基于密度峰值的聚类集成

褚睿鸿 [1] 王红军 [1] 杨 燕 [1] 李天瑞 [1]

摘 要 聚类集成的目的是为了提高聚类结果的准确性、稳定性和鲁棒性. 通过集成多个基聚类结果可以产生一个较优的结果. 本文提出了一个基于密度峰值的聚类集成模型, 主要完成三个方面的工作: 1) 在研究已有的各聚类集成算法和模型后发现各基聚类结果可以用密度表示; 2) 使用改进的最大信息系数 (Rapid computation of the maximal information coefficient, RapidMic) 表示各基聚类结果之间的相关性, 使用这种相关性来衡量原始数据在经过基聚类器聚类后相互之间的密度关系; 3) 改进密度峰值 (Density peaks, DP) 算法进行聚类集成. 最后, 使用一些标准数据集对所设计的模型进行评估. 实验结果表明, 相比经典的聚类集成模型, 本文提出的模型聚类集成效果更佳.

关键词 聚类集成, 近邻传播, 密度峰值, 相似性矩阵

引用格式 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

DOI 10.16383/j.aas.2016.c150864

Clustering Ensemble Based on Density Peaks

CHU Rui-Hong[1] WANG Hong-Jun[1] YANG Yan[1] LI Tian-Rui[1]

Abstract Clustering ensemble aims to improve the accuracy, stability and robustness of clustering results. A good ensemble result is achieved by integrating multiple base clustering results. This paper proposes a clustering ensemble model based on density peaks. First, this paper discovers that the base clustering results can be expressed with density after studying and analyzing the existing clustering algorithms and models. Second, rapid computation of the maximal information coefficient (RapidMic) is introduced to represent the correlation of the base clustering results, which is then used to measure the density of these original datasets after base clustering. Third, the density peak (DP) algorithm is improved for clustering ensemble. Furthermore, some standard datasets are used to evaluate the proposed model. Experimental results show that our model is effective and greatly outperforms some classical clustering ensemble models.

Key words Clustering ensemble, affinity propagation, density peaks, similarity matrix

Citation 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

1 绪论

1.1 研究背景和研究意义

类簇可以描述为一个包含密度相对较高的点集的多维空间中的连通区域, 一个类簇内的实体是相似的, 不同类簇的实体是不相似的, 同一类簇任意两个点间的距离小于不同类簇任意两个点间的距离[1]. 近几年, 有不少新的聚类算法被提出. 周晨曦等[2] 设计出一种基于动态更新约束的半监督凝聚层次聚类方法, 其更新过程可以保证最终结果的有效性. 为了处理混合属性的数据, 陈晋音等[3] 提出了基于自动确定密度聚类中心的聚类算法. 王卫卫等[4] 提出了 SSC (Sparse subspace clustering), 能够揭示高维数据真实子空间结构的表示模型. Taşdemir 等[5] 通过对高空间分辨率遥感影像进行无监督聚类来识别土地覆盖. Parvin 等[6] 提出了一种模糊加权聚类算法 (Fuzzy weighted locally adaptive clustering, FWLAC), 能够处理不平衡的聚类.

单一的聚类方法普遍存在局限性, 例如聚类结果很大程度上取决于参数及其初始化, 无法准确判断数据集的真实类簇个数等. 另外, 真实的数据集往往具有不同的结构和大小, 因此, 任何一种聚类方法都无法在全部的数据集上获得好的聚类效果.

收稿日期 2015-12-25 录用日期 2016-04-18
Manuscript received December 25, 2015; accepted April 18, 2016
国家科技支撑计划课题 (2015BAH19F02), 国家自然科学基金 (61262058, 61572407), 教育部在线教育研究中心在线教育研究基金 (全通教育) (2016YB158), 西南交通大学中央高校基本科研业务费专项基金 (A0920502051515-12) 资助
Supported by National Science and Technology Support Program (2015BAH19F02), National Natural Science Foundation of China (61262058, 61572407), Online Education Research Center of the Ministry of Education Online Education Research Fund (Full Education) (2016YB158) and Fundamental Research Funds for the Central Universities of Southwest Jiaotong University (A0920502051515-12)
本文责任编委 王立威
Recommended by Associate Editor WANG Li-Wei
1. 西南交通大学信息科学与技术学院 成都 611756
1. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756

集成学习是指通过集成多个不同学习器来解决同一问题. Strehl 等[7] 最早明确提出聚类集成: 通过计算和比较多个基聚类结果的相关关系以及信息熵, 集成多个基聚类结果可以获得一个综合性的结果. 聚类集成有很多优点, 不仅能够提高聚类结果的准确性、稳定性和鲁棒性, 还能并行处理数据集[7−9].

近年来, 聚类集成在原有的基础上获得了极大的发展, 技术也日趋成熟, 越来越多的方法被应用到数据挖掘、生物信息、医学等领域. 与此同时, 一些问题也显露出来, 例如从不同结构的数据源中获得的基聚类结果往往具有不同的结构, 如何确定最具代表性的聚类结构或是构建一个新的聚类集成结构显得尤为重要[10−11].

本文在对各类已有的聚类集成算法和模型研究的基础上, 分析基聚类算法对原始数据进行聚类后其数据结构上的改变. 探索了基聚类结果是否可以用原始数据的密度关系进行表示. 尝试用改进的密度峰值 (Density peaks, DP) 算法[12] 进行聚类集成.

1.2 主要研究内容

本文主要有三个方面的研究内容: 1) 本文把基聚类结果看成是原始数据的新增属性, 研究结果发现只使用这些新增属性就可以发现数据的密度关系; 2) 采用改进的最大信息系数 (Rapid computation of the maximal information coefficient, RapidMic)[13−14] 来表示这些数据的密度关系, 通过计算数据之间的相关关系矩阵, 可以把基聚类结果转换成一个最大相关系数矩阵; 3) 改进 DP 算法, 确定聚类数量这一参数, 自动选取峰值最大的 K 个点作为聚类中心, 然后使用这个改进的 DP 算法对前面得到的最大相关系数矩阵进行聚类集成. 最后, 使用标准的数据集对所设计的聚类集成算法和经典的聚类集成算法、以及 K 均值 (K-means) 算法进行比较, 采用准确率和纯度值这两个评价指标对聚类集成结果进行评价, 并对评价结果进行统计检验等工作.

在上述研究内容中, 本文有两个方面的创新: 1) 对聚类结果进行分析, 使用 RapidMic 进行运算, 发现基聚类结果可以表示成为原始数据的密度关系, 通过这个密度关系可以得到一个最大相关系数矩阵; 2) 改进了 DP 算法, 增加了 DP 算法的一个参数 K, 然后使用这个改进的 DP 算法进行聚类集成.

1.3 本文内容安排

本文的其余部分安排如下. 第 2 节介绍聚类集成的相关工作. 第 3 节首先例证基聚类结果可以表示成为原始数据的密度关系, 随后介绍了基于改进的 DP 算法的聚类集成方法. 第 4 节展示实验步骤及实验结果. 第 5 节得出结论, 并介绍了未来可能的工作.

2 相关工作

近年来, 集成学习获得越来越多的关注. 大多数集成学习方法可分为三类: 监督学习、半监督学习以及无监督学习.

无监督学习方法的集成, 也就是聚类集成, 其基本思想是用多个独立的基聚类器分别对原始数据集进行聚类, 然后使用某种集成方法进行处理, 获得一个最终的集成结果[15]. 这类算法在第一阶段应尽可能地使用多种方式来获取基聚类结果, 第二阶段应选择一个最合适的集成解决方案来处理这些结果. 依据算法中解决问题的重点不同, 现有的聚类集成方法可大致分为三类.

