Pattern Recognition 72 (2017) 108–122

Contents lists available at ScienceDirect

Pattern Recognition

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

Model-based co-clustering for the effective handling of sparse data

Melissa Ailem*, François Role, Mohamed Nadif

LIPADE, Paris Descartes University, Sorbonne Paris Cité, Paris F-75006, France

ARTICLE INFO

Article history:
Received 10 December 2016
Revised 26 May 2017
Accepted 1 June 2017
Available online 3 July 2017

Keywords:
Mixture models
Poisson distribution
Latent block model
Co-clustering
Variational EM
Text data
Sparse data

ABSTRACT

With the exponential growth of text documents on the web, there is a genuine need for techniques that organize terms and documents, simultaneously, into meaningful clusters, thereby making large datasets easier to handle and interpret. Several block diagonal clustering algorithms have proven successful in identifying co-clusters of documents and words. However, despite their effectiveness, most of the existing methods do not provide a parameterizable model for tackling the problem of block diagonal identification. In this paper, we rely on mixture models, which offer strong theoretical foundations and considerable flexibility. More precisely, we propose a parsimonious latent block model based on the mixture of Poisson distributions and tailored for sparse high dimensional data such as document-term matrices. In order to efficiently estimate the model parameters, we derive two scalable co-clustering algorithms based on variational inference. Empirical results obtained on several real-world text datasets highlight the advantages of the proposed model and the corresponding co-clustering algorithms.

1. Introduction

Nowadays, the co-clustering techniques are undoubtedly a considerable asset in numerous application domains such as bioinformatics, text mining and collaborative filtering. Given a data matrix $\mathbf{x}$ of size $n \times d$ where I is the set of n rows, and J the set of d columns, a co-cluster $k\ell$ is a sub-matrix $I_k \times J_\ell$ ($I_k \subseteq I$, $J_\ell \subseteq J$) where rows and columns follow some consistent patterns (see Fig. 1). Since the seminal work of Hartigan [23], a variety of co-clustering methods have been proposed and applied in different areas; see, e.g., [6,9,11,15,16,28,29,32,35,39].

In this paper, we are concerned with text co-clustering, a field where the volume of available information is exponentially growing and data synthesizing techniques are necessary. In this context, the data is often represented by a sparse, high-dimensional document-term matrix. When dealing with such matrices, the classical co-clustering algorithms (see Fig. 1(b)) which seek homogeneous blocks might be deceived by data sparsity. In fact, due to this sparsity, several co-clusters are primarily composed of zeros; such co-clusters are homogeneous but not meaningful. Therefore, a post-processing step is necessary to filter out these blocks and keep only the relevant ones. An effective way to solve this problem is to use block-diagonal algorithms. These algorithms seek an optimal block diagonal clustering, meaning that objects (documents) and features (terms) have the same number of clusters and that, after a suitable permutation of the rows and columns, the algorithm reveals a block diagonal structure (see Fig. 1(c)). Examples of such algorithms include those described in [1,13,27]. Dhillon [13] finds co-clusters by minimizing the cut objective function in a bipartite *document* × *term* graph. Labiod and Nadif [27] proposed a co-clustering algorithm which tries to maximize bipartite modularity using a spectral approach. In a previous work [1], we proposed an alternative, more direct way of maximizing graph modularity for co-clustering, using an iterative alternating optimization procedure.

However, the above-mentioned block-diagonal graph-based algorithms have some limitations. The main one is that, they do not provide a parameterizable model for tackling the problem of block diagonal identification. Moreover, these approaches do not present a generative process of data and do not take into account some data properties such as cluster proportions.

Herein, we consider a radically different approach to discover a block diagonal structure. We rely on the Latent Block Model (LBM) [17–20] which is undoubtedly a useful contribution to co-clustering; see, e.g., [26,34,37,43]. In addition to its parsimony, LBM offers strong theoretical foundations, since it defines its own generative process of data. Besides, LBM is flexible, leading to various models which are able to take into account the nature[1] of the data and a variety of challenging situations, such as het-

* Corresponding author.
E-mail address: melissa.ailem@parisdescartes.fr (M. Ailem).

