Hybrid Clustering of multi-view data via MLSVD

Xinhai Liu, Lieven De Lathauwer, Wolfgang Glänzel, Bart De Moor

ESAT-SCD
Katholieke Universiteit Leuven

TDA, September, 14, 2010, Bari, Italy

Outline

Outline

Introduction

Motivation

- Booming demand: grouping multi-view data for better partition (Web mining, Social network, Literature analysis).
- Clustering methods
  - Most methods: single-view data
  - Hybrid clustering: multi-view data
- Tensor methods
  - powerful tool to handle multi-way data sources.
  - multi-linear singular value decomposition (MLSVD) (Tucker, 1964 & 1966; De Lathauwer et al, 2000a)

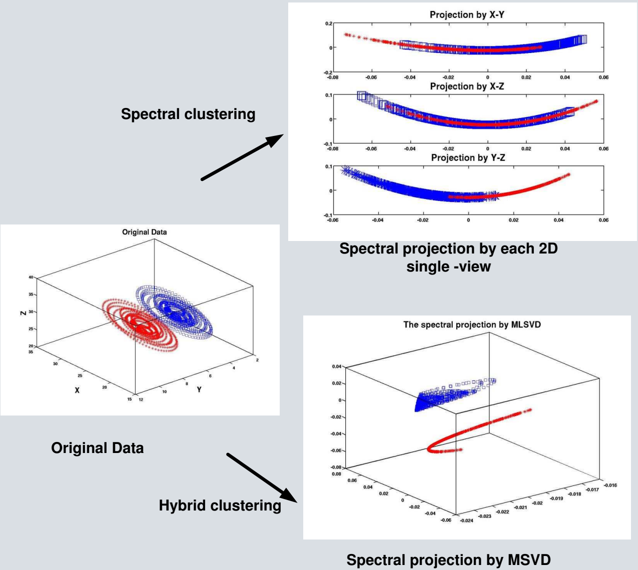


Figure: Demo of a hybrid clustering by MLSVD on synthetic 3D data sets

Introduction

Related work

- Hybrid clustering: multiple kernel fusion (MKF)(Joachims et al, 2001) and clustering ensemble (Strehl & Ghosh, 2002)
- MLSVD based clustering on image recognition (Huang & Ding, 2008)
- Multi-way latent semantic analysis (Sun et al, 2006)
- CANDECOMP/PARAFAC (CP): Scientific publication data with multiple linkage (Dunlavy, Kolda, et al, 2006; Selee, Kolda et al, 2007)

Introduction

Main contributions

- An extendable framework of hybrid clustering based on MLSVD
  - Modelling the multi-view data as a tensor
  - Seeking a joint optimal subspace by tensor analysis
- Two novel clustering algorithms: AHC-MLSVD and WHC-HOOI.
- Experiments on both synthetic data and real Application on Web of Science (WoS) journal database.

Outline

Hybrid clustering

Spectral clustering

Given $S \in \mathbb{R}^{N \times N}$, the affine matrix (similarity matrix) of a graph G; D, the degree matrix; our Laplacian matrix

$$L = D^{-1/2} S D^{-1/2} \tag{1}$$

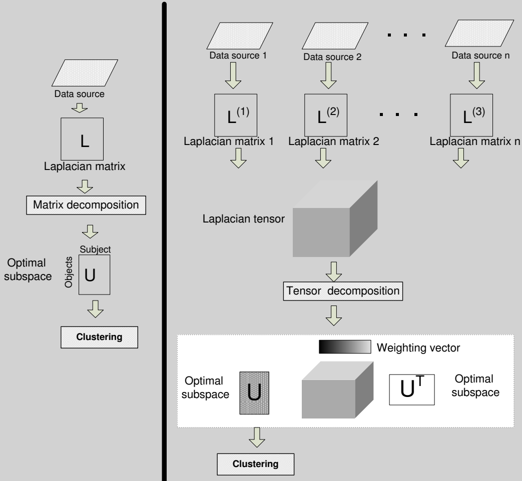
Let an relaxed indicator matrix be $U, \ U \in \mathbb{R}^{N \times M}$, M is the number of clusters

$$\begin{aligned} \max_{U} \ & \mathrm{tr}(U^T L U), \\ \text{s.t. } & U^T U = I. \end{aligned} \tag{2}$$

Eigenvalue decomposition of matrix L: the solution of spectral clustering (Luxburg, 2007)

