Feature co-shrinking for co-clustering

Qi Tan\textsuperscript{a}, Pei Yang\textsuperscript{b,c,}\dagger, Jingrui He\textsuperscript{c}

\textsuperscript{a}Department of Computer Science and Engineering, South China Normal University, Guangzhou 510630, China
\textsuperscript{b}Department of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China
\textsuperscript{c}Department of Computer Science and Engineering, Arizona State University, Tempe 85281, USA

A R T I C L E   I N F O

Article history:
Received 4 July 2017
Revised 30 October 2017
Accepted 9 December 2017
Available online 11 December 2017

Keywords:
Co-clustering
Non-negative matrix tri-factorization
Co-sparsity
Co-feature-selection

A B S T R A C T

Many real-world applications require multi-way feature selection rather than single-way feature selection. Multi-way feature selection is more challenging compared to single-way feature selection due to the presence of inter-correlation among the multi-way features. To address this challenge, we propose a novel non-negative matrix tri-factorization model based on co-sparsity regularization to facilitate feature co-shrinking for co-clustering. The basic idea is to learn the inter-correlation among the multi-way features while shrinking the irrelevant ones by encouraging the co-sparsity of the model parameters. The objective is to simultaneously minimize the loss function for the matrix tri-factorization, and the co-sparsity regularization imposed on the model. Furthermore, we develop an efficient and convergence-guaranteed algorithm to solve the non-smooth optimization problem, which works in an iteratively update fashion. The experimental results on various data sets demonstrate the effectiveness of the proposed approach.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

Feature selection plays an important role in many machine learning and data mining applications \[1,2\]. However, most of the previous work focus on single-way feature selection, which may not be the most appropriate for the scenarios in need of two-way or multi-way feature selection. In this paper, we study the two-way feature selection in the context of co-clustering \[3\], which is the problem of simultaneously clustering rows and columns of a data matrix. Co-clustering has been widely used in many practical applications such as the simultaneous clustering of documents and words in text mining \[3\], reviews and ratings in social media analysis \[4\], users and movies in recommender systems \[5\], etc. Co-clustering is desirable over single-sided clustering since simultaneous grouping of rows and columns is more informative and digestible \[5\]. In this context, we are interested in conducting co-feature-selection for co-clustering, which could lead to better performance by co-shrinking the irrelevant features.

To address the problem of two-way feature selection, we propose a novel Non-negative Matrix Triplet Factorization model based on feature Co-Shrinking (NMTFCoS) to conduct the co-feature-selection in co-clustering. It aims to simultaneously minimize the loss function for the matrix tri-factorization, and the co-sparsity regularization imposed on the model. The key idea is to co-shrink the features by encouraging the co-sparsity of the model parameters. The motivation behind the proposed method is that we aim to mimic the success in co-clustering which seeks the co-clusters with inter-correlation. To this end, we co-learn the latent features that are inter-related with each other to effectively leverage the intertwining information between row-clusters and column-clusters. Taking the co-clustering of documents and words as an example, it can be formulated as a matrix tri-factorization problem. The co-sparsity constraint imposed on the model is to weigh the latent topics for both document clusters and word clusters. The resulting optimization problem is challenging due to the non-convex and non-smooth properties of the objective function. We transform the non-smooth objective to a smooth one by making use of an auxiliary function. Then, an efficient algorithm is proposed to solve the reformulated problem. We prove the convergence of the algorithm. The experimental results on various data sets demonstrate the effectiveness of the proposed approach.

The main contributions can be summarized as follows:

- We propose a novel non-negative matrix tri-factorization model based on co-sparsity regularization to enable the co-feature-selection for co-clustering. It aims to learn the inter-correlation among the multi-way features while co-shrinking the irrelevant ones by encouraging the co-sparsity of the model parameters.
- We propose an efficient algorithm to solve the non-smooth optimization problem. It works in an iteratively update fashion, and is guaranteed to converge.
• Experimental results on various data sets show the effectiveness of the proposed approach.

The rest of the paper is organized as follows. After a brief review of the related work in Section 2, we present the NMTFCos model and the optimization algorithm in Section 3. Section 4 shows the experimental results. Finally, we conclude the paper in Section 5.

2. Related work

Since the proposed co-feature-selection approach for co-clustering is based on matrix decomposition, we review the related work on feature selection, non-negative matrix factorization, and co-clustering.