[1] By nature we mean the type of data (binary data, contingency tables, continuous data, categorical data ...) and some characteristics such as the proportions of clusters.

http://dx.doi.org/10.1016/j.patcog.2017.06.005
0031-3203/

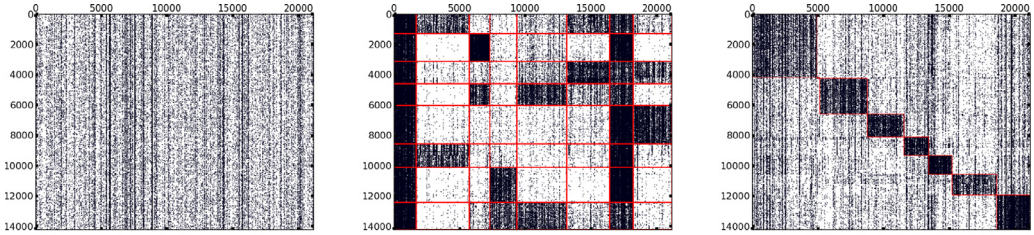

Fig. 1. (a) Original data. (b) General partitional co-clustering. (c) Diagonal co-clustering.

erogeneous populations and unbalanced cluster sizes. Moreover, the associated estimators of posterior probabilities give rise to a soft or hard clustering using the maximum a posterior principle (MAP), which is an advantage compared to classical methods inferring a hard partition, only. To deal with the co-clustering of *document* × *term* matrices, we use the Poisson LBM [21,22] which is adequate for count data such as document-term matrices. Although this model provides a setting for treating this type of data, it does not directly take into account the above-described sparsity-related issues. Hence, we present Sparse Poisson Latent Block Model (SPLBM) which assumes that each block or co-cluster is generated according to the Poisson distribution with some specific parameters. As opposed to PLBM, SPLBM has the advantages of being more parsimonious and able to reveal the most meaningful co-clusters. Thereby, SPLBM is more suitable than PLBM when dealing with high dimensionality and sparsity. The SPLBM model was first introduced in our previous work [2]. In the latter, we relied on the *classification maximum likelihood* approach [31,40,41], i.e., the maximization of the complete-data likelihood, for inference and parameter estimation. Although the derived algorithm has proven effective on several real world data sets, it is quite restrictive as it yields hard clustering, of both documents and words, at each stage, i.e., each word/document is assigned to a single cluster at each step. Indeed, in real-world data sets, a document (resp. word) typically may belong to multiple clusters with different degrees. Therefore, soft co-clustering algorithms, where each document (resp., word) has a probability of being generated from all clusters, are expected to be more appropriate.

In this paper, we rely on the *mixture maximum likelihood* approach [31] and derive a variational mean-field EM procedure to fit the parameters of our model, which yields a soft and stochastic SPLBM-based co-clustering algorithms. Furthermore, we show that by combining the aforementioned algorithms we can leverage the benefits of both the soft and stochastic variants, simultaneously. The experimental section is divided into two parts, the first part is devoted to document clustering evaluation and the second part to term clusters assessment. As will be clearly shown by results obtained from experiments on real-world text data, the proposed algorithms allow to get high quality document clustering results compared to some clustering and co-clustering algorithms devoted to this task. Furthermore, we observe that the obtained term clusters are often meaningful and semantically coherent vis-à-vis document clusters.

The rest of the paper is organized as follows. After reviewing the Poisson Latent Block Model in Section 2, we introduce SPLBM in Section 3. We show how co-clustering algorithms can be derived from this model taking into account a diagonal structure of data. Section 4 is devoted to comparative numerical experiments that demonstrate the effectiveness of the proposed algorithm on several real-world datasets. In Section 5, we conclude and suggest paths for future research.

Notation. Herein we present the notations that we will use all along this paper.

