文章编号:1009-3087(2014)02-0105-06

基于图划分的网状高阶异构数据联合聚类算法

杨欣欣,黄少滨
(哈尔滨工程大学 计算机科学与技术学院,哈尔滨 150001)

摘 要:目前已有的高阶联合聚类算法主要集中于分析星型高阶异构数据,然而实际应用中,存在大量网状高阶异构数据。为了有效挖掘网状高阶异构数据内部隐藏的结构,提出一种基于图划分的高阶联合聚类算法(简称为GPHCC),该算法将网状高阶异构数据的聚类问题转化为多对二部图的最小正则割划分问题。为了降低计算复杂度,将此优化问题转化为半正定问题求解。实验结果表明GPHCC算法优于目前已有的5种2阶联合聚类算法和5种高阶联合聚类算法。

关键词:网状结构;高阶异构数据;联合聚类;谱聚类
中图分类号:TP181 **文献标志码**:A

A Net-structure High-order Heterogeneous Data Co-clustering Algorithm Based on Graph Partitioning

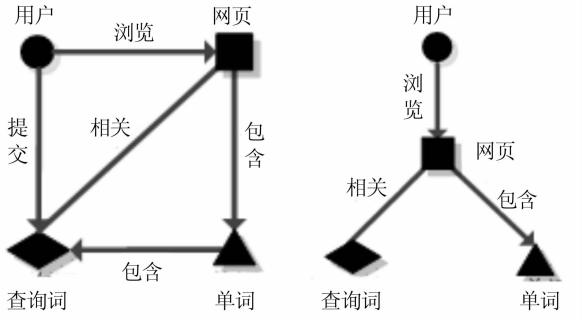
YANG Xinxin, HUANG Shaobin
(College of Computer Sci. and Technol., Harbin Eng. Univ., Harbin 150001, China)

Abstract: Existing high-order co-clustering algorithm just can be suitable for analyzing star-structure high-order heterogeneous data. In order to analyze net-structure high-order heterogeneous data, a high-order co-clustering algorithm based on graph partitioning was proposed. The problem of high-order co-clustering was converted to optimal problem of graph partitioning of minimum normal cut. In order to reduce computational complexity, the optimal problem was converted to semi-definite problem. Experimental studies showed that the qualities of clustering results of GPHCC are superior five pair-wise coclustering algorithms and five high-order co-clustering algorithms.

Key words: net-structure; high-order heterogeneous data; co-clustering; spectral clustering

聚类的目标是将相似的数据划分到同一个聚簇中,将不相似的数据划分到不同的聚簇中。传统的聚类算法主要集中于分析单一类型的同构数据(homogeneous data)[1-3]。近年来,信息技术的发展使得包含多种异质数据的数据集广泛出现[1-3]。

例如,在Web搜索过程中至少包含如图1(a)所示的4种类型数据,分别是单词、页面、查询词和用户。一般将这种异质的、相互关联的数据称为异构数据(heterogeneous data)[1-2]。



(a) 网状结构 (b) 星型结构

图1 Web搜索系统中高阶异构数据示例

Fig. 1 Example of high-order heterogeneous data of Web search system

依据异质数据类型的个数,异构数据分为2阶异构数据和高阶异构数据,数据类型为2种时称为2阶异构数据(pair-wise heterogeneous data)[4];数据类型多于2种时称为高阶异构数据(high-order heterogeneous data)[1]。依据不同类型数据之间关联模

收稿日期:2013-08-27
基金项目:国家自然科学基金资助项目(71272216;60903080;60093009);博士后科学基金资助项目(2012M5100480);国家科技支撑计划资助项目(2009BAH42B02;2012BAH08B02);中央高校基本科研业务费专项基金资助项目(HEUCFZ1212;HEUCFT1208)
作者简介:杨欣欣(1987—),男,博士生. 研究方向:数据挖掘. E-mail:yangxinxin051131@126.com