High-dimensionality of data has posed serious challenges (e.g., the curse of dimensionality) to many data mining algorithms. Feature selection aims to select a subset of relevant features according to certain relevance evaluation criterion, which usually leads to better learning performance, lower computational cost, and better model interpretability. Some recent survey papers have provided the comprehensive reviews of feature selection in the settings of classification [2], semi-supervised learning [6], and clustering [7]. Feature selection methods can also be divided into three categories [1]: filter methods based on various criteria such as generalized Fisher score, mutual information and F-statistic [8]; wrapper methods such as expectation-maximization based method FSSEM [9] and SVM based method SVM-RFE [10]; and embedded methods based on various regularizations such as Lasso penalized methods [11], elastic net penalized method [12], $\ell_2,1$-norm penalized method [13], and sparse discriminative method [14]. However, most of the existing methods are limited to single-way feature selection.

Non-negative matrix factorization (NMF) aims to extract data-dependent non-negative basis functions, which has been given much attention due to its part-based and easy interpretable representation [15,16]. Incorporating extra constraints such as smoothness [17], sparsity [18], or orthogonality [19] was shown to improve the performance [20]. Various extensions of NMF have been proposed, including two-factor decomposition methods such as Semi-NMF [21], Convex-NMF [21], Robust NMF [18], sparse NMF [22], weighted NMF [23], versatile sparse NMF [24], and tri-factorization methods such as ONMF [19] and PMTF [25]. Some recent work extended NMF to multi-layer structure, such as DeepSemiNMF [26] which applied the concept of Semi-NMF [21] to a multi-layer structure, and a non-negative deep network architecture [27] resulting from unfolding the NMF iterations and untying its parameters.

Different with two-factor decomposition, Orthogonal non-negative matrix tri-factorization (ONMF) [19] simultaneously clustered rows and columns of a data matrix, which can be viewed a matrix implementation of co-clustering [3]. Co-clustering [3] is a powerful tool for unsupervised data analysis in that it is able to effectively intertwine row and column information. The orthogonality constraint was imposed to enhance sparsity in the ONMF model. The FNMTF [28] approach imposed hard-clustering constraint on the model by requiring the factor matrices to be cluster indicator, and regularized by the manifold structures in both data and feature spaces. The probabilistic matrix tri-factorization (PMTF) [25] allowed for the soft-clustering of both instances and features. GTSC [29] is a tensor spectral co-clustering method that simultaneously clusters the rows, columns, and slices of a non-negative tensor. The multi-view co-clustering method [30] performs co-clustering in each view to identify both row clusters and column clusters simultaneously and constrains the row clusters from the different views to be the same. Co-ClusterD [31] is a distributed framework for data co-clustering. Co-clustering has shown superiority over single-sided clustering on various applications, such as text mining [3], social media analysis [4], recommender systems [5], semiconductor manufacturing [32], and cargo pricing optimization [33].

In summary, although some two-factor factorization methods such as SparseNMF [34] and VSMF [24] impose $\ell_1$ sparsity regularization on both factors, they are not able to leverage the intercorrelation among the co-clusters. Matrix tri-factorization methods such as ONMF [19], by contrast, could take advantage of such intrinsic correlations. Therefore, it is a promising way to promote the performance of matrix tri-factorization methods by encouraging the co-sparsity of model parameters to enjoy the benefits resulting from both co-clustering and co-sparsity. Inspired by this idea, our proposed NMTFCos approach moves forward along non-negative tri-factorization by using co-sparsity regularization to simultaneously conduct co-feature-selection for co-clustering, which is distinctive from the existing work.

3. Feature co-shrinking for co-clustering

We first introduce the proposed feature co-shrinking approach for co-clustering, then present an efficient algorithm to solve the resulting optimization problem.

3.1. The proposed NMTFCos model

Let $P \in \mathbb{R}^{n \times m}$ be the input data matrix, $X \in \mathbb{R}^{r \times k}$ the row encoding matrix, $Y \in \mathbb{R}^{m \times p}$ the column encoding matrix, $W \in \mathbb{R}^{k \times p}$ the correlation matrix. We try to use $X$, $W$, and $Y$, to recover $P$, i.e., $P=XWY^T$.

