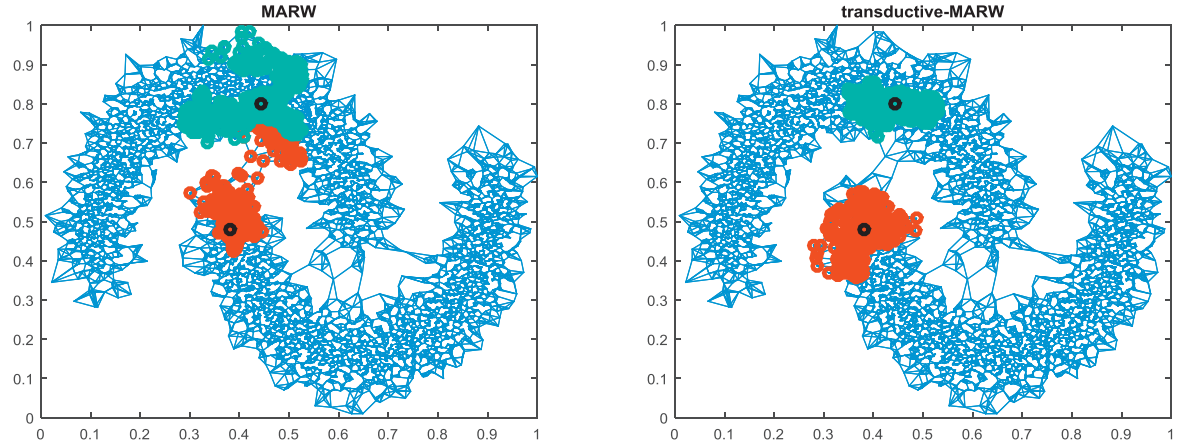

Fig. 4. illustration of local cluster on two moon dataset with MARW and Transductive-MARW. The black circle denotes the start vertex and the red circle denote the visited vertex of agent.

topic inference on pseudo-document. A common practice is to estimate θ_d by fixing the φ to the ones estimated during model fitting on discriminative subspace. Even with a fixed φ, the estimation is intractable due to the interaction between a large space of discrete random variables, and an efficient estimation method is proposed based on Markov Chain Monte Carlo sampling, such as collapsed Gibbs sampling. The sampling method involves iterating over a set of observed words, updating the topic assignment for each word conditioned on the topic assignments for the remaining words.

Formally, let $N_{s|q}$ indicate the count that topic z_s is assigned in document d_q, and $N_{t|s}$ indicate the count that word w_t is generated by topic z_s. Combining the Dirichlet prior, the joint probability of the given subspace $w \in \boldsymbol{V}_i$ and a set of corresponding latent topics $z \in \boldsymbol{K}_i$ is:

$$\mathrm{P}(w, z|\alpha, \beta) = \prod_{q \in \aleph_i^l} \frac{\Delta\left(N_{\cdot|q} + \alpha\right)}{\Delta(\alpha)} \cdot \prod_{s \in \boldsymbol{K}_i} \frac{\Delta\left(N_{\Delta|s} + \beta\right)}{\Delta(\beta)} \text{ where}$$
$$N_{\cdot|q} = \left\{N_{s|q}\right\}_{s=1}^{\boldsymbol{K}_i} \text{ and } N_{\cdot|s} = \left\{N_{t|s}\right\}_{t=1}^{\boldsymbol{V}_i}, \tag{16}$$

Based on the expression (16), we can derive the update equation of Gibbs sampler that the topic assignment of one word conditioned on the topic assignment of remaining observed words,

$$\mathrm{P}\left(z_i = k|z_{\neq i}, w\right) = \frac{\mathrm{P}(w|z)}{\mathrm{P}\left(w_{\neq i}|z_{\neq i}\right)\mathrm{P}(w_i)} \cdot \frac{\mathrm{P}(z)}{\mathrm{P}\left(z_{\neq i}\right)} \propto \frac{N_{t|s,\neq i} + \beta_t}{\sum_{t \in \boldsymbol{V}_i} N_{t|s,\neq i} + \beta_t} \cdot \frac{N_{s|q,\neq i} + \alpha_s}{\left[\sum_{s \in \boldsymbol{K}_i} N_{s|q} + \alpha_s\right] - 1} \tag{17}$$

Based on the expectation of the Dirichlet distribution $Dir(a) = a_i / \sum_i a_i$, the point estimation for the word proportions $\hat{\varphi}_{t|s}$ during the model fitting is

$$\hat{\varphi}_{t|s} = \frac{N_{t|s} + \beta_t}{\sum_{t \in \boldsymbol{V}_i} N_{t|s} + \beta_t} \tag{18}$$

After fitting the model, the topic inference for unseen document d could be produced with a fixed $\hat{\varphi}_{t|s}$, the sampling update equation thus be modified as

$$\mathrm{P}\left(z_i = k|z_{\neq i}, w, \hat{\varphi}_{t|s}\right) \propto \hat{\varphi}_{t|s} \cdot \frac{N_{s|d,\neq i} + \alpha_s}{\left[\sum_{s \in \boldsymbol{K}_i} N_{s|d} + \alpha_s\right] - 1} \tag{19}$$

Similar to the expression (19), the point estimation of the topic inference for given unseen document $\hat{\theta}_{s|d}$ is

$$\hat{\theta}_{s|d} = \frac{N_{s|d} + \alpha_s}{\sum_{s \in \boldsymbol{K}_i} N_{s|d} + \alpha_s} \tag{20}$$

3.3.2. Learning an affine map from the regularized training of AEs

Given a document $\boldsymbol{x}_i$, let $\boldsymbol{h}_i \in \mathbb{R}^{\boldsymbol{H} \times 1}$ denotes the intermediate output of hidden layer of AEs, where $\boldsymbol{H}$ is the dimension of hidden layer, let $\boldsymbol{r}_i \in \mathbb{R}^{\boldsymbol{V} \times 1}$ denotes the recovery of $\boldsymbol{x}_i$ and $\tilde{\boldsymbol{s}}_i \in \mathbb{R}^{\boldsymbol{V}_i \times 1}$ indicates the recovery related to the pseudo-document $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{\boldsymbol{V}_i \times 1}$, where $\boldsymbol{V}_i \in \boldsymbol{V}$ denotes the vocabulary consisting of $\aleph_i^l$ and $\boldsymbol{x}_i$. Based on the view of AEs, Disc-LDE consists of two components: *encoder* and *decoder*. The *encoder* converts an input vector $\boldsymbol{x}_i$ into embeddings $\boldsymbol{h}_i$, whose form is an affine map followed by a nonlinearity:

$$\boldsymbol{h}_i = f_{\boldsymbol{W},\boldsymbol{b}}(\boldsymbol{x}_i) = \sigma\left(\boldsymbol{W} \cdot x_i + \boldsymbol{b}\right). \tag{21}$$

While the *decoder* converts the embeddings $\boldsymbol{h}_i$ back to a recovery $\boldsymbol{r}_i$ related to the input vector and an approximation $\tilde{s}_i$ related to the pseudo-document, whose form is an affine map followed by a nonlinearity:

$$[\boldsymbol{r}_i;\ \tilde{\boldsymbol{s}}_i] = g_{\boldsymbol{W}^\dagger,\boldsymbol{W}^\ddagger,\boldsymbol{c}^\dagger,\boldsymbol{c}^\ddagger}(\boldsymbol{h}_i) = \sigma\left(\left[\boldsymbol{W}^\dagger; \boldsymbol{W}^\ddagger\{\boldsymbol{l} \in \boldsymbol{V}_i\}\right] \cdot \boldsymbol{h}_i + \left[\boldsymbol{c}^\dagger;\ \boldsymbol{c}^\ddagger\{\boldsymbol{l} \in \boldsymbol{V}_i\}\right]\right). \tag{22}$$

where $\sigma(\ \cdot\ )$ is sigmoid function $\sigma(a) = 1 + \exp\{a\}^{-1}$, $\boldsymbol{b} \in \mathbb{R}^{\boldsymbol{H} \times 1}$ is bias vector of hidden layer, and $\boldsymbol{W} \in \mathbb{R}^{\boldsymbol{H} \times \boldsymbol{V}}$ is encoder weights connecting input layer and hidden layer. $\boldsymbol{c}^\dagger,\ \boldsymbol{c}^\ddagger \in \mathbb{R}^{\boldsymbol{V} \times 1}$ is bias vector of output layer and $\boldsymbol{W}^\dagger = \boldsymbol{W}^\ddagger = \boldsymbol{W}^{\mathrm{T}} \in \mathbb{R}^{\boldsymbol{V} \times \boldsymbol{H}}$ is weight matrix of decoder, but $\boldsymbol{W}\dagger$ is part of the weights for input recovery in decoder and $\boldsymbol{W}\ddagger$ is part of the weights for pseudo-document approximation in decoder. This definition reduce the scale of estimated parameters and can make it harder for the encoder to stay in the linear regime of its nonlinearity without paying a high price in reconstruction error [35]. Note that, due to the synthetic pseudo-document consists of a partial vocabulary that appeared in the given document and its subspace, for example, given a pseudo-document $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^{\boldsymbol{V}_i \times 1}$ where $\boldsymbol{V}_i \subseteq \boldsymbol{V}$, thus $\tilde{\boldsymbol{x}}_i$ can be approximated using a thinner part-based network. Let $\boldsymbol{W}\ddagger\{\boldsymbol{l} \in \boldsymbol{V}_i\}$ and $\boldsymbol{c}\ddagger\{\boldsymbol{l} \in \boldsymbol{V}_i\}$ denote the corresponding weights, where $\boldsymbol{l}$ indicates the index of row vector and $\boldsymbol{Matrix}\{condition\}$ indicates a submatrix satisfied the condition. While the remaining weights $\boldsymbol{W}\ddagger\{\boldsymbol{l} \notin \boldsymbol{V}_i\}$ and $\boldsymbol{c}\ddagger\{\boldsymbol{l} \notin \boldsymbol{V}_i\}$ will keep silent and stop update during the training, more details will be discussed below.

