doi:10.11887/j.cn.201805022 http://journal.nudt.edu.cn

网络异构信息的张量分解聚类方法*

吴继冰,黄宏斌,邓 苏
(国防科技大学 系统工程学院 信息系统工程重点实验室,湖南 长沙 410073)

摘 要:提出基于张量分解的聚类算法,能够同时处理网络中多类型、多语义关系的异构信息。网络信息体系中的各种异构信息被建模为一个多维张量,异构信息之间丰富的语义关系建模为张量中的元素。提出有效的张量分解方法,将不同类型的信息对象一次性划分到不同的簇中。在人工合成的数据集和真实数据集上的实验结果表明:该聚类方法可以很好地处理网络信息体系中的异构信息聚类问题,并且性能优于现有的聚类方法。

关键词:聚类;异构信息;张量分解;信息网络
中图分类号:TN95 **文献标志码**:A **文章编号**:1001-2486(2018)05-146-07

Tensor decomposition based clustering method for heterogeneous information in networks

WU Jibing, *HUANG Hongbin*, *DENG Su*
(Science and Technology on Information Systems Engineering Laboratory, College of Systems Engineering, National University of Defense Technology, Changsha 410073, China)

Abstract: A tensor decomposition based clustering method was proposed for heterogeneous information in networks. This clustering method can cluster multiple types of objects and rich semantic relationships simultaneously. The multi-types of information objects in networks were modeled as a high-dimensional tensor, and the rich semantic relationships among different types of objects were modeled as elements in the tensor. Based on an effective tensor decomposition method, the multi-types of objects were partitioned into different clusters simultaneously. The experimental results on both synthetic datasets and real-world dataset show that the proposed clustering method can deal with the heterogeneous information in networks well, and can outperform the state-of-the-art clustering algorithms.

Key words: clustering; heterogeneous information; tensor decomposition; information networks

信息的处理和分析在现代网络信息体系中占有重要的地位,扮演着网络信息体系中的大脑角色。信息的高效分析处理决定了网络信息体系的智能化程度。通过各种途径汇聚而来的信息,只有经过高效的分析挖掘才能更好地服务于用户。而在信息分析处理中,聚类分析是信息分析挖掘的一种重要手段。聚类分析是挖掘数据中蕴含的语义信息及拓扑结构的有效方法。

网络信息体系中包含着语义丰富、来源多样的各类信息,而这些信息对象之间存在着丰富的语义和交互关系。也就是说,网络信息体系中包含着丰富的异构信息对象。传统的聚类方法往往忽略了这些信息对象之间的异构性,采用统一的度量方法来计算异构信息之间的距离或者相似度。显然,这种方法对于异构信息对象是不适用的。例如,在电子商务网络中,包含着庞大的用户信息和商品信息,这两种信息对象之间又存在着收藏、购买等语义信息。无法直接度量某一个用户与某一件商品之间的相似性。

在科学论著发表网络中,存在的异构信息包括:作者、论文、期刊、会议、主题等。这些异构对象之间存在着复杂的语义关系,例如作者著作论文、论文发表于期刊、作者参加会议、论文包含主题词等。传统的聚类方法试图将不同类型的对象转换到统一的欧氏空间,且只能聚类某一种类型的对象,如作者,而将其他类型的对象经过转换、归一化后作为属性来度量不同作者之间的相似度。这种转换忽视了不同类型对象之间的语义差别,就必然导致信息的丢失。

由于无法直接度量不同类型对象或者不同类

* **收稿日期**:2017-06-02
基金项目:国家自然科学基金资助项目(61401482, 61401483)
作者简介:吴继冰(1987—),男,江苏南通人,博士研究生,E-mail:wujibing@nudt.edu.cn;
邓苏(通信作者),男,教授,博士,博士生导师,E-mail:sudeng@sohu.com

型关系之间的相似度,异构信息的聚类要比单一类型对象的聚类困难许多。最近几年里,有越来越多的学者开始关注于网络异构信息的聚类方法研究。Sun 等将聚类和排序算法相结合,提出了 RankClus[1] 和 NetClus[2] 算法对异构信息进行挖掘分析。这些算法假设网络中的异构信息符合某些特定的网络模式,在此基础上获得了很好的实验结果。其中,RankClus 只能处理二元网络模式,NetClus 只能处理星型网络模式。最近的一项研究成果 FctClus[3] 在计算速度和准确度上都获得了较大的提升,但是与 NetClus 类似,FctClus 只能处理星型网络模式的异构信息。

还有部分学者提出了基于元路径的聚类方法。元路径[4] 是一条定义在网络模式上的连通路径,表达了两个信息对象之间的某种复合语义关系。PathSim[4] 算法是一种基于元路径的信息对象相似度度量方法。文献[5 -7]将 PathSim 与用户的先验知识相结合,提出 PathSelClus 算法对网络中的异构信息进行聚类分析,但是该算法需要用户给定每个簇的种子对象作为聚类的起始条件。

1　问题描述

首先介绍一些关于张量的概念和符号,关于张量代数的详细内容可以参考文献[8 -9]。张量,即多维数组;张量的阶或者称为张量的模是张量具有的维度的数量。

用到的张量代数的符号见表 1。

表 1　张量代数符号描述

Tab. 1　Description of symbols

符号	说明
$\boldsymbol{A}$	矩阵(黑体大写字母)
$\mathcal{X}$	张量(花体字母)
$x_{i_1,i_2,\cdots,i_N}$	N 维张量 $\mathcal{X}$ 的元素
$\mathcal{X}_{(n)}$	张量 $\mathcal{X}$ 沿着 n 维的矩阵化
$\vec{\mathcal{X}}$	张量 $\mathcal{X}$ 的向量化,$\vec{\mathcal{X}}\equiv\mathcal{X}_{(\varnothing)}$
$\mathcal{X}*\mathcal{Y}$	两个相同维度张量的 Hadamard 积
$\boldsymbol{A}\otimes\boldsymbol{B}$	两个矩阵的 Kronecker 积
$\|\mathcal{X}\|_F$	张量的 Frobenius 范数
$\mathcal{X}\times_n\boldsymbol{U}$	张量与矩阵的 Mode-n 矩阵积

