Computational Statistics and Data Analysis 66 (2013) 193–201

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Two-way incremental seriation in the temporal domain with three-dimensional visualization: Making sense of evolving high-dimensional datasets

Peter Wittek

Swedish School of Library and Information Science, University of Borås, Allegatan 1, Borås, S-501 90, Sweden

ARTICLE INFO

Article history:
Received 25 June 2012
Received in revised form 11 February 2013
Accepted 26 March 2013
Available online 1 April 2013

Keywords:
Two-way seriation
Gaussian filtering
Landscape visualization
High-dimensional data
Hamiltonian path

ABSTRACT

Two-way seriation is a popular technique to analyze groups of similar instances and their features, as well as the connections between the groups themselves. The two-way seriated data may be visualized as a two-dimensional heat map or as a three-dimensional landscape where colour codes or height correspond to the values in the matrix. To achieve a meaningful visualization of high-dimensional data, a compactly supported convolution kernel is introduced, which is similar to filter kernels used in image reconstruction and geostatistics. This filter populates the high-dimensional space with values that interpolate nearby elements and provides insight into the clustering structure. Ordinary two-way seriation is also extended to deal with updates of both the row and column spaces. Combined with the convolution kernel, a three-dimensional visualization of dynamics is demonstrated on two datasets, a news collection and a set of microarray measurements.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Biclustering and heat maps are visual approaches to demonstrate the clustering structure in a set of data with a given feature set. As the dataset is updated, the existing distributional pattern of instances and features changes, their relations and relative importance shift. This leads to a dynamic pattern that we wish to visualize to gain insight into evolving structures.

Several methods exist that achieve a similar purpose. For instance, self-organizing maps (SOMs) have been used to provide a visual overview of the clusters in expanding document collections (Kohonen et al., 2000). A SOM is a two-dimensional grid of a neural network in which the nodes are adjusted as subsequent document vectors stimulate the network. Eventually clusters of documents will be assigned to nodes and the overall structure can be visualized.

In a neural network, the two dimensional layout is somewhat arbitrary and only relative positions matter. The x and y axes are not meaningful. The axes in a matrix layout, however, are typically easily interpreted. For instance, the row space may refer to features and the column space to instances that the features describe. In a term-document matrix the features are index terms (terms that are semantically charged to be significant for marking up a document), whereas instances are documents (text, a web page, a blog entry, etc.). If we rearrange the rows and columns in the matrix such that related terms will be near one another, and clusters of similar documents will also be nearby column vectors, we arrive at a visual reflection of a statistical model, a heat map of the matrix (Wilkinson and Friendly, 2009).

The underlying idea of a heat map is a *seriation* of both the row and the column space. We regard seriation as "sequencing objects along a continuum that rely upon a symmetric proximity measure defined between the objects to be seriated" (Hubert, 1974). Seriation is widely used in the visualization of binary matrices (Chen, 2002; Liiv et al., 2011), and it is also used in genetic research to understand which genes are activated simultaneously (Eisen et al., 1998; Caraux and Pinloche, 2005).

E-mail address: peterwittek@acm.org.

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.03.026

In the context of text mining, Wittek et al. (2009) introduced a seriation using a distance function that did not only rely on the distributional patterns of words, but also included an external lexical database encoding word relations. This method translated seriation to a graph problem in which an approximate solution to a minimum-weight Hamiltonian path was sought (Rosenkrantz et al., 1977). The weights in the path corresponded to distances, and the nodes to words. While many other methods exist to find a one-way seriation (Liiv, 2010), this was proved to be efficient when applied to text classification in conjunction with compactly supported basis functions, B-spline kernels in particular (Wittek and Tan, 2011).