式的特点,高阶异构数据分为网状结构和星型结构[1],分别参见图1(a)和(b)。与星型结构相比,网状结构具有更好的普适性,星型结构是网状结构的一种简化。

为了有效分析异构数据,提出了异构数据联合聚类方法,同时对多种类型数据进行聚类[2]。初期的研究主要是针对2阶异构数据开展2阶联合聚类(pairwise co-clustering)[3]。典型的算法包括:模糊方法CoDoK[5]、图划分方法BSGB[6]和ICA[7]、信息论方法ITC[8]和矩阵分解方法NBVD[9]等。为了有效探测高阶异构数据内部隐藏的聚类模式,近年来,在2阶聚类方法的基础上,发展了一类针对高阶异构数据的高阶联合聚类方法(high-order co-clustering)[2]。已有的高阶联合聚类算法包括:基于图论的方法[10-11],主要有CBGC[10]和CIHC[11];基于信息论的方法[12-13],例如CIT[12]和AD-HOCC[13];基于k部图学习的方法[3,14],包括RSN[3]和SKGC[14];基于矩阵分解的方法[2,4,15],代表算法有NMTF[2]和SS-NMF[4];基于Goodman Kruskal的方法CoStar[16];Han等[17-18]提出基于排名的聚类方法,代表算法有NetClus[17]和RankClus[18]。这些高阶联合聚类算法仅仅适用于分析星型高阶异构数据,为了分析网状高阶异构数据,需要将其转化为星型结构,例如将图1(a)网状结构数据转化为图1(b)星型结构数据,在此过程中造成用户与查询词的提交关系信息以及查询词与单词的包含关系信息的丢失,影响聚类效果。

为了更有效挖掘网状高阶异构数据内部隐藏的聚类模式,提出一种基于图划分的网状高阶异构数据的联合聚类算法(GPHCC),该算法将网状高阶异构数据聚类问题转化为多对二部图的最小正则割[19]划分问题,但此最优化函数为2次优化问题,时间复杂度较高,为了降低时间复杂度,将此2次最优化问题转化为半正定问题。通过解半正定问题,得出每个数据点的嵌入值,将嵌入值看作空间中的点,运用k-means算法聚类,得出高阶异构数据的聚类结果。

1 基于图划分的高阶联合聚类算法

利用正则割图划分思想,提出适用于分析网状高阶异构数据的聚类算法。首先将高阶联合聚类问题转化为多对二部图的最小正则割划分问题,提出GPHCC算法的目标函数。为了降低时间复杂度将此最优化问题转化为半正定问题,通过求解半正定问题,获得聚类结果。

1.1 目标函数

由 $X^1=\{x_1^1,\cdots,x_{n_1}^1\},X^2=\{x_1^2,\cdots,x_{n_2}^2\},\cdots,X^m=\{x_1^m,\cdots,x_{n_m}^m\}$ 共 m 种类型数据组成的网状高阶异构数据,x_p^i 代表第 i 种类型数据的第 p 个数据,x_q^j 代表第 j 种类型数据的第 q 个数据,$n_i(1\le i\le m)$ 表示第 i 种类型数据 X_i 的数据个数。X^i 与 X^j 的关系矩阵是 $\boldsymbol{R}^{ij}=(r_{pq}^{ij})_{n_i\times n_j}$,$p$ 行 q 列元素 r_{pq}^{ij} 表示 x_p^i 与 x_q^j 之间的关系强度。X^i 与 X^j 的邻接矩阵

$$\boldsymbol{M}^{ij}=\begin{bmatrix}\boldsymbol{0} & \boldsymbol{R}^{ij}\\ \boldsymbol{R}^{ji} & \boldsymbol{0}\end{bmatrix}_{n_i+n_j} \tag{1}$$

其中,$\boldsymbol{0}$ 表示所有元素均为0的块矩阵。设 $\boldsymbol{D}^{ij}\in\mathbb{R}^{(n_i+n_j)\times(n_i+n_j)}$ 为对角矩阵,第 l 行对角元素 $d_{ll}^{ij}=M_{l1}^{ij}+M_{l2}^{ij}+\cdots+M_{l(n_i+n_j)}^{ij}$。从最优化理论的角度出发,首先分别给出任意2种类型数据联合聚类问题的目标函数,再把它们组合成一个全局的目标函数进行线性加权从而构造出全局目标函数。具体地说,由任意2种类型数据 X^i 与 X^j 以及它们之间的关系构建的二部图 $G^{ij}(V^i,V^j,E^{ij})$,数据 X^i 与 X^j 表示为节点集 V^i 和 V^j,X^i 与 X^j 之间的关系表示为边集 E^{ij} 的权重。设 $\boldsymbol{f}_i=[f_1^i,f_2^i,\cdots,f_{n_i}^i]^{\mathrm{T}}$ 为 X^i 的嵌入向量,二部图 $G^{ij}(V^{ij},V^j,E^{ij})$ 的最小正则割划分为最优化问题[19]:

$$\begin{aligned}&\min\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}R_{pq}^{ij}(f_p^i-f_q^j)^2\\ &\text{s.t. }(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})\boldsymbol{D}^{ij}\boldsymbol{e}=0,\\ &(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})\boldsymbol{D}^{ij}(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})^{\mathrm{T}}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{D}^{ij}\boldsymbol{e}\end{aligned} \tag{2}$$