Different from the basic AEs, Disc-LDE adopts the pseudo-document as an additional recovery object, thus the optimization objective is a minimization of Empirical risk on observed corpus $\mathbf{X} = \{\boldsymbol{x}_i\}_1^m$ using the method of regularization.

$$\mathcal{J}(\theta|\mathbf{X}) = \underbrace{\frac{1}{m}\sum_{i\in\mathbf{X}}\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{r}_i; \theta)}_{Empirical\ \ cost} + \underbrace{\frac{\lambda}{2m}\sum_{i\in\mathbf{X}}\|\tilde{\boldsymbol{s}}_i - \tilde{\boldsymbol{x}}_i\|^2}_{Regularizer} \tag{23}$$

where $\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{r}_i; \theta)$ denotes a non-negative real-valued loss function to measure how different the recovery $\boldsymbol{r}_i$ from the expected output $\boldsymbol{x}_i$. Here, the squared error $\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{r}_i; \theta) = \boldsymbol{r}_i - \boldsymbol{x}_i{}^2$ is used to measure the reconstruction loss. The alternative choice is the cross-entropy loss $\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{r}_i; \theta) = -\sum_{s=1}^{|\boldsymbol{x}_i|} x_{is}\log(r_{is}) + (1 - x_{is})\log(1 - r_{is})$ when expected output is binary.

The addition of the *regularizer* is intended to smooth the affine map to the documents-generating structure. Since Transductive-MARW can provide subspaces with a large ratio overlap for some documents closely among neighborhood, thus facilitates the efficient sharing of documents-generating structures between their pseudo-document. Hence, the regularization ensures the smoothness of the encoding affine map for capturing the variation along the documents-generating structure of the neighborhood. Through an appropriate choice of the regularization parameter λ, the learnt affine map achieves a tradeoff between "fidelity" of the observation and "smoothness" of the documents-generating structure. In particular, for the limiting case $\lambda \to 0$, it implies the affine map is completely dependent on the observation like basic AEs. On the contrary, when $\lambda \to \infty$, it implies that the affine map focus on the "smoothness" of the documents-generating structures shared by neighboring samples.

The ultimate goal of Disc-LDE is to find the parameters $\theta^* \in \{\boldsymbol{W}^\dagger, \boldsymbol{W}^\ddagger, \boldsymbol{W}, \boldsymbol{c}^\dagger, \boldsymbol{c}^\ddagger, \boldsymbol{b}\}$ to build an affine map so that the optimization objective $\mathcal{J}(\theta|\mathbf{X})$ is minimal. Given a observed corpus $\mathbf{X} = \{\boldsymbol{x}_i\}_1^m$, the parameters estimation problem can be solved via stochastic gradient descent (SGD) in the form of error Back-propagation. Here, we compute the true gradient of the cost function on a small random segment of training samples (mini-batch) rather the entire training samples (batch), which is a common compromise between convergence stability and time-efficiency. To avoid confusion, we give the following notation for the description of parameters estimation at first.

d_i	size of input layer
d_h	size of hidden layer
d_r	size of reconstruction
d_s	size of approximation
$x_t(j)$	the jth value of specific input $\boldsymbol{x}_t$
$r_t(j)$	the jth value of reconstruction $\boldsymbol{r}_t$
$\tilde{s}_t(j)$	the jth value of approximation $\tilde{\boldsymbol{s}}_t$
$\tilde{x}_t(j)$	the jth value of pseudo-document $\tilde{\boldsymbol{x}}_t$
$h_t(j)$	the jth value of hidden layer
$\delta_j^{(l)}$	local gradient of jth neuron node in lth layer
$o_i^{(l)}$	The output of ith neuron in lth layer
$w_{ij}^{(l)}$,	weight connecting ith neuron in $(l-1)^{th}$ layer to jth neuron in lth layer
b_i	the bias connecting jth neuron in hidden layer (or 1st layer)
c_j	the bias connecting jth neuron in output layer (or 2nd layer)
λ	non-negative regularization hyper-parameter

Assuming a mini-batch consisting of n training instances, the parameters are modified according to SGD algorithm as follows.

$$\theta_{new}^{(l)} = \theta_{old}^{(l)} - \eta\left[\frac{1}{n}\Delta\theta^{(l)}\right] = \theta_{old}^{(l)} - \eta\left[\frac{1}{n}\sum_{i=1}^{n}\nabla(\mathcal{J}(\theta|\mathbf{X}))\right] \tag{24}$$

where η is learning rate, and $\nabla(\cdot)$ is partial derivatives of corresponding parameters. Therefore, the key step of parameters estimation is to compute the partial derivatives of the object function. The back-propagation algorithm provides an efficient way using the chain rule to iteratively compute gradients for each layer. Specifically, for a specific sample $\boldsymbol{x}_t$, the objective function is

$$\mathcal{J}(\theta|\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_t) = \frac{1}{2}\boldsymbol{r}_t - \boldsymbol{x}_t{}^2 + \frac{\lambda}{2}\tilde{\boldsymbol{s}}_t - \tilde{\boldsymbol{x}}_t^2 \tag{25}$$

According to expressions (21) and (22) and Chain rule, we have $\nabla(\theta_{ij}) = \delta_j^{(l)} \times o_i^{(l-1)}$, where $\delta_j^{(l)}$ reflects the required changes in synaptic weights, and its computation depends whether neuron is from hidden layer or output layer.

$$\delta_j^{(l)} = \begin{cases} (r_t(j) - x_t(j))(1 - r_t(j))r_t(j) + 1\{\mathrm{j} \in d_s\}\lambda(\tilde{s}_t(j) \\ \quad - \tilde{x}_t(j))(1 - \tilde{s}_t(j))\tilde{s}_t(j), \ \ l = 2 \\ \left(\sum_{k\in\{d_r, d_s\}}\delta_k^{(l)}w_{jk}^{(l)}\right)(1 - h_t(j))h_t(j), \ \ l = 1 \end{cases}$$

where $1\{\mathrm{j} \in d_s\}$ is indicator function, so that $1\{\text{a true statement}\}=1$, and $1\{\text{a false statement}\}=0$. As we can see that, when the neuron is from output layer, the required changes may derive from two aspects: the input reconstruction and the pseudo-document approximation, since, in *decoder*, the weights $\boldsymbol{W}\dagger$ and $\boldsymbol{c}\dagger$ that are "responsible" for input reconstruction are shared by a weights submatrix like $\boldsymbol{W}\ddagger\{\boldsymbol{l} \in \boldsymbol{V_i}\}$ and $\boldsymbol{c}\ddagger\{\boldsymbol{l} \in \boldsymbol{V_i}\}$. In particular, for a specific sample $\boldsymbol{x}_t$, the word that not appear in $\tilde{\boldsymbol{x}}_i$ (or $1\{\mathrm{j} \in d_s\} = 0$), and the $\delta_j^{(2)}$ just depends on the difference between $\boldsymbol{x}_t$ and $\boldsymbol{r}_t$. Hence, this procedure gives the parameters an inequitable chance for modification, which produces various thinner part-based networks to govern the approximation of specific pseudo-documents and enrich the diversity of global embedding for the document manifold. Besides, the part-based networks usually contain less variables need to be approximated, which may suppress irrelevant variations and relieve the entanglement and redundancy of variation around the local geometry. Considering the case of tied weights, the parameters modification is accumulation of each layer, thus we have,

$$\begin{aligned}\nabla \mathrm{w}_{ij}^{(2)} = \nabla \mathrm{w}_{ij}^{(1)} = x_t(i)&\left(\sum_{k\in\{d_r, d_s\}}\delta_k^{(l)}w_{jk}^{(l)}\right)(1 - h_t(j))h_t(j) \\ &+ h_t(i)[(r_t(j) - x_t(j))r_t(j)(1 - r_t(j)) \\ &+ 1\{\mathrm{j} \in d_s\}\lambda(\tilde{s}_t(j) - \tilde{x}_t(j))\tilde{s}_t(j)(1 - \tilde{s}_t(j))]\end{aligned} \tag{27}$$

$$\begin{aligned}\nabla c_j = &(r_t(j) - x_t(j))(1 - r_t(j))r_t(j) + 1\{\mathrm{j} \in d_s\}\lambda(\tilde{s}_t(j) \\ &-\tilde{x}_t(j))(1 - \tilde{s}_t(j))\tilde{s}_t(j)\end{aligned} \tag{28}$$

$$\nabla b_i = \left(\sum_{k\in\{d_r, d_s\}}\delta_k^{(l)}w_{jk}^{(l)}\right)(1 - h_t(i))h_t(i) \tag{29}$$

Now, taken together, the procedure of parameters estimation Algorithm II can be described as follow.

4. Experiments

Our experimental goal is to evaluate the improvements of the built smooth affine map for uncovering discriminative document embeddings. For this purpose, we compared the performance of the proposed method for clustering and classification to the state-of-the-arts methods by three experiments setting.

In the 1st experiment, we built the smooth affine map of Disc-LDE by training set and extract various dimensions embeddings (50, 100, 150, 200, 250 and 300) as representation of test documents; then conducted the intra-class compactness evaluating by

Algorithm II. Parameters estimation for Disc-LDE.