The $i$th row and $j$th column vectors of a matrix $W$ are represented by $W_i$ and $W_j$, respectively. diag($v$) returns a square diagonal matrix with the elements of vector $v$ on the main diagonal. $\|W\|_F$ is the Frobenius norm of matrix $W$. The $\ell_1,q$ norm of a matrix $W$ is defined as $\|W\|_{1,q} = \sum_{i=1}^r \left( \sum_{j=1}^m \|W_{ij}\|^q \right)^{1/q}$. For a mixed-sign matrix $A$, we separate the positive and negative parts of $A$ as $A^+ = \frac{1}{2} (A + |A|)$ and $A^- = \frac{1}{2} (A - |A|)$, respectively. For two matrices $X$ and $Y$, let $X \odot Y$, $X \circ Y$, and $\frac{X}{Y}$ be the Hadamard product (or entrywise product), Kronecker product, and Hadamard division, respectively.

Let $x = \text{vec}(X)$ be vectorization of the matrix $X$ into a vector $x$.

In this paper, we conduct co-clustering under the non-negative matrix tri-factorization framework, which simultaneously divides rows and columns of a data matrix into their corresponding clusters. The main idea of the proposed NMTFCos approach is to use co-sparsity regularization to co-select the most informative features while shrinking the irrelevant ones. The objective is to minimize the reconstruction loss resulting from using three matrices, $X$, $Y$, and $W$, to recover the data matrix $P$. Meanwhile, we place the co-sparsity constraint on the correlation matrix $W$, and the non-negative constraint on both $X$ and $Y$. Therefore, it is to minimize,

$$\min_{X \geq 0, Y \geq 0} \min_{W \geq 0} \min_{\alpha, \beta} \left\| P - XWY^T \right\|_F + \alpha \left\| W \right\|_{2,1} + \beta \left\| W^2 \right\|_{2,1},$$

(1)

where $\alpha$ and $\beta$ are non-negative trade-off parameters to control the co-sparsity of $W$. Note that we use the square loss function here. But other kinds of loss functions such as KL-divergence are also applicable. The non-negative constraints on both $X$ and $Y$ allow for clustering interpretation. No non-negative constraint is imposed on $W$, enabling the model to be applicable to a mixed-sign data matrix $P$.

Fig. 1 shows a graphical interpretation of the NMTFCos model. The $\|W\|_{2,1}$ sparsity regularization encourages certain rows (say $i$ and $j$) of $W$ to be sparse in order to filter out the corresponding columns in $X$. Likewise, the $\|W^2\|_{2,1}$ sparsity regularization encourages certain columns (say $s$ and $t$) of $W$ to be sparse, so as to filter out irrelevant rows in $Y^T$ (i.e., corresponding columns in...
Proof. The original objective function in Eq. (1) is rewritten into:

\[ F(W) = \| P - XWY^T \|_F^2 + \alpha \sum_i \| W_i \|_2 + \beta \sum_j \| W_j \|_2 \]

The key step is to find an auxiliary function which should meet two conditions: 1) \( G(W, W) = F(W) \); 2) \( G(W, W^{(t)}) \geq F(W) \). Define

\[
G(W, W^{(t)}) = \| P - XWY^T \|_F^2 + \alpha \sum_i \| W_i \|_2^2 + \frac{\| W_i \|_2^2}{2\| W_i \|_2} + \beta \sum_j \frac{\| W_j \|_2^2}{2\| W_j \|_2}
\]

where \( W^{(t)} \) is the value of \( W \) at \( t \)-iteration. It is easy to verify \( G(W, W) = F(W) \). Also, the second condition follows from \( a^2 + b^2 \geq 2ab \) for any scalars \( a \) and \( b \). Since \( G(W, W^{(t)}) \) is an auxiliary function of \( F(W) \), \( F(W) \) is non-increasing under the update \( W^{(t+1)} = \arg \min_W G(W, W^{(t)}) \)

We can obtain the derivative of \( G(W, W^{(t)}) \):

\[
\frac{\partial}{\partial W} G(W, W^{(t)}) = \frac{\partial}{\partial W} \left( \| P - XWY^T \|_F^2 + \alpha \sum_i \frac{\| W_i \|_2^2}{2\| W_i \|_2} + \beta \sum_j \frac{\| W_j \|_2^2}{2\| W_j \|_2} \right)
\]

Since \( F(W^{(t)}) = G(W, W^{(t)}) \geq \min_{W} G(W, W^{(t)}) = G(W^{(t+1)}) \), the objective function \( F(W) \) is non-increasing under the above update. \( \square \)

Theorem 1 provides an iterative update rule to obtain the optimal solution for \( W \). Note that the objective in Eq. (2) is an unconstrained quadratic optimization problem. We can obtain the analytic solution for \( W \), which is shown in Theorem 2.

