Sentiment Classification with Graph Co-Regularization

Guangyou Zhou, Jun Zhao, and Daojian Zeng
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
95 Zhongguancun East Road, Beijing 100190, China
{gyzhou,jzhao,djzeng}@nlpr.ia.ac.cn

Abstract

Sentiment classification aims to automatically predict sentiment polarity (e.g., positive or negative) of user-generated sentiment data (e.g., reviews, blogs). To obtain sentiment classification with high accuracy, supervised techniques require a large amount of manually labeled data. The labeling work can be time-consuming and expensive, which makes unsupervised (or semi-supervised) sentiment analysis essential for this application. In this paper, we propose a novel algorithm, called graph co-regularized non-negative matrix tri-factorization (GNMTF), from the geometric perspective. GNMTF assumes that if two words (or documents) are sufficiently close to each other, they tend to share the same sentiment polarity. To achieve this, we encode the geometric information by constructing the nearest neighbor graphs, in conjunction with a non-negative matrix tri-factorization framework. We derive an efficient algorithm for learning the factorization, analyze its complexity, and provide proof of convergence. Our empirical study on two open data sets validates that GNMTF can consistently improve the sentiment classification accuracy in comparison to the state-of-the-art methods.

1 Introduction

Recently, sentiment classification has gained a wide interest in natural language processing (NLP) community. Methods for automatically classifying sentiments expressed in products and movie reviews can roughly be divided into supervised and unsupervised (or semi-supervised) sentiment analysis. Supervised techniques have been proved promising and widely used in sentiment classification (Pang et al., 2002; Pang and Lee, 2008; Liu, 2012). However, the performance of these methods relies on manually labeled training data. In some cases, the labeling work may be time-consuming and expensive. This motivates the problem of learning robust sentiment classification via unsupervised (or semi-supervised) paradigm.

A traditional way to perform unsupervised sentiment analysis is the lexicon-based method (Turney, 2002; Taboada et al., 2011). Lexicon-based methods employ a sentiment lexicon to determine overall sentiment orientation of a document. However, it is difficult to define a universally optimal sentiment lexicon to cover all words from different domains (Lu et al., 2011a). Besides, most semi-automated lexicon-based methods yield unsatisfactory lexicons, with either high coverage and low precision or vice versa (Ng et al., 2006). Thus it is challenging for lexicon-based methods to accurately identify the overall sentiment polarity of users generated sentiment data. Recently, Li et al. (2009) proposed a constrained non-negative matrix tri-factorization (CNMTF) approach to sentiment classification, with a domain-independent sentiment lexicon as prior knowledge. Experimental results show that CNMTF achieves state-of-the-art performance.

From the geometric perspective, the data points (words or documents) may be sampled from a distribution supported by a low-dimensional manifold embedded in a high-dimensional space (Cai et al., 2011). This geometric structure, meaning that two words (or documents) sufficiently close to each other tend to share the same sentiment polarity, should be preserved during the matrix factorization. Research studies

Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1331–1340, Dublin, Ireland, August 23-29 2014.

have shown that learning performance can be significantly enhanced in many real applications (e.g., text mining, computer vision, etc.) if the geometric structure is exploited (Roweis and Saul, 2000; Tenenbaum et al., 2000). However, CNMTF fails to exploit the geometric structure, it is not clear whether this geometric information is useful for sentiment classification, which remains an under-explored area. This paper is thus designed to fill the gap.

In this paper, we propose a novel algorithm, called graph co-regularized non-negative matrix tri-factorization (GNMTF). We construct two affinity graphs to encode the geometric information underlying the word space and the document space, respectively. Intuitively, if two words or documents are sufficiently close to each other, they tend to share the same sentiment polarity. Taking these two graphs as co-regularization for the non-negative matrix tri-factorization, leading to the better sentiment polarity prediction which respects to the geometric structures of the word space and document space. We also derive an efficient algorithm for learning the tri-factorization, analyze its complexity, and provide proof of convergence. Empirical study on two open data sets shows encouraging results of the proposed method in comparison to state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 introduces the basic concept of matrix tri-factorization. Section 3 describes our graph co-regularized non-negative matrix tri-factorization (GNMTF) for sentiment classification. Section 4 presents the experimental results. Section 5 introduces the related work. In section 6, we conclude the paper and discuss future research directions.

2 Preliminaries

2.1 Non-negative Matrix Tri-factorization

Li et al. (2009) proposed a matrix factorization based framework for unsupervised (or semi-supervised) sentiment analysis. The proposed framework is built on the orthogonal non-negative matrix tri-factorization (NMTF) (Ding et al., 2006). In these models, a term-document matrix $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$ is approximated by three factor matrices that specify cluster labels for words and documents by solving the following optimization problem:

$$\min_{\mathbf{U},\mathbf{H},\mathbf{V} \geq 0} \mathcal{O} = \left\|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\right\|_F^2 + \sigma_1 \left\|\mathbf{U}^T\mathbf{U} - \mathbf{I}\right\|_F^2 + \sigma_2 \left\|\mathbf{V}^T\mathbf{V} - \mathbf{I}\right\|_F^2 \tag{1}$$

where σ_1 and σ_2 are the shrinkage regularization parameters, $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_k] \in \mathbb{R}_+^{m \times k}$ is the word-sentiment matrix, $\mathbf{V} = [\mathbf{v}_1, \cdots, \mathbf{v}_n] \in \mathbb{R}_+^{n \times k}$ is the document-sentiment matrix, and k is the number of sentiment classes for documents. Our task is polarity sentiment classification (positive or negative), i.e., $k = 2$. For example, $\mathbf{V}_{i1} = 1$ (or $\mathbf{U}_{i1} = 1$) represents that the sentiment polarity of document i (or word i) is positive, and $\mathbf{V}_{i2} = 1$ (or $\mathbf{U}_{i2} = 1$) represents that the sentiment polarity of document i (or word i) is negative. $\mathbf{V}_{i*} = 0$ (or $\mathbf{U}_{i*} = 0$) represents unknown, i.e., the document i (or word i) is neither positive or negative. $\mathbf{H} \in \mathbb{R}_+^{k \times k}$ provides a condensed view of $\mathbf{X}$; $\|\cdot\|_F$ is the Frobenius norm and $\mathbf{I}$ is a $k \times k$ identity matrix with all entries equal to 1. Based on the shrinkage methodology, we can approximately satisfy the orthogonality constraints for $\mathbf{U}$ and $\mathbf{V}$ by preventing the second and third terms from getting too large.

