计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2016.20148284
53(2):459-466, 2016

基于非负矩阵分解的大规模异构数据联合聚类

申国伟 杨 武 王 巍 于 淼 董国忠
(哈尔滨工程大学信息安全研究中心 哈尔滨 150001)
(shenguowei@hrbeu.edu.cn)

Large-Scale Heterogeneous Data Co-Clustering Based on Nonnegative Matrix Factorization

Shen Guowei, Yang Wu, Wang Wei , Yu Miao, and Dong Guozhong
(*Research Center of Information Security*, *Harbin Engineering University*, *Harbin* 150001)

Abstract Heterogeneous information network contains multi-typed entities and interactive relations. Some co-clustering algorithms have been proposed to mine underlying structure of different entities. However, with the increase of data scale, the scale of different class entities are growing unbalanced, and heterogeneous relational data are becoming extremely sparse. In order to solve this problem, we propose a two steps co-clustering algorithm FNMTF-CM based on correlation matrix decomposition. In the first step, the correlation matrix is built with the correlation relationship of smaller-typed entities and decomposed into indicating matrix of smaller-typed entity based on symmetric nonnegative matrix factorization. Correlation matrix has higher dense degree and smaller size compared with the original heterogeneous relationship matrix, so our algorithm can process large-scale heterogeneous data and maintain a high precision. After that, the indicating matrix of smaller-typed can be used as the input directly, so the heterogeneous relational matrix tri-factorization is very fast. Experiments on artificial and real-world heterogeneous data sets show that the accuracy and performance of FNMTF-CM algorithm are superior to the traditional co-clustering algorithms based on nonnegative matrix factorization.

Key words heterogeneous network; co-clustering; nonnegative matrix factorization; large-scale data; correlation matrix

摘 要 异构信息网络中包含多类实体和关系.随着数据规模增大时,不同类实体规模增长不平衡,异构关系数据也变得异常稀疏,导致聚类算法的时间复杂度高、准确率低.针对上述问题,提出了一种基于关联矩阵分解的2阶段联合聚类算法FNMTF-CM.第1阶段,抽取规模较小的一类实体中的关联关系构建关联矩阵,通过对称非负矩阵分解得到划分指示矩阵.与原始关系矩阵相比,关联矩阵的稠密度更高,规模更小.第2阶段,将划分指示矩阵作为关系矩阵三分解的输入,进而快速求解另一类实体的划分指示矩阵.在标准测试数据集和异构关系数据集上的实验表明,算法准确率和性能整体优于传统的基于非负矩阵分解的联合聚类算法.

收稿日期:2014-11-24;**修回日期**:2015-03-26
基金项目:国家"八六三"高技术研究发展计划基金项目(2012AA012802);国家自然科学基金项目(61170242)
This work was supported by the National High Technology Research and Development Program of China (863 Program) (2012AA012802) and the National Natural Science Foundation of China (61170242).
通信作者:杨武(yangwu@hrbeu.edu.cn)

关键词 异构网络;联合聚类;非负矩阵分解;大规模数据;关联矩阵

中图法分类号 TP391

随着微博、社交网络等异构信息网络的兴起,异构信息挖掘已经成为当前数据挖掘领域中的一个研究热点.异构网络中包含多类实体,实体之间存在着复杂的交互关系.例如微博中包含用户、消息、标签、词等实体,用户发布消息,消息由词语组成,消息中还包含标签等.通过抽取实体间的关系数据进行聚类分析,能够挖掘出异构网络中不同实体间的潜在结构关系.

联合聚类能够针对不同的实体同时进行聚类分析[1-2],因而应用广泛.传统的联合聚类算法包括基于信息理论的算法 ITCC[3]、基于矩阵谱信息[4]和矩阵分解的方法.由于关系数据中一般都是非负元素,非负矩阵分解方法[5]成为目前最常用的方法.

传统的非负矩阵分解仅仅处理同类节点之间的同质关系聚类问题,Long 等人[6]首次在二元关系矩阵上运用块值分解法实现矩阵分解.在此基础上,提出了一系列改进的非负矩阵分解方法实现联合聚类[7-9].采用半监督的非负矩阵分解方法实现联合聚类[10-12],算法 SS-NMF[12]中融合肯定链接或否定链接等约束信息提高联合聚类算法的准确度,但是真实数据中通常很难获取约束先验知识.

在处理关系数据时,Wang 等人[13]提出了快速的非负矩阵三分解方法 FNMTF 实现快速的矩阵分解,进而实现联合聚类.非负矩阵分解在联合聚类算法取得了很好的效果[14],但是数据本身的几何结构会影响聚类的准确性[15-16].当待分析的异构数据规模增大时,关系数据结构呈现明显变化.主要存在以下问题:

1) 非平衡问题.待分析的异构数据规模增大时,异构数据中不同类实体的规模并不呈现统一的增长模式.例如微博消息数量呈线性增长时,用户、词和标签等实体并不呈现线性增长模式.传统的非负矩阵分解方法的时间复杂度都与矩阵的行和列规模相关,因此处理大规模异构数据时计算时间复杂度较高.

