Computational Statistics and Data Analysis 123 (2018) 32–49

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Clustering sparse binary data with hierarchical Bayesian Bernoulli mixture model

Mao Ye, Peng Zhang *, Lizhen Nie

School of Mathematical Science, Zhejiang University, 38 Zheda Road, Hangzhou, Zhejiang, 310021, China

ARTICLE INFO

Article history:
Received 19 December 2016
Received in revised form 31 October 2017
Accepted 6 January 2018
Available online 13 February 2018

Keywords:
Bayes factor
Categorical data
Defining features
Model selection

ABSTRACT

Sparsity in features presents a big technical challenge to existing clustering methods for categorical data. Hierarchical Bayesian Bernoulli mixture model (HBBMM) incorporates constrained empirical Bayes priors for model parameters, so the resulting Expectation Maximization (EM) algorithm of estimator searching is confined in a proper region. The EM algorithm enables to obtain the maximum a posterior (MAP) estimation, in which cluster labels are simultaneously assigned. Three criteria are proposed to identify defining features of individual clusters, leading to understanding of the underlying data structures. Information based model selection criterion is applied to determine the number of clusters. Estimation consistency and performance of model selection criteria are investigated. Two real-world sparse categorical datasets are analyzed with the proposed method.

1. Introduction

Cluster analysis plays an important role in exploring data structure in many applied fields. Many existing algorithms focus on clustering continuous or numeric data, including K-means (Hochbaum and Shmoys, 1985), metrics-based hierarchical clustering schemes (Johnson, 1967; Murtagh and Contreras, 2012) and spectral clustering method such as Ng et al. (2002), Krzakala et al. (2013) and Lei et al. (2015). However, these methods cannot be directly applied to categorical data whose domain values are discrete with no defined ordering. As a result, some clustering algorithms have been proposed to deal with categorical data. K-modes algorithm is a modified version of K-means for clustering categorical data with two basic extensions proposed separately by Chaturvedi et al. (2001) and Huang (1998). In hierarchical clustering schemes, new metric spaces are proposed for categorical data (Zhang et al., 2006; Iam-On et al., 2012). Labiod and Nadif (2011) and David and Averbuch (2012) develop spectral-based methods for clustering categorical data. Population based methods take a certain distribution assumption for the data. One of popular population based methods is finite mixture model (McLachlan and Peel, 2004), which represents heterogeneity in a finite number of latent classes (Loh et al., 2015; Moser et al., 2015). One extension of the mixture model is to introduce a prior distribution to the mixing proportions (Rasmussen, 1999; Medvedovic and Sivaganesan, 2002; Yerebakan et al., 2014; Miller and Harrison, 2017). EM algorithm, Monte Carlo Markov Chain (MCMC) and Variational Bayes (VB) are widely applied to estimate the parameters of mixture models (White and Murphy, 2014a). Cheeseman et al. (1993) propose the so-called AutoClass algorithm that relaxes the distributional assumption of the mixture model. Many researchers (Pizzuti and Talia, 2003; McLachlan and Krishnan, 2007; Achcar et al., 2009) extend this method to various applications. See Aggarwal and Reddy (2013) for an overview of various clustering techniques.

Recently, many big data projects generate volumions of sparse datasets in that identification of sub-populations is of primary interest. Clustering methods are proposed to deal with the sparsity in data clustering. For example, Dhillon and

* Corresponding author.
E-mail address: pengz@zju.edu.cn (P. Zhang).

https://doi.org/10.1016/j.csda.2018.01.020
0167-9473/

Modha (2001) propose a spherical K-means algorithm to cluster high-dimensional and sparse text documents; Elhamifar and Vidal (2009), Elhamifar and Vidal (2011) and Wang and Xu (2016) propose a subspace clustering algorithm for geometrical sparse embedded data. Chen et al. (2012) propose a clustering algorithm for sparse unweighted graphs based on intra-clusters density and inter-clusters density. Krzakala et al. (2013) present a spectral algorithm to track a nonbacktracking walk on directed edges of a graph to cluster sparse networks. Azizyan et al. (2015) apply a mixture of two non-spherical Gaussian distributions to cluster high-dimensional continuous data. Zhang and Lu (2016) propose a two-step optimization strategy to cluster large-scale sparse continuous data. Arias-Castro and Pu (2017) focus on clustering of sparse continuous data with many noise features. Floriello and Vitelli (2017) consider the development of clustering algorithm for functional data while jointly selecting most relevant features. Dimensionality reduction methods, such as ZIFA (Pierson and Yau, 2015) and t-SNE (van der Maaten and Hinton, 2008), cluster text and gene sequencing data by embedding the data matrix (Macosko et al., 2015; Grün et al., 2015).

