A Matrix-based Approach for Semi-supervised Document Co-clustering

Yanhua Chen, Lijun Wang and Ming Dong

{chenyah, ljwang, mdong}@wayne.edu

Department of Computer Science, Wayne State University

Introduction

Co-clustering Algorithms

SS-NMF (Non-negative Matrix Factorization) Co-clustering

1. NMF tri-factorization:

Given a Heterogeneous Relational Data (HRD) set \( \{ x_i | i = 1, 2, \ldots, n \} \) with clusters numbers \( k \), each representing one data type, our goal is to simultaneously cluster these data into proper clusters.

Let \( R^{(p)} \in R^{n \times k} \) represent the relations between \( x_i \) and \( x_j \) \( (1 \leq p, q \leq k) \), then the task of co-clustering an optimization problem with nonnegative tri-factorization of \( R^{(p)} \):

\[
J = \min_{G^{(p)},G^{(q)},S^{(pq)}} \sum_{i,j} (R^{(p)}(i,j) - G^{(p)}(i,:) \cdot S^{(pq)} \cdot G^{(q)}(:,j))^2
\]

where \( G^{(p)} \in R^{n \times k} \) and \( G^{(q)} \in R^{k \times k} \) are the cluster indicator matrices, and \( S^{(pq)} \in R^{k \times k} \) is the cluster association matrix which gives the relation among the clusters of \( x_i \) and \( x_j \).

2. Define set of pairwise constraints:

a) Must-Link constraints: \( M = \{(i,j) \} \), where \( (x_i, x_j) \in M \) implies that \( x_i \) and \( x_j \) are labeled as belonging to the same cluster.

b) Cannot-Link constraints: \( C = \{(i,j) \} \), where \( (x_i, x_j) \in C \) implies that \( x_i \) and \( x_j \) are labeled as belonging to different clusters.

Related Work

Semi-supervised Clustering

Semi-supervised clustering uses class labels or pairwise constraints on examples to aid unsupervised clustering. Based on source information, existing methods for semi-supervised clustering generally fall into two categories: Constraint-based and Distance-based methods.

SS-NMF for Document Co-clustering

1. Distance Matrix Learning

The objective of pairwise co-clustering is to cluster the \( n \) documents along with the \( n \) words while constraint violations are kept to a minimum. Let \( R^{(n)} \in R^{n \times n} \) denote the word-document matrix, then the distance matrix can be learned by solving an optimization problem:

\[
\text{max } g(L^{(n)}) = \sum_{i,j} (d(i,j) - \langle c(i,j), L^{(n)} \rangle)^2
\]

where \( \langle . \rangle \) is the Frobenius matrix norm. This maximization problem is equivalent to the generalized Semi-Supervised Linear Discriminant Analysis (SS-LDA) problem as follows:

\[
J = \min_{R^{(n)}} \frac{\text{trace}(L^{(n)} W^{(n)} L^{(n)^T})}{\text{trace}(L^{(n)^T} B^{(n)})}
\]

where \( W^{(n)} \) is within-distance matrix and \( B^{(n)} \) is between-distance matrix.

Experiments and Results

Experiment setup

1. Datasets: We primarily utilized the different text data used in the University of Minnesota. We selected the top 100 words with mutual information for each document and mixed up some of the data sets mentioned above.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Data structure</th>
<th>No of documents</th>
<th>No of words</th>
<th>No of categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>Documents</td>
<td>10161</td>
<td>1637</td>
<td>214</td>
</tr>
<tr>
<td>Text</td>
<td>Documents</td>
<td>10032</td>
<td>1637</td>
<td>214</td>
</tr>
<tr>
<td>Text</td>
<td>Documents</td>
<td>10032</td>
<td>1637</td>
<td>214</td>
</tr>
<tr>
<td>Text</td>
<td>Documents</td>
<td>10032</td>
<td>1637</td>
<td>214</td>
</tr>
<tr>
<td>Text</td>
<td>Documents</td>
<td>10032</td>
<td>1637</td>
<td>214</td>
</tr>
<tr>
<td>Text</td>
<td>Documents</td>
<td>10032</td>
<td>1637</td>
<td>214</td>
</tr>
<tr>
<td>Text</td>
<td>Documents</td>
<td>10032</td>
<td>1637</td>
<td>214</td>
</tr>
<tr>
<td>Text</td>
<td>Documents</td>
<td>10032</td>
<td>1637</td>
<td>214</td>
</tr>
</tbody>
</table>

We also used eight datasets from Kent Ridge Biomedical Data Set Repository for gene expression clustering. In our experiment, all datasets are reduced to 2000 features by PCA transform.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Data structure</th>
<th>No of documents</th>
<th>No of words</th>
<th>No of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>BT1</td>
<td>Array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT1</td>
<td>Array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT1</td>
<td>Array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT1</td>
<td>Array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT1</td>
<td>Array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT1</td>
<td>Array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT1</td>
<td>Array</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT1</td>
<td>Array</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Evaluation accuracy metric:

\[
AC = \frac{\sum_{i} \delta(y_i, \hat{y}_i)}{R}
\]

SS-NMF for Word-Document Co-clustering

1. Updating Rules

Project \( R^{(p)} \) into new space \( \hat{R}^{(p)} = \sqrt{\text{trace}(L^{(p)} W^{(p)})} \), then perform:

\[
J = \min_{G^{(p)},G^{(q)},S^{(pq)}} \sum_{i,j} (R^{(p)}(i,j) - G^{(p)}(i,:) \cdot S^{(pq)} \cdot G^{(q)}(:,j))^2
\]

we get:

\[
G^{(p)} = G^{(p)} \frac{(R^{(p)2}G^{(q)2}S^{(pq)2})}{(G^{(p)2}G^{(q)2}S^{(pq)2})}
\]

where \( \langle . \rangle \) is the Frobenius matrix norm. This maximization problem is equivalent to the generalized Semi-Supervised Linear Discriminant Analysis (SS-LDA) problem as follows:

\[
J = \min_{R^{(n)}} \frac{\text{trace}(L^{(n)} W^{(n)} L^{(n)^T})}{\text{trace}(L^{(n)^T} B^{(n)})}
\]

We perform comparisons of other unsupervised co-clustering methods: KK, BSGP, CMRF, NMF, SS-KK, SS-CMRF with SS-NMF.

Table: Comparison of clustering accuracy between unsupervised KK, BSGP, CMRF, NMF, SS-KK, SS-CMRF, SS-NMF with 10% constraints on text pairwise (DOCUMENT-WORD) CO-CLUSTERING (DATA SETS CT1 - CT8) AND GENE EXPRESSION PAIRWISE (CONDITION-GENE) CO-CLUSTERING (DATA SETS BT1 - BT7).

![Figure: Comparison of clustering accuracy between SS-KK, SS-CMRF, and SS-NMF with different amounts of constraints on (a) text pairwise (i.e., document-word) co-clustering and (b) gene expression pairwise (i.e., condition-gene) co-clustering.

From the experimental results, the superior performance of SS-NMF is evident across all the data sets. AC values of BSGP or CMRF, on average, are about 10% lower than NMF, which is the best amongst the unsupervised methods. Moreover, SS-NMF outperforms SS-KK and SS-CMRF, especially in the data sets having more than 2 clusters, i.e., data sets CT5 to CT8 and BT7. It is also worth pointing out that the AC value of SS-NMF is as high as 98% on the data sets CT2, CT5, CT7 and BT6.