Knowledge-Based Systems 74 (2015) 106–118

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Random spatial subspace clustering

Yi Guo [a,*], Junbin Gao [b], Feng Li [c]

[a] CSIRO Digital Productivity and Services Flagship, North Ryde, NSW 1670, Australia
[b] School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia
[c] Earth Observation Technology Application Department, Academy of Opto-Electronics, CAS, China

ARTICLE INFO

Article history:
Received 13 December 2013
Received in revised form 16 September 2014
Accepted 9 November 2014
Available online 15 November 2014

Keywords:
Spatial clustering
Subspace learning
Hyperspectral linear mixing
Sparse models
Randomization

ABSTRACT

Strong spatial or time correlation exists in many types of data, for example, the hyperspectral data acquired by a spectrometer scanning through rock samples from a drill hole. It is of practical interests to identify spatially continuous segments in a given data set where we know a priori that the samples are strongly correlated spatially. Recently, a novel method called spatial subspace clustering (SpatSC) was proposed to address this problem. However, due to the subspace learning nature of the SpatSC model, this method becomes intractable when the number of samples to be processed is very large. To alleviate computational intensity, we proposed a method called random spatial subspace clustering or RSSC for short. In RSSC, only a subset of data is segmented by SpatSC and an overall solution is obtained through propagation. This reduces the computational cost significantly. Yet a very important question to answer is to what extent the RSSC solution differs from that of SpatSC. In this paper, we analyse the propagation procedure and derive an average error rate of RSSC solution compared to SpatSC solution on the whole data set. The results show that the RSSC clustering result is close to SpatSC result under mild conditions. This provides a theoretic performance guarantee of RSSC. Our analysis also reveals the guided random sampling implemented by crude spatial clustering is crucial in improving RSSC results. We evaluate RSSC quantitatively on various data sets to assess its effectiveness under different settings. The results show that RSSC has similar performance to SpatSC as indicated by the theory while its computational cost is only a fraction of that of SpatSC.

1. Introduction

Many data sets exhibit strong spatial or time correlation among their samples. For example, the environmental data such as remote sensing images and soil mapping data [28] with very high spatial correlation. In fact, we concern the type of data in general that is acquired subjected to one external variable such as time, location and temperature. The data evolve along the direction of that variable and "neighboring" data very likely share similar pattern. In this paper, we focus on a very special kind of spatial data, that is the hyperspectral data of rocks from a drill hole, which are acquired by spectrometer scanning through the rock samples up to some depth. Fig. 1 illustrates a section of the drill hole and its corresponding spectra. The middle panel shows the image of rock samples labeled by their depths in the drill hole. The left panels plots several spectra associated with those samples. The spatial correlation among samples is obvious in the plot as the spectra from spatially close rock samples share similar features. This research is originated from this data. Therefore, in the following discussion throughout this paper, we constrain to spatial data to simplify the explanation. However, the algorithms to be detailed can be applied to more general cases.

It is of practical interests to identify spatially continuous segments in a given data set where we know a priori that the samples are strongly correlated spatially. For example the segments of many drill holes, often called domains in geology, from different locations can be aligned to map the mineral deposit of a certain area for exploration and mining purposes. A novel method called spatial subspace clustering (SpatSC) was proposed in [15] to address this problem. However, due to the subspace learning nature of the SpatSC model, the computational complexity of this method scales up with N^2, where N is the number of samples in a data set. Although SpatSC adopts every efficient convex programming scheme [24], when N is very large, say tens of thousands for a typical drill hole, SpatSC becomes intractable because of its $\mathcal{O}(N^2)$ memory consumption and computational cost, where $\mathcal{O}(x)$ means at the order of x.

* Corresponding author.
E-mail addresses: yi.guo@csiro.au (Y. Guo), jbgao@csu.edu.au (J. Gao), lifeng@aoe.ac.cn (F. Li).

http://dx.doi.org/10.1016/j.knosys.2014.11.006
0950-7051/

To alleviate computational intensity, one often takes two strategies, i.e. parallelization and reduction of complexity. The former requires separability so that a large amount of threads can be spawned out. In our particular case, because the data are coupled in spatial context, the separability is not very clear. Therefore we turn to the other strategy, that is to reduce the complexity of the large scale problem. We proposed a method called random spatial subspace clustering or RSSC for short in [14]. RSSC approaches our goal by exploiting the redundancy within data as the following. A small subset of the whole date set called *calibration set* is formed through random sampling from crude spatial segments. SpatSC finds the accurate clusters of the data in calibration set. Finally, the clustering solution for the remaining data is obtained by propagation. As the calibration set contains only a fraction of the data, the computational cost can be reduced significantly compared to the original SpatSC. However, a follow-up question is to what extent RSSC solution differs from that of SpatSC. We try to answer this question here as a main contribution of this paper. To this end, we discuss the crude spatial clustering schemes and analyse the propagation procedure employed in RSSC in detail. Based on the worst scenario, namely, choosing samples totally by random without any guidance, we derive a very conservative average error rate of RSSC solution compared to SpatSC solution. Not surprisingly, RSSC clustering result depends on the size of the calibration set. The larger, the better. The result also suggests that another possible way to improve RSSC performance is to include as many true segment boundaries as possible in calibration set.

