Co-clustering: Model Based or Model Free Approaches

Organizer: Ch. Biernacki [1], Discussant: Ch. Keribin [2]

[1] Université de Lille
INRIA - Lille - Modal

[2] Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay
INRIA - Saclay Ile de France - Celeste

62nd ISI WSC 2019, Kuala Lumpur

Plan

1 Introduction

2 Regularized spectral co-clustering (M. Nadif)

3 Discussion

Clustering

Unsupervised ML method

- data set $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ with $dim(\mathcal{X}) = d$
- to partition into K homogeneous clusters $G_1, \ldots, G_k$

Construct a map f from $D = \{x \in \mathcal{X}^n\}$ to $\{1, \ldots, K\}$ where K is a number of classes : $f : x_i \mapsto z_i$

- classification matrix $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$: $z_{ik} = \mathbb{1}_{i \in G_k}$

Motivations

- data summary
- data exploratory
- preprocessing for more flexibility of a forthcoming prediction step

Many applications

- marketing, assurance, ecology, bioinformatics, social sciences,...
- consumption curves, texts, images, internet logs, graphs ...

Clustering

Challenges

- No ground truth (no given labels) ($\neq$ classification)
- No obvious measure of the quality of a cluster
- Choice of K?

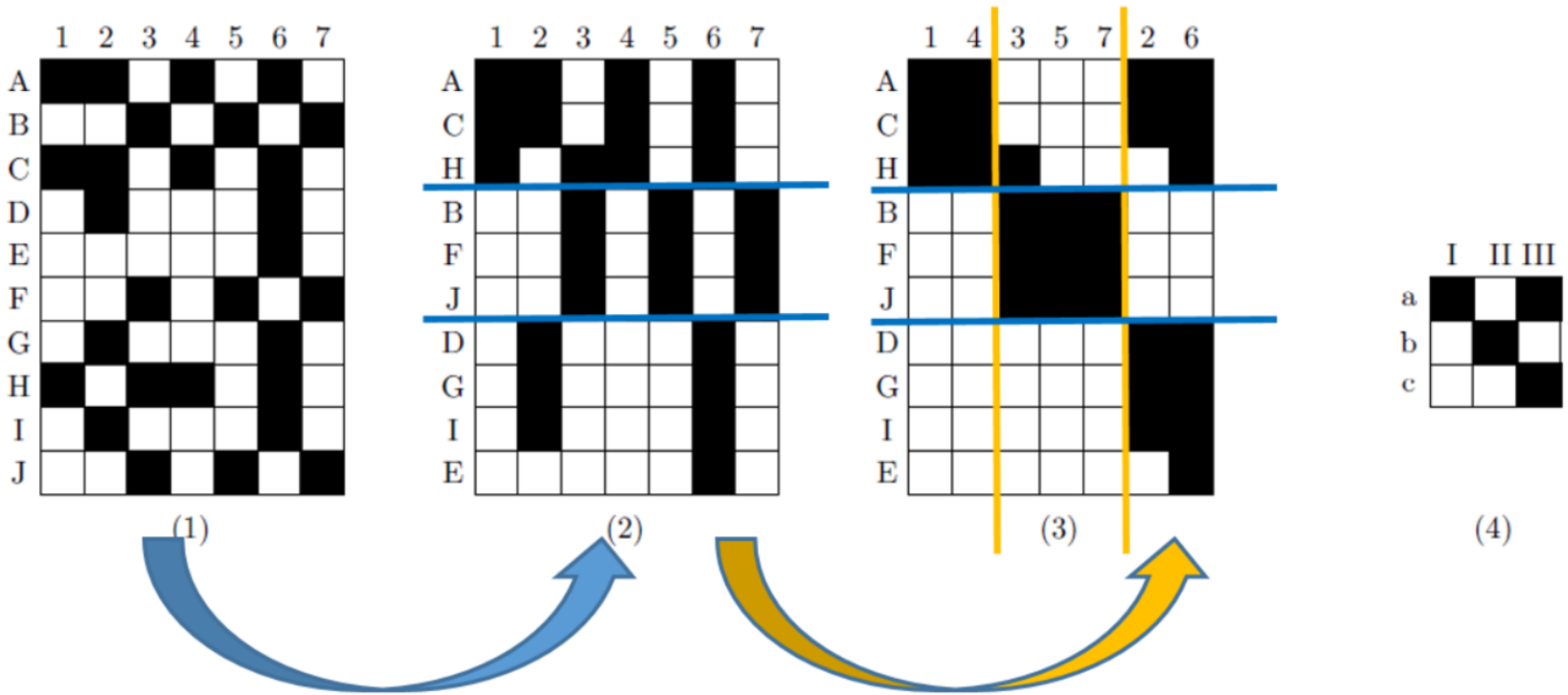
Methods

- Model free
  - $\hookrightarrow$ homogeneity based (k-means), density based (dbscan), agglomerative (hierarchical clustering)
  - $\hookrightarrow$ Non-negative Matrix Factorization (NMF)
  - $\hookrightarrow$ Spectral clustering with Laplacian of graph

  k-means and spectral clustering are able to be expressed as certain canonical forms of NMF [Ding et al. SIAM'05]