Theorem 2 (Optimum for \( W \)). The optimal solution for Eq. (2) is as follows:

\[
\text{vec}(W) = \left( (Y^T Y) \otimes (X^T X) + D^{-1} \text{vec}(X^T PY) \right)
\]

where \( D = \alpha I_p \otimes D_{\ell_1} + \beta D_{\ell_1} \otimes I_{k \times k} \).

Proof. The zero gradient condition of Eq. (2) with respect to \( W \) gives

\[
X^T X W Y + \alpha D_{\ell_1} W + \beta W D_{\ell_1} = X^T P Y
\]

Based on the property of Kronecker product, \( \text{vec}(XWY^T) = (Y \otimes X) \text{vec}(W) \). We have

\[
\left( (Y^T Y) \otimes (X^T X) \right) \text{vec}(W) + D \text{vec}(W) = \text{vec}(X^T PY)
\]

\[
\Rightarrow \text{vec}(W) = \left( (Y^T Y) \otimes (X^T X) + D \right)^{-1} \text{vec}(X^T PY)
\]

which completes the proof. \( \square \)

Second, we derive the update rule for \( Y \). When both \( X \) and \( W \) are fixed, the objective in Eq. (1) with respect to \( Y \) is a quadratic optimization problem with non-negative constraint. Theorem 3 shows the multiplicative updating rule for \( Y \).

Theorem 3 (Update Rule for \( Y \)). The objective in Eq. (1) with respect to \( Y \) is non-increasing under the update

\[
Y \leftarrow Y \odot \left( \frac{(P^T X W)^+ + Y (W^T X^T X) W}{(P^T X W)^+ + Y (W^T X^T X) W} \right)
\]

Also, the solution satisfies the Karush-Kuhn-Tucker complementarity condition.

Proof. Denote \( F = X W \), and

\[
J(Y) = \| P - FY^T \|_F^2 = tr(P^T P - 2P^T FY^T + YF^TF^T)
\]
s.t. \( Y \geq 0 \)

We introduce the Lagrangian function:

\[
L(Y) = tr\left[-2P^TFY^T + YF^FY^T - 2\mu Y^T\right]
\]

s.t. \( \mu \geq 0 \)

where \( \mu \) are the Lagrangian multipliers. The zero gradient condition of \( L(Y) \) with respect to \( Y \) gives

\[
\frac{\partial}{\partial Y} L(Y) = -2P^TF + 2YF^FY^T - 2\mu = 0 \Rightarrow \mu = -P^TF + YF^TY
\]

Based on the Karush-Kuhn-Tucker (KKT) complementarity condition, we have

\[
\mu \odot Y = 0 \Rightarrow (-P^TF + YF^TY) \odot Y = 0
\]

Similar to [21], we can obtain the update rule for \( Y \):

\[
Y \leftarrow Y \odot \left[ \frac{(P^T F)^+ + Y(F^T F)^+}{(P^T F)^+ + Y(F^T F)^+} \right]
\]

Next, we demonstrate when the updating algorithm converges, the solution satisfies the complementarity condition. At convergence, \( Y^{(\infty)} = Y^{(t+1)} = Y^{(t)} \) where \( t \) is the iteration index, i.e.,

\[
(Y \odot Y) \odot \left[ (P^T F)^- + Y(F^T F)^+ \right] = (Y \odot Y) \odot \left[ (P^T F)^+ + Y(F^T F)^- \right]
\]

\[
\Leftrightarrow (Y \odot Y) \odot \left[ -P^TF + YF^TY \right] = 0 \Leftrightarrow \left[ -P^TF + YF^TY \right] \odot Y = 0
\]

Finally, we can obtain the update rule,

\[
Y \leftarrow Y \odot \left[ \frac{(P^TWX)^+ + Y(W^T Y^TXW) \odot Y}{(P^TWX)^+ + Y(W^T Y^TXW)} \right]
\]

which completes the proof. \( \square \)

Third, we derive the update rule for \( X \). Likewise, when both \( Y \) and \( W \) are fixed, the objective in Eq. (1) with respect to \( X \) is a constrained quadratic optimization problem. The multiplicative updating rule for \( X \) is derived in Theorem 4, which can be proved in a way similar to Theorem 3.