2) 稀疏性问题.真实异构网络中的关系数据比较稀疏,随着待分析异构数据规模进一步增大,关系数据变得异常稀疏.例如微博中的消息内容最多包含 140 个字,构建的消息和词之间的关系矩阵非常稀疏.当消息规模进一步增大时,由于中文常用词的数量是一定的,因此消息和词之间的关系矩阵变得异常稀疏,消息和用户、标签的关系矩阵同样如此.传统的非负矩阵分解方法针对异常稀疏的关系矩阵进行分解时得到的聚类效果并不理想.

本文针对大规模异构数据分析时出现的非平衡和稀疏性 2 个问题进行解决.针对非平衡增长问题,在非负矩阵分解时提出了 2 阶段分解方法.首先,仅对关系矩阵中的规模较小的一类实体进行分析.异构实体之间的关系矩阵非常稀疏,但是同一类实体之间的关联性比较强[17],通过同类实体之间的关联关系构造的关联矩阵能够明显提高矩阵的稠密度.其次,以较小规模的实体聚类结果直接作为第 2 阶段的输入,在确保大规模实体聚类结果的同时提高了整体处理效率.

综上所述,本文将针对大规模异构关系数据提出一种基于关联矩阵的 2 阶段快速联合聚类算法,同时解决非平衡问题和稀疏性问题.

1 问题定义

异构关系数据中包括多类实体,目前的联合聚类算法主要针对二阶异构关系进行联合聚类分析,因此,本文以 2 类实体之间的异构关系为例叙述.二阶异质关系数据采用二部图 $G=(V,E,W)$ 进行建模,其中 $V=X_1 \cup X_2$,X_1 和 X_2 为异构关系中的 2 类实体,实体 X_1 和 X_2 的数量分别为 m 和 n,E 为异构关系对应的边集合,W 为边的权重.进一步可将二部图 G 表示成 $m \times n$ 的异构关系矩阵 $\boldsymbol{R}$,由于大规模数据中的非平衡问题,可假设 $m \gg n$.

传统的联合聚类算法中将 X_1 和 X_2 分别划分到 k_1 和 k_2 类(通常 $k_1=k_2$),本文将针对 X_1 和 X_2 的联合聚类问题转换成针对关系矩阵 $\boldsymbol{R}$ 的行和列同时进行划分的问题.

2 2 阶段非负矩阵分解框架

针对大规模异构关系数据中的非平衡问题和稀疏性问题,本文提出了一个 2 阶段的非负矩阵分解框架,如图 1 所示.

对关系矩阵 $\boldsymbol{R}$ 的行和列同时聚类可将关系矩阵 $\boldsymbol{R}$ 分解为 $\boldsymbol{F}$,$\boldsymbol{S}$,$\boldsymbol{B}$ 三个矩阵,如图 1(a)所示,其中

矩阵 $\boldsymbol{F},\boldsymbol{B}$ 分别为 2 类目标实体的聚类指示矩阵,$\boldsymbol{S}$ 为联合类之间相关矩阵. 本文不直接针对关系矩阵 $\boldsymbol{R}$ 进行分解,而是分 2 阶段实现.

第 1 阶段针对实体数较少的一类实体 X_2 进行处理,其数量为 n. 从关系矩阵 $\boldsymbol{R}$ 中抽取同类实体间的关联关系,进而构建关联矩阵 $\boldsymbol{C}$,矩阵 $\boldsymbol{C}$ 的规模为 $n\times n$. 对矩阵 $\boldsymbol{C}$ 进行对称非负矩阵分解,得到指示矩阵 $\boldsymbol{B}$,如图 1(b)所示. 由于采用同类实体的关联关系,构建的同质关系矩阵 $\boldsymbol{C}$ 比原来的关系矩阵 $\boldsymbol{R}$ 要稠密,在某种程度上能够避免非负矩阵分解中的稀疏性问题,进而提高非负矩阵分解的准确性.

在第 2 阶段中,将关联矩阵 $\boldsymbol{C}$ 分解得到的指示矩阵 $\boldsymbol{B}$ 直接作为关系矩阵 $\boldsymbol{R}$ 三分解的指示矩阵,如图 1(c)所示. 在矩阵 $\boldsymbol{B}$ 已知的情况下,可以很容易计算指示矩阵 $\boldsymbol{F}$ 和矩阵 $\boldsymbol{S}$. 由问题定义可知,矩阵 $\boldsymbol{C}$ 的规模小于原始关系矩阵 $\boldsymbol{R}$ 的规模 $m\times n$,因此该框架能够处理大规模异构关系数据.

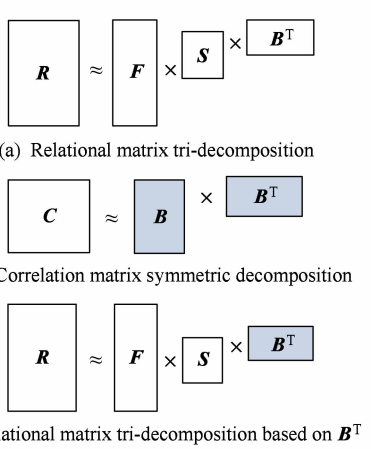

Fig. 1 The framework of heterogeneous data co-clustering.

