基于张量的方法及应用综述

夏 虹[1,2],张雅倩[1],靳晓东[1],陈彦萍[1,2],高 聪[1,2],王忠民[1,2]
(1. 西安邮电大学 计算机学院,陕西 西安 710121;
2. 西安邮电大学 陕西省网络数据分析与智能处理重点实验室,陕西 西安 710121)

摘 要:大数据时代的不断发展促使传感及移动互联设备所产生数据的规模和复杂度快速增长,呈现出多源、异构、海量的特点。因此对这些复杂数据的统一表示、降维处理以及缺失值补全等问题受到研究人员的广泛关注。张量具有对高维数据强大的表示和降维能力并能挖掘元素值之间的潜在关系,被普遍应用于这些问题的研究中。张量分解方法获取高维复杂数据的低维特征,在降低计算复杂度的同时还能够保持原有数据的内在结构,解决“维度灾难”问题。张量补全方法根据已有数据的全局结构获取低秩模型来估计缺失条目。该文从张量分解与补全的视角出发,分别总结相关经典方法的基本思想并分析各自的优缺点。从多源异构大数据分析、人脸识别、数据压缩三方面对张量分解的最新算法进行总结。针对 QoS 缺失数据预测、短时交通流量预测、图像恢复三个场景介绍了张量补全的最新应用。最后对未来张量研究发展中可能存在的问题与挑战进行展望。

关键词:统一表示;张量分解;张量补全;降维;缺失值预测
中图分类号:TP31 **文献标识码**:A **文章编号**:1673-629X(2022)06-0001-08
doi:10.3969/j.issn.1673-629X.2022.06.001

Review on Tensor-based Methods and Applications

XIA Hong[1,2],ZHANG Ya-qian[1],JIN Xiao-dong[1],CHEN Yan-ping[1,2],
GAO Cong[1,2],WANG Zhong-min[1,2]
(1. School of Computer Science & Technology,Xi'an University of Posts and Telecommunications,
Xi'an 710121,China;
2. Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing,Xi'an University of
Posts and Telecommunications,Xi'an 710121,China)

Abstract:The continuous development of the big data era has led to a dramatic increase in the scale and complexity of data generated by sensing and mobile Internet devices,showing the characteristics of multi-source,heterogeneous,and massive. Therefore,the unified representation,dimensionality reduction processing,and missing value completion of these complex data have received extensive attention from researchers. Tensors have powerful representation and dimensionality reduction capabilities for high-dimensional data,and can mine potential relationships between element values. They are widely used in the research of these problems. Tensor decomposition method obtains the low-dimensional features of high-dimensional complex data,which can reduce the computational complexity while maintaining the internal structure of original data,and can solve the "dimension disaster" problem. Tensor completion method obtains a low-rank model based on the global structure of the existing data to estimate missing items. From the perspective of tensor decomposition and completion,the basic ideas of related classic approaches as well as their advantages and disadvantages are analyzed. The latest algorithms of tensor decomposition are summarized from three aspects of multi-source heterogeneous big data analysis,face recognition,and data compression. The latest application of tensor completion is introduced from three scenarios of QoS missing data prediction,short-term traffic flow prediction,and image restoration. Finally,the problems and challenges in the future development of tensor research are prospected.

Key words:unified representation;tensor decomposition;tensor completion;dimensionality reduction;missing values prediction

收稿日期:2021-06-22 **修回日期**:2021-10-22
基金项目:陕西省重点研发计划-重点产业创新链-工业领域项目(2019ZDLGY07-08);陕西省自然科学基础研究计划面上项目(2020JM-582);西安市科技计划项目(2019218114 GXRC017CG018-GXYD17.9)
作者简介:夏 虹(1977-),女(蒙古),博士,讲师,CCF 会员(2134OM),研究方向为服务计算、工业大数据服务;张雅倩(1997-),女,硕士研究生,CCF 会员(E1564G),研究方向为工业大数据处理。

0 引 言

传感设备和移动互联设备的普遍应用使得其产生的数据量飞快增长,关系错综复杂[1]。类型的多样化和结构的复杂化导致挖掘数据背后隐藏的信息变得困难,亟需引入一种高效的融合模型对这些多源异构数据进行统一表示。现有的数据融合统一表示方法有本体论、语义网、大图模型等,这些方法虽然在很多领域已经有了广泛的应用,但其表示形式较为简单,难以对高维空间中的跨领域异构数据进行表示。因此,基于张量的异构数据统一表示方法开始被研究人员关注[2]。作为向量和矩阵向高阶的扩展,张量可以保持复杂数据的潜在结构,能够很好的描述和表示高维度、多样化的海量数据。

数据的高维特征使计算处理的时间和空间复杂度急剧增加,随之带来的是“维度灾难”问题,所以在应用之前如何对这些高维数据进行降维处理是一个重要环节。已有的缩减维数的方法有相关系数矩阵、独立成分分析、主成分分析、线性判别分析等,这些方法有一定的降维效果,但是会不可避免的破坏原始数据的内在结构,造成一定的损失。基于张量对高维数据强大的表示能力,一些工作[3-6]引入张量分解来获取这些数据的低维特征,在降低计算复杂度的同时还可以维持原有数据的几何结构,减少损失。

由于各种不可避免的因素,如错误操作、缺少权限等,在实际应用中获取到的多维数据集往往是不完整的,如何利用已获得的数据来估计缺失值是张量补全所要解决的问题。虽然矩阵补全[7-8]在过去的几十年已经得到广泛的研究,含缺失值的张量也可以被展开为矩阵从而通过矩阵补全方法进行恢复。但是这些方法会破坏张量的多维内在结构,并且当引入更高维的大数据时会带来指数级增长的计算复杂度,无法进行高效的数据修补,因此需要利用张量全局结构中的多维信息来补全缺失值。