2.2 Constrained NMTF

Lexical knowledge in the form of the polarity of words in the lexicon can be introduced in matrix tri-factorization. By partially specifying word polarity via $\mathbf{U}$, the lexicon influences the sentiment prediction $\mathbf{V}$ over documents. Following the literature (Li et al., 2009), let $\mathbf{U}_0$ represent lexical prior knowledge about sentiment words in the lexicon, e.g., if word i is positive $(\mathbf{U}_0)_{i1} = 1$ while if it is negative $(\mathbf{U}_0)_{i2} = 1$, and if it does not exist in the lexicon $(\mathbf{U}_0)_{i*} = 0$. Li et al. (2009) also investigated that we had a few documents manually labeled for the purpose of capturing some domain-specific connotations. Let $\mathbf{V}_0$ denote the manually labeled documents, if the document expresses positive sentiment $(\mathbf{V}_0)_{ii} = 1$, and $(\mathbf{V}_0)_{i2} = 1$ for negative sentiment. Therefore, the semi-supervised learning with lexical knowledge can be written as:

$$\min_{\mathbf{U},\mathbf{H},\mathbf{V} \geq 0} \mathcal{O} + \alpha \mathrm{Tr}\big[(\mathbf{U} - \mathbf{U}_0)^T \mathbf{C}^u (\mathbf{U} - \mathbf{U}_0)\big] + \beta \mathrm{Tr}\big[(\mathbf{V} - \mathbf{V}_0)^T \mathbf{C}^v (\mathbf{V} - \mathbf{V}_0)\big] \tag{2}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix, $\alpha > 0$ and $\beta > 0$ are the parameters which control the contribution of lexical prior knowledge and manually labeled documents. $\mathbf{C}^u \in \{0,1\}^{m\times m}$ is a diagonal matrix whose entry $\mathbf{C}^u_{ii} = 1$ if the category of the i-th word is known and $\mathbf{C}^u_{ii} = 0$ otherwise. $\mathbf{C}^v \in \{0,1\}^{n\times n}$ is a diagonal matrix whose entry $\mathbf{C}^v_{ii} = 1$ if the category of the i-th document is labeled and $\mathbf{C}^v_{ii} = 0$ otherwise.

3 Graph Co-regularized Non-negative Matrix Tri-factorization

In this section, we introduce our proposed graph co-regularized non-negative matrix tri-factorization (GNMTF) algorithm which avoids this limitation by incorporating the geometrically based co-regularization.

3.1 Model Formulation

Based on the manifold assumption (Belkin and Niyogi, 2001), if two documents $\mathbf{x}_i$ and $\mathbf{x}_j$ are sufficiently close to each other in the intrinsic geometric of the documents distribution, then their sentiment polarity $\mathbf{v}_i$ and $\mathbf{v}_j$ should be close. In order to model the geometric structure, we construct a document-document graph G^v. In the graph, nodes represent documents in the corpus and edges represent the affinity between the documents. The affinity matrix $\mathbf{W}^v \in \mathbb{R}^{n\times n}$ of the graph G^v is defined as

$$\mathbf{W}^v_{ij} = \begin{cases} \cos(\mathbf{x}_i, \mathbf{x}_j) & \text{if } \mathbf{x}_i \in \mathcal{N}_p(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_p(\mathbf{x}_i) \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{N}_p(\mathbf{x}_i)$ represents the p-nearest neighbors of document $\mathbf{x}_i$. Many matrices, e.g., 0-1 weighting, textual similarity and heat kernel weighting (Belkin and Niyogi, 2001), can be used to obtain nearest neighbors of a document, and further define the affinity matrix. Since $\mathbf{W}^v_{ij}$ in our paper is only for measuring the closeness, we only use the simple textual similarity and do not treat the different weighting schemes separately due to the limited space. For further information, please refer to (Cai et al., 2011).

Preserving the geometric structure in the document space is reduced to minimizing the following loss function:

$$\begin{aligned} \mathcal{R}^v &= \frac{1}{2}\sum_{i,j=1}^{n} \left\|\mathbf{v}_i - \mathbf{v}_j\right\|_2^2 \mathbf{W}^v_{ij} = \sum_{i=1}^{n} \mathbf{v}_i^T \mathbf{v}_i \mathbf{D}^v_{ii} - \sum_{i,j=1}^{n} \mathbf{v}_i^T \mathbf{v}_j \mathbf{W}^v_{ij} \\ &= \mathrm{Tr}(\mathbf{V}^T \mathbf{D}^v \mathbf{V}) - \mathrm{Tr}(\mathbf{V}^T \mathbf{W}^v \mathbf{V}) = \mathrm{Tr}(\mathbf{V}^T \mathbf{L}^v \mathbf{V}) \end{aligned} \tag{4}$$

where $\mathbf{D}^v \in \mathbb{R}^{n\times n}$ is a diagonal matrix whose entries are column (or row, since $\mathbf{D}^v$ is symmetric) sums of $\mathbf{W}^v$, $\mathbf{D}^v_{ii} = \sum_{j=1}^{n} \mathbf{W}^v_{ij}$, and $\mathbf{L}^v = \mathbf{D}^v - \mathbf{W}^v$ is the Laplacian matrix (Chung, 1997) of the constructed graph G^v.