第一类侧重于设计新的聚类集成方法. 为了改善最终单聚类算法的结果, Strehl 等[7] 提出基于三种集成技术的聚类集成框架. Fred 等[16] 提出了 EAC (Evidence accumulation clustering) 算法. Topchy 等[17] 提出了 EM (Expectation maximization) 算法并通过对比其他算法分析了其在聚类集成中的性能表现. Ayad 等[18] 提出了基于累积投票方法的聚类集成框架. Zheng 等[19] 设计出一种在分层聚类过程中考虑超度量距离分层的聚类集成框架. Wang 等[20] 在聚类集成框架中引入贝叶斯理论, 并设计出基于贝叶斯网络的聚类集成方法. 周林等[21] 提出了基于谱聚类的聚类集成算法, 既能利用谱聚类算法的优越性能, 又能避免精确选择尺度参数的问题. Banerjee 等[22] 提出了一种能够为 EM 算法产生更好聚类参数近似值的聚类集成算法. Lingras 等[23] 提出了一种基于粗糙集的聚类集成方法用以维护固有的聚类顺序. Wahid 等[24] 提出了一种基于 SPEA (Strength pareto evolutionary algorithm) 的新的多目标聚类集成方法: SPEA-II. 结合不同的聚类集成方法, Goswami 等[25] 提出了一种基于遗传算法的聚类集成算法. Wei 等[26] 设计出一个能够同时考虑成对约束和度量学习的半监督聚类集成框架. Liu 等[27] 提出一种选择性的聚类集成算法以及自适应加权策略. Hao 等[28] 设计的算法能够用以改善基于链接相似性度量的数据聚类关联矩阵. Huang 等[29] 通过在聚类集成中引入超对象的概念, 提出一种新的方法: ECFG (Ensemble clustering using factor graph).

第二类主要探索聚类集成方法的属性.

Kuncheva 等[30] 研究了聚类的多样性与准确性之间的关系. Kuncheva 等[31] 探索出如何利用合适的多样性提高聚类准确性. Topchy 等[32] 重点研究了聚类集成方法的收敛性. Amasyali 等[33] 就不同因素对聚类集成性能的影响进行了研究. Zhang 等[34] 提出了一个广义调整的兰德指数来衡量数据集中两个基分区之间的一致性. Wang[35] 设计出基于 CA 树的分层数据聚类结构, 可加速聚类形成, 提升聚类集成效率. 为了提高聚类集成算法的鲁棒性, Zhou 等[36] 提出在捕获到稀疏和对称的错误后, 将其整合到强大和一致的框架下用以学习低秩矩阵. Zhong 等[37] 认为证据积累是一种有效的框架能够将基分区转换为关联矩阵, 从而充分利用每个基分区的集群结构信息. Wahid 等[38] 研究出的聚类集成方法能够解决两个不同但相互关联的问题: 从数据集中产生多个聚类集成结果, 同时产生一个最终的聚类集成结果.

第三类聚类集成方法主要探索其应用领域. 通过检测基因表达数据集的基础聚类结构, Yu 等[39] 提出的聚类集成框架可用于发现癌症基因. Zhang 等[40] 提出的聚类集成方法可应用于 SAR 图像分割. Hu 等[41] 研究了如何使用聚类集成从基因表达数据集中确定基因簇的问题. 徐森等[42] 在聚类集成中引入谱聚类思想, 以解决文本聚类问题. Ye 等[43] 融合了聚类集成框架与领域知识, 用以实现恶意软件的自动分类. Zhang 等[44] 探索出基于聚类集成对流数据进行数据挖掘的方法. Yu 等[45] 借助新的聚类集成方法 BAE (Bagging-Adaboost ensemble) 实现了对真核细胞蛋白质磷酸化位点的预测. 在从基因表达数据集发现癌症的过程中, 为了降低噪声基因的影响, Yu 等[46] 提出两种新的共识聚类框架: 三谱聚类为基础的共识聚类 (SC3) 和双谱聚类为基础的共识聚类 (SC2). Ammour 等[47] 提出的聚类集成方法可应用于图像分割领域, 方法中包含了模糊 C 均值聚类 (Fuzzy C-means, FCM) 算法和具有不同邻居效应值的本地信息 FCM 算法 FCM_S1, FCM_S2. 为了解决大规模社会媒体网络中的隐身术检测问题, Li 等[48] 提出了高阶共同特征和聚类集成的方法. 受 Chameleon 理念的启发, Xiao 等[49] 设计出一种半监督的聚类集成模型用于高速列车行进过程中传动装置的故障诊断. Teng 等[50] 提出用基于数据处理分组方法的聚类集成框架 (Cluster ensemble framework based on the group method of data handling, CE-GMDH) 提升数据处理技术.

本文提出一种基于改进的 DP 算法的聚类集成模型, 获得基聚类结果后, 使用 RapidMic 衡量各基聚类结果之间的相关性, 通过计算得到最大相关系数矩阵后, 使用改进的 DP 算法进行聚类集成, 获得最终的聚类集成结果.

3 基于改进的 DP 算法的聚类集成

3.1 聚类集成问题

聚类集成可以分为两个步骤进行. 第一步是使用基聚类器对原始数据集进行多次聚类, 得到多个基聚类结果. 这一步可选择两种方式达成: 1) 使用某一种算法重复运算多次获得基聚类结果; 2) 选用多种不同的算法进行运算获得基聚类结果. 第二步是基聚类结果集成, 选取一种适当的聚类集成方法或者框架, 使之能够最大限度地分析这些结果, 得到一个对原始数据集最好的集成结果.

3.2 基聚类结果的产生

近邻传播 (Affinity propagation, AP)[51] 算法是 2007 年在 *Science* 上被提出的. 本文选用 AP 算法作为基聚类算法, 与其他算法不同, AP 算法不需要在一开始指定聚类个数, 所有的数据点均作为潜在的聚类中心. 通过计算原始数据集的相似度矩阵, 使用 AP 算法进行聚类, 产生基聚类结果. 假设原始数据集有 n 个数据点, 选用欧式距离作为相似度的测度指标, 则任意两点之间的相似度为两点距离平方的负数, 例如对于点 x_i 和点 x_k, 有 $G(i,k) = -\parallel x_i - x_k \parallel^2$. 通过计算所有数据点的相似度, 得到 $n \times n$ 维的相似度矩阵 G. AP 算法初始设定所有 $G(k,k)$ 为相同值 p. 通过参考度 p 的值来判断某个点是否能成为聚类中心, 参考度 p 直接影响了最终的聚类数量.

AP 算法传递两种类型的消息: 吸引度值 (Responsibility) 和归属度值 (Availability). 吸引度值 $r(i,k)$ 表示从点 i 发送到候选聚类中心 k 的数值消息, 反映了 k 点是否适合作为 i 点的聚类中心. 而归属度值 $a(i,k)$ 表示从候选聚类中心 k 发送到 i 的数值消息, 反映了 i 点是否选择 k 作为其聚类中心. $r(i,k)$ 与 $a(i,k)$ 越强, k 点作为聚类中心的可能性就越大, 并且 i 点隶属于以 k 点为聚类中心的聚类的可能性也越大. 算法运行过程中, 通过迭代过程不断更新每一个点的吸引度值和归属度值, 直到产生 K 个高质量的聚类中心, 随后将其余的数据点分配到相应的聚类中.

通过选取不同的 p 值, 重复使用 AP 算法计算次, 最终可获得 m 个不同的基聚类结果 $P = [P_1, P_2, P_3, \cdots, P_m]$. P 为一个 $n \times m$ 维的矩阵, 矩阵的每一行代表每一个数据点在 m 种不同

聚类算法中被分配到的类别标签, 而矩阵的每一列则代表一次基聚类运算的结果 (对应某个参考度 p).

3.3 基聚类结果的相似性矩阵

互信息 (Mutual information) 是信息论里一种有用的信息度量, 可以看成是一个随机变量中包含的关于另一个随机变量的信息量. 两个变量之间的互信息越大, 说明两者的相关性越大. 反之, 则越小. 衡量简单的离散变量之间的关联度大小可以通过计算互信息实现, 然而, 互信息无法衡量混合类型数据之间的相关性. 2011 年, 最大信息系数 (Maximal information coefficient, MIC) 被提出[13], MIC 能够对不同类型的庞大数据集进行关联关系的评估. 其算法原理是: 通过计算一个数据集中两两变量之间的互信息, 找出每两个变量之间互信息最大的值, 通过归一化后构成一个特征矩阵. 与互信息相比, MIC 有下面两大优势: 1) 除了能够对本身是离散型的数据进行处理以外, 还能够通过对连续型数据进行离散处理, 实现对混合类型数据的处理. 2) 通过构建互信息特征矩阵来寻找变量之间的最大信息系数, 可以更精确地表示出数据属性间关联性的大小.

