doi:10.3969/j.issn.1000-436x.2014.06.003

高阶异构数据模糊联合聚类算法

黄少滨，杨欣欣，申林山，李艳梅
（哈尔滨工程大学 计算机科学与技术学院，黑龙江 哈尔滨 150001）

摘 要：为了更有效地分析聚簇重叠部分高阶异构数据的聚簇结果，提出了一种高阶异构数据模糊联合聚类（HFCC）算法，该算法最小化每个特征空间中对象与聚簇中心的加权距离。推导出对象隶属度和特征权重的迭代更新公式，设计出聚类过程的迭代算法，并且从理论上证明了该迭代算法的收敛性。另外，通过泛化 *XB* 指标，提出适用于评估高阶异构数据聚类质量的指标 *GXB*，用于判断聚簇数目。实验表明，HFCC 算法能够有效探测数据内部隐藏的重叠聚簇结构，并且 HFCC 算法聚类效果明显优于 5 种有代表性的硬划分算法，此外 *GXB* 指标能够有效判定高阶异构数据的聚簇数目。

关键词：高阶异构数据；联合聚类；模糊聚类
中图分类号：TP391 **文献标识码**：A **文章编号**：1000-436X(2014)06-0015-10

Fuzzy co-clustering algorithm for high-order heterogeneous data

HUANG Shao-bin, YANG Xin-xin, SHEN Lin-shan, LI Yan-mei
(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

Abstract: In order to analyze the clustering results of high-order heterogeneous data at the overlaps of different clusters more efficiently, a fuzzy co-clustering algorithm was developed for high-order heterogeneous data (HFCC). HFCC algorithm minimized distances between objects and centers of clusters in each feature space. The update rules for fuzzy memberships of objects and weights of features were derived, and then an iterative algorithm was designed for the clustering process. Additionally, convergence of iterative algorithm was proved. In order to estimate the number of clusters, *GXB* validity index was proposed by generalizing the *XB* validity index, which could measure the quality of high-order clustering results. Finally, experimental results show that HFCC can efficiently mine the overlapped clusters and the qualities of clustering results of HFCC are superior five classical hard high-order co-clustering algorithms. Additionally, *GXB* validity index can efficiently estimate the number of high-order clusters.

Key words: high-order heterogeneous data; co-clustering; fuzzy clustering

1 引言

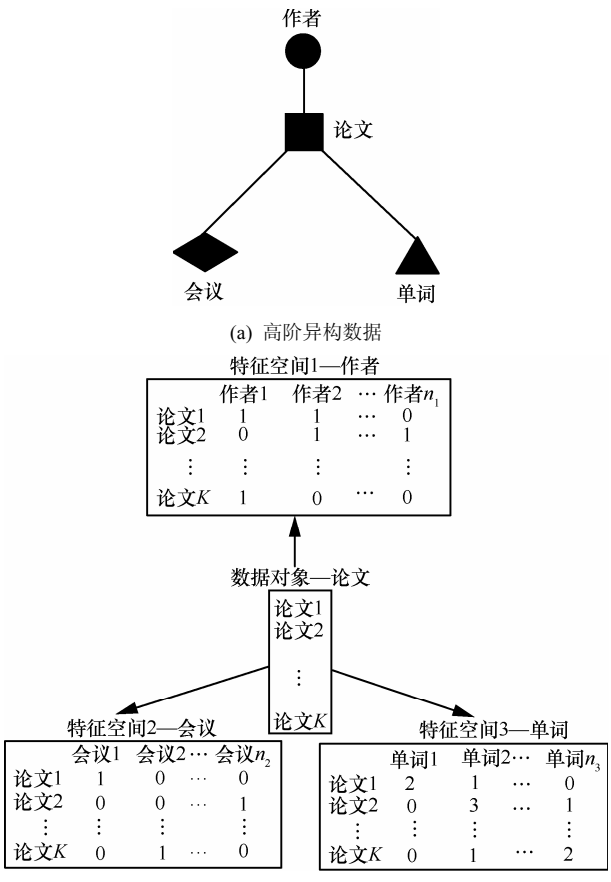
传统的聚类算法集中于分析同构数据(homogeneous data)，同构数据利用单一的特征空间进行描述[1,2]。然而近年来，随着现代信息技术特别是因特网技术的发展，产生了大量需要利用多种特征空间进行描述的数据。例如在学术出版系统中，对论文的描述涉及 3 种特征空间[2]，分别是作者、会议和单词，如图 1 所示。将这种利用多种特征空间描述的数据称为高阶异构数据[3~5]。

与同构数据不同，高阶异构数据包含了多种特征空间信息，从多种视角描述数据[2]。如何有效利用多种特征空间信息，以提高聚类效果成为了新的问题[2,6]。为此发展了一类针对高阶异构数据的联合

收稿日期：2013-01-11；**修回日期**：2013-08-30
基金项目：国家自然科学基金资助项目(71272216, 60903080, 60093009)；国家科技支撑计划基金资助项目(2009BAH42B02, 2012BAH08B02)；博士后科学基金资助项目(2012M510480)；中央高校基本科研业务费专项基金资助项目(HEUCFZ1212, HEUCFT1208)
Foundation Items: The National Natural Science Foundation of China (71272216, 60903080, 60093009); The National Key Technology R&D Program (2009BAH42B02, 2012BAH08B02); The Science Foundation for Post Doctorate Research (2012M510480); The Fundamental Research Funds for the Central University (HEUCFZ1212, HEUCFT1208)

聚类方法，同时对数据对象以及多种特征空间中的特征进行聚类，将此聚类方法称为高阶异构数据联合聚类[3~5]，也称为高阶联合聚类。

与同构数据聚类算法相比，高阶联合聚类具有以下优点[1,2]。首先，大量数据具有高维性，例如，图 1 数据集往往包括上千作者和单词。高维数据包含大量无关的特征，影响聚类效果。同构算法采用所有特征，容易陷入维数灾难。而高阶联合聚类算法在聚类过程中形成特征簇，自动实现降维，一定程度避免了维数灾难，有效地减弱了高维性对聚类效果的影响。另外，数据对象与特征之间的聚类结果相互影响，存在交互性。例如，发表同一主题论文的作者往往研究同一主题，所以论文的聚类结果影响作者聚类结果；同时，研究同一主题的作者往往发表同一主题的文章，所以作者的聚类结果影响论文聚类结果。同构数据聚类算法采用静态的特征向量描述数据，失去对象与特征之间聚类结果的交互性。