Similarly to document-document geometric structure, if two words $\mathbf{w}_i = [\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in}]$ and $\mathbf{w}_j = [\mathbf{x}_{j1}, \cdots, \mathbf{x}_{jn}]$ are sufficiently close to each other in the intrinsic geometric of the words distribution, then their sentiment polarity $\mathbf{u}_i$ and $\mathbf{u}_j$ should be close. In order to model the geometric structure in the word space, we construct a word-word graph G^u. In the graph, nodes represent distinct words and edges represent the affinity between words. The affinity matrix $\mathbf{W}^u \in \mathbb{R}^{m\times m}$ of the graph G^u is defined as

$$\mathbf{W}^u_{ij} = \begin{cases} \cos(\mathbf{w}_i, \mathbf{w}_j) & \text{if } \mathbf{w}_i \in \mathcal{N}_p(\mathbf{w}_j) \text{ or } \mathbf{w}_j \in \mathcal{N}_p(\mathbf{w}_i) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\mathcal{N}_p(\mathbf{w}_j)$ represents the p-nearest neighbor of word $\mathbf{w}_j$. Here, we represent a term $\mathbf{w}_j$ as a document vector $[\mathbf{x}_{j1}, \cdots, \mathbf{x}_{jn}]$. To measure the closeness of two words, a common way is to calculate the similarity of their vector representations. Although there are several ways (e.g., co-occurrence information, semantic similarity computed by WordNet, Wikipedia, or search engine have been empirically studied in NLP literature (Hu et al., 2009)) to define the affinity matrix $\mathbf{W}^u$, we do not treat the different ways separately and leave this investigation for future work.

Preserving the geometric structure in the word space is reduced to minimizing the following loss function:

$$\mathcal{R}^u = \frac{1}{2}\sum_{i,j=1}^{m} \left\|\mathbf{u}_i - \mathbf{u}_j\right\|_2^2 \mathbf{W}^u_{ij} = \mathrm{Tr}(\mathbf{U}^T \mathbf{L}^u \mathbf{U}) \tag{6}$$

where $\mathbf{L}^u = \mathbf{D}^u - \mathbf{W}^u$ is the Laplacian matrix of the constructed graph G^u, and $\mathbf{D}^u \in \mathbb{R}^{m\times m}$ is a diagonal matrix whose entries are $\mathbf{D}^u_{ii} = \sum_{j=1}^{m} \mathbf{W}^u_{ij}$.

Finally, we treat unsupervised (or semi-supervised) sentiment classification as a clustering problem, employing lexical prior knowledge and partial manually labeled data to guide the learning process. Moreover, we introduce the geometric structures from both document and word sides as co-regularization. Therefore, our proposed unsupervised (or semi-supervised) sentiment classification framework can be mathematically formulated as solving the following optimization problem:

$$\begin{aligned}\min_{\mathbf{U},\mathbf{H},\mathbf{V}\geq 0} \mathcal{L} = &\left\|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\right\|_F^2 + \sigma_1\left\|\mathbf{U}^T\mathbf{U} - \mathbf{I}\right\|_F^2 + \sigma_2\left\|\mathbf{V}^T\mathbf{V} - \mathbf{I}\right\|_F^2 \\ &+ \alpha\mathrm{Tr}\big[(\mathbf{U}-\mathbf{U}_0)^T\mathbf{C}^u(\mathbf{U}-\mathbf{U}_0)\big] + \gamma\mathrm{Tr}(\mathbf{U}^T\mathbf{L}^u\mathbf{U}) \\ &+ \beta\mathrm{Tr}\big[(\mathbf{V}-\mathbf{V}_0)^T\mathbf{C}^v(\mathbf{V}-\mathbf{V}_0)\big] + \delta\mathrm{Tr}(\mathbf{V}^T\mathbf{L}^v\mathbf{V})\end{aligned} \tag{7}$$

where $\delta > 0$ and $\gamma > 0$ are parameters which control the contributions of document space and word space geometric information, respectively. With the optimization results, the sentiment polarity of a new document $\mathbf{x}_i$ can be easily inferred by $f(\mathbf{x}_i) = \arg\max_{j\in\{p,\, n\}} \mathbf{V}_{ij}$.

3.2 Learning Algorithm

We present the solution to the GNMTF optimization problem in equation (7) as the following theorem. The theoretical aspects of the optimization are presented in the next subsection.

Theorem 3.1. *Updating* $\mathbf{U}$*,* $\mathbf{H}$ *and* $\mathbf{V}$ *using equations (8)∼(10) will monotonically decrease the objective function in equation (7) until convergence.*

$$\mathbf{U} \leftarrow \mathbf{U} \circ \frac{\left[\mathbf{X}\mathbf{V}\mathbf{H}^T + \sigma_1\mathbf{U} + \alpha\mathbf{C}^u\mathbf{U}_0 + \gamma\mathbf{W}^u\mathbf{U}\right]}{\left[\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + \sigma_1\mathbf{U}\mathbf{U}^T\mathbf{U} + \alpha\mathbf{C}^u\mathbf{U} + \gamma\mathbf{D}^u\mathbf{U}\right]} \tag{8}$$

$$\mathbf{H} \leftarrow \mathbf{H} \circ \frac{\left[\mathbf{U}^T\mathbf{X}\mathbf{V}\right]}{\left[\mathbf{U}^T\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\right]} \tag{9}$$