高阶联合聚类算法同时划分多对二部图 G^{ij},使得每对二部图的正则割均最小,显然为多目标优化问题。利用加权组合方法将多目标优化问题转化为单目标优化问题,把每对二部图的正则割损失函数组合成的全局目标函数,有如下最优化问题:

$$\begin{aligned}&\min\sum_{i,j=1}^{m}\beta_{ij}\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}R_{pq}^{ij}(f_p^i-f_q^j)^2\\ &\text{s.t. }(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})\boldsymbol{D}^{ij}\boldsymbol{e}=0;\\ &(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})\boldsymbol{D}^{ij}(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})^{\mathrm{T}}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{D}^{ij}\boldsymbol{e},1\le i,j\le m\end{aligned} \tag{3}$$

其中,β_{ij} 表示 X^i 与 X^j 之间关系的权重,$\sum_{ij}\beta_{ij}=1$。此最优化问题为二次优化问题,时间复杂度较高,文献[20]中证明了二次优化问题可以近似转化为半正定优化问题(semi-definitee problem,SDP)求解,半正定优化问题的一般形式为:

$$\begin{aligned}&\min\boldsymbol{C}\cdot\boldsymbol{W}\\ &\text{s.t. }\boldsymbol{A}_i\cdot\boldsymbol{W}=\boldsymbol{b}_i,\ i=1,2,\cdots,m;\\ &\boldsymbol{W}\succeq 0\end{aligned} \tag{4}$$

其中,$\boldsymbol{C}$ 为目标函数对称系数矩阵,$\boldsymbol{W}$ 为求解的对称矩阵,符号 $\geq$ 表示矩阵为半正定矩阵,$\boldsymbol{A}_i(i=1,\cdots,m)$ 和 $\boldsymbol{b}_i$ 为约束条件的系数矩阵,矩阵内积定义如下:$\boldsymbol{C}\cdot\boldsymbol{W}=\sum_{ij}C_{ij}W_{ij}$。

1.2　优化问题求解

为了降低计算复杂度,将优化问题(3)转化为半正定问题,详细过程如下。设 $n=\sum_{i=1}^{m}n_i$,由 $X^1,\cdots,X^m$ 中任意 2 种类型数据之间的关系矩阵 $\boldsymbol{R}^{ij}$ 组成矩阵 $\boldsymbol{M}\in\mathbb{R}^{s\times s}$:

$$\boldsymbol{M}=\begin{bmatrix}\boldsymbol{0} & \beta_{12}\boldsymbol{R}^{12} & \cdots & \beta_{1m}\boldsymbol{R}^{1m}\\ \beta_{21}\boldsymbol{R}^{21} & \boldsymbol{0} & \cdots & \beta_{2m}\boldsymbol{R}^{2m}\\ \vdots & \vdots & \ddots & \vdots\\ \beta_{m1}\boldsymbol{R}^{m1} & \beta_{m2}\boldsymbol{R}^{m2} & \cdots & \boldsymbol{0}\end{bmatrix}_{n\times n}。$$

其中,$\boldsymbol{0}$ 表示所有元素均为0的块矩阵。设 $\boldsymbol{D}\in\mathbb{R}^{n\times n}$ 为对角矩阵,第 l 行对角元素 $d_{ll}=M_{l1}+M_{l2}+\cdots+M_{ln}$。矩阵 $\boldsymbol{L}=\boldsymbol{M}-\boldsymbol{D}$。设 $\boldsymbol{w}=[\boldsymbol{f}_1^{\mathrm{T}},\boldsymbol{f}_2^{\mathrm{T}},\cdots,\boldsymbol{f}_m^{\mathrm{T}}]^{\mathrm{T}}$,则

$$\sum_{i,j=1}^{m}\beta_{ij}\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}R_{pq}^{ij}\,(f_p^i-f_q^j)^2=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{w}\tag{5}$$