Hybrid clustering

Concept overview



Spectral Clustering based on SVD

Hybrid clustering based on MLSVD

Hybrid clustering

Laplacian tensor

From a set of K Laplacian matrices $L^{(i)} \in \mathbb{R}^{N \times N}, i = 1, ..., K$ to a Laplacian tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times K}$

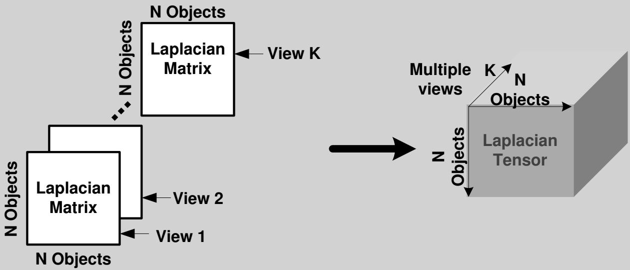


Figure: The formulation of a Laplacian tensor

Hybrid clustering

AHC-MLSVD

Averaging multi-view data for joint analysis:

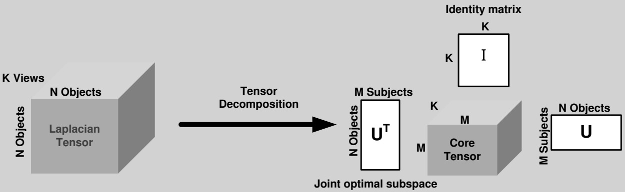


Figure: Average hybrid clustering of multi-view data

$U \in \mathbb{R}^{N \times M}$, the joint optimal subspace
$I \in \mathbb{R}^{K \times K}$, an indentity matrix.

Hybrid clustering

AHC-MLSVD

The optimization of average hybrid clustering,

$$\max_{U} \|\mathcal{A} \times_1 U^T \times_2 U^T \times_3 I\|_F^2, \quad \text{s.t. } U^T U = I. \tag{3}$$

The solution of MLSVD (Tucker, 1964 & 1966; De Lathauwer et al, 2000a)

- An approximate solution
- Usually satisfied results
- An upper bound on the approximation error

Hybrid clustering

WHC-HOOI

Taking the effect of each single-view data into account

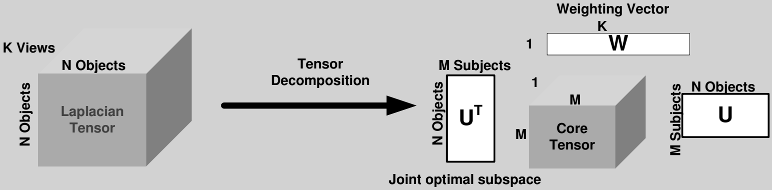


Figure: Weighted hybrid clustering of multi-view data

$W = \{\alpha_1, \alpha_2, \cdots, \alpha_K\}^T$: the weighting factor of each view.

Hybrid clustering

WHC-HOOI

The equivalent optimization of weighted hybrid clustering

$$\max_{U,W} \|\mathcal{A} \times_1 U^T \times_2 U^T \times_3 W^T\|_F^2, \quad \text{s.t. } U^T U = I \text{ and } W^T W = 1. \tag{4}$$

The solution of higher-order orthogonal iteration (HOOI) (Kroonenberg & De Leeuw, 1980; De Lathauwer et al, 2000b)

- An optimal solution
- An appropriate weight for each view data
- Other tensor methods

Outline

Experiments

Clustering of a multiplex network

Multiplex network: a group of networks which share the same nodes but multiple types of links (Mucha et al, 2010)
The synthetic multiplex network:

- Three clusters with each having 50,100, 200 members respectively
- Three views generated by different noise
- Three interaction matrices from each view $\Longrightarrow$ a tensor

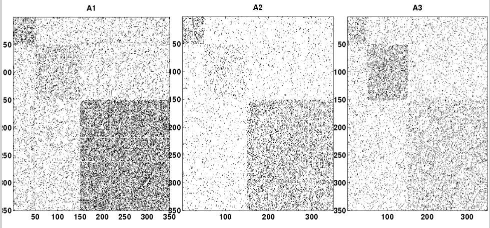
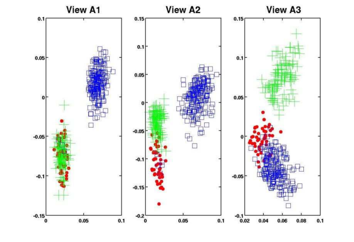


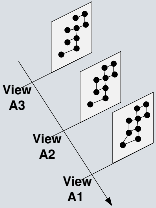
Figure: The adjacent matrices from a synthetic multiplex network