Most of the above-mentioned methods are developed for sparse but relatively informative data such as sparse continuous data, sparse gene data, sparse text data, geometrical sparse embedded data, sparse graphic data and sparse functional data. However, being pervasive in practice, sparse categorical data with binary features challenges these existing methods, since information sparsity and lack of numerical granularity violate some of the assumptions used by these algorithms. This motivates us to develop a new clustering algorithm for sparse categorical data to address the technical needs of performing cluster analysis of, for example, the following two real-world datasets.

The first motivating dataset, CNAE-9 (Ciarelli and Oliveira, 2009), downloaded from the UCI Repository, contains 1080 documents of free text business descriptions of Brazilian companies, which are divided into 9 categories by experts. The documents have gone through pre-processing after which each document is represented by a vector of word frequencies. Since the number of words is very large whereas the frequency of each word's appearance in each document is small, features in the dataset are highly sparse. Actually, over 99% entries of the data matrix are filled with zeros, which severely challenges any existing clustering algorithms.

The second motivating example pertains to a search engine marketing (SEM) data analysis. SEM, by its definition, is a form of internet marketing practice, through which companies promote their products to potential customers based on keywords people search with search engines. The logic behind this marketing strategy is that users' search behaviors could more or less reveal their interest in a particular product. This strategy has been widely used by many search engine companies. However, one challenge in SEM is that both the number of ads-seekers and the total number of keywords searches are huge, while frequencies of keyword hits are extremely low, making this dataset exhibits ultra sparseness. Having no specific techniques to handle such excessively zero-inflated features, many clustering methods lose analytic power in SEM data analysis.

To develop a clustering algorithm that overcomes the aforementioned technical difficulties, we propose a hierarchical Bayesian Bernoulli mixture model (HBBMM). Due to the fact that attributes with extremely low frequencies of appearance are routinely eliminated in a pre-processing step, we attempt to introduce certain prior information to account for such attributes with limited effects. This idea may be formulated by a constrained empirical Bayesian estimation in the context of EM algorithm. The use of constraints allows us to cluster sparse binary data efficiently. A critical issue in the cluster analysis is how to choose the number of clusters, or equivalently, the number of mixture components. We explore some model selection criteria and study their asymptotic properties. Identification of important attributes specific to individual clusters is important in clustering applications. Thus, we propose using Bayes factor to detect defining features of individual clusters.

The rest of this paper is organized as follows. We present the HBBMM and associated assumptions in Section 2. In Section 3, we propose a Bayes factor and three model selection criteria to identify defining features of individual clusters. Section 4 discusses asymptotic properties of the maximum a posterior (MAP) estimator and the selection consistency of Bayesian Information Criteria (BIC) and Hannan Quinn information criterion (HQC). Section 5 illustrates extensive simulations to evaluate the performance of the proposed clustering method in four aspects: the accuracy of parameter estimation, the accuracy of clustering, the performance of model selection and the performance of the Bayes factor in identifying defining features. Two real datasets are analyzed with the proposed algorithm in Section 6. Conclusions and discussions are included in Section 7.

2. Clustering algorithm

2.1. Assumption and notation

Let 1 indicate hit or presence of a binary attribute and 0 otherwise. Let $\mathbf{X}_i,\ i = 1, \ldots, n$ be d-dimensional random vectors identically and independently drawn from a population with K clusters of the following Bernoulli mixture distribution:

$$p(\mathbf{X}_i \mid \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k P(\mathbf{X}_i \mid \boldsymbol{\mu}_k),$$

where $\mathbf{X}_i = (X_{i1}, \ldots, X_{id}) \in \mathbb{R}^d,\ i = 1, \ldots, n$, $(\mu_{k1}, \ldots, \mu_{kd})^T = \boldsymbol{\mu}_k \in \mathbb{R}^d$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_K^T)^T$, $P(\mathbf{X}_i \mid \boldsymbol{\mu}_k) = \prod_{\ell=1}^{d} \mu_{k\ell}^{X_{i\ell}} (1 - \mu_{k\ell})^{1 - X_{i\ell}}$, π_k are the mixture proportions, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$ and $\sum_{k=1}^{K} \pi_k = 1$. Denote $\mathbf{X} = (\mathbf{X}_1^T, \ldots, \mathbf{X}_n^T)^T$. For the model specification, we assume the following prior distribution for the parameters of the Bernoulli distributions in each component. For $k = 1, 2, \ldots, K$ and $\ell = 1, 2, \ldots, d$, $\mu_{k\ell}$ are independent $Beta(\alpha, \beta)$ variables. Since we usually do not