Knowledge-Based Systems 000 (2016) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Clustering time-stamped data using multiple nonnegative matrices factorization

Xiaohui Huang[a,b,*], Yunming Ye[b], Liyan Xiong[a], Shaokai Wang[b], Xiaofei Yang[b]

[a] School of Information Engineering Department, East China Jiaotong University, Nanchang, 330013, China
[b] Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China

ARTICLE INFO

Article history:
Received 29 March 2016
Revised 29 September 2016
Accepted 4 October 2016
Available online xxx

Keywords:
Clustering
Time-stamped data set
Matrix factorization
Social media

ABSTRACT

Time-stamped data are ubiquitous in our daily life, such as twitter data, academic papers and sensor data. Finding clusters and their evolutionary trends in time-stamped data sets are receiving increasing attention from researchers. Most existing methods, however, can only tackle the clustering problem of a data set without time-stamped information which is inherent in almost all the data objects. Actually, not only the performance can be improved by effectively incorporating the time-stamped information in the clustering process on most data sets, but also we can find the evolutionary trends of the clusters with time information. In this paper, we introduce an approach for clustering time-stamped data and discovering the evolutionary trends of the clusters by using Multiple Nonnegative Matrices Factorization (MNMF) with smooth constraint over time. To utilize time-stamped information in the clustering process, an extra object-time matrix is constructed in our proposed method. Then, we jointly factorize multiple feature matrices using smooth constraint to perform the object-time matrix to obtain the clusters and their evolutionary trends. Experimental results on real data sets demonstrate that our proposed approach outperforms the comparative algorithms with respect to Fscore, NMI or Entropy.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

In the last decade, tremendous research efforts have been devoted to clustering techniques to analyze time-stamped data. Especially, with the development of the Internet, large-scale time-stamped data, such as News event data sets, Microblog data sets and Sensor data sets etc., are generated in our daily life. We see a growing demand for finding the clusters hidden in a time-stamped data set. Another demand is to find the evolutionary trends of the clusters with the help of the time-stamped information of every object. The analysis of time-stamped data is extremely useful for some applications, such as time prediction [1,2], community evolution [3] and trend analysis of researching interests [4,5], to name just a few. However, most existing clustering methods do not consider the time-stamped feature of a data set as a useful factor in the clustering process. Take document clustering as an instance, the representative vector space model [6] employs only term frequency matrix in the clustering process. In fact, time stamp as a feature of the document can also play an important role in clustering tasks from real applications. For example, in topic analysis of U.S. Presidential State-of-the-Union addresses, the methods that do not consider time-stamped feature may confound Mexican-American War (1846–1848) with some aspects of World War I (1914–1918), because these methods are unaware of the 70-year separation between two events [4].

In this paper, we introduce a new method for analyzing time-stamped data, named Multiple Nonnegative Matrices Factorization (MNMF). It starts by using multiple matrices to represent a data set and each matrix represents a type of feature in a data set. In comparison to traditional methods, e.g. single matrix factorization [7], MNMF is able to fuse multiple types of features to improve the clustering performance. We then decompose these matrices simultaneously for utilizing different types of features information in a data set. For example, we can employ document-term frequency matrix, document-time stamp matrix and document-author simultaneously in document clustering. Moreover, since multiple matrices factorization is able to incorporate time-stamped information, the evolutionary trends of the clusters over time can be discovered by our proposed approach. Generally, the evolutionary trend of a cluster has continuity along the time dimension to some extent, i.e. the trend curve of a cluster should be smooth to a certain extent. Therefore, smooth constraint to the time dimension is used in our model. Experimental results on various data sets suggest that our proposed approach outperforms the state-of-the-art

* Corresponding author.
E-mail address: hxh016@gmail.com (X. Huang).

http://dx.doi.org/10.1016/j.knosys.2016.10.007
0950-7051/© 2016 Elsevier B.V. All rights reserved.

algorithms with respect to various metrics. The main contributions of this paper are:

- We propose a nonnegative multiple matrices factorization approach for clustering the data which has multiple types of features.
- Since the evolutionary trend of a cluster has continuity, we propose the smooth constraint to time stamp dimension in the process for discovering the evolutionary trend of a cluster.
- We develop an objective function for our proposed algorithm and give the updating rules through optimizing the objective function.

The remaining sections of this paper are organized as follows: a brief overview of related works about clustering time-stamped data is given in Section 2 and the preliminary of our method is presented in Section 3. Section 4 introduces our proposed time-stamped data clustering method using multiple nonnegative matrices factorization with smooth constraint. Experiments on real data sets are presented in Section 5. We discuss the features of our proposed algorithm in Section 6 and conclude the paper in Section 7.

2. Related work

In this section, we give a brief survey of clustering time-stamped data and detecting the evolutionary trends of clusters on time-stamped data from two aspects: Discrete-time Model and Continuous-time Model.

