计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):220−225 [doi: 10.15888/j.cnki.csa.007429]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于教与学优化改进的近邻传播聚类算法①

马翩翩, 张新刚, 梁晶晶

(南阳师范学院 计算机与信息技术学院, 南阳 473061)

摘　要: 针对近邻传播聚类 (AP) 中偏向参数和阻尼因子设定导致聚类效果有一定局限性的问题, 提出了一种基于教与学优化算法 (TLBO) 的近邻传播聚类. 首先确定偏向参数 p 的搜索空间, 然后使用教与学优化算法在搜索空间中寻找最优参数值, 同时在聚类过程中自适应调整阻尼因子防止发生震荡, 从而提高 AP 算法的聚类质量. 实验表明, 该算法能有效的解决偏向参数和阻尼因子对聚类结果造成的局限性, 提高了聚类的轮廓系数, 并降低了聚类错误率.

关键词: 近邻传播; 偏向参数; 教与学优化; 搜索空间; 自适应调整

引用格式: 马翩翩,张新刚,梁晶晶.基于教与学优化改进的近邻传播聚类算法.计算机系统应用,2020,29(5):220–225. http://www.c-s-a.org.cn/1003-3254/7429.html

Affinity Propagation Clustering Based on Teaching Learning-Based Optimization

MA Pian-Pian, ZHANG Xin-Gang, LIANG Jing-Jing

(School of Computer and Information Technology, Nanyang Normal University, Nanyang 473061, China)

Abstract: Aiming at the limitation of the clustering effect caused by the preference and damping factors in Affinity Propagation (AP), a Teaching and Learning Based Optimization (TLBO) algorithm is proposed. First, the search space of parameter p is determined, and then the TLBO algorithm is used to find the optimal parameter value in the search space. At the same time, the damping factor is automatically adjusted to prevent numerical oscillations during the clustering process, so as to improve the clustering quality of AP algorithm. The experimental results show that the algorithm can effectively solve the problem caused by preference and damping factors, improve the contour coefficient of clustering, and reduce the clustering error rate.

Key words: Affinity Propagation (AP); preference; teaching learning; search space; adaptive

近邻传播聚类算法是 2007 由 Frey BJ 和 Dueck D 提出的一种新型的发展较快的聚类算法[1], 与 K-means 算法相比它把所有的数据点作为潜在的簇中心, 通过消息传递不停迭代进而确定最终的簇中心. 该算法对点与点之间的相似性没有绝对的要求, 可以根据具体的应用来选择对应的相似性度量方法, 相似性的值可以是正的或负的, 也可以是非对称的. 该算法已经被广泛应用在数据流聚类、图像处理、文本聚类、基因序列等领域[2−5]. 但是该算法在实际应用中也有一定的局限性: 1) 该算法属于无监督聚类, 不能利用已有的标记信息来进行聚类; 2) 偏向参数的取值对最终簇的结果有较大的影响, 取值过大, 最终划分的簇集偏多, 取值过小, 最终簇的集合偏小; 3) 对形状复杂的数据集的聚类效果不佳、时间复杂度较高.

针对以上问题, 不同的学者从不同的角度进行了改进. 为了利用已有的标记信息, 文献[6]根据半监督聚类的思想, 使用少量已知的标签数据和成对点约束对数据集形成的相似度矩阵进行调整, 进而达到提高

① 基金项目: 河南省科技攻关项目 (182102210114); 南阳师范学院校级科研项目 (2019QN020)
Foundation item: Science and Technology Program of Henan Province (182102210114); Science and Technology Project of Nanyang Normal University (2019QN020)
收稿时间: 2019-10-23; 修改时间: 2019-11-20; 采用时间: 2019-12-05; csa 在线出版时间: 2020-05-07

