一种改进的自适应聚类集成选择方法

徐 森 [1] 皋 军 [1, 2] 花小朋 [1] 李先锋 [1] 徐 静 [1]

摘 要 针对自适应聚类集成选择方法 (Adaptive cluster ensemble selection, ACES) 存在聚类集体稳定性判定方法不客观和聚类成员选择方法不够合理的问题, 提出了一种改进的自适应聚类集成选择方法 (Improved ACES, IACES). IACES 依据聚类集体的整体平均归一化互信息值判定聚类集体稳定性, 若稳定则选择具有较高质量和适中差异性的聚类成员, 否则选择质量较高的聚类成员. 在多组基准数据集上的实验结果验证了 IACES 方法的有效性: 1) IACES 能够准确判定聚类集体的稳定性, 而 ACES 会将某些不稳定的聚类集体误判为稳定; 2) 与其他聚类成员选择方法相比, 根据 IACES 选择聚类成员进行集成在绝大部分情况下都获得了更佳的聚类结果, 在所有数据集上都获得了更优的平均聚类结果.

关键词 机器学习, 聚类分析, 聚类集成, 聚类集成选择

引用格式 徐森, 皋军, 花小朋, 李先锋, 徐静. 一种改进的自适应聚类集成选择方法. 自动化学报, 2018, **44**(11): 2103−2112

DOI 10.16383/j.aas.2018.c170376

An Improved Adaptive Cluster Ensemble Selection Approach

XU Sen[1] GAO Jun[1, 2] HUA Xiao-Peng[1] LI Xian-Feng[1] XU Jing[1]

Abstract Adaptive cluster ensemble selection (ACES) is not only non-objective in judging the stability of cluster ensemble but also unreasonable in selecting cluster members. To overcome such drawbacks, an improved adaptive cluster ensemble selection (IACES) approach is proposed. First, IACES judges the stablity of cluster ensemble according to its total average normalized mutual information. Second, if cluster ensemble is stable, then cluster members with high quality and moderate diversity are selected, else, cluster members with high quality are selected. We evaluate the proposed method on several benchmark datasets and the results show that IACES can judge the stability of cluster ensemble correctedly while ACES misjudges some unstable cluster ensemble as stable. Besides, ensembling the cluster members selected by IACES produces better final solutions than other cluster member selection methods in almost all cases, and is superior average results in all cases.

Key words Machine learning, cluster analysis, cluster ensemble, cluster ensemble selection

Citation Xu Sen, Gao Jun, Hua Xiao-Peng, Li Xian-Feng, Xu Jing. An improved adaptive cluster ensemble selection approach. *Acta Automatica Sinica*, 2018, **44**(11): 2103−2112

聚类分析的目标是依据对象之间的相似度对其自动分组/簇, 使得簇内对象彼此相似度尽量高, 而簇间对象彼此相似度尽量低[1−2]. 虽然已有上千种聚类算法, 但没有一种能有效识别不同大小, 不同形状, 不同密度甚至可能包含噪声的簇[1]. 已有聚类方法主要分为: 1) 基于划分的方法, 将聚类问题转化为优化问题, 并采用不同的优化策略, 例如, K 均值算法 (K-means, KM)[3]、非负矩阵分解 (Non-negative matrix factorization, NMF)[4]、近邻传播算法 (Affinity propagation, AP)[5]、子空间聚类算法[6] 以及基于深度学习[7] 的聚类算法[8]; 2) 层次聚类[2], 例如单连接 (Single linkage, SL)、全连接 (Complete linkage, CL)、组平均 (Average linkage, AL); 3) 基于模型的方法[1], 将聚类问题转化为模型的充分统计量估计问题; 4) 基于密度的方法, 通过寻找被低密度区域分离的高密度区域来进行聚类, 例如密度峰值 (Density peaks, DP) 算法[9]、谱聚类方法[10]; 5) 基于谱图理论将聚类问题转化为图划分问题.

收稿日期 2017-03-17 录用日期 2017-11-06
Manuscript received March 17, 2017; accepted November 6, 2017
国家自然科学基金 (61105057, 61375001), 江苏省自然科学基金 (BK 20151299), 江苏省政策引导类计划 (产学研合作)–前瞻性联合研究项目 (BY2016065-01), 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 (江南大学) 开放课题 (18ST0201), 江苏省 "333 工程", 江苏省高校 "青蓝工程" 资助
Supported by National Natural Science Foundation of China (61105057, 61375001), Natural Science Foundation of Jiangsu Province (BK20151299), the Industry-Education-Research Prospective Project of Jiangsu Province (BY2016065-01), the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (18KJB520050), Open Project of Jiangsu Key Laboratory of Media Design and Software Technology (18ST0201), the "333 Project" of Jiangsu Province, and Jiangsu Province Qing Lan Project
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 盐城工学院信息工程学院 盐城 224051 2. 江苏省媒体设计与软件技术重点实验室 (江南大学) 无锡 214122
1. School of Information Engineering, Yancheng Institute of Technology, Yancheng 224051 2. Jiangsu Key Laboratory of Media Design and Software Technology (Jiangnan University), Wuxi 214122

2002 年, 文献 [11] 正式提出聚类集成 (Cluster ensemble), 通过组合多个不同的聚类结果能够获得更加优越的最终结果, 具有传统聚类算法无可比拟的优势[12]. 早期的聚类集成研究主要围绕聚类成员生成和共识函数设计问题展开, 目前已有较多聚类成员生成方法及共识函数设计方法[13−26]. 受 "选择性集成学习" 研究启发, 文献 [27] 于 2008 年开启了选择性聚类集成研究, 其关键问题为聚类成员选择问题, 即如何从所有聚类成员集合 (称为聚类集体) 中选出部分聚类成员用于后续集成, 获得比对所有聚类成员进行集成更加优越的结果. 文献 [27] 提出了质量和多样性综合准则 (Joint criterion, JC)、聚类并选择 (Cluster and select, CAS) 和凸包 (Convex hull, CH) 三种方法. 文献 [28] 提出了自适应聚类集成选择 (Adaptive cluster ensemble selection, ACES), 依据聚类成员与初始一致划分 π^* 的归一化互信息 (Normalized mutual information, NMI) 将聚类集体分为稳定和不稳定两类, 并选择不同的聚类成员子集. ACES 方法存在两个问题: 1) 判定聚类集体稳定性的方法不客观, 稳定性与初始一致划分 π^* 有关, 在某些情况下容易将不稳定的聚类集体误判为稳定. 当聚类成员之间差异性较低时, NMI 值较大, 且 NMI 值大于 0.5 的比例较高, 聚类成员与 π^* 的 NMI 值也较大, 此时聚类集体稳定; 当聚类成员之间差异性较高时, NMI 值较低, 平均 NMI 值低于 0.5, 且 NMI 值大于 0.5 的比例低于 50 %, 但仍然有绝大多数的聚类成员与 π^* 的 NMI 值大于 0.5, 此时虽然聚类集体不稳定, 但 ACES 方法却判定聚类集体是稳定的. 2) 聚类成员子集的选择方法不够合理. 当聚类集体稳定时, ACES 简单地选择 Full 集并输出 π^*, 而没有进一步选择差异性较高的聚类成员来提高集成效果; 当聚类集体不稳定时, ACES 简单地选择 High 集, 可能会选出少量聚类质量较差的聚类成员.