Tucker 分解是张量分析中最流行的一种模型[10]。给定一个 N 阶张量 $\mathcal{X}\in\mathbb{R}^{I_1\times I_2\times\cdots\times I_N}$,Tucker 分解是将 $\mathcal{X}$ 分解为一个指定规模的核张量 $\mathcal{G}\in\mathbb{R}^{J_1\times J_2\times\cdots\times J_N}$,$J_n\leqslant I_n$ 和一系列的因子矩阵 $\boldsymbol{U}^{(n)}\in\mathbb{R}^{I_n\times J_n}$,$n=1,2,\cdots,N$ 的矩阵积,即 $\mathcal{X}\approx\mathcal{G}\times_1\boldsymbol{U}^{(1)}\times_2\boldsymbol{U}^{(2)}\times_3\cdots\times_N\boldsymbol{U}^{(N)}\equiv[\![\mathcal{G};\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\cdots,\boldsymbol{U}^{(N)}]\!]$。

目前大部分关于 Tucker 分解的求解方法都是基于因子矩阵正交的假设,即 $\boldsymbol{U}^{(n)}\boldsymbol{U}^{(n)\mathrm{T}}=\boldsymbol{U}^{(n)\mathrm{T}}\boldsymbol{U}^{(n)}=\boldsymbol{E}$,$n=1,2,\cdots,N$,$\boldsymbol{E}$ 为单位矩阵。

假设在网络信息体系中存在着 T 种信息对象类型 $V=\{V_t\}_{t=1}^{T}$,各种信息对象之间存在着 S 种语义交互关系 $E=\{R_s\}_{s=1}^{S}$。那么整个网络信息体系可以用一个图 $S=(V,E)$ 来描述。图 S 也称为网络信息体系的模式,它是该网络信息体系的一个元模版。一个网络模式 $S=\{V,E\}$ 表示了网络信息体系中各种对象类型及不同类型对象之间可能存在的边的类型。图 1 给出了计算机科学文献库(Digital Bibliography & Library Project, DBLP)网络的模式,其中 DBLP 是著名的计算机领域内的科学论著发表网络,也是一个典型的星型网络模式。

记每一种类型的信息对象 V_t 表示为 $V_t=\{v_n^t\}_{n=1}^{N_t}$,其中 N_t 是类型 V_t 中的对象总数,即 $N_t=|V_t|$,$t=1,2,\cdots,T$。网络中的对象总数记为 $N=\sum_{t=1}^{T}N_t$。

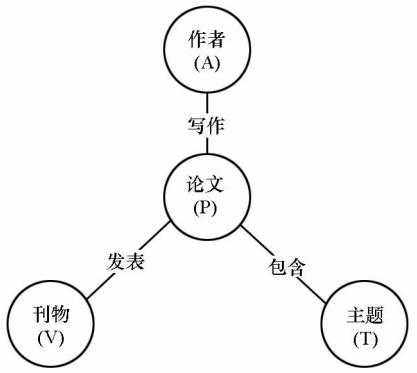


图 1　DBLP 网络的网络模式

Fig. 1　Network schema of DBLP

定义　基因网络。网络信息体系中所有符合网络模式 $S=\{V,E\}$ 的子网络中最小的子网络称为基因网络,记为 $G=(V,E)$。

显而易见,基因网络是所有子网络集合中 $S=\{V,E\}$ 的一个最小实例。例如,DBLP 网络中包含四种类型的对象 $\{A,P,V,T\}$,它的一个基因网络记为 $G'=(\{v_i^A,v_j^p,v_m^V,v_n^T\},\{\langle v_i^A,v_j^p\rangle,\langle v_j^p,v_m^V\rangle,\langle v_m^V,v_n^T\rangle\})$,表示一条语义关系:一个作者 v_i^A 写了一篇论文 v_i^A,发表在刊物 v_m^V 上,这篇论文包含了主题词 v_n^T。简单起见,可以用 G' 中每个对象

的下标来标记这个基因网络，例子中的基因网络 G 可以表示为 $G_{i,j,m,n}$。

令 $\mathcal{X}$ 表示一个规模为 $N_1 \times N_2 \times \cdots \times N_T$ 的 T 维张量，$\mathcal{X}$ 的每一个维度表示网络信息体系中的一种信息对象类型。任意元素 $x_{n_1 n_2 \cdots n_T} \in \{0,1\}$，$n_t = 1,2,\cdots,N_t$ 标识了对应的基因网络 $G_{n_1,n_2,\cdots,n_T}$ 是否存在，即

$$x_{n_1 n_2 \cdots n_T} = \begin{cases} 1, \text{if } \exists\, G_{n_1,n_2,\cdots,n_T} \\ 0, \text{otherwise} \end{cases} \tag{1}$$

从而网络信息体系可以建模为一个张量 $\mathcal{X}$。基于张量分解的方法可以将网络中的信息对象划分到不同的簇中去。令 $\boldsymbol{U}^{(t)} \in \mathbb{R}^{N_t \times K}$ $(t = 1,2,\cdots,T)$ 为第 t 种对象类型的簇标记矩阵，即元素 $u_{i,k}^{(t)} \in \boldsymbol{U}^{(t)}$ $(i = 1,2,\cdots,N_t; t = 1,2,\cdots,T; k = 1,2,\cdots,K)$ 是类型 V_t 中的第 i 个对象 v_i^t，划分到第 k 个簇的概率。一个小规模的张量 $\mathcal{G} \in \mathbb{R}^{\overbrace{K \times K \times \cdots \times K}^{T}}$ 作为张量各个维度和簇指示矩阵之间的调节系数。令 $\mathcal{G}$ 为核张量，$\boldsymbol{U}^{(t)}$ $(t = 1,2,\cdots,T)$ 为特征矩阵，从而可以利用 $[\![\mathcal{G};\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\cdots,\boldsymbol{U}^{(T)}]\!]$ 来逼近 $\mathcal{X}$。异构信息聚类问题可以形式化为一个类似 Tucker 分解的模型：

$$\min_{\mathcal{G},\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\cdots,\boldsymbol{U}^{(T)}} \left\| \mathcal{X} - [\![\mathcal{G};\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\cdots,\boldsymbol{U}^{(T)}]\!] \right\|_{\mathrm{F}}^2$$

$$\text{s.t.} \begin{cases} \forall t, \sum_{k=1}^{K} u_{ik}^{(t)} = 1 \\ \forall t, \forall i, \forall k, u_{ik}^{(t)} \in [0,1] \\ \forall t, \operatorname{rank}(\boldsymbol{U}^{(t)}) = K \end{cases} \tag{2}$$