Input: The entire training set $\mathbf{X}$.
initialize parameters $\{\boldsymbol{W}, \boldsymbol{c}\dagger, \boldsymbol{c}\ddagger, \boldsymbol{b}\}$
While not stopping criterion **do**
  Set $\Delta w^{(l)} = 0$, $\Delta b^{(l)} = 0$ and $\Delta c^{(l)} = 0$ for all l
  Randomly choose a mini-batch $\{\boldsymbol{x}_i\}_1^n$
  For each $\boldsymbol{x}_i$ in mini-batch $\{\boldsymbol{x}_i\}_1^n$
  Perform a feed-forward pass, computing the activations of the hidden layer and output layer;
  Compute the partial derivatives with respect to the input as Eqs. (27), (28) and (29)
  Compute the modification with respect to the parameters: $\Delta\theta^{(l)} = \Delta\theta^{(l)} + \nabla\theta^{(l)}$
  End for
  Update parameters: $\theta_{new}^{(l)} = \theta_{old}^{(l)} - \eta[\frac{1}{n}\Delta\theta^{(l)}]$
End while
Return the parameter of embedding map $\{\boldsymbol{W}, \boldsymbol{c}\dagger, \boldsymbol{c}\ddagger, \boldsymbol{b}\}$

K-means and 1-nearest neighbor (1-NN),[6] and the inter-class separability evaluating by normalized cut (NCut) and Support Vector Machine (SVM); meanwhile, offered a visualization of some better embeddings by t-Distributed Stochastic Neighbor Embedding (t-SNE) [45]. In particular, when building the knn graph, we selected the 30-nearest neighbors based on the average performance from five choice ($k = 5,\ 15,\ 30,\ 45,\ 55$). For the Transductive-MARW, the initial separating hyperplane $H_0 : \mathbf{K}w^{\mathrm{T}} + b = 0$ of TSVM was founded with incremental training, i.e. Disc-LDE-10%, Disc-LDE-30%, Disc-LDE-50%, where the percentage indicates the ratio of labeled set from training set, then execute a transduction inference for next labeling. Besides, for the pseudo-document inference, the topic number is determinate based on the measure in [46]. At last, the training of AEs is regularized with mini-batch ($\mathrm{n} = 100$), tied weights and sigmoid activation function, where the regularization parameter is fixed $\lambda = 100$ based on the average performance with four choice of the regularization parameter ($\lambda = 1,\ 10,\ 100$ and 1000).

During the 2nd experiment, to better present the contribution of each stage of Disc-LDE, we also try to investigate the improvement of discrimination as the first setting by performing an alternative procedure of each basic stage. In particular, when constructing knn graph, we employed the cam weighted distance to weight all documents pairs (denoted as cam-Disc-LDE), where the cam weighted distance is also a robust similarity metric that could describe the attraction and repulsion effects from their deformed neighbors by a deflective cam contours [47]. For discriminative neighborhood selection, we utilized the 30-nearest neighbors of the same class based on the knn graph to characterize local geometry (denoted as nn-Disc-LDE). Finally, the regularization of parallel recovery of all documents of discriminative subspace [48] is used to encode the local geometrical information (denoted as pr-Disc-LDE). Here, we compared the above alternative methods with the case Disc-LDE-50%, thus, for pr-Disc-LDE and cam-Disc-LDE, we utilize 50% labeled set to start our Transductive-MARW for a fair comparison. Nevertheless, for nn-Disc-LDE, it is a type of fully-supervised setting that all documents need to be labeled.

In the 3rd experiment, we aim to show the availability of the Transductive-MARW for other neighborhood-based manifold learning methods. Hence, the LLE is extended by Transductive-MARW (denoted as Disc-LLE) to compare the supervised-LLE (SLLE) [49]. Since LLE is a basis of neighborhood-based methods, we believe that the usage of Transductive-MARW can be suitable for other neighborhood-based manifold learning methods by investigating the performance of two extensions of LLE in this experiment. The comparison also includes clustering (K-means and Ncut) and classification (1-NN and SVM). In particular, we employed the inclusive approach to solve the OOS problem for classification. For SLLE, we computed the embeddings under various hyper-parameters of controlling the distance matrix as $\alpha = 0.1,\ 0.5,\ 1$. For Disc-LDE, we also built the initial separating hyperplane $H_0 : \mathbf{K}w^{\mathrm{T}} + b = 0$ of TSVM with incremental training, denoted as Disc-LLE-10%, Disc-LLE-30%, Disc-LLE-50%.

[6] The K-means clustering algorithm is based on cosine similarity, while 1-NN classification algorithm based on the Euclidean distance.

4.1. Datasets

We offer empirical evidence on three different scales real-world dataset (20 newsgroups, Amazon reviews, and RCV1). The 20 newsgroups[7] is a collection of newsgroup, and has become a popular corpus in documents applications, like documents clustering and classification. It consists of 20 different newsgroups, and some of the newsgroups are very closely related to each other, like *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*, while some others are unrelated, such as *soc.religion.christian* and *misc.forsale*. The Amazon reviews[8] consists of reviews of various products crawled from amazon, including 142.8 million reviews spanning May 1996 - July 2014. What we are interested is to predicate the category of products according to the content of the reviews, thus we selected 8 product category for our evaluation, such as “Books”, “Electronics”, “Movies and TV”, “Kindle Store”, “Video Games”, “Apps for Android”, “Digital Music” and “Musical Instruments”. Most of product category are imbalanced and include some noises. The RCV1[9] is a large multi-class dataset, which is an archive of over 800,000 manually categorized newswire stories recently made available by Reuters [50]. We extracted 2 sub-categories of documents from topic “MARKETS”, such as M131 (INTERBANK) and M132 (FOREX). Table 1 presents some statistical information about three data sets, where C is number of categories and B denotes the mean and standard deviation of the number of samples for one category. V is the vocabulary size, $\bar{D}$ denotes the mean and standard deviation of the number of words consisting of a document, and $\mathbf{X}_{train}$ is a number of training set, while $\mathbf{D}_{test}$ is number of test set.

4.2. Evaluation metrics

The clustering results for 5 shuffled data set were evaluated by normalized mutual information (NMI). Formally, the definition of NMI is as follows,

$$\mathrm{NMI}(C, T) = \frac{\mathrm{MI}(C, T)}{\max(H(C), H(T))} \tag{30}$$

where $C = \{C_i\}$ and $T = \{T_i\}$ be the set of clusters obtained from the ground truth and the assigned clusters by the clustering algorithm,

[7] https://sites.google.com/site/renatocorrea02/textcategorizationdatasets/.
[8] http://jmcauley.ucsd.edu/data/amazon/.
[9] http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm.

Table 1
statistical information of 3 corpus.

Corpus	C	B	$\mathbf{X}_{train}$	$\mathbf{X}_{test}$	V	$\bar{D}$
20 newsgroups	20	940.85 ± 96.44	12,000	6817	8156	75.31 ± 65.29
Amazon reviews	8	12,875 ± 4980	90,000	13,000	85,685	81.32 ± 72.36
RCV1	2	25,000 ± 0	40,000	10,000	47,236	71.24 ± 67.6

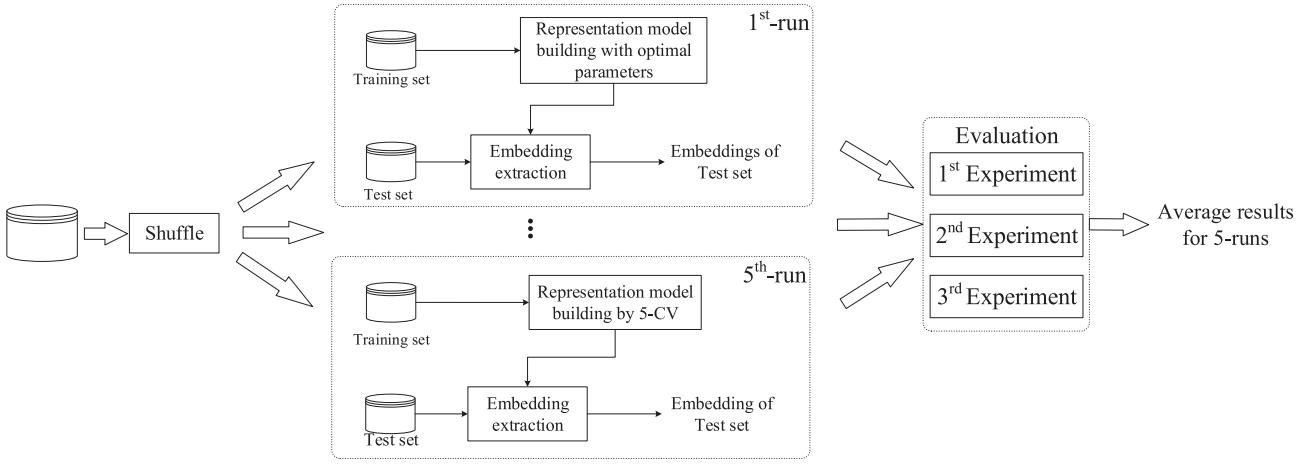

Fig. 5. The scheme of the experimental setup.

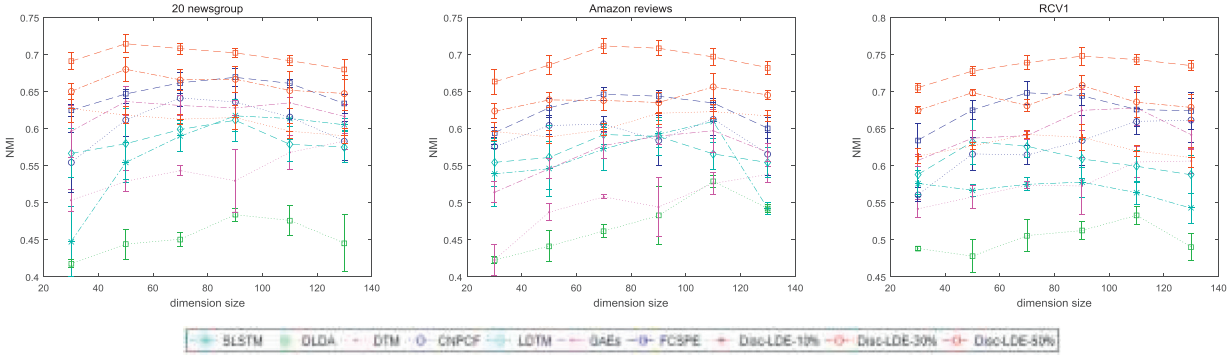

Fig. 6. NMI of K-means on three datasets under various embeddings, where each point consist of the mean value and standard deviations over 5 runs.

