基于流形正则化的多类型关系数据联合聚类方法

黄梦婷 张 灵 姜文超
(广东工业大学计算机学院 广州 510006)

摘 要 随着大数据应用的发展,通过非线性流形采样得到的多类型关系数据规模越来越大,数据几何结构更加复杂,异构关系数据变得异常稀疏,导致数据挖掘难度增大且准确率降低。针对上述问题,提出一种基于流形非负矩阵三分解的多类型关系数据联合聚类方法:首先,对于较小规模的实体,根据其自然关系或内容相关性构造关联矩阵,对其分解后得到该类实体的聚类指示矩阵,将其作为非负矩阵三分解的输入;然后,在快速非负矩阵三分解(FNMTF)的基础上加入流形正则化处理,实现数据类型间关系与类型内部关系的联合聚类,进一步提高聚类的准确率。实验表明:在准确率和整体性能方面,流形非负矩阵三分解算法优于传统的基于非负矩阵分解的联合聚类算法。

关键词 多类型关系数据,流形正则化,非负矩阵分解,关联矩阵
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.06.008

Multi-type Relational Data Co-clustering Approach Based on Manifold Regularization

HUANG Meng-ting ZHANG Ling JIANG Wen-chao
(School of Computers, Guangdong University of Technology, Guangzhou 510006, China)

Abstract With the development of big data applications, the size of multi-type relational data sampled from nonlinear manifolds is getting larger. The data geometric structure is more complicated, and the heterogeneous relational data are becoming extremely sparse. As a result, data mining becomes more difficult and less accurate. In order to solve this problem, this paper proposed a manifold nonnegative matrix tri-factorization (MNMTF) approach for multi-type relational data co-clustering. First of all, the correlation matrix is constructed with the natural relationship or content relevance of smaller-scale entities and it is decomposed into indicating matrix. The indicating matrix is used as the input of nonnegative matrix tri-factorization. Then, the manifold regularization is added on the basis of fast nonnegative matrix tri-factorization (FNMTF) to simultaneously cluster data inter-type relationships and intra-type relationships, improving the accuracy of clustering. Experiments show that the accuracy and performance of MNMTF algorithm are superior to the traditional co-clustering algorithms based on nonnegative matrix factorization.

Keywords Multi-type relational data, Manifold regularization, Nonnegative matrix factorization, Correlation matrix

1 引言

快速发展的移动互联网相关技术导致大数据爆炸式增长,其数据结构更加丰富多样,同时产生彼此相关的多种类型的数据实体。例如,在搜索引擎系统中,存在4种不同类型的数据实体,包括单词、网页、搜索查询和用户。这4类数据实体分别拥有各自的属性,不同类型数据之间存在复杂的相互关联。区别于单一类型的同质关系数据,多类型关系数据包含了结构更加丰富的信息,考虑了数据集的两种形式,即类型间关系和类型内部关系。类型间关系指的是不同类型实体间的异构关系;类型内部关系指的是同类型实体间的相关性。传统的聚类算法仅关注异构数据中的类型间关系。但是,已有研究[1-2]表明许多真实世界的数据实际上是从非线性流形中采样而来。数据几何结构不仅仅体现在数据的类型间关系,而且还包括类型内部关系。

联合聚类的目标是针对不同类型的实体同时进行聚类分析。Ailem等[3]提出算法Coclus来聚类单词和文档,通过迭代交替优化使模块最大化,从而获得更好的聚簇;Honda等[4]把增量算法推广到基于共生矩阵的模糊联合聚类,通过把单程和在线方法应用到分类多元数据模糊聚类和文档与关键字的模糊联合聚类,使共簇的聚合度最大化,证明了增量方法在模糊联合聚类中的适用性;Lee等[5-6]提出了非负矩阵分解

到稿日期:2018-05-06 返修日期:2018-09-14 本文受广东省自然科学基金项目(2016A030313703),广东省科技计划项目(2016B030305002,2017B030305003,2017B010124001),广东省产学研合作项目(2017B090901005)资助。

黄梦婷(1994-),女,硕士生,CCF会员,主要研究方向为数据挖掘与分析;**张 灵**(1968-),女,博士,教授,主要研究方向为智能化信息处理、自动化装备、人工智能和计算机视觉等;**姜文超**(1977-),男,博士,讲师,主要研究方向为云计算、高性能计算、分布式系统等,E-mail:june4567@21cn.com(通信作者)。