其中：$i = 1,2,\cdots,N_t; t = 1,2,\cdots,T; k = 1,2,\cdots,K$，并且 $K \leqslant \min\{N_1,N_2,\cdots,N_T\}$ 为簇的总数。

异构信息聚类不需要特征矩阵 $\{\boldsymbol{U}^{(t)}\}_{t=1}^{\mathrm{T}}$ 为正交矩阵。式(2)中的第一个约束条件保证了每个对象属于所有簇的概率之和为1；第二个约束条件保证概率的范围是[0,1]；最后一个约束确保了每一个因子矩阵都是列满秩的，也就是说，对于张量的每一阶而言，没有空簇，也没有任意两个簇是相同的。

2 算法

提出的异构信息聚类算法由两部分组成：特征矩阵更新和核张量更新。本节利用到的张量代数的知识和特性请参考文献[9]。

2.1 特征矩阵更新阶段

每一个特征矩阵 $\boldsymbol{U}^{(t)}$ 依次进行更新，保持核张量和其他特征矩阵不变。式(2)中的目标函数可以写为 $\mathcal{X}$ 沿着第 t 维的矩阵化形式。

$$\min_{\boldsymbol{U}^{(t)}} \left\| \mathcal{X}_{(t)} - \boldsymbol{U}^{(t)} [\![\mathcal{G};\boldsymbol{U}^{(1)},\cdots,\boldsymbol{U}^{(t-1)},\boldsymbol{U}^{(t+1)},\cdots,\boldsymbol{U}^{(T)}]\!]_{(t)} \right\|_{\mathrm{F}}^2 \tag{3}$$

式中，$\boldsymbol{X}_{(t)} \in \mathbb{R}^{N_t \times (N_1 \times \cdots \times N_{t-1} \times N_{t+1} \times \cdots \times N_T)}$。

假设最优解 $\boldsymbol{U}^{(t)}$ 满足式(2)中的所有约束，则式(3)可以转化为：

$$\begin{aligned} \mathcal{X}_{(t)} &= \boldsymbol{U}^{(t)} [\![\mathcal{G};\boldsymbol{U}^{(1)},\cdots,\boldsymbol{U}^{(t-1)},\boldsymbol{U}^{(t+1)}\cdots,\boldsymbol{U}^{(T)}]\!]_{(t)} \\ &= \boldsymbol{U}^{(t)} \mathcal{G}_{(t)} (\boldsymbol{U}^{(T)} \otimes \cdots \otimes \boldsymbol{U}^{(t+1)} \otimes \boldsymbol{U}^{(t-1)} \otimes \cdots \otimes \boldsymbol{U}^{(1)})^{\mathrm{T}} \end{aligned} \tag{4}$$

记：

$$\mathcal{S}_{(t)} = \mathcal{G}_{(t)} (\boldsymbol{U}^{(T)} \otimes \cdots \otimes \boldsymbol{U}^{(t+1)} \otimes \boldsymbol{U}^{(t-1)} \otimes \cdots \otimes \boldsymbol{U}^{(1)})^{\mathrm{T}} \tag{5}$$

式(4)转化为一个类似于文献[11-12]中的非负矩阵分解(Nonnegative Matrix Factorization，NMF)问题，即

$$\mathcal{X}_{(t)} = \boldsymbol{U}^{(t)} \mathcal{S}_{(t)} \tag{6}$$

式(6)与 NMF 具有相同的形式。因此，文献[12]中的 NMF 更新方法可以用来求解 $\boldsymbol{U}^{(t)}$：

$$\boldsymbol{U}^{(t)} \leftarrow \boldsymbol{U}^{(t)} * \frac{\mathcal{X}_{(t)} \mathcal{S}_{(t)}^{\mathrm{T}}}{\boldsymbol{U}^{(t)} \mathcal{S}_{(t)} \mathcal{S}_{(t)}^{\mathrm{T}}} \tag{7}$$

由式(7)得到的特征矩阵并不满足式(2)中的第一和第二个约束条件。为了满足这两个约束条件，可以对特征矩阵进行正规化处理：

$$u_{i,k}^{(t)} \leftarrow \frac{u_{i,k}^{(t)}}{\sum_{k=1}^{K} u_{i,k}^{(t)}} \tag{8}$$

2.2 核张量更新阶段

保持所有特征矩阵不变，式(2)中的目标函数可以变换为：

$$\begin{aligned} &\left\| \mathcal{X} - [\![\mathcal{G};\boldsymbol{U}^{(1)},\boldsymbol{U}^{(2)},\cdots,\boldsymbol{U}^{(T)}]\!] \right\|_{\mathrm{F}}^2 \\ &= \left\| \vec{\mathcal{X}} - (\boldsymbol{U}^{(T)} \otimes \cdots \otimes \boldsymbol{U}^{(1)}) \vec{\mathcal{G}} \right\|_{\mathrm{F}}^2 \end{aligned} \tag{9}$$

由式(9)可得以下线性方程

$$\vec{\mathcal{X}} = (\boldsymbol{U}^{(T)} \otimes \cdots \otimes \boldsymbol{U}^{(1)}) \vec{\mathcal{G}} \tag{10}$$

记：

$$\boldsymbol{Q} = \boldsymbol{U}^{(T)} \otimes \cdots \otimes \boldsymbol{U}^{(1)} \tag{11}$$

