基于理想点的星型高阶联合聚类一致融合策略

黄少滨[1] 杨欣欣[1] 吕天阳[1),2),3)] 郑纬民[3]

[1)](哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)

[2)](中华人民共和国审计署审计科研所 北京 100086)

[3)](清华大学计算机科学与技术系 北京 100084)

摘 要 高阶联合聚类一般被转化为多对二阶联合聚类结果的一致融合问题,将多个二阶聚类目标函数的加权线性组合作为高阶联合聚类的目标函数,通过交替迭代方法得到聚类结果.然而,现有算法仍根据专家经验预设权值,自动的确定线性组合的最优权值仍是一个经典难题.文中针对星型高阶异构数据,提出一种基于理想点的自动确定权值的一致融合策略,将各二阶聚类目标函数的最优值构成的空间中的点称为理想点.通过将二阶聚类结果与其理想结果间的相对距离作为聚类质量的度量标准,解决了各二阶聚类质量不可公度的问题,最终使得高阶聚类目标函数与理想点的相对距离最小.基于理想点的方法能够解决多种星型高阶联合聚类算法的一致融合问题,因此具有一定的普适性.实验结果表明该方法有效地提高了 5 种经典高阶聚类算法的效果.

关键词 异构数据;高阶联合聚类;理想点;一致融合

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2015.01460

Research on Consistent Ensemble of Star-Structure High-Order Co-Clustering Based on Ideal Point

HUANG Shao-Bin[1] YANG Xin-Xin[1] LV Tian-Yang[1),2),3)] ZHENG Wei-Min[3]

[1)](*College of Computer Science and Technology, Harbin Engineering University, Harbin* 150001)

[2)](*Audit Research Institute, National Audit Office of the People's Republic of China, Beijing* 100086)

[3)](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract The problem of high-order co-clustering is converted in to the problem of consistent ensemble of multiple pair two-order co-clustering. Clustering results are obtained by an alternate iterative method, which is used to optimize a weighted combination of objective functions of each pair of two-order co-clustering. However, existing algorithms set the weights according to artificial expert expertise. So far how to automatically determine the optimal weights is a classic problem. Based on ideal point which is the point in space that composed of optimal value of each two-order co-clustering objective function, a strategy of consistent ensemble which can automatically determine the weights is developed for star-structure high-order heterogeneous data. By taking the relative distance between two-order co-clustering results and ideal results as criterion, we solve the problem of incommensurability, and finally minimum the relative distance between high-order co-clustering objective function value and ideal point. Because the strategy based on ideal point

收稿日期:2013-11-13;最终修改稿收到日期:2014-12-08.本课题得到国家自然科学基金(71272216,60903080,60093009)、国家科技支撑计划(2009BAH42B02,2012BAH08B02)、博士后科学基金(2012M510480)、中央高校基本科研业务费专项资金资助项目(HEUCFZ1212,HEUCFT1208)资助.**黄少滨**,男,1965 年生,博士,教授,博士生导师,主要研究领域为数据挖掘、模型检测.E-mail:huangshaobin@hrbeu.edu.cn.**杨欣欣**(通信作者),男,1987 年生,博士研究生,主要研究方向为数据挖掘、社会网络、复杂网络.E-mail:yangxinxin051131@126.com.**吕天阳**,男,1979 年生,博士,副教授,主要研究方向为数据挖掘、社会网络、复杂网络.**郑纬民**,男,1946 年生,博士,教授,博士生导师,主要研究领域为集群计算、高性能存储系统和生物信息学.

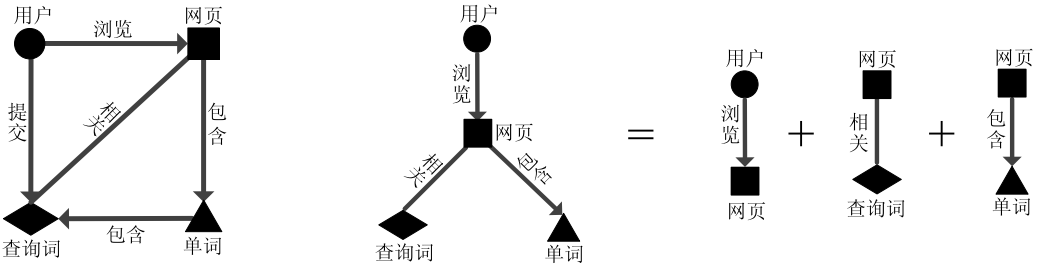
can solve the problem of consistent ensemble of multiple algorithms of high-order co-clustering, it is a general method. Experimental results show that the method can improve the clustering effect of five algorithms of high-order co-clustering.

Keywords　heterogeneous data; high-order co-clustering; ideal point; consistent ensemble

1　引　言

传统的聚类算法主要分析同构数据(Homogeneous data)[1]. 同构数据的数据类型以及数据之间关联关系的类型是单一、同质的[2]. 近年来,信息技术的发展使得包含多种类型数据的数据集广泛出现[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,而且不同类型数据之间的关系也存在明显差异[4].

一般将这种相互关联的多种类型数据称为异构数据(Heterogeneous data),当数据类型为两种时称为二阶异构数据(pair-wise heterogeneous data),当数据类型多于两种时称为高阶异构数据(High-order heterogeneous data)[5]. 依据拓扑结构的特点,异构数据的关联模式被分为网状结构和星型结构[2],分别参见图 1(a)和图 1(b). 星型结构的拓扑特点决定了居于核心类型的数据为分析要点,例如图 1(b)中的论文类型数据,本文针对星型结构展开讨论.



(a) 网状结构　　(b) 星型结构　　(c) 分解为多对二阶数据

图 1　Web 搜索系统中高阶异构数据示例

与同构数据相比,高阶异构数据对分析对象的刻画更丰富,明显有助于提高聚类效果[4]. 因此,针对异构数据的高阶联合聚类方法逐渐成为研究热点[5]. 初期的研究主要是针对二阶异构数据开展二阶联合聚类(pair-wise co-clustering),后文简称二阶聚类. 典型的算法包括谱图划分方法[6]、信息论方法[7]等. 近年来,在二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[5],后文简称高阶聚类,并在 KDD、ICML 等会议和期刊中出现很多重要的研究成果[1,3-5,8-12].

这些高阶聚类方法主要针对星型结构的异构数据. 在聚类过程中,通常将星型结构的异构数据分解为多对二阶数据后分别进行聚类,参见图 1(c),最终合并聚类结果,得到高阶聚类结果. 其中,如何融合并优化各对二阶聚类结果,同时确保星型结构中核心类型数据的聚类结果一致性,就成为高阶聚类面临的核心问题之一,本文将此问题称为一致融合问题. 一致融合问题面临着不同二阶聚类结果间不可公度且矛盾的难题. 不可公度是指缺乏各二阶聚类效果的统一度量标准;矛盾则是指对任意二阶聚类效果的调优,会使得其他二阶聚类结果恶化.

现有的高阶聚类算法主要采用加权线性组合方法解决一致融合问题[1,3,5,8-9,11]. 通过将每个二阶聚类结果的目标函数进行加权线性组合,衡量高阶聚类的全局质量,从而将一致融合问题转化为二阶聚类目标函数的线性组合的最优化问题. 很明显,基于加权线性组合的方法面临如何确定最优权值的问题. 目前仍是根据专家经验预设权值[1,5,8-9,11],自动的确定线性组合的最优权值则是一个经典难题.

由于不可公度性和矛盾性也是多目标线性规划中的典型问题[13],因此高阶聚类的一致融合问题与多目标线性规划方法存在着较强的关联. 例如,文献[10]采用多目标优化中的极大极小方法解决高阶聚类的一致融合问题,但是其方法仅适用于基于信息论的高阶聚类算法,不利于解决其他类型高阶聚类算法的一致融合问题.

为此,本文提出一种基于理想点的一致融合策略. 理想点是多目标优化领域寻找全局妥协解的一

个重要概念,被广泛应用于规划、决策和工程设计等领域[13-14].理想点是由 N 个子目标函数的最优解构成的 R^N 解空间的一个点[13].通过最优化各分目标函数与其理想点之间的距离,从而将多目标优化问题转化为单目标优化问题,为获取全局妥协解提供了方案.

首先给出了高阶聚类中"理想点"的含义.由 $N+1$ 种类型的数据构成的星型结构被分解为 N 对二阶数据,则这 N 个二阶聚类目标函数的最优值构成的 N 维空间的点即为高阶聚类解空间 R^N 中的理想点.由于计算理想点的复杂度极高,因此文中基于高阶聚类迭代过程的特点,提出了近似理想点的概念.通过在第 t 次迭代时第 i 对二阶聚类质量的评价结果与目标函数当前最优值的相对距离作为度量标准,解决了各二阶聚类质量不可公度的问题,使得每次迭代获得的高阶聚类结果的质量与近似理想点的相对距离最小.最终获得的高阶聚类结果对于第 i 对二阶数据可能是次优解,却是解空间中极其逼近理想点的全局妥协解.