图 1 异构数据联合聚类框架

3 基于关联矩阵的稀疏联合聚类

根据 2 阶段非负矩阵分解框架,在异构关系矩阵的基础上,联合聚类主要包括关联矩阵构造、关联矩阵分解、基于关联矩阵的异构关系矩阵三分解 3 部分.

3.1 关联矩阵构造

在异构关系数据中,选择规模较小的一类实体 X_2,通过异构关系矩阵 $\boldsymbol{R}$ 构造 X_2 对应的关联矩阵 $\boldsymbol{C}$. 文中利用关联强度 $W_{i,j}$ 度量实体 X_2 中任意 2 个实体 x_i,x_j 的关联关系,其可通过 2 个实体 x_i,x_j 基于 X_1 中实体的同现概率进行计算,其计算方法如式(1)所示:

$$W_{i,j}=\max\left(\lg\frac{P(x_i,x_j)}{P(x_i)P(x_j)},0\right). \tag{1}$$

其中概率 $P(x_i,x_j)$ 和 $P(x_i)$ 计算分别如式(2)(3)所示:

$$P(x_i,x_j)=\frac{N(x_i,x_j)}{\sum\limits_{s,t}N(x_s,x_t)}; \tag{2}$$

$$P(x_i)=\frac{\sum\limits_{j}N(x_i,x_j)}{\sum\limits_{s,t}N(x_s,x_t)}. \tag{3}$$

式(2)和式(3)中,$N(x_i,x_j)$ 为 X_2 中的实体 x_i,x_j 基于 X_1 中实体同时出现次数.

3.2 关联矩阵分解

通过关联关系构造的关联矩阵 $\boldsymbol{C}$,采用对称非负矩阵分解方法进行分解,其对应的目标函数为式(4)所示:

$$J_1=\|\boldsymbol{C}-\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\|^2 \quad \text{s.t. } B_{ij}\geqslant 0, \tag{4}$$

其中 $\|\cdot\|^2$ 为矩阵 F-范数.

针对目标函数 J_1,可通过非负最小二乘法进行计算,其计算公式如式(5)所示. 基于关联矩阵 $\boldsymbol{C}$ 的分解结果为聚类指示矩阵 $\boldsymbol{B}$.

$$\boldsymbol{B}_{t+1}=\max(\boldsymbol{C}\boldsymbol{B}_t(\boldsymbol{B}_t^{\mathrm{T}}\boldsymbol{B}_t)^{-1},0). \tag{5}$$

由于聚类指示矩阵中每一个实体只属于一个聚类标签,因此,对矩阵 $\boldsymbol{B}$ 进行二元化,即 $\boldsymbol{B}$ 中每一行的最大值对应的聚类结果为 1,其余对应的都为 0. 二元化后的矩阵 $\boldsymbol{B}$ 将作为关系矩阵 $\boldsymbol{R}$ 三分解的输入.

3.3 基于关联矩阵的异构关系矩阵三分解

传统的非负矩阵分解通常采用的目标函数如式(6)所示,该目标函数中采用两因子分解法.

$$\begin{aligned}&J_2=\|\boldsymbol{R}-\boldsymbol{F}\boldsymbol{B}^{\mathrm{T}}\|^2,\\&\text{s.t. } F_{ij}\geqslant 0,B_{ij}\geqslant 0,\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}=\boldsymbol{I},\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}=\boldsymbol{I}.\end{aligned} \tag{6}$$

两因子分解法得到的近似低秩矩阵效果较差,因此 Ding 等人[18]提出了正交非负矩阵三分解,其对应的目标函数为式(7)所示,在目标函数中引入了矩阵 $\boldsymbol{S}$,使得分解得到的矩阵 $\boldsymbol{F},\boldsymbol{B}$ 具有实际意义.

$$\begin{aligned}&J_3=\|\boldsymbol{R}-\boldsymbol{F}\boldsymbol{S}\boldsymbol{B}^{\mathrm{T}}\|^2,\\&\text{s.t. } F_{ij}\geqslant 0,S_{ij}\geqslant 0,B_{ij}\geqslant 0,\\&\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F}=\boldsymbol{I},\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B}=\boldsymbol{I}.\end{aligned} \tag{7}$$

由于正交约束条件在某些情况下过于严格,本文中将采用无正交约束的目标函数,如式(8)所示:

$$\begin{aligned}&J_4=\|\boldsymbol{R}-\boldsymbol{F}\boldsymbol{S}\boldsymbol{B}^{\mathrm{T}}\|^2,\\&\text{s.t. } F_{ij}\geqslant 0,B_{ij}\geqslant 0,S_{ij}\geqslant 0.\end{aligned} \tag{8}$$

对于目标函数 J_4,现有的方法中常采用乘法更新的迭代求解方法实现,但是其收敛速度较慢.本文将采用快速的迭代求解方法实现,关联矩阵 $\boldsymbol{C}$ 对称分解得到的矩阵 $\boldsymbol{B}$ 经过二元化后,直接作为目标函数 J_4 的输入,因此,只需迭代求解矩阵 $\boldsymbol{F}$ 和 $\boldsymbol{S}$.