因此，式(10)可以转换为一个 NMF 模型，即

$$\vec{\mathcal{X}} = \boldsymbol{Q} \vec{\mathcal{G}} \tag{12}$$

从而，根据文献[12]中的 NMF 更新方法来更新 $\vec{\mathcal{G}}$。

$$\vec{\mathcal{G}} \leftarrow \vec{\mathcal{G}} * \frac{\boldsymbol{Q}^{\mathrm{T}} \vec{\mathcal{X}}}{\boldsymbol{Q}^{\mathrm{T}} \boldsymbol{Q} \vec{\mathcal{G}}} \tag{13}$$

根据式(11)和式(13)，得

$$\begin{aligned} \vec{\mathcal{G}} &\leftarrow \vec{\mathcal{G}} * \frac{\overrightarrow{[\![\mathcal{X};(\boldsymbol{U}^{(1)})^{\mathrm{T}},\cdots,(\boldsymbol{U}^{(T)})^{\mathrm{T}}]\!]}}{\overrightarrow{[\![\mathcal{G};(\boldsymbol{U}^{(1)})^{\mathrm{T}}\boldsymbol{U}^{(1)},\cdots,(\boldsymbol{U}^{(T)})^{\mathrm{T}}\boldsymbol{U}^{(T)}]\!]}} \\ &= \mathcal{G} * \frac{\overrightarrow{[\![\mathcal{X};(\boldsymbol{U}^{(1)})^{\mathrm{T}},\cdots,(\boldsymbol{U}^{(T)})^{\mathrm{T}}]\!]}}{[\![\mathcal{G};(\boldsymbol{U}^{(1)})^{\mathrm{T}}\boldsymbol{U}^{(1)},\cdots,(\boldsymbol{U}^{(T)})^{\mathrm{T}}\boldsymbol{U}^{(T)}]\!]} \end{aligned} \tag{14}$$

关于 Kronecker 积和向量化操作的特性可以参考文献[13]。根据式(14)可得核张量 $\mathcal{G}$ 的更新方法。

$$\mathcal{G} \leftarrow \mathcal{G} * \frac{[\![\mathcal{X}; (\boldsymbol{U}^{(1)})^{\mathrm{T}}, \cdots, (\boldsymbol{U}^{(T)})^{\mathrm{T}}]\!]}{[\![\mathcal{G}; (\boldsymbol{U}^{(1)})^{\mathrm{T}}\boldsymbol{U}^{(1)}, \cdots, (\boldsymbol{U}^{(T)})^{\mathrm{T}}\boldsymbol{U}^{(T)}]\!]} \tag{15}$$

基于张量分解的聚类算法的伪代码见算法 1。

算法 1　基于张量分解的聚类

Alg. 1　Tensor decomposition based on clustering

已知:$\mathcal{X}, K, \{\boldsymbol{U}^{(t)}\}_{t=1}^{\mathrm{T}}$ 和 $\mathcal{G}$ 的初始值,收敛阈值 ε

```
repeat
    for t←1 to T:
        Update U^(t) according to (7);
        Normalize U^(t) according to (8);
    end for
    Update G according to (15);
until ‖X − [[G; U^(1), U^(2), ..., U^(T)]]‖_F^2 ≤ ε
```

2.3　算法分析

定理 1　基于张量分解的聚类方法相当于同时对网络中所有类型的对象进行了 K-means 聚类。

证明:首先证明基于张量分解的聚类方法中的特征矩阵 $\boldsymbol{U}^{(t)}$ 的迭代更新等价于对网络中第 t 种类型的对象进行 K-means 聚类。

对于网络中的第 $t\,|_{t=1}^{T}$ 种类型的对象,$\mathcal{X}_{(t)}$ 表示第 t 种类型的对象与其他类型对象的邻接矩阵,$\mathcal{X}_{(t)}$ 中的每一行 $\boldsymbol{x}_{n_t:}$ 表示第 t 种类型的一个对象。K-means 算法将第 t 种类型的对象划分到 K 个簇 $\{C_1, C_2, \cdots, C_K\}$ 中,对应的簇中心记为 $\{c_1, c_2, \cdots, c_K\}$,使得每一个对象与对应簇中心之间的距离最小,即

$$\min \sum_{k=1}^{K} \sum_{n_t=1}^{N_t} \| \boldsymbol{x}_{n_t:} - p_{n_t k} c_k \|_{\mathrm{F}}^2 \tag{16}$$

式中,$p_{n_t k}$ 为对象 $\boldsymbol{x}_{n_t:}$ 属于簇 C_k 的标识,若 $\boldsymbol{x}_{n_t:} \in C_k$,则 $p_{n_t k} = 1$,否则 $p_{n_t k} = 0$。将式(16)写成矩阵形式,有

$$\min_{P,C} \| \mathcal{X}_{(t)} - \boldsymbol{P}\boldsymbol{C} \|_{\mathrm{F}}^2 \tag{17}$$

将式(5)代入式(3),得

$$\min_{U^{(t)}} \| \mathcal{X}_{(t)} - \boldsymbol{U}^{(t)} \mathcal{S}_{(t)} \|_{\mathrm{F}}^2 \tag{18}$$

令 $\boldsymbol{U}^{(t)} = \boldsymbol{P}$,$\mathcal{S}_{(t)} = \boldsymbol{C}$,从而式(18)与式(17)等价,即基于张量分解的聚类算法中的特征矩阵 $\boldsymbol{U}^{(t)}$ 的迭代更新等价于对网络中第 t 种类型的对象进行 K-means 聚类。

其次证明基于张量分解的聚类可以同时获得网络中所有不同类型对象的聚类结果。

由算法 1 可知,基于张量分解的聚类算法一次运行可以同时获得张量每个维度上的特征矩阵以及核张量,即 $\boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \cdots, \boldsymbol{U}^{(T)}$ 和 $\mathcal{G}$。而前面已经证明每一个特征矩阵 $\boldsymbol{U}^{(t)}$ 相当于异构信息网络中第 t 种类型对象的簇标识矩阵,而 $\mathcal{S}_{(t)} = \mathcal{G}_{(t)}(\boldsymbol{U}^{(T)} \otimes \cdots \otimes \boldsymbol{U}^{(t+1)} \otimes \boldsymbol{U}^{(t-1)} \otimes \cdots \otimes \boldsymbol{U}^{(1)})^{\mathrm{T}}$ 是第 t 种类型对象的簇中心矩阵。所以基于张量分解的聚类可以同时获得网络中所有不同类型对象的聚类结果。□