(NMF)方法，但在实际应用中，两因子分解法得到的近似低秩矩阵的效果较差；Ding 等[7-8]发现了非负矩阵分解与 K-means/谱聚类之间的关系，因此提出了非负矩阵三分解(NMTF)方法用于单词和文档的联合聚类。数据中一般都是非负元素，因此非负矩阵分解方法已成为目前最常用的方法之一。

非负矩阵三分解能够将不同类型的数据进行联合聚类，已经得到了广泛应用。在此基础上，研究者们提出了一系列改进的非负矩阵三分解方法来实现联合聚类[9-10]，但是这些算法都忽略了数据中的几何结构。Gu 等[11]提出了双正则联合聚类(DRCC)方法，即把数据构建成图，基于流形正则化探索其几何结构。已有的研究[12-13]表明，通过流形正则化挖掘数据内部关联的信息，能够提高联合聚类的准确率。然而，基于非负矩阵三分解的联合聚类存在着计算速度慢的问题，算法中每个迭代步骤都涉及到大量的矩阵乘法，因此很难运用到大规模的实际应用数据中。Wang 等[14]提出了快速的非负矩阵三分解(FNMTF)方法来实现快速的矩阵分解，进而实现联合聚类。当异构数据规模继续增大时，不同实体的规模并不呈现统一的增长模式，关系数据也比较稀疏。为了解决数据增长的非平衡问题和稀疏问题，申国伟等[15]提出了 FNMTF-CM 算法。但是此算法却忽略了数据中的几何结构，丢失了数据类型内部关系这一有价值的信息。

通过考虑数据中的几何结构，提出一种基于流形正则化的多类型关系数据联合聚类算法。首先，构造一个关联矩阵进行非负矩阵分解，得到聚类指示矩阵作为关系矩阵分解的输入；然后，在快速非负矩阵三分解的基础上加入流形正则约束，实现数据类型间关系与类型内部关系的联合聚类，进一步提高聚类的准确率。

2　相关工作

2.1　问题描述

真实世界的数据是从非线性低维流形中采样，然后嵌入到高维空间中。其中，二类型关系数据是最常见的多类型关系数据，因此，本文以二类型关系数据为例进行叙述。对数据中的两类实体 $X_1=\{x_{1\cdot},\cdots,x_{m\cdot}\}$ 和 $X_2=\{x_{\cdot 1},\cdots,x_{\cdot n}\}$ 分别构建两个图 G_f 和 G_g，其中实体 X_1 和 X_2 的样本数量分别为 m 和 n，基于大规模数据中的非平衡问题[15]，可假设 $m\gg n$。关系矩阵 $\boldsymbol{X}_{m\times n}$ 描述两类实体的类型间关系，关联矩阵 $\boldsymbol{W}_f$ 和 $\boldsymbol{W}_g$ 分别描述实体类型内部关系。$\boldsymbol{D}_f$ 和 $\boldsymbol{D}_g$ 分别为实体 X_1 和 X_2 的度矩阵，$\boldsymbol{L}_f$ 和 $\boldsymbol{L}_g$ 分别为实体 X_1 和 X_2 的拉普拉斯矩阵。联合聚类算法中将 X_1 和 X_2 分别划分成 c 类和 d 类(通常 $c=d$)，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 $\boldsymbol{X}$ 的行和列同时进行划分的问题。其中，聚类指示矩阵 $\boldsymbol{F}\in\{0,1\}^{m\times c}$ 描述实体 X_1 的聚类结果，如果 $x_{i\cdot}$ 属于聚类 c_j，$\boldsymbol{F}_{ij}=1$，否则 $\boldsymbol{F}_{ij}=0$。同理可得 X_2 的聚类指示矩阵 $\boldsymbol{G}\in\{0,1\}^{m\times d}$。

3　基于流形正则化的联合聚类方法

基于流形正则化的联合聚类方法(MNMTF)方法分为两个阶段，分别对关联矩阵和关系矩阵进行分解。首先，将较小规模的实体构造成关联矩阵 $\boldsymbol{W}_g$，基于非负矩阵三分解得到聚类指示矩阵 $\boldsymbol{G}$，将其作为关系矩阵 $\boldsymbol{X}$ 基于非负矩阵三分解的输入；然后，在关系矩阵分解的基础上加入流形正则化处理，最终分解得到聚类指示矩阵 $\boldsymbol{F}$。