本文针对 ACES 存在的问题, 提出了一种改进的自适应聚类集成选择方法 (Improved adaptive cluster ensemble selection, IACES). 本文把所有聚类成员的整体平均 NMI 值 (Total average NMI, TANMI) 作为判定聚类集体是否稳定的依据, 若 TANMI 大于 0.5, 则聚类集体稳定; 否则, 不稳定. 有效解决了上述第一个问题. 当聚类集体稳定时, 聚类成员提供相似的聚类结构, 差异可能由聚类算法陷入局部最优值引起, π^* 能够在一定程度上减小聚类成员之间差异引起的方差, 可能比聚类成员更加接近于真实分类结果. 与 ACES 选择 Full 集不同, 本文首先选择与 π^* 的差异性最低 (NMI 值最大) 的 1/4 的聚类成员, 降低平均误差; 然后选择与 π^* 的差异性最高 (NMI 值最小) 的 1/4 的聚类成员, 增加平均差异性; 另外, 为了避免选出离群点, 通过约束聚类成员的平均 NMI 值 (Average NMI, ANMI) 排名高于某一阈值. 此时, 选出的聚类成员既具有较高质量, 又具有适中的差异性, 往往能够获得比 π^* 更加优越的结果. 当聚类集体不稳定时, 聚类成员提供了不同的聚类结构, 差异可能由聚类算法本身的偏置或数据集的复杂结构引起, 此时 π^* 偏离真实分类结果的可能性较大, 因此应该尽量降低聚类成员的平均误差, 选择与 π^* 差异性大的聚类成员 (High 集) 往往会得到更好的结果. 但 High 集里会存在一些质量较差的聚类成员, 此时通过约束聚类成员 ANMI 值排名高于某一阈值, 能够在很大程度上避免选出质量较差的聚类成员. 有效解决了上述第二个问题.

1 聚类集成相关研究

本文在文献 [11] 提出的聚类集成框架上, 增加聚类成员选择模块, 形成选择性聚类集成系统框架, 如图 1 所示, 其中聚类成员选择是本文的研究重点. 选择性聚类集成分为三步: 第一步将数据集作为输入, 运行聚类算法, 输出聚类集体 $P = \{P^{(1)}, \cdots, P^{(l)}\}$, 这一步称为聚类成员生成; 第二步将聚类集体作为输入, 输出若干聚类成员构成的集合 $E = \{E^{(1)}, \cdots, E^{(r)}\} \subseteq P$, 这一步称为聚类成员选择; 第三步将 E 作为输入, 对它们进行组合, 输出最终的聚类结果, 这一步称为聚类组合 (Combination)/集成 (Ensemble)/融合 (Fusion), 也称为共识函数 (Consensus function) 设计. 下面对聚类集成相关研究予以阐述.

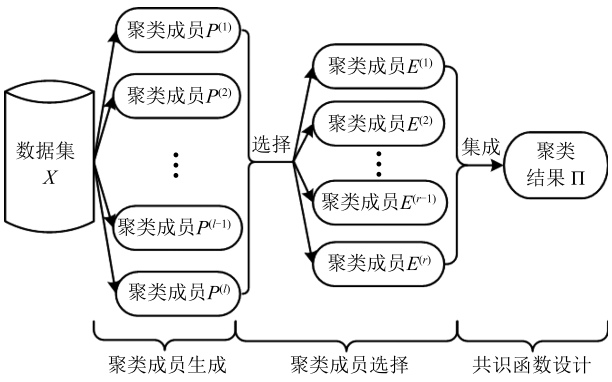


图 1 选择性聚类集成系统框架

Fig. 1 Framework of selective cluster ensemble system

1.1 聚类成员生成

研究人员从聚类模型和数据集等角度入手提出了不同的聚类成员生成方法: 1) 采用同一种聚类算法[11−16, 18−22, 24−30]. 由于采用随机初始化的 KM 每次运行会得到不同的聚类结果, 因此可通过

多次运行来生成聚类成员. 该方法计算复杂度仅为 O($lknd$), 其中 l 为聚类成员的个数, k 为簇个数, n 为数据集大小, d 为特征数, 因此成为产生聚类成员最常见的方法[31]. 2) 对不同的数据子集进行聚类, 例如随机投影、投影到不同的子空间、采用不同的采样技术、选择不同的特征子集等[11, 23]. 3) 采用不同的聚类个数, 例如设置多个不同的 k 值或在指定的区间随机选择 k[17].

1.2 聚类成员选择

文献 [32] 指出当聚类成员多样性较高时能够获得更好的集成效果. 与此不同, 文献 [33] 指出适中的多样性能够获得最佳的集成效果. 文献 [28] 认为不同数据集产生的聚类成员具有不同的特点, 应该区别对待, 提出了 ACES: 1) 采用 AL 对聚类集体 (Full 集) 进行集成, 获得一致划分 π^*; 2) 计算所有聚类成员与 π^* 的 NMI 值, 若平均 NMI 值 (Mean NMI, MNMI) 大于 0.5, 则聚类集体稳定 (Stable, S), 否则, 聚类集体不稳定 (Non-stable, NS); 若聚类集体稳定, 则输出 π^* 为最终的聚类集成结果, 若不稳定, 则选择与 π^* 差异大的一半子集 High (具体地, 将所有聚类成员与 π^* 的 NMI 值按照降序排列, 选择 NMI 值小的一半子集), 采用 AL 对 High 集进行集成得到最终的聚类集成结果.

文献 [27] 提出了 JC, CAS 和 CH 三种聚类成员选择方法, 其中每个聚类成员的质量与该成员和其他聚类成员的归一化互信息之和 (Sum of normalized mutual information, SNMI) 成正比, 而聚类集体的多样性与所有聚类成员与其他成员的 SNMI 之和成反比. JC 首先选择质量最高的聚类成员, 然后逐一选择使得目标函数值最大的聚类成员, 直到选出预设的 K 个聚类成员为止. CAS 使用 NJW 谱算法[34] 将聚类集体划分为 K 个分组, 并从每个分组中选出一个质量最高的聚类成员. CH 首先根据聚类集体绘制质量–多样性图, 然后通过凸包创建该图的简要概括, 其中包括质量最高、多样性最大的聚类成员所对应的点, 最后选出有凸包内的点对应的聚类成员. 该文使用 CSPA (Cluster-based similarity partitioning algorithm)[11] 作为一致性函数, 对 JC, CAS 和 CH 进行了实验比较, 总体来看, CAS 获得了最佳聚类集成结果, 但需要人工设置聚类成员个数.

文献 [29] 提出最有效一致划分 (Best validated consensus partition, BVCP), 采用不同的聚类成员选择方法选出不同子集, 并对每个子集进行集成, 获得多个候选一致划分 (Candidate consensus partition, CCP), 最后使用多个相对有效性指标评价每个 CCP, 得到最佳评价指标的 CCP 即为最终的一致划分. 近期, 文献 [30] 基于证据空间扩展有效性指标 Davies-Bouldin (DB), 利用聚类成员的类别相关矩阵度量差异性, 以较高有效性和较大差异性为目标选择聚类成员.