The purpose of the present work is to introduce an incremental two-way seriation on a non-binary matrix, that is, seriating both the row and column spaces thereof, and subsequently visualize the resulting matrix. The two-way seriation must be able to accommodate updates of the matrix, as either the feature space or the number of instances or both can expand. Eventually we arrive at a surface-type visualization in three dimensions based on the seriated matrix that shows the structure of the expanding collection. We demonstrate the algorithm using two datasets. The first one is a large collection of news, which, when processed, results in a sparse vector space with a very high number of dimensions. The second collection is a series of experiments on yeast; the instances refer to experiment types, whereas the features are derived from microarray measurements. The key contributions of the paper are as follows:

- Large-scale two-way seriation of sparse data.
- Incremental updates of both the row and column space seriation.
- Application of filtering for better visualization of sparse data.
- 3D visualization of dynamics.

The rest of the paper is organized as follows. We quickly overview the relevant work on seriation and two-way seriation in Section 2. Section 3 introduces the problem of temporal evolution of datasets and proposes a solution to extend two-way seriation to such cases. The high-dimensional data might be sparse, or the number of features can be so high that visualizing the result of the two-way seriation can be difficult. Section 4 offers a solution by applying a convolution filter to the output of the two-way seriation. Using high-dimensional sparse datasets, Section 5 explores the parameters of the proposed method, and Section 6 applies the method to another collection. Finally, Section 7 concludes the paper.

2. Seriation and two-way seriation

Seriation is a combinatorial data analysis method that reorders instances into a sequence along a one-dimensional continuum. The basis of the reordering is a pairwise distance between the instances. Seriation algorithms place instances next to each other if the distance between them is small, and the eventual order reveals regularity and patterns in the whole series (Liiv, 2010). Seriation is different from clustering: clustering groups nearby objects together, but it does not necessarily reveal relations among the groups. Yet both approaches are NP-hard problems (Wilkinson and Wills, 2005, p.525).

Seriation as described above is uncommon. Two-way seriation is the typical application. Two-way seriation assumes a matrix representation, or a two-dimensional layout of the data, where rows correspond to instances and columns correspond to features that describe the individual instances (this assignment of rows and columns is arbitrary and can happen the other way). This approach has a history of over a hundred years (see Liiv, 2010 for an overview), and it is commonly used in a wide variety of information visualization methods, including microarray data (Eisen et al., 1998; Caraux and Pinloche, 2005), binary matrices (Liiv et al., 2011), and others (Chen, 2002). Seriation is similar to clustering, and two-way seriation is similar to biclustering: instances with similar feature subsets are grouped together in regular patterns. The difference is that the overall structure of the seriated two-dimensional array is meaningful, the groups follow each other in an optimized order. In other words the overall structure is not an arbitrary ordering of row and column cluster trees. Certain types of two-way seriation are also called heat maps (Wilkinson and Friendly, 2009).

Many seriation algorithms have implementations that are available for free, and some are also available under an open source licence (Caraux and Pinloche, 2005; Hahsler and Hornik, 2007; Wu et al., 2010). These implementations deal with dense data.

While sparse data asks for similar seriation methods, visualizing the result is harder. Berry et al. (1996) tested different methods to improve the browsing experience of hypertext documents with promising results, although the plots of the sparse matrices resulted from document indexing were not immediately useful for visual analysis. Minimizing the Hamiltonian path length results in a seriation optimal with respect to dissimilarities between neighbouring objects (Rosenkrantz et al., 1977; Hahsler et al., 2007). Taking this approach as its basis, Wittek et al. (2009) dealt with one-way seriation of sparse data to improve the performance of text classification. Potential visual applications were not considered.

We extend a Hamiltonian path-based seriation to deal with two-way sparse data that are dynamically updated, and we explore some potential visualization methods that reveal patterns in the collection (see Fig. 1).

3. Dynamic seriation

The seriation method described in Wittek et al. (2009) works well with a wide range of distance functions on sparse data. The key idea of the algorithm is to study the weighted graph that is described by the pairwise distances between feature vectors. Nodes correspond to features; the weight of an edge is the distance between the two features. More formally, let V denote a set of instances $\{x_1, x_2, \ldots, x_n\}$, where n is the number of instances; the instances are in an arbitrary order.

Download English Version:

https://daneshyari.com/en/article/417460

Download Persian Version:

https://daneshyari.com/article/417460

Daneshyari.com