(a) 高阶异构数据

(b) 特征空间

图 1 高阶异构数据及其特征空间示例

由于以上原因，近年来高阶联合聚类逐渐成为研究热点，并取得了很多重要的研究成果[1~16]。Long 等[1,7]提出基于 k 部图学习的方法，将对象和特征建模为 k 部图节点，将特征值建模为边的权重。将对象簇和特征簇以及它们之间的关联关系建模为关系概要网络。聚类问题转化为寻找最相似于原始 k 部图的关系概要网络的优化问题。Dino 等[2]提出 Goodman Kruskal 方法，利用对象和特征值之间的预测能力估计对象簇和特征簇的质量，采用启发式的方法寻找最强预测能力的对象簇和特征簇。高斌和 Rege 等[3,4,8~10]提出基于图划分的方法，图节点表示对象和不同特征空间的特征，边的权值表示特征值，将聚类问题转化为最优图划分问题。高斌和 Greco 等[5,11~13]提出基于信息论的方法，将对象簇以及特征簇看作随机变量，用互信息衡量它们之间的相关性，求解具有最高相关性的聚类结果，即具有最大互信息值的聚类结果。韩家炜[14,15]提出基于排名的聚类方法，对象和特征在不同的聚簇中具有不同的排名结果，准确的排名结果有助于聚类，同时，准确的聚类结果有助于排名，排名和聚类互相指导，共同实现。Hua 和 Lijun 等[6,16]提出基于矩阵分解的方法，用低维空间中聚簇指示矩阵与基矩阵重构对象在每个特征空间中的特征矩阵，聚类问题转化为求解最接近特征矩阵的矩阵分解问题。

以上聚类算法都是硬划分方法，一个对象属于或者不属于一个聚簇。然而在实际问题中，有些对象同时属于多个聚簇，聚簇之间存在重叠的边界。比如，在科学论文领域存在交叉学科的论文及作者，这些论文和作者应属于多个聚簇。硬划分方法并没有考虑有些数据可能同时属于多个聚簇的聚类问题，所以如何更有效地分析聚簇重叠部分数据的聚簇结果，成为高阶联合聚类面对的新问题。

与硬化分聚类方法相比，模糊聚类方法作为一种软聚类方法，利用隶属度描述对象属于聚簇的程度，能够更好地描述聚簇之间的重叠边界。为此提出一种高阶异构数据模糊联合聚类 (HFCC, fuzzy co-clustering for high-order heterogeneous data) 算法。HFCC 算法利用对象隶属度描述对象与聚簇的隶属关系，用权值描述每个特征空间中特征与聚簇的关系，将聚类问题转化为最小化每个特征空间中对象与聚簇中心的加权距离的优化问题。推导出求解优化问题的隶属度和特征权重迭代更新公式，设计聚类过程的迭代算法。并且还从理论和实验两方面验证了此迭代算法的收敛性。另外，已有的高阶联合聚类算法需要预先指定聚簇数目，如何自动判

断聚簇数目是另一需要解决的问题。*XB* 指标是模糊聚类质量的指标[17]，用于判断同构数据的聚簇数目。本文泛化了 *XB* 指标，使其适用于评估模糊高阶聚类质量，判断高阶异构数据聚簇数目。

2　高阶异构数据模糊联合聚类算法

本节首先给出了 HFCC 算法目标函数的定义式；然后推导出隶属度迭代的更新表达式，给出了 HFCC 算法的计算流程；从理论上证明了 HFCC 算法的收敛性；最后提出高阶模糊聚类质量评价指标。

2.1　HFCC 算法目标函数

由对象 $X=\{x_1,\cdots,x_K\}$ 和 N 种特征空间 $Y^1=\{y_1^{(1)},\cdots,y_{n_1}^{(1)}\},\cdots,Y^N=\{y_1^{(N)},\cdots,y_{n_N}^{(N)}\}$ 组成的高阶异构数据。图 2 是由对象 X、特征空间 Y^1 和 Y^2 组成的高阶异构数据，其中，图节点表示对象和特征，边的权重表示相应的特征值。

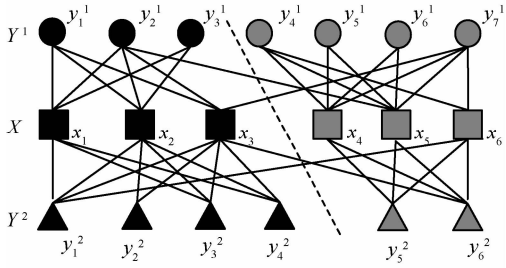


图 2　2 种特征空间组成的高阶异构数据示意

HFCC 算法利用隶属度 $u_{c\mu}\in[0,1]$ 描述对象 x_μ 属于第 c 个聚簇的程度，u_{ij} 值越接近 1，表示 x_μ 属于第 c 个聚簇的程度越高。特征权重 $v_{cp}^{(i)}\in[0,1]$ 表示特征 $y_p^{(i)}$ 在第 c 个聚簇的权重，$v_{cp}^{(i)}$ 值越接近 1，表示 $y_p^{(i)}$ 属于第 c 个聚簇的程度越高。首先，最小化每个特征空间中对象 x_μ 与原型 $o_c^{(i)}$ 的加权距离 $\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2$，其中，$q_i\in(0,+\infty]$ 表示模糊指数，用于调节特征权重 $v_{cp}^{(i)}$ 的模糊度。其次，若对象 x_μ 与原型 $o_c^{(i)}$ 的特征加权距离较小，则对象 x_μ 应赋予更大的隶属度值 u_{ij}。最小化隶属度 u_{ij} 与特征加权距离 $\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2$ 的乘积 $(u_{c\mu})^m\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2$，其中，$m\in(0,+\infty]$ 为模糊指数，m 值越大，u_{ij} 的模糊度越高，即越接近于 $1/C$，越难以判别对象 x_μ 属于哪一个聚簇。高阶异构数据联合聚类问题转化为同时最小化每个特征空间中 u_{ij} 与加权距离乘积的优化问题，即最小化加权线性组合 $\sum_{i=1}^{N}\beta_i(\sum_{c=1}^{C}\sum_{\mu=1}^{K}(u_{c\mu})^m\sum_{p=1}^{n_i}(v_{cp}^{(i)})^q(d_{c\mu p}^{(i)})^2)$，其中，$\beta_i$ 为特征空间 $Y^{(i)}$ 的权值，满足 $\sum_{i=1}^{N}\beta_i=1$。基于以上思想，HFCC 算法的目标函数定义如下

$$J=\sum_{i=1}^{N}\beta_i(\sum_{c=1}^{C}\sum_{\mu=1}^{K}(u_{c\mu})^m\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2) \tag{1}$$