近年来,张量研究在数据挖掘、图像处理、服务质量预测、交通流量预测等领域得到广泛的关注。基于以上对张量分解和补全问题的简单介绍,下文主要对相关方法及应用进行总结,为同行研究人员提供一个参考,以促进未来的工作和应用。

1 张量分解方法

最先提出的张量分解方法为CP分解和Tucker分解,它们将一个N阶张量分解为若干因子矩阵来减小存储空间并挖掘其核心要素。之后为了增强解释能力并提高计算效率,张量网络作为对传统分解方法的推广被提出,主要包括张量链和张量环,思想是将一个高阶张量表示为一组稀疏的互相连接的低阶张量,从而一个大规模高阶张量优化问题可以被转化为小规模低阶张量的处理问题,从而能够减轻维度灾难的影响。本节介绍CP分解、Tucker分解、张量链分解以及张量环分解的过程及各自的优缺点。

1.1 CP分解

CP分解[9]将高阶张量表示为有限个秩为1的张量之和。三阶张量的CP分解过程如图1所示。对于一个N阶张量,CP分解可表示为:

$$\chi \approx [\lambda;\boldsymbol{A}^{(1)},\boldsymbol{A}^{(2)},\cdots,\boldsymbol{A}^{(N)}] \equiv \sum_{r=1}^{R}\lambda_r a_r^{(1)}\circ a_r^{(2)}\circ a_r^{(N)} \tag{1}$$

其中,$\circ$表示向量外积,$\boldsymbol{R}$是张量的秩且分解后的因子矩阵$\boldsymbol{A}^{(n)} \in R^{I_n\times R}(n=1,2,\cdots,N)$。

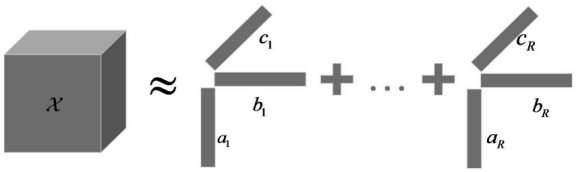

图1 三阶张量$\chi \in \boldsymbol{R}^{I_1\times I_2\times I_3}$的CP分解过程

除了排列以及缩放的不确定性外,CP分解仅存在唯一可能的秩1张量的组合,数学表达形式简单,分解过程简便,可以大规模处理分布式数据集,但张量秩R的定义对分解结果影响较大,很难找到最优解。

1.2 Tucker分解

Tucker分解[10]是将高阶张量表示为一个核心张量和若干个张量模对应的伴随矩阵的乘积。三阶张量Tucker分解过程如图2所示。一个N阶张量的Tucker分解表示为:

$$\chi \approx [\boldsymbol{G};\boldsymbol{A}^{(1)},\boldsymbol{A}^{(2)},\cdots,\boldsymbol{A}^{(N)}] = G\times_1\boldsymbol{A}^{(1)}\times_2\boldsymbol{A}^{(2)}\cdots\times_N\boldsymbol{A}^{(N)} \tag{2}$$

其中,核心张量$\boldsymbol{G}$包含原张量中的主要信息,分解得到的每个因子矩阵$\boldsymbol{A}^{(n)}(n=1,2,\cdots,N)$表示第$n$阶上的主成分。

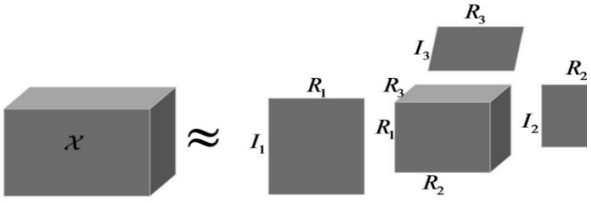

图2 三阶张量$\chi \in \boldsymbol{R}^{I_1\times I_2\times I_3}$ Tucker分解过程

分析以上两种分解过程可以明显看出,CP分解是Tucker分解的一种特殊情况,即当Tucker分解中的核心张量$\boldsymbol{G}$是对角的且每一阶对应的维数相等时,它也就变成了CP分解。与CP分解方法相比,Tucker分解更容易捕获目标张量中的潜在联系,得到的核心张量的存储空间相对于原始张量会大大减小,但其参数的数量与张量阶成指数关系,时间复杂度与核心张量大

小密切相关,因此不适合大规模数据的处理。

1.3　张量链分解

对于张量链分解(TT 分解)[11],其主要思想是使用张量的缩并运算将一连串稀疏的低阶张量(大多是 2 阶或 3 阶)进行互连来表示一个高阶张量,张量核的个数与张量的阶数保持一致。图 3 是一个 N 阶张量的张量链分解过程,其数学表达式为:

$$\chi = \boldsymbol{X}^{(1)} \cdot \boldsymbol{X}^{(2)} \cdot \cdots \cdot \boldsymbol{X}^{(n)} \cdot \cdots \cdot \boldsymbol{X}^{(N)} \tag{3}$$

其中,·表示张量的缩并,即张量的单模乘。$\{r_0, r_1, \cdots, r_N\}$ 代表张量链的秩,且 $r_0 = r_N = 1$。$X^{(n)} \in R^{r_{n-1} \times I_n \times r_n}$($n = 1, 2, \cdots, N$) 称为核心张量(或张量核)。张量链中的元素定义为 $\chi(i_1, i_2, \cdots, i_N) = \boldsymbol{X}^{(1)}(:, i_1, :)\boldsymbol{X}^{(2)}(:, i_2, :)\cdots\boldsymbol{X}^{(N)}(:, i_N, :)$。