$H(C)$ and $H(T)$ are the entropies of C and T, respectively. MI(C, T) denotes their mutual information, which is measured as,

$$\mathrm{MI}(C,T)=\sum_{C_i\in C,T_i\in T} p(C_i,T_i)\log\frac{p(C_i,T_i)}{p(C_i)p(T_i)} \tag{31}$$

where $p(C_i, T_i)$ is the joint probability that $x_i \in C_i$, T_i at the same time, and $p(C_i)$ and $p(T_i)$ indicate the probability that $x_i \in C_i$ and $x_i \in T_i$, respectively.

Considering the imbalance of various categories, we used the weighted F-measure $\bar{F}$ to estimate the final accuracy of both classification models, which is calculated as follows,

$$\bar{F}=\frac{\sum_i c_i F_i}{C} \tag{32}$$

where c_i indicates the ratio of samples in categories i in test set, and C is the size of test set. The F_i is the F-measure of categories i, it is closely related to the precision P_i and recall R_i, which are defined as follows.

$$R_i=\frac{|\{relevant\ documents\}|\cap|\{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \tag{33}$$

$$R_i=\frac{|\{relevant\ documents\}|\cap|\{retrieved\ documents\}|}{|\{relevant\ documents\}|} \tag{34}$$

$$F_i=2\frac{P_i\cdot R_i}{P_i+R_i} \tag{35}$$

The $\bar{F}$ represents a weighted average of the classes' F-measure, where the higher score indicates the better classification performance.

4.3. Experimental details

To get a fair view of the performance, we performed 5 runs on the three datasets. We first randomly shuffled all corpus 5 times and divided each corpus into $\mathbf{X}_{train}$ and $\mathbf{X}_{test}$ as Table 1, where $\mathbf{X}_{train}$ is used for learning embedding map then capture the embeddings of $\mathbf{X}_{test}$ for the following evaluation, let $\mathbf{Y}_{test}$ denote the embeddings of $\mathbf{X}_{test}$. In particular, in the 1st run, 5-fold cross-validation (CV) was executed with $\mathbf{X}_{train}$ for selection of the optimal parameters of all comparative methods. At last, another 4 runs were conducted on remaining 4 shuffled datasets with the optimal parameters estimated in the 1st run. For clustering, we utilized K-means and 1-NN to group $\mathbf{Y}_{test}$ by fixing the cluster numbers as same as

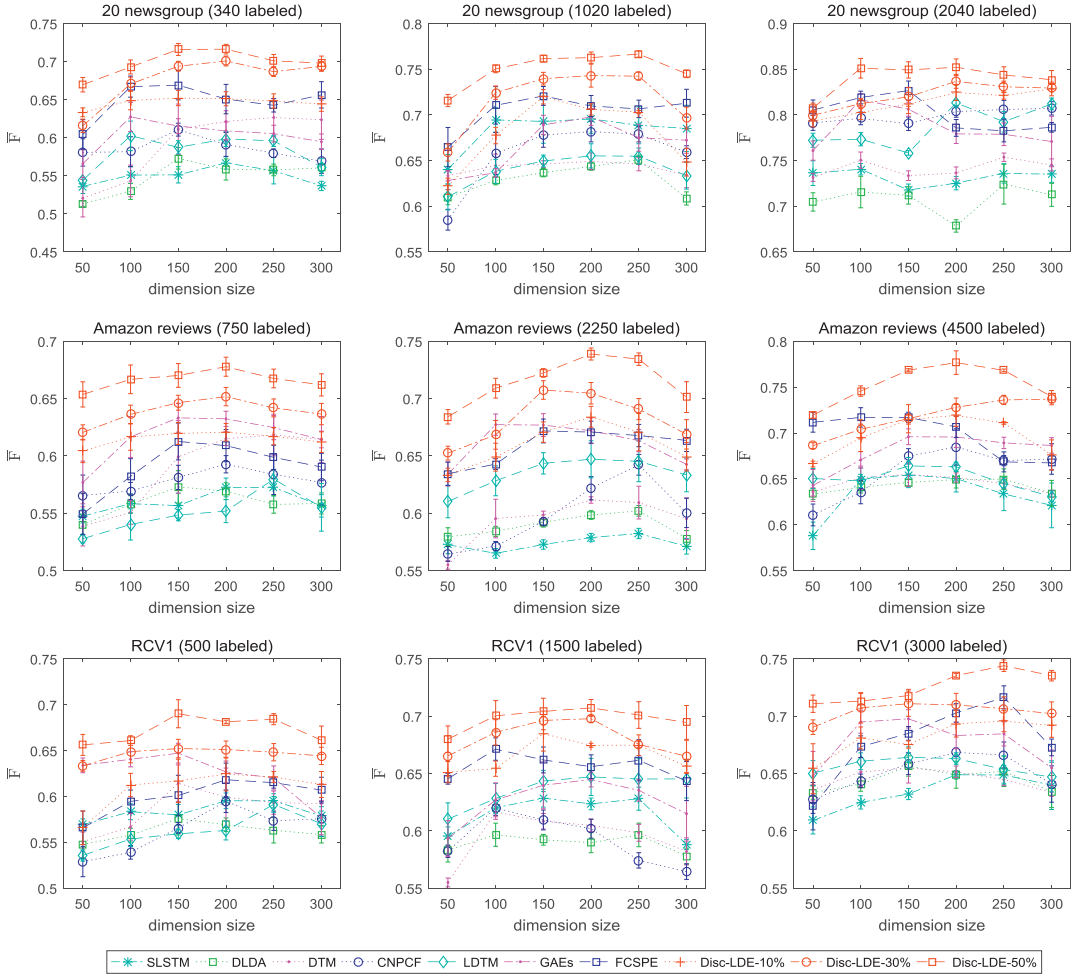

Fig. 7. Average performance of 1-NN on three datasets under various embeddings. The 1-NN classifer is built with incremental training, where each point consists of the mean value and standard deviations over 5 runs.

true classes so that avoid the impact bringing by cluster numbers. For classification, we randomly divided $\mathbf{Y}_{test}$ into two halves, i.e. test set $\mathbf{Y}_{test}(test)$ and training set $\mathbf{Y}_{test}(train)$, and built classifier by incremental training like 10%, 30%, 60%, which means the ratio of labeled set from the training set $\mathbf{Y}_{test}(train)$.The scheme of the experimental setup is shown as Fig. 5.

In the 1st experiment, we evaluated the performance of Disc-LDE as above setting and provided a comparison with two traditional methods, i.e. SLSTM [15] and DLDA [51]; three manifold-inspired embedding methods, i.e. CNPCF [21], LDTM [23] and DTM [25]; and two embedding methods by building a parameterized mapping function, i.e. GAEs [28] and FCSPE [30]. Both DLDA and SLSTM are fully-supervised embedding methods, thus they took advantage of the labels of all samples on the training set for their models learning in each run. On the contrary, CNPCF and FCSPE are semi-supervised methods, they were permitted to utilize half of label information of the training set for a fair comparison to Disc-LDE. For LDTM, DTM and GAEs, a implementation inspired by MFA was adopted to incorporate the information of class label into their construction of knn graph [52]. In detail, a connecting edge was introduced between documents of the same category and an edge was eliminated between documents of different categories. In particular, DTM further consider the information of non-neighboring pairs by eliminating an edge between documents of the same category and adding between documents from different categories. Besides, LDTM, DTM and CNPCF cannot give a specific mapping function for the test documents, we thus employed the inclusive approach that rebuilds proximity matrices with test documents, and retrained the model based on the matrices [25].

4.4. Results and discussions

4.4.1. Discriminative embeddings evaluating

Intra-class compactness evaluating: Fig. 6 present three line charts of the clustering results (NMI) of K-means using three datasets, where each point consist of the mean value and standard deviations over 5 runs. As we can see that Disc-LDE-50% can consistently outperform others state-of-the-arts methods (GAEs and FCSPE) on various dimensions of embeddings, and the standard deviations of NMI varies smaller compared with other manifold-inspired methods (LDTM, DTM and CNPCF). Besides, when we utilize 30% labeled set to start our transductive SVM for discriminative subspace selection, Disc-LDE-30% also has the better NMI on a specific dimension than others document embedding methods, i.e. 68.04% in 50-dimension for 20 newsgroups, 65.64% in 110-dimension for Amazon reviews, 70.84% in 90-dimension for RCV1. As we all known, K-means is a common clustering method that partition some samples into one cluster according to the similarity to their mean. Hence, the above evidence demonstrates that Disc-LDE is capable of capturing some better embeddings to improve the similarity metric for documents of the same class. Finally, for an imbalanced and noises dataset, like Amazon reviews, Disc-LDE also achieves satisfactory results, which demonstrate the robustness of our method.