文献[12]中已经证明了 NMF 算法的收敛性,在此直接引入。

定理 2[12]　$\| \mathcal{X}_{(t)} - \boldsymbol{U}^{(t)} \mathcal{S}_{(t)} \|_{\mathrm{F}}^2$ 在更新规则式(7)下是非递增的,当且仅当 $\boldsymbol{U}^{(t)}$ 是局部最小值时,$\| \mathcal{X}_{(t)} - \boldsymbol{U}^{(t)} \mathcal{S}_{(t)} \|_{\mathrm{F}}^2$ 的值保持不变。

证明:请参见文献[12]的定理 1 证明。□

将定理 2 扩展到高维空间,可以得到基于张量分解的聚类算法是收敛的。

引理　式(2)中的目标函数 $\| \mathcal{X} - [\![\mathcal{G}; \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \cdots, \boldsymbol{U}^{(T)}]\!] \|_{\mathrm{F}}^2$ 在更新规则(7)下是非递增的,当且仅当 $\boldsymbol{U}^{(t)}$ 是局部最小值时,$\| \mathcal{X} - [\![\mathcal{G}; \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \cdots, \boldsymbol{U}^{(T)}]\!] \|_{\mathrm{F}}^2$ 的值保持不变。

证明:记 $\boldsymbol{U}_{\mathrm{iter}+1}^{(t)}$ 及 $\boldsymbol{U}_{\mathrm{iter}}^{(t)}$ 分别为相邻两次迭代求得的特征矩阵 $\boldsymbol{U}^{(t)}$ 的结果,即 $\boldsymbol{U}_{\mathrm{iter}+1}^{(t)} = \boldsymbol{U}_{\mathrm{iter}}^{(t)} * \frac{\mathcal{X}_{(t)} \mathcal{S}_{(t)}^{\mathrm{T}}}{\boldsymbol{U}^{(t)} \mathcal{S}_{(t)} \mathcal{S}_{(t)}^{\mathrm{T}}}$。根据定理 2 可知,

$$\| \mathcal{X}_{(t)} - \boldsymbol{U}_{\mathrm{iter}+1}^{(t)} \mathcal{S}_{(t)} \|_{\mathrm{F}}^2 \leqslant \| \mathcal{X}_{(t)} - \boldsymbol{U}_{\mathrm{iter}}^{(t)} \mathcal{S}_{(t)} \|_{\mathrm{F}}^2 \tag{19}$$

当且仅当 $\boldsymbol{U}_{\mathrm{iter}+1}^{(t)} = \boldsymbol{U}_{\mathrm{iter}}^{(t)}$,$\boldsymbol{U}_{\mathrm{iter}}^{(t)}$ 为局部最小值时等号成立。

将式(5)代入式(19),得

$$\begin{aligned} &\| \mathcal{X}_{(t)} - \boldsymbol{U}_{\mathrm{iter}+1}^{(t)} \mathcal{G}_{(t)} (\boldsymbol{U}^{(T)} \otimes \cdots \otimes \boldsymbol{U}^{(t+1)} \otimes \boldsymbol{U}^{(t-1)} \otimes \cdots \otimes \boldsymbol{U}^{(1)})^{\mathrm{T}} \|_{\mathrm{F}}^2 \\ &\leqslant \| \mathcal{X}_{(t)} - \boldsymbol{U}_{\mathrm{iter}}^{(t)} \mathcal{G}_{(t)} (\boldsymbol{U}^{(T)} \otimes \cdots \otimes \boldsymbol{U}^{(t+1)} \otimes \boldsymbol{U}^{(t-1)} \otimes \cdots \otimes \boldsymbol{U}^{(1)})^{\mathrm{T}} \|_{\mathrm{F}}^2 \end{aligned} \tag{20}$$

将式(20)写为张量形式,得

$$\begin{aligned} &\| \mathcal{X} - [\![\mathcal{G}; \boldsymbol{U}^{(1)}, \cdots, \boldsymbol{U}_{\mathrm{iter}+1}^{(t)}, \cdots, \boldsymbol{U}^{(T)}]\!] \|_{\mathrm{F}}^2 \\ &\leqslant \| \mathcal{X} - [\![\mathcal{G}; \boldsymbol{U}^{(1)}, \cdots, \boldsymbol{U}_{\mathrm{iter}}^{(t)}, \cdots, \boldsymbol{U}^{(T)}]\!] \|_{\mathrm{F}}^2 \end{aligned} \tag{21}$$

当且仅当 $\boldsymbol{U}_{\mathrm{iter}+1}^{(t)} = \boldsymbol{U}_{\mathrm{iter}}^{(t)}$,$\boldsymbol{U}_{\mathrm{iter}}^{(t)}$ 为局部最小值时,等号成立。□

同理可证 $\| \mathcal{X} - [\![\mathcal{G}; \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \cdots, \boldsymbol{U}^{(T)}]\!] \|_{\mathrm{F}}^2$ 在核张量更新规则式(13)下也是非递增的,并且当且仅当核张量 $\mathcal{G}$ 为局部最小值时,$\| \mathcal{X} - [\![\mathcal{G}; \boldsymbol{U}^{(1)}, \boldsymbol{U}^{(2)}, \cdots, \boldsymbol{U}^{(T)}]\!] \|_{\mathrm{F}}^2$ 的值保持不变。

3　实验和结果

本节在人工合成数据集和真实数据集上对所

提基于张量分解的聚类方法进行了实验验证,并且将实验结果与其他先进的算法进行了比较。实验中采用标准互信息(Normalized Mutual Information, NMI)[1]和准确度(ACcuracy, AC)[14]作为性能评价指标。其中,NMI用来度量聚类结果和基准集之间的标准互信息;AC用来计算聚类准确度。所有的实验结果都是将算法在对应的数据集上运行10次后取平均值得到的。