$$\mathbf{V} \leftarrow \mathbf{V} \circ \frac{\left[\mathbf{X}^T\mathbf{U}\mathbf{H} + \sigma_2\mathbf{V} + \beta\mathbf{C}^v\mathbf{V}_0 + \delta\mathbf{W}^v\mathbf{V}\right]}{\left[\mathbf{V}\mathbf{H}^T\mathbf{U}^T\mathbf{U}\mathbf{H} + \sigma_2\mathbf{V}\mathbf{V}^T\mathbf{V} + \beta\mathbf{C}^v\mathbf{V} + \delta\mathbf{D}^v\mathbf{V}\right]} \tag{10}$$

where operator $\circ$ *is element-wise product and* $\frac{[\cdot]}{[\cdot]}$ *is element-wise division.*

Based on Theorem 3.1, we note that the multiplicative update rules given by equations (8)∼(10) are obtained by extending the updates of standard NMTF (Ding et al., 2006). A number of techniques can be used here to optimize the objective function in equation (7), such as alternating least squares (Kim and Park, 2008), the active set method (Kim and Park, 2008), and the projected gradients approach (Lin, 2007). Nonetheless, the multiplicative updates derived in this paper has reasonably fast convergence behavior as shown empirically in the experiments.

3.3 Theoretical Analysis

In this subsection, we give the theoretical analysis of the optimization, convergence and computational complexity. Without loss of generality, we only show the optimization of $\mathbf{U}$ and formulate the Lagrange function with constraints as follows:

$$\mathcal{L}(\mathbf{U}) = \left\|\mathbf{X} - \mathbf{U}\mathbf{H}\mathbf{V}^T\right\|_F^2 + \sigma_1\left\|\mathbf{U}^T\mathbf{U} - \mathbf{I}\right\|_F^2 + \alpha\mathrm{Tr}\big[(\mathbf{U}-\mathbf{U}_0)^T\mathbf{C}^u(\mathbf{U}-\mathbf{U}_0)\big] + \mathrm{Tr}(\Psi\mathbf{U}^T) \tag{11}$$

where Ψ is the Lagrange multiplier for the nonnegative constraint $\mathbf{U} \geq \mathbf{0}$.

The partial derivative of $\mathcal{L}(\mathbf{U})$ w.r.t. $\mathbf{U}$ is

$$\begin{aligned}\bigtriangledown_{\mathbf{U}}\mathcal{L}(\mathbf{U}) = &-2\mathbf{X}\mathbf{V}\mathbf{H}^T + 2\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + 2\sigma_1\mathbf{U}\mathbf{U}^T\mathbf{U} - 2\sigma_1\mathbf{U} \\ &+ 2\alpha\mathbf{C}^u\mathbf{U} - 2\alpha\mathbf{C}^u\mathbf{U}_0 + 2\gamma\mathbf{D}^u\mathbf{U} - 2\gamma\mathbf{W}^u\mathbf{U} + \Psi\end{aligned}$$

Using the Karush-Kuhn-Tucker (KKT) (Boyd and Vandenberghe, 2004) condition $\Psi \circ \mathbf{U} = \mathbf{0}$, we can obtain

$$\begin{aligned}\nabla_{\mathbf{U}}\mathcal{L}(\mathbf{U}) \circ \mathbf{U} = &\left[\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + \sigma_1\mathbf{U}\mathbf{U}^T\mathbf{U} + \alpha\mathbf{C}^u\mathbf{U} + \gamma\mathbf{D}^u\mathbf{U}\right] \circ \mathbf{U} \\ &- \left[\mathbf{X}\mathbf{V}\mathbf{H}^T + \sigma_1\mathbf{U} + \alpha\mathbf{C}^u\mathbf{U}_0 + \gamma\mathbf{W}^u\mathbf{U}\right] \circ \mathbf{U} = \mathbf{0}\end{aligned}$$

This leads to the update rule in equation (8). Following the similar derivations as shown above, we can obtain the updating rules for all the other variables $\mathbf{H}$ and $\mathbf{V}$ in GNMTF optimization, as shown in equations (9) and (10).

3.3.1 Convergence Analysis

In this subsection, we prove the convergence of multiplicative updates given by equations (8)$\sim$(10). We first introduce the definition of auxiliary function as follows.

Definition 3.1. *$\mathcal{F}(\mathbf{Y}, \mathbf{Y}')$ is an auxiliary function for $\mathcal{L}(\mathbf{Y})$ if $\mathcal{L}(\mathbf{Y}) \leq \mathcal{F}(\mathbf{Y}, \mathbf{Y}')$ and equality holds if and only if $\mathcal{L}(\mathbf{Y}) = \mathcal{F}(\mathbf{Y}, \mathbf{Y})$.*

Lemma 3.1. *(Lee and Seung, 2001) If $\mathcal{F}$ is an auxiliary function for $\mathcal{L}$, $\mathcal{L}$ is non-increasing under the update $\mathbf{Y}^{(t+1)} = \arg\min_{\mathbf{Y}} \mathcal{F}(\mathbf{Y}, \mathbf{Y}^{(t)})$*

Proof. By Definition 3.1, $\mathcal{L}(\mathbf{Y}^{(t+1)}) \leq \mathcal{F}(\mathbf{Y}^{(t+1)}, \mathbf{Y}^{(t)}) \leq \mathcal{F}(\mathbf{Y}^{(t)}, \mathbf{Y}^{(t)}) = \mathcal{L}(\mathbf{Y}^{(t)})$ □

