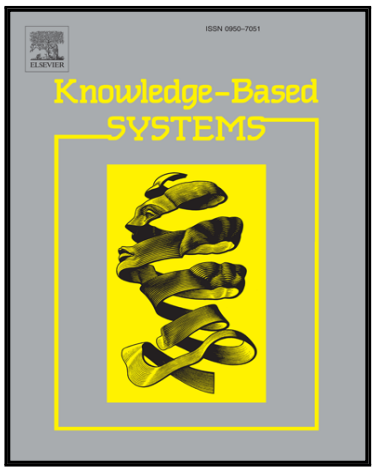

Adaptive Local Structure Learning for Document Co-clustering

Shudong Huang, Zenglin Xu, Jiancheng Lv

PII: S0950-7051(18)30070-4
DOI: 10.1016/j.knosys.2018.02.020
Reference: KNOSYS 4228

To appear in: *Knowledge-Based Systems*

Received date: 20 September 2017
Revised date: 6 February 2018
Accepted date: 9 February 2018

Please cite this article as: Shudong Huang, Zenglin Xu, Jiancheng Lv, Adaptive Local Structure Learning for Document Co-clustering, *Knowledge-Based Systems* (2018), doi: 10.1016/j.knosys.2018.02.020

Adaptive Local Structure Learning for Document Co-clustering

Shudong Huang[a], Zenglin Xu[a,*], Jiancheng Lv[b]

[a] *SMILE Lab, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.*

[b] *College of Computer Science, Sichuan University, Chengdu 610065, China.*

Abstract

The goal of document co-clustering is to partition textual data sets into groups by utilizing the duality between documents (i.e., data points) and words (i.e., features). That is, the documents can be grouped based on their distribution on words, while words can be grouped based on their distribution on documents. However, traditional co-clustering methods are usually sensitive to the input affinity matrix since they partition the data based on the fixed data graph. To address this limitation, in this paper, based on nonnegative matrix tri-factorization, we propose a new framework of co-clustering with adaptive local structure learning. The proposed unified learning framework performs intrinsic structure learning and tri-factorization (i.e., 3-factor factorization) simultaneously. The intrinsic structure is adaptively learned from the results of tri-factorization, and the factors are reformulated to preserve the refined local structures of the textual data. In this way, the local structure learning and factorization can be mutually improved. Furthermore, considering the duality between documents and words, the proposed framework explores not only the adaptive local structure of the data space, but also the adaptive local structure of the feature space. In order to solve the optimization problem of our method, an efficient iterative updating algorithm is proposed with guaranteed convergence. Experiments on benchmark textual data sets demonstrate the effectiveness of the proposed method.

Keywords: Adaptive local structure learning, Graph regularization, Document co-clustering, Nonnegative matrix tri-factorization.

1. Introduction

Clustering is one of the most important unsupervised learning solutions and has been widely applied in text mining, computer vision, biology and so on. From a traditional view point, clustering aims at partitioning a dataset into groups of similar objects [1, 2]. Many clustering methods, such as K-means [3], spectral clustering [4, 5], spectral embedded clustering [6] and Nonnegative Matrix Factorization (NMF) [7, 8, 9], have been proposed up to now.

In recent years, co-clustering has received widespread attention in algorithm development and applications. It overcomes several limitations associated with traditional clustering methods by allowing automatic discovery of similarity based on a subset of attributes. Co-clustering methods have been studied intensively through many different theories and methodologies [10, 11], including co-clustering based on Bayesian models [12, 13, 14, 15, 16, 17], Nonnegative Matrix Tri-Factorization (NMTF) based co-clustering methods [18], spectral co-clustering [19, 20, 21], and so on. In particular, co-clustering methods based on graph theory have abstracted a lot of attentions [22], since intrinsic geometrical structure of data graph have been proved to be useful in a number of machine learning methods [23, 24, 25, 26, 27, 28, 29, 30]. However, data graphs of these methods are usually constructed by considering the K-Nearest Neighbors (KNN) which may mislead the clustering process since the nearest neighbors may belong to different clusters [31, 32]. Furthermore, these methods are sensitive to the input affinity matrix since they partition data based on the fixed data graph. In other words, the similarity measurement and data clustering are often conducted in two separated steps, the learned data graph may not be the optimal one for clustering and lead to the suboptimal results.

To address this issue, we propose a new co-clustering method with adaptive local structure learning based on nonnegative matrix tri-factorization. Instead of performing similarity measurement and data clustering in two separated steps, the proposed model learns the affinity matrix and tri-factorization simultaneously to achieve the optimal clustering results. Meanwhile, both the data graph and the feature graph are constructed by selecting the adaptive and optimal neighbors for each data point and feature respectively. It is based on the assumption that the data points (or features) with a smaller distance should have a larger probability to be the optimal neighbors. We also apply the proposed method to the problem of document clustering using the benchmark textual data sets. The experimental results show that our method has several fa-

*Corresponding author
Email address: zlxu@uestc.edu.cn (Zenglin Xu)