设对角矩阵模板

$$\boldsymbol{J}=\begin{bmatrix}\boldsymbol{J}^{11}_{n_1\times n_1} & 0 & 0 & 0\\ 0 & \boldsymbol{J}^{22}_{n_2\times n_2} & 0 & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & \boldsymbol{J}^{mm}_{n_m\times n_m}\end{bmatrix}_{n\times n}。$$

为了得到矩阵 $\boldsymbol{\Pi}^{ij}$,将对角矩阵块 $\boldsymbol{J}$ 中的对角矩阵 $\boldsymbol{J}^{ii}$ 和 $\boldsymbol{J}^{jj}$ 中的对角元素赋值如下,

$$J_{pp}^{ii}=\sum_{q=1}^{n_j}R_{pq}^{ij},J_{qq}^{jj}=\sum_{p=1}^{n_i}R_{pq}^{ij}\tag{6}$$

矩阵 $\boldsymbol{J}$ 中其余元素为零,将矩阵 $\boldsymbol{J}$ 的值赋予 $\boldsymbol{\Pi}^{ij}$,得到矩阵 $\boldsymbol{\Pi}^{ij}$,则

$$(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})\boldsymbol{D}^{ij}\boldsymbol{e}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}\boldsymbol{e}=0,$$
$$(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})\boldsymbol{D}^{ij}(\boldsymbol{f}_i^{\mathrm{T}},\boldsymbol{f}_j^{\mathrm{T}})^{\mathrm{T}}=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}\boldsymbol{w}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{D}^{ij}\boldsymbol{e}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}\boldsymbol{e}\tag{7}$$

所以最优化问题(3)转换如下最优化问题:

$$\begin{aligned}&\min\ \boldsymbol{w}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{w}\\ &\text{s.t. }\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}\boldsymbol{e}=0;\\ &\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}\boldsymbol{w}=\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}\boldsymbol{e},\ 1<i,j<m\end{aligned}\tag{8}$$

优化问题(8)为二次约束的二次优化问题(quadratically constrained quadratic programming,简写为 QCQP)[20]。由于矩阵 $\boldsymbol{\Pi}^{ij}$ 是半正定矩阵,所以不难证明(8)的所有约束条件都是凸的。文献[21]证明凸约束的 QCQP 问题可以放松为 SDP 问题,SDP 问题的解近似于 QCQP 问题的最优解,产生的误差较小。设 $\boldsymbol{\Omega}_{ij}=w_iw_j$,则优化问题(8)转化为如下最优化问题:

$$\begin{aligned}&\min_{w,\Omega}\begin{bmatrix}0 & 0\\ 0 & \boldsymbol{L}\end{bmatrix}\cdot\begin{bmatrix}1 & \boldsymbol{w}^{\mathrm{T}}\\ \boldsymbol{w} & \boldsymbol{\Omega}\end{bmatrix}\\ &\text{s.t. }\begin{bmatrix}-\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}\boldsymbol{e} & 0\\ 0 & \boldsymbol{\Pi}^{ij}\end{bmatrix}\cdot\begin{bmatrix}1 & \boldsymbol{w}^{\mathrm{T}}\\ \boldsymbol{w} & \boldsymbol{\Omega}\end{bmatrix}=0;\\ &\begin{bmatrix}0 & \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}/2\\ \boldsymbol{\Pi}^{ij}\boldsymbol{e}/2 & 0\end{bmatrix}\cdot\begin{bmatrix}1 & \boldsymbol{w}^{\mathrm{T}}\\ \boldsymbol{w} & \boldsymbol{\Omega}\end{bmatrix}=0,1\leq i,j\leq m;\\ &\begin{bmatrix}1 & \boldsymbol{w}^{\mathrm{T}}\\ \boldsymbol{w} & \boldsymbol{\Omega}\end{bmatrix}\geq 0\end{aligned}\tag{9}$$