**Theorem 4** (Update Rule for \( X \)). The objective in Eq. (1) with respect to \( X \) is non-increasing under the update

\[
X \leftarrow X \odot \left[ \frac{(PYWT)^+ + X(Y^TWY^TW)^+}{(PYWT)^+ + X(Y^TWY^TW)} \right].
\]

Also, the solution satisfies the Karush-Kuhn-Tucker complementarity condition.

**Proof.** Denote \( G = Y^TW^T \), and

\[
J(X) = \|P - XG^T\|^2 = tr(P^TP - 2PGX^T + XG^T GX)
\]

s.t. \( X \geq 0 \)

Likewise, we can obtain the update rule for \( X \):

\[
X \leftarrow X \odot \left[ \frac{(PG)^+ + X(G^T G)^+}{(PG)^+ + X(G^T G)^+} \right] \odot \left[ \frac{(PYWT)^+ + X(Y^TWY^TW)^+}{(PYWT)^+ + X(Y^TWY^TW)} \right].
\]

Also, the solution satisfies the KKT complementarity condition. \( \square \)

**Algorithm 1** NMTFCoS Algorithm.

**Input:** Data matrix \( P \in \mathbb{R}^{n \times m} \), dimensions \( k \) and \( p \), parameters \( \alpha \), \( \beta \).

**Output:** \( W \in \mathbb{R}^{k \times p} \), \( X \in \mathbb{R}^{m \times k} \), \( Y \in \mathbb{R}^{m \times p} \).

1. Set \( t = 0 \);
2. Initialize \( D_f^{(t)} \) and \( D_c^{(t)} \) as identity matrices;
3. Initialize \( X^{(t)} \) and \( Y^{(t)} \) by using traditional clustering algorithm;
4. repeat
5. Update \( W^{(t+1)} \) by Eq. (5);
6. Update \( X^{(t+1)} \) by Eq. (7);
7. Update \( Y^{(t+1)} \) by Eq. (6);
8. Update \( D_f^{(t+1)} \) and \( D_c^{(t+1)} \) by Eqs. (3) and (4), respectively;
9. Set \( t = t + 1 \);
10. until converged

The proposed NMTFCoS algorithm is shown in Algorithm 1. First, it initializes the encoding matrices, \( X \) and \( Y \), by using traditional clustering algorithms. In our experiment, we use PLSA\(^1\) to obtain the initialized clusters for both \( X \) and \( Y \). The diagonal matrices, \( D_f \) and \( D_c \), are initialized as identity matrices. Then, the algorithm updates \( X \), \( Y \), \( W \), \( D_f \), and \( D_c \) in an iterative way until convergence. We prove the convergence of the proposed method in Theorem 5.

**Theorem 5** (Convergence). The proposed NMTFCoS algorithm is guaranteed to converge to the local optimum.

**Proof.** The objective is non-increasing under the updates:

- When both \( X \) and \( Y \) are fixed, the objective in Eq. (1) is non-increasing under the updating of \( W \) according to Theorem 1;
- When both \( X \) and \( W \) are fixed, the objective in Eq. (1) is non-increasing under the updating of \( Y \) according to Theorem 3;
- When both \( W \) and \( Y \) are fixed, the objective in Eq. (1) is non-increasing under the updating of \( X \) according to Theorem 4.

Since the objective is block-wise convex, according to Theorem 3.1 in [35], the iterative updating algorithm converges to a local optimum. \( \square \)

In each iteration, the computations involved in Step 6 and 7 are mainly matrix multiplications which can be solved efficiently. In Step 5, the algorithm complexity is related to \( k \) and \( p \), which are the sizes of dimensions of the correlation matrix \( W \). Usually we have \( k \ll m \) and \( p \ll n \). Since \( W \) is sparse, we can also take advantage of this property to save space and speed up the computation.