Theorem 3.2. *Let function*

$$\begin{aligned}\mathcal{F}(\mathbf{U}_{ij}, \mathbf{U}_{ij}^{(t)}) = &\mathcal{L}(\mathbf{U}_{ij}^{(t)}) + \mathcal{L}'(\mathbf{U}_{ij}^{(t)})(\mathbf{U}_{ij} - \mathbf{U}_{ij}^{(t)}) \\ &+ \frac{\left[\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + \sigma_1\mathbf{U}\mathbf{U}^T\mathbf{U} + \alpha\mathbf{C}^u\mathbf{U} + \gamma\mathbf{D}^u\mathbf{U}\right]_{ij}}{\mathbf{U}_{ij}}(\mathbf{U}_{ij} - \mathbf{U}_{ij}^{(t)})\end{aligned} \tag{12}$$

be a proper auxiliary function for $\mathcal{L}(\mathbf{U}_{ij})$, where $\mathcal{L}'(\mathbf{U}_{ij}) = [\nabla_{\mathbf{U}}\mathcal{L}(\mathbf{U})]_{ij}$ is the first-order derivatives of $\mathcal{L}(\mathbf{U}_{ij})$ with respect to $\mathbf{U}_{ij}$.

Theorem 3.2 can be proved similarly to (Ding et al., 2006). Due to limited space, we omit the details of the validation. Based on Lemmas 3.1 and Theorem 3.2, the update rule for $\mathbf{U}$ can be obtained by minimizing $\mathcal{F}(\mathbf{U}_{ij}^{(t+1)}, \mathbf{U}_{ij}^{(t)})$. When setting $\nabla_{\mathbf{U}_{ij}^{(t+1)}}\mathcal{F}(\mathbf{U}_{ij}^{(t+1)}, \mathbf{U}_{ij}^{(t)})$, we can obtain

$$\mathbf{U}_{ij}^{(t+1)} = \mathbf{U}_{ij}^{(t)} \frac{\left[\mathbf{X}\mathbf{V}\mathbf{H}^T + \sigma_1\mathbf{U} + \alpha\mathbf{C}^u\mathbf{U}_0 + \gamma\mathbf{W}^u\mathbf{U}\right]_{ij}}{\left[\mathbf{U}\mathbf{H}\mathbf{V}^T\mathbf{V}\mathbf{H}^T + \sigma_1\mathbf{U}\mathbf{U}^T\mathbf{U} + \alpha\mathbf{C}^u\mathbf{U} + \gamma\mathbf{D}^u\mathbf{U}\right]_{ij}}$$

By Lemma 3.1 and Theorem 3.2, we have $\mathcal{L}(\mathbf{U}^{(0)}) = \mathcal{F}(\mathbf{U}^{(0)}, \mathbf{U}^{(0)}) \geq \mathcal{F}(\mathbf{U}^{(1)}, \mathbf{U}^{(0)}) \geq \mathcal{F}(\mathbf{U}^{(1)}, \mathbf{U}^{(1)}) = \mathcal{L}(\mathbf{U}^{(1)}) \geq \cdots \geq \mathcal{L}(\mathbf{U}^{(Iter)})$, where $Iter$ denotes the number of iteration number. Therefore, $\mathbf{U}$ is monotonically decreasing. Since the objective function $\mathcal{L}$ is lower bounded by 0, the correctness and convergence of Theorem 3.1 is validated.

3.3.2 Time Complexity Analysis

In this subsection, we discuss the time computational complexity of the proposed algorithm GNMTF. Besides expressing the complexity of the algorithm using big O notation, we also count the number of arithmetic operations to provide more details about running time. We show the results in Table 1, where $m \gg k$ and $n \gg k$.

Based on the updating rules summarized in Theorem 3.1, it it not hard to count the arithmetic operators of each iteration in GNMTF. It is important to note that $\mathbf{C}^u$ is a diagonal matrix, the nonzero elements on each row of $\mathbf{C}^u$ is 1. Thus, we only need zero addition and mk multiplications to compute $\mathbf{C}^u\mathbf{U}$. Similarly, for $\mathbf{C}^u\mathbf{U}_0$, $\mathbf{C}^v\mathbf{V}$, $\mathbf{C}^v\mathbf{V}_0$, $\mathbf{D}^u\mathbf{U}$ and $\mathbf{D}^v\mathbf{V}$, we also only need zero addition and mk multiplications for each of them. Besides, we also note that $\mathbf{W}^u$ is a sparse matrix, if we use a p-nearest neighbor graph, the average nonzero elements on each row of $\mathbf{W}^u$ is p. Thus, we only need mpk additions and mpk multiplications to compute $\mathbf{W}^u\mathbf{U}$. Similarly, for $\mathbf{W}^v\mathbf{V}$, we need the same operation counts as $\mathbf{W}^u\mathbf{U}$. Suppose the multiplicative updates stop after $Iter$ iterations, the time cost of multiplicative updates then becomes $O(Iter \times mnk)$. Therefore, the overall running time of GNMTF is similar to the standard NMTF and CNMTF.

	addition	multiplication	division	overall
GNMTF: $\mathbf{U}$	$2k^3 + (2m+n)k^2 + m(n+p)k$	$2k^3 + (2m+n)k^2 + m(n+p+7)k$	mk	$O(mnk)$
GNMTF: $\mathbf{H}$	$2k^3 + (m+n+2)k^2 + mnk$	$2k^3 + (m+n+1)k^2 + mnk$	k^2	$O(mnk)$
GNMTF: $\mathbf{V}$	$2k^3 + (2n+m)k^2 + n(m+p)k$	$2k^3 + (2n+m)k^2 + n(m+p+7)k$	nk	$O(mnk)$

Table 1: Computational operation counts for each iteration in GNMTF.

