Computational Statistics and Data Analysis 105 (2017) 217–228

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A simple approach to sparse clustering[☆]

Ery Arias-Castro, Xiao Pu [*]

Department of Mathematics, University of California, San Diego 9500 Gilman Drive # 0112, La Jolla, CA 92093-0112, USA

ARTICLE INFO

Article history:
Received 24 February 2016
Received in revised form 4 August 2016
Accepted 7 August 2016
Available online 21 August 2016

Keywords:
Sparse clustering
Hill-climbing
High-dimensional
Feature selection

ABSTRACT

Consider the problem of sparse clustering, where it is assumed that only a subset of the features are useful for clustering purposes. In the framework of the COSA method of Friedman and Meulman, subsequently improved in the form of the Sparse K-means method of Witten and Tibshirani, a natural and simpler hill-climbing approach is introduced. The new method is shown to be competitive with these two methods and others.

1. Introduction

Consider a typical setting for clustering n items based on pairwise dissimilarities, with $\delta(i,j)$ denoting the dissimilarity between items $i,j \in [n] := \{1,\dots,n\}$. For concreteness, we assume that $\delta(i,j) \ge 0$ and $\delta(i,i) = 0$ for all $i,j \in [n]$. In principle, if we want to delineate κ clusters, the goal is (for example) to minimize the average within-cluster dissimilarity. In detail, a clustering into κ groups may be expressed as an assignment function $C : [n] \mapsto [\kappa]$, meaning that $C(i)$ indexes the cluster that observation $i \in [n]$ is assigned to. Let $\mathcal{C}_\kappa^n$ denote the class of clusterings of n items into κ groups. For $C \in \mathcal{C}_\kappa^n$, its average within-cluster dissimilarity is defined as

$$\Delta[C] := \sum_{k\in[\kappa]} \frac{1}{|C^{-1}(k)|} \sum_{i,j\in C^{-1}(k)} \delta(i,j). \tag{1}$$

This dissimilarity coincides with the *within-cluster sum of squares* commonly used in k-means type of clustering algorithms, with $\delta(i,j) = \|x_i - x_j\|^2$. If under the Euclidean setting, we further define cluster centers

$$\mu_k = \frac{1}{n} \sum_{i\in C^{-1}(k)} x_i \quad \text{with } k \in [\kappa],$$

then the within-cluster dissimilarity can be rewritten as follows,

$$\Delta[C] = \sum_{k\in[\kappa]} \frac{1}{|C^{-1}(k)|} \sum_{i,j\in C^{-1}(k)} \|x_i - x_j\|^2 = \sum_{k\in[\kappa]} \sum_{i\in C^{-1}(k)} \|x_i - \mu_k\|^2.$$

[☆] Reproducible research. The R code for the numerical experiments is available at https://github.com/victorpu/SAS_Hill_Climb.

[*] Corresponding author. Fax: +858 534 5273.
E-mail addresses: eariasca@ucsd.edu (E. Arias-Castro), xipu@ucsd.edu (X. Pu).
URLs: http://www.math.ucsd.edu/~eariasca/ (E. Arias-Castro), http://www.math.ucsd.edu/~xipu/ (X. Pu).

http://dx.doi.org/10.1016/j.csda.2016.08.003

Since this paper deals with non-Euclidean settings also, we will use the more general within-cluster dissimilarity defined in (1). The resulting optimization problem is the following:

$$\text{Given } (\delta(i,j) : i,j \in [n]), \ \text{minimize } \Delta[C] \text{ over } C \in \mathcal{C}^n_\kappa. \tag{2}$$

This problem is combinatorial and quickly becomes computationally too expensive, even for small datasets. A number of proposals have been suggested (Hastie et al., 2009), ranging from hierarchical clustering approaches to K-medoids.

Following the footsteps of Friedman and Meulman (2004), we consider a situation where we have at our disposal not 1 but $p \geq 2$ measures of pairwise dissimilarities on the same set of items, with $\delta_a(i,j)$ denoting the ath dissimilarity between items $i,j \in [n]$. Obviously, these measures of dissimilarity could be combined into a single measure of dissimilarity, for example,

$$\delta(i,j) = \sum_a \delta_a(i,j). \tag{3}$$

Our working assumption, however, is that only a few of these measures of dissimilarity are useful for clustering purposes, but we do not know which ones. This is the setting of sparse clustering, where the number of useful measures is typically small compared to the whole set of available measures.

We assume henceforth that all dissimilarity measures are equally important (for example, when we do not have any knowledge a priori on the relative importance of these measures) and that they all satisfy

$$\sum_{i,j \in [n]} \delta_a(i,j) = 1, \quad \forall a \in [p], \tag{4}$$

which, in practice, can be achieved via normalization, meaning,

$$\delta_a(i,j) \leftarrow \frac{\delta_a(i,j)}{\sum_{i,j} \delta_a(i,j)}.$$

This assumption is important when combining measures in the standard setting (3) and in the sparse setting (5) below.

Suppose for now that we know that at most s measures are useful among the p measures that we are given. For $S \subset [p]$, define the S-dissimilarity as

$$\delta_S(i,j) = \sum_{a \in S} \delta_a(i,j), \tag{5}$$

and the corresponding average within-cluster S-dissimilarity for the cluster assignment C as

$$\Delta_S[C] := \sum_{k \in [\kappa]} \frac{1}{|C^{-1}(k)|} \sum_{i,j \in C^{-1}(k)} \delta_S(i,j).$$

If the goal is to delineate κ clusters, then a natural objective is the following:

$$\begin{aligned} &\text{Given } (\delta_a(i,j) : a \in [p], i,j \in [n]), \\ &\text{minimize } \Delta_S[C] \text{ over } S \subset [p] \text{ of size } s \text{ and over } C \in \mathcal{C}^n_\kappa. \end{aligned} \tag{6}$$

In words, the goal is to find the s measures (which play the role of features in this context) that lead to the smallest optimal average within-cluster dissimilarity. The problem stated in (6) is at least as hard as the problem stated in (2), and in particular, is computationally intractable even for small item sets.

Remark 1. In many situations, but not all, p measurements of possibly different types are taken from each item i, resulting in a vector of measurements $x_i = (x_{ia} : a \in [p])$. This vector is not necessarily in a Euclidean space, although this is an important example — see Section 2.2. We recover our setting when we have available a dissimilarity measure $\delta_a(i,j)$ between x_{ia} and x_{ja}. This special case justifies our using of the terms 'feature' and 'attribute' when referring to a dissimilarity measure.

2. Related work

The literature on sparse clustering is much smaller than that of sparse regression or classification. Nonetheless, it is substantial and we review some of the main proposals in this section. We start with the contributions of Friedman and Meulman (2004) and Witten and Tibshirani (2010), which inspired this work.

2.1. COSA, sparse K-means and regularized K-means

Friedman and Meulman (2004) propose clustering objects on subsets of attributes (COSA), which (in its simplified form) amounts to the following optimization problem

$$\begin{aligned} &\text{minimize } \sum_{k \in [\kappa]} \alpha(|C^{-1}(k)|) \sum_{i,j \in C^{-1}(k)} \sum_{a \in [p]} (w_a \delta_a(i,j) + \lambda w_a \log w_a), \\ &\text{over any clustering } C \text{ and any weights } w_1, \ldots, w_p \geq 0 \text{ subject to } \sum_{a \in [p]} w_a = 1. \end{aligned} \tag{7}$$