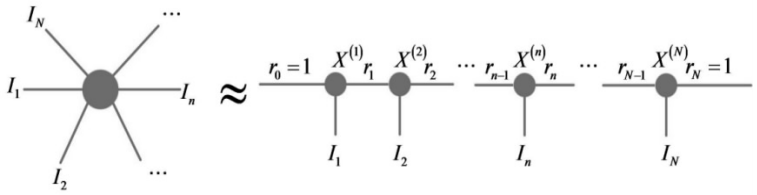

图 3　N 阶张量 $\chi \in \boldsymbol{R}^{I_1 \times \cdots \times I_N}$ 张量链分解过程

TT 分解的原理是通过矩阵的序列乘积来近似张量中的每个元素,其中第一个和最后一个矩阵是向量,以确保标量输出。所以 TT 分解完全基于矩阵的 QR 或 SVD 分解序列,在处理一个高阶张量时是稳定的,且不需要进行任何递归操作,分解后的参数级别和 CP 分解相同,所以它能够解决维度灾难的问题。但是 TT 分解会受到以下几点限制:(1)TT 秩的限制,即 $r_0 = r_N = 1$ 导致了有限的表示能力和灵活性。(2)TT 秩总是有一个固定的模式,即边界核越小,中间核越大,这不是特定张量的最佳模式。(3)张量核的多线性乘积有严格的顺序约束,这使得优化的张量核高度依赖于张量维数的排列,但寻找最优排列是一个较为困难的问题。

1.4　张量环分解

一些学者考虑解决上述张量链分解的局限性。首先,需要针对 TT 秩的限制即 $r_0 = r_N = 1$ 放宽条件,从而提高数据的表示能力。其次,应减轻多线性乘积在张量核之间受排列顺序的影响。第三,通过使模型对称实现对张量核的循环移位和等价处理。为此,文献[12]发现这些目标可以通过使用迹操作来实现。由于这些张量核是循环互连的,看起来像一个环结构,所以这种分解模式被称为是张量环分解(TR 分解)。

TR 分解将一个高阶张量表示为一连串循环相乘的三阶张量。N 阶张量的张量环分解过程如图 4 所示。其数学表达式为:

$$\chi = \mathrm{tr}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots, \boldsymbol{X}^{(n)}, \cdots, \boldsymbol{X}^{(N)}) \tag{4}$$

其中与 TT 分解类似,$\boldsymbol{X}^{(n)} \in \boldsymbol{R}^{r_{n-1} \times I_n \times r_n}$ 称为核心张量,$\{r_0, r_1, \cdots, r_N\}$ 表示张量的秩,不同之处在于 TR 分解的秩约束为 $r_0 = r_N$,而不需要为 1。张量环中的元素可以定义为 $\chi(i_1, i_2, \cdots, i_N) = \mathrm{tr}(\boldsymbol{X}^{(1)}(:, i_1, :)\boldsymbol{X}^{(2)}(:, i_2, :)\cdots\boldsymbol{X}^{(N)}(:, i_N, :))$。

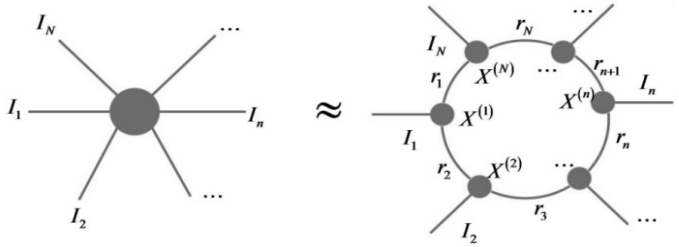

图 4　N 阶张量 $\chi \in \boldsymbol{R}^{I_1 \times \cdots \times I_N}$ 张量环分解过程

TR 分解后得到的张量核都是三阶张量,表示形式更加统一,存储的变量数量较少,其分解格式是张量对张量的分解,因此它可以更好地保持原始数据的内在结构。TR 分解的特点在于每个张量核在迹操作下都可以进行循环移位和等价处理,而其他分解方式未能保持这一优势。

2　张量补全方法

在采集数据的过程中,一些无法避免的外在条件的干扰导致获取的数据往往是不完整的。张量补全用来填补缺失的或未观测到的条目。重构张量分解后的低秩因子也可进行张量补全,但是张量分解目的是从高维数据中提取出一个低秩结构来降低计算复杂度,而张量补全是利用获取到的低秩模型来估计缺失的数据。两者目的的不同限制了一些传统的分解方法的应用,需要对其进行改进,并且这些方法需要手工设置张量秩。张量补全需要明确目标张量中哪些数据是已观测到的,在每次迭代过程中都需要估计丢失的信息。

除了基于分解的方法外，Liu[13] 在矩阵迹范数的基础上对张量迹范数进行定义，将张量秩最小化问题表述为凸优化问题，提出了几种低秩张量补全模型。虽然这些展开矩阵由于具有多线性相关性而不能独立优化，但这些方法可以在预先不设置张量秩的情况下来解决张量补全问题，减轻自定义张量秩的影响，这在实际操作中更易于处理。本节对几种基于分解的和基于迹范数的补全方法进行介绍。

2.1 基于分解的方法

基于分解的方法根据已观测到的条目将目标函数定义为最小化原始张量与分解重构之间的误差，然后通过对目标函数进行多次迭代优化得到理想的潜在因子，利用这些潜在因素来预测缺失条目。下面列举几种基于分解的方法需要优化的问题。

