Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Clustering of multivariate binary data with dimension reduction via L_1-regularized likelihood maximization

Michio Yamamoto [a,*], Kenichi Hayashi [b,1]

[a] Department of Biomedical Statistics and Bioinformatics, Kyoto University Graduate School of Medicine, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan

[b] Osaka University Graduate School of Medicine, 2-2 Yamadaoka, Suita, Osaka 565-0871, Japan

ARTICLE INFO

Article history:
Received 25 June 2014
Received in revised form
27 May 2015
Accepted 29 May 2015

Keywords:
Binary data
Clustering
Dimension reduction
EM algorithm
Latent class analysis
Sparsity

ABSTRACT

Clustering methods with dimension reduction have been receiving considerable wide interest in statistics lately and a lot of methods to simultaneously perform clustering and dimension reduction have been proposed. This work presents a novel procedure for simultaneously determining the optimal cluster structure for multivariate binary data and the subspace to represent that cluster structure. The method is based on a finite mixture model of multivariate Bernoulli distributions, and each component is assumed to have a low-dimensional representation of the cluster structure. This method can be considered as an extension of the traditional latent class analysis. Sparsity is introduced to the loading values, which produces the low-dimensional subspace, for enhanced interpretability and more stable extraction of the subspace. An EM-based algorithm is developed to efficiently solve the proposed optimization problem. We demonstrate the effectiveness of the proposed method by applying it to a simulation study and real datasets.

1. Introduction

Binary data are commonly observed and analyzed in many application fields: behavioral and social research, biosciences, document classification, and inference on binary images. For example, Ekholm et al. [1] analyzed biomedical data including five unequally spaced binary self-assessment measurements of arthritis and obesity data on the presence or absence of obesity in five cohorts of children. Also, the binarized data of the MovieLens 100K and the Netflix dataset, which are popular datasets for collaborative filtering tasks, have been analyzed by Kozma et al. [2]. One of the purposes of analyzing binary data, as well as continuous data, is the partitioning of objects which have binary features into several unpredetermined homogeneous groups (clusters). For clustering of objects with many variables, it is quite important to know if some of the variables do not contribute much to the structure of clusters because the inclusion of redundant information can reduce the performance of the cluster analysis [3]. Also, a lower-dimensional (say two or three dimensional) representation of the cluster structure, based on the most significant information, is very useful for evaluating and interpreting the results of the cluster analysis [4].

Hence, what is needed is a procedure that constructs a low-dimensional representation of the multivariate binary data, such that the cluster structure in the data is maximally revealed. For this purpose, researchers often carry out a preliminary dimension reduction technique (e.g., [5–10]). Among the references, [5,6] developed principal component analysis (PCA) models for binary data, while the other references have developed more general PCA models to handle exponential family data. Cluster analysis is then performed on the object scores on the first few principal components. Although it is easy to implement, this two-step sequential approach, also called the tandem approach, provides no assurance that the components extracted in the first step are optimal for the subsequent cluster analysis, because the two steps are implemented separately by optimizing a different loss function [4,11–15]. For multivariate continuous data, instead of the two-step tandem clustering procedure, several methods that simultaneously perform cluster analysis and dimension reduction have been proposed [4,13,15–17].

On the other hand, for multivariate binary data, a few methods can conduct the analysis for simultaneously obtaining a cluster structure and a subspace for the cluster structure. Patrikainen and Mannila [18] have developed a subspace clustering method of binary data that can be used in high-dimensional settings. Cagnone and Viroli [19] have proposed a factor mixture analysis model for multivariate binary data, in which latent variables are distributed as a finite mixture of multivariate Gaussian distributions.

In general, there are two types of clustering techniques with finding subspaces: one intends to find a subspace that is common

* Corresponding author. Tel.: +81 75 751 4755; fax: +81 75 751 4732.
E-mail addresses: michyama@kuhp.kyoto-u.ac.jp (M. Yamamoto), kenichi@medstat.med.osaka-u.ac.jp (K. Hayashi).
[1] Tel.: +81 6 6879 3597; Fax: +81 6 6879 3598.

http://dx.doi.org/10.1016/j.patcog.2015.05.026
0031-3203/

to all clusters [14], while the other aims to find a subspace specific to each group [20]. These two techniques can be used for different purposes. The former has a strong point in helping researchers to understand the configuration of objects and cluster centers in a single low-dimensional space. We need this technique if we want to analyze the data at hand using a component-based approach like the ordinary factor analysis and principal component analysis. The illustration shown in [4] is useful for understanding how to analyze the data using the common subspace clustering. On the other hand, the latter approach is needed for analyzing the data based on the assumption that the data points could be drawn from multiple subspaces. For example, a video sequence could contain several moving objects, and different subspaces might be needed to describe the motion of different objects in the scene [20]. In this paper, we focus on the common subspace clustering.