3.1　非负矩阵分解框架

基于流形正则化的联合聚类方法分为两个阶段，分别对关联矩阵 $\boldsymbol{W}_g$ 和关系矩阵 $\boldsymbol{X}$ 进行分解，分解框架如图 1 所示。

$$\boxed{\boldsymbol{W}_g}\approx\boxed{\boldsymbol{G}}\times\boxed{\boldsymbol{H}}\times\boxed{\boldsymbol{G}^{\mathrm{T}}}$$

(a)关联矩阵分解

$$\boxed{\boldsymbol{X}}\approx\boxed{\boldsymbol{F}}\times\boxed{\boldsymbol{S}}\times\boxed{\boldsymbol{G}^{\mathrm{T}}}$$

(b)关系矩阵分解

图 1　关联矩阵与关系矩阵分解示意图

Fig. 1　Schematic diagram of correlation matrix and relationship matrix decomposition

先将样本数较小的一类实体 X_2 所构造的关联矩阵 $\boldsymbol{W}_g$ 对称分解为 $\boldsymbol{G},\boldsymbol{H},\boldsymbol{G}^{\mathrm{T}}$ 3 个矩阵，使得 $\boldsymbol{W}_g\approx\boldsymbol{G}\boldsymbol{H}\boldsymbol{G}^{\mathrm{T}}$，如图 1(a)所示。其中，矩阵 $\boldsymbol{G}$ 为小规模实体的聚类指示矩阵，矩阵 $\boldsymbol{H}$ 为具有一定自由度的平衡矩阵，当对 $\boldsymbol{W}_g$ 进行分解时，$\boldsymbol{H}$ 提供的自由度可以保证低维矩阵表示的准确性[8]。

然后将关联矩阵 $\boldsymbol{W}_g$ 分解所得的聚类指示矩阵 $\boldsymbol{G}$ 作为关系矩阵 $\boldsymbol{X}$ 分解的输入。将关系矩阵 $\boldsymbol{X}$ 分解为 $\boldsymbol{F},\boldsymbol{S},\boldsymbol{G}$ 3 个矩阵，使得 $\boldsymbol{X}\approx\boldsymbol{F}\boldsymbol{S}\boldsymbol{G}^{\mathrm{T}}$，如图 1(b)所示。其中，矩阵 $\boldsymbol{F}$ 和 $\boldsymbol{G}$ 分别为两类实体的聚类指示矩阵，矩阵 $\boldsymbol{S}$ 为具有一定自由度的平衡矩阵。

3.2　基于异构相关性的关联矩阵构造

关系矩阵 $\boldsymbol{X}$ 的构造遵循两类实体之间的自然关系。例如对于文档和单词的关系矩阵 $\boldsymbol{X}$ 的构造，如果文档 $x_{i\cdot}$ 中出现单词 $x_{\cdot j}$，则 $\boldsymbol{X}_{ij}=1$，否则 $\boldsymbol{X}_{ij}=0$。

关联矩阵 $\boldsymbol{W}$ 的构造分两种情况：1)实体样本之间存在显著的类型内部关系，例如网页之间的链接，此时关联矩阵 $\boldsymbol{W}$ 就按照样本之间的自然关系构造；2)实体样本之间不存在显著的类型内部关系，例如文档中的单词，此时关联矩阵 $\boldsymbol{W}$ 的构造就需要借助异构实体来实现。

两个样本之间的关联强度的计算方法如式(1)所示。

$$w_{ij}=S(x_i,x_j) \tag{1}$$

其中，$S(x_i,x_j)$ 为实体 X_1(或 X_2)中的样本 x_i 和 x_j 在异构实体 X_2(或 X_1)中同时出现的次数。

由于不同实体样本的基数可能会存在很大的差异，构造出来的关联矩阵的值也会存在数量级的差别，因此需要进一步对关联矩阵进行标准化。改进后的关联强度的计算方法如式(2)所示。

$$w_{ij}=\frac{S(x_i,x_j)}{\sum\limits_{a,b\in X}S(x_a,x_b)} \tag{2}$$

3.3　关联矩阵对称三分解

通过同类实体的类型内部关系构造的关联矩阵 $\boldsymbol{W}$ 比关系矩阵 $\boldsymbol{X}$ 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[15]。

