

一种基于二部图谱划分的聚类集成方法

徐 森^{1†}, 皋 军^{1,2}, 徐秀芳¹, 花小鹏¹, 徐 静¹, 安 晶¹

(1. 盐城工学院 信息工程学院, 江苏 盐城 224001;

2. 江苏省媒体设计与软件技术重点实验室(江南大学), 江苏 无锡 214122)

摘 要: 将二部图模型引入聚类集成问题中, 使用二部图模型同时建模对象集和超边集, 充分挖掘潜藏在对象之间的相似度信息和超边提供的属性信息. 设计正则化谱聚类算法解决二部图划分问题, 在低维嵌入空间运行 K -means++ 算法划分对象集, 获得最终的聚类结果. 在多组基准数据集上进行实验, 实验结果表明所提出方法不仅能获得优越的结果, 而且具有较高的运行效率.

关键词: 机器学习; 聚类分析; 二部图模型; 聚类集成; 谱聚类算法

中图分类号: TP391

文献标志码: A

A cluster ensemble approach based on bipartite spectral graph partitioning

XU Sen^{1†}, GAO Jun^{1,2}, XU Xiu-fang¹, HUA Xiao-peng¹, XU Jing¹, AN Jing¹

(1. School of Information Engineering, Yancheng Institute of Technology, Yancheng 224001, China; 2. Jiangsu Key Laboratory of Media Design and Software Technology(Jiangnan University), Wuxi 214122, China)

Abstract: A bipartite graph model is brought into the cluster ensemble problem. The object set and hyperedge set are modeled simultaneously via a bipartite graph formulation considering the similarity among instances and the information provided by hyperedges collectively. A normalized spectral clustering algorithm is proposed to solve the bipartite graph partitioning problem, and the final clustering result is attained by performing K -means++ algorithm to partition object set embedded in low dimensional space. Experimental results on several baseline datasets show that the proposed approach is not only well-performed but also high efficient.

Keywords: machine learning; cluster analysis; bipartite graph model; cluster ensemble; spectral clustering algorithm

0 引 言

聚类集成具有传统聚类算法无可比拟的优势, 其关键在于如何将聚类成员组合为更加优越的聚类结果^[1]. 解决聚类集成问题最常见的方法是引入超图的邻接矩阵 H 表示对象之间的两两关系, 避免簇标签对应问题. 根据处理的矩阵可以分为对 H 进行处理的方法^[1-3] 和对相似度矩阵 S 进行处理的方法^[1,4-10]. 对 H 进行处理的方法虽然利用了簇标签提供的属性信息, 但忽略了对象之间的关系; 对 S 进行处理的方法揭示了对象之间的关系, 但未能有效利用簇标签提供的属性信息.

本文引入二部图模型^[11] 解决聚类集成问题, 使

用二部图同时建模对象集和超边集, 充分挖掘潜藏在对象之间的相似度信息和超边/簇提供的属性信息. 通过正则化谱聚类算法解决二部图划分问题, 得到对象的低维嵌入, 在低维空间使用 K -means++ (KM++)^[12] 算法划分对象集, 从而获得最终的聚类结果. 在多组基准数据集上的实验结果验证了所提出算法的有效性.

1 聚类集成相关研究

聚类集成主要研究以下两个问题: 1) 聚类成员生成问题, 即如何生成具有差异性的聚类成员; 2) 共识函数设计问题/聚类集成问题, 即如何将聚类成员

收稿日期: 2017-07-27; 修回日期: 2017-12-26.

基金项目: 国家自然科学基金项目(61375001); 江苏省自然科学基金项目(BK20151299); 江苏省333工程项目; 江苏省高等学校自然科学研究项目(18KJB520050); 江苏省媒体设计与软件技术重点实验室开放课题项目(18ST0201).

责任编委: 陈虹.

作者简介: 徐森(1983—), 男, 副教授, 博士, 从事机器学习及其应用等研究; 皋军(1971—), 男, 教授, 博士, 从事机器学习及其应用等研究.

[†]通讯作者. E-mail: xusen@ycit.cn

集成/组合为更加优越的聚类结果.

由于采用随机初始化的 K 均值算法 (K -means, KM) 每次运行会得到不同的聚类结果,可多次运行 KM 获得聚类成员. 该方法因为简单高效而成为产生聚类成员最常见的方法^[1,10].