1.3 共识函数设计

聚类分析过程中, 对象标签是未知的, 因此不同聚类成员得到的簇标签没有显式的对应关系. 另外, 聚类成员可能包含不同的簇个数, 使得簇标签对应问题极具挑战[11]. 根据是否显式解决簇标签对应问题, 聚类集成方法可分为以下两类: 1) 组对法 (Pairwise approach), 引入超图的邻接矩阵 H 将表示对象之间的两两关系, 有效避免了簇标签对应问题. 根据处理的矩阵不同, 可以分为: a) 对 H (或其加权矩阵) 进行处理, 包括 HGPA (Hypergraph partitioning algorithm)[11] 和 MCLA (Meta-clustering algorithm)[11]、基于 NMF 的方法[15]、基于 KM 的方法[19]、结合 KM 与拉普拉斯矩阵的方法[26]、基于矩阵低秩近似的方法[35] 等; b) 对相似度矩阵 S (或 S 的加权矩阵) 进行处理, 包括基于图划分算法的 CSPA[11]、混合模型方法[12]、二部图模型方法[13]、层次聚类法[14]、链接法[17]、加权共现矩阵 (Weighted co-association matrices) 方法[20]、使用多蚁群算法的方法[22]、基于 AP 的方法[24]、基于 DP 的方法[25]、基于谱聚类的方法[36] 等. 组对法因其思想简单而成为解决共识函数设计问题的主要方法. 2) 重新标注法 (Re-labeling approach), 包括累积投票 (Cumulative voting)[16]、PRI (Probabilistic rand index)[18]、选择性投票[21] 等.

2 本文方法

2.1 聚类成员之间的差异性计算

在没有先验知识的情况下, 衡量聚类成员差异性大小的一种思路是依据聚类成员彼此之间的 “相似” 程度: 两个聚类成员越相似, 差异越小, 反之差异越大. 本文采用源自信息论的 NMI 值[11] 来度量聚类成员之间的相似度, NMI 值越大, 两个聚类结果的匹配程度越高, 其相似度越大, 差异性越小. 通过计算聚类成员两两之间的 NMI 值, 即可得到聚类成员之间的相似度矩阵.

2.2 聚类集体稳定性判定方法

假设 l 个聚类成员构成的聚类集体 $P = \{P^{(1)}, \cdots, P^{(l)}\}$, ACES 首先采用 AL 对聚类集体进行集成, 获得一致划分 π^*; 然后计算所有聚类成员与 π^* 的 NMI 值, 若 MNMI 大于 0.5, 则聚类集体稳定, 否则, 不稳定. MNMI 的计算方法如下:

$$\mathrm{MNMI}=\frac{1}{l}\sum_{i=1}^{l}\mathrm{NMI}(P^{(i)},\pi^{*}) \tag{1}$$

与 ACES 不同, 本文根据所有聚类成员之间的相似程度判定聚类集体的稳定性, 当所有聚类成员的 TANMI 大于 0.5 (或 NMI 值大于 0.5 的比例 Proportion 较高, 例如 Proportion $\geq$ 50 %) 时, 说明聚类集体的整体相似度较高, 差异性较低, 聚类集体稳定; 否则, 不稳定. TANMI 的计算方法如下:

$$\mathrm{TANMI}=\frac{2}{l(l-1)}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l}S_{ij} \tag{2}$$

其中, S_{ij} 表示聚类成员 $P^{(i)}$ 和 $P^{(j)}$ 之间的 NMI 值, S_{ij} 越大, 其相似度越大, 差异性越小. Proportion 的计算方法如下:

$$\mathrm{Proportion}=\frac{2}{l(l-1)}\sum_{i=1}^{l-1}\sum_{j=i+1}^{l}F(S_{ij}) \tag{3}$$

其中, $F(x)$ 为指示函数: 当 $x>0.5$ 时, $F(x)$ 为 1, 否则为 0.

由式 (1) 可知, MNMI 的大小不仅与聚类集体有关, 还与 π^{*} 有关. 由式 (2) 和式 (3) 可知, TANMI 统计聚类集体的整体平均 NMI 值, Proportion 统计聚类成员之间 NMI 值大于 0.5 的比例, 对于给定的聚类集体, TANMI 和 Proportion 是固定不变的. 聚类集体稳定性分为两种情况: 1) 聚类集体稳定, 此时, 多数聚类成员之间的相似度较高, 差异性较低 (NMI 值大于 0.50), Proportion > 0.5, TANMI > 0.5, 多数聚类成员与 π^{*} 的 NMI 值也大于 0.5, 故 MNMI > 0.5, 因此, ACES 与 IACES 方法都能正确判定聚类集体为稳定. 2) 聚类集体不稳定, 此时, 多数聚类成员之间的相似度较低, 差异性较高 (NMI 值小于 0.5), Proportion < 0.5, TANMI < 0.5, IACES 方法能够正确判定聚类集体为不稳定, 而 ACES 判定聚类集体是否稳定与 π^{*} 有关. 当多数聚类成员与 π^{*} 的 NMI 值大于 0.5 时, MNMI > 0.5, ACES 方法会将聚类集体误判为稳定; 当多数聚类成员与 π^{*} 的 NMI 值小于 0.5 时, MNMI < 0.5, ACES 方法判定聚类集体不稳定.

2.3 聚类成员选择方法

根据集成学习理论[37], 集成的泛化误差 E 等于集成中各基学习器的平均泛化误差 $\overline{E}$ 与平均差异性 $\overline{A}$ 之差, 即 $E=\overline{E}-\overline{A}$. 因此, 要提高集成学习的性能, 可从两个方面着手: 1) 尽量降低各基学习器的误差; 2) 尽量增加基学习器之间的差异性.

文献 [28] 通过实验对比了 4 种聚类成员集合, 分别是所有聚类成员构成的集合 Full, 与 π^{*} 差异性最低的一半聚类成员构成的集合 Low, 与 π^{*} 差异性最高的一半聚类成员构成的集合 High, 与 π^{*} 差异性适中的一半聚类成员构成的集合 Medium, 实验结果显示, 当聚类集体稳定时, 选择 Full 集合进行集成获得了最佳结果; 当聚类集体不稳定时, 选择 High 集合进行集成获得了最佳结果.