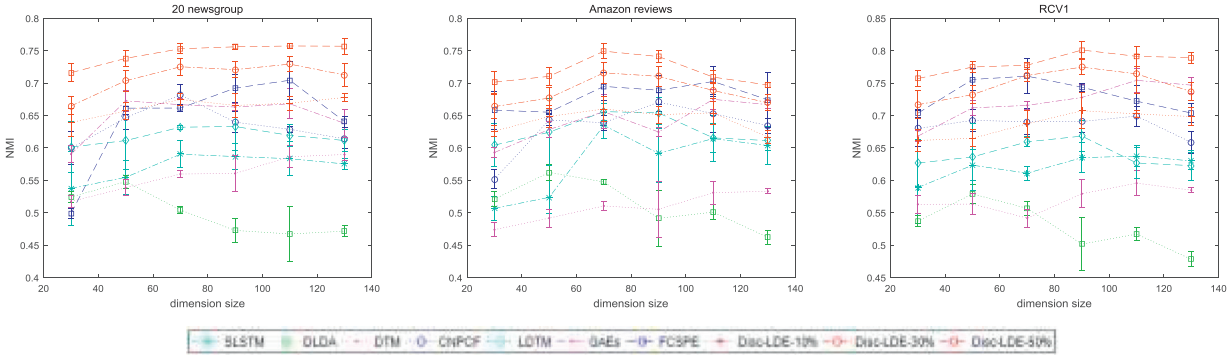

Fig. 8. NMI of Ncut on three datasets under various embeddings, where each point consist of mean value and standard deviations over 5 runs.

Fig. 7 indicate the line charts of $\bar{F}$ of 1-NN classifier with incremental training as 10% (left), 30% (middle), 60% (right). From Fig. 7, we see some similar performance as the above K-means. Specifically, Disc-LDE-30% can outperform almost all comparative methods on three datasets. As a type of instance-based learning, or lazy learning classification algorithm, 1-NN predicates the label of a sample to the category of its closest neighbor in the feature space. Hence, the better $\bar{F}$ is sufficient to indicate that some documents from the same category are assigned a more similar representation in the embedding space. Besides, the variance of the standard deviations of various dimensions are also smaller than other neighborhood-based methods when the samples for training classifier is small, like 10%, which also demonstrate that the proposed method has good robustness in practical application.

Inter-class separability evaluating: Fig. 8 shows the average NMI of Ncut over 5 runs on three datasets, where the point indicates the mean value as well as standard deviations. As we can see, it presents a significant improvement of NMI than other document embedding methods, which demonstrate that the goodness of partition for document embeddings of test set. Recalling the definition of Ncut, it is a disassociation measure and its computation depends on the total edge within a cluster and the total edge between various clusters [53]. The optimal partition criteria of spectral clustering is to minimize the connections weights between the clusters and maximize the connections weights within the clusters. Therefore, the goodness of partition may derive from the enhancement of the compactness of intra-class on one hand, which is beneficial for maximizing the connections weights within the clusters. On the other hand, the improvement presented in Fig. 8 may further derive from the minimization of the connections weights between clusters by enlarging inter-class separability.

Additionally, from the average performance shown in Fig. 9, we can see that when we utilize 10% labeled set (Disc-LDE-10%) to start our transductive SVM for neighborhood selection, Disc-LDE could achieve the comparable performance to the other document embedding methods. Considering the fully-supervised setting of other comparative methods, i.e. DLDA, SLSTM, LDTM, DTM and GAEs, this evidence shows the superiority of our method. Besides, the variation of standard deviations of various dimensions is also less compared with other neighborhood-based methods (LDTM and DTM). Since the construction of SVM classifier is to seek a separating hyperplane with a large margin, the better performance demonstrates that the better $\bar{F}$ of SVM derives from the improvement to inter-class separability by our method.

Visualization of some better embeddings: Besides, to assess the discrimination of document embeddings intuitively, we also presented the distributions of some better embeddings (GAEs, FCSPE and Disc-LDE-50%) by producing their 2D scatter plot using, shown in Fig. 10. As the visualization illustrated, for all datasets, our method could present the majority of the categories in a cluster more clearly than other comparative methods, whereas, for Amazon reviews, due to the imbalance of various categories, FCSPE and GAEs mix some categories with different categories, like "Musical Instruments" marked by pink cross, but Disc-LDE still provide a clear cluster presentation.

Based on all of the above evaluations, a conclusion is that Disc-LDE could provide discriminative embeddings so that enhance inner-class compactness as well as enlarging inter-class separability. Moreover, our methods could also provide an improvement in a semi-supervised setting, which is of great value in practical application. The evidence is sufficient to demonstrate that our approach can capture the embeddings of intrinsic semantic information. The superiority of Disc-LDE derives from our smooth affine map efficiently incorporate the document-generating structure behind the discriminative subspace. Specifically, the recovery of specific pseudo-document will regularize the parameters updating of AEs, thereby guide various part-based networks to keep "sensitive" to the documents-generating information of discriminative subspace. Hence, when handling such observations that locate closely along the structure of the documents-generating, the affine map tends to assign more similar embeddings, due to their neighborhoods (or subspaces) have a large ratio overlap, and may arouse more related thinner networks to work for capturing the embeddings. Besides, the accumulation of various thinner networks could enrich the diversity of global embedding. For example, for such observations that are far away along the structure of the documents-generating, due to the less sharing of the mutual latent topic pattern, the built affine map may assign various embeddings.

4.4.2. Contribution from each component

Figs. 11 and 12 indicate the average results of clustering and classification on all corpus by performing an alternative procedure of each basic stage for Disc-LDE-50%. From the comparison between Disc-LDE-50% and cam-Disc-LDE in Figs. 11 and 12, we can see the performance deriving from the NMD measure is better than cam weight distance, which demonstrates the NMD could provide a more reasonable connectivity structure to depict the semantic relationships. Since it incorporates the semantic knowledge encoded in the word embeddings, and could uncover the similarity of given pairs even they do not have the same word. Moreover, pr-Disc-LDE achieves a worse performance than our method, since the parallel reconstruction of all observations of discriminative subspace just plays a role of averaging the geometrical information of the observations, like location in the bag-of-word space, which fails to capture the intrinsic structure of the document-generating of the discriminative subspace, such as latent topic. The evidence is sufficient to demonstrate that, except for the geometrical information, the documents-generating structure of discriminative sub-

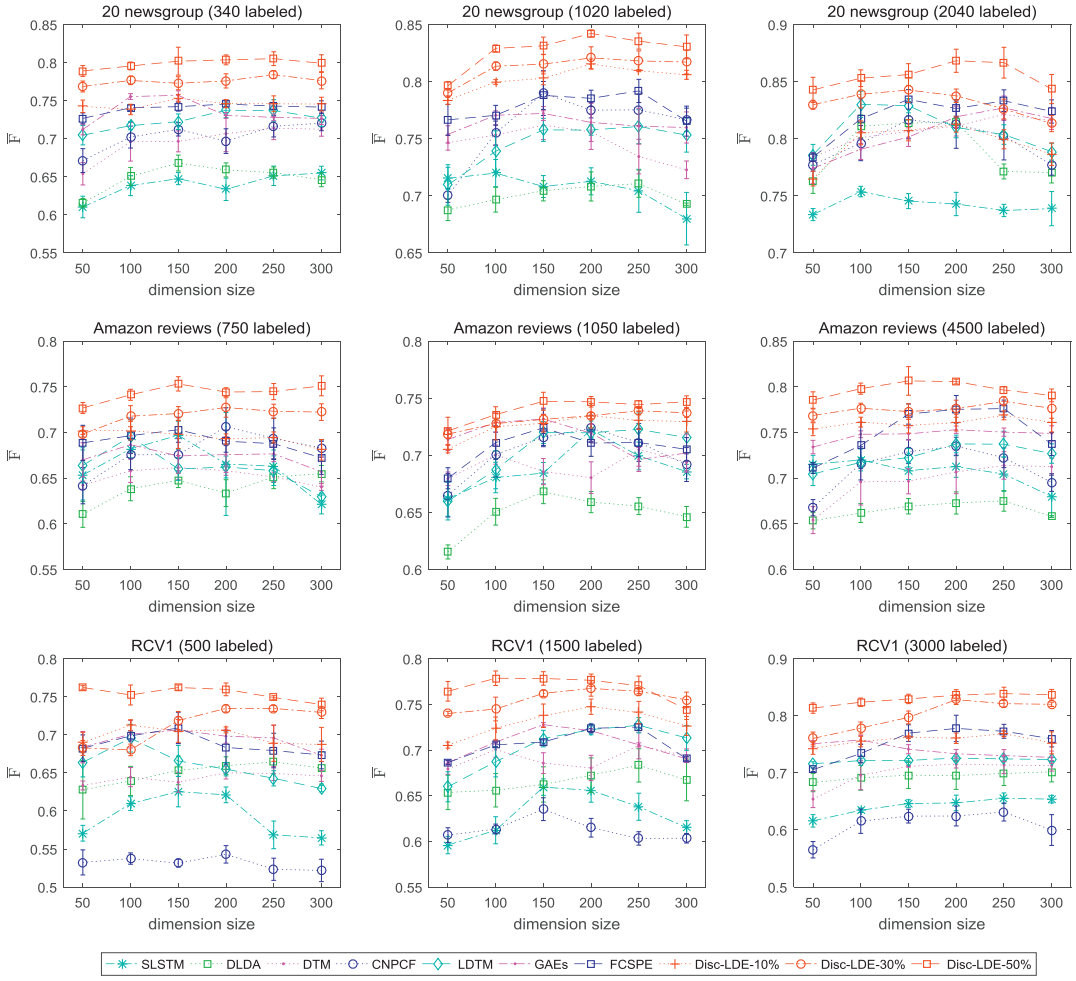

Fig. 9. Average performance of SVM on three datasets under various embeddings. The SVM classifer is built with incremental training and each point consists of mean value and standard deviations over 5 runs.

space is also worthy of attention for capturing the intrinsic semantic information.

