Knowledge-Based Systems 000 (2017) 1–14

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation

Xiaobo Chen[a,*], Zhongjie Wei[b], Zuoyong Li[c], Jun Liang[a], Yingfeng Cai[a], Bob Zhang[d]

[a] Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China
[b] School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China
[c] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350108, China
[d] Department of Computer and Information Science, University of Macau, Macau, China

ARTICLE INFO

Article history:
Received 24 August 2016
Revised 4 June 2017
Accepted 6 June 2017
Available online xxx

Keywords:
Missing data
Low-rank matrix completion
Nearest neighbor
Pearson's correlation
Ensemble learning

ABSTRACT

Low-rank matrix completion (LRMC) is a recently emerging technique which has achieved promising performance in many real-world applications, such as traffic data imputation. In order to estimate missing values, the current LRMC based methods optimize the rank of the matrix comprising the whole traffic data, potentially assuming that all traffic data is equally important. As a result, it puts more emphasis on the commonality of traffic data while ignoring its subtle but crucial difference due to different locations of loop detectors as well as dates of sampling. To handle this problem and further improve imputation performance, a novel correlation-based LRMC method is proposed in this paper. Firstly, LRMC is applied to get initial estimations of missing values. Then, a distance matrix containing pairwise distance between samples is built based on a weighted Pearson's correlation which strikes a balance between observed values and imputed values. For a specific sample, its most similar samples based on the distance matrix constructed are chosen by using an adaptive K-nearest neighboring (KNN) search. LRMC is then applied on these samples with much stronger correlation to obtain refined estimations of missing values. Finally, we also propose a simple but effective ensemble learning strategy to integrate multiple imputed values for a specific sample for further improving imputation performance. Extensive numerical experiments are performed on both traffic flow volume data as well as standard benchmark datasets. The results confirm that the proposed correlation-based LRMC and its ensemble learning version achieve better imputation performance than competing methods.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

Intelligent Transportation System (ITS) is an effective way to alleviate traffic congestion and improve transportation efficiency. It is an integrated comprehensive system, which synthesizes a variety of technologies, including information, computer, data communication, sensor, electronic control, automatic control theory, operations research, and artificial intelligence. Data is one of the most important factors for intelligent transportation system, where the most popular parameters include average speed, real-time vehicle volume, average lane occupancy rate, etc. By collecting and analyzing massive amounts of traffic data, ITS can manage and predict better. For example, it can (1) find traffic anomalies quickly and make traffic management convenient, (2) discover inherent rules and knowledge from traffic data, so as to improve the operational efficiency of traffic management and road traffic capacity. Based on the predicted traffic flow a few hours ahead, users are able to adjust their route plans in advance in order to avoid congested roads. Thus, traffic data will play a fundamental role in the construction of ITS.

In actual traffic environment, the data collected by traffic equipment, e.g., loop detectors, are usually not completed where many missing values may occur because of a variety of reasons, such as the failures of loop detectors or transmission network. In the case of incomplete traffic data, it is insufficient to express traffic information accurately. More importantly, it prevents the applications of many classic data mining algorithms, such as support vector machine [1–3], neural networks [4], sparse learning [5], etc., because these algorithms generally require a complete set of data. Therefore, the imputation of missing values in a loop detection system is of great value.

Besides traffic data, missing values also occur frequently in other real-life processes, such as physical measurements, commercial surveys, business reports, etc. In the data mining and machine learning community, missing values imputation has attracted

* Corresponding author.
E-mail address: xbchen82@gmail.com (X. Chen).

http://dx.doi.org/10.1016/j.knosys.2017.06.010
0950-7051/© 2017 Elsevier B.V. All rights reserved.

Please cite this article as: X. Chen et al., Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.06.010

much attention during the last several decades because of its benefits in various practical applications [6]. Here, imputation means a procedure that replaces the missing values in a dataset with some plausible values [7,8]. Until now, researchers have proposed many imputation approaches, some of which are general-purpose and suitable for different types of applications. Typical methods belonging to this category include singularity value decomposition (SVD) [9], local least squares regression (LLS) [10], probabilistic principle component analysis (PPCA) [11–13], K-nearest neighbors (KNN) [8], etc. More imputation methods can be found in [7]. Besides the above approaches, some imputation methods are specifically developed for traffic volume data [14–16] with inherent spatial or temporal correlations [15,17] due to road network connectivity. These methods typically assume that missing data is localized to certain links and time intervals. As a result, the historical data can be used to infer the relationship between the target road and its neighbors or past states of that road.