AP 算法的聚类性能; 文献[7]在半监督聚类的基础上利用惩罚参数添加软约束来调整相似性度量, 从而提高聚类质量. 为了克服预设偏向参数和阻尼系数对聚类结果的影响, 有学者利用群体智能优化算法对偏向参数和阻尼系数寻优: 文献[8]提出了将偏向参数和阻尼系数作为粒子群算法中的粒子, 使用粒子群算法对系数寻优, 从而提高聚类质量; 为了寻找较优的偏向参数值, 文献[9]通过改进的布谷鸟搜索对偏向参数和阻尼系数进行调整, 从而提高聚类效果; 文献[10]使用量子优化中量子叠加态和旋转门对偏向参数编码和寻优以提高 AP 算法聚类的精确度和时间. 为了处理复杂形状的数据集: 文献[11]通过对数据集构建概率图模型, 分析数据集的概率密度, 并将其注入偏向参数进而进行近邻传播聚类; 文献[12]提出了一种基于全局核与高斯核的混合核函数的相似性度量, 增强了泛化能力, 从而提高了提高聚类质量; 文献[13]在密度分布的基础上, 利用最近邻搜索的方法进行相似性度量来对球面和非球面分布的数据进行聚类. 也有学者使用其他方法来对近邻传播聚类进行改进, 文献[14]使用迁移思想, 在综合考虑源数据域和目标数据域的数据特性及几何特征的基础上改进 AP 算法中的消息传递机制提高聚类效果; 文献[15]提出了一种进化亲和传播聚类方法, 考虑潜在的动态和保存时间平滑性同时对多个时间点收集的数据进行聚类. 本文则使用教与学的群体智能优化算法来对近邻传播聚类算法中的偏向参数和阻尼因子寻优, 进而提高聚类质量.

1 教与学的自适应近邻传播聚类

1.1 教与学优化算法

教与学优化算法 (Teaching Learning-Based Optimization algorithms, TLBO) 是由 Rao RV 等[16]于 2010 年提出的新的群体智能优化算法, 具有简单、无参、快速收敛、易于实现等特点. 它被广泛应用在机电工程、结构优化、模式识别、图片处理、人工神经网络等领域.

该算法模拟一个班级的教师与学生的学习过程. 其中班级相当于智能优化算法中的种群, 教师和学生相当于种群中的个体, 所学科目为智能优化算法中的变量, 各科成绩为目标函数的适应度值, 全局最优即为目标函数的极值. 假设最优化问题用式 $z=\max F(x)$ 表示, 则 $X=(x_1, x_2, \cdots, x_i, \cdots, x_d)$ 表示自变量, 对应学生个体; X_i 表示任意一个决策变量, 对应学生的某一个科目; d 表示决策变量的维度, 对应学生所学科目数; $S=\{X|X_i^L \le X_i \le X_i^H\}$ 表示有个体构成的种群, 即由学生构成的班级; 其中 X_i^L 和 X_i^H 分别代表个体 X_i 的下界和上界.

教师是从学生种群中选取出来的最优秀个体, 即 $z=\text{argmax}F(x)$. 为了找到在决策空间中的目标函数极值, 需要经历两个阶段. (1) 教学阶段. 该阶段主要借鉴在学习场景中教师是最优学习者, 其他学生在教师的教授下得到进步, 所以首先找到在该次迭代中的最优学习者即具有最优目标函数值的参数作为教师, 其他参数向此轮最优参数学习; (2) 学习阶段. 该阶段主要借鉴于学生与学生之间互相学学习的场景, 从众多参数中随机选取和自己相异的参数进行学习, 以改变自己的参数从而提高目标函数值.

1.1.1 教师阶段

在此阶段, 在整个搜索空间中随机生成解. 最优函数值将在所有随机生成解中进行选择, 并把最优搜索解称为教师 X_{teacher}, 而教师通过教授学生 X_i, 从而提高群体的整体知识水平, 使平均成绩得到一定提高. 具体如式 (1)、式 (2) 所示:

$$Difm = r(X_{\text{teacher}} - T_F \cdot X_{\text{mean}}) \tag{1}$$

$$X_{\text{new},i} = \begin{cases} X_{\text{old},i} + Difm, & \text{if } F_i(X_{\text{old},i}) < F_i(X_{\text{new},i}) \\ X_{\text{old},i}, & \text{otherwise} \end{cases} \tag{2}$$

其中, X_{mean} 代表的是整群体的平均水平, 而老师讲授知识, 学生接受新知识的过程是一个随机的过程, 具体学习效果取决于学习补长 r 和教学因子 T_F, r 为取值范围为[0,1]的随机数, $T_F=roud[1+rand(0,1)]$. 式 (2) 中 $X_{\text{new},i}$ 代表的是第 i 个学生学习之后的知识水平, $X_{\text{old},i}$ 代表的是学习前的知识水平, 知识水平是否得到提升取决于个体的目标函数是否朝着优化方向前进.