组对法是解决共识函数设计问题的主要方法,其关键思想在于引入超图的邻接矩阵 H 表示对象之间的两两关系,有效避免簇标签对应问题. 由处理的矩阵可分为: 1) 对 H 进行处理,包括 H GPA(Hypergraph partitioning algorithm)^[1]、MCLA(Meta-clustering algorithm)^[1]、基于矩阵低秩近似的算法^[2]、基于 KM 的算法^[3] 等; 2) 对相似度矩阵 S 进行处理,包括 CSPA(Cluster-based similarity partitioning algorithm)^[1]、基于证据积累(Evidence accumulation, EA)的层次聚类方法^[4-5]、谱聚类方法^[6]、链接法^[7]、基于近邻传播的方法^[8]、基于密度峰值的方法^[9]、加权共现矩阵方法^[10] 等.

2 本文方法

2.1 基于二部图模型的聚类集成方法

设数据集 $D = \{d_1, d_2, \dots, d_n\}$, 由 r 个聚类成员构成的集合 $P = \{P^{(1)}, \dots, P^{(r)}\}$, 超图的邻接矩阵 $H = H^{(1,2,\dots,r)} = (H^{(1)}, \dots, H^{(r)})$. 假设每个聚类成员均包含 k 个簇, 则超图包含 n 个顶点, $t = rk$ 条超边(每条超边对应于一个簇), 即 H 的大小为 $n \times t$.

文献[11]引入二部图模型同时聚类文本和词,其关键思想在于词和文本聚类的双重性(duality),即词簇能导出文本簇,文本簇能导出词簇. 本文引入该思想,同时聚类对象和超边,原因在于对象和超边聚类也满足双重性,即对象簇能够导出超边簇,超边簇也能够导出对象簇. 下面举例说明本文方法的关键思想,最终结果也验证了对象和超边聚类满足双重性.

例1 假设对象集 $D = \{d_1, d_2, d_3, d_4, d_5\}$, 3个聚类成员(cluster member, CM)提供的聚类标签构成集合 $P = \{P^{(1)}, P^{(2)}, P^{(3)}\}$. 其中: CM_1 提供的标签为 $P^{(1)} = \{1, 1, 1, 2, 2\}^T$, 包含2个簇 $C_1 = \{d_1, d_2, d_3\}$ 和 $C_2 = \{d_4, d_5\}$, 对应的超边分别为 $h_1 = (1, 1, 1, 0, 0)^T$ 和 $h_2 = (0, 0, 0, 1, 1)^T$; CM_2 提供的标签为 $P^{(2)} = \{2, 2, 2, 1, 1\}^T$, 包含2个簇 $C_3 = \{d_1, d_2, d_3\}$ 和 $C_4 = \{d_4, d_5\}$, 对应的超边分别为 $h_3 = (1, 1, 1, 0, 0)^T$ 和 $h_4 = (0, 0, 0, 1, 1)^T$; CM_3 提供的标签为 $P^{(3)} = \{1, 1, 2, 2, 2\}^T$, 包含2个簇 $C_5 = \{d_1, d_2\}$ 和 $C_6 = \{d_3, d_4, d_5\}$, 对应的超边分别为 $h_5 = (1, 1, 0, 0, 0)^T$ 和 $h_6 = (0, 0, 1, 1, 1)^T$. 3个聚类成员对应的超图的邻接矩阵分别为

$$H^{(1)} = H^{(2)} = [h_1, h_2] = [h_2, h_1], H^{(3)} = [h_5, h_6].$$

图1给出了基于二部图模型的聚类集成方法. 图中圆点表示对象, 包含多个圆点的虚线表示簇/超边. 图1(a)~图1(c)分别为3个聚类成员提供的聚类结果, 图1(d)为对象集(对象用圆点表示)和簇集(簇用菱形表示)所构建的二部图模型. 对象与对象之间没有边, 簇与簇之间没有边, 只有当某个簇包含某个对象时, 对应的簇与对象之间有边.