We evaluated RSSC quantitatively on controlled semi-simulated thermal infrared data sets against other state-of-the-art clustering algorithms including SpatSC. The results of RSSC are satisfactory when SNR is normal (around 40 db), which confirms our findings. We applied this algorithm to an entire thermal infrared drill hole data set with a comparison with a linear unmixing model called TSA (The Spectral Assistant) [4]. They produce similar results. However, TSA requires a significant amount of human intervention and a spectral library while RSSC is a fully automated unsupervised procedure. To further demonstrate its usefulness to other types of data, we extended our range of tests to include *in situ* X-ray Diffraction (XRD) data for material science research and mitochondrial calcium overload (MCO) data for a study on mouse heart cells. The former is a time/temperature revolving X-ray counts patterns and the latter is a typical functional data. The evaluation shows that RSSC is superior to other clustering methods in terms of cluster quality and its performance is similar to or better than that of SpatSC.

The rest of this paper is organized as follows. We will give a brief introduction to SpatSC in the next section, which is followed by a thorough discussion on RSSC. In Section 4, we evaluate RSSC on various data sets and we draw a conclusion in the last section.

2. Spatial subspace clustering

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ be the matrix of the hyperspectral data taken from a drill hole, and $\mathbf{x}_i \in \mathbb{R}^D$ the i-th individual spectrum. Note that index $i \in \{1 \ldots N\}$ corresponds to the physical location of a particular sample which is the depth. D is determined by the spectral resolution of a spectrometer. We write matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_M]$, $\mathbf{a}_i \in \mathbb{R}^D$, as the spectral library of pure materials [12], which are the bases of the spectra in $\mathbf{X}$. Precisely,

$$\mathbf{x}_i = \mathbf{A}\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{\beta}_i$ is the coefficient vector and $\boldsymbol{\epsilon}_i$ is the error. This is called the linear mixing model [12]. $\boldsymbol{\beta}_i$ is supposed to be very sparse, i.e. a lot of its elements are zero because a rock sample can only contain a few materials. In subspace learning literature, $\mathbf{A}$ is called dictionary. There are some cases where $\mathbf{A}$ has to be estimated, which is often referred as dictionary learning [2].

Spatial subspace clustering algorithm (SpatSC) is a member of the family of subspace learning algorithms [6,19,26,32]. The model is built on the assumption that the rocks are stratified and hence the drill hole spectral samples are congregated as continuous segments. The purpose of the spatial subspace clustering algorithm is to recover these spatially continuous segments *without referring to a spectral library*. Therefore it is an unsupervised learning method. As we mentioned in Section 1, although the spatial subspace clustering is designed for spatial data, it can be applied to any type of data where some smoothness is associated with the indices.

SpatSC has two major components in its formulation. The first is the data self-reconstruction. It works with the sparsity constraint

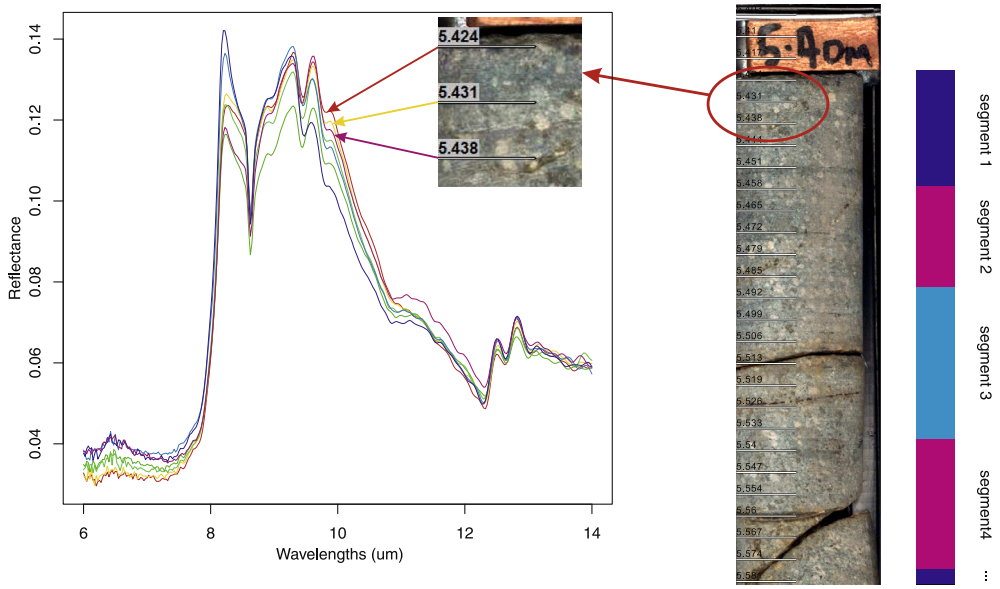

Fig. 1. An example of a small section of a drill hole data. The middle panels is the image of the rock samples with the numbers indicating the depths (in meters). The left panel shows the spectra of the rock samples acquired by a thermal infrared spectrometer with 321 wavelengths from 6 μm to 14 μm. The right panel is the illustrative spatial segments of these rock samples. Note that one subspace may have several physically separated segments, e.g. segment 2 and segment 4.