由于基于理想点的一致融合策略并不拘泥于具体的评价函数和聚类过程,因此该方法具有较强的普适性,适合于多种高阶聚类算法,包括基于信息论的方法[9]、基于 k 部图的方法[1]、基于矩阵分解的方法[8]、基于一致二部图的方法[5]、基于模糊的方法[11].为了验证这一点,本文提出了基于理想点的解决一致融合问题的通用算法架构.并以模糊高阶聚类方法为例,给出了基于理想点的一致融合方法的实例.实验中,针对 1 个模拟数据集和 3 个真实数据集,比较了 5 种高阶聚类方法采用基于理想点和基于线性组合的一致融合方法的实验效果.结果表明:对于上述数据集,基于理想点的方法均优于最优权值下的加权线性组合方法.

本文第 2 节介绍高阶聚类的相关工作及一致融合问题;第 3 节给出高阶聚类中理想点的定义,并给出基于理想点的高阶聚类一致融合策略框架;第 4 节给出将基于理想点的一致融合策略融入到模糊高阶聚类的具体实例;第 5 节实验与分析;最后总结全文.

2 高阶聚类过程中的一致融合问题

当聚类算法由分析同构数据转向分析二阶异构数据时,分析更高阶的异构数据就成为较为自然的进展.现有的高阶聚类算法主要针对如图 1(b)所示的星型结构的异构数据.

不妨设,对于由 $N+1$ 种类型的数据对象集合 $X=\{x_1,\cdots,x_m\}$, $Y^1=\{y_1^1,\cdots,y_{n_1}^1\}$, $\cdots$, $Y^i=\{y_1^i,\cdots,y_{n_i}^i\}$, $\cdots$, $Y^N=\{y_1^N,\cdots,y_{n_N}^N\}$ 构成如图 1 所示的星型结构.每类非核心类型的数据对象 Y^i 与核心类型数据对象 X 关联构成一对二阶异构数据 $\{X,Y^i\}$.设 M^i 是 X 与 Y^i 关联关系矩阵,其中第 p 行 q 列的元素表示 x_p 与 y_q^i 的关系强度.设 $f_{X,Y^i}(C_X,C_{Y^i})$ 为衡量二阶异构数据 $\{X,Y^i\}$ 聚类效果的评价函数,C_X 和 C_{Y^i} 分别表示 X 和 Y^i 的聚类结果,其中 k 是 X 的聚簇个数、l_i 是 Y^i 的聚簇个数.

现有的算法主要将高阶聚类问题转化为对多对二阶聚类结果加权线性组合的最优化问题,具体地说:

(1)每对二阶数据 $\{X,Y^i\}$ 的聚类质量用评价函数 $f_{X,Y^i}(C_X,C_{Y^i})$ 进行衡量,不同方法采用的函数有所不同,但是同一方法使用的评价函数相同.其中,基于信息论的方法(后文简称 CIT 方法)利用互信息损失[7,9];基于模糊的方法(后文简称 FCMR 方法)利用聚集度[11,17];基于二部图划分的方法(后文简称 CBGC 方法)利用正则割[5-6];基于 k 部图的方法(后文简称 RSN 方法)利用原始数据 k 部图和关系概要网络之间的差距[1];基于矩阵分解的方法(后文简称 SRC 方法)利用原关系矩阵和低维空间重构矩阵之间的差别[8,15].

(2)"全局"的高阶聚类质量将每对二阶聚类结果的评价函数进行加权线性组合,作为统一的目标函数进行最优化.以图 2 为例,CBGC 算法中的加权线性组合方法就是将分别衡量 $\{X,Y^1\}$、$\{X,Y^2\}$ 聚类质量的评价函数进行线性组合 $\beta f_{X,Y^1}+(1-\beta)f_{X,Y^2}$ 的最优化问题[5,16],其中因子 $\beta>0$ 用于表示评价函数的权重.更高阶的情况与之类似.

(3)采用交互迭代的方法求解最优的高阶聚类结果.在第 t 次迭代中,首先在 $X^{(t)}$ 的聚类结果 $C_X^{(t)}$ 一定的情况下,求解 Y^i 的聚类结果 $C_{Y^i}^{(t)}$,使得对 $\{X,Y^i\}$ 的聚类质量评价最优;而后在 $C_{Y^i}^{(t)}$ 一定的情况下,求解 $(C_X^{(i)})^{(t)}$;最后调整不同二阶聚类结果中 $(C_X^{(i)})^{(t)}$,使得对核心对象 X 的聚类结果一致,记为 $C_X^{(t+1)}$,$C_X^{(t+1)}$ 使得每对二阶聚类评价函数加权和最优.反复迭代,直至高阶聚类结果达到收敛或达到指定迭代次数.

很明显,各对二阶聚类的质量存在冲突性,提高

一对二阶聚类的质量会降低另一对二阶聚类的效果，因此需要设定恰当的加权线性组合的权值，才能获得较好的全局聚类结果. 而且 β 取值对聚类结果的影响较大. 仍以图 2 为例，分析 5 种高阶聚类算法的效果，若数据间存在边，则二者的关系强度为 1；否则为 0. 当 β 取值为 0.9、0.8 或 0.7 时，CIT、CBGC 和 FCMR 算法得到如图 2 实线划分的聚类结果，此结果对 Y^2 做出不合适的划分；当 β 取值为 0.6、0.5 或 0.4 时，CIT、RSN、CBGC、SRC 和 FCMR 得到如图 2 虚线划分的聚类结果；β 取值为 0.3、0.2 或 0.1 时，CIT、RSN、CBGC 和 SRC 得到如图 2 点画线划分的聚类结果，此结果对 Y^1 做出不合适的划分.

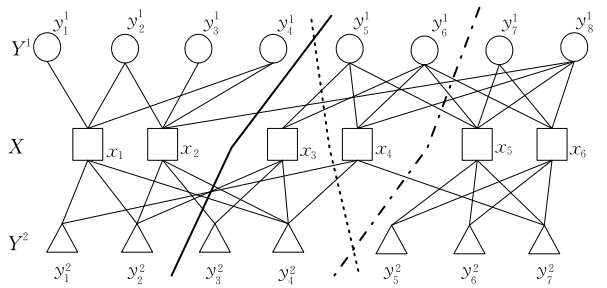


图 2　3 种类型数据构成的星型结构数据的高阶聚类结果示例

综上，高阶聚类中每对二阶聚类结果的一致融合方法需要对评价函数的加权线性组合选取合适的权值. 但是，现有研究中并未深入探索加权线性组合中权值的确定问题，而是由用户根据经验指定[1,5,8-9,11]. 所以如何寻找一种自动确定权值的一致融合方法成为急需解决的问题.

3　基于理想点的高阶聚类一致融合策略

为了解决星型结构高阶异构数据在聚类过程中各二阶聚类结果线性组合的权值确定问题，本节提出基于理想点的一致融合策略. 首先给出高阶聚类中理想点的含义；其次给出高阶聚类过程中获取理想点的策略；最后给出基于理想点的一致融合策略的基本架构.

很明显，高阶聚类问题可视为由 N 个评价函数对应的 N 个子目标函数 $f_{X,Y^i}(C_X,C_{Y^i})$ 组成的多目标函数的最优化问题. 为使各子目标函数的取值均尽可能最优，可以先分别求出各子目标函数的最优值，然后使得高阶聚类结果尽量接近其最优值. 由于不同二阶聚类结果对于核心对象 X 的聚类通常存在冲突，因此各子目标函数的最优值是高阶聚类结果希望得到、却通常无法得到的理想值，故称之为理想点，具体定义如下.

定义 1.　称 $(f^*_{X,Y^1}(C_X^{(1)},C_{Y^1}),\cdots,f^*_{X,Y^N}(C_X^{(N)},C_{Y^N}))$ 为高阶聚类在解空间 R^N 中的理想点，如果对任意聚类结果 C'_X 和 C'_{Y^i} 满足

$$f^*_{X,Y^i}(C_X^{(i)},C_{Y^i})\leq f_{X,Y^i}(C'_X,C'_{Y^i}),\ i=1,\cdots,N \tag{1}$$

在理想点的定义中，并没有要求各二阶聚类结果中的 $C_X^{(i)}$ 是一致的，因此在绝大多数情况下高阶聚类的结果只能逼近理想点.

但是，获得理想点的计算复杂度却极高. 由于各种二阶聚类方法均采用启发式过程，其最终结果仅为子目标函数 $f_{X,Y^i}(C_X,C_{Y^i})$ 最优值的近似. 因此，只能在 C_X 和 C_{Y^i} 均为变量的情况下求解 $f_{X,Y^i}(C_X,C_{Y^i})$ 得到的最优值，直接导致状态空间爆炸问题. 例如，图 2 所示的简单数据集就有 240831 种解. 真实数据集数据量庞大，显然从指数级增长的状态空间中选取最优解的复杂度极高，并不适合直接用于高阶聚类过程.