3.1 人工合成数据集

3.1.1 数据集描述

利用人工合成数据集的目的是检测基于张量分解的聚类算法是否能够达到预期的目标,因为人工合成数据集中的簇结构都是已知的。设置了以下参数来评估算法的性能。

1) T:异构信息网络中对象的类型数量。

2) K:簇的数量。

3) D:网络规模, $D = N_1 \times N_2 \times \cdots \times N_T$。

实验A:为了定量地评价算法在不同 T 和 D 上的性能,将网络中簇的数量固定为 $K=2$。然后构造四种不同规模的数据集,分别是 $D=5\times10^6$、$D=5\times10^7$、$D=5\times10^8$ 和 $D=5\times10^9$。针对每一种规模的网络,分别设置不同的对象类型数 $T=2$、$T=4$、$T=6$ 和 $T=8$,详见表2,共有4种不同规模的16个数据集。

表2 人工合成数据集

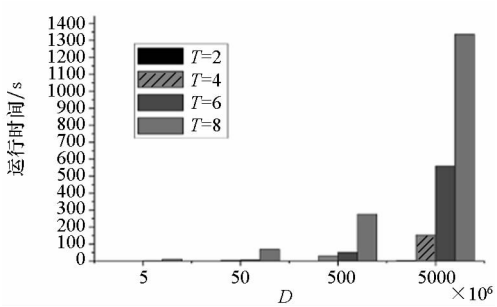
Tab. 2 Synthetic datasets

	$T=2$	$T=4$	$T=6$	$T=8$
$D=5\times10^6$	5000×1000	50×100×10×100	50×10×10×10×10×10	5×10×10×10×5×5×5×8
$D=5\times10^7$	5000×10 000	50×100×100×100	50×100×10×10×10×10	10×10×10×10×5×10×10×10
$D=5\times10^8$	50 000×10 000	500×100×100×100	50×100×100×10×10×10	10×10×10×10×50×10×10×10
$D=5\times10^9$	50 000×100 000	500×1000×100×100	50×100×100×100×10×10	100×10×10×10×50×10×10×10

实验B:为了测试算法中分解维度的大小,即网络中存在簇的数量 K 对算法性能的影响,固定网络中对象的类型数 $T=4$,网络规模 $D=1\times10^8=100\times100\times100\times100$。分别设置 $K=2$、$K=4$、$K=6$和 $K=8$。

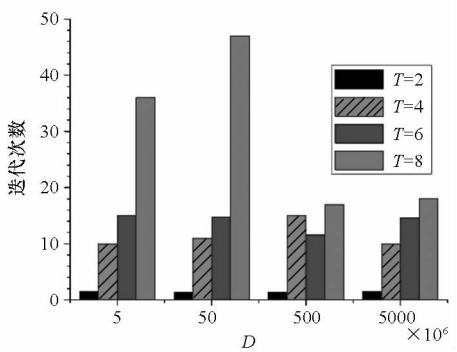
3.1.2 实验结果

实验A中,由随机初始化方法产生的特征矩阵和核张量偶尔会导致算法的收敛速度较慢,甚至出现在迭代次数达到用户设定的最大迭代次数时还是没有达到收敛条件,但是这种现象比较少,移除了结果中不收敛的情况。图2给出了运行时间和迭代次数的结果,图3给出了AC和NMI的结果。



(a) 不同 T 和 D 情况下的运行时间

(a) Running time with different T and D



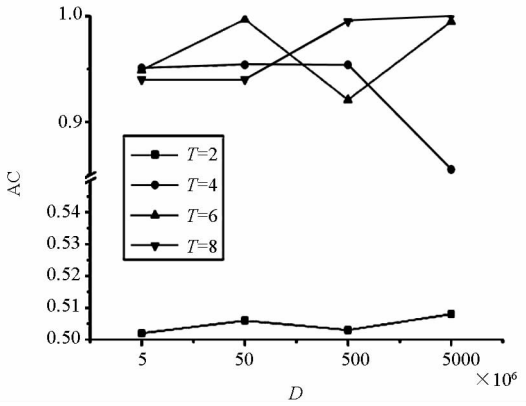
(b) 不同 T 和 D 情况下的迭代次数

(b) Iterations with different T and D

图2 不同 T 和 D 情况下的运行时间和迭代次数

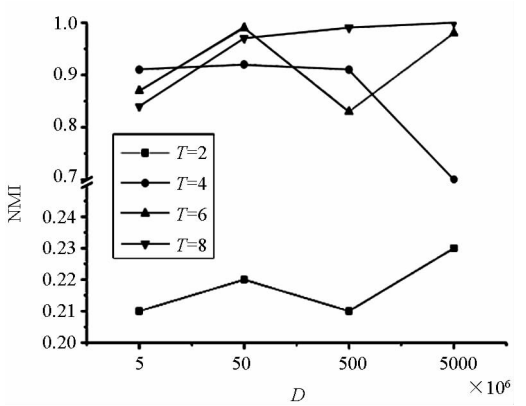
Fig. 2 Running time and iterations with different T and D

在相同网络规模的情况下,随着对象类型数的增加,运行时间和迭代次数也几乎是增加的,直至AC和NMI趋近于1。当 $T=2$ 时,在不同网络规模的情况下,运行时间和迭代次数几乎都是相同的,AC和NMI也都是比较小的。这也就意味着,聚类的结果在对象类型较少的情况下是不佳的。原因是每一个对象出现在基因网络中的频率,也就是网络中有用的语义关系较少。当对象类型变多时,每一个对象出现在基因网络中的频率也就相应变大,聚类过程中可利用的语义关系就增加了。这也是AC和NMI同时趋近于1的原因。