满足如下限制条件

$$\sum_{p=1}^{n_i}v_{cp}^{(i)}=1 \tag{2}$$

$$\sum_{c=1}^{C}u_{c\mu}=1 \tag{3}$$

$d_{c\mu p}^{(i)}$ 表示在特征空间 $Y^{(i)}$ 中第 p 维特征上对象 x_μ 与第 c 个聚簇中心的距离，计算形式如下

$$d_{c\mu p}^{(i)}=|x_{\mu p}^{(i)}-o_{cp}^{(i)}| \tag{4}$$

其中，$x_{\mu p}^{(i)}$ 表示在特征空间 $Y^{(i)}$ 中 x_μ 的第 p 维特征值，$o_{cp}^{(i)}$ 表示在特征空间 $Y^{(i)}$ 中第 c 个聚簇中心向量 $o_c^{(i)}$ 的第 p 维值。

2.2　迭代更新规则

通过拉格朗日乘子法求解 $u_{c\mu}$ 和 $v_{cp}^{(i)}$ 的更新规则，用于设计求解目标函数的迭代算法。首先构造如下拉格朗日函数

$$L=\sum_{i=1}^{N}\beta_i(\sum_{c=1}^{C}\sum_{\mu=1}^{K}(u_{c\mu})^m\sum_{p=1}^{n_i}(v_{cp}^{(i)})^q(d_{c\mu p}^{(i)})^2)-\sum_{\mu=1}^{K}\phi_\mu(\sum_{c=1}^{C}u_{c\mu}-1)-\sum_{i=1}^{N}\sum_{c=1}^{C}\lambda_c^{(i)}(\sum_{p=1}^{n_i}v_{cp}^{(i)}-1) \tag{5}$$

其中，$\lambda_c^{(i)}$ 和 ϕ_μ 分别是相应于约束条件式(2)和式(3)的拉格朗日因子。由拉格朗日最优解的必要条件，求 L 关于 $\lambda_c^{(i)}$ 的偏导数，并令其为零

$$\frac{\partial L}{\partial\lambda_c^{(i)}}=\sum_{p=1}^{n_i}v_{cp}^{(i)}-1=0 \tag{6}$$

同理求 L 关于 $v_{cp}^{(i)}$ 的偏导数，并令其为零

$$\frac{\partial L}{\partial v_{cp}^{(i)}}=\beta_i(q_i(v_{cp}^{(i)})^{q_i-1}\sum_{\mu=1}^{K}(u_{c\mu})^m(d_{c\mu p}^{(i)})^2-\lambda_c^{(i)}=0 \tag{7}$$

所以

$$v_{cp}^{(i)}=\left[\frac{\lambda_c^{(i)}}{\beta_i(q_i\sum_{\mu=1}^{K}(u_{c\mu})^m(d_{c\mu p}^{(i)})^2)}\right]^{1/(q_i-1)} \tag{8}$$

将式(8)代入式(6)得

$$\sum_{p=1}^{n_i}v_{cp}^{(i)}=\left[\frac{\lambda_c^{(i)}}{q_i}\right]^{1/(q_i-1)}\sum_{p=1}^{n_i}\left[\frac{1}{\beta_i(\sum_{\mu=1}^{K}(u_{c\mu})^m(d_{c\mu p}^{(i)})^2)}\right]^{1/(q_i-1)}=1 \tag{9}$$

因而

$$[\lambda_c^{(i)}/q_i]^{1/(q_i-1)}=\frac{1}{\sum_{p=1}^{n_i}[1/(\beta_i(\sum_{\mu=1}^{K}(u_{c\mu})^m(d_{c\mu p}^{(i)})^2))]^{1/(q_i-1)}} \tag{10}$$

将式(10)代入式(8)得

$$v_{cp}^{(i)}=\frac{[1/(\beta_i(\sum_{\mu=1}^{K}(u_{c\mu})^m(d_{c\mu p}^{(i)})^2))]^{1/(q_i-1)}}{\sum_{t=1}^{n_i}[1/(\beta_i(\sum_{\mu=1}^{K}(u_{c\mu})^m(d_{c\mu t}^{(i)})^2))]^{1/(q_i-1)}} \tag{11}$$

式(11)表明 $\lim_{q_i\to+\infty}v_{cp}^{(i)}=\frac{1}{n}$，可见模糊指数 q_i 调节 $v_{cp}^{(i)}$ 的模糊度。q_i 越大，$v_{cp}^{(i)}$ 的模糊度越高，$v_{cp}^{(i)}$ 越接近于 $\frac{1}{n}$，越难以判别 $y_{cp}^{(i)}$ 属于哪一个聚簇。求 L 关于 ϕ_μ 的偏导数，并令其为零

$$\frac{\partial L}{\partial\phi_\mu}=\sum_{c=1}^{C}u_{c\mu}-1=0 \tag{12}$$

求 L 关于 $u_{c\mu}$ 的偏导数，并令其为零

$$\frac{\partial L}{\partial u_{c\mu}}=\sum_{i=1}^{N}\beta_i(m(u_{c\mu})^{m-1}\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)-\phi_\mu=0 \tag{13}$$

由式(13)得

$$u_{c\mu}=\left[\frac{\phi_\mu}{\sum_{i=1}^{N}\beta_i(m\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)}\right]^{1/(m-1)} \tag{14}$$

代入式(12)得

$$\sum_{c=1}^{C}u_{c\mu}=(\phi_\mu)^{1/(m-1)}\sum_{c=1}^{C}\left[\frac{1}{\sum_{i=1}^{N}\beta_i(m\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)}\right]^{1/(m-1)}=1 \tag{15}$$

由式(15)得

$$(\phi_\mu)^{1/(m-1)}=\frac{1}{\sum_{c=1}^{C}\left[1/\sum_{i=1}^{N}\beta_i(m\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)\right]^{1/(m-1)}} \tag{16}$$

代入式(14)得

$$u_{c\mu}=\frac{1}{\sum_{t=1}^{C}\left[\frac{\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)}{\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{t\mu p}^{(i)})^2)}\right]^{1/(m-1)}} \tag{17}$$

求 L 关于 $o_{cp}^{(i)}$ 的偏导数，并令其为零

$$\frac{\partial L}{\partial o_{cp}^{(i)}}=-2\sum_{\mu=1}^{K}(u_{c\mu})^m(v_{cp}^{(i)})^{q_i}(x_{\mu p}^{(i)}-o_{cp}^{(i)})=0 \tag{18}$$

由式(18)得

$$o_{cp}^{(i)}=\frac{(v_{cp}^{(i)})^{q_i}\sum_{\mu=1}^{K}(u_{c\mu})^m(x_{\mu p}^{(i)})}{(v_{cp}^{(i)})^{q_i}\sum_{\mu=1}^{K}(u_{c\mu})^m} \tag{19}$$