本文在 ACES 基础上进行了改进, 提出以下聚类成员选择方法: 当聚类集体稳定时, 本文首先选择与 π^{*} 的 NMI 值最大的 1/4 的聚类成员, 尽量降低聚类成员的平均误差; 然后选择与 π^{*} 的差异性最高 (NMI 值最小) 的 1/4 的聚类成员, 尽量增加平均差异性, 构成聚类成员集合 LaH (Low and high); 另外, 为了避免选出精度较低的聚类成员 (离群点), 在 LaH 集合中剔除部分平均 NMI 值较低的聚类成员. 具体地, 本文对每个聚类成员与其他聚类成员的 ANMI 进行排名, 限制选出的聚类成员排名在前 θ 以内, $0\,\%<\theta\leq 100\,\%$. 不妨假设聚类成员 $P^{(i)}$ 的 ANMI_i 排名为 $\mathrm{Rank}^{(i)}$, 则符合条件的聚类成员集合为 $C=\{P^{(i)}|\ \mathrm{Rank}^{(i)}/l\geq\theta,\ 1\leq i\leq l\}$, 因此, 选择的聚类成员集合为 LaH $\cap$ C, 其中 $\cap$ 表示集合的交运算. 此时, LaH $\cap$ C 既具有较高质量, 又具有适中的差异性, 对其进行集成往往能够获得比 π^{*} 更加优越的结果. 当聚类集体不稳定时, 聚类成员提供了不同的聚类结构, 差异可能由聚类算法本身的偏置或数据集的复杂结构引起, 此时 π^{*} 很可能偏离了真实分类结果, 因此应该尽量降低聚类成员的平均误差, 选择与 π^{*} 差异性大的聚类成员 (High 集) 可能会得到更好的结果. 但 High 集里可能会存在一些质量较差的聚类成员, 此时可通过约束聚类成员的 ANMI 排名在某一范围内. 本文对每个聚类成员与其他聚类成员的 ANMI 进行排名, 限制选出的聚类成员排名在前 θ 以内, $0\,\%<\theta\leq 100\,\%$, 因此, 选择的聚类成员集合为 High $\cap$ C.

为了确定合理的 θ, 本文首先采用不同的 θ 选出不同的聚类成员子集, 然后对每个子集进行集成, 获得多个 CCP, 最后使用内部有效性指标 DB 对每个 CCP 进行评价, 得到最低 DB 值的 CCP 即为最终的一致划分. 考虑到要确定最佳的参数 θ 需要运行 l 次, 而 DB 的计算复杂度较高, 当聚类成员个数 l 较大时, 需要耗费极其高昂的计算代价. 因此, 为了提高算法效率, 本文仅设置 $\theta=10\,\%, 20\,\%, \cdots, 100\,\%$, 比较这 10 种情况下获得的最佳结果, 获得一个较优解. 在本文实验中, 绝大多数情况下, 当 θ 等于 70 %, 80 % 或 90 % 时获得了最低的 DB 值.

3 实验

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 MAT-

LAB2016b 下运行.

3.1 实验数据集和评价指标

实验采用 13 组公共文本测试集, 具体描述如表 1 所示, 数据集中的文本数 (n_d) 从 204～8 580 不等, 特征数 (n_w) 从 5 832～41 681 不等, 类别数 (k^*) 从 3～10 不等, 平均每个类别包含的文本数 (n_c) 从 34～1 774 不等, 平衡因子 (Balance) 从 0.037～0.998 不等 (Balance 等于最小类别包含的文本数除以最大类别包含的文本数, 值越小, 数据集越不平衡, 反之越平衡). 对于每个数据集, 使用停用词表移去停用词, 并且去掉出现在少于两个文本中的词. 数据集 tr11, tr23, tr41 和 tr45 取自 TREC-5TREC-6 和 TREC-7 数据集 la1, la2 和 la12 取自 TREC-5, 由洛杉矶时报 (*Los Angles Times*) 上的文章构成. 数据集 hitech, reviews 和 sports 取自报纸 San Jose Mercury, 它们是 TREC 文本集的一部分. 数据集 classic 由用于评估信息检索系统的四种摘要构成, 每个摘要集合构成单独的一类. 数据集 k1b 来自于 WebACE project[38], 每个文本对应于 Yahoo! 主题层次下的一个网页. ng3 为 NG20 的子集, 包含了有关政治的 3 个不同方面, 每方面分别包含约 1 000 条信息.

表 1 实验数据集描述

Table 1 Description of datasets

Dataset	n_d	n_w	k^*	n_c	Balance
tr11	414	6 429	9	46	0.046
tr23	204	5 832	6	34	0.066
tr41	878	7 454	10	88	0.037
tr45	690	8 261	10	69	0.088
la1	3 204	31 472	6	534	0.290
la2	3 075	31 472	6	543	0.274
la12	6 279	31 472	6	1 047	0.282
hitech	2 301	10 080	6	384	0.192
reviews	4 069	18 483	5	914	0.098
sports	8 580	14 870	7	1 226	0.036
classic	7 094	41 681	4	1 774	0.323
k1b	2 340	21 839	6	390	0.043
ng3	2 998	15 810	3	999	0.998

因为文本类别标签已知, 本文采用 NMI 值量化聚类结果和已知类别的匹配程度. 当两个类别标签一一对应时, NMI 值达到最大值 1. 另外, 本文还采在信息检索领域常用的综合指标, F 值 (F-measure). F 值越大, 聚类质量越高, 反之越低.

3.2 实验设计与结果分析

本节通过实验对文献 [28] 提出的 ACES、本文提出的 IACES、文献 [27] 提出的 CAS 进行比较, 其中 ACES 和 IACES 根据聚类集体的稳定性自适应选择不同的聚类成员, CAS 通过人工设置的方法选择不同个数的聚类成员 (分别设置为聚类集体大小的 5 %～25 %, 以 5 % 递增, 本文沿用该方法). 实验分为两部分: 1) 聚类集体稳定性判定方法对比, 分别根据 ACES 和 IACES 判定出聚类集体的稳定性结果, 并进行分析比较, 验证本文提出的聚类集体稳定性判定方法的有效性; 2) 聚类成员选择方法对比, 分别根据 ACES, IACES 和 CAS 选择不同的聚类成员集合并采用不同的共识函数设计方法进行集成, 比较不同算法获得的 NMI 值和 F 值, 验证本文的聚类成员选择方法的有效性. 下面对实验中采用的聚类成员生成策略和共识函数设计方法进行介绍.

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 l 次, 每次生成 k_0 个簇, 采用如下两种不同的策略分别生成 l = 1 000 个聚类成员: 1) $k_0 = k^*$; 2) k_0 随机选自区间 $[2, 2 \times k^*]$, 由此分别构建聚类集体 P_1 和 P_2. 策略 1 是聚类集成研究中最常见的方法, 由于采用了相同的聚类算法, 每个算法生成的簇个数相等, 因此聚类成员的差异性仅由不同初始聚类中心引起, 聚类集体往往会缺乏多样性. 策略 2 试图通过约束聚类成员具有不同的簇个数来提高聚类成员多样性.

常见的共识函数设计方法有基于图划分算法的 CSPA, HGPA, MCLA (其中 CSPA 总体聚类效果最好), 基于层次聚类 SL, CL, AL, WL 的方法 (其中 AL 总体聚类效果最好), 基于谱聚类 (Spectral clustering, SC) 的方法, 基于 KM 的方法. 因此, 本文采用 CSPA, AL, SC 和 KM 进行集成. CSPA 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. AL 获得了稳定的聚类结果. 谱聚类方法由于调用了 KM 算法, 在部分数据集上获得的聚类结果不够稳定, 本文重复运行 KM 算法 10 次取最优结果. 基于 KM 的方法获得的聚类结果极不稳定, 受初始聚类中心影响较大. 为了提高聚类结果的稳定性和聚类质量, 本文引入 K-means++ (KM++) 算法, 运行 KM++ 10 次取最优结果.