先对样本数较少的实体 X_2 构造的关联矩阵 $\boldsymbol{W}_g$ 进行对称非负矩阵三分解，得到聚类指示矩阵 $\boldsymbol{G}$。由于正交条件的约束，传统的非负矩阵三分解得到的目标矩阵不是只包含 0 和 1 的聚类指示矩阵，还需要额外的后期处理，并且在每次交替迭代计算中涉及大量的矩阵乘法，计算代价太高。为了解决这些问题，本文采用 Wang 等[14]提出的快速非负矩阵三分解方法来实现矩阵的快速分解，如式(3)所示。

$$J_1=\|\boldsymbol{W}_g-\boldsymbol{GHG}^{\mathrm{T}}\|^2,\ \mathrm{s.t.}\ \boldsymbol{G}\in\{0,1\}^{n\times d} \tag{3}$$

对于式(3)的求解，本文采用选择性求解变量 $\boldsymbol{G},\boldsymbol{H}$ 的方式。首先，求解矩阵 $\boldsymbol{H}$ 的过程中，把矩阵 $\boldsymbol{G}$ 看作已知条件，矩阵 $\boldsymbol{H}$ 的求解如式(4)所示。

$$\boldsymbol{H}=(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{W}_g\boldsymbol{G}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1} \tag{4}$$

然后，求解矩阵 $\boldsymbol{G}$ 的过程中把矩阵 $\boldsymbol{GH}$ 当作一个整体，对矩阵 $\boldsymbol{G}$ 的求解问题进行优化，如式(5)所示。

$$\min_{\boldsymbol{G}\in\{0,1\}}\|\boldsymbol{W}_{\cdot i}-\boldsymbol{GH}\boldsymbol{g}_{i\cdot}^{\mathrm{T}}\|^2 \tag{5}$$

其中，向量 $\boldsymbol{w}_{\cdot i}$ 表示矩阵 $\boldsymbol{W}_g$ 的第 i 列，向量 $\boldsymbol{g}_{i\cdot}$ 表示矩阵 $\boldsymbol{G}$ 的第 i 行。在向量 $\boldsymbol{g}$ 中，有且只有一个元素为 1，其余为 0，因此式(5)的进一步处理如式(6)所示。

$$\boldsymbol{g}_{ij}=\begin{cases}1, & j=\arg\min_k\|\boldsymbol{w}_{\cdot i}-\widetilde{\boldsymbol{h}}_{\cdot k}\|^2\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

其中，$\widetilde{\boldsymbol{H}}=\boldsymbol{GH}$，向量 $\widetilde{\boldsymbol{h}}_{\cdot k}$ 表示矩阵 $\widetilde{\boldsymbol{H}}$ 的第 k 列。式(6)简单地列举了向量 $\widetilde{\boldsymbol{h}}$ 的模，并找到其中值最大的一个。

3.4 流形异构关系矩阵三分解

基于 FNMTF 方法的关系矩阵分解如式(7)所示。

$$\begin{aligned}&J_2=\|\boldsymbol{X}-\boldsymbol{FSG}^{\mathrm{T}}\|^2,\\&\mathrm{s.t.}\ \boldsymbol{F}\in\{0,1\}^{m\times c},\boldsymbol{G}\in\{0,1\}^{n\times d}\end{aligned} \tag{7}$$

算法 FNMTF 假定样本数据是从欧氏空间中采样得来的，忽略了数据中的几何结构。根据流形假设，如果两个样本 $x_{i\cdot}$ 和 $x_{j\cdot}$ 在几何结构中相近，那么这两个样本的现实意义也相近，在聚类中体现为两个样本的聚类标签相近。

对实体 X_1 构建图 G_f，样本 $x_{i\cdot}$ 和 $x_{j\cdot}$ 的聚类标签分别为 $\boldsymbol{f}_{i\cdot}$ 和 $\boldsymbol{f}_{j\cdot}$。实体 X_1 中所有样本的标签距离总和如式(8)所示。

$$\begin{aligned}&\frac{1}{2}\sum_{i,j}\|\boldsymbol{f}_{i\cdot}-\boldsymbol{f}_{j\cdot}\|^2\boldsymbol{w}_{ij}^f\\&=\sum_{i,j}\boldsymbol{f}_{i\cdot}\boldsymbol{w}_{ij}^f\boldsymbol{f}_{i\cdot}^{\mathrm{T}}-\sum_{i,j}\boldsymbol{f}_{i\cdot}\boldsymbol{w}^{ij}\boldsymbol{f}_{j\cdot}^{\mathrm{T}}\\&=\sum_i\boldsymbol{f}_{i\cdot}\boldsymbol{D}_{ii}^f\boldsymbol{f}_{i\cdot}^{\mathrm{T}}-\sum_{i,j}\boldsymbol{f}_{i\cdot}\boldsymbol{w}_{ij}^f\boldsymbol{f}_{j\cdot}^{\mathrm{T}}\\&=\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}(\boldsymbol{D}_f-\boldsymbol{W}_f)\boldsymbol{F})\\&=\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_f\boldsymbol{F})\end{aligned} \tag{8}$$