本文使用改进的 MIC, 也就是 RapidMic[14] 算法计算基聚类结果的相关性. 基聚类过程中计算得到的 $n \times m$ 维的矩阵 P 包含所有的基聚类结果. P 中有 n 个数据点, 把基聚类结果看成是原始数据的新增属性, 每个数据点有 m 种属性. 计算新的数据对象之间的互信息. 可以用 $I(\alpha_i, \alpha_j)$ 表示 α_i 和 α_j 之间的互信息, 其中, $H(\alpha_i)$ 表示变量 α_i 的熵, $I(\alpha_i, \alpha_j)$ 的值没有上界限. 为了更好地比较变量之间的相关性, 可以采用标准化的 $I(\alpha_i, \alpha_j)$ 值, 范围在 0 到 1 之间, 0 表示两个变量互相独立, 而 1 表示两个变量有无噪的关系. 已有的几种标准化方法以 $I(\alpha_i, \alpha_j) \le \min(H(\alpha_i), H(\alpha_j))$ 为依据, 需要计算 $H(\alpha_i)$ 和 $H(\alpha_j)$ 的算术和几何平均数. 标准化后, 当 $i = j$ 时, $H(\alpha_i) = I(\alpha_i, \alpha_j)$, 且是在 Hilbert 空间中对数据进行标准化的, 所以一般更倾向于采用几何平均值的方法来对数据进行标准化. 归一化互信息 (Normalized mutual information, NMI) 公式如式 (1) 所示:

$$NMI(\alpha_i, \alpha_j) = \frac{I(\alpha_i, \alpha_j)}{\sqrt{H(\alpha_i)H(\alpha_j)}} \tag{1}$$

根据式 (1) 可以计算出新的数据对象之间的互信息大小. 在得到所有的互信息值后, 构建 $n \times n$ 维特征矩阵 S. 该相似性矩阵为一个对称阵, 主对角线上的值是属性自身的互信息, 值为 1, 其余的为两两属性间的互信息值, 即 $NMI(\alpha_i, \alpha_j)$.

3.4 基聚类结果的密度关系

Isomap 算法[52] 是一种非线性降维方法, 首先创建能够正确表达数据邻域结构的邻域图, 接着用最短路径法计算各数据点间的最短路径, 逼近相应的测地距离, 最后使用经典的多维标度分析 (Multi-dimensional scaling, MDS) 算法在低维可视空间重建数据.

本文使用 Iris 数据集进行例证, Iris 有 150 个数据点, 即 n 取值 150, 每个数据点 4 种维度. 实验中以 AP 算法作为基聚类器, 重复运算 30 次, 即 m 取值 30, 得到包含所有基聚类结果的 $n \times m$ 维的 P 矩阵. 接着使用 RapidMic 算法获得基聚类结果的相似性矩阵 S. 最后通过 Isomap 算法获得原始数据的基聚类结果的二维关系映射. 具体过程如图 1 所示, 最终结果如图 2 所示.

将这些基聚类结果从 150 维降到 2 维之后, 得到一个二维关系映射图, 观察图片可以发现, 基聚类结果的二维映射以某种规律聚成了几类. 由此推论, 如果把基聚类结果看成是原始数据的新增属性, 那么只使用这些新增属性就可以发现数据的密度关系.

3.5 改进基于密度峰值的 DP 算法用于聚类集成

DP 算法是 2014 年在 *Science* 上被提出的聚类

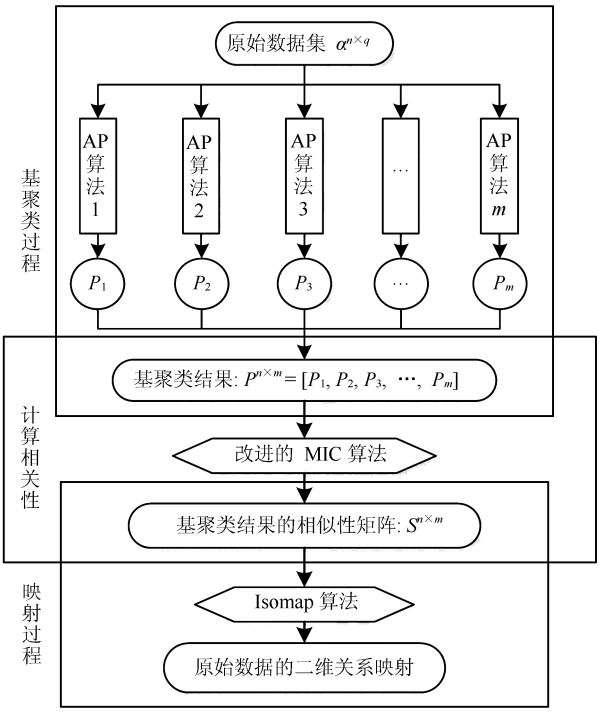


图 1 基于基聚类结果获取原始数据二维关系映射图的过程

Fig. 1 The process of obtaining the two-dimensional relational mapping of original dataset based on clustering results

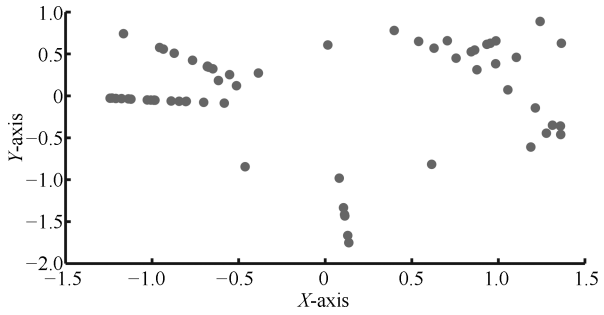


图 2　原始数据的基聚类结果的二维关系映射

Fig. 2　The two-dimensional relational mapping of the base clustering results of original dataset

算法, 算法有很好的鲁棒性并且对于各种数据集都能达到很好的聚类效果[12].

DP 算法基于的假设是: 类簇中心被具有较低局部密度的邻居点包围, 且与具有更高密度的任何点有相对较大的距离. 对于每一个数据点 i, 要计算两个量: 数据点的局部密度 ρ_i 和该点到具有更高局部密度的点的距离 δ_i, 而这两个值都取决于数据点间的距离 d_{ij}, d_c 是一个截断距离. 数据点 i 的局部密度 ρ_i 如式 (2) 所示:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{2}$$

其中, 令 $x = d_{ij} - d_c$, 如果 $x < 0$, 那么 $\chi(x) = 1$; 否则 $\chi(x) = 0$. 基本上, ρ_i 等于与点 i 的距离小于 d_c 的点的个数. 算法只对不同点的 ρ_i 的相对大小敏感, 这意味着对于大数据集, 分析结果对于 d_c 的选择有很好的鲁棒性.

δ_i 是数据点 i 到任何比其密度大的点的距离的最小值, 计算公式如式 (3) 所示:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{3}$$

对于密度最大的点, 我们可以得到 $\delta_i = \min_j(d_{ij})$.

只有具有高 δ 和相对较高 ρ 的点才是类簇中心. 而有高 δ 值和低 ρ 值的, 往往就是异常点. 类簇中心找到后, 剩余的每个点被归属到有更高密度的最近邻所属类簇. 类簇分配只需一步即可完成, 不像其他算法要对目标函数进行迭代优化.

本文改进了经典的 DP 算法, 在原有的算法中增加了一个参数 K, 具体的修改方法见算法 1. 经过改进之后, 算法能够自动选取拥有最大密度峰值的前 K 个点作为聚类中心. 接着使用这个改进的 DP 算法进行聚类集成, 将相似性矩阵 $S^{n\times n}$ 作为目标数据输入算法, 计算出局部密度 ρ_i 和该点到具有更高局部密度的点的距离 δ_i, 自动选取具有高 δ 和相对较高 ρ 的点作为类簇中心, 将其余的数据点归到各个类簇, 要求分配到的类簇与比其密度高并且与其距离近的节点所属的类一样. 最终的聚类集成结果可以看成是针对原始数据集所获得的最好聚类结果. 整个聚类集成过程如图 3 所示.