1.1.2 学生阶段

如上所述, 根据教与学的过程, 学生可以学习知识; 通过学生之间的互动, 学生也可以增加他们的知识. 因此, 一个搜索空间中的解 (学生 X_i) 与总体中的其他解 (学生 X_j) 随机互动, 肯定有一方会学到新的东西. 如果 X_i 优于 X_j, 则 X_j 将向 X_i 移动, 如式 (3) 所示; 反之 X_i 将向 X 移动, 如式 (4) 所示. 具体描述如下:

$$X_{\text{new},i} = X_i + r\left(X_i - X_j\right), \text{ if } F(X_i) \ge F(X_j) \tag{3}$$

$$X_{\text{new},i} = X_i + r\left(X_j - X_i\right), \text{ if } F(X_i) \le F(X_j) \tag{4}$$

1.2 近邻传播聚类算法

聚类算法是寻找数据内在特征的一种划分过程. 在一组含有 N 个样本 $D=\{x_1, x_2, \cdots, x_i, \cdots, x_N\}$, 维度为 d, $x_{id}=\{x_{i1}, x_{i2}, \cdots, x_{id}\}$的数据集中, 聚类算法是将样本数据集划分为 k 个不相交的簇$\{C_l \mid l=1, 2, \cdots, k\}$其中 $C_L \cap_{L' \neq L} C_L=\phi$ 且 $D=\cup_{l=1}^{k}$. 近邻传播是近年来发展较快的聚类算法, 它按照聚类定义找出簇中心, 使簇内距离最小.

$$F(x,c)=\sum_{i=1}^{N}\min\left\{\|x_i-c_i\|^2\right\} \tag{5}$$

其中, $i=1,\cdots, k$,代表该聚类结果中有 k 个簇. C_i 为簇中心, $C_i\neq\phi$, 且 $C_i\cap C_j=\phi$.

在近邻传播聚类算法执行过程中需输入任意两点 i 和 k 之间的相似性 $s(i,k)$ 所构造的相似度矩阵, 默认为负的欧式距离, 即 $s(i,k)=-\|x_i-x_k\|^2$; 设定自相似性 $s(k,k)$ 又叫偏向参数 p, 因为近邻传播聚类算法不需要指定簇的个数, 而是通过参数 p 来影响最终的簇个数.

其次近邻传播聚类算法的聚类过程是点与点之间的消息传递来实现的. 信息有两类, 吸引度 $r(i,k)$ 和归属度 $a(i,k)$. $r(i,k)$ 代表的是从簇成员 i 发送到簇中心 k 的消息, 表示数据点 i 作为以 k 为簇中心的簇成员的度量; $a(i,k)$ 代表的是从簇中心 k 发送到簇成员 i 的消息, 表示数据点 k 作为数据点 i 的簇中心的度量; 结合归属度和吸引度, 对于数据点 i, 找到使 $a(i,k)+r(i,k)$ 达到最大化值的 k 时, 如果 $k=i$, 则点 k 为簇中心; 如果 $k\neq i$ 时, 则表示数据点 k 作为数据点 i 的簇中心. 初始时 $r(i,k)=0$, $a(i,k)=0$; 当循环终止时寻找 $(diag(A)+diag(R))>0$ 的点 k 作为簇中心. $r(i,k)$ 和 $a(i.k)$ 更新公式如下:

$$r(i,k)\leftarrow s(i,k)-\max_{k'\neq k}\{a(i,k')+s(i,k')\} \tag{6}$$

如果 $i=k$, 则:

$$r(k,k)\leftarrow s(k,k)-\max_{k'\neq k}\{a(i,k')+s(i,k')\} \tag{7}$$

$$a(i,k)\doteq\min\left\{0, r(k,k)+\sum_{i'\notin\{i,k\}}\max\{0,r(i',k)\}\right\} \tag{8}$$

如果 $i=k$, 则:

$$a(k,k)\leftarrow\sum\max\{0,r(i,k)\} \tag{9}$$

根据式 (8) 可知归属度 $a(i,k)$ 被设置为吸引度 $r(k,k)$ 加上潜在簇中心 k 从其他点获得吸引度的总和, 而且只添加传入吸引度为正值的那一部分, 再次验证了一个好的簇中心只需要表明他很适合作为一些数据点的簇中心即可 (所以只添加正值的吸引度), 而不管他多么不适合做为其他点的簇中心 (负值的吸引度). 如果自吸引度 $r(k,k)$ 是负的, 则表明点 k 不适合做为簇中心; 如果 $r(i,k)$ 为正值, 那么 $a(i,k)$ 也会增加. 为了限制强势吸引度 $r(i,k)$ 的影响, 对总和进行阈值, 使其不能超过零.