CP 分解需要优化的问题如下：

$$\begin{aligned}&\min_{\chi,A_1,A_2,\cdots,A_N}:\frac{1}{2}\|\chi-\boldsymbol{A}_1\circ\boldsymbol{A}_2\circ\cdots\circ\boldsymbol{A}_N\|_F^2\\&\text{s.t.}:\chi_\Omega=\boldsymbol{S}_\Omega\end{aligned}\tag{5}$$

其中，χ 和 $\boldsymbol{S}$ 是每个维度具有相同大小的 n –模张量，在集合 Ω 中的元素是已给定的。

Tucker 分解需要优化的问题如下：

$$\begin{aligned}&\min_{\chi,\boldsymbol{G},A_1,A_2,\cdots,A_N}:\frac{1}{2}\|\chi-\boldsymbol{G}\times_1\boldsymbol{A}_1\times_2\boldsymbol{A}_2\times_3\cdots\times_N\boldsymbol{A}_N\|_F^2\\&\text{s.t.}:\chi_\Omega=\boldsymbol{S}_\Omega\end{aligned}\tag{6}$$

张量链分解需要优化的问题如下：

$$\begin{aligned}&\min_{\chi,\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N}:\frac{1}{2}\|\chi-\boldsymbol{X}_1\cdot\boldsymbol{X}_2\cdot\cdots\cdot\boldsymbol{X}_N\|_F^2\\&s.t.:\chi_\Omega=\boldsymbol{S}_\Omega\end{aligned}\tag{7}$$

张量环分解需要优化的问题如下：

$$\begin{aligned}&\min_{\chi,X_1,X_2,\cdots,X_N}:\frac{1}{2}\|\chi-\mathrm{tr}(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N)\|_F^2\\&s.t.:\chi_\Omega=S_\Omega\end{aligned}\tag{8}$$

对以上几种优化问题的求解可以分为两种情况。第一种情况是缺失值的输入和模型参数的交叉估计同时完成，通常使用交替投影优化来解决，如 ALS 优化。这种方法操作简单，但如果缺失比率逐渐增加，收敛速度可能会降低。第二种情况是忽略缺失值，只根据观测到的部分数据来建立模型，通常使用梯度优化或概率方法来求解，对以上几个目标函数进行加权优化，在缺失率较高时有很好的补全效果。

2.2 基于迹范数的方法

基于迹范数最小化的方法被称为低秩张量补全，是低秩矩阵补全的推广，定义了张量迹范数，并将基于张量的非凸秩最小化问题定义为凸张量迹范数极小化问题。下面介绍基于迹范数的补全方法需优化的问题及相应求解算法。

基于秩最小化的张量补全模型需要优化的问题如下：

$$\begin{aligned}&\min_{\chi}:\sum_{i=1}^{n}\alpha_i\mathrm{rank}(\chi_{(i)})\\&\text{s.t.}:\chi_\Omega=\boldsymbol{S}_\Omega\end{aligned}\tag{9}$$

其中，$\mathrm{rank}(\chi_{(i)})$ 表示张量 χ 的 i –模秩，在其尽可能小的情况下来确定 χ 中丢失的元素。α_i 是常量，需满足 $\alpha_i\geqslant 0$ 并且 $\sum_{n=1}^{N}\alpha_i=1$。但是函数 $\mathrm{rank}(\chi_{(i)})$ 是非凸的，此问题为非凸优化问题。由于迹范数 $\|\ \|_*$ 是秩函数的凸包络，因此解决该问题一种常见的方法是使用迹范数来近似矩阵的秩，这转化为以下凸优化问题：

$$\begin{aligned}&\min_{\chi}:\sum_{i=1}^{n}\alpha_i\|\chi_{(i)}\|_*\\&\text{s.t.}:\chi_\Omega=\boldsymbol{S}_\Omega\end{aligned}\tag{10}$$

本质上，张量迹范数是沿所有模态展开所得的矩阵迹范数的凸组合。

Liu 等人在文献[13]中提出了几种解决上述迹范数最小化问题的算法。2.2.1 和 2.2.2 节分别介绍了简单低秩张量补全算法和高精度低秩张量补全算法及各自的优缺点。

2.2.1 简单低秩张量补全

引入 n 个辅助矩阵 $\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_n$，可以将优化问题转化如下：

$$\begin{aligned}&\min_{\chi,\boldsymbol{M}_i}:\sum_{i=1}^{n}\alpha_i\|M_i\|_*\\&\text{s.t.}:\chi_{(i)}=M_i,\chi_\Omega=S_\Omega,i=1,2,\cdots,n\end{aligned}\tag{11}$$

由于 $\chi_{(i)}=\boldsymbol{M}_i$ 的约束限制，矩阵迹范数项依然相互依赖难以解决。下式通过引入惩罚项 $\beta_i(i=1,2,\cdots,n)$ 来独立地解决每个子问题。

$$\begin{aligned}&\min_{\chi,\boldsymbol{M}_i}:\sum_{i=1}^{n}\alpha_i\|\boldsymbol{M}_i\|_*+\frac{\beta_i}{2}\|\boldsymbol{M}_i-\chi_{(i)}\|_F^2\\&\text{s.t.}:\chi_\Omega=\boldsymbol{S}_\Omega\end{aligned}\tag{12}$$

这是一个凸但不可微的优化问题，能够使用块坐标下降法来解决，在优化其中一个块变量的同时对其余块变量进行固定，因此将问题(12)转化为对 $\chi,\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_n$ 共 $n+1$ 个块的优化求解。