情况 1 $v_{cp}^{(i)}=0$。这种情况下，特征 $y_p^{(i)}$ 与第 c 个聚簇无关，$o_{cp}^{(i)}$ 的值不会影响特征加权距离。这种情况下，$o_{cp}^{(i)}$ 可以为任何值，所以将 $o_{cp}^{(i)}$ 简单设为 0

$$o_{cp}^{(i)}=0 \tag{20}$$

情况 2 $v_{cp}^{(i)}\neq 0$。将式(19)化简为

$$o_{cp}^{(i)}=\frac{\sum_{\mu=1}^{K}(u_{c\mu})^m(x_{\mu p}^{(i)})}{\sum_{\mu=1}^{K}(u_{c\mu})^m} \tag{21}$$

综上，聚簇中心向量更新为

$$o_{cp}^{(i)}=\begin{cases}0 & ,f\ v_{cp}^{(i)}\neq 0\\ \dfrac{\sum_{\mu=1}^{K}(u_{c\mu})^m(x_{\mu p}^{(i)})}{\sum_{\mu=1}^{K}(u_{c\mu})^m} & ,f\ v_{cp}^{(i)}\neq 0\end{cases} \tag{22}$$

以下以图 2 数据集为例，说明 HFCC 算法的第一次迭代过程。

1) 随机产生并根据式(3)归一化隶属度 $u_{c\mu}$，将对象 X 划分到隶属度最大的聚簇中，如图 3(a)所示，其中黑色表示对象被划分到第 1 个聚簇中，灰色表示对象被划分到第 2 个聚簇中。

2) 根据式(22)和式(4)分别计算聚簇中心 $o_{cp}^{(i)}$ 和距离 $d_{c\mu p}^{(i)}$，然后根据式(11)计算 Y^1 和 Y^2 的权重值，并根据式(3)对隶属度归一化处理，将 Y^1 和 Y^2 数据划分到隶属度最大的聚簇中，如图 3(b)所示。

3) 根据式(17)计算 X 的隶属度，并根据式(2)对隶属度归一化处理，将对象 X 划分到隶属度最大的聚簇中，如图 3(c)所示。

4) 重复上述过程直到 X 隶属度值达到收敛或达到指定迭代次数。图 3(d)为每次迭代目标函数值，经过 8 次左右迭代，算法达到收敛状态。

综上，高阶异构数据模糊联合聚类算法步骤描述如下。

初始化：确定聚簇数目 C，模糊指数 m 和 q_i，终止条件阈值 ε 和最大迭代次数 $\tau_{\max}$。随机产生并归一化 X 隶属度 $(u_{c\mu})^{\tau}$。

Repeat

For each *i*

1) 利用式(22)更新聚簇中心 $(o_{cp}^{(i)})^{\tau+1}$；

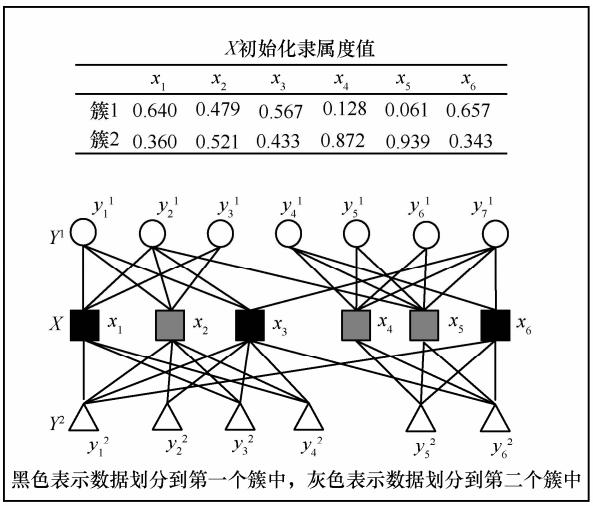
2) 利用式(4)更新距离 $(d_{c\mu p}^{(i)})^{\tau+1}$；

3) 利用式(11)更新 $Y^{(i)}$的权重 $(v_{cp}^{(i)})^{\tau+1}$；

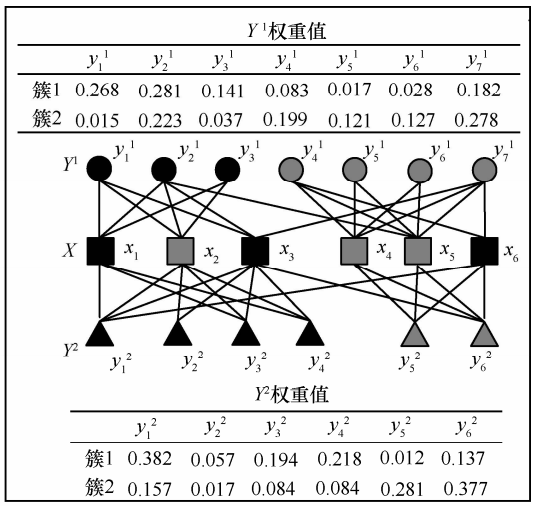
End

4) 利用式(17)更新 X 隶属度 $(u_{c\mu})^{\tau+1}$；

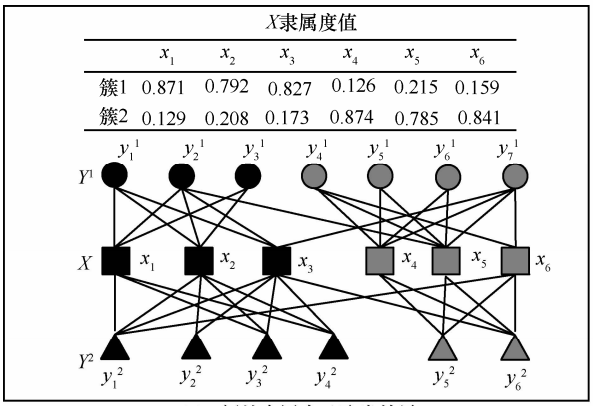
5) $\tau=\tau+1$；



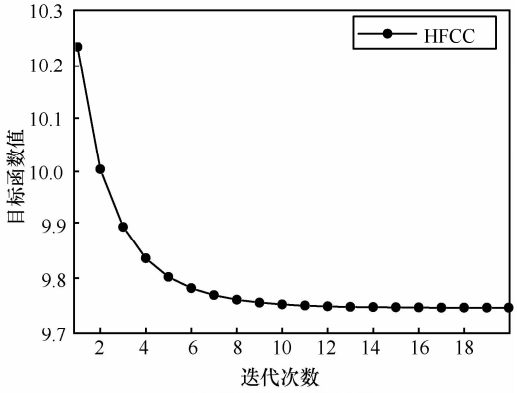
(a) 初始化X隶属度及聚类结果



(b) 第一次迭代Y^1和Y^2权重值及聚类结果



(c) X新的隶属度及聚类结果



(d) 算法迭代收敛过程

图 3　针对图 2 数据集 HFCC 算法计算步骤示例

Until $\max_{c,\mu}|(u_{c\mu})^{(\tau+1)}-(u_{c\mu})^{(\tau+1)}|<\varepsilon$ 或 $\tau=\tau_{\max}$。