In the related works to the subspace clustering, there are several works on the problem of multi-task learning in which multiple tasks share a low-dimensional subspace. In the multi-task problem, parameters to be estimated are assumed to share some common structure in the tasks. For example, parameters are divided into two parts: one is common to all tasks and another is specific to each task [21]. Also, Argyriou et al. [22] assume that tasks' structure is summarized by a positive definite matrix which is linked to the covariance matrix between the tasks. For supervised learning, [23] uses the formulation in which tasks share a linear low-dimensional subspace, and [24] proposes an optimization problem regularized by the projection distance of task-related parameters from the manifold shared by all tasks. In addition, for semi-supervised learning, there are some works that formulate the subspace shared by multiple tasks [25,26].

As described above, Patrikainen and Mannila's [18] method allows for obtaining a cluster structure and a subspace for the cluster structure simultaneously. However, their method is rather cluster-specific subspace clustering. In addition, in the past few decades, because of technical advances in storing and processing data, we can obtain a large dataset that includes a large number of variables. Thus, we need to take into account such high-dimensional data. Cagnone and Viroli's [19] method, which is a common subspace clustering technique, cannot be used for a high-dimensional setting straightforwardly and may need strict restrictions for their parameters because of the identifiability problem.

Thus, we propose a new method to simultaneously find a cluster structure of multivariate binary data and an optimal low-dimensional space for clustering. The proposed model is based on the framework of latent class analysis (LCA) [27], which is used not only for analyzing the relation between categorical variables and discrete latent factors but for clustering objects with categorical features (e.g., [28]). Furthermore, our proposed method can deal with high-dimensional data.

The remainder of this paper is structured as follows. In Section 2, we introduce a new method to cluster multivariate binary data with dimension reduction. Section 3 describes an algorithm for the proposed optimization problem. Section 4 is devoted to studying the working of the clustering method using artificial and real data examples. Finally, we sum up our findings and set out directions for future expansion in Section 5.

2. Proposed method

Let $\tilde{y}=(\tilde{y}_1,\ldots,\tilde{y}_D)'$ be a random vector of D binary variables. Suppose there are K latent (unobservable) classes in a population and let $\tilde{u}_k$, $k=1,\ldots,K$, be an allocation variable that takes "1" if an observation belongs to class k, and "0" otherwise. We write $\tilde{\mathbf{u}}=(\tilde{u}_1,\ldots,\tilde{u}_K)'$. We assume that the allocation variable follows a multinomial distribution, i.e., the probability that $\tilde{\mathbf{u}}$ takes the value $\mathbf{u}=(u_1,\ldots,u_K)'$ is

$$f(\tilde{\mathbf{u}}=\mathbf{u})=\prod_{k=1}^{K}\xi_k^{u_k},$$

where $\xi_k=Pr(\tilde{u}_1=0,\ldots,\tilde{u}_k=1,\ldots,\tilde{u}_K=0)$.

Given that an observation is in the kth latent class, the probability that the random vector $\tilde{y}$ takes the value $\mathbf{y}=(y_1,\ldots,\mathbf{y}_D)'$, where each y_d takes 0 or 1, is represented as $Pr(\tilde{y}=y|\,\tilde{u}_k=1)$. The unconditional probability of the response $\mathbf{y}$ when we do not know the latent class of the observation is

$$\mathbf{Pr}(\tilde{y}=\mathbf{y})=\sum_{k=1}^{K}\xi_k Pr(\tilde{\mathbf{y}}=\mathbf{y}|\,\tilde{u}_k=1). \tag{2.1}$$

Here, we need to specify how the probability $Pr(\tilde{y}=\mathbf{y}|\,\tilde{u}_k=1)$ depends on parameters. We postulate that, given the latent class to which an observation belongs, the responses on the binary variables are independent:

$$Pr(\tilde{y}=\mathbf{y}|\,\tilde{u}_k=1)=\prod_{d=1}^{D}Pr(\tilde{y}_d|\,\tilde{u}_k=1). \tag{2.2}$$

This assumption of *conditional independence* has been widely used in latent class modeling in sociology [29], and is directly analogous to the assumption in the factor analysis model that observed variables are conditionally independent given the factors [27].