4. Experiments

In order to evaluate the effectiveness of the proposed approach, we conduct the detailed experiments on the benchmark data sets in comparison with the related algorithms. In particular, we focus on using the proposed algorithm to improve the clustering performance.

4.1. Setup

Three text data sets are used in our experiments, including Reuters21578, 20Newsgroups, and TDT2 data\(^2\). These data sets are widely used for the evaluation of the clustering algorithms.

The 20Newsgroups data set is a collection of approximately 20,000 documents. The documents are organized into 20 different newsgroups, each corresponding to a different topic. Some of

---

the newsgroups are very closely related to each other, whereas others are highly unrelated. Reuters21578 corpus contains 21,578 documents in 135 categories. The documents in Reuters 21578 collection are Reuters newswire stories. There are 8293 documents in 65 categories after removing those documents with multiple category labels. This corpus contains 18,933 distinct terms after data preprocessing. The TDT2 corpus (Nist Topic Detection and Tracking corpus) comprises 11,201 on-topic documents which are classified into 96 semantic categories. We use a subset of TDT2 corpus, in which those documents appearing in two or more categories were discarded, and only the largest 30 categories were kept. It left us with 9394 documents in total. Table 1 shows the statistics information including the numbers of instance, feature, and category in each data set.

NMF based methods have been widely used for clustering. We follow the approach [36] to apply the NMF based methods to clustering. Consider \( P \in \mathbb{R}^{n \times m} \) as an instance-feature matrix with \( n \) instances and \( m \) features. Given the tri-factorization \( P \approx X Y F \) or two-factor factorization \( P \approx X F \), we use matrix \( X \in \mathbb{R}^{n \times k} \) to group the \( n \) instances into \( k \) clusters. Each instance is placed into a cluster, i.e., the \( i \) instance is placed in cluster \( j \) if \( X_{ij} \) is the largest entry in the \( i \)th row, where \( 1 \leq i \leq n \) and \( 1 \leq j \leq k \).

We compare NMTFCoS with various algorithm such as two-factor factorization methods including NMF [16], SparseNMF2 [34], and versatile sparse NMF [24] (or VSMF for short), and tri-factorization method ONMF [19]. We also compare NMTFCoS with DeepSemiNMF [26] which is a deep two-factor factorization model. SparseNMF2 [34] imposes \( \ell_1 \) norm sparsity regularization on both factors, while VSMF [24] imposes the elastic net [37] regularization on both factors. VSMF [24] is a generic model. The standard NMF [16], Sparse NMF [22], and Semi-NMF [21] are the specific cases of VSMF. SparseNMF2 [34] and VSMF [24] are closely related to our method since both of them encourage the sparsity of the two factors. The source code for DeepSemiNMF [26] is available online\(^3\). The other comparison algorithms are implemented in the NMF toolbox [34], which is publicly available\(^4\). NMF, SparseNMF2, VSMF, and ONMF correspond to the functions “nmfrule”, “sparsenmf2rule”, “vsmf”, and “orthnmfrule”, respectively.

Normalized mutual information (NMI) [38] is a popular metric for determining the clustering performance. Let \( \lambda^{(a)} \) and \( \lambda^{(b)} \) be the cluster labelings with \( k^{(a)} \) and \( k^{(b)} \) groups respectively. Note that \( n \) is the number of total instances. Let \( n_{h,l}^{(a)} \) be the number of instances in cluster \( C_h \) according to cluster labeling \( \lambda^{(a)} \), and \( n_{h,l}^{(b)} \) the number of instances in cluster \( C_l \) according to cluster labeling \( \lambda^{(b)} \). Let \( n_{h,l} \) denote the number of instances that are in cluster \( C_h \) according to \( \lambda^{(a)} \) as well as in group \( C_l \) according to \( \lambda^{(b)} \). Then, NMI is defined as:

\[
\text{NMI} = \frac{\sum_{h=1}^{k^{(a)}} \sum_{l=1}^{k^{(b)}} n_{h,l} \log \left( \frac{n_{h,l}^{(a)} n_{h,l}^{(b)}}{n_h^{(a)} n_h^{(b)}} \right)}{\sqrt{\left( \sum_{h=1}^{k^{(a)}} n_h^{(a)} \log \frac{n_h^{(a)}}{n} \right) \left( \sum_{l=1}^{k^{(b)}} n_l^{(b)} \log \frac{n_l^{(b)}}{n} \right)}}
\]

Note that \( 0 \leq \text{NMI} \leq 1 \). In general, the larger the value of NMI, the better the clustering quality. In our experiments, \( \lambda^{(a)} \) and \( \lambda^{(b)} \) correspond to the cluster labelings given by the clustering algorithm and the true labelings known as the document categories, respectively.