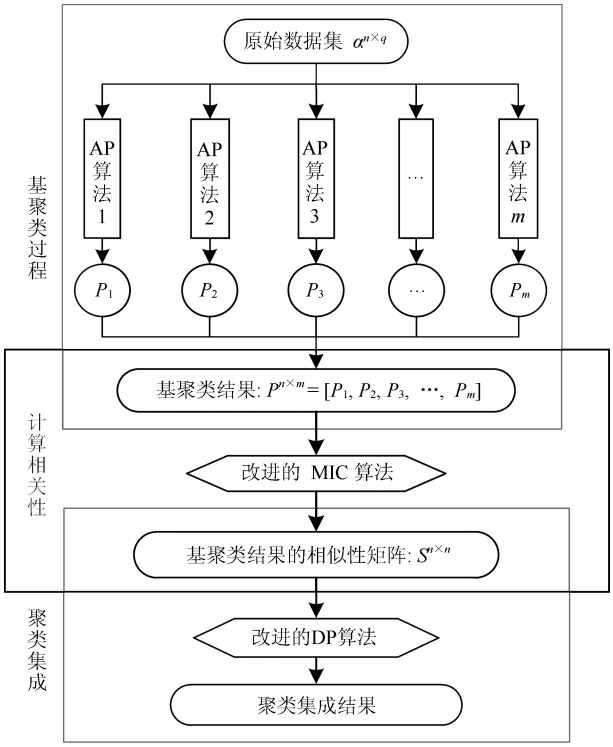


图 3　基于改进的 DP 算法的聚类集成过程

Fig. 3　The process of cluster ensembling based on improved DP algorithm

3.6　算法描述

根据图 3 展示的聚类集成过程, 算法 1 描述如下:

算法 1. DP_ensemble

输入. 实验数据集 $\alpha^{n\times q}$, 基聚类器运算次数 m, 实验数据集的总类簇数 K

输出. α 的聚类集成标签 L

步骤 1. 重复运算 m 次 AP 算法获得数据集 α 的基聚类结果 $P^{n\times m}$.

步骤 2. 根据式 (1) 使用 RapidMic 计算 $P^{n\times m}$ 的相似性矩阵 $S^{n\times n}$.

步骤 3. 使用改进的 DP 算法对 $S^{n\times n}$ 进行聚类集成:

1) 根据式 (2) 和 (3) 计算 S 中每个数据点的局部密度 ρ_i 和该点到具有更高局部密度的点的距离 δ_i (本文使用的是 RapidMic 算法, 取值范围是 0 到 1. 相似度越大, 表示距离越近, 具体转换关系是: 距离 = 1 − 相似度);

2) 自动选取具有高 δ_i 和相对较高 ρ_i 的前 K 个点作为类簇中心;

3) 对除类簇中心外的所有数据点进行划分, 获得 α 的聚类集成标签 L.

4 实验

4.1 数据集和评价标准

本文使用 UCI (University of California Irvine) 机器学习库中的 20 个数据集作为实验数据集. 表 1 列出了数据集的样本、属性和类别数量. 有很多标准可以用来衡量聚类集成算法, 本文以这些数据的真实类别标签为标准, 选用 Micro-precision (MP)[53−54] 标准和 Purity[55] 标准来衡量聚类结果的好坏.

表 1 实验数据集的样本、属性和类别数量

Table 1 The number of instances, features and classes of datasets

ID	Datasets	Number of instances	Number of features	Number of classes
1	Aerosol	905	892	3
2	Amber	880	892	3
3	Ambulances	930	892	3
4	Aquarium	922	892	3
5	Balloon	830	892	3
6	Banner	860	892	3
7	Baobab	900	892	3
8	Basket	892	892	3
9	Bateau	900	892	3
10	Bathroom	924	892	3
11	Bed	888	892	3
12	Beret	876	892	3
13	Beverage	873	892	3
14	Bicycle	844	892	3
15	Birthdaycake	932	892	3
16	Blog	943	892	3
17	Blood	866	892	3
18	Boat	857	892	3
19	Bonbon	874	892	3
20	Bonsai	867	892	3

MP 标准的计算如式 (4) 所示:

$$\mathrm{MP} = \frac{1}{n}\sum_{h=1}^{K} a_h \tag{4}$$

其中, a_h 表示对数据某一类分类正确的数量, n 表示数据集中数据对象的数量, K 表示此数据集中类别的数量. MP 值越大, 代表聚类的准确率越高. 为了获得更好的准确率需要进行重复的实验, 采用平均 MP 值来衡量结果将更精确. 具体计算公式如式 (5) 所示:

$$\mathrm{AMP} = \frac{1}{m \times n}\sum_{t=1}^{m}\sum_{h=1}^{K} a_h \tag{5}$$

其中, m 为重复实验的次数, 本文在基聚类中 m 为 10, 在聚类集成中 m 为 3.

Purity 标准的计算如式 (6) 所示:

$$\mathrm{Purity} = \frac{1}{n}\sum_{k=1}^{r}\max_{1<l<q} n_k^l \tag{6}$$

其中, n_k^l 是原始类簇的样本数. 一个较大的纯度值代表较好的聚类性能.

4.2 实验步骤和实验结果

本文选取 AP 作为基聚类器算法, 通过选取 10 个不同的参考度 p 值, 使用 AP 算法重复运算 10 次, 获得基聚类结果 $P = [P_1, P_2, P_3, \cdots, P_{10}]$. 随后使用 CSPA (Cluster-based similarity partitioning algorithm)[7]、HGPA (Hypergraph partitioning algorithm)[7]、MCLA (Meta-clustering algorithm)[7]、DP、EM 和 QMI (Quadratic mutual information)[9] 六种算法对基聚类结果的相似性矩阵进行聚类集成, 获得最终的聚类集成标签. 最后将这些标签与数据集的真实标签进行对比, 采用准确率和纯度值标准进行评价.

此外, 同时使用经典的 K-means 算法对数据集进行聚类, 获取标签, 对比真实标签后, 采用准确率和纯度值标准进行评价. 使用 K-means 与各聚类集成算法进行比照实验, 是为了证明本文所提出的聚类集成算法是有效的, 实际上并没有太大的意义, 因为聚类集成更关注在已有的基聚类基础上的效果的提升, 而非进行单一聚类算法的比较.

表 2 展示了实验结果的准确率及其标准差. 表中第一列数据给出了实验中用到的数据集 ID. 第二列和第三列给出了使用基聚类器 AP 算法对 20 个数据集进行 10 次聚类之后得到的平均准确率 AP-average、最大准确率 AP-max, 之后的六列则依次给出使用 CSPA、HGPA、MCLA、DP、EM 和 QMI 六种聚类算法对基聚类结果的相似性矩阵进行集成运算, 以及使用 K-means 进行聚类获得的准确率. 从表 2 可以观测得到以下结论:

1) 从 AP-average 和 AP-max 两列数据可以看出, 使用 AP 算法对 20 个数据集进行聚类运算, 准确率极低, 平均准确率仅为 0.023. 正确率极低的原因在于, AP 算法产生的聚类标签与真实的标签很不一样, 某种程度上来说, 这两列结果是无意义的.

表 2 平均准确率和标准差 (每个数据集的最大准确率加粗显示.)

Table 2 Average MPs and standard deviations (The highest MP among different algorithms on each dataset is bolded.)