Recently, low-rank matrix completion (LRMC) was applied to traffic data imputation and achieved promising results [18,19]. LRMC consists of finding or approximating a low-rank matrix based on a few observable entries of this matrix. It essentially depends on the correlation between rows and columns of a matrix. For example, empirical results [18] have shown that in the case of road network, LRMC is able to estimate missing values with better accuracy in comparison with other imputation methods. However, the classic LRMC potentially ignores the specificity of the recorded traffic volumes, because this method simply treats traffic data as a whole and imputes missing values by imposing low-rank constraint in a global way. In fact, the variation trends of traffic flow are not fully consistent across the whole samples, depending on different road segments and/or different weekdays. It indicates that the degree of linear dependence between traffic flow volumes is subject to large variation. However, the current LRMC-based imputation ignores those subtle but crucial distinctions and will inevitably impede the performance of LRMC. This issue is not only limited to traffic data, but also extends to many scenarios where LRMC is applied.

Aiming to address the above shortcomings of imputation methods based on LRMC, we propose in this paper a novel method called correlation-based LRMC (CLRMC) and its enhanced version with ensemble learning (CLRMC-EN). Specifically, the proposed methods are significant from the following aspects:

(1) CLRMC concentrates more on each sample and its most similar nearest neighbors [20] which suit the assumption of LRMC very well, owing to stronger temporal and spatial correlation. Therefore, it is expected to improve the imputation performance of LRMC.
(2) CLRMC-EN fully makes use of multiple imputation results for each sample within the framework of ensemble learning [21,22]. Each imputation result can provide valuable information for missing entries. By integrating them properly, the imputation performance can be further enhanced.
(3) The proposed methods are evaluated on both real traffic volume data and standard machine learning datasets from the UCI repository [23,24]. Experiments show that CLRMC and CLRMC-EN outperform other competing methods in most cases, thus confirming their effectiveness.

The rest of this paper is organized as follows. Section 2 presents some preliminaries including related works, the problem definition of missing data imputation, and the mathematical model of LRMC. In Section 3, we present correlation-based LRMC imputation framework and an ensemble learning strategy. In Section 4, we evaluate different imputation methods through extensive experiments on real-world traffic volume data and the standard UCI datasets. At last, we conclude this paper in Section 5 and discuss some future works.

2. Preliminaries

2.1. Related works

In the literatures, many researchers have studied the problem of missing data occurring in different fields. A quick approach is to replace missing values with mean, median or model of the observed values corresponding to the same variable. K-nearest neighbors (KNN) algorithm [8] and its weighted version were also used for imputation because of its simplicity. The above methods are easy to implement, however, with limited imputation performance.

In maximum likelihood estimation (MLE) based methods [13,25,26], such as Probabilistic Principal Component Analysis (PPCA), a specific parametric model, e.g., multivariate Gaussian mixture model, is adopted to describe the distribution of data. Then, model fitting and missing data imputation can be implemented simultaneously. It takes advantage of the relevance between observed data and latent variables and adopts the expectation-maximization (EM) algorithm [27] to estimate the parameters of the model. However, PPCA may perform poorly when the missing ratio is high [28] due to the intrinsic characteristic of EM algorithm [27].

Regression-based imputation methods [10,29,30] attempt to construct a mapping function from known attributes to missing attributes. Local least squares (LLS) imputation [10] is a typical regression-based method which was applied to DNA microarray gene expression data. In this method, a target gene with missing values is represented as a linear combination of similar genes. LLS uses a local similarity measure based on Pearson's correlation to choose similar genes [31] and applies least squares optimization to estimate missing values. More imputation methods as well as their performance comparison in supervised learning scenario can be found in [7].

For traffic data, some methods make explicit use of spatial and temporal correlation caused by prior topology of road network as well as the periodicity of human activity. For example, two adjacent loop detectors tend to have similar traffic flow pattern. Temporal correlation based imputation solves missing data problem by replacing missing values with the average of the temporally nearest observable values on the same loop detector at the same day or other days. Spatial correlation based imputation estimates missing values by taking the average of observed values on other loop detectors with similar traffic flow variation, at the same sampling time point. Spatial correlation prefers large road networks with numerous loop detectors, which also restricts its application in limited or even single detector situations. To measure the similarity between traffic samples, different distance metrics can be applied, such as Euclidean distance, Pearson's correlation, Cosine distance, etc. Due to the simple structure, these imputation approaches have the advantage of high computational efficiency [16], their imputation performance, however, is often limited.

LRMC [18,19,32] has drawn much attention recently because of its promising performance in many applications, such as traffic data imputation. This method formulates traffic data in matrix form. Considering the temporal and spatial correlation between traffic volume data, the rows and columns of the corresponding matrix are assumed to be linearly dependent, which leads the matrix to have a low-rank structure. Under this condition, missing values can be imputed reliably by finding a completed matrix with smallest rank. However, the rank minimization problem has been proved to be NP-hard and very difficult to solve in practice. Fortunately, E. J. Candès and B. Recht showed that [33] the rank of most matrices can be perfectly recovered by the nuclear norm of

Please cite this article as: X. Chen et al., Ensemble correlation-based low-rank matrix completion with applications to traffic data imputation, Knowledge-Based Systems (2017), http://dx.doi.org/10.1016/j.knosys.2017.06.010

Download English Version:

https://daneshyari.com/en/article/4946139

Download Persian Version:

https://daneshyari.com/article/4946139

Daneshyari.com