其中，拉普拉斯矩阵 $\boldsymbol{L}_f=\boldsymbol{D}_f-\boldsymbol{W}_f$，度矩阵 $\boldsymbol{D}_f$ 为对角矩阵，$\boldsymbol{D}_{ii}^f=\sum_j\boldsymbol{w}_{ij}^f$。

在流形中，聚类标签越平滑，式(8)的值就越小。同理可得实体 X_2 的流形正则约束 $\mathrm{tr}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}_g\boldsymbol{G})$。

在式(7)的基础上加入流形正则约束，不仅考虑两类实体的类型间关系，而且还考虑同类实体的类型内部关系。基于流形正则化的关系矩阵分解如式(9)所示。

$$\begin{aligned}&J_3=\|\boldsymbol{X}-\boldsymbol{FSG}^{\mathrm{T}}\|^2+\lambda\mathrm{tr}(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{L}_f\boldsymbol{F})+\varphi\mathrm{tr}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{L}_g\boldsymbol{G})\\&\mathrm{s.t.}\ \boldsymbol{F}\in\{0,1\}^{m\times c},\boldsymbol{G}\in\{0,1\}^{n\times d}\end{aligned} \tag{9}$$

其中，λ 和 $\varphi>0$ 为正则化参数，用于平衡式(9)第一项因聚类重构产生的误差和第二、三项聚类标签的平滑度。标准化的拉普拉斯矩阵 $\boldsymbol{L}_f=\boldsymbol{I}-\boldsymbol{D}_f^{-1}\boldsymbol{W}_f$，$\boldsymbol{L}_g=\boldsymbol{I}-\boldsymbol{D}_g^{-1}\boldsymbol{W}_g$。

对于式(9)的求解，本文采用选择性求解变量 $\boldsymbol{F},\boldsymbol{S}$ 的方式。矩阵 $\boldsymbol{G}$ 的求解如式(6)所示，此处 $\boldsymbol{G}$ 直接作为 J_3 的输入。

首先，在求解矩阵 $\boldsymbol{S}$ 的过程中，把矩阵 $\boldsymbol{F}$ 和 $\boldsymbol{G}$ 看作已知条件。矩阵 $\boldsymbol{S}$ 的求解如式(10)所示。

$$\boldsymbol{S}=(\boldsymbol{F}^{\mathrm{T}}\boldsymbol{F})^{-1}\boldsymbol{F}^{\mathrm{T}}\boldsymbol{XG}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1} \tag{10}$$

然后，在求解矩阵 $\boldsymbol{F}$ 的过程中把矩阵 $\boldsymbol{SG}^{\mathrm{T}}$ 当作一个整体，对矩阵 $\boldsymbol{F}$ 的求解问题进行优化，如式(11)所示。

$$\min_{\boldsymbol{F}\in\{0,1\}}\|\boldsymbol{x}_{j\cdot}-\boldsymbol{f}_{j\cdot}\boldsymbol{SG}^{\mathrm{T}}\|^2 \tag{11}$$

其中，向量 $\boldsymbol{x}_{j\cdot}$ 表示矩阵 $\boldsymbol{X}$ 的第 j 行，向量 $\boldsymbol{f}_{j\cdot}$ 表示矩阵 $\boldsymbol{F}$ 的第 j 行。在向量 $\boldsymbol{f}$ 中，有且只有一个元素为 1，其余为 0，因此式(11)的进一步处理如式(12)所示。

$$\boldsymbol{f}_{ji}=\begin{cases}1, & i=\arg\min_p\|\boldsymbol{x}_{j\cdot}-\widetilde{\boldsymbol{g}}_{p\cdot}\|^2\\ 0, & \text{otherwise}\end{cases} \tag{12}$$

其中，$\widetilde{\boldsymbol{G}}=\boldsymbol{SG}^{\mathrm{T}}$，向量 $\widetilde{\boldsymbol{g}}_{p\cdot}$ 表示矩阵 $\widetilde{\boldsymbol{G}}$ 的第 p 行。式(12)简单地列举了向量 $\widetilde{\boldsymbol{g}}$ 的模，并找到其中值最大的一个。