4 Experiments

4.1 Data Sets

Sentiment classification has been extensively studied in the literature. Among these, a large majority proposed experiments performed on the benchmarks made of Movies Reviews (Pang et al., 2002) and Amazon products (Blitzer et al., 2007).

Movies data This data set has been widely used for sentiment analysis in the literature (Pang et al., 2002), which consists of 1000 positive and 1000 negative reviews drawn from the IMDB archive of rec.arts.movies.reviews.newsgroups.

Amazon data This data set is heterogeneous, heavily unbalanced and large-scale, a smaller version has been released. The reduced data set contains 4 product types: Kitchen, Books, DVDs, and Electronics (Blitzer et al., 2007). There are 4000 positive and 4000 negative reviews.[1]

For these two data sets, we select 8000 words with highest document-frequency to generate the vocabulary. Stopwords[2] are removed and a normalized term-frequency representation is used. In order to construct the lexical prior knowledge matrix $\mathbf{U}_0$, we use the sentiment lexicon generated by (Hu and Liu, 2004). It contains 2,006 positive words (e.g., "beautiful") and 4,783 negative words (e.g., "upset").

4.2 Unsupervised Sentiment Classification

Our first experiment is to explore the benefits of incorporating the geometric information in the unsupervised paradigm (that is $\mathbf{C}^v = \mathbf{0}$). Therefore, the third part in equation (7) will be ignored. For this unsupervised paradigm of GNMTF, we empirically set $\alpha = \delta = \gamma = 1$, $\sigma_1 = \sigma_2 = 1$, $Iter = 100$ and run GNMTF 10 repeated times to remove any randomness caused by the random initialization. Due to limited space, we do not present the impacts of the parameters on the learning model. Now we compare our proposed GNMTF with the following four categories of methods:

(1) Lexicon-Based Methods (LBM in short): Taboada et al. (2011) proposed to incorporate intensification and negation to refine the sentiment score for each document. This is the state-of-the-art lexicon-based method for unsupervised sentiment classification.

(2) Document Clustering Methods: We choose the most representative cluster methods, K-means, NMTF, Information-Theoretic Co-clustering (ITCC) (Dhillon et al., 2003), and Euclidean Co-clustering method (ECC) (Cho et al., 2004). We set the number of clusters as two in these methods. Note that all these methods do not make use of the sentiment lexicon.

(3) Constrained NMTF (CNMTF in short): Li et al. (2009) incorporated the sentiment lexicon into NMTF as a domain-independent prior constraint.

(4) Graph co-regularized Non-negative Matrix Tri-factorization (GNMTF in short): It is a new algorithm proposed in this paper. We use cosine similarity for constructing the p-nearest neighbor graph for its simplicity. The number of nearest neighbor p is set to 10 empirically both on document and word spaces.

4.2.1 Sentiment Classification Results

The experimental results are reported in Table 2. We perform a significant test, i.e., a t-test with a default significant level of 0.05. From Table 2, we can see that (1) Both CNMTF and GNMTF consider the lexical prior knowledge from off-the-shelf sentiment lexicon and achieve better performance than NMTF. This suggests the importance of the lexical prior knowledge in learning the sentiment classification (row

[1]The data set can be freely downloaded from http://www.cs.jhu.edu/ mdredze/datasets/sentiment/.

[2]http://truereader.com/manuals/onix/stopwords1.html

#	Methods	Movies	Amazon
1	LBM	0.632	0.580
2	K-means	0.543 (-8.9%)	0.535 (-4.5%)
3	NMTF	0.561 (-7.1%)	0.547 (-3.3%)
4	ECC	0.678 (+4.6%)	0.642 (+6.2%)
5	ITCC	0.714 (+8.2%)	0.655 (+7.5%)
6	CNMTF	0.695 (+6.3%)	0.658 (+7.8%)
7	**GNMTF**	**0.736 (+10.4%)**	**0.705 (+12.5%)**

Table 2: Sentiment classification accuracy of unsupervised paradigm on the data sets. Improvements of K-means, NMTF, ITCC, ECC, CNMTF and GNMTF over baseline LBM are shown in parentheses.

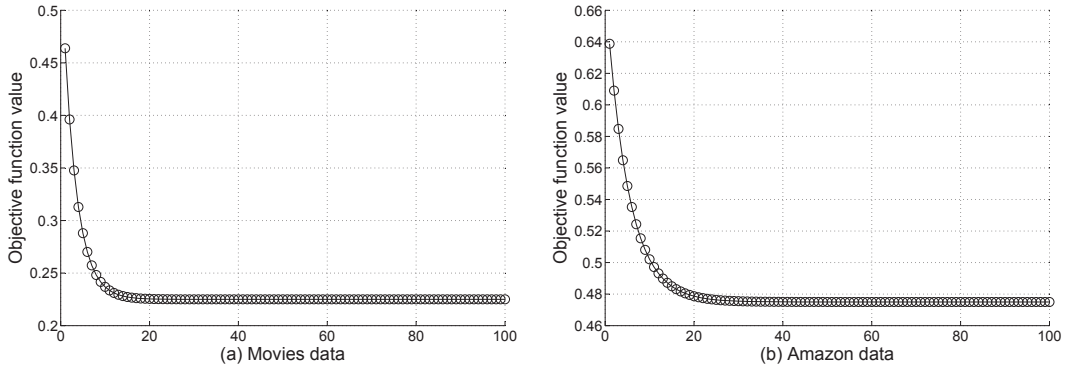

Figure 1: Convergence curves of GNMTF on both data sets.

3 vs. row 6 and row 7); (2) Regardless of the data sets, our GNMTF significantly outperforms state-of-the-art CNMTF and achieves the best performance. This shows the superiority of geometric information and graph co-regularization framework (row 4 vs. row 5, the improvements are statistically significant at $p < 0.05$).