4.2. Performance study

Figs. 2–4 show the clustering performances for different algorithms on 20Newsgroups, Reuters-21578, and TDT2 data sets, respectively. In each figure, y-axis represents NMI, and x-axis represents the number of document clusters \( k \). We repeat the experiments 10 times on each data set, and report the average NMI and the standard deviation.

From these figures, we can see that the proposed NMTFCoS algorithm performs better than the other methods on the three data sets. In contrast, the performance of the standard NMF [16] is limited due to the issues of scale-variance and non-unique solutions. It implies that the non-negativity constrained on the factorization model is insufficient in some cases. VSMF [24] performs better than NMF [16] and SparseNMF2 [34] on the three data sets. VSMF provides a flexible mechanism to control sparsity, smoothness, and non-negativity on both basis matrix and coefficient matrix. It shows that the elastic net regularization [37] adopted by VSMF helps improve the performance by encouraging sparsity and enhancing smoothness of the model simultaneously. Different from two-factor NMF methods, ONMF [19] is a tri-factorization model. ONMF shows the comparable performance with VSMF on both

---

\(^3\) https://github.com/trigeorgis/deep-semi-nmf.
\(^4\) https://sites.google.com/site/nmfool/.

---

<table>
<thead>
<tr>
<th>Data set</th>
<th>Instances</th>
<th>Features</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>20Newsgroups</td>
<td>18,846</td>
<td>26,214</td>
<td>20</td>
</tr>
<tr>
<td>Reuters-21578</td>
<td>8293</td>
<td>18,933</td>
<td>65</td>
</tr>
<tr>
<td>TDT2</td>
<td>9394</td>
<td>36,771</td>
<td>30</td>
</tr>
</tbody>
</table>
20Newsgroups and Reuters21578 data. However, the disjoint property on ONMF may be too restrictive for some applications.

In [26], the authors suggested to use the two hidden layer architectures for DeepSemiNMF. However, we have a different observation in our experiments. We found that DeepSemiNMF with the five hidden layer architecture performed better than the one with the two hidden layer architecture in these data sets. We also experimented with a high-number of layers, but the additional number of layers did not seem to have a significant impact on the results. Therefore, we empirically adopt the five hidden layer architecture for DeepSemiNMF: 1600-800-400-200-k, where k is the number of features in the top layer, which ranges in [20, 40, 60, 80, 100], corresponding to the number of clusters. The results show that DeepSemiNMF [26] improves upon the single-layered SemiNMF [21] by capturing the hierarchical representations of data. However, DeepSemiNMF does not impose the non-negativity constraint on its base matrix, which would not be well suited for the document-word data used in our experiment. Also, DeepSemiNMF is a two-factorization method, which is not able to leverage the inter-correlations among clusters. These might account for the results that DeepSemiNMF underperforms our proposed method in the three data sets.

The performance superiority of NMTFCoS over the comparison methods demonstrates the effectiveness of the proposed approach. The competency of NMTFCoS is twofold. First, NMTFCoS takes advantage of co-clustering to cluster rows and columns simultaneously. Unlike single-sides clustering methods which seeks similar either rows or columns only, co-clustering seeks blocks (or co-clusters) of rows and columns that are inter-related. Second, NMTFCoS uses co-sparsity constraint to conduct co-feature-selection for co-clustering. In the text data sets, NMTFCoS co-clusters documents and words, while truncating the irrelevant latent topics from both document clusters and word clusters. Analogous to co-clustering, co-sparsity regularization encourages the model to co-learn the latent features that are inter-related with each. Intertwining both row and column information can effectively leverage the correlation among them, and thus yield better quality clusters.

4.3. Parameter sensitivity

We first examine how the number of word clusters $p$ affects the performance of NMTFCoS. The results on 20Newsgroups data set are shown in Fig. 5. We vary $p$ from 20 to 100. The algorithm performs better when $p$ gradually increases. But a relatively large $p$ (e.g., $p = 100$) may hurt the performance. Nonetheless, the performance curve of NMTFCoS is relatively flat, and not very sensitive to the number of word clusters.