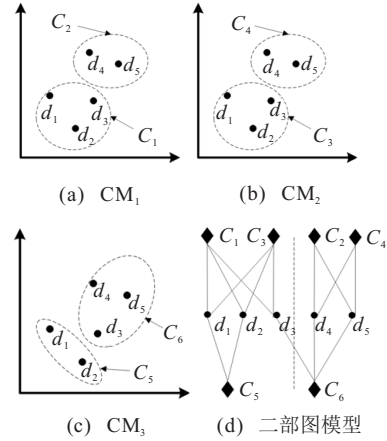


图1 基于二部图模型的聚类集成方法

图1(d)中的虚线表示对二部图的一个划分, 同时划分了对象集和簇集. 要获得最终的聚类结果, 有以下两种方案: 1) 直接取顶点集的划分结果 $\{\{d_1, d_2, d_3\}, \{d_4, d_5\}\}$; 2) 参照 MCLA^[1] 间接获得顶点集的划分结果. 具体地, 首先获得对簇集的划分结果 $\{\{C_1, C_3, C_5\}, \{C_2, C_4, C_6\}\}$, 其中 $\{C_1, C_3, C_5\}$ 和 $\{C_2, C_4, C_6\}$ 称为元簇(meta-cluster); 然后将每个元簇中的所有超边断裂并合并为一条元超边(meta-hyperedge), 用一个包含每个对象与该元超边关联程度的向量进行表示, 关联程度等于元簇中包含的所有超边向量的平均值, 元簇 $\{C_1, C_3, C_5\}$ 对应的元超边 $\{d_1, d_2, d_3\}$, 其关联向量为 $(1, 1, 2/3, 0, 0)^T$, 元簇 $\{C_2, C_4, C_6\}$ 对应的元超边 $\{d_3, d_4, d_5\}$, 其关联向量为 $(0, 0, 1/3, 1, 1)^T$; 最后每个元簇根据关联向量竞争每个对象, 每个对象被分到具有最高关联向量值的元簇中. 例如, 在竞争对象 d_3 时, 元簇 $\{C_1, C_3, C_5\}$ 以 $2/3 > 1/3$ 的优势赢了元簇 $\{C_2, C_4, C_6\}$, 因此, 最终得到的两个对象簇为 $\{d_1, d_2, d_3\}$ 和 $\{d_4, d_5\}$, 该结果与采用第1种方案得到的结果一致.

2.2 二部图谱划分算法设计

根据二部图模型, 同时建模对象集 D 和超边集 W , 得到拉普拉斯矩阵 $L = \begin{bmatrix} D_W & -H \\ -H^T & D_D \end{bmatrix}$, 其中 D_W 和 D_D 为对角矩阵, 对角元素分别为

$$D_W(i, i) = \sum_{j=1}^t H_{ij}, D_D(j, j) = \sum_{i=1}^n H_{ij}.$$

下面将正则化拉普拉斯矩阵 $D^{-1}L$ 的特征值分解 (Eigenvalue decomposition, EVD) 问题转换为规模较小的矩阵 EVD 问题.

定理1 设

$$H_n = D_W^{-1/2} H D_D^{-1/2}, D = \begin{bmatrix} D_W & 0 \\ 0 & D_D \end{bmatrix},$$

则 $D^{-1}L$ 的前 l 个最小特征向量可通过求解 t 阶方阵 $H_n^T H_n$ 的前 l 个最大特征值及对应的特征向量获得.

证明 考虑广义特征值分解问题 $Lz = \lambda Dz$, 设 $z = \begin{bmatrix} w \\ d \end{bmatrix}$, 则有

$$\begin{bmatrix} D_W & -H \\ -H^T & D_D \end{bmatrix} \begin{bmatrix} w \\ d \end{bmatrix} = \lambda \begin{bmatrix} D_W & 0 \\ 0 & D_D \end{bmatrix} \begin{bmatrix} w \\ d \end{bmatrix},$$

展开得

$$\begin{aligned} D_W w - H d &= \lambda D_W w, \\ H^T w + D_D d &= \lambda D_D d. \end{aligned}$$

因为 D_W, D_D 为非奇异矩阵, 分别将上面两式左乘 $D_W^{-1/2}$ 和 $D_D^{-1/2}$, 得到

$$\begin{aligned} D_W^{1/2} w - D_W^{-1/2} H d &= \lambda D_W^{1/2} w, \\ -D_D^{-1/2} H^T w + D_D^{1/2} d &= \lambda D_D^{1/2} d. \end{aligned}$$

设 $u = D_W^{1/2} w, v = D_D^{1/2} d$, 经过整理得

$$\begin{aligned} D_W^{-1/2} H D_D^{-1/2} v &= (1 - \lambda) u, \\ D_D^{-1/2} H^T D_W^{-1/2} u &= (1 - \lambda) v. \end{aligned}$$

设 $\sigma = 1 - \lambda$, 则 $H_n v = \sigma u, H_n^T u = \sigma v$. 可见, u, v 分别为 H_n 的左、右奇异向量, σ 为相应的奇异值. 因此, $D^{-1}L$ 的前 l 个最小特征向量可以通过求解 H_n 的前 l 个最大奇异值对应的左、右奇异值向量获得.