针对式(9)中的目标函数 J_3 的计算过程，MNMTF 算法设计如下。

算法 1 流形异构数据联合聚类算法 MNMTF

输入：关系矩阵 $\mathbf{X}$，关联矩阵 $\mathbf{W}_f$ 和 $\mathbf{W}_g$，正则化参数 λ 和 φ，最大迭代次数 Niter，收敛阈值 δ

输出：实体 X_1 的聚类指示矩阵 $\mathbf{F}$，实体 X_2 的聚类指示矩阵 $\mathbf{G}$

1. 初始化 $\mathbf{F},\mathbf{G}$；
2. while(iter<Niter，$\Delta J<\delta$)
3. 根据 $\mathbf{H}=(\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{W}_g\mathbf{G}(\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}$ 计算 $\mathbf{H}$；
4. 根据 $\mathbf{g}_{ij}=\begin{cases}1, & j=\arg\min_k\|\mathbf{w}_{\cdot i}-\widetilde{\mathbf{h}}_{\cdot k}\|^2\\ 0, & \text{otherwise}\end{cases}$ 计算 $\mathbf{G}$；
5. iter=iter+1；
6. 求解 $J_1=\|\mathbf{W}_g-\mathbf{GHG}^{\mathrm{T}}\|^2$；
7. end while
8. while(iter<Niter，($J<\delta$)
9. 根据 $\mathbf{S}=(\mathbf{F}^{\mathrm{T}}\mathbf{F})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{XG}(\mathbf{G}^{\mathrm{T}}\mathbf{G})^{-1}$ 计算 $\mathbf{S}$；
10. 根据 $\mathbf{f}_{ji}=\begin{cases}1, & i=\arg\min_p\|\mathbf{x}_{j\cdot}-\widetilde{\mathbf{g}}_{p\cdot}\|^2\\ 0, & \text{otherwise}\end{cases}$ 计算 $\mathbf{F}$；
11. iter=iter+1；
12. 求解 $J_3=\|\mathbf{X}-\mathbf{FSG}^{\mathrm{T}}\|^2+\lambda\mathrm{tr}(\mathbf{F}^{\mathrm{T}}\mathbf{L}_f\mathbf{F})+\varphi\mathrm{tr}(\mathbf{G}^{\mathrm{T}}\mathbf{L}_g\mathbf{G})$；
13. end while

4 实验及分析

实验分别对 FNMTF，FNMTF-CM 和本文算法 MNMTF 进行测试分析与比较。每一组实验分别运行 50 次，采用随机初始化方式，最终实验结果取平均值。

4.1 实验数据集

本文将在 Webkb5 和 TTC 两个异构稀疏数据集上测试 MNMTF 方法。下面从数据集大小和数据来源等方面分别介绍这两个数据集。

Webkb5 数据提供了从 4 所大学收集的网页信息，构建了网页和单词之间的异构关系数据集。对数据进行预处理，删除了停用词和文档频率小于 10 的单词，最终留下了 1703 个单词。TTC 数据是来自土耳其 6 个知名门户网站的新闻数据集。对其进行预处理，删除了停用词，最终留下了 4813 个单词。两个数据集的详细信息如表 1 所列。

表 1　异构关系数据集

Table 1　Heterogeneous relational dataset

数据集	实体 1	实体 2	聚类数	稀疏度
Webkb5	877	1703	5	0.05
TTC	3600	4813	6	0.01

为了使算法 MNMTF 取得最优聚类性能，首先在稀疏的测试数据集上对算法进行评估，选取合适的正则化参数 λ 和 φ。测试数据集取 Webkb5 数据中的一所大学的网页信息，在原数据集的基础上删除了文档频率为 0 的单词，最终留下了 1588 个单词，详细信息如表 2 所列。