- Model based reformulates cluster analysis in a well-posed estimation problem
  - $\hookrightarrow$ Mixture Models

From clustering to co-clustering

simultaneous clustering of data points (observations) and features (variables)



$(n = 10) \times (d = 7)$ $(K = 3) \times (d = 7)$ $(K = 3) \times (L = 3)$

[*Govaert (2011)*]

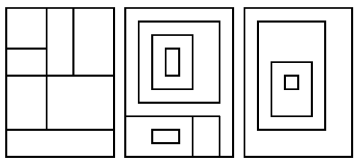
Co-clustering

In scope for this session

- blocks as a cartesian product of a row clustering by a column clustering
- one related case: rows and columns are the same objects (graph clustering)

out of scope (but interesting)

- nested co-clustering



- overlapping co-clustering

1	1	1	1	4	5
1	1	1	1	5	4
1	1	1	1	4	5
1	1	1	1	1	1
4	5	1	1	1	1
4	4	1	1	1	1
5	5	1	1	1	1

1	1	1	1	4	5
1	1	1	1	5	4
1	1	1	1	4	5
1	1	3	3	2	2
4	5	2	2	2	2
4	4	2	2	2	2
5	5	2	2	2	2

Co-clustering

Applications

- recommendation systems (George and Merugu, 2005; Deodhar and Ghosh, 2010; Xu *et al.*, 2012)
- gene expression analysis (Cheng *et al.*, 2008; Hanisch *et al.*, 2002)
- text mining (Dhillon *et al.*, 2003; Wang *et al.*, 2009).
- Netflix challenge ((Bennett and Lanning, 2007)
- ...

Co-clustering

Methods

- Model free, metric based
  - ↪ alternated double k-means (Govaert, 1977 for contingency table)
  - ↪ Non-negative Matrix Tri- Factorization (NMTF) M. Nadif
  - ↪ Spectral co-clustering (Dhillon et al (2001)) C. Laclau
  - ↪ Information theory (Dhillon et al (2003)) C. Laclau

- Model based reformulates cluster analysis in a well-posed estimation problem
  - ↪ Finite Mixture Models: Latent Block Model or Stochastic block model
    How to select a model N. Frial
  - ↪ non-parametric Bayesian model.
    How to cope with heterogenous data? T. Tokuda

How to compare?