ID	AP-average	AP-max	CSPA	HGPA	MCLA	DP	EM	QMI	K-means
1	0.022 ± 0.015	0.113 ± 0.072	0.379 ± 0.020	0.384 ± 0.003	0.395 ± 0.020	$\mathbf{0.484 \pm 0.019}$	0.354 ± 0.004	0.466 ± 0.053	0.369 ± 0.008
2	0.023 ± 0.016	0.111 ± 0.054	0.472 ± 0.045	0.502 ± 0.020	0.493 ± 0.004	0.571 ± 0.001	0.387 ± 0.025	0.526 ± 0.050	$\mathbf{0.595 \pm 0.008}$
3	0.026 ± 0.016	0.119 ± 0.056	0.392 ± 0.026	0.394 ± 0.021	0.384 ± 0.008	$\mathbf{0.596 \pm 0.041}$	0.370 ± 0.022	0.497 ± 0.019	0.442 ± 0.029
4	0.021 ± 0.013	0.095 ± 0.040	0.401 ± 0.013	0.394 ± 0.027	0.381 ± 0.026	$\mathbf{0.653 \pm 0.042}$	0.353 ± 0.011	0.580 ± 0.057	0.361 ± 0.006
5	0.026 ± 0.020	0.161 ± 0.118	0.410 ± 0.037	0.468 ± 0.017	0.393 ± 0.009	$\mathbf{0.554 \pm 0.010}$	0.358 ± 0.032	0.541 ± 0.035	0.445 ± 0.004
6	0.027 ± 0.017	0.124 ± 0.058	0.346 ± 0.002	0.365 ± 0.010	0.346 ± 0.004	$\mathbf{0.805 \pm 0.158}$	0.358 ± 0.004	0.791 ± 0.112	0.462 ± 0.001
7	0.018 ± 0.015	0.108 ± 0.098	0.432 ± 0.026	0.474 ± 0.008	0.427 ± 0.017	$\mathbf{0.534 \pm 0.068}$	0.389 ± 0.050	0.482 ± 0.024	0.503 ± 0.004
8	0.022 ± 0.017	0.133 ± 0.099	0.362 ± 0.018	0.394 ± 0.020	0.394 ± 0.008	$\mathbf{0.538 \pm 0.025}$	0.357 ± 0.023	0.483 ± 0.049	0.409 ± 0.000
9	0.028 ± 0.018	0.135 ± 0.066	0.401 ± 0.020	0.445 ± 0.039	0.423 ± 0.023	$\mathbf{0.511 \pm 0.050}$	0.367 ± 0.017	0.510 ± 0.020	0.441 ± 0.003
10	0.019 ± 0.012	0.088 ± 0.045	0.351 ± 0.014	0.369 ± 0.001	0.361 ± 0.014	$\mathbf{0.756 \pm 0.059}$	0.355 ± 0.011	0.662 ± 0.031	0.394 ± 0.001
11	0.035 ± 0.021	0.173 ± 0.045	0.371 ± 0.002	0.401 ± 0.025	0.382 ± 0.024	$\mathbf{0.617 \pm 0.046}$	0.347 ± 0.006	0.542 ± 0.029	0.459 ± 0.004
12	0.020 ± 0.012	0.101 ± 0.048	0.368 ± 0.005	0.372 ± 0.018	0.373 ± 0.017	$\mathbf{0.629 \pm 0.015}$	0.354 ± 0.007	0.575 ± 0.094	0.417 ± 0.009
13	0.031 ± 0.023	0.173 ± 0.101	0.419 ± 0.014	0.425 ± 0.019	0.400 ± 0.013	$\mathbf{0.531 \pm 0.027}$	0.353 ± 0.004	0.511 ± 0.015	0.396 ± 0.000
14	0.020 ± 0.015	0.113 ± 0.089	0.410 ± 0.020	0.412 ± 0.025	0.407 ± 0.006	$\mathbf{0.522 \pm 0.017}$	0.357 ± 0.008	0.478 ± 0.028	0.452 ± 0.008
15	0.017 ± 0.013	0.092 ± 0.063	0.392 ± 0.044	0.446 ± 0.032	0.450 ± 0.014	$\mathbf{0.576 \pm 0.008}$	0.372 ± 0.005	0.563 ± 0.042	0.491 ± 0.001
16	0.023 ± 0.015	0.119 ± 0.056	0.347 ± 0.005	0.383 ± 0.010	0.369 ± 0.006	$\mathbf{0.685 \pm 0.069}$	0.357 ± 0.003	0.627 ± 0.077	0.411 ± 0.001
17	0.018 ± 0.011	0.088 ± 0.025	0.349 ± 0.011	0.352 ± 0.013	0.375 ± 0.003	$\mathbf{0.776 \pm 0.059}$	0.354 ± 0.017	0.687 ± 0.095	0.473 ± 0.004
18	0.022 ± 0.014	0.106 ± 0.052	0.388 ± 0.007	0.368 ± 0.020	0.377 ± 0.012	$\mathbf{0.587 \pm 0.038}$	0.354 ± 0.008	0.537 ± 0.025	0.404 ± 0.001
19	0.016 ± 0.011	0.090 ± 0.049	0.397 ± 0.022	0.411 ± 0.019	0.402 ± 0.012	$\mathbf{0.508 \pm 0.012}$	0.357 ± 0.013	0.481 ± 0.020	0.465 ± 0.002
20	0.021 ± 0.012	0.095 ± 0.032	0.383 ± 0.005	0.385 ± 0.024	0.355 ± 0.011	$\mathbf{0.642 \pm 0.045}$	0.380 ± 0.039	0.587 ± 0.051	0.443 ± 0.003
AVG	0.023 ± 0.015	0.117 ± 0.063	0.389 ± 0.018	0.407 ± 0.019	0.394 ± 0.012	$\mathbf{0.604 \pm 0.040}$	0.362 ± 0.016	0.556 ± 0.046	0.442 ± 0.005

2) 从 CSPA、HGPA、MCLA、DP、EM 和 QMI 六列数据可以观测到, 这六种算法的聚类集成准确率都超过 0.340. 说明聚类集成算法可以获得比单纯聚类更好的结果.

3) 最高的准确率已在表中加粗显示, 在所有的聚类集成算法中, DP 获得的准确率最高, 平均准确率超过 0.600. 准确率越高, 说明聚类效果越好, 实验中, DP 获得了最好的聚类效果.

4) 表 2 中各准确率的标准差相对都比较小, 绝大部分不超过 0.100. 标准差越小, 说明结果越稳定, 实验中, 各算法均具有良好的稳定性.

5) 对比 K-means 和 DP 两列数据可以发现, 比起单纯使用 K-means 进行聚类运算, 使用 DP 进行聚类集成能够获得更高的准确率.

表 3 展示了实验结果的纯度值及其标准差. 表中第一列数据给出了实验中用到的数据集 ID. 第二列和第三列给出了使用基聚类器 AP 算法对 20 个数据集进行 10 次聚类之后得到的平均纯度值 AP-average、最大纯度值 AP-max, 之后的六列则依次给出使用 CSPA、HGPA、MCLA、DP、EM 和 QMI 六种聚类算法对基聚类结果的相似性矩阵进行集成运算, 以及使用 K-means 进行聚类获得的纯度值.

从表 3 可以观测得到以下结论:

1) 从 AP-average 和 AP-max 两列数据可以看到, 使用 AP 算法对 20 个数据集进行聚类获得的平均纯度值较低, 仅为 0.277, 而最大纯度值却很高, 达到 0.672. 纯度值不高, 且波动很大的原因在于 AP 算法产生的聚类标签与真实的标签很不一样, 某种程度上来说, 这两列结果是无意义的.

2) 从 CSPA、HGPA、MCLA、DP、EM 和 QMI 六列数据可以观测到, 这六种算法的聚类集成纯度值远远高于基聚类结果的平均纯度值, 说明聚类集成算法可以获得比单纯聚类算法更好的结果.

3) 最高的纯度值已在表中加粗显示, 在所有的聚类集成算法中, DP 和 EM 获得的纯度值最高, 两者在 20 个数据集上的平均纯度值均超过 0.740. 纯度值越高, 说明聚类效果越好. 实验中, DP 和 EM 能够获得较好的聚类效果.

4) 表中除了 AP-average 和 AP-max 的标准差比较大, 六种聚类集成算法的纯度值的标准差都比较小, 特别是 DP 和 EM 两种算法, 平均标准差不到 0.010, 标准差越小, 说明结果越稳定. 实验中, 聚

表 3 平均纯度值和标准差 (每个数据集的最大纯度值加粗显示.)

Table 3 Average purities and standard deviations (The highest purity among different algorithms on each dataset is bolded.)