(a) 不同 T 和 D 情况下的 AC

(a) AC with different T and D



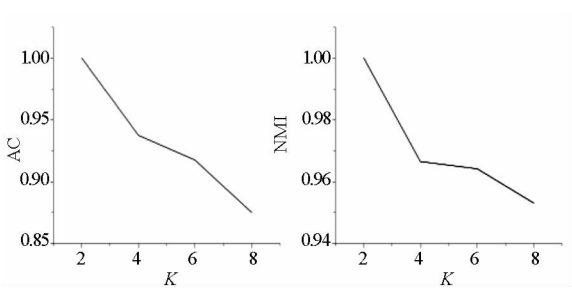
(b) 不同 T 和 D 情况下的 NMI

(b) NMI with different T and D

图 3　不同 T 和 D 情况下的 AC 和 NMI

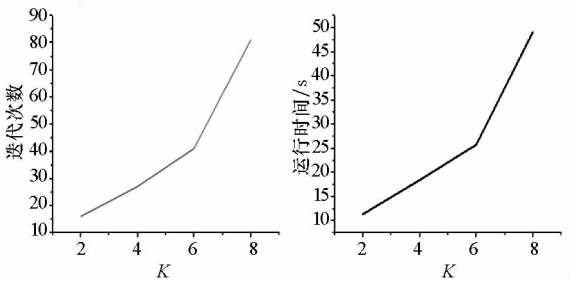
Fig. 3　AC and NMI with different T and D

实验 B 中,保持网络规模和网络中对象的类型数不变,随着网络中簇数的增大,AC 及 NMI 下降,而迭代次数及运行时间有所增加,实验结果如图 4 所示。显然地,随着 K 的增加,特征矩阵 $\boldsymbol{U}^{(t)} \in \mathbb{R}^{N_t \times K}$ 和核张量 $\mathcal{G} \in \mathbb{R}^{\overbrace{K \times K \times \cdots \times K}^{T}}$ 的规模成倍增加,从而导致算法求解过程中消耗的运算量上升。同时,在网络规模不变的情况下,构造的张量规模不变,随着网络中簇的数量变多,张量分解的维度越大,算法效率越低。



(a) 准确度　(b) 标准互信息

(a) Accuracy　(b) Normalized mutual information



(c) 迭代次数　(d) 运行时间

(c) Number of Iterations　(d) Running time

图 4　簇数 K 对算法性能的影响

Fig. 4　Effect of K on the performance of the algorithm

显然地,网络规模越大,对象类型越多,需要的迭代次数和运行时间都相应增加,同时 AC 和 NMI 也越大。有两个原因导致迭代次数和运行时间的增加。第一,网络中对象类型越多,构造的张量维度越多。也就是说特征矩阵的数量和核张量的规模越大。第二,当网络规模固定的时候,随着对象类型的增加,每一种类型的对象数目却在减少。

3.2　真实数据集

3.2.1　数据集描述

真实数据集上的实验是为了比较基于张量分解的异构信息聚类算法和现有算法的性能。真实数据集是从 DBLP 网络中抽取的 DBLP - four - areas dataset,下载地址为 http://web. cs. ucla. edu/ ~ yzsun/data/DBLP _ four _ area. zip。这是 DBLP 中的 4 个研究领域的数据集,用于文献[2 -5, 7, 15 -16]。4 个研究领域分别是数据库、数据挖掘、机器学习和信息检索。每一个研究领域中挑选除了 5 个具有代表性的会议,相关的作者和他们发表在这些会议上的论文,以及论文中使用的主题词均包含在数据集中。DBLP - four - areas 数据集包含14 376 篇论文,其中 100 篇标记了簇标签;14 475 个作者,其中 4057 个标记了簇标签,20 个标记了簇标签的会议和 8920 个主题词。由此构造了一个 4 阶张量,其规模为 14 376 × 14 475 × 20 × 8920。

3.2.2　对比算法

1) NetClus[2]:RankClus[1] 在星型网络模式上的扩展。

2) PathSelClus[5, 7]:基于预定义元路径和用户知识的聚类方法。在 PathSelClus 中,相同类型对象之间的距离用 PathSim[15] 来度量,该方法需要用户给定每个簇的种子。

3) FctClus[3]:最近提出的一种异构信息网络

聚类方法。与 NetClus 一样,FctClus 只能处理星型网络模式。

3.2.3 实验结果

由于这些基准方法都只能处理指定模式的异构信息,实验为这些方法构造了不同的网络。对于 NetClus 和 FctClus,网络被组织为星型模式,与文献[2-3]类似,论文(P)作为中心类型,作者(A)、会议(C)和主题词(T)作为属性类型。对于 PathSelClus,选择元路径 P-T-P,A-P-C-P-A和 C-P-T-P-C 分别来聚类论文(P)、作者(A)和会议(C),并且给定每个簇一个种子。

将 DBLP-four-areas 数据集建模为一个4维张量,其中每一个维度代表一种对象类型。4个维度分别是作者(A)、论文(P)、会议(C)和主题词(T)。实际上,对象类型的顺序是无关紧要的。

表3~5 分别给出了基于张量分解的异构信息聚类和其他方法在真实数据集上的实验结果。从 DBLP-four-areas 数据集上的实验结果来看,基于张量分解的异构信息聚类方法在 AC 和 NMI 上的表现都优于其他方法,而 PathSelClus 在运行时间上表现最优。

表3 真实数据集上的 AC 结果

Tab. 3 AC on the real-world dataset

	基于张量分解的聚类	NetClus	PathSelClus	FctClus
论文(P)	0.769 9	0.715 4	0.755 1	0.788 7
作者(A)	0.825 4	0.717 7	0.795 1	0.800 8
会议(C)	0.999 8	0.917 2	0.995 0	0.903 1
avg(AC)	0.825 0	0.718 6	0.795 1	0.801 0

表4 真实数据集上的 NMI 结果

Tab. 4 NMI on the real-world dataset

	基于张量分解的聚类	NetClus	PathSelClus	FctClus
论文(P)	0.704 4	0.540 2	0.614 2	0.715 2
作者(A)	0.854 9	0.548 8	0.677 0	0.601 2
会议(C)	0.999 4	0.885 8	0.990 6	0.824 8
avg(NMI)	0.852 0	0.550 3	0.677 0	0.605 0

尽管基于张量分解的异构信息聚类方法的运行时间是最长的,但是它可以一次性同时得到所有类型对象的聚类结果,这是其他方法所不能达到的。这也是为什么表5中基于张量分解的异构信息聚类方法的结果只有总时间的原因。NetClus 在 AC 和 NMI 上表现最差,只达到71.86% 和55.03%,但是 NetClus 有一个重要的优势就是其可以在聚类的同时得到这些对象的重要性排序。PathSelClus 在 AC 和 NMI 上表现优于 NetClus,它也有一个优点,就是基于 PathSim,PathSelClus 可以快速度量两个相同类型对象之间的相似度。PathSelClus 在运行时间上的表现也是最好的,但是 PathSelClus 的结果严重依赖于元路径的选择和用户预先给定的种子。