时间复杂度分析：对于特征空间 $Y^{(i)}$，步骤 1 每次循环的时间复杂度为 $O(CKn_i)$，共 N 个特征空间，所以步骤 1 每次循环时间复杂度为 $O(CK\sum_{i=1}^{N}n_i)$。同理，步骤 2 和步骤 3 每次循环时间复杂度为 $O(CK\sum_{i=1}^{N}n_i)$。步骤 4 利用式(17)计算 $u_{c\mu}$ 时 $\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)$ 和 $\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{t\mu p}^{(i)})^2)$ 可以独立计算。计算 $\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)$ 的时间复杂度为 $O(\sum_{i=1}^{N}n_i)$，一次循环共计算 CK 项，所以计算 $\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)$ 的时间复杂度为 $O(CK\sum_{i=1}^{N}n_i)$。计算 $\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{t\mu p}^{(i)})^2)$ 的时间复杂度为 $O(\sum_{i=1}^{N}n_i)$，一次循环共计算 CK 项，所以计算 $\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{t\mu p}^{(i)})^2)$ 的时间复杂度为 $O(CK\sum_{i=1}^{N}n_i)$。综上，HFCC 算法的时间复杂度为 $O(CKn\tau)$，其中 $n=\sum_{i=1}^{N}n_i$，τ 为迭代次数。

2.3 收敛性证明

由单调有界原理（即单调有界函数必收敛）知，为证明式(1)目标函数 J 通过式(22)、式(4)、式(11)、式(17)迭代更新收敛，只需证明 J 为迭代次数 τ 的单调递减有界函数。以下通过引理 1~3 证明 J 为迭代次数 τ 的单调递减函数，通过引理 4 证明 J 为有界函数。

引理 1 根据式(17)迭代更新 $u=(u_{11},\cdots,u_{CK})$ 不会增加式(1)中的目标函数值 J。

证明 因为式(17)更新 u 时把上次迭代获得的 $v_{cp}^{(i)}$ 和 $d_{c\mu p}^{(i)}$ 看作常数，所以把目标函数 J 看作 u 的函数 $J(u)$。由拉格朗日乘子法，根据式(17)计算的 u^* 值为 $J(u)$ 的一个驻点。为了证明 u^* 为极大值点，只需证明海森矩阵 $\nabla^2\boldsymbol{J}(\boldsymbol{u}^*)$ 在点 u^* 处为正定矩阵。

$$\nabla^2\boldsymbol{J}(\boldsymbol{u})=\begin{bmatrix}\frac{\partial J(u)^2}{\partial u_{11}u_{11}} & \cdots & \frac{\partial J(u)^2}{\partial u_{11}u_{CK}}\\ \vdots & \ddots & \vdots\\ \frac{\partial J(u)^2}{\partial u_{CK}u_{11}} & \cdots & \frac{\partial J(u)^2}{\partial u_{CK}u_{CK}}\end{bmatrix} \tag{23}$$

对式(1)求取关于 $u_{c\mu}$ 的二阶偏导数

$$\frac{\partial J(u)^2}{\partial u_{c\mu}\partial u_{c\mu}}=m(m-1)(u_{c\mu})^{m-2}\sum_{i=1}^{N}\beta_i(\sum_{p=1}^{n_i}(v_{cp}^{(i)})^{q_i}(d_{c\mu p}^{(i)})^2)$$

$\nabla^2\boldsymbol{J}(\boldsymbol{u}^*)$ 在 u^* 处的值如上式所示，因为 $u_{c\mu}$ 为正值，所以海森矩阵 $\nabla^2\boldsymbol{J}(\boldsymbol{u}^*)$ 为正定矩阵。综上，u^* 是目标函数 $J(u)$的驻点并且海森矩阵 $\nabla^2\boldsymbol{J}(\boldsymbol{u}^*)$ 为正定矩阵。由多元函数极值存在充分必要条件知 u^* 是 $J(u)$ 的局部最小值，所以 $J(u^\tau)\geqslant J(u^*)=J(u^{\tau+1})$，其中 u^τ 为根据式(17)第 τ 次迭代结束后 $\sum_{l=1}^{n_i}\sum_{\mu=1}^{m}u_{k\mu}d_{\mu l}^{(i)}$ 的值，$O(mn_i)$ 为根据式(17)第 $\tau+1$ 次迭代中的 u 值。所以根据式(17)迭代更新 $u=(u_{11},\cdots,u_{CK})$ 不会提高式(1)中的目标函数值 J。

引理 2 根据式(11)迭代更新 $v^{(i)}=(v_{11}^{(i)},\cdots,v_{Cn_i}^{(i)})$ 不会增加式(1)中的目标函数值 J。

证明 与引理 1 相同的方法证明。

引理 3 根据式(22)迭代更新 $o^{(i)}=(o_{11}^{(i)},\cdots,o_{Cn_i}^{(i)})$ 不会增加式(1)中的目标函数值 J。

证明 与引理 1 相同的证明方法证明。

引理 4 式(1)中的目标函数 J 有界，即存在常数 M，使得 $J<M$。

证明 对于 $\forall i\in\{1,\cdots,N\}$，$\forall\mu\in\{1,\cdots,K\}$，$\forall c\in\{1,\cdots,C\}$，$\forall p\in\{1,\cdots n_i\}$ 有 $0<u_{c\mu}<1$，$0<v_{cp}^{(i)}<1$，w_i 和 $d_{c\mu p}^{(i)}$ 为常数，由此可见式(1)目标函数必有界。

定理 1 HFCC 算法收敛于局部最小值。

证明 引理 1~引理 3 表明通过式(22)、式(4)、式(11)、式(17)迭代更新不会增加目标函数 J 的值。引理 4 表明目标函数 J 具有有限的极限。根据单调有界原理，式(22)、式(4)、式(11)、式(17)迭代更新必定使 J 收敛。