ID	AP-average	AP-max	CSPA	HGPA	MCLA	DP	EM	QMI	K-means
1	0.315 ± 0.159	0.743 ± 0.251	0.796 ± 0.006	0.796 ± 0.004	0.795 ± 0.004	0.797 ± 0.005	$\mathbf{0.803 \pm 0.000}$	0.790 ± 0.009	0.795 ± 0.002
2	0.233 ± 0.127	0.591 ± 0.231	0.704 ± 0.046	0.667 ± 0.011	0.684 ± 0.007	$\mathbf{0.769 \pm 0.007}$	0.764 ± 0.007	0.741 ± 0.017	0.633 ± 0.005
3	0.274 ± 0.149	0.707 ± 0.290	0.753 ± 0.009	0.755 ± 0.005	0.755 ± 0.002	$\mathbf{0.767 \pm 0.001}$	0.764 ± 0.003	0.754 ± 0.012	0.749 ± 0.013
4	0.252 ± 0.137	0.651 ± 0.244	0.702 ± 0.013	0.706 ± 0.006	0.711 ± 0.008	0.717 ± 0.003	$\mathbf{0.719 \pm 0.002}$	0.707 ± 0.003	0.697 ± 0.002
5	0.316 ± 0.170	0.802 ± 0.317	0.863 ± 0.016	0.843 ± 0.007	0.868 ± 0.003	0.878 ± 0.004	$\mathbf{0.881 \pm 0.004}$	0.876 ± 0.005	0.840 ± 0.008
6	0.079 ± 0.043	0.202 ± 0.077	0.215 ± 0.001	0.212 ± 0.000	0.215 ± 0.002	$\mathbf{0.216 \pm 0.002}$	$\mathbf{0.216 \pm 0.001}$	$\mathbf{0.216 \pm 0.000}$	0.213 ± 0.000
7	0.270 ± 0.140	0.678 ± 0.240	0.764 ± 0.007	0.761 ± 0.001	0.779 ± 0.005	$\mathbf{0.806 \pm 0.004}$	0.801 ± 0.010	0.800 ± 0.007	0.764 ± 0.002
8	0.336 ± 0.179	0.809 ± 0.325	0.861 ± 0.002	0.853 ± 0.005	0.857 ± 0.001	$\mathbf{0.868 \pm 0.001}$	0.866 ± 0.002	0.861 ± 0.006	0.842 ± 0.001
9	0.339 ± 0.177	0.775 ± 0.300	0.836 ± 0.023	0.839 ± 0.013	0.825 ± 0.012	$\mathbf{0.866 \pm 0.002}$	0.864 ± 0.005	0.849 ± 0.003	0.837 ± 0.003
10	0.239 ± 0.129	0.551 ± 0.223	0.575 ± 0.003	0.578 ± 0.002	0.576 ± 0.000	$\mathbf{0.581 \pm 0.000}$	0.580 ± 0.002	0.576 ± 0.002	0.578 ± 0.000
11	0.345 ± 0.177	0.726 ± 0.284	0.769 ± 0.010	0.752 ± 0.003	0.775 ± 0.006	0.782 ± 0.004	$\mathbf{0.786 \pm 0.001}$	0.775 ± 0.009	0.752 ± 0.007
12	0.250 ± 0.134	0.658 ± 0.237	0.716 ± 0.003	0.717 ± 0.003	0.718 ± 0.001	0.721 ± 0.003	$\mathbf{0.722 \pm 0.001}$	0.718 ± 0.003	0.705 ± 0.004
13	0.325 ± 0.166	0.731 ± 0.274	0.776 ± 0.007	0.777 ± 0.004	0.786 ± 0.003	$\mathbf{0.801 \pm 0.000}$	0.800 ± 0.001	0.789 ± 0.014	0.766 ± 0.000
14	0.325 ± 0.166	0.731 ± 0.274	0.776 ± 0.007	0.777 ± 0.004	0.786 ± 0.003	0.801 ± 0.000	0.800 ± 0.001	0.789 ± 0.014	$\mathbf{0.877 \pm 0.003}$
15	0.289 ± 0.150	0.713 ± 0.270	0.820 ± 0.027	0.779 ± 0.012	0.784 ± 0.020	$\mathbf{0.839 \pm 0.001}$	$\mathbf{0.839 \pm 0.000}$	0.818 ± 0.015	0.735 ± 0.002
16	0.265 ± 0.141	0.654 ± 0.265	0.680 ± 0.002	0.674 ± 0.002	0.675 ± 0.002	0.680 ± 0.003	$\mathbf{0.681 \pm 0.001}$	0.679 ± 0.002	0.673 ± 0.000
17	0.171 ± 0.090	0.425 ± 0.158	0.465 ± 0.000	0.460 ± 0.006	0.464 ± 0.001	$\mathbf{0.471 \pm 0.001}$	$\mathbf{0.471 \pm 0.001}$	0.468 ± 0.001	0.459 ± 0.001
18	0.306 ± 0.167	0.784 ± 0.329	0.817 ± 0.006	0.823 ± 0.005	0.817 ± 0.005	0.827 ± 0.002	$\mathbf{0.828 \pm 0.001}$	0.824 ± 0.002	0.807 ± 0.002
19	0.316 ± 0.167	0.807 ± 0.316	0.845 ± 0.010	0.844 ± 0.006	0.848 ± 0.005	0.862 ± 0.002	$\mathbf{0.863 \pm 0.002}$	0.861 ± 0.000	0.835 ± 0.001
20	0.295 ± 0.156	0.709 ± 0.276	0.750 ± 0.001	0.750 ± 0.007	0.755 ± 0.001	$\mathbf{0.759 \pm 0.001}$	0.755 ± 0.003	0.752 ± 0.004	0.752 ± 0.000
AVG	0.277 ± 0.146	0.672 ± 0.259	0.724 ± 0.010	0.718 ± 0.005	0.724 ± 0.005	$\mathbf{0.740 \pm 0.002}$	$\mathbf{0.740 \pm 0.002}$	0.732 ± 0.006	0.715 ± 0.003

类集成算法具有良好的稳定性.

5) 对比 K-means 和 DP 两列数据可以发现, 比起单纯使用 K-means 进行聚类运算, 使用 DP 进行聚类集成能够获得更高的纯度值.

为了对六种聚类集成算法做一个更详细的比较, 本文使用 Friedman aligned ranks[56] 方法来统计和检验纯度值.

表 4 展示了六种算法针对 20 个数据集的调整后观察值, 括号中是对应的调整秩次. 表中第一列数据给出了实验中用到的数据集 ID. 之后的六列则依次给出使用 CSPA、HGPA、MCLA、DP、EM 和 QMI 六种聚类集成算法的调整后观察值.

从表 4 中可以观察到: EM 以 24.25 的平均指标排名第一; DP 以 25.55 排名第二, 由于 DP 与 EM 的纯度值实验结果非常接近 (两种算法均占据 11 个最佳结果, 且平均值相等), 因此使用 Friedman aligned ranks 测试计算得到的调整后观察值也非常接近, 差距非常小, 仅为 1.30; QMI 以 51.85 位列第三, 与第一和第二的差距较大; MCLA、CSPA、HGPA 分别以 80.65、86.25、94.45 排在第四、五、六名. 使用这一方法进行测试的目的是检查测得的调整秩次的总和是否与预期的对准零假设下的总的调整秩次 $\widehat{R}_j = 1\,210$ 有显著不同. 从全局对比来看, DP 的调整秩次与 EM 非常接近, 因此可以近似认为 EM 和 DP 调整秩次均最小, 调整秩次越小, 结果越好.

在 20 个数据集和 6 种算法中, T 依据 $6-1=5$ 自由度的卡方分布而分布, T 值计算方法如式 (7) ∼ (9) 所示. 使用 $\chi^2(5)$ 分布求得的单尾检验概率为 0.00000001, 求得的双尾检验的概率也为 0.00000001. 一般来说, 如果概率值大于 0.05, 可认为差异不显著. 实验中, 单尾和双尾检验的概率值都远远小于 0.05, 因此差异非常显著, 拒绝零假设.

$$\sum_{j=1}^{k} \widehat{R}_{\cdot,j}^2 = 1\,725^2 + 1\,889^2 + 1\,613^2 + 511^2 + 485^2 + 1\,037^2 = 10\,717\,430 \tag{7}$$

表 4 实验选择的 6 种算法的调整后观察值 (括号中的调整秩次用于 Friedman 调整秩和检验的计算. 最小代表最好.)

Table 4 Aligned observations of six algorithms selected in the experimental study (The ranks in the parentheses are used in the computation of the Friedman aligned ranks test. The smallest one is the best.)