3.2.1 聚类集体稳定性判定方法对比

表 2 给出了分别根据 ACES 和 IACES 方法判定的聚类集体稳定性结果, 其中 MNMI 值根据式 (1) 计算, Number 表示与 π^* 的 NMI 值大于 0.5 的聚类成员个数, TANMI 根据式 (2) 计算, Propor-

表 2 分别根据 ACES 和 IACES 判定的聚类集体稳定性结果

Table 2 Stability results of cluster ensemble according to ACES and IACES

	聚类集体 P_1						聚类集体 P_2					
	ACES			IACES			ACES			IACES		
Dataset	MNMI	Number	Stability	TANMI	Proportion	Stability	MNMI	Number	Stability	TANMI	Proportion	Stability
tr11	0.655	989	S	0.539	0.7498	S	0.682	940	S	0.574	0.8384	S
tr23	0.663	991	S	0.607	0.9361	S	0.712	904	S	0.649	0.8736	S
tr41	0.731	999	S	0.642	0.9939	S	0.732	959	S	0.649	0.8922	S
tr45	0.718	1 000	S	0.640	0.9917	S	0.705	922	S	0.616	0.8121	S
la1	0.597	863	S	0.514	0.5553	S	0.592	894	S	0.541	0.6879	S
la2	0.593	934	S	0.524	0.6296	S	0.539	735	S	0.489	0.4374	NS
la12	0.634	973	S	0.558	0.7586	S	0.570	838	S	0.493	0.4938	NS
hitech	0.551	727	S	0.475	0.3251	NS	0.537	654	S	0.458	0.2602	NS
reviews	0.683	940	S	0.610	0.8480	S	0.672	958	S	0.608	0.7622	S
sports	0.736	998	S	0.652	0.9637	S	0.651	958	S	0.585	0.7443	S
classic	0.801	966	S	0.692	0.8375	S	0.709	945	S	0.594	0.7500	S
k1b	0.673	994	S	0.585	0.8992	S	0.654	969	S	0.555	0.7811	S
ng3	0.541	664	S	0.451	0.3791	NS	0.525	648	S	0.467	0.4441	NS

tion 根据式 (3) 计算, Stability 为 “S” 表示聚类集体稳定, Stability 为 “NS” 表示聚类集体不稳定.

根据表 2, 可以进行以下比较:

1) 因为在所有数据集上 MNMI 值都大于 0.5, 所以 ACES 将由不同聚类成员生成策略产生的聚类集体都判定为稳定. 当 $k_0 = k^*$ 时, 在 hitech 和 ng3 上, 聚类集体的 TANMI 值小于 0.5, 所以 IACES 将其判定为不稳定, 而在其他 11 个数据集上, 聚类集体的 TANMI 值大于 0.5, 所以 IACES 将其判定为稳定; 当 $k_0 \in [2, 2 \times k^*]$ 时, 在 la2, la12, hitech 和 ng3 上, 聚类集体的 TANMI 值小于 0.5, 所以 IACES 将其判定为不稳定, 而在其他 9 个数据集上, 聚类集体的 TANMI 值都大于 0.5, 所以 IACES 将其判定为稳定.

2) 如果以 Proportion 是否高于 0.5 来判定聚类集体稳定与否, 那么在所有数据集上的判定结果都与依据 TANMI 判定的结果一致, 而在部分数据集上与依据 MNMI 判定的结果不一致. 究其原因, 当半数以上的聚类成员之间的 NMI 值大于 0.5 时, 聚类集体的差异性相对较低, 聚类集体稳定. 此时, TANMI 值也大于 0.5, 绝大多数聚类成员 (Number $>$ 500) 与 π^* 的 NMI 值大于 0.5, 故 MNMI 也大于 0.5. 因此, 根据 TANMI 和 ANMI 判定的结果都是聚类集体稳定. 然而, 当半数以下的聚类成员之间的 NMI 值大于 0.5 时, 聚类集体的差异性相对较高, 聚类集体不稳定. 此时, TANMI 值也小于 0.5, 但仍然有绝大多数聚类成员 (Number $>$ 500) 与 π^* 的 NMI 值大于 0.5, 故 MNMI 仍然大于 0.5. 因此, 根据 TANMI 判定的结果是聚类集体不稳定, 而根据 ANMI 判定的结果依然是聚类集体稳定.

综上, 由于 IACES 依据 TANMI 判定聚类集体稳定性, 只客观地依赖于聚类集体本身的特性, 因此能够准确判定其稳定性; 而由于 ACES 依据 MNMI 判定聚类集体稳定性, 与初始一致划分 π^* 有关, 因此会将某些不稳定的聚类集体误判为稳定.

3.2.2 聚类成员选择方法对比

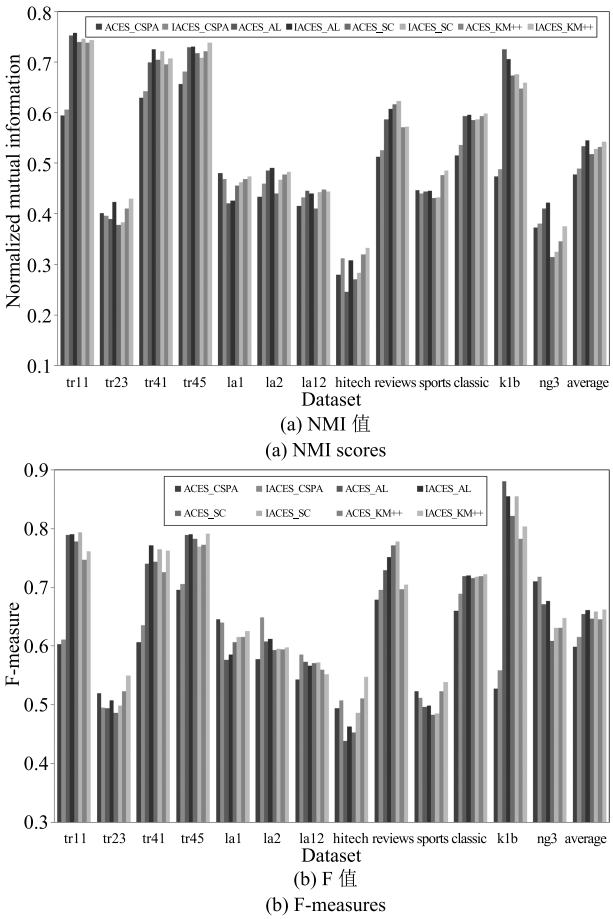
1) 分别根据 ACES 和 IACES 选择聚类成员并进行集成获得的结果对比.

图 2 和图 3 分别显示了采用聚类集体 P_1 和 P_2 时根据 ACES 和 IACES 选择聚类成员, 并采用 CSPA, AL, SC, KM++ 进行集成获得的 NMI 值和 F 值, 其中 Average 统计了 8 种算法 (以 “聚类成员选择方法_共识函数设计方法” 命名) 在 13 组数据集上的平均结果.

根据图 2, 分别比较不同共识函数根据 ACES 和 IACES 选择聚类成员并进行集成获得的 NMI 值和 F 值, 可以发现:

a) 对于聚类集体不稳定的 2 种情况 (hitech 和 ng3), IACES_CSPA, IACES_AL, IACES_SC 和

IACES_KM++ 获得的 NMI 值和 F 值都分别高于 ACES_CSPA, ACES_AL, ACES_SC 和 ACES_KM++.