综上，已经证明根据迭代更新式(22)、式(4)、

式(11)、式(17)，HFCC 算法的目标函数 J 必定收敛于满足约束条件式(2)、式(3)的局部最小值。

2.4 最佳聚簇数目确定

HFCC 算法需要领域专家根据经验确定数据集的聚簇数目 c，然而数据的聚簇数目往往是未知的，那么如何自对确定聚簇数目成为聚类过程面对的问题。为此，Xie 和 Beni[17,18] 提出 *XB* 指标评价聚簇质量，以此确定最佳聚簇数目。*XB* 指标具体定义如下

$$XB(c)=\frac{\sum_{c=1}^{C}\sum_{\mu=1}^{K}(u_{c\mu})^{m}\parallel x_{\mu}-o_{c}\parallel^{2}}{K\min_{s,t}\parallel o_{s}-o_{t}\parallel^{2}}$$

然而 *XB* 指标仅仅适用于分析同构数据，为了评价高阶联合聚类质量，确定高阶联合聚类最佳聚簇数目，需要对 *XB* 指标进行泛化，提出适用于高阶联合聚类的评价指标 *HXB*，具体定义如下

$$HXB(c)=\frac{\sum_{i=1}^{N}\beta_{i}(\sum_{c=1}^{C}\sum_{\mu=1}^{K}(u_{c\mu})^{m}\sum_{p=1}^{n_{i}}(v_{cp}^{(i)})^{q}(d_{c\mu p}^{(i)})^{2})}{K\min_{s,t}\sum_{i=1}^{N}\beta_{i}\parallel o_{s}^{(i)}-o_{t}^{(i)}\parallel^{2}}$$

HXB 指标分子表示在每个特征空间中对象与聚簇中心的加权距离，分母表示聚簇之间最短距离。分子越小，分母越大，*HXB* 指标值越小，表示聚簇质量越好。在不同的聚簇数目情况下，当 *HXB* 指标值达到最小时，聚簇个数为最佳聚簇数目。

3 实验分析

3.1 数据集介绍

本节使用 3 个标准数据集测试算法的性能，介绍如下。

Cora dataset 是科技文献数据集。从表 1 所示的主题中随机选择 1 000 篇论文，由论文、作者和单词构建高阶异构数据集 *T*1 和 *T*2。

Corel dataset 包含 5 000 张图像，其中每个图像包含图像分割块和文字标注信息[2]。从表 1 所示的每个主题中选取 100 张图像，采用文献[2]的方法，由图像、单词和图像分割组构建高阶异构数据集 *I*1 和 *I*2。

IAPR TC-12 Benchmark dataset 包含 20 000 张照片和相应的文本标注信息。从表 1 所示的每个主题中随机选取 300 张照片。采用文献[9]的方法，由图像、特征和单词构建高阶异构数据集 *P*1。

3.2 结果分析

3.2.1 聚簇重叠结构分析

首先，图 4 所示为 *T*1、*T*2、*I*1 和 *P*1 数据集在 2 种特征空间中对象相似性可视化结果，灰度越深表示相似性越高。结果表明数据集中存在明显的聚簇重叠结构，*I*2 数据集也具有相似的现象，不再列出。由此可见，“硬划分”高阶联合聚类算法“硬性”地将对象划分到某个聚簇中，难以描述重叠部分对象的聚类结果。

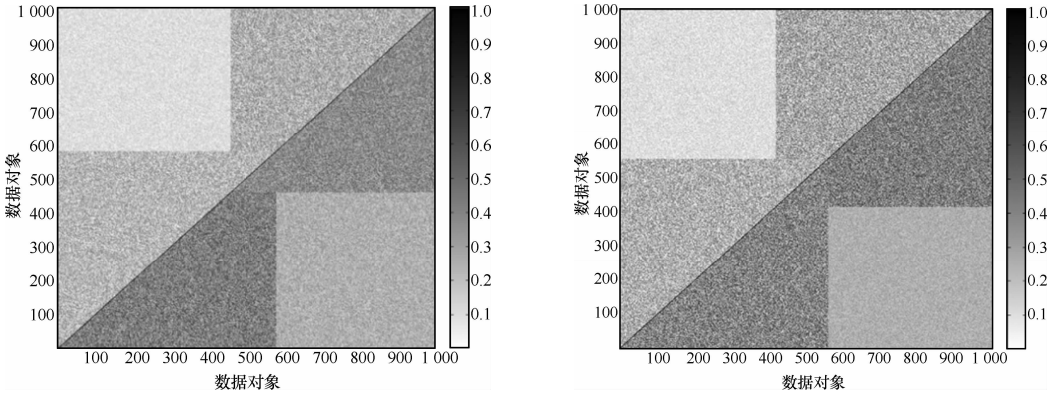
其次，HFCC 算法引入模糊概念，利用隶属度描述对象属于某聚簇的程度。表 2 为 *P*1 数据集中聚簇重叠部分对象的隶属度值示例。这些对象既属于 animal 聚簇又属于 traveler 聚簇，难以严格地将对象划分到 traveler 聚簇或 animal 聚簇中。这些对象的隶属度值在 0.5 左右，聚类结果具有较强的模糊度，难以确定这些对象属于 animal 聚簇或 traveler 聚簇。由此可见，与“硬划分”高阶联合聚类算法相比，HFCC 算法更客观地描述了现实世界对象的聚类结果，能够有效地发现和分析聚簇重叠部分的对象。

3.2.2 HFCC 算法准确率分析

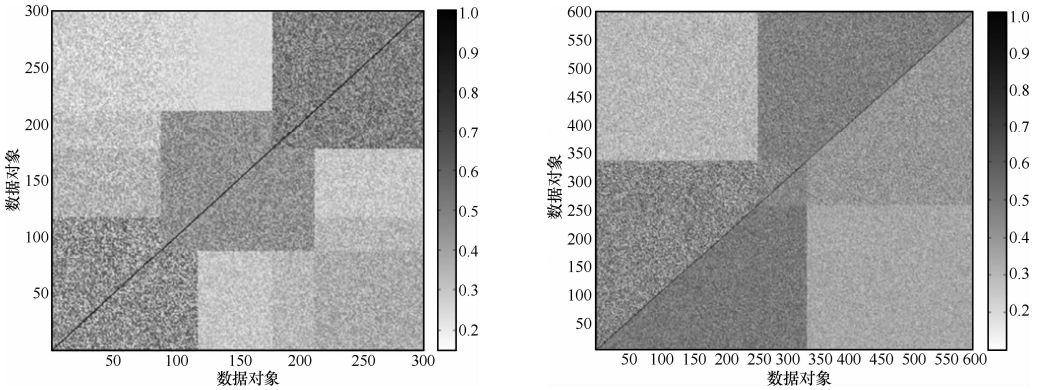
首先，HFCC 算法与目前已有的 5 种“硬划分”高阶联合聚类算法进行准确率比较，包括一致二部图划分算法(CBGC)[4]、一致信息论算法(CIT)[5]、*k* 部图学习算法(RSN)[1]、无监督矩阵分解方法(NMF)[16]和基于 Goodman Kruskal 的算法(CoStar)[2]，聚类结果的 *NMI* 值[19]如图 5 所示。图 5 表明 HFCC 算法聚类效