Additionally, from Figs. 11 and 12, we observe that nn-Disc-LDE presents the most serious of degradation, which reveals the improvement heavily depends on the selection of neighborhood. Furthermore, the difference in performance between nn-Disc-LDE and Disc-LDE-50% implies a long-range propagation of the semantic information of discriminative subspace plays a positive role in providing global discriminative embeddings. Although Disc-LDE takes a semi-supervised strategy for subspace selection, but the actual purpose of Transductive-MARW and knn is to find a discriminative neighborhood where the elements are of the same category. Hence, the difference of both methods is the ratio of overlap of subspaces. Specifically, nn-Disc-LDE adopts an alternative solution of subspace selection that takes the nearest neighbors of the same class based on the knn graph. The selection strategy just considers the direct relation between the neighbors and given document $\boldsymbol{x}_i$, but fails to capture the indirect relation between the neighbor's neighbor and given document $\boldsymbol{x}_i$, thus, there is little overlap for two neighborhood, which leads to a limitation of long-range propagation of the semantic information of discriminative subspace. Whereas, Transductive-MARW guarantee a long-range propagation of the local semantic information, since it could adaptively provide a large overlap subspace by taking into account the indirect relation via Random Walk. For two adjacent documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ on the local field of manifold, a large overlap subspace provides them an easy chance to share local semantic pattern and may produce a similar pseudo-document for the following regularized training of AEs, so that inject more smoothness of the built affine map. Considering two nearby documents $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ from *comp.graphics* of 20 newsgroup, we illustrate their pseudo-document $\tilde{\boldsymbol{x}}_i$ or $\tilde{\boldsymbol{x}}_i$ as an image using the full range of colors in the colormap and pick the top word based on the value, shown in Fig. 13. From the drawing, we observe that the pseudo-document based on Transductive-MARW (bottom) consists of more words than knn (top), and they share more semantic.

4.4.3. Discussion of discriminative neighborhood selection

Figs. 14 and 15 present the average results of clustering and classification on all corpus by Disc-LLE and SLLE. We see that Disc-LLE gives a good performance, in particular, Disc-LLE-50% outperforms SLLE consistently. This evidence demonstrates our Transductive-MARW could be used for LLE framework, or as a separate module to promote the effect of neighborhood-dependent methods. Besides, although SLLE allows the class label information give a partially supervised impact by a varying hyper-parameter $\alpha \in [0, 1]$, it requires the entire training data set is labeled. Furthermore, the performance is sensitive to the selection of the value of α. For example, the 0.5-SLLE achieves better performance in most case, but for Amazon reviews, the 0.1-SLLE outperform than 0.5-SLLE. This may be related to the imbalance of category of Amazon reviews, since the imbalance give a bias of large category and produce an unreliable measure of various categories. On the contrary, our Transductive-MARW allows a semi-supervised setting, and re-

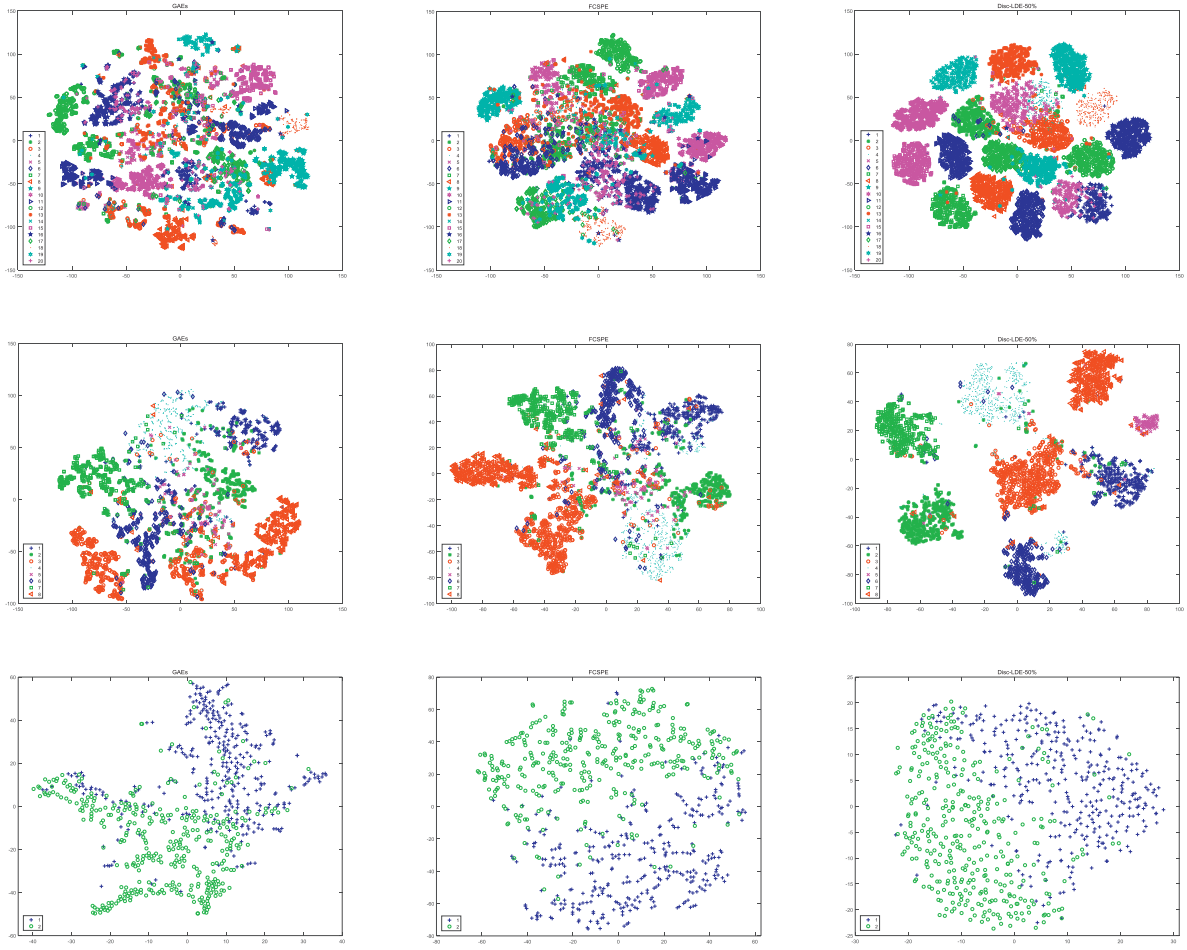

Fig. 10. 2D scatter plot of various embeddings on three datasets. Each dot represents a document and each marker indicates a category. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

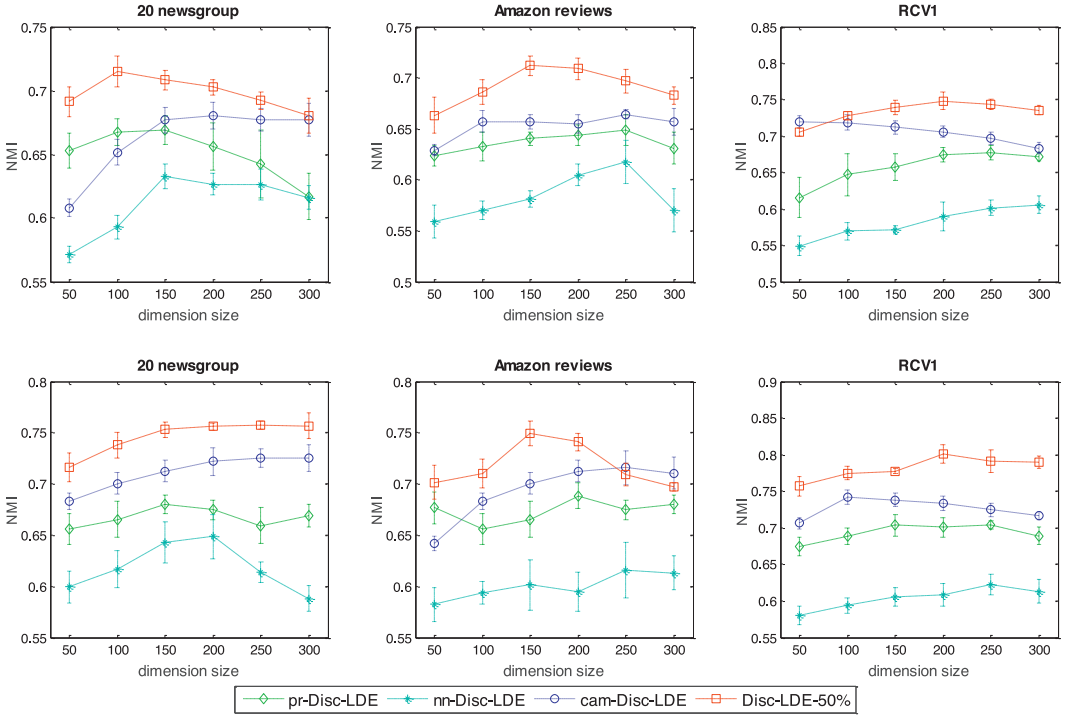

Fig. 11. NMI of K-means (top) and Ncut (bottom) on three datasets using various embeddings, where each point consist of mean value and standard deviations over 5 runs.