Clustering of a multiplex network

Spectral clustering

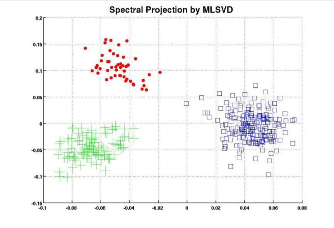


Spectral projection by
each single -view



A Multiplex network

Hybrid clustering



Spectral projection by
MLSVD

Experiments

Application on Web of Science (WoS) journal database

- Objective: Obtain a good scientific mapping from the WoS journals
- Integrating two view data: textual information and journal cross-citations. $N = 8,305$ and $d_{text} = 669,700$
- Cosine similarity matrix of both text and cross-citation

Experiments

Clustering evaluation measures

- Standard categories: Essential Science Indicator (ESI) from WoS
- Normalized mutual information (NMI)

$$NMI = \frac{2 \times H(\{c_i\}), \{l_i\}}{H(\{c_i\})H(\{l_i\})} \tag{5}$$

  where $H(\{c_i\}, \{l_i\})$ is the mutual information between clustering labels $\{c_i\}_{i=1}^{n}$ and reference category indicators $l_{i\,i=1}^{\,n}$, $H(\{c_i\})$ and $H(\{l_i\})$ are their entropies.
- Cognitive analysis by a bibliometrist

Experiments

Clustering performance

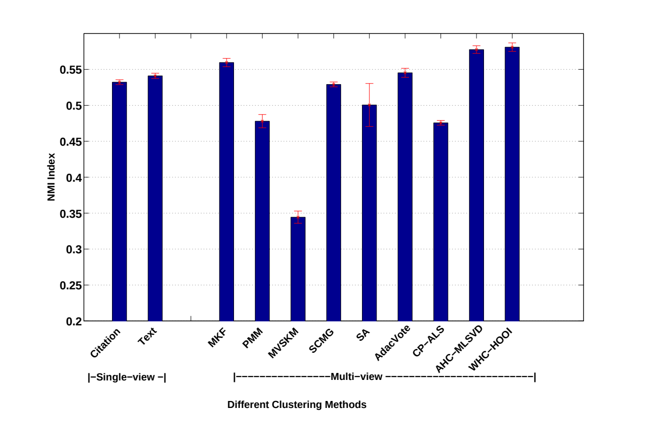


Figure: NMI validation of various clustering methods on WoS journal database (Cluster number:22)

Experiments

Visualization of the journal clusters obtained by HC-MLSVD

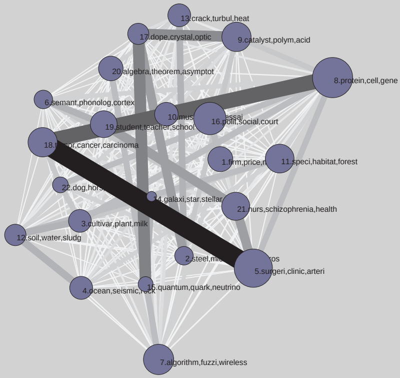


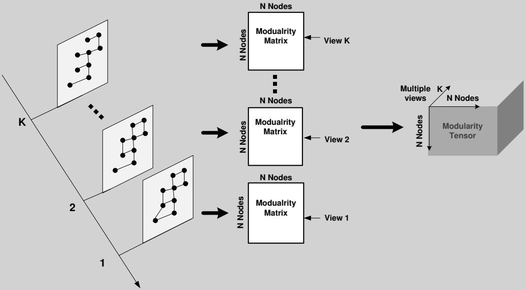
Figure: Visualization of 22 clusters on the WoS journal database (**the node**: the journal clusters where the circle size is proportional to its scale; **the edge**: cross-citation between two journal clusters; **the annotated terms**: the top three text terms within each journal clusters)

Outline

Discussion and outlook

Discussion

- Extendable hybrid clustering framework:
  - Other learning tasks of multi-view data (classification, spectral embedding, collaborative filtering)
  - Other tensor based solutions
  - Other matrices (similarity matices, modularity matrices)



Discussion and outlook

Outlook

- Scalable issue: large-scale database and efficient implementation
- Multiple-model tensor (Currently 3-model): dynamic data analysis
- Other potential tensor methods (CP, INDSCAL,DEDICOM)

Outline

Acknowledgement

Research supported by (1) KUL ESAT SISTA research group; (2) China Scholarship Council (CSC, No. 2006153005); (3) Thanks for discussion with Dr. Carlos Alzate in K.U.Leuven.

Thank you for your attending!