根据近邻传播聚类算法的消息传递过程可知, 在迭代过程中, 当一些点被分配到合适的簇中心时, 它们的归属度 $a(i, k)$ 如式 (8) 所示将会降到 0 及其以下. 根据式 (6) 可知这些负归属度可减少一些输入相似性 $s(i,k)$ 的有效值, 从而从竞争中删除相应的候选范例. 在式 (7) 中如果 $i=k$ 时, 吸引度 $r(k,k)$ 的值为 $s(k,k)$ 减去点 i 和所有其他候选簇中心之间的相似性的最大值, 这种“自我吸引”即偏向参数 p 反映了在聚类迭代过程中 k 点是否适合做为一个簇中心, 即偏向参数 p 的设定会影响最终的聚类效果. 较高的偏向参数值表示成为簇中心的点较多, 结果导致划分的簇集较多; 较低的偏向参数值代表成为簇中心的点较少, 结果导致划分的簇偏少.

同时在更消息更新过程中, 为避免在某些情况下出现振荡, 需对消息即吸引度和归属度进行阻尼. 具体如下所示:

$$\begin{cases} R_{\text{new}}=\lambda*R_{\text{old}}+(1-\lambda)*R_{\text{new}} \\ A_{\text{new}}=\lambda*A_{\text{old}}+(1-\lambda)*A_{\text{new}} \end{cases} \tag{10}$$

1.3 基于教与学优化的近邻传播聚类

教与学优化的近邻传播聚类 (TLBOAP) 是一种基于 AP 的改进聚类方法, 是在计算目标函数过程中进行近邻传播聚类, 即通过群体智能优化中的教与学优化算法来寻找最优偏向参数 p 从而提高 AP 的聚类效果, 同时该算法在聚类过程中会调整阻尼因子 λ 防止发生震荡. 算法 1 为基于 TLBO 的 AP 算法.

算法 1. TLBOAP 算法

1) 加载数据集.

2) 对数据集进行归一化处理. 算法种群规模为 n, 对应数据集中样本个数, 决策变量为偏向参数 p, 最大迭代次数为 $T_{\max}$.

3) 采用负的欧式距离建立相似度矩阵 S.

4) 根据式 (6)~(9) 执行近邻传播聚类, 同时根据聚类目标函数 (5) 计算目标函数值.

5) 找出目标函数值最大的值作为群体内的教师 $X_{teacher}$, 计算群体平均值 X_{mean}.
6) 教师阶段: 按照式 (1) 和式 (2) 产生新个体, 如果新个体的目标函数值优于原有个体, 则保留新个体的值.
7) 学生阶段: 个体根据式 (3) 和式 (4) 产生新个体, 如果新个体的目标函数值优于原有个体, 则保留新个体的值.
8) 循环执行直到最大迭代次数为 T_{max}, 找出最小目标函数值下的对应参数.
9) 找出簇中心并进行分配, 输出最终的簇.

在步骤 2) 中为了提高教与学寻找最优参数的效率, 需指定参数 p 的搜索空间. 在文献[17]中验证了当 p 的上限取值为 $pm/2$; 文献[18]中根据净相似度来求出 p 值的下限 p=$dpsim1-dpsim2$. 净相似性指的是数据集中任一点和它所对应的簇中心的相似性之和. 式 (11) 中 $dpsim1$ 代表的是聚类后簇个数为 1 的的净相似度值,

$$dpsim1 = \max\left\{\sum_{j} s(i,j)\right\} \tag{11}$$

式 (12) 中 $dpsim2$ 代表的是聚类后簇个数为 2 的的净相似度值.

$$dpsim2 = \max_{i\neq j}\left\{\sum_{k}\max\{s(i,k),s(j,k)\}\right\} \tag{12}$$

在步骤 4) 进行近邻传播聚类过程中, 通过监控窗口的设置监控震荡. 设置监控窗口大小为 v=40, 如果窗口中长度有三分之二簇个数不断发生变化, 则认为发生震荡, 此时以步长 0.05 的速度调整阻尼因子 λ 的值, λ 默认值为 0.5, 同时阻尼因子 λ 有上限, 需小于 1. 因为当阻尼因子过大时, 会导致聚类速度过慢.