设 $\text{rank}(H_n) = p \leq n$, 其奇异值分解 (Singular value decomposition, SVD) 定义为 $H_n = U \Sigma V^T$. 其中: U, V 分别为 n 阶和 t 阶酉矩阵, $U^T U = I_n, V^T V = I_t$; $\Sigma \in R^{n \times t}$ 为对角阵, 且对角元素 σ_i 为 H_n 的奇异值, $1 \leq i \leq n$, 当 $1 \leq i \leq p$ 时, $\sigma_i > 0$, 当 $i \geq p + 1$ 时, $\sigma_i = 0$. 易得 $H_n^T = V \Sigma U^T, H_n^T H_n = V \Sigma^2 V^T$. 可见, H_n 的奇异值等于 $H_n^T H_n$ 的特征值的非负平方根, 对应的右奇异向量等于 $H_n^T H_n$ 的非零特征值对应的特征向量. 另外, 将 H_n 的 SVD 两边同时右乘 $V \Sigma^\dagger$, 其中 $\Sigma^\dagger$ 为 Σ 的广义逆, 得到 $U = H_n V \Sigma^\dagger$. 设 $H_n^T H_n$ 的特征值及对应的特征向量为 δ_i 和 q_i , 有

$$\sigma_i = \delta_i^{1/2}, v_i = q_i, u_i = H_n q_i \delta_i^{-1/2}.$$

因此, H_n 的前 l 个最大奇异值对应的左、右奇异向量可通过求解 t 阶方阵 $H_n^T H_n$ 的前 l 个最大特征值 δ_i 对应的特征向量 q_i 得到.

设 $\Lambda_l = \text{diag}(\delta_1^{-1/2} \cdots \delta_l^{-1/2}), Q_l = [q_1, q_2, \cdots, q_l]$, $D^{-1}L$ 的前 l 个最大特征向量构成的矩阵为

$$Z = \begin{bmatrix} D_W^{-1/2} U_l \\ D_D^{-1/2} V_l \end{bmatrix} = \begin{bmatrix} D_W^{-1/2} H_n Q_l \Lambda_l \\ D_D^{-1/2} Q_l \end{bmatrix}.$$

即 $D^{-1}L$ 前 l 个最小特征向量可通过求解 t 阶方阵 $H_n^T H_n$ 前 l 个最大特征值及对应的特征向量获得. $\square$

在获得对象与超边的低维表示 Z 后, 考虑到 Z 的第 $1 \sim n$ 行对应于对象集, 第 $n+1 \sim n+t$ 行对应于超边集, 因此可以采用 KM 算法对 Z 的第 $1 \sim n$ 行聚类, 直接获得最终的聚类结果. KM 采用 K 个均值向量表示数据集, 其目标函数为最小化误差平方和 (Sum of squared errors, SSE), 有

$$J = \sum_{k=1}^K \sum_{z_i \in C_k} \|z_i - m_k\|^2 = C - \sum_k \frac{1}{n_k} \sum_{z_i, z_j \in C_k} z_i^T z_j.$$

其中: $m_k = \sum_{i \in C_k} z_i / n_k$ 为簇 C_k 的质心向量, n_k 为簇 C_k 包含的对象个数, $C = \sum_i \|z_i\|^2$ 为常量.