在优化求解矩阵 $\boldsymbol{S}$ 的过程中,固定矩阵 $\boldsymbol{F}$,矩阵 $\boldsymbol{S}$ 的求解方法如式(9)所示:

$$\boldsymbol{S}=(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{B}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}. \tag{9}$$

在优化求解矩阵 $\boldsymbol{F}$ 的过程中,固定矩阵 $\boldsymbol{S}$.由于矩阵 $\boldsymbol{F}$ 为关系矩阵 $\boldsymbol{R}$ 的行划分指示矩阵,$\boldsymbol{F}$ 中的每一行有且只有一个元素为 1,其余为 0,因此求解矩阵 $\boldsymbol{F}$ 的优化问题可按照行进行处理,其转换为式(10)的优化问题.

$$J_5=\min_{\boldsymbol{F}}\|\boldsymbol{r}_{j\cdot}-\boldsymbol{f}_{j\cdot}\boldsymbol{S}\boldsymbol{B}^{\mathrm{T}}\|^2, \tag{10}$$

其中 $\boldsymbol{f}_{j\cdot}$ 为行聚类指示向量,在该向量中,有且只有一个元素为 1,其余的元素都为 0,因此,式(10)的优化问题可通过式(11)进行快速求解.

$$f_{ij}=\begin{cases}1, & i=\arg\min_k\|\boldsymbol{r}_{j\cdot}-\tilde{\boldsymbol{b}}_{k\cdot}\|;\\ 0, & \text{otherwise}.\end{cases} \tag{11}$$

其中 $\tilde{\boldsymbol{b}}_{k\cdot}$ 为 $\boldsymbol{S}\boldsymbol{B}^{\mathrm{T}}$ 对应的第 k 个行向量.式(10)可通过向量范式枚举法快速求解,避免了使用矩阵乘法迭代更新求解,提高了算法处理速度.

面向大规模异构数据的联合聚类算法的整个过程总结如算法 1 所示.在关联矩阵 $\boldsymbol{C}$ 对称分解的基础上对关系矩阵 $\boldsymbol{R}$ 进行非负矩阵三分解,能够同时解决非平衡和稀疏性问题.算法 1 中第④~⑨步为异构关系矩阵迭代求解过程.

算法 1. 异构数据联合聚类算法 FNMTF-CM.

输入:$\boldsymbol{R}$ 为关系矩阵,聚类数目 k,N_{iter} 为最大迭代次数,δ 为收敛阈值;

输出:$\boldsymbol{F}$ 为实体 X_1 聚类指示矩阵,$\boldsymbol{B}$ 为实体 X_2 聚类指示矩阵.

① 初始化 $\boldsymbol{F}_0$,$\boldsymbol{S}$,$\boldsymbol{B}_0$;

② 根据式(1)计算关联矩阵 $\boldsymbol{C}$;

③ 根据式(5)得到分解矩阵 $\boldsymbol{B}$,二元化后得到聚类指示矩阵 $\boldsymbol{B}$;

④ while ($iter<N_{\mathrm{iter}}$,$\Delta J<\delta$)

⑤ 根据式(9)计算 $\boldsymbol{S}$;

⑥ 根据式(11)计算 $\boldsymbol{F}$;

⑦ $iter=iter+1$;

⑧ 计算 ΔJ;/ * 2 次迭代的 F-范数差值 * /

⑨ end while

4 实验及分析

本文所有实验都在 Matlab 下实现,硬件平台为曙光 8 core 服务器、8 GB 内存.

实验中将分别对比算法 SS-NMF,FNMTF 和本文算法 FNMTF-CM.每一组实验分别运行 10 次,采用随机初始化,最终实验结果中给出平均值.

4.1 实验数据集

本文首先将在联合聚类算法的标准测试数据集[19]①上对算法进行全面的评估.该数据集给出了 2 类实体的聚类标签,不仅能够针对算法的准确率等指标值进行对比分析,还能对算法在不同聚类难度等级的数据集下进行对比分析.

数据集中共计 36 组数据,通过贝叶斯错误率 *Error* 作为数据集的难度控制参数,包括 5%,12%,20%共 3 个难度等级,其中 5%是最容易聚类的数据集,20%是最难聚类的数据集.每 1 个难度等级分别对应 50,100,200,500 共 4 种规模(行和列的规模相同),可针对节点规模进行聚类算法对比分析.每一类节点规模的数据集分别对应 3,5,10 共 3 种聚类数目,可针对不同的聚类数进行对比分析.