为了解决这一问题，本文借助高阶聚类在迭代过程中的特点，给出理想点的近似值，用于指导聚类过程. 如前所述，在高阶聚类的第 t 次迭代中，先在 $C_X^{(t)}$ 一定的情况下，求解 $C_{Y^i}^{(t)}$，而后在 $C_{Y^i}^{(t)}$ 一定的情况下，求解 $(C_X^{(i)})^{(t)}$. 此时，二阶异构数据 $\{X,Y^i\}$ 的聚类结果 $C_{Y^i}^{(t)}$ 和 $(C_X^{(i)})^{(t)}$，是在 $C_{Y^i}^{(t)}$ 一定的情况下 f_{X,Y^i} 的最佳解. 由于不同二阶聚类结果 $(C_X^{(i)})^{(t)}$ 中核心对象 X 的聚类结果通常不一致，因此将 $C_{Y^i}^{(t)}$ 一定的情况下的最佳解构成的 R^N 解空间中的点称为第 t 次迭代时的近似理想点，定义如下.

定义 2.　称 $(f^{(t)*}_{X,Y^1}((C_X^{(1)})^{(t)},C_{Y^i}^{(t)}),\cdots,f^{(t)*}_{X,Y^N}((C_X^{(N)})^{(t)},C_{Y^i}^{(t)}))$ 为高阶聚类在第 t 次迭代中在解空间 R^N 中的近似理想点，如果在 $C_{Y^i}^{(t)}$ 确定的情况下，$\forall C'_X$ 满足

$$f^{(t)*}_{X,Y^i}((C_X^{(i)})^{(t)},C_{Y^i}^{(t)})\leq f_{X,Y^i}(C'_X,C_{Y^i}^{(t)}),\ i=1,\cdots,N \tag{2}$$

为了与近似理想点相对应，又将理想点称为绝对理想点. 需要注意的是，在 $C_{Y^i}^{(t)}$ 一定的情况下 f_{X,Y^i} 的最佳解在迭代过程中已经获取，并不需要重新计算. 同时它可能不是 f_{X,Y^i} 的全局最优值或极值，仅是本次迭代中在 $C_{Y_i}^t$ 约束下 f_{X,Y^i} 最优解的近

似.因此在迭代初始和迭代过程中,近似理想点可能与绝对理想点相距较远,但是近似理想点将在迭代过程中逐渐逼近绝对理想点.这是由二阶聚类的交互迭代过程的特点决定的:每次迭代时最优化评价函数 $f_{X,Y^i}^{(t)}((C_X^{(i)})^{(t)},C_{Y^i}^{(t)})$ 的取值,都会提高 Y^i 的聚类效果[7],并逐渐逼近绝对理想点所对应的 C_{Y^i},最终使得近似理想点逼近绝对理想点.

虽然多个目标函数之间具有固有的不可公度性,即不同的 f_{X,Y^i} 之间没有统一的度量标准,但是各二阶聚类结果在每次迭代时,趋近于近似理想点所对应的最佳解的程度却可以相互比较.为此,在近似理想点的基础上提出相对距离的概念.

定义 3. 称 $R_{X,Y^i}^{(t)}(C'_X,C_{Y^i}^{(t)})$ 为二阶聚类结果 $(C'_X,C_{Y^i}^{(t)})$ 相对其本次迭代最佳解间的相对距离

$$R_{X,Y^i}^{(t)}(C'_X,C_{Y^i}^{(t)})=\frac{|f_{X,Y^i}^{(t)}(C'_X,C_{Y^i}^{(t)})-f_{X,Y^i}^{(t)*}((C_X^{(i)})^{(t)},C_{Y^i}^{(t)})|}{f_{X,Y^i}^{(t)*}((C_X^{(i)})^{(t)},C_{Y^i}^{(t)})} \tag{3}$$

相对距离能够解决不同二阶聚类结果质量的不可公度问题,最终使得各二阶数据与近似理想点相对差距之和最小,从而获得较好的全局妥协解 $(C_X)^{(t)}$,$(C_X)^{(t)}$ 满足

$$\sum_{i=1}^{N}R_{X,Y^i}^{(t)}((C_X)^{(t)},C_{Y^i}^{(t)})\leq\sum_{i=1}^{N}R_{X,Y^i}^{(t)}(C'_X,C_{Y^i}^{(t)})\ \forall C'_X \tag{4}$$

因此在第 t 次迭代中,最优化的目标函数为

$$\min\sum_{i=1}^{N}R_{X,Y^i}^{(t)}((C_X)^{(t)},C_{Y^i}^{(t)})=\sum_{i=1}^{N}\frac{f_{X,Y^i}^{(t)}((C_X)^{(t)},C_{Y^i}^{(t)})}{f_{X,Y^i}^{(t)*}((C_X^{(i)})^{(t)},C_{Y^i}^{(t)})} \tag{5}$$

与加权线性组合策略相比,基于理想点的一致融合策略相当于在第 t 次迭代中设定子目标函数 f_{X,Y^i} 的线性组合权值 $\beta_i^{(t)}$ 为

$$\beta_i^{(t)}=\frac{1/f_{X,Y^i}^{(t)*}((C_X^{(i)})^{(t)},C_{Y^i}^{(t)})}{1/f_{X,Y^1}^{(t)*}((C_X^{(i)})^{(t)},C_{Y^i}^{(t)})+\cdots+1/f_{X,Y^N}^{(t)*}((C_X^{(i)})^{(t)},C_{Y^i}^{(t)})},\quad 1\leq i\leq N \tag{6}$$

由于基于理想点的方法并不拘泥于特定的二阶聚类子目标函数,而且不同高阶聚类算法的迭代过程具有共同特征,因此该方法适用于多种高阶聚类算法,从而提供了一种具普遍意义的解决一致融合问题的策略.至此,提出的基于理想点的一致融合高阶聚类框架如下.

框架 1. 基于理想点的一致融合高阶聚类框架.

输入:关系矩阵 $\boldsymbol{M}^{(i)}$,X 聚簇数目 k,Y^i 聚簇数目 l_i

输出:高阶聚类划分函数 $(C_X,C_{Y^1}^0,\cdots,C_{Y^N}^0)$

1. 初始化:随机初始化 $(C_X,C_{Y^1}^0,\cdots,C_{Y^N}^0)$,设 $t=0$;

2. 对于每对二阶数据 (X,Y^i),依据核心类型数据 X 的聚类结果 $(C_X^{(i)})^{(t)}$ 得出 Y^i 的聚类结果 $C_{Y^i}^{(t+1)}$;

3. 对于 (X,Y^i),依据 Y^i 的聚簇结果 $C_{Y^i}^{(t+1)}$ 得出 X 的聚类结果 $(C_X^{(i)})^{(t+1)}$;

4. 将 (X,Y^i) 的聚类结果 $((C_X^{(i)})^{(t+1)},C_{Y^i}^{(t+1)})$ 代入评价函数,得到理想点 $(f_{X,Y^1}^{(t)*},\cdots,f_{X,Y^N}^{(t)*})$,基于式(6)计算 $\beta^{(t)}$;

5. 根据式(4)得出一致融合后核心类型数据 X 的聚簇结果 $(C_X)^{(t+1)}$;

6. $t=t+1$,重复步 2~5 直到收敛.

基于理想点的一致融合聚类策略的时间复杂度分析如下:

(1)步 2~3 为二阶聚类过程,复杂度取决于具体采用的二阶聚类算法,对二阶异构数据 $\{X,Y^i\}$,CIT 方法的时间复杂度为 $O((k+l_i)mn_i)$,SRC 方法为 $O((k+l_i)(m+n_i)^2)$,RSN 方法为 $O((k+l_i)(m+n_i)^2)$,FCMR 方法为 $O(kmn_i)$,CBGC 方法为 $O((m+n_i)^3)$.其中,m 是核心数据对象 X 的数据量,k 是 X 的聚簇个数,n_i 是非核心数据对象 Y^i 的数据量,l_i 是 Y^i 的聚簇个数.

(2)步 4 计算本次迭代的近似理想点以及各子目标函数的权值大小,采用 CIT 方法的时间复杂度为 $O(kl_i)$,SRC 方法为 $O(kl_i(m+n_i))$,RSN 方法为 $O(kl_i(m+n_i))$,FCMR 方法为 $O(kmn_i)$,CBGC 方法为 $O(kmn_i)$.可见,计算近似理想点和各子目标函数权值的时间复杂度均低于相应的二阶聚类算法的时间复杂度.

(3)步 5 为一致融合过程.在目标函数权值确定的情况下,可以利用加权线性组合的方法计算核心类型数据的聚类结果.加权线性组合一致融合方法同样采用了步 2 和步 5,与之相比,理想点方法增加了步 3 和步 4,步 3 的时间复杂度与步 2 相同,步 4 的时间复杂度低于步 2.所以基于理想点的一致融合过程的时间复杂度不会高于基于加权线性组合的一致融合过程的时间复杂度.