表 1　　标准数据集描述

简称	数据集	单词数	对象数	聚类个数	主题
*T*1	Cora	2 000	1 000	2	database, artificial intelligence
*T*2	Cora	2 000	1 000	2	operating systems, architecture
*I*1	Corel	116	300	3	cow, grass, horses
*I*2	Corel	114	300	3	tree, bird, sky
*P*1	IAPR TC-12	1000	600	2	traveler, animal



(a) *T*1 数据集　　(b) *T*2 数据集



(c) *I*1 数据集　　(d) *P*1 数据集

图 4　聚簇结构可视化结果

表 2　　*P*1 数据集重叠聚簇部分数据隶属度值示例

簇	隶属度值															
																
簇 1	0.518	0.492	0.481	0.521	0.494	0.513	0.517	0.494	0.518	0.503	0.519	0.521	0.517	0.485	0.489	0.502
簇 2	0.482	0.508	0.519	0.479	0.506	0.487	0.483	0.506	0.482	0.497	0.481	0.479	0.483	0.515	0.511	0.498

果明显超出了这 5 种“硬划分”高阶联合聚类算法，这在一定程度上是由于 HFCC 能更好地描述聚簇重叠部分对象的聚类结果所致。

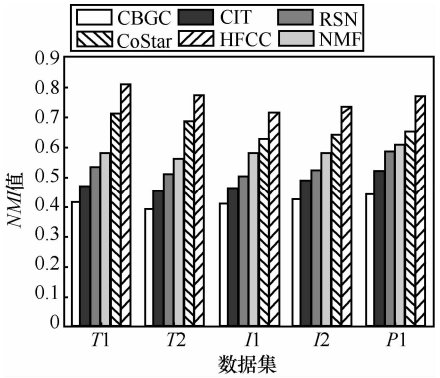


图 5　HFCC 与 CBGC、CIT、RSN、NMF 和 CoStar 算法聚类结果 *NMI* 值比较

其次，图 6 反映了模糊指数 *m* 对聚簇结果的影响。当 *m* 取值在 1 和 3 之间时，聚簇结果没有明显的变化；当 *m* 值大于 3 时，聚簇结果具有明显的下降趋势，所以多数情况下 *m* 的最佳取值在 1 和 3 之间。

3.2.3　*HXB* 指标分析

表 3~表 5 是在不同聚簇数目 *c* 和模糊指数 *m* 时数据集 *T*1、*I*1 和 *P*1 的 *HXB* 指标值。表 3 表明在不同 *m* 值情况下，*HXB* 指标值均能准确地确定 *T*1 数据集的聚簇数目。表 3 表明 *m*=1 时，*HXB* 指标没有正确地确定 *I*1 数据集聚簇数目，但 *m* 取值为 1.5、2、2.5 和 3 时，*HXB* 指标值能够准确地确定 *I*1 数据集的聚簇数目。表 4 表明 *m*=1.5 时，*HXB* 指标没有正确地确定 *P*1 数据集聚簇数目，但 *m* 取

表 3　　***T*1 数据集在不同聚簇数目 *c* 和模糊指数 *m* 时 *HXB* 值**

c	*HXB* 值				
	m=1	*m*=1.5	*m*=2	*m*=2.5	*m*=3
2	2.53×10^{-1}	1.28×10^{-1}	1.26×10^{-1}	1.34×10^{-1}	1.67×10^{-1}
3	1.24×10^{-5}	5.79×10^{-5}	7.98×10^{-4}	1.45×10^{-4}	3.47×10^{-1}
4	3.18×10^{-6}	3.26×10^{-6}	8.02×10^{-8}	1.67×10^{-8}	1.27×10^{-1}
5	5.49×10^{-7}	3.48×10^{-7}	2.63×10^{-9}	2.39×10^{-9}	2.89×10^{-2}
6	7.49×10^{-8}	2.76×10^{-8}	1.08×10^{-10}	2.78×10^{-9}	1.03×10^{-1}

表 4　　***I*1 数据集在不同聚簇数目 *c* 和模糊指数 *m* 时 *HXB* 值**

c	*HXB* 值				
	m=1	*m*=1.5	*m*=2	*m*=2.5	*m*=3
2	8.79×10^{-1}	3.19×10^{-2}	7.89×10^{-2}	2.09×10^{-2}	1.09×10^{-2}
3	2.20×10^{-2}	8.76×10^{-1}	4.13×10^{-2}	9.45×10^{-1}	9.39×10^{-1}
4	1.93×10^{-3}	2.54×10^{-3}	4.69×10^{-3}	1.83×10^{-4}	2.13×10^{-4}
5	3.63×10^{-4}	1.65×10^{-5}	9.29×10^{-5}	6.50×10^{-6}	6.20×10^{-5}
6	5.29×10^{-5}	8.40×10^{-6}	3.49×10^{-5}	7.28×10^{-9}	8.29×10^{-6}

表 5　　***P*1 数据集在不同聚簇数目 *c* 和模糊指数 *m* 时 *HXB* 值**

c	*HXB* 值				
	m=1	*m*=1.5	*m*=2	*m*=2.5	*m*=3
2	3.79×10^{-2}	9.38×10^{-2}	6.52×10^{-1}	3.29×10^{-3}	6.90×10^{-2}
3	9.67×10^{-2}	1.16×10^{-2}	2.40×10^{-2}	9.89×10^{-3}	5.59×10^{-3}
4	6.09×10^{-5}	5.39×10^{-4}	3.80×10^{-4}	5.30×10^{-5}	6.60×10^{-5}
5	4.58×10^{-8}	7.39×10^{-6}	5.09×10^{-3}	3.59×10^{-7}	1.47×10^{-7}
6	9.28×10^{-9}	5.60×10^{-7}	4.39×10^{-5}	8.92×10^{-6}	5.40×10^{-6}

值为 1、2、2.5 和 3 时，*HXB* 指标值能够准确地确定 *P*1 数据集的聚簇数目。由此可见，*HXB* 指标多数情况下能够确定最佳的聚簇数目，*m* 取值在 2 和 3 之间时具有较好的效果。