(a) NMI 值

(a) NMI scores

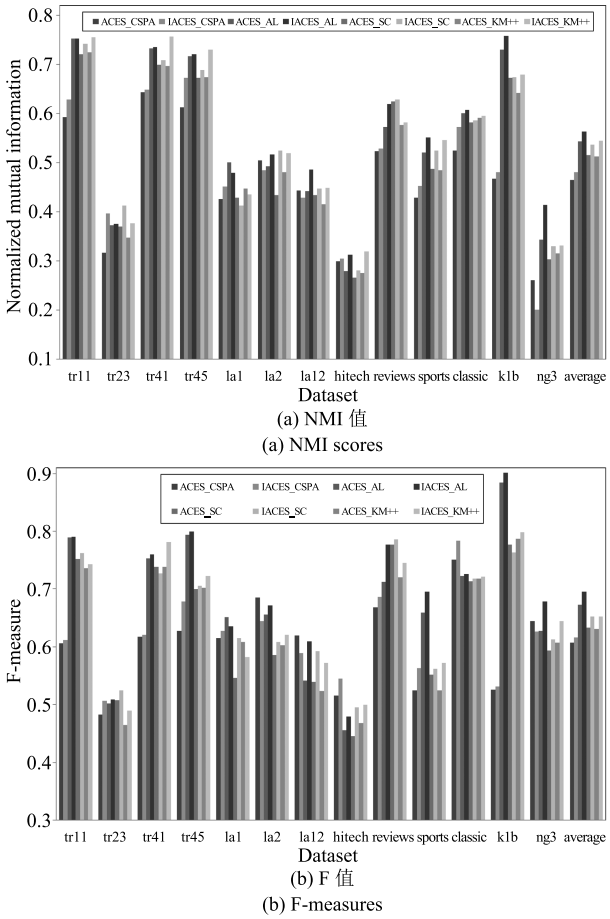
(b) F 值

(b) F-measures

图 2　采用聚类集体 P_1 时获得的聚类结果 (NMI 值和 F 值)

Fig. 2　Clustering results obtained when using cluster ensemble P_1 (NMI scores and F measures)

b) 对于其他 11 种聚类集体稳定的情况, IACES_CSPA 仅在 tr23, la1, sports 上获得了比 ACES_CSPA 低的 NMI 值和 F 值, 而在其他 8 组数据集上都获得了高于 ACES_CSPA 的 NMI 值和 F 值; IACES_AL 仅在 la12 和 k1b 上获得了比 ACES_AL 低的 NMI 值和 F 值, 而在其他 9 组数据集上都获得了高于 ACES_AL 的 NMI 值和 F 值; IACES_SC 仅在 tr45 上获得了比 ACES_SC 低的 NMI 值, 而在其他 10 组数据集上都获得了高于 ACES_SC 的 NMI 值, IACES_SC 仅在 tr45 和 k1b 上获得了比 ACES_SC 低的 F 值, 而在其他 9 组数据集上都获得了高于 ACES_SC 的 F 值; IACES_KM++ 仅在 la12 上获得了比 ACES_KM++ 低的 NMI 值和 F 值, 而在其他 10 组数据集上都获得了高于 ACES_KM++ 的 NMI 值和 F 值.



(a) NMI 值

(a) NMI scores

(b) F 值

(b) F-measures

图 3　采用聚类集体 P_2 时获得的聚类结果 (NMI 值和 F 值)

Fig. 3　Clustering results obtained when using cluster ensemble P_2 (NMI scores and F measures)

c) 总体来看, 当采用聚类集体 P_1 时, 与 ACES 相比, 采用 IACES 进行成员选择, CSPA 分别以 10/13 和 10/13 的比例提高了 NMI 值和 F 值; AL 分别以 11/13 和 11/13 的比例提高了 NMI 值和 F 值; SC 分别以 12/13 和 11/13 的比例提高了 NMI 值和 F 值; KM++ 分别以 12/13 和 12/13 的比例提高了 NMI 值和 F 值; CSPA, AL, SC, KM++ 获得的平均 NMI 值和 F 值都有不同程度的提高.

根据图 3, 分别比较不同共识函数根据 ACES 和 IACES 选择聚类成员并进行集成获得的 NMI 值和 F 值, 可以发现:

a) 对于聚类集体不稳定的 4 种情况 (la2, la12, hitech 和 ng3), IACES_CSPA 在 la2, la12 和 ng3 上获得的 NMI 值和 F 值低于 IACES_CSPA, 在 hitech 上获得了比 ACES_CSPA 高的 NMI 和 F 值; IACES_AL, IACES_SC 和 IACES_KM++ 获得的 NMI 值和 F 值都分别高于 ACES_AL, ACES_SC 和 ACES_KM++.

b) 对于其他 7 种聚类集体稳定的情况,

IACES_CSPA 在所有 7 组数据集上都获得了高于 ACES_CSPA 的 NMI 值和 F 值; IACES_AL 仅在 tr11 和 la1 上获得了比 ACES_AL 低的 NMI 值, 而在其他 5 组数据集上都获得了高于 ACES_AL 的 NMI 值, IACES_AL 仅在 la1 上获得了比 ACES_AL 低的 F 值, 而在其他 6 组数据集上都获得了高于 ACES_AL 的 F 值; IACES_SC 仅在 la1 上获得了比 ACES_SC 低的 NMI 值, 而在其他 6 组数据集上都获得了高于 ACES_SC 的 NMI 值, IACES_SC 仅在 tr41 上获得了比 ACES_SC 低的 F 值, 而在其他 6 组数据集上都获得了高于 ACES_SC 的 F 值; IACES_KM++ 仅在 la1 上获得了比 ACES_KM++ 低的 NMI 值和 F 值, 而在其他 6 组数据集上都获得了高于 ACES_KM++ 的 NMI 值和 F 值.

c) 总体来看, 当采用聚类集体 P_2 时, 与 ACES 相比, 采用 IACES 进行成员选择, CSPA 分别以 10/13 和 10/13 的比例提高了 NMI 值和 F 值; AL 分别以 11/13 和 12/13 的比例提高了 NMI 值和 F 值; SC 分别以 12/13 和 11/13 的比例提高了 NMI 值和 F 值; KM++ 分别以 12/13 和 12/13 的比例提高了 NMI 值和 F 值; CSPA, AL, SC, KM++ 获得的平均 NMI 值和 F 值都有不同程度的提高.

综上, 与 ACES 相比, 根据 IACES 选择聚类成员进行 CSPA, AL, SC 和 KM++ 集成在绝大部分情况下都获得了更高的 NMI 值和 F 值, 每个共识函数设计方法在所有数据集上获得的平均 NMI 值和 F 值都更高.