综上,基于理想点的高阶聚类方法的时间复杂度不高于对应的基于加权线性组合的高阶聚类方法的时间复杂度.

4 基于理想点的模糊高阶聚类算法

4.1 基于理想点的模糊高阶聚类算法

本节以模糊高阶聚类为例,说明如何将基于理

想点的一致融合策略融入到具体的高阶聚类算法中. 而后简单论述基于理想点的策略对于其他高阶聚类算法的适用性.

模糊高阶聚类的目标是最大化聚类内部数据之间的“凝聚”程度[17]. 对于高阶数据$(X,\cdots,Y^N)$，模糊高阶聚类用隶属度 u_{kp}，$v_{kq}^{(i)}\in[0,1]$分别描述数据 x_p 和 $y_q^{(i)}$ 属于聚簇 k 的程度，隶属度越接近 1，表示属于聚簇 k 的程度越高[17]. 模糊高阶聚类希望将关系较强的数据划分到同一个簇中，即若 x_p 和 $y_q^{(i)}$ 之间的关系强度 $d_{pq}^{(i)}$ 较大，则对于某聚簇 k，希望 x_p 和 $y_q^{(i)}$ 具有较大隶属度 u_{kp} 和 $v_{kq}^{(i)}$，那么关系强度 $d_{pq}^{(i)}$ 与隶属度 u_{kp} 和 $v_{kq}^{(i)}$ 的乘积 $u_{kp}v_{kq}^{(i)}d_{pq}^{(i)}$ 较大. 对于数据集中所有数据 $\sum_{p=1}^{m}\sum_{q=1}^{n_1}u_{kp}v_{kq}^{(i)}d_{pq}^{(i)}$ 较大，其中 $\sum_{p=1}^{m}\sum_{q=1}^{n_1}u_{kp}v_{kq}^{(i)}d_{pq}^{(i)}$ 称为聚簇 k 的聚集度. 聚集度反映了聚簇内部数据之间的“凝聚”程度.

对于二阶数据$\{X,Y^i\}$，二阶模糊聚类的目标函数如下：

$$J_i=\sum_{k=1}^{K}\sum_{p=1}^{m}\sum_{q=1}^{n_i}u_{kp}v_{kq}^{(i)}d_{pq}^{(i)}-T_u\sum_{k=1}^{K}\sum_{p=1}^{m}u_{kp}^2-T_v^{(i)}\sum_{k=1}^{K}\sum_{q=1}^{n_i}v_{kq}^2$$

$$\text{s.t.}\ \sum_{k=1}^{K}u_{kp}=1,\ \sum_{q=1}^{n_i}v_{kq}^{(i)}=1 \tag{7}$$

其中，参数 T_u 和 $T_v^{(i)}$ 分别用于调节 u_{kp} 和 $v_{kq}^{(i)}$ 的模糊程度. T_u 越大，$u_{kp}(1\leq k\leq K)$ 越接近于平均值 $1/K$，越难以判别 x_p 属于哪一个簇，模糊程度越高.

由于很难直接求得此类目标函数的最优解，因此采用交互优化(Alternative optimization)求解策略. 在 u_{kp} 一定的情况下利用 $v_{kq}^{(i)}$ 的更新规则求解 $v_{kq}^{(i)}$；再在 $v_{kq}^{(i)}$ 一定的情况下利用 u_{kp} 的更新规则求解 u_{kp}，如此交替进行迭代，直至目标函数收敛到最优值.

模糊高阶聚类算法希望各对二阶数据“凝聚”程度达到最大，尽量最大化各对二阶数据的聚集度，即同时最大化各对二阶模糊聚类结果的目标函数. 采用基于加权线性组合一致融的方法的目标函数为 $\sum_{i=1}^{N}\beta_iJ_i$，其中 β_i 表示各子目标函数的权重，相应的隶属度 u_{cp} 的更新规则为

$$u_{kp}=\frac{1}{K}+\frac{1}{2T_u}\left[\sum_{i=1}^{N}\beta_i\sum_{q=1}^{n_i}v_{kq}^{(i)}d_{pq}^{(i)}-\frac{1}{K}\sum_{l=1}^{K}\sum_{i=1}^{N}\beta_i\sum_{q=1}^{n_i}v_{lq}^{(i)}d_{pq}^{(i)}\right] \tag{8}$$

与之相比，基于理想点的一致融合方法希望各对二阶数据聚集度尽量接近绝对理想点. 因此在每次迭代中，由核心类型数据隶属度计算非核心类型数据隶属度，再由非核心类型数据隶属度计算核心类型数据隶属度，进而计算本次迭代的近似理想点. 每次迭代中，一致融合后的核心类型数据隶属度结果使得各对二阶目标函数值与近似理想点的相对距离最小；随着迭代的进行，每次迭代的近似理想点将逐渐逼近绝对理想点.

以下以图 2 数据集为例，说明基于理想点的模糊高阶聚类算法的第一次迭代过程，其中参数 $T_u=0.5$，$T_v^{(1)}=2$，$T_v^{(2)}=2$.

(1) 随机初始化隶属度 u_{kp}，将核心类型 X 的数据划分到隶属度最大的聚簇中，参见图 3(a)，其中黑色表示数据被聚到第 1 个聚簇中，灰色表示数据被聚到第 2 个聚簇中. 此过程对应基于理想点一致融合框架的步 1；

(2) 对于每对二阶数据$\{X,Y^i\}$，在 u_{kp} 一定的情况下，根据二阶模糊聚类 v_{kq} 的更新规则计算 $v_{kq}^{(i)}$，此过程对应基于理想点一致融合框架的步 2；而后在 $v_{kq}^{(i)}$ 一定的情况下，根据二阶模糊聚类 u_{kp} 的更新规则计算 $u_{kp}^{(i)}$. 每对二阶数据$\{X,Y^i\}$得到的 $u_{kp}^{(i)}$ 和聚类结果参见图 3(b)，此过程对应基于理想点一致融合框架的步 3. 由图可见，不同二阶模糊聚类结果对于核心类型数据的聚簇存在明显的不一致性，相应的聚类结果如图 3(b)虚线框和实线框所示，对于数据 x_3 和 x_4 的聚簇结果截然不同；

(3) 将每对二阶数据$\{X,Y^i\}$的隶属度 $u_{kp}^{(i)}$ 和权重 $v_{kq}^{(i)}$ 代入式(7)，得到的近似理想点为(1.097，0.8931)，利用式(6)计算的到权值 $\beta=0.4484$. 此过程对应基于理想点一致融合框架的步 4；

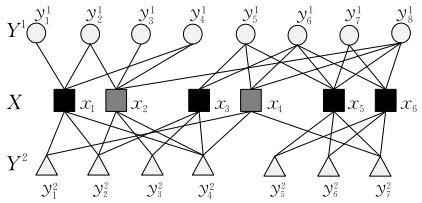
(4) 由式(8)计算一致融合后的隶属度 u_{kp}，并获得一致融合后的聚类结果，依据式(3)计算其与近似理想点的相对距离为 0.2622，聚类结果和新的隶属度 u_{kp} 取值参见图 3(c). 新的隶属度 u_{kp} 将作为下一次迭代的初始隶属度. 此过程对应基于理想点一致融合框架的步 5.

重复上述过程直到聚类结果达到收敛或达到指定迭代次数. 图 3(d)将该数据集在解空间中的全部可能解进行了可视化，其中对应解距离绝对理想点越近越趋向于深黑色，同时给出了每次迭代后的解及其近似理想点，可以看出基于理想点的方法每次迭代中获取的近似理想点在逐渐逼近绝对理想点.