在其余 n 个块变量都固定的情况下，通过优化求解问题(13)来得到最优 χ。

$$\begin{aligned}&\min_{\chi}:\sum_{i=1}^{n}\frac{\beta_i}{2}\|\boldsymbol{M}_i-\chi_{(i)}\|_F^2\\&\text{s.t.}:\chi_\Omega=\boldsymbol{S}_\Omega\end{aligned}\tag{13}$$

此问题的最优解如下：

$$\chi_{i_1,\cdots,i_n}=\begin{cases}\left(\dfrac{\sum_i \beta_i \operatorname{fold}_i(\boldsymbol{M}_i)}{\sum_i \beta_i}\right)_{i_1,\cdots,i_n}(i_1,\cdots,i_n)\notin\Omega\\ \boldsymbol{S}_{i_1,\cdots,i_n}\qquad (i_1,\cdots,i_n)\in\Omega\end{cases}\tag{14}$$

简单低秩张量补全算法分割矩阵迹范数项之间的依赖关系,使它们能够独立求解,并通过使用块坐标下降法寻找全局最优解。虽然该算法较易实现,但在实际操作过程中收敛速度较慢,并且扩展到一般张量迹范数最小化问题也很困难,不具备通用性。

2.2.2　高精度低秩张量补全

引入 n 个辅助张量 $\{\boldsymbol{M}_i\in\boldsymbol{R}^{I_1\times I_2\times\cdots\times I_n}\}_{i=1}^n$,可以得到问题(10)的等价求解形式如下:

$$\begin{aligned}&\min_{\chi,\boldsymbol{M}_1,\cdots,\boldsymbol{M}_n}:\sum_{i=1}^n\alpha_i\|\boldsymbol{M}_{i(i)}\|_*\\&\text{s.t.}:\chi_\Omega=\boldsymbol{S}_\Omega,\chi=\boldsymbol{M}_i,\ i=1,2,\cdots,n\end{aligned}\tag{15}$$

上式对应的拉格朗日函数如下:

$$\begin{aligned}L(\chi,\boldsymbol{M}_1,\cdots,\boldsymbol{M}_n,\boldsymbol{Y}_1,\cdots,\boldsymbol{Y}_n,\rho)=&\sum_{i=1}^n\alpha_i\|\boldsymbol{M}_{i(i)}\|_*\\&+\langle\chi-\boldsymbol{M}_i,\boldsymbol{Y}_i\rangle+\frac{\rho}{2}\|\boldsymbol{M}_i-\chi\|_F^2\end{aligned}\tag{16}$$

由于交替方向乘子法(ADMM)能够有效解决大规模目标函数中含多个非光滑项的优化问题,因此被用来求解函数(16)的极值。

更新 $\boldsymbol{\chi}$ 需要求解以下优化问题:

$$\begin{aligned}&\min_\chi:\sum_{i=1}^n\frac{1}{2}\|\chi-\boldsymbol{M}_i+\boldsymbol{Y}_i/\rho\|_F^2\\&\text{s.t.}:\chi_\Omega=\boldsymbol{S}_\Omega\end{aligned}\tag{17}$$

最优 χ 可表示如下:

$$\chi_\Omega=\frac{1}{n}\left(\sum_{i=1}^n\boldsymbol{M}_i-\frac{1}{\rho}\boldsymbol{Y}_i\right)_{\bar\Omega}\tag{18}$$

其中,$\bar\Omega$ 是集合 Ω 的补集。

高精度低秩张量补全算法旨在解决没有观测噪声的张量补全问题,使用 ADMM 方法直接处理等式约束,而不采用任何松弛技术,可以更快地实现较高的补全精度。

3　张量分解的应用

目前各个领域所产生的数据规模逐渐变大,呈现出高维的特征。直接操作这些高维数据会导致计算复杂度急剧增加,并且可能出现“维度灾难”的问题。根据第 1 节中几种方法的分析,张量分解的目的是对高维数据进行降维处理和特征提取,解决传统降维方法会破坏内在信息和结构的缺陷。第 1 节介绍的分解方法及推广目前被应用于很多领域,本节主要介绍张量分解方法在多源异构大数据分析[3,4,14]、人脸识别[5,15-16]和数据压缩[6,17-19]这三种应用场景的最新进展。

3.1　多源异构大数据分析

日渐丰富的传感设备及移动互联设备从多个维度进行数据的采集,得到海量的多源异构数据。在数据量快速增长的时代,如何对海量、多源、异构数据进行统一表示以及降维处理是亟需解决的问题。目前基于本体论、语义网、大图模型的表示方法仅在一定领域有好的效果,但不能在多数场景中通用,并且由于数据的高维性会丢失部分特征。为了减少存储空间及后续的计算成本,解决“维度灾难”问题,还需要对这些统一表示后的高维数据进行降维处理以获取高质量的核心数据。传统的主成分分析、因子分析、独立成分分析等方法仅适合处理低维数据,在高维数据处理方面效果不佳并会损坏内部结构,并且不能有效处理增量产生的流式数据。张量模型由于具有良好的统一表示能力和降维效果而在多源异构大数据分析中受到关注。

Kuang[3]首先提出把从多个维度采集到的复杂异构数据使用张量模型进行统一表示,为结构化、半结构化和非结构化数据分别建立子张量,然后利用张量扩展算子进行融合。王在文献[4]中提出了多种基于雅可比正交化的分布式增量高阶奇异值分解方法,包括单模分解方法、树形多模分解方法、基于 RoundRobin 环的分解方法以及嵌入式树形多模分解方法。这些方法用来解决分布式以及增量处理中的张量展开问题。文献[14]还将张量链分解引入大数据增量式降维问题中,分别介绍了基于奇异值分解和 QR 分解的增量式张量链分解方案,仅存储更新后得到的低阶核心张量来减小存储空间并提高处理效率。