ID	CSPA	HGPA	MCLA	DP	EM	QMI	Total
1	0.000 (68)	0.000 (60)	−0.001 (70)	0.001 (57)	0.007 (26)	−0.006 (97)	378
2	−0.017 (111)	−0.054 (120)	−0.038 (119)	0.048 (1)	0.042 (2)	0.019 (7)	360
3	−0.005 (93)	−0.003 (83)	−0.002 (79)	0.009 (21)	0.006 (29)	−0.004 (88)	393
4	−0.009 (104)	−0.004 (91)	0.001 (55)	0.007 (28)	0.009 (19)	−0.003 (84)	381
5	−0.005 (94)	−0.025 (116)	0.000 (65)	0.01 (17)	0.013 (11)	0.007 (23)	326
6	0.000 (63)	−0.003 (80)	0.000 (61)	0.001 (56)	0.001 (52)	0.000 (58)	370
7	−0.021 (112)	−0.025 (115)	−0.007 (99)	0.021 (5)	0.016 (9)	0.015 (10)	350
8	0.000 (67)	−0.008 (102)	−0.004 (90)	0.007 (24)	0.005 (32)	0.000 (62)	377
9	−0.010 (106)	−0.008 (101)	−0.022 (114)	0.019 (6)	0.018 (8)	0.003 (40)	375
10	−0.003 (82)	0.000 (59)	−0.002 (74)	0.004 (37)	0.002 (41)	−0.001 (71)	364
11	−0.004 (92)	−0.021 (113)	0.002 (42)	0.009 (18)	0.013 (12)	0.002 (43)	320
12	−0.003 (81)	−0.001 (73)	0.000 (66)	0.002 (44)	0.003 (38)	−0.001 (69)	371
13	−0.012 (109.5)	−0.011 (107.5)	−0.002 (77.5)	0.013 (13.5)	0.012 (15.5)	0.001 (53.5)	377
14	−0.012 (109.5)	−0.011 (107.5)	−0.002 (77.5)	0.013 (13.5)	0.012 (15.5)	0.001 (53.5)	377
15	0.007 (27)	−0.034 (118)	−0.029 (117)	0.026 (4)	0.026 (3)	0.005 (33)	302
16	0.001 (49)	−0.004 (89)	−0.003 (87)	0.002 (45)	0.003 (39)	0.001 (50)	359
17	−0.001 (72)	−0.007 (100)	−0.002 (76)	0.004 (35)	0.004 (36)	0.001 (48)	367
18	−0.006 (96)	0.000 (64)	−0.006 (95)	0.004 (34)	0.005 (31)	0.002 (47)	367
19	−0.008 (103)	−0.010 (105)	−0.006 (98)	0.008 (22)	0.009 (20)	0.007 (25)	373
20	−0.003 (86)	−0.003 (85)	0.001 (51)	0.006 (30)	0.002 (46)	−0.002 (75)	373
Total	1 725	1 889	1 613	**511**	**485**	1 037	
AVG	86.25	94.45	80.65	**25.55**	**24.25**	51.85	

$$\sum_{i=1}^{k} \widehat{R}_{i,\cdot}^2 = 378^2 + 360^2 + 393^2 + \cdots + 367^2 + 373^2 + 373^2 = 2\,645\,040 \tag{8}$$

$$T = \frac{(6-1)(10\,717\,430 - (6\times 20^2/4)(6\times 20+1)^2)}{((6\times 20(6\times 20+1)(2\times 6\times 20+1)) - 2\,645\,040)/6} = 314.669 \tag{9}$$

5 结论

本文提出一种基于改进的 DP 算法的聚类集成模型. 在 20 个实验数据集上与单纯的 AP 算法、CSPA、HGPA、MCLA、EM 和 QMI 聚类集成算法, 以及 K-means 算法进行对比实验后发现, 用改进的 DP 聚类集成算法处理数据集, 可以获得最高的聚类准确率、纯度值, 以及相对较小的调整秩次. 尽管 EM 和 DP 的纯度值都很高, DP 的调整秩次与 EM 也非常接近, 但是在准确率上, DP 远远高于 EM. 因此综合来说, 基于改进的 DP 算法的聚类集成效果可以认为是最佳的. 下一步的工作重点是尝试将半监督信息加入 DP 算法中, 研究基于改进的 DP 算法的半监督聚类集成方法.

References

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

2 Zhou Chen-Xi, Liang Xun, Qi Jin-Shan. A semi-supervised agglomerative hierarchical clustering method based on dynamically updating constraints. *Acta Automatica Sinica*, 2015, **41**(7): 1253−1263
(周晨曦, 梁循, 齐金山. 基于约束动态更新的半监督层次聚类算法. 自动化学报, 2015, **41**(7): 1253−1263)

3 Chen Jin-Yin, He Hui-Hao. Research on density-based clustering algorithm for mixed data with determine cluster centers automatically. *Acta Automatica Sinica*, 2015, **41**(10): 1798−1813
(陈晋音, 何辉豪. 基于密度的聚类中心自动确定的混合属性数据聚类算法研究. 自动化学报, 2015, **41**(10): 1798−1813)

4 Wang Wei-Wei, Li Xiao-Ping, Feng Xiang-Chu, Wang Si-Qi. A survey on sparse subspace clustering. *Acta Automatica Sinica*, 2015, **41**(8): 1373−1384

(王卫卫, 李小平, 冯象初, 王斯琪. 稀疏子空间聚类综述. 自动化学报, 2015, **41**(8): 1373−1384)

5 Taşdemir K, Moazzen Y, Yildirim I. An approximate spectral clustering ensemble for high spatial resolution remote-sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2015, **8**(5): 1996−2004

6 Parvin H, Minaei-Bidgoli B. A clustering ensemble framework based on selection of fuzzy weighted clusters in a locally adaptive clustering algorithm. *Pattern Analysis and Applications*, 2015, **18**(1): 87−112

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

8 Gionis A, Mannila H, Tsaparas P. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2007, **1**(1): Article No. 4

9 Topchy A, Jain A K, Punch W. Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(12): 1866−1881

10 Bhattacharjee A, Richards W G, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark E J, Lander E S, Wong W, Johnson B E, Golub T R, Sugarbaker D J, Meyerson M. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 2001, **98**(24): 13790−13795

11 Lindblad-Toh K, Tanenbaum D M, Daly M J, Winchester E, Lui W O, Villapakkam A, Stanton S E, Larsson C, Hudson T J, Johnson B E, Lander E S, Meyerson M. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnology*, 2000, **18**(9): 1001−1005

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

13 Reshef D N, Reshef Y A, Finucane H K, Grossman S R, McVean G, Turnbaugh P J, Lander E S, Mitzenmacher M, Sabeti P C. Detecting novel associations in large data sets. *Science*, 2011, **334**(6062): 1518−1524

14 Tang D M, Wang M W, Zheng W F, Wang H J. RapidMic: rapid computation of the maximal information coefficient. *Evolutionary Bioinformatics Online*, 2014, **10**: 11−16

15 Tang Wei, Zhou Zhi-Hua. Bagging-based selective clusterer ensemble. *Journal of Software*, 2005, **16**(4): 496−502
(唐伟, 周志华. 基于 Bagging 的选择性聚类集成. 软件学报, 2005, **16**(4): 496−502)

16 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

17 Topchy A, Jain A K, Punch W. A mixture model for clustering ensembles. In: Proceedings of the 2004 SIAM International Conference on Data Mining. Florida, USA: SIAM, 2004.