综上，在基于理想点的高阶聚类框架下，提出基于理想点一致融合的模糊高阶聚类算法，其计算步

核心类型数据的初始化隶属度值

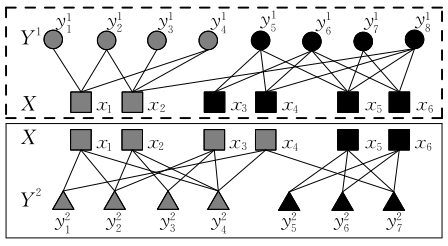
	x_1	x_2	x_3	x_4	x_5	x_6
簇1	**0.578**	0.272	**0.612**	0.457	**0.866**	0.552
簇2	0.422	**0.728**	0.388	**0.543**	0.134	0.448



(a) 初始化聚类结果

核心类型数据的隶属度

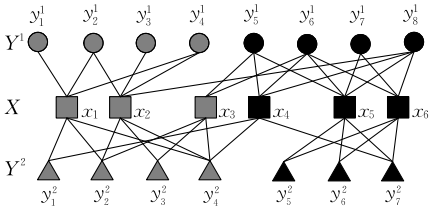
		x_1	x_2	x_3	x_4	x_5	x_6
$u_{kp}^{(1)}$	簇1	0.431	0.411	**0.549**	**0.576**	**0.609**	**0.575**
	簇2	**0.569**	**0.589**	0.451	0.424	0.391	0.426
$u_{kp}^{(2)}$	簇1	0.431	0.405	0.405	0.447	**0.609**	**0.610**
	簇2	**0.569**	**0.595**	**0.595**	**0.553**	0.391	0.391



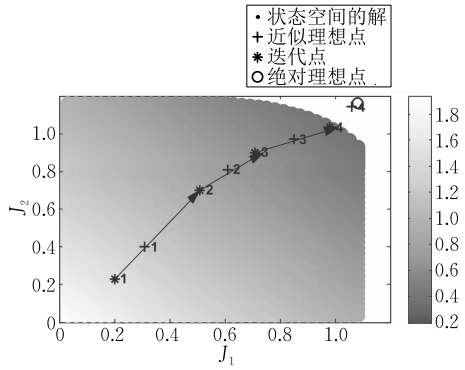
(b) 第1次迭代各二阶数据的模糊聚类结果及$u_{kp}^{(i)}$取值

核心类型数据的隶属度

	x_1	x_2	x_3	x_4	x_5	x_6
簇1	0.431	0.408	0.469	**0.505**	**0.609**	**0.594**
簇2	**0.569**	**0.592**	**0.531**	0.495	0.391	0.406



(c) 一致融合后核心类型数据聚类结果及新的隶属度u_{kp}



(d) 模糊高阶聚类解空间的形态及历次迭代获得的解及近似理想点在解空间的位置

图 3 针对图 2 所示数据的基于理想点的模糊高阶聚类算法实例

骤与基于理想点的高阶聚类框架一一对应,具体描述如下.

算法 1. 基于理想点一致融合的模糊高阶聚类算法.

输入:数据关系 $d_{pq}^{(i)}$,聚簇数目 K,模糊度参数 T_u 和 $T_v^{(i)}$,终止条件阈值 ε 和最大迭代次数 $t_{\max}$

输出:X 隶属度 u_{kp} 和 Y^i 的权重 $v_{kp}^{(i)}$

1. 随机产生 X 隶属度$(u_{kp})^{(0)}$;

2. 对于每对二阶数据$(X,Y^{(i)})$,在$(u_{kp})^{(t)}$一定的情况下,根据二阶模糊聚类$(v_{kq}^{(i)})^{(t)}$的更新规则计算$(v_{kq}^{(i)})^{(t+1)}$;

3. 在$(v_{kq}^{(i)})^{(t+1)}$一定的情况下,根据二阶模糊聚类 u_{kp} 的更新规则计算$(u_{kp}^{(i)})^{(t+1)}$;

4. 根据式(7)计算近似理想点$(f_{X,Y^1}^{(t)*},\cdots,f_{X,Y^N}^{(t)*})$,利用式(6)计算权值 $\beta_i^{(t)}$;

5. 根据由式(8)计算一致融合后的隶属度$(u_{kp})^{(t+1)}$;

6. 如果 $\max_{k,p}|(u_{kp})^{(t+1)}-(u_{kp})^{(t)}|\leq\varepsilon$ 或 $t=t_{\max}$,则算法结束;否则 $t=t+1$,跳转第 2 步.

基于理想点法的一致融合方法具有较强的普遍性,不仅适于模糊聚类算法,还适合于多种典型的高阶聚类算法.其基本原理都是在每次迭代中依据核心类型数据 X 计算非核心类型数据的聚类结果,再由非核心类型数据计算核心类型数据结果.然后计算本次迭代的近似理想点,不同的算法计算方法如下:CIT 方法计算各对二阶数据的信息损失;CBGC 方法计算各对二阶数据的正则割[5-6];RSN 方法计算各对二阶数据原始数据 k 部图和关系概要网络之间的差距[1];SRC 方法计算各对二阶数据原关系矩阵和低维空间重构矩阵之间的差别[8,15].而后即根据式(6)计算本次迭代中各子目标函数的权值,再采用加权线性组合的方法获取核心类型数据的聚类结果.

4.2 收敛性证明

以下以基于理想点一致融合的模糊高阶聚类算法为例证明其收敛性,同理可以证明其他基于理想点的高阶联合聚类算法的收敛性,不再一一赘述.

单调有界定理指出单调有界函数必收敛,因此为证明基于理想点一致融合的模糊高阶聚类算法收敛,只需证明目标函数 $J=\sum_{i=1}^{N}\beta_i^{(t)}J_i^{(t)}$ 是关于迭代次数 t 的单调递增函数.根据式(8)更新 $u=(u_{11},\cdots,u_{Km})$时,$v^{(i)}=(v_{11}^{(i)},\cdots,v_{Kn_i}^{(i)})$为常数,目标函数 J 是以 u 为自变量的函数,用 $J(u)$表示.由朗格拉日乘子法,根据式(8)计算的 u^* 值为 $J(u)$在约束条件(7)下的一个驻点.海森矩阵 $\nabla^2J(u)$在点 u^* 处为负定矩阵如下:

$$\nabla^2 J(u)=\begin{bmatrix}\dfrac{\partial J(u)^2}{\partial u_{11}u_{11}} & \cdots & \dfrac{\partial J(u)^2}{\partial u_{11}u_{Km}}\\ \vdots & \ddots & \vdots\\ \dfrac{\partial J(u)^2}{\partial u_{Km}u_{11}} & \cdots & \dfrac{\partial J(u)^2}{\partial u_{Km}u_{Km}}\end{bmatrix}=\begin{bmatrix}-2T_u & & 0\\ & \ddots & \\ 0 & & -2T_u\end{bmatrix} \tag{9}$$

T_u为正值，所以海森矩阵$\nabla^2 J(u^*)$为负定矩阵，u^*是$J(u)$的极大值，有$J(u^t)\leq J(u^*)=J(u^{t+1})$，其中$u^t$表示第$t$次迭代结束后$u$的值，$u^*=u^{t+1}$表示第$t+1$次迭代中$u$的值. 因此，依据式(8)迭代更新$u=(u_{11},\cdots,u_{Km})$不会使$J$值下降. 同理可证明根据式(10)迭代更新$v^{(i)}=(v_{11}^{(i)},\cdots,v_{Kn_i}^{(i)})$不会使式(7)$J_i$值下降.

$$v_{kq}^{(i)}=\frac{1}{n_i}+\frac{1}{2T_v^{(i)}}\left[\beta_i\sum_{p=1}^{m}u_{kp}d_{pq}^{(i)}-\frac{1}{n_i}\beta_i\sum_{q=1}^{n_i}\sum_{p=1}^{m}u_{kp}d_{pq}^{(i)}\right] \tag{10}$$

综上，基于理想点一致融合的模糊高阶聚类算法的目标函数J必定收敛于约束条件(7)下的全局最大值.