进一步忽略 $\boldsymbol{\Omega}_{ij}=w_iw_j$,则优化问题(9)转化为如下最优化问题:

$$\begin{aligned}&\min_{w,\Omega}\begin{bmatrix}0 & 0\\ 0 & \boldsymbol{L}\end{bmatrix}\cdot\boldsymbol{W}\\ &\text{s.t. }\begin{bmatrix}-\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}\boldsymbol{e} & 0\\ 0 & \boldsymbol{\Pi}^{ij}\end{bmatrix}\cdot\boldsymbol{W}=0;\\ &\begin{bmatrix}0 & \boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Pi}^{ij}/2\\ \boldsymbol{\Pi}^{ij}\boldsymbol{e}/2 & 0\end{bmatrix}\cdot\boldsymbol{W}=0,\ 1\leq i,j\leq m;\\ &\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\cdot\boldsymbol{W}=1;\\ &\begin{bmatrix}0 & \boldsymbol{e}\\ \boldsymbol{e} & 0\end{bmatrix}\cdot\boldsymbol{W}=\theta_1;\\ &\begin{bmatrix}0 & 0\\ 0 & \boldsymbol{E}\end{bmatrix}\cdot\boldsymbol{W}=\theta_2;\\ &\boldsymbol{W}\geq 0\end{aligned}\tag{10}$$

其中,$\boldsymbol{E}$ 表示所有元素为 1 的矩阵块,约束条件 $\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}\cdot\boldsymbol{W}=1$ 使得 $W_{11}=1$,限制条件 $\begin{bmatrix}0 & \boldsymbol{e}\\ \boldsymbol{e} & 0\end{bmatrix}\cdot\boldsymbol{W}=\theta_1$ 和 $\begin{bmatrix}0 & 0\\ 0 & \boldsymbol{E}\end{bmatrix}\cdot\boldsymbol{W}=\theta_2$ 通过参数 θ_1 和 θ_2 控制优化问题的边界。至此,将二次最优化问题(3)转化为半正定问题的标准形式。最优解 $\boldsymbol{W}$ 的第 1 列(除了 W_{11})表示嵌入向量 $\boldsymbol{w}$。内点法 SDPA[21]是一种高效求解 SDP 问题的算法,利用 SDPA 求解优化问题(10)。综上,GPHCC 算法的具体计算过程描述如下:

输入:关系矩阵 $\boldsymbol{R}^{(ij)}$,权值 $\beta_{ij}(1\leq i,j\leq m)$,参数 θ_1 和 θ_2,聚簇个数为 K;

输出:$X^1,X^2,\cdots,X^m$ 的聚类结果。

1)根据关系矩阵 $\boldsymbol{R}^{(ij)}$ 计算矩阵 $\boldsymbol{L}$ 和 $\boldsymbol{\Pi}^{ij}(1\leq i,j\leq m)$;

2)利用迭代算法比如 SDPA 解最优化问题

(11),得到矩阵 $\boldsymbol{W}$;

3) 从矩阵 $\boldsymbol{W}$ 中抽取出嵌入向量 $\boldsymbol{w}$;

4) 将 $\boldsymbol{w}$ 的每一列看成是 $\mathbb{R}^1$ 空间内的一个点,利用 k-means 算法将其聚为 k 类;

时间复杂度分析:该算法的运行时间主要消耗在半正定问题(10)的求解,求解半正定问题的 SDPA 算法的时间复杂度为 $O(n\mathrm{lb}(1/\varepsilon))$[21],其中 n 为各种类型数据个数的总和,ε 为算法精度。

2 实验结果与分析

2.1 评价方法

利用 *F-score* 评价聚类效果,*F-score* 由准确率和召回率计算而来。*F-score* 的值越大,说明聚类效果越好。*F-score* 的最优值为1,最差值为0,计算方法如下:

$$F\text{-}Score = \frac{2 \times Precision \times Recall}{Precision + Recall}。$$

另外,由于本实验涉及的算法具有一定的随机性,将30次聚类结果的 *F-score* 的平均值作为对比实验结果。

2.2 数据集介绍

本实验采用一组来自布告栏系统(BBS)[22]的数据集。在 BBS 系统中,用户使用他们注册的 ID 读取其他用户发布的留言,并且可以通过自己的留言发表自己的意见。整个系统由许多讨论社区(field)组成,每个社区包含很多类似主题的板块(board),每个块板包含多个主题,每个主题通过文章连接多个 IDs。利用 BBS[22] 系统中的数据构建数据集1和2,选取的社区和板块如表1所示,并从每个块板中随机选取80个主题。