Often, only row clustering quality is measured

- clustering accuracy : $ACC = \frac{1}{n} \sum_i \mathbb{1}_{map(z_i)=z_i^0}$
- normalized mutual information (NMI),

$$NMI = \frac{\sum_{k=1}^{K} \sum_{\ell=1}^{K} |G_k \cap \mathcal{L}_\ell| \log \frac{|G_k \cap \mathcal{L}_\ell|}{|G_k||\mathcal{L}_\ell|}}{\sqrt{(\sum_k |G_k| \log \frac{|G_k|}{n})(\sum_\ell \mathcal{L}_\ell \log \frac{\mathcal{L}_\ell}{n})}}$$

- adjusted rand index (ARI)
- purity

Plan

1 Introduction

2 Regularized spectral co-clustering (M. Nadif)

3 Discussion

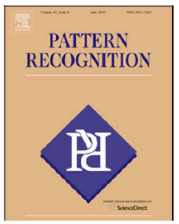
introduction

Pattern Recognition 64 (2017) 386–398

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr



Multi-manifold matrix decomposition for data co-clustering

Kais Allab*, Lazhar Labiod, Mohamed Nadif

LIPADE, University of Paris Descartes, 45, Rue des Saints-Peres, Paris, France

defines a novel regularized spectral co-clustering algorithm by considering the intrinsic geometric structure in the data which is essential for data co-clustering on manifolds.

Basic principle

- Non-negative Matrix Tri-Factorization (NMTF) is popular to deal with co-clustering
- NMTF rests on a global Euclidean geometry only
- a manifold learning technique can be used to map a set of high-dimensional data into a low-dimensional space, while preserving the intrinsic structure of the data.
- Use a set of dimensionality reduction methods and merge the output of the different methods.

Non-negative Matrix Tri-Factorization

Co-clustering as an approximation problem: minimize the approximation error[1]. can also user

$$\min_{S\geq 0, F\geq 0, S\geq 0} ||X - FSG'||^2$$

between the original data $X = (x_{ji}) \in \mathbb{R}^{d\times n}$ and the reconstructed matrix based on column clustering (classification matrix G), row clustering (classification matrix F) and a summary matrix S

[1] Frobenius norm

Non-negative Matrix Tri-Factorization

Equivalent expression

$$1.\left\|X - FSG^T\right\|^2 = \|X - FZ\|^2 + \left\|Z - SG^T\right\|_{D_f}^2$$

$$\text{where } Z := \left\{\mathbf{z}_{qi} = \frac{\sum_j f_{jq} x_{ji}}{\#\mathcal{Q}_q}; q = 1, \ldots, \ell; i = 1, \ldots, n\right\}$$

$$2.\left\|X - FSG^T\right\|^2 = \left\|X - WG^T\right\|^2 + \|W - FS\|_{D_g}^2$$

$$\text{where } W := \left\{\mathbf{w}_{jp} = \frac{\sum_i g_{jp} x_{ji}}{\#\mathcal{P}_p}; p = 1, \ldots, k; j = 1, \ldots, d\right\}$$

with $D_f = F'F$, $D_g = G'G$ and

$$Z = (F^TF)^{-1}F^TX, \tag{4}$$

$$W = XG(G^TG)^{-1}, \tag{5}$$

$$S = (F^TF)^{-1}F^TXG(G^TG)^{-1}. \tag{6}$$

↪ Two NMFs (for Z and W) and an optimization (for S)

Non-negative Matrix Tri-Factorization

Double k-means on intermediate reduced matrices Z and W

Algorithm 1. : *double kmeans* algorithm.

1. Start from an initial position $(G^{(0)}, F^{(0)})$; t=0;
2. Compute $S^{(0)}$ by using Eq.(6);

repeat

(a) - Compute $(Z)^{(t)}$ by using Eq. (4). Then Update $G^{(t+1)}$ by

$$g_{ip}^{(t+1)} = \begin{cases} 1 & p = \arg\min_{p'} \parallel (\mathbf{z}_{.i})^{(t)} - \mathbf{s}_{.p'}^{(t)} \parallel_{D_g}^2 ; \\ 0 & \text{otherwise.} \end{cases}$$