18 Ayad H G, Kamel M S. On voting-based consensus of cluster ensembles. *Pattern Recognition*, 2010, **43**(5): 1943−1953

19 Zheng L, Li T, Ding C. Hierarchical ensemble clustering. In: Proceedings of the 10th IEEE International Conference on Data Mining (ICDM). Sydney, NSW: IEEE, 2010. 1199−1204

20 Wang H J, Shan H H, Banerjee A. Bayesian cluster ensembles. *Statistical Analysis and Data Mining*, 2011, **4**(1): 54−70

21 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)

22 Banerjee B, Bovolo F, Bhattacharya A, Bruzzone L, Chaudhuri S, Mohan B K. A new self-training-based unsupervised satellite image classification technique using cluster ensemble strategy. *IEEE Geoscience and Remote Sensing Letters*, 2015, **12**(4): 741−745

23 Lingras P, Haider F. Partially ordered rough ensemble clustering for multigranular representations. *Intelligent Data Analysis*, 2015, **19**(s1): S103−S116

24 Wahid A, Gao X Y, Andreae P. Multi-objective clustering ensemble for high-dimensional data based on strength pareto evolutionary algorithm (SPEA-II). In: Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Paris: IEEE, 2015. 1−9

25 Goswami J P, Mahanta A K. A genetic algorithm based ensemble approach for categorical data clustering. In: Proceedings of the 2015 Annual IEEE India Conference (INDICON). New Delhi, India: IEEE, 2015. 1−6

26 Wei S T, Li Z X, Zhang C L. A semi-supervised clustering ensemble approach integrated constraint-based and metric-based. In: Proceedings of the 7th International Conference on Internet Multimedia Computing and Service. New York, USA: ACM, 2015. Article No. 26

27 Liu L M, Liao Z F, Liao Z N. An efficient clustering ensemble selection algorithm. *International Journal of Autonomous and Adaptive Communications Systems*, 2015, **8**(2−3): 200−212

28 Hao Z F, Wang L J, Cai R C, Wen W. An improved clustering ensemble method based link analysis. *World Wide Web*, 2015, **18**(2): 185−195

29 Huang D, Lai J H, Wang C D. Ensemble clustering using factor graph. *Pattern Recognition*, 2016, **50**: 131−142

30 Kuncheva L I, Whitaker C J. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 2003, **51**(2): 181−207

31 Kuncheva L I, Hadjitodorov S T. Using diversity in cluster ensembles. In: Proceedings of the 2004 IEEE International Conference on Systems, Man and Cybernetics. The Hague: IEEE, 2004. 1214−1219

32 Topchy A P, Law M H C, Jain A K, Fred A L. Analysis of consensus partition in cluster ensemble. In: Proceedings of the 4th IEEE International Conference on Data Mining. Brighton, UK: IEEE, 2004. 225−232

33 Amasyali M F, Ersoy O. The performance factors of clustering ensembles. In: Proceedings of the 16th IEEE Communication and Applications Conference on Signal Processing. Aydin: IEEE, 2008. 1−4

34 Zhang S H, Wong H S. ARImp: a generalized adjusted rand index for cluster ensembles. In: Proceedings of the 20th International Conference on Pattern Recognition (ICPR). Istanbul: IEEE, 2010. 778−781

35 Wang T. CA-Tree: a hierarchical structure for efficient and scalable coassociation-based cluster ensembles. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2011, **41**(3): 686−698

36 Zhou P, Du L, Wang H M, Shi L, Shen Y D. Learning a robust consensus matrix for clustering ensemble via Kullback-Leibler divergence minimization. In: Proceedings of the 24th International Conference on Artificial Intelligence. Washington, USA: AAAI, 2015. 4112−4118

37 Zhong C M, Yue X D, Zhang Z H, Lei J S. A clustering ensemble: two-level-refined co-association matrix with path-based transformation. *Pattern Recognition*, 2015, **48**(8): 2699−2709

38 Wahid A, Gao X Y, Andreae P. Multi-objective multi-view clustering ensemble based on evolutionary approach. In: Proceedings of the 2015 IEEE Congress on Evolutionary Computation (CEC). Sendai, Japan: IEEE, 2015. 1696−1703

39 Yu Z W, Wong H S. Class discovery from gene expression data based on perturbation and cluster ensemble. *IEEE Transactions on NanoBioscience*, 2009, **8**(2): 147−160

40 Zhang X R, Jiao L C, Liu F, Bo L F, Gong M G. Spectral clustering ensemble applied to SAR image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 2008, **46**(7): 2126−2136

41 Hu X H, Park E K, Zhang X D. Microarray gene cluster identification and annotation through cluster ensemble and EM-based informative textual summarization. *IEEE Transactions on Information Technology in Biomedicine*, 2009, **13**(5): 832−840

42 Xu Sen, Lu Zhi-Mao, Gu Guo-Chang. Two spectral algorithms for ensembling document clusters. *Acta Automatica Sinica*, 2009, **35**(7): 997−1002
(徐森, 卢志茂, 顾国昌. 解决文本聚类集成问题的两个谱算法. 自动化学报, 2009, **35**(7): 997−1002)

43 Ye Y F, Li T, Chen Y, Jiang Q S. Automatic malware categorization using cluster ensemble. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM, 2010. 95−104

44 Zhang P, Zhu X Q, Tan J L, Guo L. Classifier and cluster ensembles for mining concept drifting data streams. In: Proceedings of the 2010 IEEE International Conference on Data Mining (ICDM). Sydney, NSW: IEEE, 2010. 1175−1180

45 Yu Z W, Deng Z K, Wong H S, Tan L R. Identifying protein-kinase-specific phosphorylation sites based on the bagging — AdaBoost ensemble approach. *IEEE Transactions on NanoBioscience*, 2010, **9**(2): 132−143

46 Yu Z W, Li L, You J, Wong H S, Han G Q. SC^3: triple spectral clustering-based consensus clustering framework for class discovery from cancer gene expression profiles. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2012, **9**(6): 1751−1765

47 Ammour N, Alajlan N. A dynamic weights OWA fusion for ensemble clustering. *Signal, Image and Video Processing*, 2015, **9**(3): 727−734

48 Li F Y, Wu K, Lei J S, Wen M, Bi Z Q, Gu C H. Steganalysis over large-scale social networks with high-order joint features and clustering ensembles. *IEEE Transactions on Information Forensics and Security*, 2016, **11**(2): 344−357

49 Xiao W C, Yang Y, Wang H J, Li T R, Xing H L. Semi-supervised hierarchical clustering ensemble and its application. *Neurocomputing*, 2016, **173**: 1362−1376

50 Teng G, He C Z, Xiao J, He Y, Zhu B, Jiang X Y. Cluster ensemble framework based on the group method of data handling. *Applied Soft Computing*, 2016, **43**: 35−46

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

52 Tenenbaum J B, de Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, **290**(5500): 2319−2323

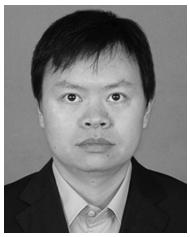
53 Zhou Z H, Tang W. Clusterer ensemble. *Knowledge-Based Systems*, 2006, **19**(1): 77−83

54 Modha D S, Spangler W S. Feature weighting in K-means clustering. *Machine Learning*, 2003, **52**(3): 217−237

55 Yang Z R, Oja E. Linear and nonlinear projective nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 2010, **21**(5): 734−749

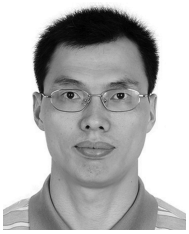
56 García S, Fernández A, Luengo J, Herrera F. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. *Information Sciences*, 2010, **180**(10): 2044−2064

褚睿鸿 西南交通大学信息科学与技术学院硕士研究生. 2014 年获得西南交通大学信息科学与技术学院学士学位. 主要研究方向为集成学习, 数据挖掘.
E-mail: rhchu@my.swjtu.edu.cn
(**CHU Rui-Hong** Master student at the School of Information Science and Technology, Southwest Jiaotong University. She received her bachelor degree from Southwest Jiaotong University in 2014. Her research interest covers ensemble learning and data mining.)



王红军 西南交通大学信息科学与技术学院副研究员. 2009 年获得四川大学计算机学院博士学位. 主要研究方向为机器学习, 集成学习与数据挖掘. 本文通信作者.
E-mail: wanghongjun@swjtu.edu.cn
(**WANG Hong-Jun** Associate professor at the School of Information Science and Technology, Southwest Jiaotong University. He received his Ph. D. degree from Sichuan University in 2009. His research interest covers machine learning, ensemble learning and data mining. Corresponding author of this paper.)

杨 燕 西南交通大学信息科学与技术学院教授. 2007 年在西南交通大学获得博士学位. 主要研究方向为数据挖掘, 集成学习, 云计算.
E-mail: yyang@swjtu.edu.cn
(**YANG Yan** Professor at the School of Information Science and Technology, Southwest Jiaotong University. She received her Ph. D. degree from Southwest Jiaotong University in 2007. Her research interest covers data mining, ensemble learning and cloud computing.)



李天瑞 西南交通大学信息科学与技术学院教授. 主要研究方向为大数据, 云计算, 粗糙集与粒计算.
E-mail: trli@swjtu.edu.cn
(**LI Tian-Rui** Professor at the School of Information Science and Technology, Southwest Jiaotong University. His research interest covers big data, cloud computing, rough set and granular computing.)