表1 选自 BBS 的数据集1和2

Tab. 1 Dataset 1 and 2 sampled from a BBS

Dataset 1		Dataset 2	
Field name	Board name	Field name	Board name
Comp. Sci	C + + Builder	Comp. Sci	Virus
Comp. Sci	Delphi	Comp. Sci	Unix
Comp. Sci	Database	Entertainment	Music
Sports	Basketball	Entertainment	Dance
Sports	Volleyball	Society	Law
Sports	Badminton	Society	Commerce

2.3 结果与分析

在实验中存在3种类型数据:主题(X^1)、用户ID(X^2)和板块(X^3)。利用每个用户发布在每个主题的留言的数目计算主题-用户的关系矩阵($\boldsymbol{R}_{12}$)元素值;如果一个主题属于某个板块,那么主题-板块矩阵($\boldsymbol{R}_{13}$)中相应的元素值为1;如果用户在该板块已发布过任何文章,那么用户-板块关系矩阵 $\boldsymbol{R}_{23}$ 相应的元素设置为1。由以上3种数据关系 $\boldsymbol{R}_{12}$、$\boldsymbol{R}_{13}$ 和 $\boldsymbol{R}_{23}$ 构建网状高阶异构数据。

2.3.1 GPHCC 算法与2阶联合聚类算法聚类效果比较

将 GPHCC 算法与目前已有的5种经典的2阶联合算法聚类结果准确性比较,2阶联合算法包括:模糊算法(简称 CoDoK)[5]、谱图划分算法(简称 BSGB)[6]、等周图划分算法(简称 ICA)[7]、信息论算法(简称 ITC)[8]和矩阵分解方法(简称 NBVD)[9]。2阶联合聚类仅仅涉及2种类型数据之间的关系,本实验利用主题-用户之间的关系信息 $\boldsymbol{R}_{12}$。

图2为 GPHCC 与 CoDoK、BSGB、ITC、ICA、NBVD 算法在数据集 Dataset 1 和 Dataset 2 上聚类结果的 *F-score* 值比较。

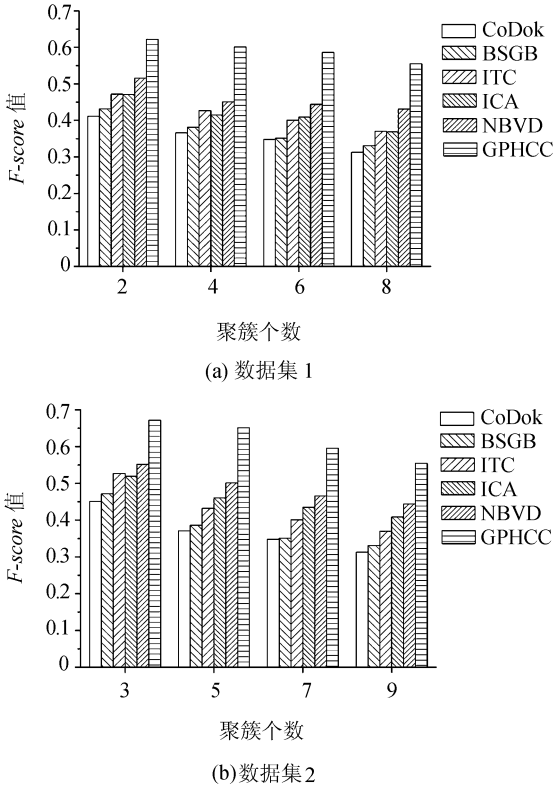


图2 在2个数据集上 CoDoK、BSGB、ITC、ICA、NBVD 与 GPHCC 算法聚类结果 *F-score* 值

Fig. 2 *F-score* value of clustering results of CoDoK, BSGB, ITC, ICA, NBVD and GPHCC algorithms for two datasets

结果表明,GPHCC 算法聚类结果明显优于目前已有的5种2阶联合聚类算法 CoDoK、BSGB、ICA、ITC 和 NBVD。原因在于,GPHCC 算法能够同时利

用主题 - 用户($\boldsymbol{R}_{12}$)、主题 - 板块($\boldsymbol{R}_{13}$) 和用户 - 板块($\boldsymbol{R}_{23}$)3 种类型关系信息, 而 CoDoK、BSGB、ICA、ITC 和 NBVD 算法仅仅利用主题 - 用户($\boldsymbol{R}_{12}$) 之间的关系信息。