为了验证 FNMTF-CM 算法在真实数据集上的效果,将在 4 个真实的异构关系数据集上进行对比实验.其中 Title 数据为 Sogou 提供的新闻标题数据集②,构建新闻标题和词之间的异构关系数据集.Weibo 数据集收集了 2012 年“闯红灯”、“丰田汽车回收”、“美国总统大选”、“莫言获得诺贝尔奖”、“我是特种兵”、“杭州烟花大会”、“中国好声音”7 个话题对应的新浪微博消息,经过预处理后得到 8 023 条微博和 374 个标签,构建消息和标签之间的异构关系数据集.DBLP1 为论文与词之间的关系数据集,DBLP2 为论文与作者之间的关系数据集[20],这 2 个数据集中分别提取了论文题目和摘要字数超过 100 个、作者出现多于 2 次对应的关系数据.4 个数据集的详细信息如表 1 所示:

① https://www.hds.utc.fr/coclustering/doku.php

② http://www.sogou.com/labs/dl/tce.html

Table 1 Heterogeneous Relational Dataset

表 1 异构关系数据集

Datasets	# Entity 1	# Entity 2	# Clusters	# Sparse Degree/%
Title	2 630	1 420	9	0.40
Weibo	8 023	374	7	0.92
DBLP1	10 184	7 529	4	0.81
DBLP2	10 184	4 590	4	0.03

4.2 评估指标

联合聚类算法的度量指标较多,本文中将采用常见的 $Purity$[21],NMI[22],ARI[23] 3 个指标作为度量标准. 对于给定的异构数据集,其中实体规模为 n,算法得到的聚类结果为 $C=\{c_1,c_2,\cdots,c_K\}$,给定的聚类标签为 $R=\{r_1,r_2,\cdots,r_L\}$,则 3 个评估指标分别定义如式(12)、式(13)和式(14).

$$Purity(C,R)=\sum_{i=1}^{k}\frac{\max\limits_{j}|c_i\cap r_j|}{n}; \tag{12}$$

$$NMI(C,R)=\frac{2I(C;R)}{H(C)+H(R)}=\frac{2\sum\limits_{i,j}\frac{|c_i\cap r_j|}{n}\lg\frac{|c_i||r_j|}{n|c_i\cap r_j|}}{\sum\limits_{i}\frac{|c_i|}{n}\lg\frac{|c_i|}{n}+\sum\limits_{j}\frac{|r_j|}{n}\lg\frac{|r_j|}{n}}; \tag{13}$$

NMI 值在 0 到 1 之间,越接近 1,则说明聚类结果越好.

$$ARI(C,R)=\left(\sum_{i,j}C_{|c_i\cap r_j|}^2-\left[\sum_i C_{|c_i|}^2\sum_j C_{|r_j|}^2\right]\Big/C_n^2\right)\Big/\left(\frac{1}{2}\left[\sum_i C_{|c_i|}^2+\sum_j C_{|r_j|}^2\right]-\left[\sum_i C_{|c_i|}^2\sum_j C_{|r_j|}^2\right]\Big/C_n^2\right); \tag{14}$$

该 ARI 值越大,则聚类结果越好.

4.3 人工数据集实验

在同一规模的数据集下评估算法受不同聚类数目 K 的影响情况,对比结果如图 2 所示. 所有的算法随着 K 值的增加,准确率都有所下降,但其他 2 个指标影响较小. 因此,在实际使用的过程中,需要根据数据中的真实情况给定聚类数据 K.

图 3 中为在不同数据规模下的对比结果. 随着规模的增加,算法的准确率等指标都随之下降. 由于该数据集中 2 类实体的数目一致,无法发挥 FNMTF-CM 算法的优势,其聚类结果接近于 FNMTF 算法.

针对标准测试数据集中不同聚类难度等级的数据集进行算法的鲁棒性对比实验,结果如图 4 所示.

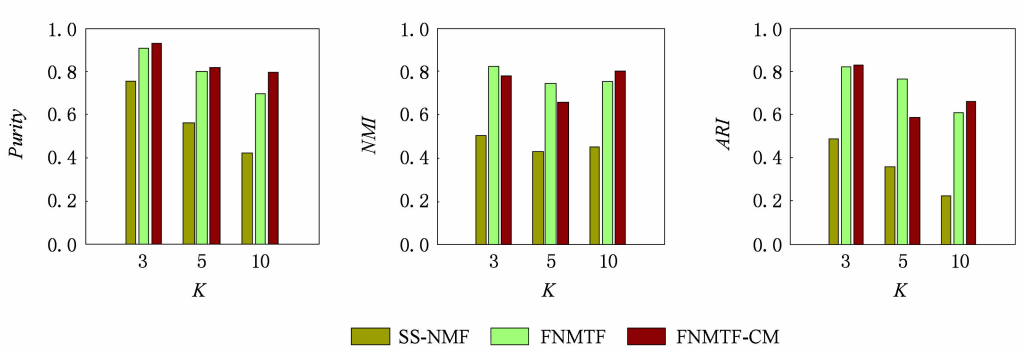

Fig. 2 The results of algorithms on the different clustering number K ($N=200$, $Error=12\%$).