不失一般性, 假设对象按照其所在的簇进行排序, 上述最小化 SSE 问题的解可以表示为 K 个指示向量构成的矩阵

$$F = (f_1, f_2, \cdots, f_K).$$

其中

$$f_k^T f_l = \delta_{kl}.$$

$$\delta_{kl} = \begin{cases} 1, & k = l; \\ 0, & \text{Otherwise.} \end{cases}$$

$$f_k = (0, \cdots, 0, \underbrace{1, \cdots, 1}_{n_k}, 0, \cdots, 0)^T / n_k^{1/2}.$$

此时上述最小化问题变为 $J = \text{Tr}(Z^T Z) - \text{Tr}(F^T Z^T Z F)$, 其中 $\text{Tr}(\cdot)$ 表示矩阵的求迹运算. 设组对相似度矩阵为 $P = Z^T Z$, 上述问题转换为迹最大化问题 $\max_{F^T F = I, F \geq 0} J_P(F) = \text{Tr}(F^T P F)$.

KM 算法简单高效, 但聚类结果受初始聚类中心 (seeds) 影响较大, 易收敛到较差的局部最优解. 为了提高聚类质量, 引入 KM++^[12] 算法, 运行 KM++ 多次, 取最优值作为最终的聚类结果. KM++ 的关键之处在于选择远离的数据点作为 seeds, 克服了 KM 算法随机选择 seeds 的盲目性. 具体地, 首先随机选择 1 个数据点作为第 1 个 seed, 然后逐步选择新的数据点作为新的 seed, 选择新 seed 的原则是, 离已有 seeds 越远的点被选为新 seed 的概率越大, 重复上述过程, 直到选出 K 个 seeds.

将本文的聚类集成方法称为二部图谱划分

(Bipartite spectral graph partitioning, BSGP), 算法主要步骤描述如下.

Input: 对象集 $D = \{d_1, d_2, \dots, d_n\} \in R^{n \times m}$, n 为对象个数, m 为特征个数, 簇个数为 K , 聚类成员个数为 r .

Step 1: 运行随机选取初始质心的 KM 算法 r 次, 每次得到 K 个簇, 生成 r 个聚类成员.

Step 2: 构建超图的邻接矩阵 H , 计算 H_n , 求解 t 阶方阵 $H_n^T H_n$ 的 EVD, 求解 Z .

Step 3: 设 $z_i \in R^l$ 为对应于 Z 的第 i 行 ($1 \leq i \leq n$) 的列向量, 使用 KM++ 算法将 $Z = \{z_i | i = 1, 2, \dots, n\}$ 划分为 K 个簇 $C_1, C_2, \dots, C_K$.

Output: 聚类结果 $\pi = \{D_1, D_2, \dots, D_K\}$, 其中 $D_k = \{d_i | z_i \in C_k\}, k = 1, 2, \dots, K$.

Step 1 时间复杂度为 $O(tmn)$, Step 2 时间复杂度为 $O(t^2n^2 + t^3)$, 其中 t 阶方阵的求解可通过高效的矩阵乘法获得, Step 3 时间复杂度为 $O(Kln)$. 通常 $t \ll n$, 因此, 本文方法的计算复杂度较低. 在下文的实验部分进一步证实了本文算法具有较高的运行效率.

3 实验分析

实验平台为: Intel Xeon E5-1650 六核处理器, 频率 3.50 GHz, 内存 16.00 GB, 程序在 Matlab2016b 下运行, 操作系统为 Windows 10 专业版.

3.1 实验数据集与评价指标

实验使用由 TREC(<http://trec.nist.gov>) 提供的 10 组基准文本数据集, 见表 1. 数据集 tr11、tr23、tr41 和 tr45 中的文本类别对应于某个特殊类别的查询. 数据集 la1、la2 和 la12 由洛杉矶时报上的文章构成, 包含 6 类文章. 数据集 hitech、reviews 和 sports 取自 San Jose Mercury 报纸, hitech 包含关于计算机、电子、健康、医疗、研究和科技方面的文章, reviews 包含关于食物、电影、音乐、广播和饭店方面的文章, sports 包含关于棒球、篮球、自行车、拳击、足球、高尔夫、曲棍球的文

章. 所有数据集中的文本类别标签唯一.

本文采用机器学习领域比较流行的规范化互信息 (Normalized mutual information, NMI)^[1] 衡量聚类结果和已知类别标签的匹配程度. 当两个类别标签一一对应时, NMI 值达到最大值 1. 另外, 本文采用在信息检索领域常用的评价文本聚类质量的综合指标 F 值 (F -measure). F 值越大, 聚类质量越高, 反之聚类质量越低.