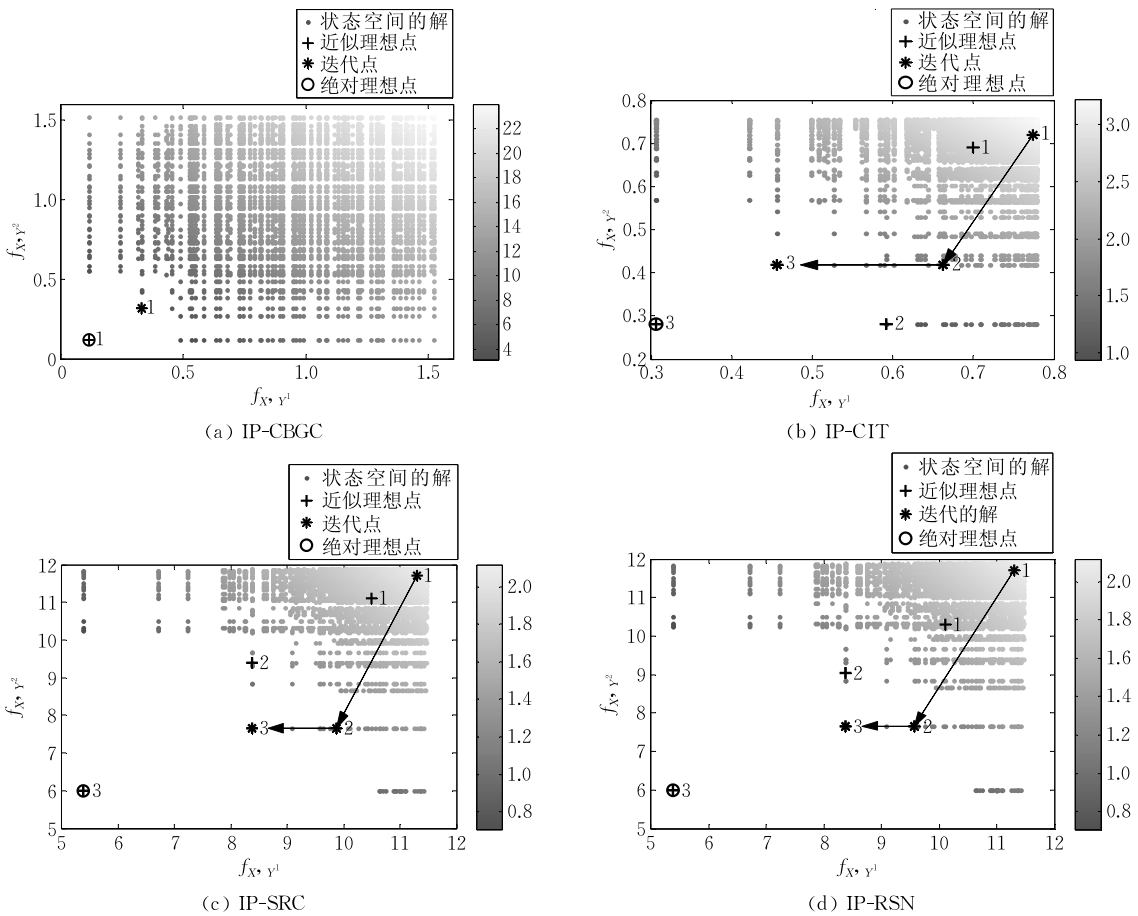
5　实验分析

实验中主要将基于线性组合传统高阶聚类算法与基于理想点的对应算法进行对比分析. 文中基于理想点法的算法简称规则是在原简称的前部添加IP-，例如基于理想点的二部图划分算法简称为IP-CBGC，其他与之类似，不再赘述. 共比对CBGC与IP-CBGC、CIT与IP-CIT、RSN与IP-RSN、SRC与IP-SRC、FCMR与IP-FCMR五组算法.

考虑到实验所涉及的算法具有一定的随机性，本实验统计20次实验结果NMI值[18]的平均值用于聚类效果对比分析. 除上述对聚类结果的评价标准外，运行时间也是算法评价的重要指标，实验运行环境：Pentium(R) Dual E2180，2.5 GHz，内存2.0 GB.

5.1　Tony数据集实验

设f_{X,Y^1}和f_{X,Y^2}分别为数据$\{X,Y^1\}$和$\{X,Y^2\}$的二阶聚类目标函数. 图3(d)和图4给出图2示例



(a) IP-CBGC　(b) IP-CIT

(c) IP-SRC　(d) IP-RSN

图4　针对图2所示数据的基于理想点的高阶聚类解空间的形态及历次迭代获得的解及近似理想点在解空间的位置

数据集基于理想点的高阶聚类解空间的形态及历次迭代获得的解及近似理想点在解空间的位置.其中点表示状态空间的解,即所有可能的高阶聚类结果,距离绝对理想点相对距离越近,颜色深度越深;加号符"+"表示每次迭代的近似理想点;星号符"*"表示某次迭代的解,即一致融合后高阶聚类结果;圆表示绝对理想点.首先,图 3(d)和图 4 表明,近似理想点逐渐逼近绝对理想点如图 4,或者最终极其逼近理想点如图 3(d).其次,选择与绝对理想点相对距离最近的状态空间中的解,作为最终的高阶聚类结果.最后,图 3(d)和图 4 反映一个有趣的现象:状态空间与绝对理想点构成一个长方形,但在绝对理想点附近形成一个空白区域,这个空白区域就是高阶聚类结果不能达到的状态空间中的解.

5.2 公开数据集实验

Corel 数据集是来自 50 Corel Stock Photo CDs,其中每个图像具有文字标注信息和该图像的图像分割组[19].从表 1 所示的每个主题中选取 90 个图像,采用文献[19]的方法,构建数据集 I1 和 I2.

表 1 公开数据集

简称	数据集	单词数	对象数	类别数目	类别
I1	Corel	114	630	7	sunsets, tigers, trains, swimmers, car, skyscrapers, war airplanes
I2	Corel	116	540	6	bears, deers, horses, cliffs, birds, bridges
P1	Cora	500	900	3	information retrieval, database, artificial intelligence
P2	Cora	500	1200	4	networking, operating systems, architecture, database
T1	Newsgroups	1000	1500	3	{comp.graphics, comp.os.ms-windows.misc} {rec.autos, rec.motorcycles}{sci.crypt, sci.electronics}
T2	Wap, oh15, re0	1000	1500	3	{Film, Television, Health}{Aden-Diph, Cell-Mov, Enzyme-Act} {interest, trade.money}

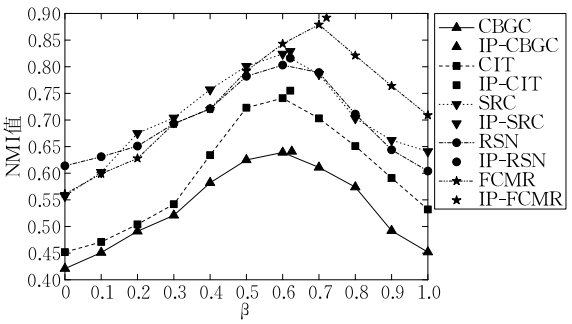
Cora 为科技论文数据集,采用文献[20]的方法,由会议/期刊、论文和单词组成高阶异构数据集 P1 和 P2,论文类别、单词数参见表 1,每个类别包含 300 篇论文.

Newsgroups 标准数据集包含来自 20 个新闻组的 20000 篇文章.oh15 是取自 MEDLINE 数据集的子集 OHSUMED 的样本,MEDLINE 数据集包含从 1987 年到 1991 年 270 个医学期刊收录的文章摘要.re0 是取自 Reuters-21578 文本分类测试集.Wap 取自 WebACE 项目的数据集,是 Yahoo!主题分层栏目的网页.采用文献[3]的方法构造如表 1 描述的数据集 T1 和 T2,建立文本-单词以及文本-主题的关系矩阵.高阶聚类的目标是将文本-主题聚类为更高层次的主题,例{comp.graphics, comp.os.ms-windows.misc}属于更高层次的主题 computer.

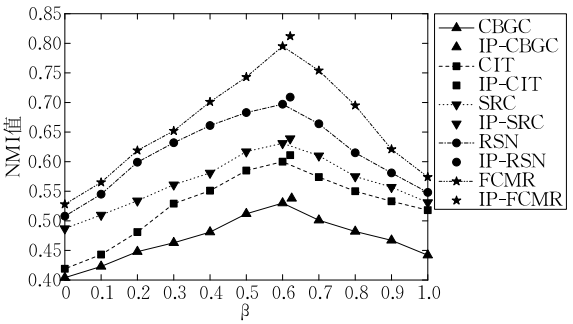
设由 X、Y^1 和 Y^2 组成的高阶异构数据集 S,其中 X 为核心类型数据,X、Y^1 和 Y^2 均包含 200 个数据,前 100 个数据属于第 1 个类别,后 100 个数据属于第 2 个类别.采用文献[8]方法生成模拟数据集 S 中 $\{X,Y^1\}$ 的关系矩阵 $\boldsymbol{M}^{(1)}$ 和 $\{X,Y^2\}$ 的关系矩阵 $\boldsymbol{M}^{(2)}$.设 $S^{(1)}=\begin{bmatrix}0.9 & 0.7\\ 0.8 & 0.9\end{bmatrix}$,第 i 行第 j 列的元素为 $S_{ij}^{(1)}$;$S^{(2)}=\begin{bmatrix}0.3 & 0.4\\ 0.7 & 0.6\end{bmatrix}$,第 i 行第 j 列的元素为 $S_{ij}^{(2)}$.X 第 i 个聚簇中的数据与 Y^1 第 j 个聚簇中的数据相互关联的概率是 $S_{ij}^{(1)}$,若两个数据相互关联,则二者的关系强度为 1;否则为 0,以此生成 X 和 Y^1 的关系矩阵 $\boldsymbol{M}^{(1)}$.采用相同的方法,利用 $S^{(2)}$ 生成 X 和 Y^2 的关系矩阵 $\boldsymbol{M}^{(2)}$.

