Chapter 23
Mining Multimodal Big Data:
Tensor Methods and Applications

Sujoy Roy
University of Memphis, USA

Michael W. Berry
University of Tennessee, USA

ABSTRACT

The last decade has witnessed exponential growth of data particularly in the fields of biomedicine, unstructured text processing and signal processing. There exist instances of data depicting simultaneous interactions amongst more than two types of entities. Such data are not readily amenable to matrix representation as matrices can show interactions between only two types of entities at a time. Tensors are multimodal extensions of matrices (a matrix can be thought of as 2-mode tensor), and tensor factorizations (decompositions) are multiway generalizations of matrix factorizations. This chapter provides an overview of tensor factorization methods as well as a literature review of selected applications in areas that are currently experiencing exponential data growth and likely of interest to a broad audience.

INTRODUCTION

Most datasets can be organized into tables (matrices) typically depicting pairwise relationships between two types of entities: objects (rows) and attributes (columns). The entries of the matrix contain attribute values for each object. Given a data matrix $A(m \times n)$ with m rows and n columns ($m << n$), the m objects can be considered n -dimensional row vectors in attribute space, while the n attributes may be interpreted as m -dimensional column vectors in object space. The objects may be prioritized against each other or clustered together by calculating similarities between their vectors. For a large data matrix such as one containing frequencies of several thousand terms (attributes) in a few hundred documents (objects), it is time consuming to compare document vectors. Matrix factorization (decomposition) methods (Golub & Van Loan, 2012) are dimensionality reduction techniques that may be utilized in part to reduce the size of the coordinate space in which to compare the vectors.

DOI: 10.4018/978-1-5225-3142-5.ch023

Singular Value Decomposition (Skillicorn, 2007) is a factorization method that expresses the original $A(m \times n)$ matrix as a scaled product of $U(m \times r)$, $(r \times r)$, and $V'(r \times n)$ component matrices, where ' denotes the transpose. The two sets of original object and attribute vectors are transformed into a new r-dimensional orthogonal space ($r \leq m$) in which the maximum variation is expressed along the first dimension axis, as much variation independent of that is expressed along an axis orthogonal to the first, and so on. The new set of axes may reveal the true dimensionality of the data if the dataset is not inherently m-dimensional. It is far less time consuming to compare object vectors in r-dimensional space than in n-dimensional space. Another factorization method known as Non-negative Matrix Factorization (Lee & Seung, 1999) constrains the component matrices to non-negative values in order to aid the interpretation of axes (columns) of the component matrices, and has been utilized in biclustering objects and attributes.

A dataset, however, may not be restricted to depicting relationships between two types of entities. There exist several scenarios where there is a need to represent simultaneous interactions amongst 3 or more types of entities. An example would be expression of genes in different body tissues over multiple time points. Such data are not readily amenable to a single matrix representation. A matrix could show 3 separate views of the data: gene $\times$ tissue (showing expression of genes varying amongst different tissues), gene $\times$ time (showing expression of genes varying over time), or tissue $\times$ time (relative activity in different tissues over time). As only two modes (rows and columns) can be represented in a matrix, looking at three modes two at a time entails that the entries across the third mode be aggregated resulting in a loss of information about the mutual dependencies among all three modes. For instance, elevated expression of certain genes in tissues during sleep may not be easy to discover in pairwise analysis of two modes at a time. Multimodal data need not be restricted to 3 modes. An example of 4-mode data would be word frequencies in emails exchanged between two distinct groups of people over discrete time periods.

Tensors are multimodal extensions of matrices (a matrix can be thought of as 2-mode tensor), and tensor factorizations (decompositions) are multiway generalizations of matrix factorizations. They are useful in analyzing the structure of such n-mode ($n \geq 3$) data in an integrated fashion, without first averaging entries and then using matrix factorizations. The aforementioned two kinds of data are more naturally represented as a gene $\times$ tissue $\times$ time tensor (3-mode array), and word $\times$ group-1 $\times$ group-2 $\times$ time tensor (4-mode array). Tensor factorizations can be used for dimensionality reduction and discovery of latent multiway associations just like their matrix factorization counterparts.

TENSOR FACTORIZATION METHODS

Several factorization techniques have been proposed to decompose a tensor into component (factor) matrices. While there are parallels between matrix and tensor factorizations, certain aspects of the former do not hold true in the latter. This section will provide an overview of the two most widely used methods. Kolda and Bader (2009) describe the methods in greater detail.