图 2 算法在不同聚类数 K 下的对比结果($N=200$,$Error=12\%$)

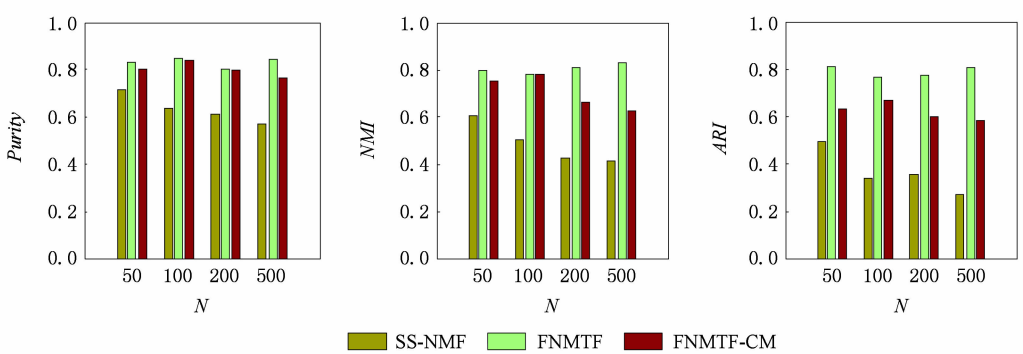

Fig. 3 The results of algorithms on the different data scale N ($K=5$, $Error=12\%$).

图 3 算法在不同节点规模 N 下的对比结果($K=5$,$Error=12\%$)

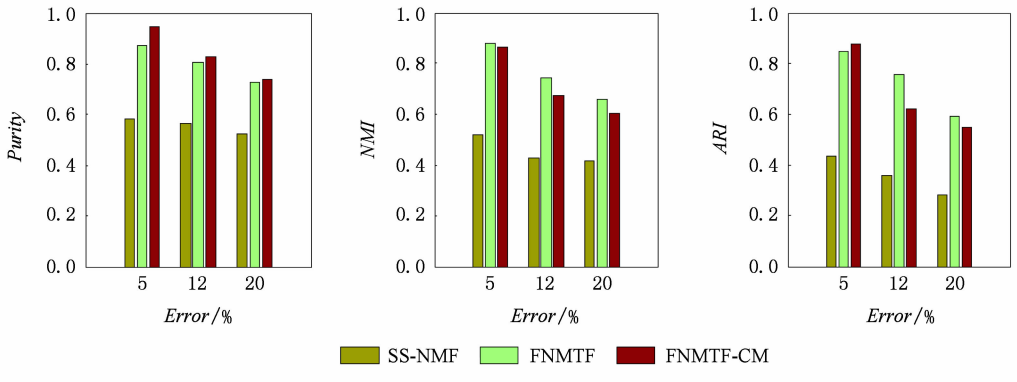

Fig. 4 The results of algorithms on the different clustering difficulty ($N=200, K=5$).

图 4 算法在不同聚类难度下的对比结果($N=200, K=5$)

本文算法在处理不同聚类难度等级的数据集时的鲁棒性都优于其他 2 种算法,这主要是在 FNMTF-CM 算法中降低了数据集本身结构的影响.

4.4 对比实验

为了验证本文算法的效果,在真实的异构稀疏数据集上进行对比实验,在该实验中,设置的聚类数目如表 1 中所示.

在 4 个不同的数据集上对比的实验结果分别如表 2 至表 5 所示,表中对应的最好结果分别加粗表示.

由实验结果可知,FNMTF-CM 算法在 4 个数据集上的结果整体优于其他 2 个算法. 在 Title 和 Weibo 两个数据集上,FNMTF-CM 算法的纯度、*NMI* 值和 *ARI* 值比其他算法都高. 这主要得益于本文算法中基于关联矩阵进行分解,提高了待分解矩阵的稠密度,进而提高了整体算法的准确率.

Table 2 The Result on Title Dataset

表 2 Title 数据集上的对比结果

Algorithms	*Purity*	*NMI*	*ARI*
SS-NMF	0.432 1	0.216 4	0.107 9
FNMTF	0.390 9	0.212 8	0.109 3
FNMTF-CM	**0.541 9**	**0.372 2**	**0.232 6**

Table 3 The Result on Weibo Dataset

表 3 Weibo 数据集上的对比结果

Algorithms	*Purity*	*NMI*	*ARI*
SS-NMF	0.545 7	0.551 2	0.452 2
FNMTF	0.601 1	0.620 1	0.463 7
FNMTF-CM	**0.795 1**	**0.806 2**	**0.7144**

Table 4 The Result on DBLP1 Dataset

表 4 DBLP1 数据集上的对比结果

Algorithms	*Purity*	*NMI*	*ARI*
SS-NMF	0.600 0	**0.389 0**	0.200 8
FNMTF	0.493 3	0.178 4	0.057 5
FNMTF-CM	**0.620 0**	0.361 3	**0.223 2**

Table 5 The Result on DBLP2 Dataset

表 5 DBLP2 数据集上的对比结果

Algorithms	*Purity*	*NMI*	*ARI*
SS-NMF	0.453 3	**0.198 7**	**0.099 9**
FNMTF	0.440 0	0.158 7	0.042 9
FNMTF-CM	**0.466 7**	0.170 0	0.052 8

在数据集 DBLP1 和 DBLP2 上,SS-NMF 算法的 *NMI* 值比本文算法要高,特别是在数据集 DBLP2 上,SS-NMF 算法的 *ARI* 值也高于本文算法. 通过分析可知,DBLP2 数据集异常稀疏,该数据集中可能包含较多的奇异点. 因此,本文算法在处理奇异点问题上仍有待进一步改进.