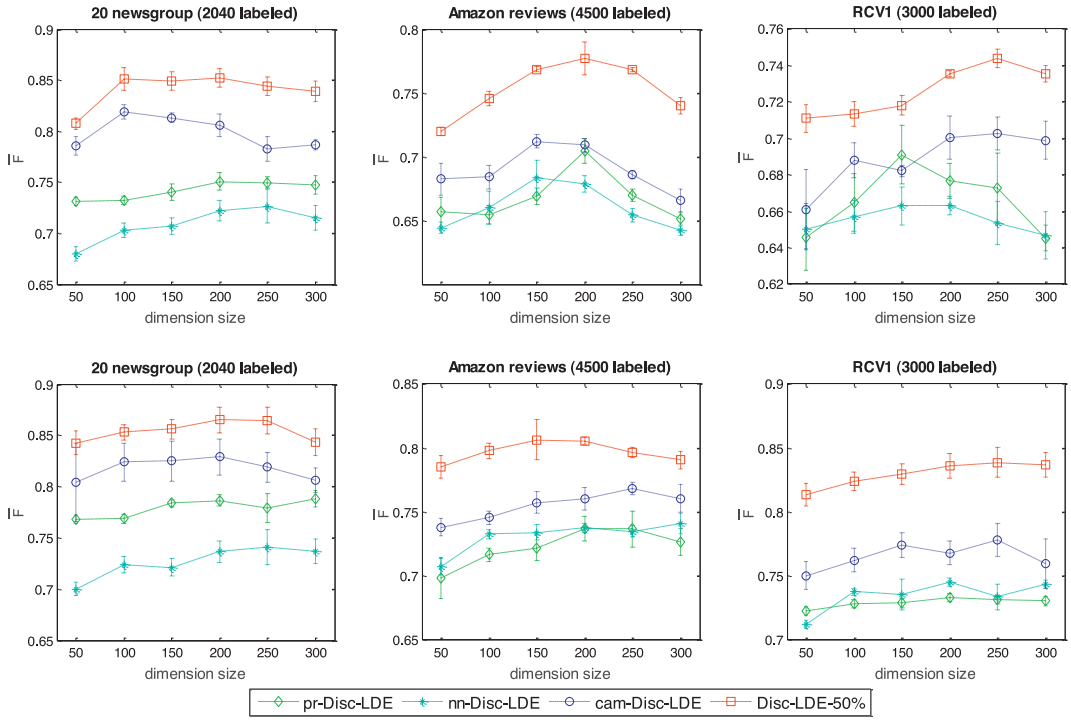

Fig. 12. Average performance of 1-NN (top) and SVM (bottom) on three datasets under various embeddings, where each point consist of mean value and standard deviations over 5 runs.

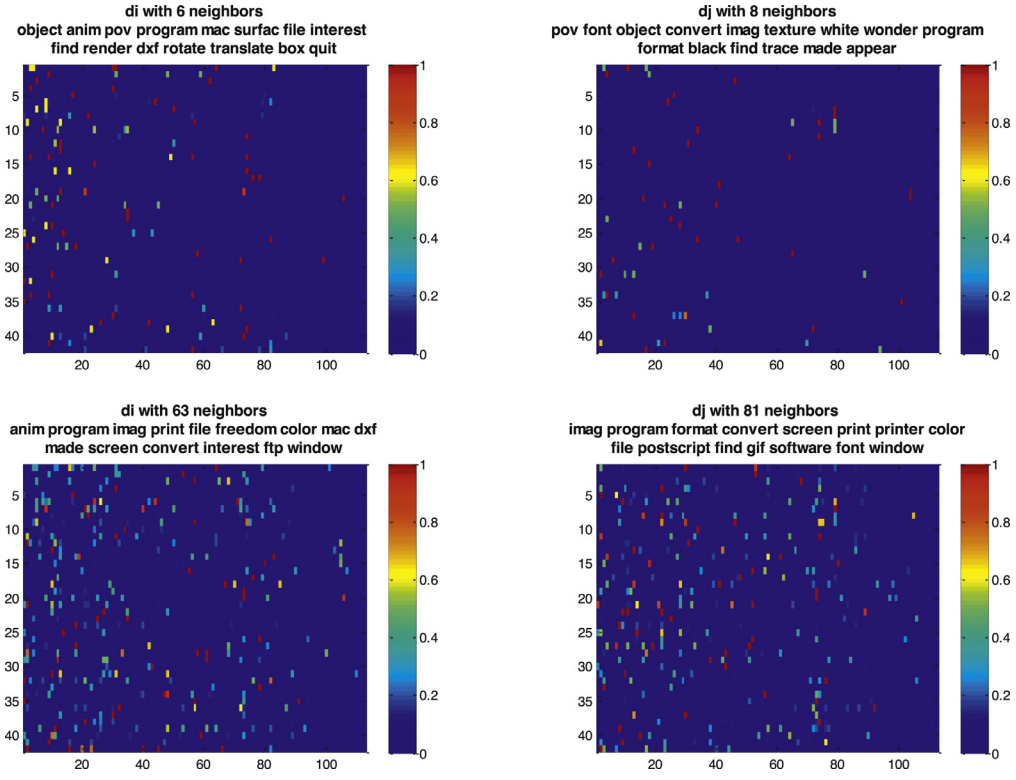

Fig. 13. Image of two pseudo-document from 20 newsgroup based on knn (top) and Transductive MARW (bottom). (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

lieves the degradation of labeling because of data imbalance, since it permits a repeat selection of the same sample at one picking, which is essentially an oversampling process and could provide balance labeling examples for next transduction.

5. Conclusions and future work

In this paper, we propose a semi-supervised local-invariant document embedding method to build a smooth affine map for document embedding by preserving documents-generating structure on a subspace, called Discriminative Locally Document Embed-

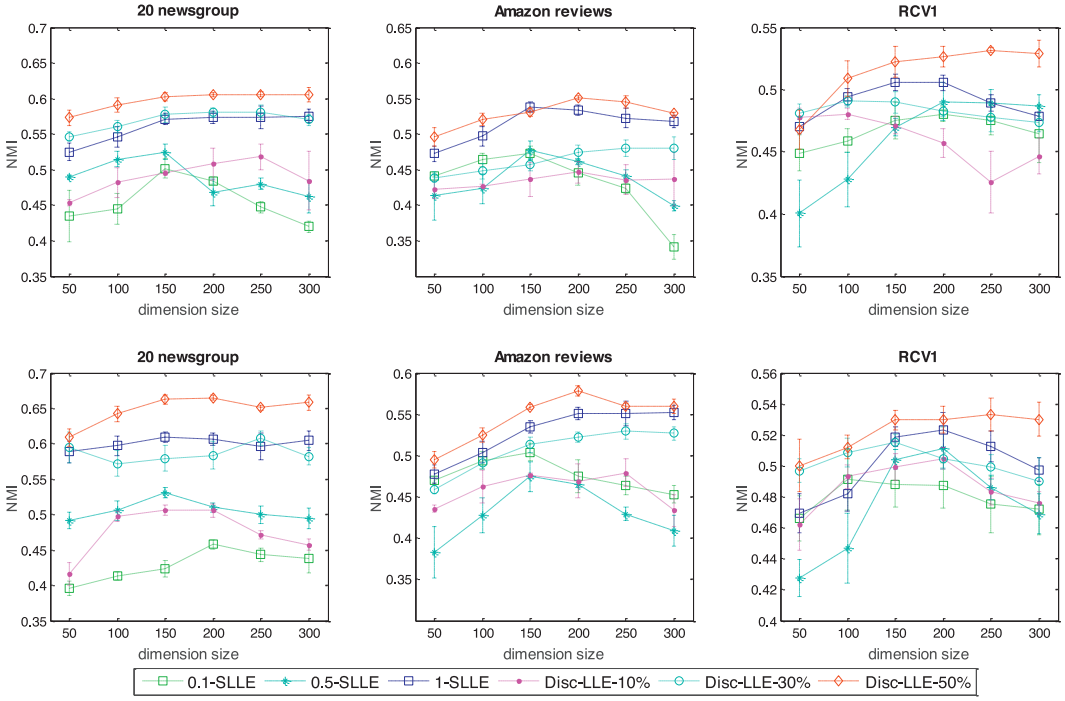

Fig. 14. NMI of K-means (top) and Ncut (bottom) on three datasets under various embeddings, where each point consist of mean value and standard deviations over 5 runs.

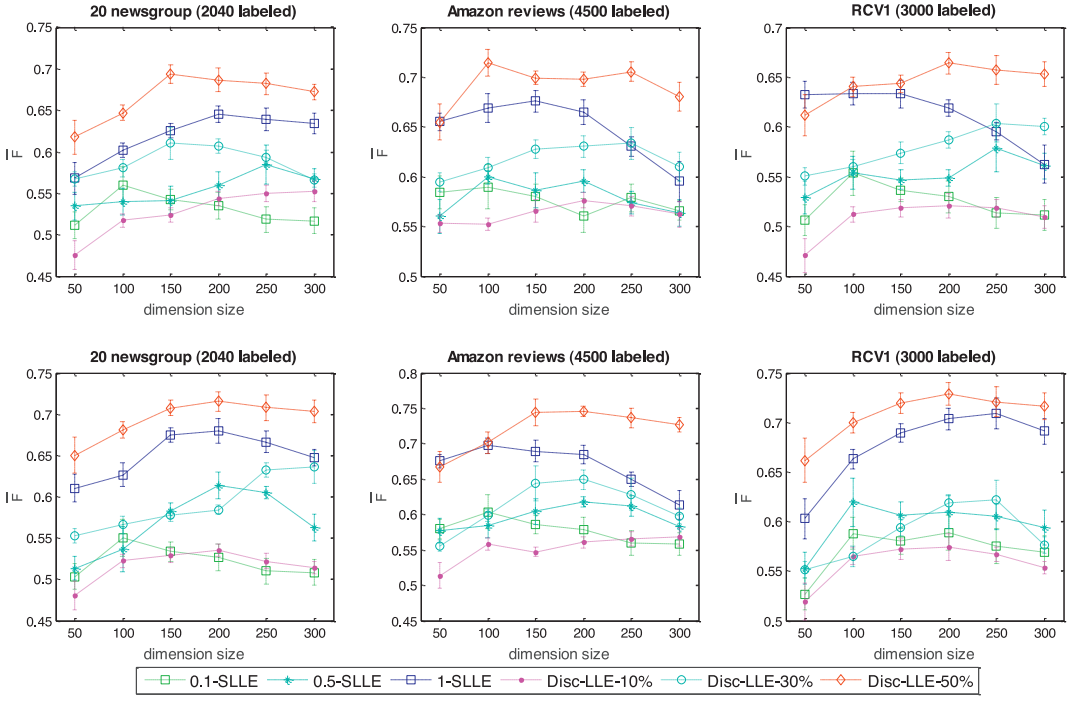

Fig. 15. Average performance of 1-NN (top) and SVM (bottom) on three datasets under various embeddings, where each point consist of mean value and standard deviations over 5 runs.

ding (Disc-LDE). Disc-LDE is developed in three stages: at first, an improved knn graph is built by using the NMD of word embeddings for weighting every documents pair; then, a discriminative subspace around given document is found by performing Transductive-MARW on the knn graph; at last, the documents-generating structure is modeled as a pseudo-document by a generative probabilistic model of subspace and the training of AEs is regularized by jointly recovering the given document as well as its pseudo-document. Under a general regularized function learning framework, the encoding affine map will keep smooth to the variation along the documents-generating structure of subspace. The experimental evaluations demonstrate the superiority of Disc-LDE for capturing informative and discriminative embeddings.