表5 真实数据集上的运行时间结果

Tab. 5 Running time on the real-world dataset

s

	基于张量分解的聚类	NetClus	PathSelClus	FctClus
论文(P)		802.6	542.3	808.4
作者(A)		743.7	681.1	774.9
会议(C)		658.4	629.3	669.8
总时间	2840.9	2204.7	1852.7	2253.1

4 结论

本文提出了一种基于张量分解的网络异构信息聚类方法。在该方法下,网络异构信息中的每一种对象类型被建模为张量的一个维度,基因网络被建模为张量中的元素。也就是说,基于张量分解的网络异构信息聚类方法可以将网络中不同类型的对象和语义关系建模为一个张量,然后基于张量分解,一次性将所有类型的对象同时进行聚类。

参考文献(References)

[1] Sun Y Z, Han J W, Zhao P X, et al. RankClus: integrating clustering with ranking for heterogeneous information network analysis [C]//Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, 2009: 565-576.

[2] Sun Y Z, Yu Y T, Han J W. Ranking-based clustering of heterogeneous information networks with star network schema [C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009: 797-806.

[3] Yang J, Chen L M, Zhang J P. FctClus: a fast clustering algorithm for heterogeneous information networks [J]. PloS One, 2015, 10(6): e0130086.

[4] Sun Y Z, Aggarwal C C, Han J W. Relation strength-aware clustering of heterogeneous information networks with incomplete attributes [J]//Proceedings of the VLDB Endowment, 2012, 5(5): 394-405.

(下转第170页)

navy vessel's development [J]. System Engineering—Theory & Practice, 2012, 32(10): 2339 – 2344. (in Chinese)

[3] Oehmen J, Olechowski A, Kenley C R, et al. Analysis of the effect of risk management practices on the performance of new product development programs [J]. Technovation, 2014, 34(8): 441 – 453.

[4] Mankins J C. Technology readiness assessments: a retrospective[J]. Acta Astronautica, 2009, 65(9/10): 1216 – 1223.

[5] Barends D M, Oldenhof M T, Vredenbregt M J, et al. Risk analysis of analytical validations by probabilistic modification of FMEA [J]. Journal of Pharmaceutical & Biomedical Analysis, 2012, 64/65(4): 82 – 86.

[6] Mohaghegh Z, Kazemi R, Mosleh A. Incorporating organizational factors into probabilistic risk assessment (PRA) of complex socio-technical systems: a hybrid technique formalization[J]. Reliability Engineering & System Safety, 2009, 94(5): 1000 – 1018.

[7] Tzeng S. Management towards success—defense business system acquisition probability of success model[D]. Fairfax, VA, USA: George Mason University, 2015.

[8] Xu Z, Li H B. Assessing performance risk for complex product development: a simulation-based model [J]. Quality & Reliability Engineering International, 2013, 29 (2): 267 – 275.

[9] Lee A. Petri net modeling of fault analysis for probabilistic risk assessment [D]. Oshawa, Canada: University of Ontario Institute of Technology, 2013.

[10] Boateng P, Chen Z, Ogunlana S O. Megaproject risk analysis and simulation: a dynamic systems approach[M]. Bingley, UK: Emerald Group Publishing Ltd., 2017.

[11] Eppinger S D, Browning T R. Design structure matrix methods and applications [M]. Massachusetts, USA: MIT Press Books, 2012.

[12] Davis P K, Tolk A. Observations on new developments in composability and multi-resolution modeling[C]//Proceeding of Winter Simulation Conference, 2007: 859 – 870.

[13] Sauser B, Gove R, Forbes E, et al. Integration maturity metrics: development of an integration readiness level [J]. Information Knowledge Systems Management, 2010, 9(1): 17 – 46.

(上接第 152 页)

[5] Sun Y Z, Norick B, Han J W, et al. Integrating meta-path selection with user-guided object clustering in heterogeneous information networks [C]//Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012: 1348 – 1356.

[6] Yu X, Sun Y Z, Norick B, et al. User guided entity similarity search using meta-path selection in heterogeneous information networks [C]// Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 2012: 2025 – 2029.

[7] Sun Y Z, Norick B, Han J W, et al. PathSelClus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks [J]. ACM Transactions on Knowledge Discovery from Data, 2013, 7(3): 11.

[8] Kolda T G, Bader B W. Tensor decompositions and applications [J]. SIAM Review, 2009, 51(3): 455 – 500.

[9] Kolda T G. Multilinear operators for higher-order decompositions: SAND2006 – 2081 [R]. USA: Sandia National Laboratories, 2006.

[10] Tucker L R. Some mathematical notes on three-mode factor analysis[J]. Psy-chometrika, 1966, 31(3): 279 – 311.

[11] Lee D D, Seung H S. Learning the parts of objects by nonnegative matrix factorization [J]. Nature, 1999, 401(6755): 788 – 791.

[12] Lee D D, Seung H S. Algorithms for nonnegative matrix factorization [C]//Proceedings of Advances in Neural Information Processing Systems, 2001: 556 – 562.

[13] Zhang Z H. The singular value decomposition, applications and beyond [J/OL]. (2015 – 10 – 29)[2017 – 04 – 28]. http://arxiv.org/abs/1510.08532v1.

[14] Cai D, He X F, Wang S H, et al. Locality preserving nonnegative matrix factorization [C]// Proceeding of International Conference on Artificial Intelligence, 2009: 1010 – 1015.

[15] Sun Y Z, Han J W, Yan X F, et al. PathSim: meta path-based top-k similarity search in heterogeneous information networks[J]. Proceedings of the VLDB Endowment, 2011, 4(1): 992 – 1003.

[16] Chen J X, Dai W, Sun Y Z, et al. Clustering and ranking in heterogeneous information networks via Gamma-Poisson model[C]// Proceedings of the SIAM International Conference on Data Mining, 2015.