2 实验与分析

为了验证 TLBOAP 算法的聚类效率, 在内存 4 GB、处理器为 Intel(R) Core(TM)i7—7700Q、Windows7 64 位的计算机上用 Matlab R2012a 来实现. 已在 6 个 UCI 数据集上进行了测试, 具体数据集如表 1 所示.

表 1 数据集性质

数据集	样本数目	属性数目	簇的数目
Iris	150	4	3
Glass	274	10	7
Wine	178	13	3
Zoo	101	17	7
Soybean	305	35	15
Sonar	208	60	2

实验中最大迭代次数为 5000, 连续收敛次数为 50. 将本文算法与原 AP 算法、自适应 AAP 从簇的个数、准确率、正则化信息、芮氏指标、时间等方面对聚类结果进行评估.

1) 准确率 (ACC)

准确率代表的是聚类正确的样本数占总样本数的比例.

$$ACC = \sum_{i=1}^{K}\frac{L_i}{NUM_i} \tag{13}$$

其中, k 为聚类后所得簇的个数, L_i 为第 i 个簇中聚类正确的样本个数, 是一个直观的度量指标.

2) 正则化互信息 (NMI)

正则化信息表示的是聚类后簇个数与真实簇之间的关联程度, 指标范围为[0,1]. 越接近 1, 表示关联程度越高, 聚类效果越好.

$$NMI = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K}N_{i,j}\log\frac{NN_{i,j}}{N_iN_j}}{\sqrt{\sum_{i=1}^{K}N_i\log\frac{N_i}{N}\sum_{j=1}^{K}N_j\log\frac{N_j}{N}}} \tag{14}$$

其中, K 为聚类后所得簇的个数; N 为样本总数; N_i、N_j 表示第 i、j 簇中样本数目; N_{ij} 表示样本在第 i 个簇中但属于第 j 个簇的数目.

3) Rand 指数 (RI)

Rand 指数是聚类性能度量中将聚类结果与外部“某个参考模型”相对比的外部指标.

$$RI = \frac{a+d}{a+b+c+d} \tag{15}$$

其中, 聚类结果用 $C=\{C_l \mid l=1, 2, \cdots, k\}$表示, 外部参考模型用 $C^*=\{C^*_l \mid l=1, 2, \cdots, s\}$表示. 其中 a 表示在 C 中隶属于同一簇且在 C^*中也隶属于同一簇的样本数; b 表示在 C 中隶属于同一簇但在 C^*属于不簇的样本数; c 表示在 C 属于不同簇但在 C^*隶属于同一簇的样本数; d 表示在 C 中不属于同一簇且在 C^*中也不属于同一簇的样本数. 是在簇 C_l 和 C^*不明确簇的对应情况下的聚类结果度量.

在近邻传播聚类过程中, 由于偏向参数 p 的取值不同, 产生的簇数目也不同. 对于原始近邻传播聚类算法 AP, p 初始值默认为相似矩阵的均值, 在传递过程中不发生改变; 对于 AAP 算法来说, 主要通过调节阻尼系数 λ 进而调节偏向参数 p, 相较于原始近邻传播算法

来说, 侧重点在于防止发生震荡, 而不在于寻找最优参数; 而在本文中所提到的 TLBOAP 算法, 侧重点在于通过群体智能优化算法教与学对方法在聚类过程中寻找最优偏向参数.

由表 2 可知在 AP、AAP 方法中产生的簇较多, 而 TLBOAP 在多数数据集中产生的正确个数的簇, 当维数过高时产生的簇的数目会多于实际产生的数目, 但是较 AP、AAP 方法来说产生的簇数目还是更接近真实簇数目.

表 2 簇个数

数据集	AP	AAP	TLBOAP
Iris	6	6	3
Glass	24	23	7
Wine	8	9	3
Zoo	8	9	7
Soybean	24	23	15
Soner	23	20	7

偏向参数 p 值的选择不但会影响簇中心的多少, 也会确定最终的簇中心, 簇中心的确定会影响最终簇聚类的质量. 原始近邻传播聚类算法 AP, 偏向参数 p 值固定且不变; 对于 AAP 算法来说, 为调整阻尼系数的基础上以固定步长调整偏向参数; TLBOAP 算法, 则通过教与学优化算法在有限有空间中寻找最优偏向参数 p, 所以会导致聚类性能指标差别也较大.