Although Disc-LDE is a type of shallow neural network, we believe that Disc-LDE has the potential as a basic module to be stacked to constitute a deep neural network. Hence, we will consider combining our approach with various deep learning approaches, so that yield more robust and generalized embeddings.

Reference

[1] C. Qimin, G. Qiao, W. Yongliang, W. Xianghua, Text clustering using VSM with feature clusters, Neural Comput. Appl. 26 (4) (2015) 995–1003.
[2] L. Bing, S. Jiang, W. Lam, Y. Zhang, S. Jameel, Adaptive concept resolution for document representation and its applications in text mining, Knowl. Based Syst. 74 (2015) 1–13.
[3] M. Dragoni, C. da Costa Pereira, A.G. Tettamanzi, A conceptual representation of documents and queries for information retrieval systems by using light ontologies, Expert Syst. Appl. 39 (12) (2012) 10376–10388.
[4] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990) 391–407.
[5] X. Wei, L. Xin, G. Yihong, Document clustering based on non-negative matrix factorization, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM, 2003, pp. 267–273.
[6] T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 1999 International Conference on Research and Development in Information Retrieval (SIGIR'99), 1999.
[7] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (1) (2003) 933–1022.
[8] Y. Bengio, Learning deep architectures for AI, Found. Trends® Mach. Learn. 2 (1) (2009) 1–127.
[9] L. Jiwei, L. Minh-Thang, D. Jurafsky, A hierarchical neural autoencoder for paragraphs and documents, in: Proceedings of Association for Computational Linguistics, ACL, 2015.
[10] H. Larochelle, S. Lauly, A neural autoregressive topic model, in: Proceedings of Advances in Neural Information Processing Systems, 2012, pp. 2717–2725.
[11] H. Liu, Z. Wu, X. Li, D. Cai, T.S. Huang, Constrained non-negative matrix factorization for image representation, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1299–1311.
[12] L. Niu, Y. Shi, Semi-supervised pLSA for document clustering, in: Proceedings of 2010 IEEE International Conference on Data Mining Workshops (ICDMW), 2010, pp. 1196–1203.
[13] H. Shan, A. Banerjee, Mixed-membership naive bayes models, Data Min. Knowl. Discov. 23 (1) (2011) 1–62.
[14] M.A. Ranzato, M. Szummer, Semi-supervised learning of compact document representations with deep networks, in: Proceedings of the 25th international conference on Machine learning, ACM, 2008, pp. 792–799.
[15] R. Johnson, T. Zhang, Supervised and semi-supervised text categorization using lstm for region embeddings, in: Proceedings of the 33rd International Conference on Machine Learning, ICML, 2016, pp. 526–534.
[16] Z. Yin, Z. Yujin, H. Larochelle, A deep and autoregressive approach for topic modeling of multimodal data, IEEE Trans. Pattern Anal. Mach. Intell. 38 (6) (2016) 1056–1069.
[17] H.S. Seung, D.D. Lee, The manifold ways of perception, Science 290 (5500) (2000) 2268–2269.
[18] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[19] J.B. Tenenbaum, V. De Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[20] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, Adv. Neural Inf. Process. Syst. 14 (6) (2002) 585–591.
[21] L. Mei, Z. Li, Z. Xiangjun, L. Fanzhang, Constrained neighborhood preserving concept factorization for data representation, Knowl. Based Syst. 102 (2016) 127–139.
[22] M. Lu, X.-J. Zhao, L. Zhang, F.-Z. Li, Semi-supervised concept factorization for document clustering, Inf. Sci. 331 (2016) 86–98.
[23] H. Wu, J. Bu, C. Chen, J. Zhu, L. Zhang, H. Liu, Locally discriminative topic modeling, Pattern Recognit. 45 (1) (2012) 617–625.
[24] D. Cai, X Wang, X. He, Probabilistic dyadic data analysis with local and global consistency, in: Proceedings of the International Conference on Machine Learning (ICML), 2009, pp. 105–112.
[25] S Huh, S.E. Fienberg, Discriminative topic modeling based on manifold learning, ACM Trans. Knowl. Discov. Data 5 (4) (2012) 653–661.
[26] H. Strange, R. Zwiggelaar, A generalised solution to the out-of-sample extension problem in manifold learning, AAAI (2011) 293–296.
[27] E. Vural, C. Guillemot, Out-of-sample generalizations for supervised manifold learning for classification, IEEE Trans. Image Process. 25 (3) (2016) 1410–1424.
[28] L. Yiyi, W. Yue, L. Yong, in: Graph regularized auto-encoders for image representation, 2016 A publication of the IEEE Signal Processing Society, doi:10.1109/TIP.2016.2605010.
[29] J. Kui, et al., Laplacian auto-encoders: an explicit learning of nonlinear data manifold, Neurocomputing 160 (2015) 250–260.
[30] L. Weng, F. Dornaika, Z. Jin, Flexible constrained sparsity preserving embedding, Pattern Recognit. 60 (2016) 813–823.
[31] D. Cai, X. He, J. Han, T.S. Huang, Graph regularized nonnegative matrix factorization for data representation, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1548–1560.
[32] D. Cai, X. He, J. Han, Locally consistent concept factorization for document clustering, IEEE Trans. Knowl. Data Eng. 23 (6) (2011) 902–913.
[33] Y Bengio, F Paiement J, P Vincent, et al., Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering, in: Proceedings of Advances in Neural Information Processing Systems, (NIPS), 16, 2004, pp. 177–184.
[34] E Vural, C. Guillemot, Out-of-sample generalizations for supervised manifold learning for classification, IEEE Trans. Image Process. 25 (3) (2016) 1410–1424.
[35] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, Manzagol, P.-A., Stacked denoising auto-encoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[36] S Rifai, P Vincent, X Muller, X Glorot, Y Bengio, Contractive Auto-Encoders: Explicit invariance during feature extraction, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 833–840.
[37] T. Jebara, J. Wang, S.F. Chang, Graph construction and b-matching for semi-supervised learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, ACM, 2009, pp. 441–448.
[38] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the Processing Systems NIPS, 2013, pp. 3111–3119.
[39] P. Jeffrey, S. Richard, C.D Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, Doha, Qatar, 2014, pp. 1532–1543.
[40] M.J. Kusner, Y. Sun, N.I. Kolkin, Nicholas, K.Q. Weinberger, From word embeddings to document distances, in: Proceedings of the International Conference on Machine Learning (ICML), 2015.
[41] Z. Zhang, J. Wang, H. Zha, Adaptive manifold learning., IEEE Trans. Pattern Anal. Mach. Intell. 34 (2) (2012) 253–265.
[42] M. Alamgir, U. Von Luxburg, Multi-agent random walks for local clustering on graphs, in: Proceedings of the IEEE 10th International Conference on Data Mining (ICDM), IEEE, 2010, pp. 18–27.
[43] R.G. Brereton, G.R. Lloyd, Support vector machines for classification and regression, Analyst 135 (2) (2010) 230–267.
[44] P. Diaconis, D. Stroock, Geometric bounds for eigenvalues of Markov chains, Ann. Appl. Probab. 1 (1) (1991) 36–61.
[45] M. Van der Maaten, G. Hilton, Visualizing data using T-SNE, J. Mach. Learn. Res 9 (2008) 2579–2605.
[46] R. Arun, V. Suresh, C.V. Madhavan, M.N. Murthy, On finding the natural number of topics with latent dirichlet allocation: some observations, Advances in Knowledge Discovery and Data Mining, Springer, Berlin, Heidelberg, 2010, pp. 391–402.
[47] C.Y. Zhou, Y.Q. Chen, Improving nearest neighbor classification with cam weighted distance, Pattern Recognit. 39 (4) (2006) 635–645.
[48] C Wei, S Luo, X Ma, et al., Locally Embedding autoencoders: a semi-supervised manifold learning approach of document representation, PLoS One 11 (1) (2016) e0146672.
[49] D. De Ridder, O. Kouropteva, O. Okun, M. Pietikäinen, R.P. Duin, Supervised locally linear embedding, in: Proceedings of the Joint International Conference Artificial Neural Networks and Neural Information Processing, CANN/ICONIP 2003, Springer, Berlin, Heidelberg, 2003, pp. 333–341.
[50] D.D. Lewis, Y. Yang, T.G. Rose, F. Li, Rcv1: a new benchmark collection for text categorization research, J. Mach. Learn. Res. 5 (2004) 361–397.
[51] H. Shan, A. Banerjee, Mixed-membership naive Bayes models, Data Min. Knowl. Discov. 23 (1) (2011) 1–62.
[52] W Wang, Y Huang, Y Wang, L Wang, Generalized auto-encoder: a neural network framework for dimensionality reduction, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2014, pp. 496–503.
[53] J. Shi, J. Malik, Normalized cuts and image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000) 888–905.