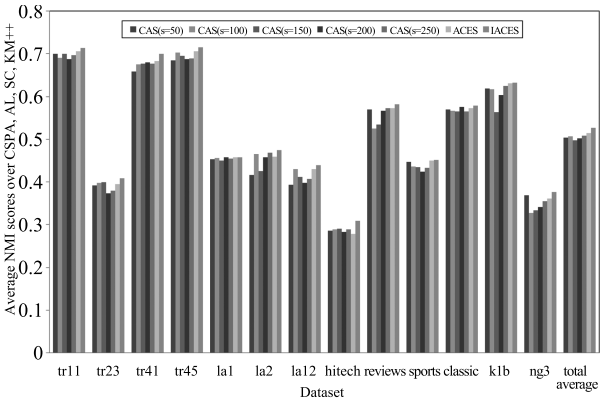
2) 聚类成员选择方法 CAS, ACES, IACES 综合比较

本文实验中聚类集体大小为 1 000, CAS 分别选择 $s = 50, 100, 150, 200, 250$ 个聚类成员, 每个 s 对应一种聚类成员选择方法, 例如 CAS ($s = 50$) 表示根据 CAS 选择 50 个聚类成员. 图 4 和图 5 分别显示了当采用聚类集体 P_1 和 P_2 时, 根据 CAS, ACES 和 IACES 选择聚类成员并采用 CSPA, AL, SC, KM++ 进行集成获得的 NMI 值和 F 值的平均值 (例如, 图 4 中的 IACES 表示 4 个聚类集成算法 IACES_CSPA, IACES_AL, IACES_SC, IACES_KM++ 获得的 NMI 值的平均值), 其中 Total average 统计了 7 种不同聚成员选择方法在 13 组数据集上的平均结果.

由图 4 和图 5 可见:

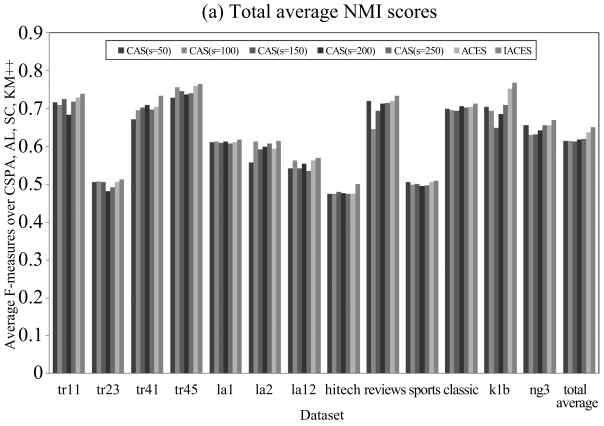
a) IACES 在每个数据集上获得的 NMI 值和 F 值的平均值都高于 ACES 和 CAS.

b) ACES 在某些数据集上获得的 NMI 值和 F 值的平均值低于 CAS (例如图 4 中的 tr23 和 ng3), 但在绝大部分情况下都高于 CAS.



(a) 平均 NMI 值

(a) Total average NMI scores



(b) 平均 F 值

(b) Total average F-measures

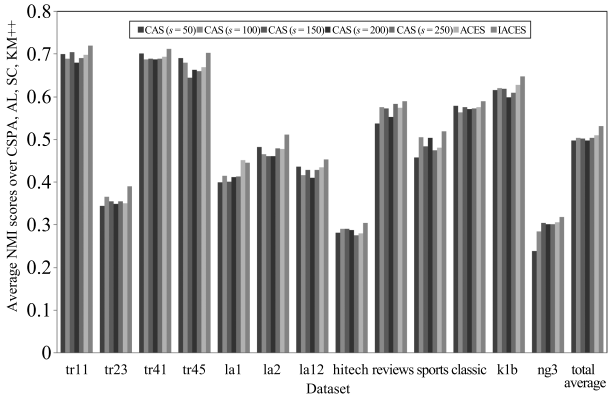
图 4 当采用聚类集体 P_1 时获得的聚类结果 (平均 NMI 值和平均 F 值)

Fig. 4 Clustering results obtained by combining cluster members selected by ACES and IACES via CSPA, AL, SC and KM++ when using cluster ensemble P_1 (Total average NMI scores and total average F measures)

c) 总体来看, IACES 在所有数据集上获得了最高的平均结果, ACES 次之, 即自适应聚类成员选择方法 ACES 优于 CAS 方法, 而本文的方法则比 ACES 更加优越.

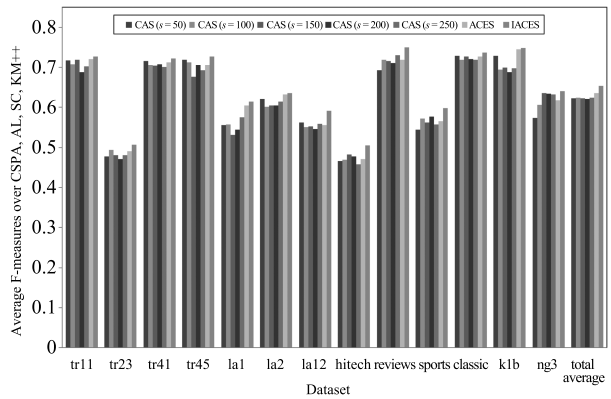
4 结论

本文提出了一种改进的自适应聚类集成选择方法 (IACES), 有效解决了 ACES 存在的聚类集体稳定性判定方法不客观和聚类成员选择方法不够合理的问题. 在多组基准数据集上进行了实验, 实验结果表明: 1) IACES 能够准确判定聚类集体的稳定性, 而 ACES 会将某些不稳定的聚类集体误判为稳定; 2) 与其他聚类成员选择方法相比, 根据 IACES 选择聚类成员进行集成在绝大部分情况下都获得了更佳的聚类结果, 在所有数据集上都获得了更优的平均聚类结果.



(a) 平均 NMI 值

(a) Total average NMI scores



(b) 平均 F 值

(b) Total average F-measures

图 5 当采用聚类集体 P_1 时获得的聚类结果 (平均 NMI 值和平均 F 值)

Fig. 5 Clustering results obtained by combining cluster members selected by ACES and IACES via CSPA, AL, SC and KM++ when using cluster ensemble P_1 (Total average NMI scores and total average F measures)

References

1 Duda R O, Hart P E, Stork D G. *Pattern Classification* (2nd edition). New York: John Wiley and Sons, 2001.

2 Jain A K, Murty M N, Flynn P J. Data clustering: a review. *ACM Computing Surveys (CSUR)*, 1999, **31**(3): 264−323

3 Jain A K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 2010, **31**(8): 651−666

4 Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999, **401**(6755): 788−791

5 Frey B J, Dueck D. Clustering by passing messages between data points. *Science*, 2007, **315**(5814): 972−976

6 Deng Z H, Choi K S, Jiang Y Z, Wang J, Wang S T. A survey on soft subspace clustering. *Information Sciences*, 2014, **348**: 84−106

7 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. *Science*, 2006, **313**(5786): 504−507

8 Xie J Y, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis. In: Proceedings of the 33rd International Conference on Machine Learning. New York City, NY, USA: International Machine Learning Society, 2016. 478−487

9 Rodriguez A, Laio A. Clustering by fast search and find of density peaks. *Science*, 2014, **344**(6191): 1492−1496

10 von Luxburg U. A tutorial on spectral clustering. *Statistics and Computing*, 2007, **17**(4): 395−416

11 Strehl A, Ghosh J. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 2002, **3**(3): 583−617

12 Topchy A, Jain A K, Punch W. A mixture model for clustering ensembles. In: Proceedings of the 4th SIAM International Conference on Data Mining. Lake Buena Vista, FL, USA: SIAM, 2004. 379−390