3.2　人脸识别

目前,在人们的日常生产生活中越来越多的用到人脸识别技术,如智能门禁、员工打卡、移动支付等。所以人脸识别研究在很多领域变得流行,并且现在已经取得了一些显著成果。人脸图像信息通常以高维的形式呈现,如何从这些高维数据中进行特征提取是人脸识别中的一个关键环节。传统的基于向量的人脸识别方法如 LDA、PCA、LPP 以及基于矩阵的人脸识别方法如 2D-LDA、2D-PCA、2D-LPP 等会破坏人脸图像原有的几何结构,并在降维的过程中会造成信息损失,所以一些研究人员考虑使用张量来建模人脸的彩色图像信息,并基于张量分解方法来减小原始人脸图像的维度数。

由于人脸图像中的数据都是非负的,梁[5]在非负矩阵分解方法的基础上提出了基于非负张量分解的人脸识别方法,能够很好地捕获人脸内部核心信息,识别效率比 NMF 或 PCA 要高。宋等人[15]为了解决非负

张量分解在提取人脸特征时冗余信息太多以及表达形式不够简单的问题,在此基础上添加正交稀疏约束来获得相关性较低的基图像从而能够获得较好的识别效果。为了避免图像增加时的重复运算,文献[16]提出了一种基于随机增量张量奇异值分解的识别算法,将随机张量模型与张量奇异值分解方法相结合,可以在已有的随机奇异值分解结果的基础上来进行下一步的更新运算。

3.3 数据压缩

各个领域所产生数据的急剧增多带来的是如何进行有效存储和传输的问题,数据压缩技术被用来去除原始大规模数据中的冗余信息,在基本不损失核心信息的情况下来减少数据量或者重新组织数据从而提高处理以及存储效率。目前需要处理的数据大都存在着天然的高维结构,传统的基于向量或矩阵的模型难以描述这些高维数据的空间结构和相关性,一些研究提出基于张量模型的方法来解决数据压缩问题,降低时间和空间复杂性,提升压缩性能。

在视频图像压缩研究中,李[6]将视频数据用张量模型进行表示,并使用张量迭代 Tucker-ALS 算法来减少原有视频数据的维数,取得了很好的压缩效果。针对配电网大数据的压缩研究,赵[17]为了保持高维数据的空间结构并解决海量异构配电网数据的存储问题,利用张量模型对智能配电网异构数据进行统一表示并提出了基于 Tucker 分解的多维配电网数据压缩方法。在此基础上,赵[18]还利用配电系统中数据量大且随时间积累的特点,提出了一种基于增量张量分解的配电网流数据压缩方法,使用实时新增数据来更新历史压缩结果。在高光谱图像的压缩问题中,Li[19]考虑高光谱图像不同维度之间的关联,提出了一种基于相关的 Tucker 分解方法来分别构造核心张量和因子矩阵,可以被应用于任何基于 Tucker 分解的 n 阶张量中,压缩性能得到提升。

4 张量补全的应用

在获取高维复杂数据的过程中,由于各种不可预测的情况,如错误操作、缺少权限等,其中某些元素会丢失,影响后续应用。所以如何捕获已有元素和缺失元素之间的潜在联系,利用已获取的数据预测缺失值成为一个亟需解决的问题。第 2 节介绍的张量补全及改进方法已被应用到很多领域的缺失值处理中。本节基于 QoS 缺失数据预测[20-23]、短时交通流量预测[24-27]、图像恢复[28-31]三个应用场景对张量补全进展进行探讨。

4.1 QoS 缺失数据预测

服务数量的急剧增多使得用户在过去只能调用有限的服务并获取对应的 QoS 值。但依据稀疏的 QoS 数据很难进行准确的服务推荐或组合,所以 QoS 缺失值预测是进行服务相关操作的重要环节。基于矩阵分解的 QoS 预测方法已经取得很好的效果,但其只用到用户和服务的二维信息,预测效果受限。在动态的互联网环境下,受服务器负载限制、网络状态波动、用户移动性需求等因素的影响,时间、位置和一些其它因素会导致 QoS 值发生变化,所以需要尽可能利用多维数据来预测未知 QoS 值。一些研究工作使用张量模型对高维 QoS 数据建模以考虑多个维度的信息来提高预测精度。

为了捕获隐藏在时间模式下的潜在信息,Zhang[20]首先将基于用户和服务的静态矩阵模型扩展到三维的用户-服务-时间动态张量模型,并考虑 QoS 数据的非负性,使用非负 CP 分解算法来处理模型中的三元关系。除了 CP 分解,Zhang[21]还将 Tucker 分解应用于 QoS 缺失数据预测问题,为了应对 QoS 数据流动态传入所带来的挑战,还引入了一种基于奇异值分解的增量张量分解方法来减少存储空间并提高计算效率。文献[22]根据用户和服务的位置将已有的 QoS 数据分别建模为局部和全局三阶张量,将位置聚类和分层张量分解相结合来实现 QoS 预测,该方法能有效提高预测精度并缓解数据稀疏性问题。上述只针对某一个维度进行预测,而没有考虑多维度(位置、时间等)之间的相关性,马[23]利用张量对多维度的 QoS 数据进行建模,通过高效求解张量分解优化算法来实现好的预测效果。