2.1. Discrete-time model

Most existing methods which study time-stamped data are Discrete-time model. It first partitions the time into several time intervals. And then, the objects in a data set are allocated into the subsets by the time intervals which the time stamps of the objects belong to. The objects in the same subsets are seen as having the same time stamp.

2.1.1. Matrix and tensor factorization for clustering time-stamped data

Since Lee and Seung [7] proposed Nonnegative Matrix Factorization (NMF) for learning the parts of objects, many variants of NMF are proposed to solve different problems. Chen and Cichocki [8] proposed NMF with temporally smooth constraint to early detection Alzheimer disease using EEG recordings. Lin et al. proposed *FacetNet* [9] to detect communities and their evolutions in dynamic temporal networks. However, single nonnegative matrix factorization can tackle a data set with only one type of feature or relationship.

Tensor, or high-order array is widely used to represent the data which have multiple types of features and Nonnegative Tensor Factorization (NTF) is a useful tool to analyze this type of data. Therefore, some researchers extended NMF to NTF [1,3,10,11] for analyzing the data sets with time stamp. Lin et al. proposed MetaGraph Factorization (MetaFac) [3] in which multiple tensors are used to represent multi-relational data. MetaFac aims at discovering community structure in rich media social networks through analysis of time-varying, multi-relational data. Matsubara et al. [1] employed a tensor to represent a "web-click logs" data set and used collapsed Gibbs sampling [12] mine time-evolving events. However, a tensor which is used to represent a data set is usually very sparse in most real applications. For example, in the reference [10], the tensor used to represent an Email data set includes only 0.0364% nonzero elements. To reduce the sparsity of data, some researchers [13,14] employed jointly multiple matrices factorization to analyze a data set with multiple types of features. However, the multiple nonnegative matrices factorization cannot directly apply for analyzing time-stamped data sets.

2.1.2. Other methods for detecting evolutionary trend of time-stamped data set

Evolutionary clustering is an emerging research topic essential to time-stamped data as it can track the changing trend of every cluster. Chakrabarti et al. [15] presented a generic framework to find the evolution of time-stamped data by fusing the data in the current time interval and that in the previous intervals based on k-means and agglomerative hierarchical clustering algorithms. Kawadia [16] developed a new measure of partition distance called *estrangement* motivated by the inertia of inter-node relationships which facilitates to detect meaningful temporal communities. Mei and Zhai [17] extracted the themes independently in each interval, and then, employed the hidden Markov model to explore the relationships between themes in different time intervals. This method is able to analyze the evolutions of the clusters, but cannot effectively employ the time-stamped information to discover the clusters. That is to say, these methods mainly contribute to explore how a topic evolve into more subtopics or more subtopics merge into a topic over time. Different to the methods mentioned above, our proposed method mainly studies how to detect the evolutionary trend of a topic, i.e. how a topic rises and fades. Leskovec et al. [18] developed a framework for tracking short, distinctive phrases through on-line text and designed scalable algorithms for clustering textual variants of such phrases. This work defines a thread associated with a given phrase cluster and tracks all threads over time considering both their individual temporal dynamics as well as their interactions with one another. Pruteanu et al. [19] proposed a hierarchical Bayesian model of topics for time-stamped documents based on hierarchical Dirichlet process [20]. Yang and Leskovec [21] proposed a K-spectral Centroid clustering algorithm to discover the temporal patterns associated with online content and how the content's popularity grows and fades over time. However, these methods cannot integrate multiple types of features in a data set to analyze its evolution. In order to track the change of the communities in network data, Du et al. [22] proposed a framework which formulates the problem of tracking temporal community strength as an optimization task by orthogonal non-negative matrix factorization and Kalyanam et al. [23] modeled the topic evolution by leveraging social context and community information using collective matrices factorization.

Recently, many works [24–26] study information propagation of time-stamped data. Iribarren and Moro [24] tracked the step-by-step email propagation of an invariable viral marketing message and found that the spreading nodes activity level is relevant to their out-degrees and active off-springs and the possibility of a node to become a spreader grows with the depth of the node in the propagation path. Haralabopoulos et al. [25] studied the lifespan and propagation of information in an on-line social network: Reddit. Different to the methods of analyzing information propagation in social media, our proposed method devotes to find the clusters in time-stamped data and their evolutionary trends.

2.2. Continuous-time model

Different to the Discrete-time model, the time is seen as continuous in continuous-time model. In Topics over Time (TOT) [4], Wang and McCallum assumed that the evolutionary trend of a cluster subjects to beta distribution. Thus, they used beta distribution with different parameters to model the evolutionary trends of the clusters and employed Latent Dirichlet Allocation (LDA) [27] model to distinguish the content of every cluster. However, it is often unreasonable to assume that the evolutionary trends of all clusters subject to beta distribution in some real applications.

Please cite this article as: X. Huang et al., Clustering time-stamped data using multiple nonnegative matrices factorization, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.10.007

Download English Version:

https://daneshyari.com/en/article/4946445

Download Persian Version:

https://daneshyari.com/article/4946445

Daneshyari.com