13 Fern X Z, Brodley C E. Solving cluster ensemble problems by bipartite graph partitioning. In: Proceedings of the 21st International Conference on Machine Learning. Banff, Alberta, Canada: ACM, 2004. 36

14 Fred A L N, Jain A K. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(6): 835−850

15 Li T, Ding C, Jordan M I. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM). Omaha, NE, USA: IEEE, 2007. 577−582

16 Ayad H G, Kamel M S. Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(1): 160−173

17 Iam-On N, Boongeon T, Garrett S, Price C. A link-based cluster ensemble approach for categorical data clustering. *IEEE Transactions on Knowledge and Data Engineering*, 2012, **24**(3): 413−425

18 Carpineto C, Romano G. Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012, **34**(12): 2315−2326

19 Wu J J, Liu H F, Xiong H, Cao J, Chen J. K-means-based consensus clustering: a unified view. *IEEE Transactions on Knowledge and Data Engineering*, 2015, **27**(1): 155−169

20 Berikov V, Pestunov I. Ensemble clustering based on weighted co-association matrices: error bound and convergence properties. *Pattern Recognition*, 2017, **63**: 427−436

21 Zhou Z H, Tang W. Clusterer ensemble. *Knowledge-Based Systems*, 2006, **19**(1): 77−83

22 Yang Y, Kamel M S. An aggregated clustering approach using multi-ant colonies algorithms. *Pattern Recognition*, 2006, **39**(7): 1278−1289

23 Luo Hui-Lan, Kong Fan-Sheng, Li Yi-Xiao. An analysis of diversity measures in clustering ensembles. *Chinese Journal of Computers*, 2007, **30**(8): 1315−1324
(罗会兰, 孔繁胜, 李一啸. 聚类集成中的差异性度量研究. 计算机学报, 2007, **30**(8): 1315−1324)

24 Yu Z W, Li L, Liu J M, Zhang J, Han G Q. Adaptive noise immune cluster ensemble using affinity propagation. *IEEE Transactions on Knowledge and Data Engineering*, 2015, **27**(12): 3176−3189

25 Chu Rui-Hong, Wang Hong-Jun, Yang Yan, Li Tian-Rui. Clustering ensemble based on density peaks. *Acta Automatica Sinica*, 2016, **42**(9): 1401−1412
(褚睿鸿, 王红军, 杨燕, 李天瑞. 基于密度峰值的聚类集成. 自动化学报, 2016, **42**(9): 1401−1412)

26 Xu S, Chan K S, Gao J, Xu X F, Li X F, Hua X P, An J. An integrated K-means-Laplacian cluster ensemble approach for document datasets. *Neurocomputing*, 2016, **214**: 495−507

27 Fern X Z, Lin W. Cluster ensemble selection. *Statistical Analysis and Data Mining*, 2008, **1**(3): 128−141

28 Azimi J, Fern X. Adaptive cluster ensemble selection. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, California, USA: ACM, 2009. 992−997

29 Naldi M C, Carvalho A C P L F, Campello R J G B. Cluster ensemble selection based on relative validity indexes. *Data Mining and Knowledge Discovery*, 2013, **27**(2): 259−289

30 Bi Kai, Wang Xiao-Dan, Xing Ya-Qiong. Cluster ensemble selection based on validity index in evidence space. *Journal on Communications*, 2015, **36**(8): 135−145
(毕凯, 王晓丹, 邢雅琼. 基于证据空间有效性指标的聚类选择性集成. 通信学报, 2015, **36**(8): 135−145)

31 Iam-On N, Boongoen T. Comparative study of matrix refinement approaches for ensemble clustering. *Machine Learning*, 2015, **98**(1−2): 269−300

32 Fern X Z, Brodley C E. Random projection for high dimensional data clustering: a cluster ensemble approach. In: Proceedings of the 20th International Conference on Machine Learning. Washington, DC, USA: ACM, 2003. 186−193

33 Hadjitodorov S T, Kuncheva L I, Todorova L P. Moderate diversity for better cluster ensembles. *Information Fusion*, 2006, **7**(3): 264−275

34 Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and an algorithm. In: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic. Vancouver, British Columbia, Canada: ACM, 2001. 849−856

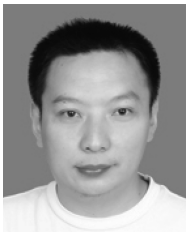
35 Xu Sen, Zhou Tian, Yu Hua-Long, Li Xian-Feng. Matrix low rank approximation-based cluster ensemble algorithm. *Acta Electronica Sinica*, 2013, **41**(6): 1219− 1224
(徐森, 周天, 于化龙, 李先锋. 一种基于矩阵低秩近似的聚类集成算法. 电子学报, 2013, **41**(6): 1219−1224)

36 Zhou Lin, Ping Xi-Jian, Xu Sen, Zhang Tao. Cluster ensemble based on spectral clustering. *Acta Automatica Sinica*, 2012, **38**(8): 1335−1342
(周林, 平西建, 徐森, 张涛. 基于谱聚类的聚类集成算法. 自动化学报, 2012, **38**(8): 1335−1342)

37 Krogh A, Vedelsby J. Neural network ensembles, cross validation and active learning. In: Proceedings of the 7th International Conference on Neural Information Processing Systems. Denver, CO, USA: ACM, 1994. 231−238

38 Han E H, Boley D, Gini M, Gross R, Hastings K, Karypis G, Kumar V, Mobasher B, Moore J. WebACE: a web agent for document categorization and exploration. In: Proceedings of the 2nd International Conference on Autonomous Agents. Minneapolis, Minnesota, USA: ACM, 1998. 408−415

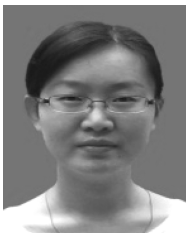
徐 森 盐城工学院信息工程学院副教授. 主要研究方向为机器学习, 人工智能, 文本挖掘. 本文通信作者.
E-mail: xusen@ycit.cn
(**XU Sen** Associated professor at the School of Information Engineering, Yancheng Institute of Technology. His research interest covers machine learning, artificial intelligence and document mining. Corresponding author of this paper.)



皋 军 盐城工学院信息工程学院教授. 主要研究方向为机器学习, 人工智能.
E-mail: gaoj@ycit.cn
(**GAO Jun** Professor at the School of Information Engineering, Yancheng Institute of Technology. His research interest covers machine learning and artificial intelligence.)

花小朋 盐城工学院信息工程学院副教授. 主要研究方向为机器学习, 人工智能.
E-mail: huaxp@ycit.cn
(**HUA Xiao-Peng** Associate professor at the School of Information Engineering, Yancheng Institute of Technology. His research interest covers machine learning and artificial intelligence.)

李先锋 盐城工学院信息工程学院副教授. 主要研究方向为机器学习, 人工智能.
E-mail: lxf@ycit.cn
(**LI Xian-Feng** Associate professor at the School of Information Engineering, Yancheng Institute of Technology. His research interest covers machine learning and artificial intelligence.)



徐 静 盐城工学院信息工程学院副教授. 主要研究方向为机器学习, 人工智能.
E-mail: xujingycit@163.com
(**XU Jing** Associate professor at the School of Information Engineering, Yancheng Institute of Technology. Her research interest covers machine learning and artificial intelligence.)