4.2 短时交通流量预测

为了缓解城市中的交通拥堵的问题,帮助出行者更好地选择出行线路,节省时间,短时交通流量预测是智能交通系统中一项重要的任务。已有的方法考虑交通数据的时间变化特性、空间相似性以及周期性进行预测,但这些方法大多基于向量或矩阵形式,数据维度的约束导致难以同时描述交通数据多模式之间的潜在联系,预测效果会受到一定的限制。所以张量开始被引入到短时交通流量预测问题中来封装交通流量数据。文献[32]将采集到的交通数据构建为天-小时-时间三阶张量的形式,在对丢失的交通数据进行补全时,获得了比向量和矩阵模型更好的效果。但在短时交通流量预测问题中,对未来数据的预测需要用到实时获得的数据,并随时间的推动来选择预测区间,静态交通数据补全方法难以描述交通数据的动态变化情况,所以研究者往往将交通数据构建为动态张量形式,不仅保留了原始数据的多模式关系,还能体现交通流的动态特征。

段[24]根据交通数据的多模式特征构建 Time ×

Day × Week × Location 四阶张量,并引入滑动窗口来选取固定长度的张量流作为历史数据,介绍了基于 CP 和 Tucker 分解的动态短时交通流量预测方法。除了基于这两种分解方式的预测方法之外,耦合张量分解也被应用到短时交通流量预测问题中。Zhou 等人[25]选取交通流数据和平均车速数据分别构造两个张量并以“路段”模式耦合,提出了一种基于矩阵和张量的加权优化算法来恢复交通流量数据,这是第一次为交通数据处理应用引入一个新的代数框架,不同于传统的单张量模型。为了进一步改进预测效果,文献[26]提出了一种基于快速低秩张量补全的增量启发式交通流量预测方法,该模型将增量张量结构与快速低秩张量补全相结合,集成了交通流数据日、周、时间、空间等多模式的特征。Li 等人[27]探讨了在各种常见的缺失场景以及不同的缺失率下几种经典的张量补全算法是否以及如何影响最终的预测精度,这是第一篇分析丢失数据及其补全对交通流预测的详细影响的论文。

4.3　图像恢复

实际应用中的图像大多都存在高维特征,如高光谱图像、医疗图像、视频等,这些多维图像中包含着比传统二维图像更加丰富的信息,但在生成、存储与传输的过程中会受到很多外界条件的影响而不能准确处理后续应用,对这些高维图像的处理问题大多是根据已观测到的数据来进行高质量的图像填充,将其还原到原来的真实图像结构。目前基于矩阵补全的方法要求图像行列之间有高度相关性,并且在展开为向量之后会破坏图像的内部结构,所以一些研究将张量补全方法应用于图像恢复问题中。

在高光谱图像超分辨率研究中,由于需要将低空间分辨率的高光谱图像和高空间分辨率的多光谱图像进行融合来捕获两者互补的因素,耦合张量分解方法常被应用至该问题的研究中。Xu[28]提出了一种基于耦合 CP 分解的方法来探讨高光谱和多光谱图像之间的关系。Xu 还使用耦合张量环分解来探讨图像的非局部自相关性和全局谱相关性,将两类图像融合表示为耦合张量环模型从而共享其中的潜在核张量之间的关系[29]。在医疗图像处理方面,文献[30]针对仅使用单一秩来解决动态 MRI 重建问题的缺陷,考虑同时最小化 CP 和 Tucker 秩来更好地利用动态 MRI 数据低秩分量的相关性。Cui 等人[31]将基于分层概率模型的 CP 分解方法用于 EEG 缺失数据的恢复问题中,张量秩能够被自动确定,而不用手工赋值。

5　存在问题与挑战

虽然张量分解及补全方法在很多领域已经取得了很好的成果,但在复杂的应用场景下还存在一些问题与挑战需要进行深入的研究。(1)大多数已有的张量分解算法都是在单机模式下运行的,但随着大数据时代的发展,传统单机模式已经无法应对复杂的海量数据,为了提高计算效率并保持分解精度,可以将已有的算法扩展到分布式环境下运行。(2)如何进一步提高张量分解的降维效率和张量补全的预测精度并将其应用于不同的场景是需要一直进行探索的。(3)将张量模型与新兴领域如边缘计算、物联网、车联网或者新兴技术如深度学习、神经网络等相结合来高效解决更多的问题也具有重要意义。相信在之后的研究发展中张量分解及补全相关算法会有更大的发展前景,能够在最大程度上发挥张量模型的应用价值。

6　结束语

张量模型具有对高维数据强大的表示和处理能力,该文总结了张量分解及补全中的几种经典方法,其中张量分解方法包括 CP 分解、Tucker 分解、张量链分解以及张量环分解,分析这些方法的基本思想及优缺点。张量补全方法从基于分解和基于迹范数两方面进行介绍,基于分解的方法利用张量分解得到的低秩结构来估计缺失值,基于迹范数的方法需要解决迹范数最小化的优化问题。接着从多源异构大数据分析、人脸识别、数据压缩三方面来总结张量分解的最新应用。针对 QoS 缺失数据预测、短时交通流量预测、图像恢复三个场景来介绍张量补全的最新算法。最后对张量分解及补全方法研究中存在的问题与挑战进行展望。

参考文献:

[1] 马　茜,谷　略,张天成,等. 一种基于数据质量的异构多源多模态感知数据获取方法[J]. 计算机学报,2013,36(10):2120-2131.

[2] WANG W,ZHANG M. Tensor deep learning model for heterogeneous data fusion in internet of things[J]. IEEE Transactions on Emerging Topics in Computational Intelligence,2020,4(1):32-41.