Finally, to specify the model completely, we need to specify a set of parameters that define the conditional probability of $\tilde{y}$, with the value of $\tilde{\boldsymbol{u}}$ given. Suppose that $\tilde{y}_1,\ldots,\tilde{y}_N$ are mutually independent random variables that have the same distribution as $\tilde{y}$, and the entries of $\boldsymbol{Y}=(y_{nd})$ are those realizations. We assume that, given the class k, $\tilde{y}_d$ follows the Bernoulli distribution with success probability π_{kd}. For the traditional LCA [27], we consider a parameter vector $\boldsymbol{\theta}_k=(\theta_{k1},\ldots,\theta_{kD})'$, where θ_{kd} is the logit transformation of π_{kd}. We define the inverse logit transformation $\pi(\theta)=(1+\exp(-\theta))^{-1}$. The success probabilities can be represented using the canonical parameters θ_{kd} as $\pi_{kd}=\pi(\theta_{kd})$. Let $\tilde{y}_{nd}$ be the dth element of $\tilde{y}_n$. The individual data-generating probability given the class then becomes

$$\begin{aligned}Pr(\tilde{y}_{nd}=y_{nd}|\,\tilde{u}_k=1)&=Pr(\tilde{y}_{nd}=y_{nd}|\,\tilde{u}_k=1,\theta_{kd})\\&=\pi(\theta_{kd})^{y_{nd}}1-\pi(\theta_{kd})^{1-y_{nd}}\\&=\pi(q_{nd}\theta_{kd}),\end{aligned}$$

with $q_{nd}=2y_{nd}-1$ since $\pi(-\theta)=1-\pi(\theta)$. Then, these representations lead to the compact form of the log likelihood as

$$\sum_{n=1}^{N}\log\left(\sum_{k=1}^{K}\xi_k\prod_{d=1}^{D}\pi(q_{nd}\theta_{kd})\right).$$

We aim to obtain a low-dimensional representation of binary data in which the true cluster structure exists. Thus, we assume that canonical parameter θ_{kd} has a low-rank representation as follows:

$$\theta_{kd}=\mu_d+\boldsymbol{f}_k'\boldsymbol{a}_d, \tag{2.3}$$

where $\mu_d\in\mathbb{R}$, and for some positive integer L, $\boldsymbol{f}_k\in\mathbb{R}^L$ and $\boldsymbol{a}_d\in\mathbb{R}^L$. Here, μ_d, $\boldsymbol{f}_k$, and $\boldsymbol{a}_d$ denote a centroid for the dth variable, a component score of the kth cluster, and a loading value for the dth variable, respectively. We write $\boldsymbol{\xi}=(\xi_1,\ldots,\xi_K)'$, $\boldsymbol{\mu}=(\mu_1,\ldots,\mu_D)'$, $\mathbf{F}=(\boldsymbol{f}_1,\ldots,\boldsymbol{f}_K)'$, and $\mathbf{A}=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_D)'$. To guarantee the determination of the decomposition for $\mathbf{F}$ and $\mathbf{A}$, we require that $\mathbf{F}$ has orthonormal columns. Then the log likelihood can be written as

$$\ell(\boldsymbol{\xi},\boldsymbol{\mu},\mathbf{F},\mathbf{A})=\sum_{n=1}^{N}\log\left(\sum_{k=1}^{K}\xi_k\prod_{d=1}^{D}\pi(q_{nd}(\mu_d+\boldsymbol{f}_k'\boldsymbol{a}_d))\right). \tag{2.4}$$

Here, to deal with the high-dimensional problem, we assume that most of the elements of the true $\mathbf{A}$ are exactly zero. A sparse

Please cite this article as: M. Yamamoto, K. Hayashi, Clustering of multivariate binary data with dimension reduction via L_1-regularized likelihood maximization, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.05.026