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

Yanhua Chen, Lijun Wang and Ming Dong
{chenyanh, ljwang, mdong}@wayne.edu
Department of Computer Science, Wayne State University

Machine Vision & Pattern Recognition Lab.

Introduction

Document co-clustering, an advancement in text mining, is a challenging problem without any prior knowledge or background information. We propose a Semi-Supervised Non-negative Matrix Factorization (SS-NMF) based framework for document co-clustering. Our method computes a new *word-document* matrix by incorporating user provided constraints through distance metric learning. Using an iterative algorithm, we perform tri-factorization of the new matrix to infer the document and word clusters. Through extensive experiments conducted on publicly available data sets, we demonstrate the superior performance of SS-NMF for document co-clustering.

Related Work

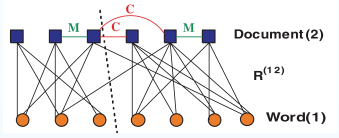
Co-clustering Algorithms

In general, co-clustering approaches can be divided into three categories:

1. *Latent semantic analysis:* LSA, multi-LSA.
2. *Information theory-based models:* Information theoretic models (such as mutual information and Bregman divergence) and Combinatorial Markov Random Field (CMRF) algorithm.
3. *Graph theoretic approaches:* Bipartitie Spectral Graph Partitioning (BSGP), Consistent Bipartite Graph Co-partitioning (CBGC), Sepectral Relational Clustering (SRC) and Relation Summary Network (RSN) model.

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: C*onstraint-based and Distance-based* methods.
Semi-supervised clustering algorithms: Semi-Supervised Kernel K-means (SS-KK), Semi-Supervised Spectral Normalize Cuts (SS-SNC), Semi-Supervised Non-Negative Matrix Factorization (SS-NMF) and Semi-Supervised CMRF (SS CMRF).



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

1. NMF tri-factorization:

Given a Heterogenous Relational Data (HRD) set , $\{\mathcal{X}_i \mid i=1,2,\cdots,l\}$ with clusters numbers $\{k_i \mid i=1,2,\cdots,l\}$, each representing one data type, our goal is to simultaneously cluster these data into proper clusters.
Let $R^{(pq)} \in R^{n_p \times n_q}$ represent the relations between $\mathcal{X}_p$ and $\mathcal{X}_q \left(1 \le p,q \le l\right)$, then the task of co-clustering as an optimization problem with nonnegative tri-factorization of $R^{(pq)}$:

$$J = \min_{G(p)\ge 0, G(q)\ge 0, S(pq)\ge 0} \sum_{1\le p,q\le l} \left\| R^{(pq)} - G^{(p)} S^{(pq)} G^{(q)} \right\|^2$$

where $G^{(p)} \in R^{n_p \times k_p}$ and $G^{(q)} \in R^{k_q \times n_q}$ are the cluster indicator matrces, and $S^{(pq)} \in R^{k_p \times k_q}$ is the cluster association matrix which gives the relation among the clusters of $\mathcal{X}_p$ and $\mathcal{X}_q$.

2. Define set of pairwise constraints:

a) Must-Link constraints: $M = \{(x_i, x_j)\}$, 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: $C = \{(x_i, x_j)\}$, where $(x_i, x_j) \in C$ implies that x_i and x_j are labeled as belonging to different clusters

SS-NMF for *Word-Document* Co-clustering

1. Distance Matrix Learning

The objective of pairwise co-clustering is to cluster the *n*2 documents along with the *n*1 words while constraint violations are kept to a minimum. Let $R^{(12)} \in R^{n_1 \times n_2}$ denote the *word-document* matrix , then the distance matrix can be learned by solving an optimization problem:

$$\max g(L^{(12)}) = \frac{\sum_{(x_i^{(12)}, x_j^{(12)}) \in C} \left\| x_i^{(12)}, x_j^{(12)} \right\|_{L^{(12)}}}{\sum_{(x_i^{(12)}, x_j^{(12)}) \in M} \left\| x_i^{(12)}, x_j^{(12)} \right\|_{L^{(12)}}}$$

where ||.|| is the Frobenius matrix norm. This maximization problem is equivalent to the generalized Semi-Supervised Linear Discriminate Analysis (SS-LDA) problem as follows,

$$J = \min \frac{trace(L^{(12)} W_M^{12})}{trace(L^{(12)} B_c^{12})}$$

Where W_M^{12} is within-distance matrix and B_c^{12} is between-distance matrix.