4.2.2 Convergence Behavior

In subsection 3.3.1, we have shown that the multiplicative updates given by equations (8)$\sim$(10) are convergent. Here, we empirically show the convergence behavior of GNMTF.

Figure 1 shows the convergence curves of GNMTF on Movies and Amazon data sets. From the figure, y-axis is the value of objective function and x-axis denotes the iteration number. We can see that the multiplicative updates for GNMTF converge very fast, usually within 50 iterations.

4.3 Semi-supervised Sentiment Classification

In this subsection, we describe our proposed GNMTF with a few labeled documents. For this semi-supervised paradigm of GNMTF, we empirically set $Iter = 100$, $\sigma_1 = \sigma_2 = 2$, $\alpha = \beta = \delta = \gamma = 1$ and $p = 10$ on document and word spaces and also run 10 repeated times to remove any randomness caused by the random initialization. Due to limited space, we do not give an in-depth parameter analysis. For CNMTF, we set $\alpha = \beta = 1$ for fair comparison. We also compare our proposed GNMTF with some representative semi-supervised approaches described in (Li et al., 2009): (1) Semi-supervised learning with local and global consistency (Consistency Method in short) (Zhou et al., 2004); (2) Semi-supervised learning using gaussian fields and harmonic functions (GFHF in short) (Zhu et al., 2003). Besides, we also compare the results of our proposed GNMTF with the representative supervised classification method: support vector machine (SVM), which has been widely used in sentiment classification (Pang et al., 2002).

The results are presented in Figure 2. From the figure, we can see that GNMTF outperforms other methods over the entire range of number of labeled documents on both data sets. By this observation, we can conclude that taking the geometric information can still improve the sentiment classification accuracy in semi-supervised paradigm.

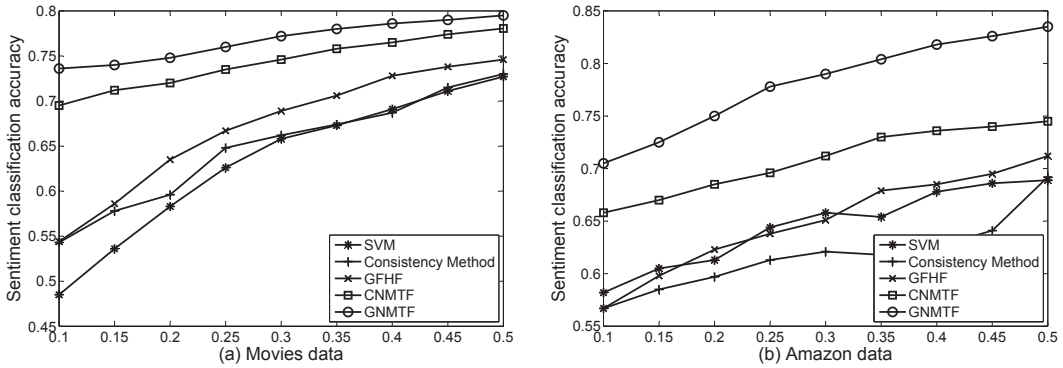

Figure 2: Sentiment classification accuracy vs. different percentage of labeled documents, where x-axis denotes the number of documents labeled as a fraction of the original labeled documents.

5 Related Work

Sentiment classification has gained widely interest in NLP community, we point the readers to recent books (Pang and Lee, 2008; Liu, 2012) for an in-depth survey of literature on sentiment analysis.

Methods for automatically classifying sentiments expressed in products and movie reviews can roughly be divided into supervised and unsupervised (or semi-supervised) sentiment analysis. Supervised techniques have been proved promising and widely used in sentiment classification (Pang et al., 2002; Pang and Lee, 2008; Liu, 2012). However, the performance of these methods relies on manually labeled training data. In some cases, the labeling work may be time-consuming and expensive. This motivates the problem of learning robust sentiment classification via unsupervised (or semi-supervised) paradigm.

The most representative way to perform semi-supervised paradigm is to employ partial labeled data to guide the sentiment classification (Goldberg and Zhu, 2006; Sindhwani and Melville, 2008; Wan, 2009; Li et al., 2011). However, we do not have any labeled data at hand in many situations, which makes the unsupervised paradigm possible. The most representative way to perform unsupervised paradigm is to use a sentiment lexicon to guide the sentiment classification (Turney, 2002; Taboada et al., 2011) or learn sentiment orientation via a matrix factorization clustering framework (Li et al., 2009; **?**; Hu et al., 2013). In contrast, we perform sentiment classification with the different model formulation and learning algorithm, which considers both word-level and document-level sentiment-related contextual information (e.g., the neighboring words or documents tend to share the same sentiment polarity) into a unified framework. The proposed framework makes use of the valuable geometric information to compensate the problem of lack of labeled data for sentiment classification. In addition, some researchers also explored the matrix factorization techniques for other NLP tasks, such as relation extraction (Peng and Park, 2013) and question answering (Zhou et al., 2013)

Besides, many studies address some other aspects of sentiment analysis, such as cross-domain sentiment classification (Blitzer et al., 2007; Pan et al., 2010; Hu et al., 2011; Bollegala et al., 2011; Glorot et al., 2011), cross-lingual sentiment classification (Wan, 2009; Lu et al., 2011b; Meng et al., 2012) and imbalanced sentiment classification (Li et al., 2011), which are out of scope of this paper.