[3] KUANG L,HAO F,YANG L T,et al. A tensor-based approach for big data representation and dimensionality reduction[J]. IEEE Transactions on Emerging Topics in Computing,2017,2(3):280-291.

[4] 王晓康. 张量分解的高效计算及其应用研究[D]. 武汉:华中科技大学,2017.

[5] 梁秋霞,何光辉,陈如丽,等. 基于非负张量分解的人脸识别算法研究[J]. 计算机科学,2016,43(10):312-316.

[6] 李鹏程. 基于张量紧凑表示的视频压缩算法[J]. 电子科技,2017,30(5):1-4.

[7] GENG J,WANG L,WANG X. Nuclear norm and indicator function model for matrix completion[J]. Journal of Inverse and Ill-posed Problems,2016,24(1):67-77.

[8] 史加荣,李金红. 融合矩阵补全与深度矩阵分解的推荐算法[J]. 计算机应用研究,2021,38(8):2376-2380.
[9] KIERS H. Towards a standardized notation and terminology in multiway analysis[J]. Journal of Chemometrics,2000,14(3):105-122.
[10] LATHAUWER L D,MOOR B D,VANDEWALLE J. Multilinear singular value tensor decompositions[J]. SIAM Journal on Matrix Analysis and Applications,2000,24(4):1253-1278.
[11] OSELEDETS I V. Tensor-train decomposition[J]. Siam Journal on Scientific Computing,2011,33(5):2295-2317.
[12] WANG W,AGGARWAL V,AERON S. Efficient low rank tensor ring completion[C]//IEEE international conference on computer vision. Venice:IEEE,2017:5698-5706.
[13] LIU J,MUSIALSKI P,WONKA P,et al. Tensor completion for estimating missing values in visual data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,35(1):208-220.
[14] CHEN Y,JIN X,XIA H,et al. Incremental QR-based tensor-train decomposition for industrial big data[J]. The Journal of China Universities of Posts and Telecommunications,2021,28(1):10-23.
[15] 宋 珊,冯 岩,徐常青. 基于正交稀疏约束非负张量分解的人脸识别算法[J]. 运筹学学报,2021,25(2):55-66.
[16] 邱子衿,陈 潇,贾志刚. 随机增量张量奇异值分解与人脸识别新算法[J]. 聊城大学学报:自然科学版,2019,32(3):23-35.
[17] 赵洪山,马利波. 基于张量 Tucker 分解的智能配电网大数据压缩[J]. 中国电机工程学报,2019,39(16):4744-4752.
[18] ZHAO H,MA L. Power distribution system stream data compression based on incremental tensor decomposition[J]. IEEE Transactions on Industrial Informatics,2020,16(4):2469-2476.
[19] LI R,PAN Z,WANG Y,et al. The correlation-based tucker decomposition for hyperspectral image compression[J]. Neurocomputing,2021,419:357-370.
[20] ZHANG W,SUN H,LIU X,et al. Temporal QoS-aware web service recommendation via non-negative tensor factorization[C]//International conference on world wide web. Seoul:ACM,2014:585-596.
[21] ZHANG W,SUN H,LIU X,et al. An incremental tensor factorization approach for web service recommendation[C]//IEEE international conference on data mining workshop. Shenzhen:IEEE,2014:346-351.
[22] CHENG T,WEN J,XIONG Q,et al. Personalized web service recommendation based on qos prediction and hierarchical tensor decomposition[J]. IEEE Access,2019,7:62221-62230.
[23] 马 友. 基于 QoS 缺失数据预测的个性化 Web 服务推荐方法研究[D]. 北京:北京邮电大学,2015.
[24] 段 炜. 基于张量实现的短期交通流量预测[D]. 长春:吉林大学,2018.
[25] ZHOU W,ZHENG H,FENG X,et al. A multi-source based coupled tensors completion algorithm for incomplete traffic data imputation[C]//2019 11th international conference on wireless communications and signal processing (WCSP). Chennai:IEEE,2019:1-6.
[26] LIAO J,TANG J,ZENG W,et al. Efficient and accurate traffic flow prediction via incremental tensor completion[J]. IEEE Access,2018,6:36897-36905.
[27] LI Q,TAN H,WU Y,et al. Traffic flow prediction with missing data imputed by tensor completion methods[J]. IEEE Access,2020,8:63188-63201.
[28] XU Y,WU Z,CHANUSSOT J,et al. Nonlocal coupled tensor CP decomposition for hyperspectral and multispectral image fusion[J]. IEEE Transactions on Geoscience and Remote Sensing,2020,58(1):348-362.
[29] XU Y,WU Z,CHANUSSOT J,et al. Hyperspectral images super-resolution via learning high-order coupled tensor ring representation[J]. IEEE Transactions on Neural Networks and Learning Systems,2020,31(11):4747-4760.
[30] WU S,LIU Y,LIU T,et al. Multiple low-ranks plus sparsity based tensor reconstruction for dynamic MRI[C]//2018 IEEE 23rd international conference on digital signal processing (DSP). Shanghai:IEEE,2018:1-5.
[31] CUI G,GUI L,ZHAO Q,et al. Bayesian CP factorization of incomplete tensor for EEG signal application[C]//IEEE international conference on fuzzy systems (FUZZ-IEEE). Vancouver:IEEE,2016:2170-2173.
[32] TAN H,FENG J,FENG G,et al. Traffic volume data outlier recovery via tensor model[J]. Mathematical Problems in Engineering,2013,1151(10):164810.1-164810.8.