3.2.4　HFCC 算法收敛性分析

图 7 表明 HFCC 算法在 5 个数据集上经过 55 次左右迭代迅速达到收敛状态。实验运行情况与 2.3 节关于 HFCC 算法收敛性的理论证明相一致。

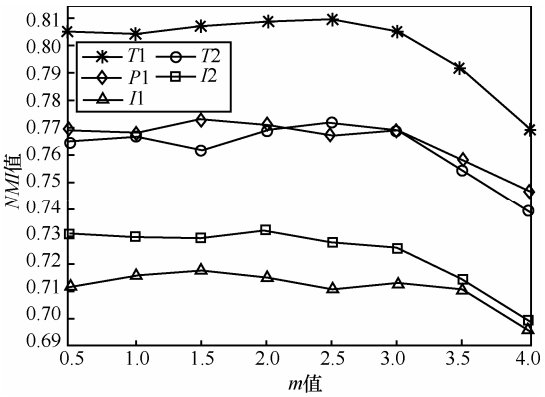


图 6　模糊指数 *m* 对聚类结果的影响

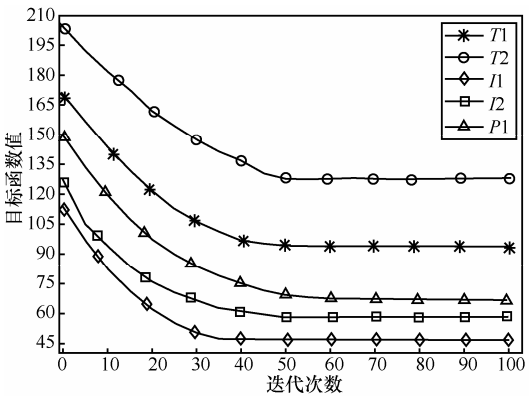


图 7　HFCC 算法的迭代收敛

4 结束语

本文提出了一种高阶异构数据模糊联合聚类算法。与硬聚类算法相比，HFCC 算法不仅能够获得较好的聚类效果，而且能够更有效地描述和处理包含聚簇重叠的高阶异构数据，获得更加符合实际情况的聚类结果。理论和实验 2 种方法验证了 HFCC 算法的收敛性。另外提出适用于确定最佳高阶聚簇数目的 *HXB* 指标，并从实验的角度验证了 *HXB* 指标的有效性。

关于 HFCC 算法仍存在一些需要进一步解决的问题：第一，HFCC 算法没有考虑离群点对聚类结果的影响，HFCC 算法的顽健性以及如何设计出具有更好顽健性的高阶联合聚类算法仍需进一步研究；第二，HFCC 算法是一种无监督的聚类算法，如何引入少量的先验知识以提高聚类效果，仍需进一步探索。

参考文献：

[1] LONG B, WU X Y, ZHANG Z F, *et al*. Unsupervised learning on *k*-partite graphs[A]. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Philadelphia, USA, 2006. 317-326.

[2] DINO I, ROBARDET C, RUGGERO G, *et al*. Parameter-less co-clustering for star-structured heterogeneous data[J]. Data Mining and Knowledge Discovery, 2013, 26(2): 217 -254.

[3] 周志华, 王珏. 机器学习及其应用[M]. 北京: 清华大学出版社, 2007.
ZHOU Z H, WANG J. Machine Learning and Application[M]. Beijing: Tsinghua University Press, 2007.

[4] GAO B, LIU T Y, ZHENG X, *et al*. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering[A]. Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining[C]. Chicago, USA, 2005. 41-50.

[5] GAO B, LIU T Y, MA W Y. Star-structured high-order heterogenous data co-clustering based on consistent information theory[A]. Proceedings of the 6th IEEE International Conference on Data Mining[C]. HongKong, China, 2006. 880-884.

[6] WANG H, NIE F P, HUANG H, *et al*. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation[A]. Proceedings of 11th IEEE International Conference on Data Mining[C]. Arlington, USA, 2011. 174-183.

[7] SHAO J, YIN W T, MA S, *et al*. Topic discovery of Web video using star-structured *k*-partite graph[A]. Proceedings of the International Conference on Multimedia[C]. Firenze, Italy, 2010. 915-918.

[8] GAO B, LIU T Y, FENG G, *et al*. Hierarchical taxonomy preparation for text categorization using consistent bipartite spectral graph co-partitioning[J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(9): 1263-1273.

[9] GAO B, LIU T Y, QIN T, *et al*. Web image clustering by consistent utilization of visual features and surrounding texts[A]. Proceedings of the 13th annual ACM International Conference on Multimedia[C]. Singapore, 2005. 112-121.

[10] REGE M, DONG M, HUA J. Graph theoretical framework for simultaneously integrating visual and textual features for efficient Web image clustering[A]. Proceedings of the 17th International Conference on World Wide Web[C]. Beijing, China, 2008. 317-326.

[11] CHIARAVALLOTI A D, GRECO G, GUZZO A, *et al*. An information-theoretic framework for high-order co-clustering of heterogeneous objects[A]. Proceedings of the 17th European Conference on Machine Learning[C]. Berlin, Germany, 2006. 598-605.

[12] GRECO G, GUZZO A. PONTIERI L.Coclustering multiple heterogeneous domains: linear combinations and agreements[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(12): 1649-1663.

[13] JING L P, YUN J L, YU J, *et al*. High-order co-clustering text data on semantics-based representation model[A]. Proceedings of the 15th Pacific-Asia Conference on Advance in Knowledge Discovery and Data Mining[C]. Shenzhen, China, 2011. 171–182.

[14] SUN Y Z, YU Y T, HAN J W. Ranking-based clustering of heterogeneous information networks with star network schema[A]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. Paris, France, 2009. 797-805.

[15] SUN Y Z, HAN J W, ZHAO P X, *et al*. RankClus: integrating clustering with ranking for heterogeneous information network analysis[A]. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology[C]. Saint Petersburg, Russia, 2009. 565-576.

[16] 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.

[17] XIE X L, BENI G. A validity measure for fuzzy clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1991, 13(8): 841-847.

[18] PAL N R, BEZDEK J C. On cluster validity for fuzzy c-means model[J]. IEEE Transactions on Fuzzy Systems, 1995, 3(3): 370-379.

[19] STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2002, 3(3): 583-617.

作者简介：

黄少滨（1965-），男，黑龙江哈尔滨人，哈尔滨工程大学教授、博士生导师，主要研究方向为数据挖掘、模型检测。

杨欣欣（1987-），男，河南开封人，哈尔滨工程大学博士生，主要研究方向为数据挖掘、社会网络和复杂网络。

申林山（1978-），男，黑龙江哈尔滨人，哈尔滨工程大学讲师，主要研究方向为分布式计算、任务调度算法及计算机审计。

李艳梅（1980-），女，黑龙江哈尔滨人，哈尔滨工程大学博士生，主要研究方向为模型检测、形式化方法和程序验证。