2. Updating Rules

Project $R^{(12)}$ into new space $\tilde{R}^{(12)} = \sqrt{L^{(12)}} R^{(12)}$, then perform:

$$J = \min_{G(1)\ge 0, G(2)\ge 0, S(12)\ge 0} \sum \left\| \tilde{R}^{(12)} - G^{(1)} S^{(12)} G^{(2)} \right\|^2$$

and we get:

$$G_{ih}^{(1)} = G_{ih}^{(1)} \frac{(\tilde{R}^{(12)} G^{(2)T} S^{(12)T})_{ih}}{(G^{(1)} S^{(12)} G^{(2)} G^{(2)T} S^{(12)T})_{ih}} \quad (1)$$

$$G_{ih}^{(2)} = G_{ih}^{(2)} \frac{(S^{(12)T} G^{(1)T} \tilde{R}^{(12)})_{ih}}{(S^{(12)} G^{(1)T} G^{(1)} S^{(12)} G^{(2)})_{ih}} \quad (2)$$

$$S_{ih}^{(12)} = S_{ih}^{(12)} \frac{(G^{(1)T} \tilde{R}^{(12)} G^{(2)T})_{ih}}{(G^{(1)T} G^{(1)} S^{(12)} G^{(2)} G^{(2)T})_{ih}} \quad (3)$$

Experiments and Results

Experiment setup

1. Datasets: We primarily utilized the different text data used in the University of Minnesota. In our experiments, we selected the top 1000 words by mutual information for each document and mixed up some of the data sets mentioned above.

Name	Data set and structure	No of clusters	No. of documents
CT1	oh15: Adenosine-Diphosphate, Blood-Vessels	2	154
CT2	oh15: Aluminum, Blood-Coagulation-Factors	2	122
CT3	re0: interest, reserves	2	261
CT4	Re0: housing, jobs	2	55
CT5	re0:housing interest, jobs	3	274
CT6	oh15: Aluminum, Blood-Vessels, Leucine	3	207
CT7	re0: cpi, housing, ipi, lei, retail	5	144
CT8	re0: bop, cpi, gnp, bousing, interest, ipi,jobs, lei,money, reserves	10	1150

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.

Name	Data set	Data structure	No of clusters	No. of conditions
BT1	AML	ALL, AML	2	72
BT2	Breastcancer	Replapse, Non-relapse	2	97
BT3	CentralNervous	Class1, Class2	2	60
BT4	ColonTumor	Positive, Negative	2	62
BT5	LunCancer	MPM, ADCA	2	181
BT6	Ovarian	Cancer, Normal	2	253
BT7	MLL	ALL, MLL, AML	3	72

2. Evaluation accuracy metric:

$$AC = \frac{\sum_{i=1}^{n} \delta(y_i, \hat{y}_i)}{n}$$

Experimental Results:

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, AND SEMI-SUPERVISED 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).

Name	KK	BSGP	CMRF	NMF	SS-KK	SS-CMRF	SS-NMF
CT1	0.7897	0.4870	0.5545	0.8052	0.9610	0.7984	0.8606
CT2	0.5164	0.6148	0.6582	0.6475	0.9444	0.9041	0.9902
CT3	0.6820	0.7510	0.7264	0.7586	0.7732	0.8682	0.8774
CT4	0.5355	0.7190	0.5419	0.4635	0.7667	0.8310	0.8248
CT5	0.4652	0.6148	0.4974	0.6364	0.871	0.7682	0.9818
CT6	0.4638	0.5072	0.5585	0.6763	1	0.7585	0.9101
CT7	0.4236	0.2778	0.5000	0.6667	0.996	0.7261	0.9944
CT8	0.2857	0.2330	0.3327	0.3774	0.7868	0.4667	0.6343
Ave	0.5191	0.5256	0.5462	0.6290	0.8194	0.7600	0.8842

Name	KK	BSGP	CMRF	NMF	SS-KK	SS-CMRF	SS-NMF
BT1	0.6050	0.8194	0.8238	0.6111	0.8606	0.9538	0.9444
BT2	0.6189	0.5155	0.6156	0.5258	0.7320	0.7426	0.7732
BT3	0.5000	0.6	0.5250	0.5833	0.6233	0.7147	0.7667
BT4	0.5000	0.7258	0.6452	0.6613	0.7613	0.8400	0.871
BT5	0.6570	0.5138	0.9118	0.8785	0.8569	1	1
BT6	0.5099	0.6522	0.5167	0.4704	0.6403	0.7393	0.996
BT7	0.3750	0.5417	0.4829	0.4306	0.4861	0.6778	0.8194
Ave	0.5380	0.6241	0.6459	0.5944	0.7086	0.8212	0.8815

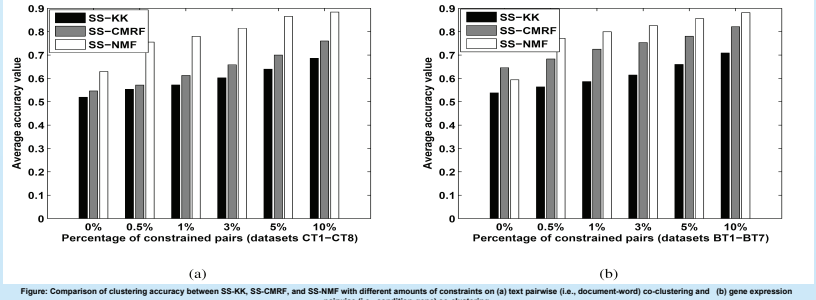


Figure: Comparison of clustering accuracy between SS-KK, SS-CMRF, and SS-NMF with different amounts of constraints on (a) text pairwise (i.e., document-word) co-clustering and (b) gene expression pairwise (i.e., condition-gene) co-clustering

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.

CIKM 2008