表 2　测试数据集

Table 2　Test dataset

数据集	实体 1	实体 2	聚类数	稀疏度
Test	195	1588	5	0.06

4.2 评估指标

将采用常见的 Purity，NMI(Normalized Mutual Information)，ARI(Adjusted Rand Index)这 3 个评估指标作为度量标准。Purity 计算正确聚类的文档数占总文档数的比例，NMI 度量两个聚类结果的相近程度，ARI 衡量聚类结果与真实情况的吻合程度，其定义分别如式(13)－式(15)所示。

$$Purity(A,B)=\frac{1}{n}\sum_i \max_j |a_i \cap b_j| \tag{13}$$

其中，$A=\{a_1,\cdots,a_n\}$ 是算法求得的聚类标签序列，$B=\{b_1,\cdots,b_n\}$ 是给定的聚类标签序列，n 表示实体样本数。

$$\begin{cases} NMI(A,B)=\dfrac{2\times I(A;B)}{H(A)+H(B)} \\ I(A;B)=H(A)+H(B)-H(A,B) \\ H(A)=-\sum\limits_{a_i\in A}[P(a_i)\lg P(a_i)] \end{cases} \tag{14}$$

其中，$I(A;B)$描述的是 A,B 两个集合之间的相关性，$H(A)$ 描述的是集合 A 中某种特定信息出现的概率，$P(a_i)=\dfrac{|a_i|}{n}$。

$$ARI(A,B)=\frac{\sum\limits_{ij}C^2_{|a_i\cap b_j|}-(\sum\limits_i C^2_{|a_i|}\sum\limits_j C^2_{|b_j|})/C^2_n}{\frac{1}{2}(\sum\limits_i C^2_{|a_i|}+\sum\limits_j C^2_{|b_j|})-(\sum\limits_i C^2_{|a_i|}\sum\limits_j C^2_{|b_j|})/C^2_n} \tag{15}$$

计算得到的 $Purity$，NMI，ARI 越大，则聚类结果越好。

4.3 实验和结果

参数寻优：MNMTF 方法把两类实体构建成两个图，其正则化参数设置为 $\lambda=\varphi$。对于每个不同 λ 值的聚类方法，我们用随机初始化的方式重复实验 50 次。图 2－图 4 显示了算法 MNMTF 在正则化参数 λ 取不同值时的对比结果。

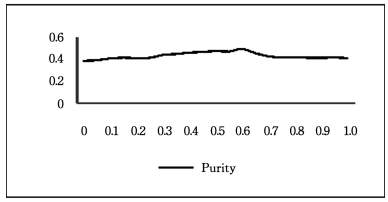

图 2　不同正则化参数 λ 对 Purity 的影响

Fig. 2　Effect of different regularization parameter λ on Purity

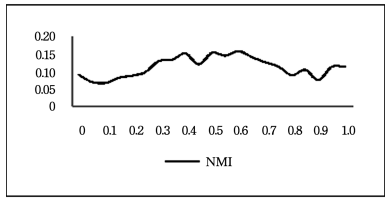

图 3　不同正则化参数 λ 对 NMI 的影响

Fig. 3　Effect of different regularization parameter λ on NMI

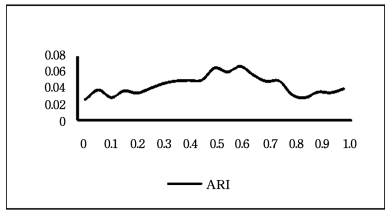

图 4　不同正则化参数 λ 对 ARI 的影响

Fig. 4　Effect of different regularization parameter λ on ARI

综合 3 个评估指标，λ 取值在 0.5～0.6 之间时，算法 MNMTF 的聚类性能比较优异，尤其在 $\lambda=0.6$ 时算法的聚类性能最优；随着 λ 逐渐增大到 1 或者减小到 0，算法的聚类性能整体呈下降趋势。根据以上结果，后续的算法对比实验中，MNMTF 的正则化参数设置为 $\lambda=0.6$。

为了验证算法 MNMTF 的效果，在两个真实的异构稀疏数据集 Webkb5 和 TTC 上进行对比实验，测试 Purity，NMI，ARI 这 3 个算法评估指标。实验结果分别如表 3 和表 4 所列，表中对应的最好结果分别加粗表示。

表 3　各算法在 Webkb5 上的实验结果

Table 3　Experimental results of each algorithm on Webkb5

算法	NMI	ARI	Purity
FNMTF	0.1238	0.0545	0.4322
FNMTF-CM	0.1459	**0.0660**	0.4523
MNMTF	**0.1655**	0.0642	**0.4658**

表 4　各算法在 TTC 上的实验结果

Table 4　Experimental results of each algorithm on TCC

算法	NMI	ARI	Purity
FNMTF	0.0615	0.0128	0.2654
FNMTF-CM	0.1144	**0.0175**	0.3407
MNMTF	**0.1713**	0.0165	**0.3781**