表 3 中的 Iris 数据集是聚类分析最常用的数据集, 有 150 个样本, 3 个类簇. 本文 TLBOAP 算法能达到 0.9457 的聚类准确率, 相较于 AP 算法、AAP 算法 *ACC* 指标分别提升 43.44%, 71.10%; TLBOAP 算法对 Iris 数据集聚类结果的 *NMI* 和 *RI* 评价指标也优于 AP、APP 算法, *NMI* 指标为 0.8322, 较 AP、AAP 算法分别提升了 20.61%、34.74%; *RI* 指标为 0.9341, 较 AP、AAP 算法分别提升了 10.92%、15.12%.

表 3 各聚类算法 ACC、NMI 和 RI 聚类评价指标比较

算法	Iris			Glass			Wine		
	ACC	*NMI*	*RI*	*ACC*	*NMII*	*RI*	*ACC*	*NMII*	*RI*
AP	0.66	0.69	0.8421	0.3645	0.4091	0.728	0.3652	0.3729	0.6959
AAP	0.5533	0.6176	0.7829	0.3878	0.4321	0.7288	0.3427	0.3949	0.6943
TLBOAP	0.9457	0.8322	0.9341	0.5607	0.3158	0.9341	0.7472	0.4815	0.7583
算法	Zoo			Soybean			Sonar		
	ACC	*NMI*	*RI*	*ACC*	*NMI*	*RI*	*ACC*	*NMI*	*RI*
AP	0.6633	0.7309	0.8505	0.1983	0.6957	0.911	0.1201	0.1617	0.5104
AAP	0.6633	0.7499	0.8560	04918	0.6877	0.9085	0.1682	0.1778	0.5102
TLBOAP	0.7624	0.7809	0.8808	0.5082	0.6287	0.8169	0.3173	0.1713	0.5176

在含有 178 个样本, 3 个类簇, 13 个特征的数据集 Wine 中, 本文 TLBOAP 算法能达到 0.7427 的聚类准确率, 相较于 AP 算法、AAP 算法 *ACC* 指标分别提升 1.04 倍, 1.18 倍; *NMI* 指标中分别提升了 29.12%、21.92%; 在 *RI* 指标上分别提升了 6.09%、6.33%

在高维特征的数据集中 Zoo、Sonar 为例, 虽然在 *NMI*、*RI* 指标上基本持平, 但是在 *ACC* 指标上有较大提升. 在 Zoo 数据集中, *ACC* 指标较 AP 算法、APP 算法均提升了 14.94%; 在 Sonar 数据集中, *ACC* 指标较 AP 算法、APP 算法分别提升了 1.64 倍、88.47%.

总之, 在度量聚类效果时, 要进行整体考虑, 因为不同的指标方法不同, 侧重点也不同. 由表 3 可知, 在 6 个 UCI 数据集中, 本文中的 TLBOAP 算法聚类效果整体上均优于 AP 算法、AAP 算法.

原始近邻传播聚类算法 AP 因为偏向参数 p 取值较大, 所以收敛速度慢; 自适应 AAP 算法, 因为迭代过程中需要针对每次阻尼因子的取值调整偏向参数 p 值, 然后在该偏向参数 p 值下再次循环调整阻尼因子; TLBOAP 算法, 因为教与学优化算法无参数, 鲁棒性强, 所以在有限有空间中寻找偏向参数 p 值较快.

由表 4 可知, 本文算法的计算效率优于 AP、AAP. 在数据集 Iris、Glass、Wine、Zoo、Soybean、Sonar 上, TLBOAP 算法的运算速率较 AP 算法分别提升了是 12.1 倍、6.7 倍、7.7 倍、3.1 倍、9.8 倍、21.2 倍; 较 AAP 算法分别提升了 61.82%、88.67%、33.03%、基本相同、44.95%、18.28%.