3.2 实验结果与分析

将本文的 BSGP 与主流聚类集成方法进行对比, 对比算法为文献 [1] 提出的 CSPA、HGPA、MCLA 和文献 [5] 提出的基于 EA 思想的 4 个层次聚类算法 (单连接 (Single linkage, SL)、全连接 (Complete linkage, CL)、组平均 (Average linkage, AL) 和 Ward, 分别记为 EASL、EACL、EAAL、EAWL).

本文首先对经过预处理的文本数据集进行 TF-IDF (Term frequency-inverse document frequency) 加权, 然后运行使用余弦相似度的 KM 算法 100 次, 每次生成 K 个簇 (K 等于真实类别个数), 生成 100 个聚类成员, 运行不同的聚类集成方法. 需要指出的是, 对于其他领域的数据集, 需要采用不同的预处理方式和相似度函数才能获得有效的聚类成员, 并进行聚类集成. CSPA 和 MCLA 调用了图划分算法 METIS, 不平衡因子 UB 取默认值 0.05, 得到稳定的聚类结果. HGPA 调用了 HMETIS 算法, 不平衡因子 UB 取默认值 0.05, 因为 HMETIS 得到局部最优解, 所以 HGPA 获得的聚类结果不够稳定, 本文运行 10 次取最佳值. 凝聚层次聚类方法递归地合并对象, 将数据集划分为嵌套的层次结构, 获得了稳定的聚类结果. BSGP 算法运行 KM++ 算法 10 次取最优结果.

图 2 和图 3 分别为不同聚类集成算法获得的 NMI 值和 F 值, 柱状图依次为 CSPA、HGPA、MCLA、EASL、EACL、EAAL、EAWL、BGSP. 由图 2 和图 3 可见: 1) BSGP 在 tr23、tr41、la12、reviews 和 sports 上获得了最高的 NMI 值, 在 tr41、la12、reviews 和 sports 上获得了最高的 F 值. 2) BSGP 在 10 组数据集上获得的平均 NMI 值和 F 值高于其他算法, 总体来看, BSGP 在多数情况下获得了最佳聚类效果, 且平均聚类质量最高. 3) 实验采用了两种常用的评价指标, 结果显示获得高 NMI 值的算法通常也可以获得高 F 值, 反之亦然, 但是也有例外情况发生, 例如 BSGP 在 tr23 上获得了最高的 NMI 值, 而 CSPA 却获得了最高的 F 值. 该现象反映了聚类结果评价的困难性, 而其根本原因在于聚类分析是不适定问题.

表 1 实验数据集描述

Dataset	文本个数	特征词个数	类别个数
tr11	414	6 429	9
tr23	204	5 832	6
tr41	878	7 454	10
tr45	690	8 261	10
la1	3 204	31 472	6
la2	307	31 472	6
la12	6 279	31 472	6
hitech	2 301	10 080	6
reviews	4 069	18 483	5
sports	8 580	14 870	7