(b) - Compute $(W)^{(t+1)}$ by using Eq. (5). Then Update $F^{(t+1)}$ by

$$f_{jq}^{(t+1)} = \begin{cases} 1 & q = \arg\min_{q'} \parallel (\mathbf{w}_{j.})^{(t+1)} - \mathbf{s}_{q'.}^{(t)}) \parallel_{D_f}^2 ; \\ 0 & \text{otherwise.} \end{cases}$$

(c) - Update $S^{(t+1)}$ by using Eq. (6).

until convergence

Dual graph regularization DRCC (Gu and Zhou, 2009)

Capture non linear low dimension manifolds embedded in high dimensional ambient space

- construct a graph to approximate the manifold in the observation space
- vertices correspond to the data samples,
- the edge weight $a_{ii'}$ represents the affinity of the data points i and i': close points are in the same between cluster.
- A regularization term $trace(GL_gG')$ uses the graph Laplacian L_g

↪ to minimize

$$||X - GSG'||^2 + \lambda\, trace(GL_gG') + \mu\, trace(FL_fF')$$

Multi-manifold matrix decomposition

Use a convex combination of several candidate manifolds (CDA, LLE, ISOMAP, MDS, ...)

$$\min_{G,F,S} \left\|X - FSG^T\right\|^2 - 2\,\alpha\ \mathrm{Tr}\left[G^T\left(\sum_{i=1}^{C}\gamma_g^c B_g^c\right)Q_g\right] + \theta_g\|\gamma_g\|^2$$
$$-2\,\beta\ \mathrm{Tr}\left[F^T\left(\sum_{c=1}^{C}\gamma_f^c B_f^c\right)Q_f\right] + \theta_f\left\|\gamma_f\right\|^2$$

s.t., $Q_g^T Q_g = I$, $Q_f^T Q_f = I$. Tr : denotes the matrix Trace.

Algorithm 3. : M3DC algorithm

Input: - Data matrix X;
- The trade-off parameters α and β;
- C sample candidate manifolds $\{B_g^1, ..., B_g^C\}$;
- C feature candidate manifolds $\{B_f^1, ..., B_f^C\}$;

Output: Partition matrices G and F;
Initialize: - G and F using a clustering algorithm; S by using (12);
repeat
(a) - Compute $Q_g^{(t)}$ and $Q_f^{(t)}$;
(b) - Compute $\gamma_g^{(t)}$ and $\gamma_f^{(t)}$ using the EMDA;
(c) - Update $G^{(t+1)}$ by (15);
(d) - Update $F^{(t+1)}$ by (16);
(e) - Update $S^{(t+1)}$ by (12);.
until convergence

Plan

1 Introduction

2 Regularized spectral co-clustering (M. Nadif)

3 Discussion

Regularized spectral co-clustering

Main features

- Uses intermediate reduced matrices Z and W and deduces a hard clustering insuring F and G orthogonality
- Nice idea to combine several spectral parsimonious representations to capture potential manifolds
  Assesses the contribution of each manifold in an unsupervised context
- only focuses on the quality of the row clustering
- better performances on single manifolds than DRCC

Regularized spectral co-clustering

Discussion

- convergence is only illustrated on numerical experiments ?
- not clear how to choose the dimension of the reduced representations
- the number of sample clusters $=$ true number $=$ number of feature clusters
- regularization parameters are taken equal

Optimal transport and rank one CC

Main features: new point of view

- Establishes links between entropy-regularized optimal transport and rank-one matrix factorization
- obtains CC partitions with an automatic detection of the number of rows and columns clusters.