进一步分析 FNMTF-CM 算法在不同聚类数目 K 值下的效果. 在 Title 数据集中,真实的聚类数目 $K=9$. 该实验中通过调整 K 值,得到的实验结果如图 5 所示. 由实验结果可知,FNMTF-CM 算法在不同的 K 值下,纯度和 *NMI* 值较为稳定,但是在真实的聚类数目下并不是最佳的结果. 因此,在实际应用中,需要根据数据选择恰当的聚类数目.

为了说明算法在处理大规模异构关系数据时的处理速度,在 Weibo 数据集上对 3 个算法的运行时间进行对比,结果如图 6 所示.

FNMTF-CM 算法的运行时间要小于其他 2 种算法,主要因为 FNMTF-CM 算法中选择规模较小的一类实体对应的关联矩阵进行分解,并且求解异构关系矩阵时无需采用乘法更新迭代求解. 在微博关系数据中,标签数目比消息数目小很多,因此关联

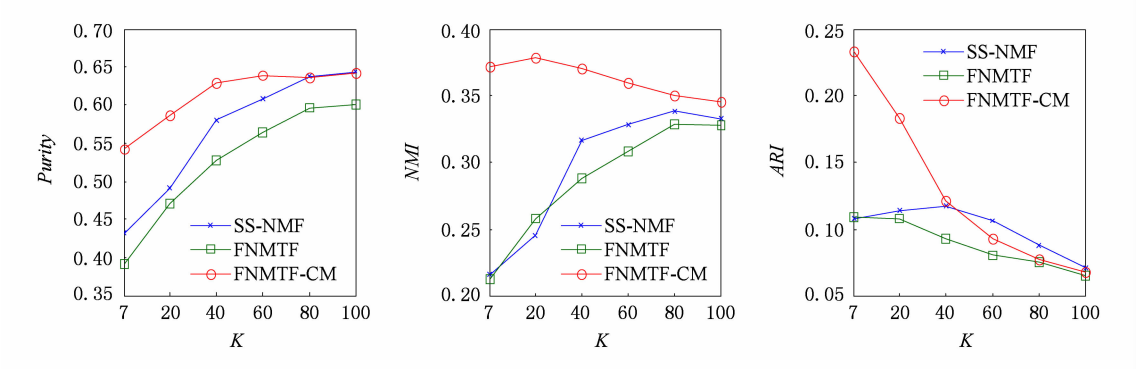

Fig. 5 The results of algorithms on different clustering number K when running on Title.

图 5 在 Title 数据集上不同聚类数目 K 的对比结果

矩阵的规模比原始异构关系矩阵小、处理速度更快.由于 SS-NMF 算法采用乘法更新迭代求解,收敛较慢,其运行时间最长.

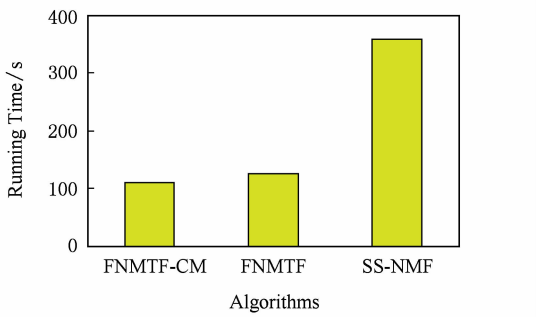

Fig. 6 The running time of algorithms on Weibo.

图 6 算法在 Weibo 数据集上的时间对比

5 结 论

本文针对大规模异构数据中的非平衡问题和稀疏性问题提出了一种基于非负矩阵分解的联合聚类算法.将传统的联合聚类算法转换成基于关联矩阵的对称分解和基于关系矩阵的三分解,进而实现快速的异构数据联合聚类.实验结果表明本文提出的算法在标准测试数据集和真实异构数据上的效果整体优于其他的算法.

本文算法主要考虑了较小规模实体对聚类的促进作用,下一步将考虑 2 类实体的相互促进作用.此外,本文只考虑了二阶异构关系,下一步将推广到高阶异构关系数据联合聚类.