由实验结果可知，MNMTF 算法在两个数据集上的结果

整体优于其他两种算法。对比 FNMTF 算法，FNMTF-CM 和 MNMTF 算法的优势在于基于关联矩阵进行分解，提高了待分解关系矩阵的稠密度。对比 FNMTF-CM 算法，MNMTF 算法的优势在于考虑了异构数据中的类型内部关系，提高了聚类结果的准确性。

由于数据预处理的方式不同，已有的研究[14]中 FNMTF 算法在公开数据集 Webkb 上的实验结果优于本文的实验结果。在网页文本的单词筛选上，本文考虑到对数据的稀疏性要求，删除 Webkb 中的停用词，提高了关系矩阵的稀疏度。为了避免矩阵过于稀疏使稀疏度趋于 0，从而造成单词间的关联性极低，删除了文档频率低于 10 的单词。在聚类数目的设定上，本文选为 5，已有研究[14]中为 4，并且无具体预处理说明，因此实验数据无法统一，为了区别于其他实验的 Webkb 数据集，本文将数据集命名为 Webkb5。由此可得，矩阵稀疏性、聚类数目都会影响同一个算法的实验结果，且稀疏度越高，聚类数目越多，实验效果越不理想。

结束语 针对大规模稀疏异构数据中存在的非线性几何结构，提出了一种基于流形正则化的联合聚类算法。在关联矩阵对称分解和关系矩阵三分解的基础上，加入流形正则化处理，不仅考虑数据中的类型间关系，而且还考虑类型内部关系，进而提高了联合聚类的准确性。实验结果表明本文提出的算法在稀疏异构数据上的效果整体优于其他算法。

算法 MNMTF 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的联合聚类。

参 考 文 献

[1] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

[2] BELKIN M, NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation [J]. Neural Computation, 2003, 15(6): 1373-1396.

[3] AILEM M, ROLE F, NADIF M. Co-clustering document-term matrices by direct maximization of graph modularity[C]// ACM International on Conference on Information and Knowledge Management. New York: ACM Press, 2015: 1807-1810.

[4] HONDA K, TANAKA D, NOTSU A. Incremental algorithms for fuzzy co-clustering of very large cooccurrence matrix[C]// IEEE International Conference on Fuzzy Systems. Piscataway: IEEE Press, 2014: 2494-2499.

[5] LEE D D, SEUNG H S. Learning the parts of objects with non-negative matrix factorization[J]. Nature, 1999, 401(21): 788-791.

[6] LEE D D, SEUNG H S. Algorithms for non-negative matrix factorization[C]// Neural Information Processing Systems. New York: NIPC Press 2000: 535-541.

[7] DING C, HE X, SIMON H D, et al. On the equivalence of non-negative matrix factorization and spectral clustering[C]// SIAM International Conference on Data Mining. Philadelphia: SIAM Press, 2005: 606-610.

[8] DING C, LI T, PENG W, et al. Orthogonal nonnegative matrix tri-factorizations for clustering[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2006: 126-135.

[9] LI Z, WU X. Weighted nonnegative matrix tri-factorization for co-clustering[C]// IEEE International Conference on TOOLS with Artificial Intelligence. Piscataway: IEEE Press, 2011: 811-816.

[10] BUONO N D, PIO G. Non-negative Matrix Tri-Factorization for co-clustering: An analysis of the block matrix[J]. Information Sciences, 2015, 301(20): 13-26.

[11] GU Q, ZHOU J. Co-clustering on manifolds[C]// ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 359-368.

[12] WANG S, HUANG A. Penalized nonnegative matrix tri-factorization for co-clustering[J]. Expert Systems with Applications, 2017, 78(C): 64-73.

[13] WANG S, GUO W. Robust co-clustering via dual local learning and high-order matrix factorization[J]. Knowledge-Based Systems, 2017, 138(15): 176-187.

[14] WANG H, NIE F, HUANG H, et al. Fast nonnegative matrix tri-factorization for large-scale data co-clustering[C]// International Joint Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2011: 1553-1558.

[15] SHEN G, YANG W, WANG W, et al. Large-scale heterogeneous data co-clustering based on nonnegative matrix factorization[J]. Journal of Computer Research and Development, 2016, 53(2): 459-466. (in Chinese)
申国伟，杨武，王巍，等. 基于非负矩阵分解的大规模异构数据联合聚类[J]. 计算机研究与发展，2016，53(2)：459-466.