Next, we study the performance of NMTFCoS varying with the parameters, $\alpha$ and $\beta$, which are used to control the importance of the co-sparsity regularization. Fig. 6 shows the result for $\alpha$ and $\beta$ on 20Newsgroups data set. We can see that NMTFCoS performs worse as $\alpha$ approaches 0 when no sparsity constraint is imposed on the model. The optimal performance is achieved at $\alpha = 0.001$. But a too large $\alpha$ (e.g., $\alpha > 10$) value will hurt the performance which indicates too much importance is put on the sparsity regularization. Nevertheless, NMTFCoS is robust over a wide range of $\alpha$ values. The figure shows a similar trend of the algorithm performance varying with $\beta$. Also, it is worth noting that NMTFCoS with $\alpha = 0$ or $\beta = 0$ corresponds to the special case of the proposed model with one-way sparsity regularization only. Both results suggest that NMTFCoS with co-sparsity regularization could yield the better performance over the one with one-sided sparsity constraint.

4.4. Convergence

We also study the convergence property of the proposed algorithm. Fig. 7 shows the performance of NMTFCoS varying with the iteration on 20Newsgroups data. We can see that the performance is poor at the first iteration, which indicates that the initialized clusters obtained by using PLSA is not satisfactory. Then, the NMI curve climbs sharply in the first several iterations and become stable after about 30 iterations. It shows that NMTFCoS could significantly improve upon the traditional clustering algorithm by conducting co-feature-selection for co-clustering. Also, the results demonstrate the fast convergence property of the proposed algorithm.
Fig. 7. NMI varies with iteration.

5. Conclusion

In this paper, we propose a novel non-negative matrix tri-factorization model based on co-sparsity regularization to co-learn the features for co-clusters by simultaneously shrinking the irrelevant ones. Then, an efficient algorithm is proposed to solve the non-convex and non-smooth optimization problem. We prove the convergence of the proposed algorithm. The effectiveness of the proposed algorithm is verified on various data sets. For future work, we will extend the proposed model from two-way to multi-way feature selection by modeling the multiple types of heterogeneous information with tensor instead of matrix.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant No. 61473123, Natural Science Foundation of Guangdong Province under Grant No. 2017A030313370, National Science Foundation under Grant No. IIS-1552654, ONR under Grant No. N00014-15-1-2821, and an IBM Faculty Award. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

References

Qi Tan is an assistant professor at South China Normal University. She received her Ph.D. from South China University of Technology. Her research interests include statistical machine learning and data mining, such as transfer learning, multi-view learning, clustering, feature selection, etc. She has published about 10 research papers on referred journals and conferences such as NCAI, IJPM, KAIS, KDD, ACL, ADMA, etc.

Pei Yang is an associate professor at South China University of Technology. He received his Ph.D. from South China University of Technology. He has been conducting postdoctoral research in Arizona State University. His research focuses on statistical machine learning and data mining such as heterogeneous learning, semi-supervised learning, transfer learning, rare category analysis, and functional data analysis, with applications in web mining, big data analytics, bioinformatics, healthcare, etc. He has published over 30 research articles on referred journals and conference proceedings, such as TKDE, TKDD, JAIR, KDD, ICDM, SDM, IJCAI, ACL, WSDM, etc. He has served on the program committee of conferences including NIPS, ICML, KDD, ICDM, SDM, IJCAI, AAAI, etc.

Jingrui He is an assistant professor in the School of Computing, Informatics and Decision Systems Engineering at Arizona State University. She received her Ph.D. in Computer Science from Carnegie Mellon University. She joined ASU in 2014 and directs the Statistical Learning Lab (STAR Lab). Her research focuses on heterogeneous machine learning, rare category analysis, semi-supervised learning and active learning, with applications in healthcare, social network analysis, semiconductor manufacturing, etc. She is the recipient of the NSF CAREER Award in 2016, IBM Faculty Award in 2015 and 2014 respectively, and has published more than 60 refereed articles. She has served on the organizing committee/senior program committee of many conferences, including ICML, KDD, IJCAI, SDM, ICDM, etc. She is also the author of the book on Analysis of Rare Categories (Springer-Verlag, 2012).