- Data will be denoted by an $n \times d$ matrix $\mathbf{x} = (x_{ij}, i \in I = \{1, \ldots, n\}; j \in J = \{1, \ldots, d\})$. We note $N = \sum_{i,j} x_{ij}$, $x_{i.} = \sum_j x_{ij}$ and $x_{.j} = \sum_i x_{ij}$.
- A partition of I into g clusters will be represented by $(z_1, \ldots, z_n)$ where $z_i \in \{1, \ldots, g\}$ and by the classification matrix $\mathbf{z} = (z_{ik}, i = 1, \ldots, n, k = 1, \ldots, g)$ where $z_{ik} = 1$ if element i arose from cluster k and $z_{ik} = 0$ otherwise. Similarly, a partition of J into m clusters will be represented by $(w_1, \ldots, w_d)$ where $w_j \in \{1, \ldots, m\}$ and by the classification matrix $\mathbf{w} = (w_{j\ell}, j = 1, \ldots, d, \ell = 1, \ldots, m)$ where $w_{j\ell} = 1$ if element j arose from cluster ℓ and $w_{j\ell} = 0$ otherwise.
- In the same way, soft partitions of I and J will be represented, respectively, by a matrix $\widetilde{\mathbf{z}}$ of elements in [0, 1] satisfying $\sum_k \widetilde{z}_{ik} = 1$ for all $i = 1, \ldots, n$ and by matrix $\widetilde{\mathbf{w}}$ of elements in [0, 1] satisfying $\sum_\ell \widetilde{w}_{j\ell} = 1$ for all $j = 1, \ldots, d$.
- The sums Σ_i, Σ_j, Σ_k and Σ_ℓ stands, respectively, for $\sum_{i=1}^{n}$, $\sum_{j=1}^{d}$, $\sum_{k=1}^{g}$, and $\sum_{\ell=1}^{m}$.

2. Poisson Latent Block Model (PLBM)

The PLBM model was introduced in [21] for the co-clustering of contingency tables. PLBM is based on the Latent Block Model (LBM), a model previously proposed by the same authors [17]. Given a data matrix $\mathbf{x}$ of size $n \times d$, where I and J represent, respectively, the set of rows and columns, LBM assumes that there is a partition $\mathbf{z}$ on I and a partition $\mathbf{w}$ on J in such a way that each element x_{ij} of matrix $\mathbf{x}$ is generated according to a probability density function (pdf) with some specific parameters. This model is based on the following assumptions:

- The univariate random variables x_{ij} are assumed to be conditionally independent given the row and column partitions $\mathbf{z}$ and $\mathbf{w}$, respectively.
- Independent latent variables: the labelings $z_1, \ldots, z_n, w_1, \ldots, w_d$ are considered as latent variables and assumed to be independent $p(\mathbf{z}, \mathbf{w}) = p(\mathbf{z})p(\mathbf{w})$ where $p(\mathbf{z}) = \prod_i p(z_i) = \prod_{i,k} \pi_k^{z_{ik}}$; $p(\mathbf{w}) = \prod_j p(w_j) = \prod_{j,\ell} \rho_\ell^{w_{j\ell}}$.
- For all i, the distribution of $p(z_i)$ is the multinomial distribution $\mathcal{M}(\pi_1, \ldots, \pi_g)$ and does not depend on i. Similarly, for all j, the distribution of $p(w_j)$ is the multinomial distribution $\mathcal{M}(\rho_1, \ldots, \rho_m)$ and does not depend on j.

The set of parameters of the latent block model is $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ and $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$ where $(\pi_k = P(z_{ik} = 1), k = 1, \ldots, g)$, $(\rho_\ell = P(w_{j\ell} = 1), \ell = 1, \ldots, m)$ are the mixing proportions, and $\boldsymbol{\alpha} = (\alpha_{k\ell}; k = 1, \ldots, g, \ell = 1, \ldots, m)$ where $\alpha_{k\ell}$ is the parameter of the distribution characterizing block $I_k \times J_\ell$. Let $\mathcal{Z}$ and $\mathcal{W}$ be the sets of all possible labels $\mathbf{z}$ for I and $\mathbf{w}$ for J, the pdf $f(\mathbf{x}; \boldsymbol{\theta})$ of $\mathbf{x}$ can be written

$$\sum_{(z,w) \in \mathcal{Z} \times \mathcal{W}} \prod_{i,k} \pi_k^{z_{ik}} \prod_{j,\ell} \rho_\ell^{w_{j\ell}} \times \prod_{i,j} f(x_{ij}; \alpha_{z_i w_j}), \tag{1}$$