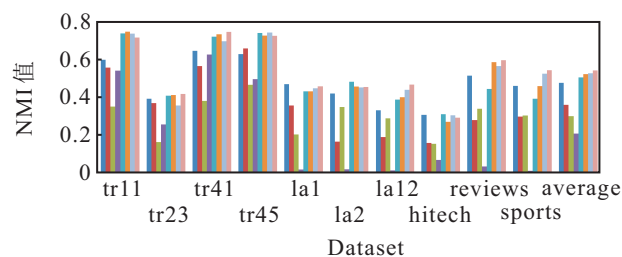


图2 不同聚类集成算法获得的NMI值

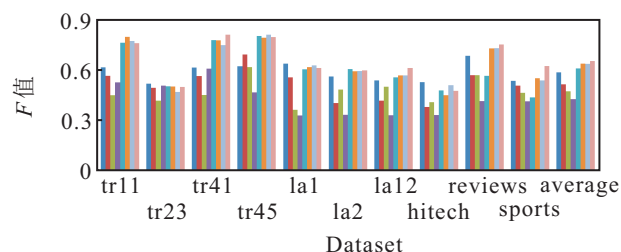


图3 不同聚类集成算法获得的F值

在面向海量数据挖掘时,聚类集成算法运行效率是衡量算法优劣的重要指标,考虑到HGPA、MCLA和EASL聚类质量较差,本文仅比较CSPA、EACL、EAAL、EAWL和BGSP的运行时间.将数据集依据其大小(显示在数据集下方)递增排列,5个聚类集成算法在聚类集成阶段的运行时间如图4所示.由图4可见,不同聚类集成算法的运行时间满足 $BGSP < CSPA < EACL \approx EAAL < EAWL$,即BGSP效率最高,CSPA次之,层次聚类算法效率最低.特别地,在规模最大的数据集sports上(大小为8 580),BGSP的运行时间不超过10s,而层次聚类算法的运行时间则超过130s.显然,当进行海量数据挖掘时,层次聚类算法高昂的计算成本让人难以接受,而BGSP较高的运行效率则使其易于扩展到大规模应用领域.

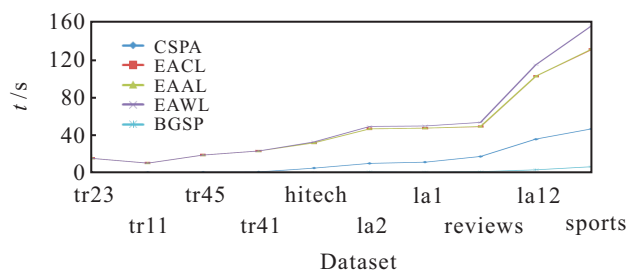


图4 不同聚类集成算法在集成阶段的运行时间

4 结论

本文将二部图模型引入聚类集成问题中,使用二部图模型同时建模对象集和超边集,并设计正则化谱聚类算法解决二部图划分问题,提出了BGSP算法.在10组基准数据集上进行的实验结果表明,与主流的聚类集成算法相比,BGSP不仅能获得更加优越的结果,而且具有较高的运行效率.

参考文献(References)

- [1] Strehl A, Ghosh J. Cluster ensembles: A knowledge reuse framework for combining multiple partitions[J]. J of Machine Learning Research, 2002, 3(3): 583-617.
- [2] 徐森, 周天, 于化龙, 等. 一种基于矩阵低秩近似的聚类集成算法[J]. 电子学报, 2013, 41(6): 1219-1224. (Xu S, Zhou T, Yu H L, et al. Matrix low rank approximation-based cluster ensemble algorithm[J]. Acta Electronica Sinica, 2013, 41(6): 1219-1224.)
- [3] Wu J, Liu H, Xiong H, et al. K-means based consensus clustering: A unified view[J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27(1): 155-169.
- [4] Fred A L, Jain A K. Combining multiple clusterings using evidence accumulation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850.
- [5] Fred A, Lourenço A. Cluster ensemble methods: From single clusterings to combined solutions[C]. Supervised and Unsupervised Ensemble Methods and their Applications. Berlin: Springer Heidelberg, 2008: 3-30.
- [6] 黄发良, 黄名选, 元昌安, 等. 网络重叠社区发现的谱聚类集成算法[J]. 控制与决策, 2014, 29(4): 713-718. (Huang F L, Huang M X, Yuan C A, et al. Spectral clustering ensemble algorithm for discovering overlapping communities in social networks[J]. Control and Decision, 2014, 29(4): 713-718.)
- [7] Iam-On N, Boongeon T, Garrett S, et al. A link-based cluster ensemble approach for categorical data clustering[J]. IEEE Trans on Knowledge and Data Engineering, 2012, 24(3): 413-425.
- [8] Yu Z, Li L, Liu J, et al. Adaptive noise immune cluster ensemble using affinity propagation[J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27(12): 3176-3189.
- [9] 褚睿鸿, 王红军, 杨燕, 等. 基于密度峰值的聚类集成[J]. 自动化学报, 2016, 42(9): 1401-1412. (Chu R H, Wang H J, Yang Y, et al. Clustering ensemble based on density peaks[J]. Acta Automatica Sinica, 2016, 42(9): 1401-1412.)
- [10] Berikov V, Pestunov I. Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties[J]. Pattern Recognition, 2017, 63: 427-436.
- [11] Dhillon I S. Co-clustering documents and words using bipartite spectral graph partitioning[C]. Acm Sigkdd Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2001: 269-274.
- [12] Arthur D, Vassilvitskii S. K-means++: The advantages of careful seeding[C]. Proc of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms. New York: Society for Industrial and Applied Mathematics, 2007: 1027-1035.

(责任编辑: 郑晓蕾)