2.3.2　GPHCC 算法与星型高阶联合聚类算法聚类效果比较

将 GPHCC 算法与目前已有的 5 种高阶联合聚类算法进行准确性对比,高阶联合聚类包括: CBGC[10]、CIT[12]、RSN[3]、NMF[4] 和 CoStar[16]。由于这 5 种算法仅仅适用于分析星型结构高阶异构数据,利用主题 - 用户($\boldsymbol{R}_{12}$) 和主题 - 板块($\boldsymbol{R}_{23}$) 之间的关系建立星型结构高阶异构数据, 其中主题是中心类型数据。图 3 所示为 GPHCC 与 CBGC、CIT、RSN、NMF 和 CoStar 算法在数据集 T1、T2、T3 和 T4 上聚类结果的 *F-score* 值比较。

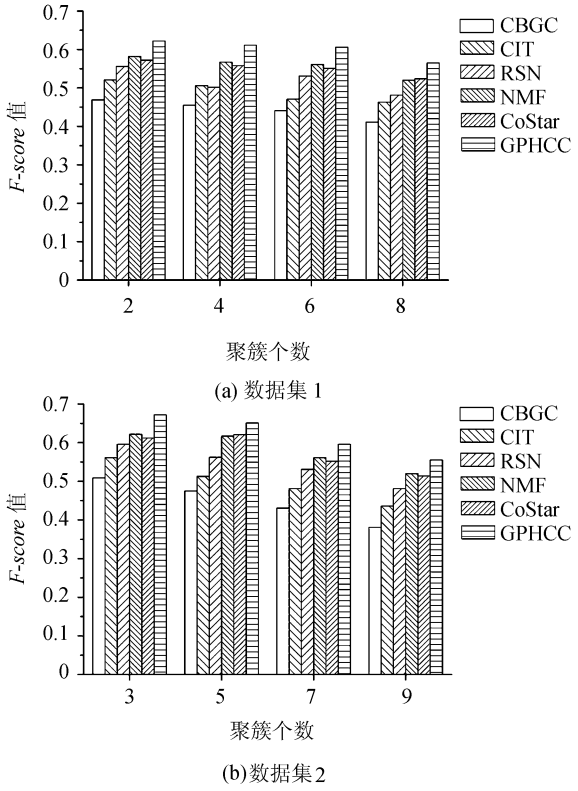


图 3　在 2 个数据集上 CBGC、CIT、RSN、NMF、CoStar 和 GPHCC 算法聚类结果 *F-score* 值

Fig. 3　*F-score* value of Clustering Results of CBGC, CIT, RSN, NMF, CoStar and GPHCC algorithms for two datasets

首先,图 3 结果表明,GPHCC 算法聚类结果明显优于目前已有的 5 种高阶联合聚类算法 CBGC、CIT、RSN、NMF 和 CoStar。原因在于,GPHCC 算法能够同时利用主题 - 用户($\boldsymbol{R}_{12}$)、主题 - 板块($\boldsymbol{R}_{13}$) 和用户 - 板块($\boldsymbol{R}_{23}$) 3 种类型关系信息, 而 CBGC、CIT、RSN、NMF 和 CoStar 算法仅仅利用主题 - 用户($\boldsymbol{R}_{12}$) 和主题 - 板块($\boldsymbol{R}_{13}$) 2 种类型关系信息。其次,对比图 2 与图 3 聚类结果的 *F-score* 值,可以发现高阶联合聚类普遍优于 2 阶联合聚类算法聚类结果。原因在于高阶联合聚类算法利用主题 - 用户($\boldsymbol{R}_{12}$) 和主题 - 板块($\boldsymbol{R}_{13}$) 2 种关系属性信息,而 2 阶联合聚类仅仅利用主题 - 用户之间单一关系信息。由此可见,有效利用不同类型数据之间丰富的关系信息能够进一步提高聚类效果。

3　结　论

提出了一种基于图划分的高阶联合聚类算法 GPHCC,该算法能够有效利用网状高阶异构数据之间丰富的关系信息挖掘网状高阶异构数据内部隐藏聚簇结构,与已有的 2 阶联合聚类算法和高阶联合聚类算法相比,GPHCC 获得了更好的聚类效果。下一步的主要工作是如何有效利用同种类型数据之间的关系,进一步提高聚类效果。

参考文献:

[1]周志华,王珏. 机器学习及其应用[M]. 北京:清华大学出版社,2007.