5.2.1 聚类效果对比分析

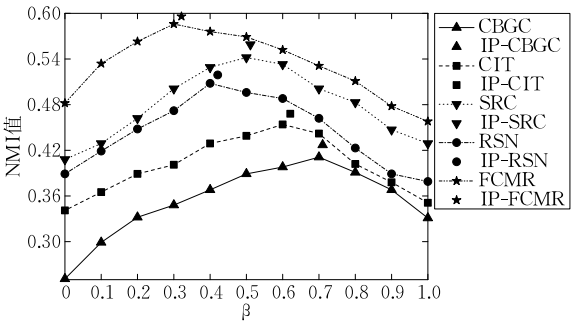
CBGC、IP-CBGC、CIT、IP-CIT、RSN、IP-RSN、SRC、IP-SRC、FCMR 与 IP-FCMR 均需要预先设定聚簇个数,本实验将其设置为数据集真实的类别个数,如表 1 所示.设置 FCMR 与 IP-FCMR 的参数 $T_u=0.01$、$T_v^{(1)}=0.1$ 和 $T_v^{(2)}=0.001$.图 5 所示为基于理想点一致融合的算法与基于加权线性组合一致融合的算法在 β 取不同值时相应的标准互信息值(NMI).首先,结果表明对于基于加权线性组合的算法其权值的选取对聚类结果有较大的影响,例如图 5(a)所示,FCMR 算法聚类结果的 NMI 最多相差 0.28 左右.另外,最优的权值不一定在 $\beta=0.5$ 附近,有时甚至与 0.5 相距较远,如图 5(e)中,CIT 最优权值在 0.8 附近.由此可见基于加权线性组合算法通常选取 $\beta=0.5$ 的方法[1,5,8-9,11]来影响聚类效果.再者,基于理想点一致融合的聚类算法近似于加权线性组合一致融合的算法在 11 个不同的权值中最优的聚类结果.如图 5(a),在 I1 数据集上,IP-RSN、IP-FCMR 和 IP-CIT 算法聚类效果分别略优于权值最优时 RSN、FCMR 和 CIT 算法,IP-CBGC 和 IP-SRC 算法聚类效果分别接近于权值最优时 CBGC



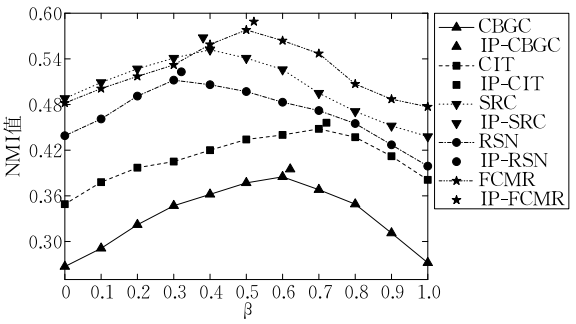
(a) I1数据集



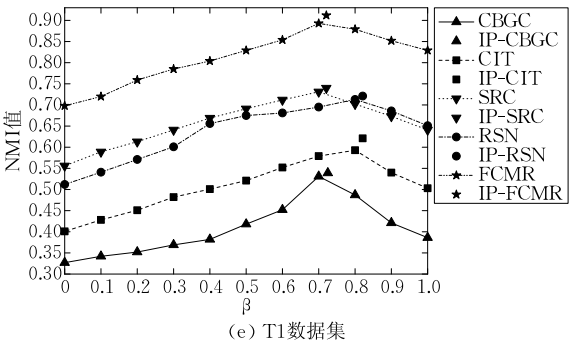
(b) I2数据集



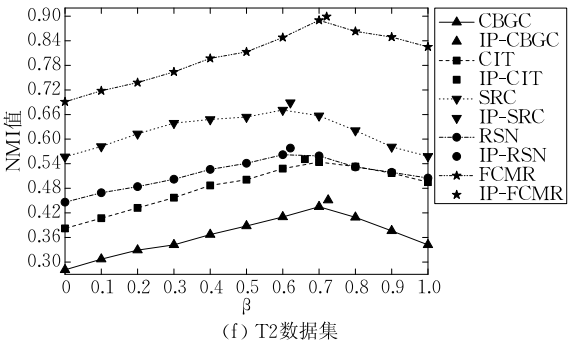
(c) P1数据集



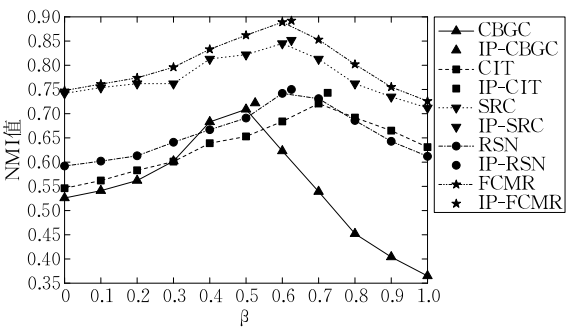
(d)P2数据集



(e) T1数据集



(f) T2数据集



(g) S模拟数据集

图 5　基于理想点与不同 β 值下基于线性组合一致融合算法聚类结果 NMI 值比较

和 SRC 算法. 最后，与已有的自动确定权值的方法 AD-HOCC[10] 相比，图 6 所示结果表明 AD-HOCC 与 IP-CIT 算法聚类效果近似，但 AD-HOCC 仅仅适用于基于信息论的方法，而本文所提出的基于信息论的方法适用于多种高阶联合聚类算法，具有较好的普适性. 由此可见，基于理想点的一致融合方法能够寻找到较好的妥协结果，具有较好的普适性，一定程度上解决了基于线性组合一致融合方法中普遍存在的权值无法确定的问题.

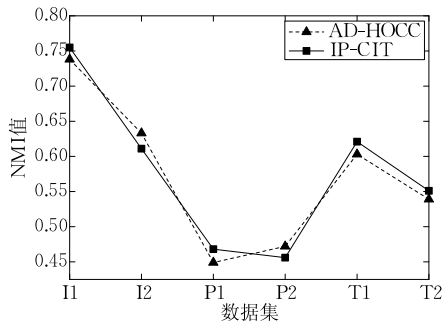
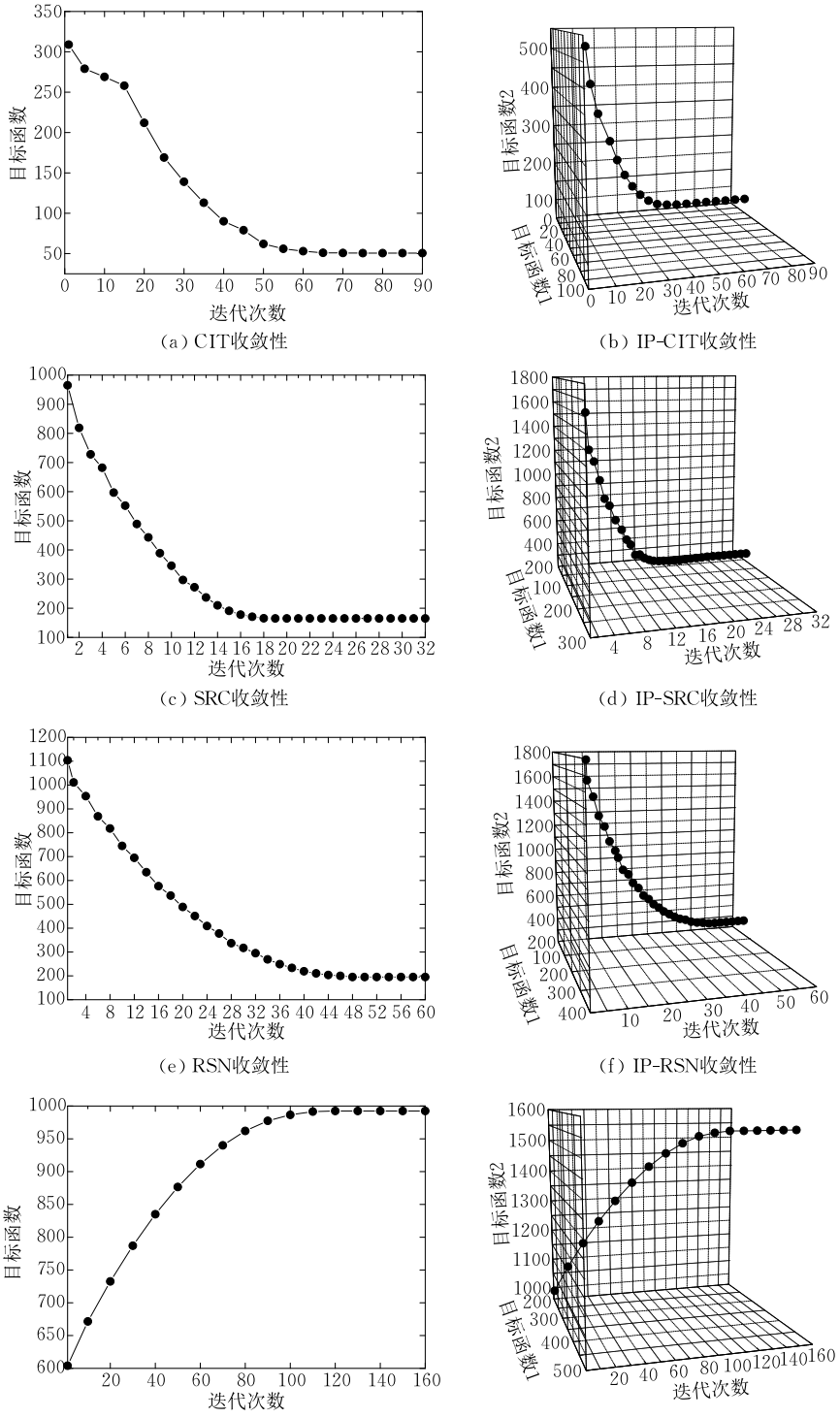


图 6　IP-CIT 与 AD-HOCC 算法聚类结果 NMI 值比较

5.2.2 收敛性与运行时间对比分析

图 7 所示为基于线性组合与基于理想点的聚类算法在数据集 P1 上目标函数随迭代次数的变化情况. 首先,图 7 结果表明基于理想点的聚类算法随着迭代的进行最终达到收敛状态;其次,基于理想点的聚类算法与基于线性组合的方法达到收敛状态时的迭代次数近似.