6 Conclusion and Future Work

In this paper, we propose a novel algorithm, called graph co-regularized non-negative matrix tri-factorization (GNMTF), from a geometric perspective. GNMTF assumes that if two words (or documents) are sufficiently close to each other, they tend to share the same sentiment polarity. To achieve this, we encode the geometric information by constructing the nearest neighbor graphs, in conjunction with a non-negative matrix tri-factorization framework. We derive an efficient algorithm for learning the factorization, analyze its complexity, and provide proof of convergence. Our empirical study on two open data sets validates that GNMTF can consistently improve the sentiment classification accuracy in comparison to state-of-the-art methods.

There are some ways in which this research could be continued. First, some other ways should be considered to construct the graphs (e.g., hyperlinks between documents, synonyms or co-occurrences between words). Second, we will try to extend the proposed framework for other aspects of sentiment analysis, such as cross-domain or cross-lingual settings.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61303180 and No. 61272332), the Beijing Natural Science Foundation (No. 4144087), CCF Opening Project of Chinese Information Processing, and also Sponsored by CCF-Tencent Open Research Fund. We thank the anonymous reviewers for their insightful comments.

References

M. Belkin and P. Niyogi. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Proceedings of NIPS*, pages 585-591.

J. Blitzer, M. Dredze and F. Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In *Proceedings of ACL*, pages 440-447.

D. Bollegala, D. Weir, and J. Carroll. 2011. Using multiples sources to construct a sentiment sensitive thesaurus. In *Proceedings of ACL*, pages 132-141.

S. Boyd and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge university press.

D. Cai, X. He, J. Han, and T. Huang. 2011. Graph regularized non-negative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8): 1548-1560.

H. Cho, I. Dhillon, Y. Guan, and S. Sra. 2004. Minimum sum squared residue co-clutering of gene expression data. In *Proceedings of SDM*, pages 22-24.

F. Chung. 1997. Spectral graph theory. *Regional Conference Series in Mathematics*, Volume 92.

I. Dhillon, S. Mallela, and D. Modha. 2003. Information-theoretic Co-clustering. In *Proceedings of KDD*, pages 89-98.

C. Ding, T. Li, W. Peng, and H. Park. 2006. Orthogonal non-negative matrix tri-factorization for clustering. In *Proceedings of KDD*, pages 126-135.

X. Glorot, A. Bordes, and Y. Bengio. 2011. Domain adaptation for larage-scale sentiment classification: a deep learning approach. In *Proceedings of ICML*.

A. Goldberg and X. Zhu. 2006. Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization. In *Proceedings of NAACL Workshop*.

Y. He, C. Lin and H. Alani. 2011. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of ACL*, pages 123-131.

M. Hu and B. Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of KDD*.

X. Hu, J. Tang, H. Gao, and H. Liu. 2013. Unsupervised sentiment analysis with emotional signals. In *Proceedings of WSDM*.

X. Hu, N. Sun, C. Zhang, and T. Chua. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowldge. In *Proceedings of CIKM*, pages 919-928.

H. Kim and H. Park. 2008. Non-negative matrix factorization based on alternating non-negativity constrained least squares and active set method. *SIAM J Matrix Anal Appl*, 30(2):713-730.

D. Lee and H. Seung. 2001. Algorithms for non-negative matrix factorization. In *Proceedings of NIPS*.

S. Li, Z. Wang, G. Zhou, and S. Lee. 2011. Semi-supervised learning for imbalanced sentiment classification. In *Proceedings of IJCAI*, pages 1826-1831.

T. Li, Y. Zhang, and V. Singhwani. 2009. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of ACL*, pages 244-252.

C. Lin. 2007. Projected gradient methods for nonnegative matrix factorization. *Neural Comput*, 19(10):2756-2779.

B. Liu. 2012. Sentiment analysis and opinion mining. *Morgan & Claypool Publishers*.

B. Lu, C. Tan, C. Cardie, and B. Tsou. 2011. Joint bilingual sentiment classification with unlabeled parallel corpora. In *Proceedings of ACL*, pages 320-330.

Y. Lu, M. Castellanos, U. Dayal, and C. Zhai. 2011. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of WWW*, pages 347-356.

X. Meng, F. Wei, X. Liu, M. Zhou, G. Xu, and H. Wang. 2012. Cross-lingual mixture model for sentiment classification. In *Proceedings of ACL*, pages 572-581.

V. Ng, S. Dasgupta, and S. Arifin. 2006. Examing the role of linguistic knowlege sources in the automatic identification and classificaton of reviews. In *Proceedings of ACL*.

S. Pan, X. Ni, J. Sun, Q. Yang, and Z. Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of WWW*.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Informaiton Retrieval*, pages 1-135.

B. Pang, L. Lee, S. Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79-86.

S. Riedel, L. Yao, A. McCallum, and B. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of NAACL*.

W. Peng and D. Park. 2011. Generative adjective sentiment dictionary for social media sentiment analysis using constrained nonnegative matrix factorization. In *Proceedings of ICWSM*.

S. Roweis and L. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323-2326.

V. Sindhwani and P. Melville. 2008. Document-word co-regulariztion for semi-supervised sentiment analysis. In *Proceedings of ICDM*, pages 1025-1030.

J. Tenenbaum, V. Silva, and J. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319-2323.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics*.

P. Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of ACL*, pages 417-424.

X. Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of ACL*, pages 235-243.

D. Zhou, Q. Bousquet, T. Lal, J. Weston, and B. Scholkopf. 2004. Learning with local and global consistency. In *Proceedings of NIPS*.

G. Zhou, F. Liu, Y. Liu, S. He, and J. Zhao. 2013. Statistical machine translation improves question retrieval in community question answering via matrix factorization. In *Proceedings of ACL*, pages 852-861.

X. Zhu, Z. Ghahramani, and J. Lafferty. 2003. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of ICML*.