Optimal transport and rank one CC

Optimal transport: find a doubly stochastic matrix γ, coupling two one-dimensional distributions, with respect to a cost matrix *M*

$$\min_{\gamma \in \Pi(\alpha,\beta)} \langle \mathbf{M}, \gamma \rangle_F - \lambda E(\gamma),$$

$$\Pi(\alpha, \beta) := \left\{ \gamma \in \mathbb{R}_+^{m \times n} : \gamma \mathbf{1}_n = \mathbf{a}, \gamma^\top \mathbf{1}_m = \mathbf{b} \right\},$$

A unique solution $\gamma_\lambda^*(a, b)$ where *u* and *v* are two non-negative scaling vectors uniquely defined up to a multiplicative factor:

$$\mathrm{diag}(\mathbf{u}) \mathrm{e}^{-\mathbf{M}/\lambda} \mathrm{diag}(\mathbf{v})$$

Application to CC Define rows and columns empirical measures

$$\boldsymbol{\kappa_r} := \sum_{i=1}^{m} \delta_{\mathbf{x}_i}/m \quad \text{and} \quad \boldsymbol{\kappa_c} := \sum_{i=1}^{n} \delta_{\mathbf{y}_i}/n.$$

[pb sur les indices?]

$\hookrightarrow$ the scaling vectors *u* and *v* can be seen as estimated rows and columns marginal distributions

Optimal transport and rank one CC

Discussion Je n'ai pas compris la fin de l'article.
Je n'ai pas vu où on donnait le nombre de classes en lignes et colonnes pour le CC
Je n'ai pas vu dans le transport optimal où on trouvait la classification

- It's important to study links between methods
  interesting for the scientific community(ies), opens new visions
  provide fundamental understanding of mechanisms behind unsupervised learning in general

A new way to perform model choice for CC

- Model Based: Latent Block Model as a generalized mixture

$$p(\mathbf{x};\theta) = \sum_{(z,w)\in\mathcal{Z}\times\mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}$$

  $\hookrightarrow$ theory for model selection

- The frequentist approach:
  - $\hookrightarrow$ Maximum likelihood estimation (VEM, VBayes, Gibbs) of the parameter $\theta = (\pi, \rho, \alpha)$
  - $\hookrightarrow$ allocation to the row $\widehat{\mathbf{z}}$ and column $\widehat{\mathbf{w}}$ clusters using a MAP rule
  - $\hookrightarrow$ selection criteria (K. et al 2015): $BIC(\widehat{\theta}, K, G)$ (double approximation) or exact $ICL(\widehat{\mathbf{z}}, \widehat{\mathbf{w}}, K, G)$
  - $\hookrightarrow$ a greedy algorithm (K, G) to explore the models (V. Robert 2017)
- The Bayesian approach: MCMC costly computations

A new way to perform model choice for CC

Nial's proposal

- directly optimize the exact $ICL(\mathbf{z}, \mathbf{w}, K, G)$
- run through the models from large K_{max}, G_{max}:
  - $\hookrightarrow$ A greedy search algorithm iteratively allocates members and clubs and merges existing clusters so as to maximize the ICL.
  - $\hookrightarrow$ sparse form if needed
  - $\hookrightarrow$ $\widehat{\theta}$ can be estimated at the end
- very efficient compared to MCMC

A new way to perform model choice for CC

Discussion

- How to choose the pruning threshold?
- Don't we loose the posterior distribution of the parameter?
- performance for (not sparse) large data?
- Is the solution not too "adjusted"?
- Link with the frequentist approach?

Multiple CC with heterogeneous marginal distributions

A nice extension to define automatically different views that characterize multiple co-clustering structures

- each variable belongs to only one view

↪ divide the variable set and is adapted to high dimensional data ($n < p$)

- individuals are common to all views, but row clusters are different from one view to another

↪ breaks the chess board structure

- variables can be of mixed types

Multiple CC with heterogeneous marginal distributions

Discussion

- variables of different type cannot be in the same cluster. How this information is taken into account?
- VBayes is sensitive to initialization. How does it affect the estimation?
- Does-it allow to recover some cluster of non-informative features?
- Dirichlet process is used to model an infinite numbers for views and clusters. Could it be also used for the greedy algorithm of Nial?
- for the case study: quel était le but ? est-ce du supervisé ou du non supervisé?

The last touch

In several talks

- use co-clustering to perform clustering with HD data
- CC is biased, but extremely regularized
- could be able to cluster redundant variable, non informative variables

Thank you for your attention !