图 8 所示为于线性组合与基于理想点的聚类算法运行时间比较,结果表明基于理想点的聚类算法运行时间略高于线性组合的方法,大约高出 5%～10%. 另外,IP-CIT 算法与 AD-HOCC 算法运行时间近似.



(a) CIT收敛性 (b) IP-CIT收敛性

(c) SRC收敛性 (d) IP-SRC收敛性

(e) RSN收敛性 (f) IP-RSN收敛性

(g) FCMR收敛性 (h) IP-FCMR收敛性

图 7 基于理想点与基于线性组合一致融合算法收敛性分析

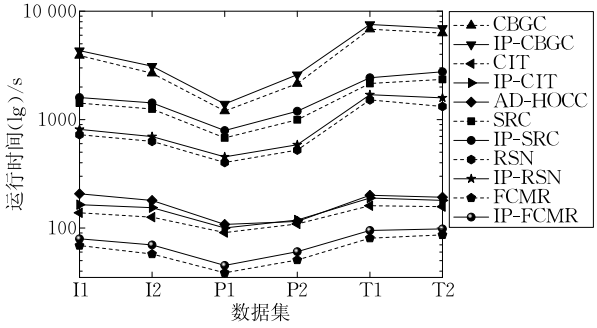


图 8　基于理想点与基于线性组合一致融合算法聚类运行时间比较

6　结　论

本文首先将高阶聚类问题看作是多对二阶聚类的一致融合问题，指出目前基于加权线性组合一致融合方法中存在的权值无法自动确定问题.针对此问题，本文提出了一种基于理想点的一致融合策略.该策略具有很好的普适性，适用于目前多种高阶聚类算法.实验结果表明基于理想点一致融合的算法聚类结果优于基于加权线性组合一致融合的算法权值最优时的聚类结果，说明基于理想点的一致融合方法能够寻找到较好的妥协聚类结果，有效地解决了目前基于加权线性组合一致融合方法中普遍存在的权值无法自动确定的问题.

参　考　文　献

[1] Long Bo, Wu Xiao-Yun, Zhang Zhong-Fei, et al. Unsupervised learning on k-partite graphs//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 317-326

[2] Zhou Zhi-Hua, Wang Jue. Machine Learning and Application 2007. Beijing: Tsinghua University Press, 2007(in Chinese)
(周志华，王珏. 机器学习及其应用 2007. 北京：清华大学出版社，2007)

[3] Chen Yan-Hua, Wang Li-Jun. Non-negative matrix factorization for semisupervised heterogeneous data coclustering. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1459-1474

[4] Wang Hua, Nie Fei-Ping, Huang Heng, et al. Nonnegative matrix tri-factorization based high-order co-clustering and its fast implementation//Proceedings of the 11th IEEE International Conference on Data Mining. Arlington, USA, 2011: 174-183

[5] Gao Bin, Liu Tie-Yan, Zheng Xin, et al. Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering//Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, USA, 2005: 41-50

[6] Dhillon I S. Co-clustering documents and words using bipartite spectral graph partitioning//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2001: 269-274

[7] Dhillon I S, Mallela S, Modha D S. Information-theoretic co-clustering//Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington, USA, 2003: 89-98

[8] Long Bo, Zhang Zhong-Fei, Wu Xiao-Yun, et al. Spectral clustering for multi-type relational data//Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 584-592

[9] Liu Tie-Yan, Ma Wei-Ying. Star-structured high-order heterogeneous data co-clustering based on consistent Information Theory//Proceedings of the 6th IEEE International Conference on Data Mining. Hong Kong, China, 2006: 880-884

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

[11] Mei Jian-Ping, Chen Li-Hui. A fuzzy approach for multitype relational data clustering. IEEE Transactions on Fuzzy Systems, 2012, 20(2): 358-371

[12] Gao Bin, Liu Tie-Yan, Feng Guang, et al. Hierarchical taxonomy preparation for text categorization using consistent bipartite spectral graph co-partitioning. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(9): 1263-1273

[13] Yu Po-Lung, Lee Yoon-Ro, Stam A. Multiple Criteria Decision Making: Concepts, Techniques and Extensions. New York: Plenum Press, 1985

[14] Feng Ying-Jun, Zhang Jie. Theory and Application of Large Scale Multiobjective Programming. Beijing: Science Press, 2004(in Chinese)
(冯英浚，张杰. 大系统多目标规划的理论及应用. 北京：科学出版社，2004)

[15] Long Bo, Zhang Zhong-Fei, Yu P S. Co-clustering by block value decomposition//Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. Chicago, USA, 2005: 635-640

[16] Zha Hong-Yuan, He Xiao-Feng, Ding C, et al. Bipartite graph partitioning and data clustering//Proceedings of the 10th International Conference on Information and Knowledge Management. Atlanta, USA, 2001: 25-32

[17] Tjhi William-Chandra, Chen Li-Hui. Dual fuzzy-possibilistic coclustering for categorization of documents. IEEE Transactions on Fuzzy Systems, 2009, 17(3): 532-543

[18] Strehl A, Ghosh J. Cluster ensembles—A knowledge reuse famework for combining multiple partitions. Journal of Machine Learning Research, 2002, 3(3): 583-617

[19] Ienco D, Robardet C, Pensa R G, et al. Parameter-less co-clustering for star-structured heterogeneous data. Data Mining and Knowledge Discovery, 2013, 26(2): 217-254

[20] Sun Yi-Zhou, Yu Yin-Tao, Han Jia-Wei. Ranking-based clustering of heterogeneous information networks with star network schema//Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Paris, France, 2009: 797-805

HUANG Shao-Bin, born in 1965, Ph. D., professor, Ph. D. supervisor. His main research interests include data mining and model checking.

YANG Xin-Xin, born in 1987, Ph. D. candidate. His main research interests include data mining, social network and complex network.

LV Tian-Yang, born in 1979, Ph. D. associate professor. His main research interests include data mining, social network and complex network.

ZHENG Wei-Min, born in 1946, professor, Ph. D. supervisor. His main research interests include high performance storage system and biology computing.

Background

The project is supported by the National Natural Science Foundation of China under Grant Nos. 71272216, 60903080, 60093009); the National Key Technology R&D Program under Grant Nos. 2009BAH42B02, 2012BAH08B02; the Science Foundation for Post Doctorate Research under Grant No. 2012M510480; the Fundamental Research Funds for the Central University under Grant Nos. HEUCFZ1212, HEUCFT1208. The goal of projects is to improve effect of retrieval of 3D Models retrieval by using multiple kinds of information. 3D model retrieval system involves multiple kinds of information, such as text annotation, feature of content, user feedback, category information. Traditional homogeneous clustering algorithms which cluster each type of objects independently not work well. First, various data types in high-order heterogeneous dataset are interrelated to each other. Clustering each data type independently will lose this interaction information. Second, in clustering a high-dimensional data set, usually not all features are important in any particular. Simultaneously clustering objects and features, we can obtain feature clusters, which play the function of dimensions reduction. This in turn can increase the accuracy of the algorithm. Last but not least, through simultaneous clustering, one can discover the hidden global structures in the heterogeneous data, which seamlessly integrates multiple data types to provide a better picture of the underlying data distribution. As a result, recent research has advanced swiftly from simple clustering of one type of data to simultaneous clustering of multiple types of data, for both two data types (pairwise co-clustering) and multiple (more than two) data types (high-order co-clustering). Many high-order co-clustering algorithms are proposed on well-known academic conferences and journals, such as consistent bipartite graph co-partitioning, information-theoretic high-order co-clustering and non-negative matrix factorization method. Abstracting the computation methods, we note that the problem of high-order co-clustering is converted in to the problem of consistent ensemble of multiple pair two-order co-clustering, by optimizing a weighted combination of objective functions of each pair of two order co-clustering, and then obtain clustering results by iterative method. However, existing algorithms set the weights according to artificial expert expertise. So far how to automatically determine the optimal weights is a classic problem. In this paper, a method of consistent ensemble which can automatically determine the weights is developed. Because the method based on ideal point can solve the problem of consistent ensemble of multiple algorithms of high-order co-clustering, it is a general method. Experimental results show that the method can improve the clustering effect of five algorithms of high-order co-clustering.