Scalable tensor factorizations for incomplete data

The problem of incomplete data—i.e., data with missing or unknown values—in multi-way arrays is ubiquitous in biomedical signal processing, network traffic analysis, bibliometrics, social network analysis, chemometrics, computer vision, communication networks, etc. We consider the problem of how to factorize data sets with missing values with the goal of capturing the underlying latent structure of the data and possibly reconstructing missing values (i.e., tensor completion). We focus on one of the most well-known tensor factorizations that captures multi-linear structure, CANDECOMP/PARAFAC (CP). In the presence of missing data, CP can be formulated as a weighted least squares problem that models only the known entries. We develop an algorithm called CP-WOPT (CP Weighted OPTimization) that uses a first-order optimization approach to solve the weighted least squares problem. Based on extensive numerical experiments, our algorithm is shown to successfully factorize tensors with noise and up to 99% missing data. A unique aspect of our approach is that it scales to sparse large-scale data, e.g., 1000 × 1000 × 1000 with five million known entries (0.5% dense). We further demonstrate the usefulness of CP-WOPT on two real-world applications: a novel EEG (electroencephalogram) application where missing data is frequently encountered due to disconnections of electrodes and the problem of modeling computer network traffic where data may be absent due to the expense of the data collection process.

General information
Publication status: Published
Organisations: Cognitive Systems, Department of Informatics and Mathematical Modeling, Gebze Institute of Technology, Sandia National Laboratories
Contributors: Acar, E., Dunlavy, D. M., KOIda, T. G., Marup, M.
Pages: 41-56
Publication date: 2011
Peer-reviewed: Yes

Publication information
Journal: Chemometrics and Intelligent Laboratory Systems
Volume: 106
Issue number: 1
ISSN (Print): 0169-7439
Ratings:
BFI (2011): BFI-level 2
Scopus rating (2011): CiteScore 2.27 SJR 0.77 SNIP 1.364
Web of Science (2011): Impact factor 1.92
ISI indexed (2011): ISI indexed yes
Web of Science (2011): Indexed yes
Original language: English
Keywords: incomplete data, optimization, CANDECOMP, tensor factorization, missing data, PARAFAC
DOIs:
10.1016/j.chemolab.2010.08.004
Source: orbit
Source-ID: 267376
Research output: Contribution to journal › Journal article – Annual report year: 2011 › Research › peer-review