[2]Wang H, Nie F P, Huang H, et al. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation[C]//Proceeding of the 11th IEEE International Conference on Data Mining. Arlington:IEEE Press,2011:174 - 183.

[3]Long B, Wu X Y, Zhang Z F, et al. Unsupervised learning on k-partite graphs[C]//Procceding of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia: ACM Press,2006:317 - 326.

[4]Chen Y H, Wang L J, Dong M. Non-negative matrix factorization for semisupervised heterogeneous data coclusteering [J]. IEEE Transactions on Knowledge and Data Engineering,2010,22(10):1459 - 1474.

[5]Kummamuuru K, Dhawale A, Krishnapuram R. Fuzzy co-clustering of documents and keywords[C]//Proceeding of the 12th IEEE International Conference on Fuzzy Systems. St Louis:IEEE Press,2003:772 - 777.

[6]Dhillon I S. Co-clustering documents and words using bipartite spectral graph partitioning[C]//Proceeding of the 7th ACM SIGKDD International Conference on Knowledge Dis-

covery and Data Mining. San Francisco, CA: ACM Press, 2001:269 - 274.

[7] Rege M, Dong M, Fotouhi F. Co-clustering documents and words using bipartite isoperimetric graph partitioning[C]//Proceeding of the 6th International Conference on Data Mining. Hong Kong: IEEE Press, 2006:532 - 541.

[8] Dhillon, Mallela S, Modha S D. Information-theoretic co-clustering[C]//Proceeding of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington CA: ACM Press, 2003:89 - 98.

[9] Long B, Zhang Z M, YU P S. Co-clustering by block value decomposition[C]//Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Chicago: ACM Press, 2005:635 - 640.

[10] Gao B, Liu T Y, Zheng X, et al. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering[C]//Proceeding of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago: ACM Press, 2005:41 - 50.

[11] Rege M, Dong M, Hua J. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering[C]//Proceeding of the 17th International Conference on World Wide Web. Beijing: ACM Press, 2008:317 - 326.

[12] Liu T Y, Ma W Y. Star-structured high-order heterogenous data co-clustering based on consistent information theory [C]//Proceeding of the 6th IEEE International Conference on Data Mining. Hong Kong: IEEE Press, 2006:880 - 884.

[13] Greco G, Guzzo A. Coclustering multiple heterogeneous domains: Linear combinations and agreements [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22 (12):1649 - 1663.

[14] Shao J, Yin W T, Ma S, et al. Topic discovery of web video using star-structured k-partite graph[C]//Proceeding of the International Conference on Multimedia. New York: ACM, 2010:915 - 918.

[15] Wang H, Nie F P, Ding C. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization[C]//Proceeding of the 20th ACM International Conference on Information and Knowledge Management. Glasgow: ACM Press, 2011:279 - 284.

[16] Dino I, Robardet C, Ruggero G, et al. Parameter-less co-clustering for star-structured heterogeneous data[J]. Data Mining and Knowledge Discovery, 2012, 26(2):217 - 254.

[17] Sun Y Z, Yu Y T, Han J W. Ranking-based clustering of heterogeneous information networks with star network schema [C]//Proceeding of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris: ACM Press, 2009:797 - 805.

[18] Sun Y Z, Han J W, et al. RankClus: integrating clustering with ranking for heterogeneous information network analysis [C]//Proceeding of the 12th International Conference on Extending Database Technology: Advances in Database Technology. Saint Petersburg: ACM Press, 2009:565 - 576.

[19] Shi J B, Malik J. Normalized cuts and image segmentation [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence, 2000, 22(8):888 - 905.

[20] Boyd S, Vandenberghe L. Convex optimization[M]. Cambridge: Cambridge University Press, 2004.

[21] Fujisawa K, Fukuda M, Kojima M, et al. Numerical evaluation of the SDPA (SemiDefinite Programming Algorithm) [R]. Dordrecht: Kluwer Academic Press, 2000:267 - 301.

[22] Kou Z, Zhang C. Reply network on a bulletin board system. Physical Review E, 2003, 67(3):0361171 - 0361176.

(编辑 杨 蓓)