表 4 运行时间

数据集	AP	AAP	TLBOAP
Iris	3.8804	0.4785	0.2957
Glass	2.5498	0.6230	0.3302
Wine	3.6557	0.6218	0.4164
Zoo	2.7029	0.6687	0.6632
Soybean	13.8592	0.4918	0.3393
Sonar	10.5411	0.5622	0.4753

3 总结

本文依据聚类的度量标准, 通过使用群体智能算法中的教与学优化在偏向参数 p 空间中寻找合适的偏向参数取值, 同时通过步长调整来防止震荡, 最终确定最终聚类结果. 同时和已有的算法 AP, AAP 进行了对比试验, 在度量指标 ACC、NMI、RI 运行时间上均有一定的提高. 但是该算法在相似性度量上采用的是欧式距离, 对于非凸数据集的聚类有一定的局限性, 下一步将结合特征提取工程和核函数来提高聚类质量.

参考文献

1 Frey BJ, Dueck D. Clustering by passing messages between data points. Science, 2007, 315(5814): 972–976. [doi: 10.1126/science.1136800]

2 Zhang XL, Furtlehner C, Germain-Renaud C, *et al*. Data stream clustering with affinity propagation. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(7): 1644–1656. [doi: 10.1109/TKDE.2013.146]

3 Devika G, Parthasarathy S. Fuzzy statistics-based affinity propagation technique for clustering in satellite cloud image. European Journal of Remote Sensing, 2018, 51(1): 754–764. [doi: 10.1080/22797254.2018.1482731]

4 Guan RC, Shi XH, Marchese M, *et al*. Text clustering with seeds affinity propagation. IEEE Transactions on Knowledge and Data Engineering, 2011, 23(4): 627–637. [doi: 10.1109/TKDE.2010.144]

5 Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: An R package for affinity propagation clustering. Bioinformatics, 2011, 27(17): 2463–2464. [doi: 10.1093/bioinformatics/btr406]

6 Xiao Y, Yu J. Semi-supervised clustering based on affinity propagation algorithm. Journal of Software, 2008, 19(11): 2803–2813.

7 Arzeno NM, Vikalo H. Semi-supervised affinity propagation with soft instance-level constraints. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(5): 1041–1052. [doi: 10.1109/TPAMI.2014.2359454]

8 谢文斌, 童楠, 王忠秋, 等. 基于粒子群的近邻传播算法. 计算机系统应用, 2014, 23(3): 103–107, 76. [doi: 10.3969/j.issn.1003-3254.2014.03.017]

9 Jia B, Yu B, Wu Q, *et al*. Adaptive affinity propagation method based on improved cuckoo search. Knowledge-Based Systems, 2016, 111: 27–35. [doi: 10.1016/j.knosys.2016.07.039]

10 苏一丹, 房骁, 覃华, 等. 量子近邻传播聚类算法的研究. 广西大学学报 (自然科学版), 2018, 43(2): 561–568.

11 覃华, 詹娟娟, 苏一丹. 基于概率无向图模型的近邻传播聚类算法. 控制与决策, 2017, 32(10): 1796–1802.

12 Sun L, Liu RN, Xu JC, *et al*. An affinity propagation clustering method using hybrid kernel function with LLE. IEEE Access, 2018, 6: 68892–68909. [doi: 10.1109/ACCESS.2018.2880271]

13 Fan ZY, Jiang J, Weng SQ, *et al*. Adaptive density distribution inspired affinity propagation clustering. Neural Computing and Applications, 2019, 31(S1): 435–445. [doi: 10.1007/s00521-017-3024-6]

14 杭文龙, 蒋亦樟, 刘解放, 等. 迁移近邻传播聚类算法. 软件学报, 2016, 27(11): 2796–2813. [doi: 10.13328/j.cnki.jos.004921]

15 Arzeno NM, Vikalo H. Evolutionary affinity propagation. Proceedings of 2017 IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, LA, USA. 2017. 2681–2685.

16 Rao RV, Savsani VJ, Vakharia DP. Teaching-learning-based optimization: A novel method for constrained mechanical design optimization problems. Computer-Aided Design, 2011, 43(3): 303–315. [doi: 10.1016/j.cad.2010.12.015]

17 Yu J, Cheng QS. The upper bound of the optimal number of clusters in fuzzy clustering. Science in China Series F: Information Sciences, 2001, 44(2): 119–125.

18 王开军, 张军英, 李丹, 等. 自适应仿射传播聚类. 自动化学报, 2007, 33(12): 1242–1246.