参 考 文 献

[1] Tanay A, Sharan R, Shamir R. Biclustering algorithms: A survey [J]. IEEE Trans on Computational Biology and Bioinformatics, 2004, 1(1): 24-45

[2] Kemal E, Mehmet D, Onur K, et al A comparative analysis of biclustering algorithms for gene expression data [J]. Briefings in Bioinformatics, 2013, 14(3): 279-292

[3] Inderjit S D, Mallela S, Modha D S. Information-theoretic co-clustering [C] //Proc of the 9th ACM SIGKDD. New York: ACM, 2003: 89-98

[4] Inderjit S D. Co-clustering documents and words using bipartite spectral graph partitioning [C] //Proc of the 7th ACM SIGKDD. New York: ACM, 2001: 269-274

[5] Li Tao, Ding Chris. Non-Negative Matrix Factorizations for Clustering: A Survey [M] //Data Clustering: Algorithms and Applications. London: Chapman & Hall/CRC, 2013: 149-176

[6] Long Bo, Zhang Zhongfei, Yu P S. Co-clustering by block value decomposition [C] //Proc of the 11th ACM SIGKDD. New York: ACM, 2005: 635-640

[7] Tjhi W C, Chen Lihui, Minimum sum-squared residue for fuzzy co-clustering [J]. Intelligent Data Analysis, 2006, 10(3): 237-249

[8] Li Zhao, Wu Xindong. Weighted nonnegative matrix tri-factorization for co-clustering [C] //Proc of the 23rd IEEE Int Conf on Tools with Artificial Intelligence. Piscataway, NJ: IEEE, 2011: 811-816

[9] Shang Fanhua, Jiao Licheng, Wang Fei. Graph dual regularization non-negative matrix factorization for co-clustering [J]. Pattern Recognition, 2002, 45(6): 2237-2250

[10] Salunke A, Liu Xumin, Rege M. Constrained co-clustering with non-negative matrix factorisation [J]. Journal of Business Intelligence and Data Mining, 2012, 7(1/2): 60-79

[11] Chen Yanhua, Rege M, Dong M, et al. Non-negative matrix factorization for semi-supervised data clustering [J]. Knowledge and Information Systems, 2008,17(3): 355-379

[12] Chen Yanhua, Wang Lijun, Dong Ming. Non-negative matrix factorization for semisupervised heterogeneous data coclustering [J]. IEEE Trans on Knowledge and Data Engineering, 2010, 22(10): 1459-1474

[13] Wang Hua, Nie Feiping, Huang Heng, et al. Fast nonnegative matrix tri-factorization for large-scale data co-clustering [C] //Proc of the 22nd Int Joint Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2011: 1553-1558

[14] Li Tao, Ding Chris. The relationships among various nonnegative matrix factorization methods for clustering [C] // Proc of the 6th Int Conf on Data Mining. Piscataway, NJ: IEEE, 2006: 362-371

[15] Gu Quanquan, Zhou Jie. Co-clustering on manifolds [C] // Proc of the 15th ACM SIGKDD. New York: ACM, 2009: 359-368

[16] Li Ping, Bu Jiajun, Chen Chun, et al. Relational co-clustering via manifold ensemble learning [C] //Proc of the 21st CIKM. New York: ACM, 2012:1687-1691

[17] Yan Xiaohui, Guo Jiafeng, Liu Shenghua, et al. Learning topics in short texts by non-negative matrix factorization on term correlation matrix [C] //Proc of the SIAM Int Conf Data Mining. Philadelphia, PA: SIAM, 2013: 749-757

[18] Ding Chris, Li Tao, Peng Wei, et al. Orthogonal nonnegative matrix tri-factorizations for clustering [C] //Proc of the 12th ACM SIGKDD. New York: ACM, 2006: 126-135

[19] Lomet A, Govaert G, Grandvalet Y. Design of artificial data tables for co-clustering analysis [R]. Compiègne, France: Université de Technologie de Compiègne, 2012

[20] Deng Hongbo, Han Jiawei, Zhao Bo, et al. Probabilistic topic models with biased propagation on heterogeneous information networks [C] //Proc of the 17th ACM SIGKDD. New York: ACM, 2011: 1271-1279

[21] Zhao Ying, Karypis G. Criterion functions for document clustering: Experiments and analysis, UMN CS 01-040 [R]. Minnesota, AK: University of Minnesota, 2001

[22] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions [J]. Journal of Machine Learning Research, 2003, 3: 583-617

[23] Hubert L, Arabie P. Comparing partitions [J]. Journal of Classification, 1985, 2(1): 193-218

Shen Guowei, born in 1986. PhD candidate at Harbin Engineering University. His main research interests include data mining, social computing, etc.

Yang Wu, born in 1974. Professor and PhD supervisor at Harbin Engineering University. His main research interests include data mining, information security, etc (yangwu@hrbeu.edu.cn).

Wang Wei, born in 1974. PhD and associate professor at Harbin Engineering University. His main research interests include data mining, information security, etc (w_wei@hrbeu.edu.cn).

Yu Miao, born in 1987. PhD candidate at Harbin Engineering University. His main research interests include data mining, social computing, etc (yumiao@hrbeu.edu.cn).

Dong Guozhong, born in 1989. PhD candidate at Harbin Engineering University. His main research interests include data mining, social computing, etc (dongguozhong@hrbeu.edu.cn).