自适应仿射传播聚类

王开军 [1] 张军英 [1] 李 丹 [1] 张新娜 [2] 郭 涛 [1]

摘 要 适合处理大类数的仿射传播聚类有两个尚未解决的问题: 一是很难确定偏向参数取何值能够使算法产生最优的聚类结果; 另一个是当震荡发生后算法不能自动消除震荡并收敛. 为了解决这两个问题, 提出了自适应仿射传播聚类方法, 具体技术包括: 自适应扫描偏向参数空间来搜索聚类个数空间以寻找最优聚类结果、自适应调整阻尼因子来消除震荡以及当调整阻尼因子方法失效时的自适应逃离震荡技术. 与原算法相比, 自适应仿射传播聚类方法性能更优, 能够自动消除震荡和寻找最优聚类结果. 对模拟和真实数据集的实验结果表明, 自适应仿射传播聚类方法十分有效, 其聚类质量优于或不低于原算法.

关键词 仿射传播聚类, 自适应聚类, 大类数的聚类算法
中图分类号 TP18

Adaptive Affinity Propagation Clustering

WANG Kai-Jun[1] ZHANG Jun-Ying[1] LI Dan[1] ZHANG Xin-Na[2] GUO Tao[1]

Abstract Affinity propagation (AP) clustering has two limitations: it is hard to know what value of parameter "preference" can yield optimal clustering solutions, and oscillations cannot be eliminated automatically if occur. This paper proposes an adaptive AP method to overcome these limitations, including adaptive scanning of preferences to search space for finding the optimal clustering solution, adaptive adjustment of damping factors to eliminate oscillations, and adaptive escaping from oscillations when the damping-factor adjustment technique fails. In comparison to AP, the adaptive AP has better performance on automatic oscillation elimination and finding of an optimal clustering solution. Experimental results on simulated and real data sets show that the adaptive AP is effective and its quality of clustering results is better than or equal to that of AP.

Key words Affinity propagation (AP) clustering, adaptive clustering, large number of clusters

1 引言

仿射传播聚类 (Affinity propagation clustering, AP)[1] 是在*Science*上提出的一种新聚类算法, 其优势体现在处理类数很多的情况时运算速度快[2]. AP 算法以 N 个数据点之间的相似度 (例如欧式距离作为测度) 组成的 $N\times N$ 相似度矩阵 S (S 中每个元素再加上负号成为负数, 例如对点 x_i 和 x_j 有 $S(i,j) = -||x_i - x_j||^2$, 以下均将 x_i 简写为点 i) 为工作基础, 并在算法开始时将所有的数据点都视为潜在的聚类中心 (称为 exemplar 或类代表). 设在数据的特征空间中存在一些比较紧密的聚类, 且聚类的能量函数为各数据点与其聚类中心的相似度之和, 即 $E(C) = -\sum_i S(i, C_i)$ (其中 $i\in C_i$, C_i 为点 i 的聚类中心)[1]. 将负的两个点之间距离设想为吸引度或归属度, 则点 k 对较近的点 i 吸引力比较大, 同样点 i 认同点 k 为其聚类中心的归属感也较大. 这样, 处于聚类中心处的数据点 k 对其他数据点的吸引力之和较大, 成为聚类中心的可能性也越大; 反之, 处于聚类边缘处的数据点对其他数据点的吸引力之和比较小, 成为聚类中心的可能性也越小. 从此角度出发, AP 算法为选出合适的类代表而不断从数据中搜集有关的证据: 为候选类代表点 k 从每个数据点 i 搜集证据 $R(i,k)$(称为点 k 对点 i 的 responsibility 或吸引度) 来描述数据点 k 适合作为数据点 i 的类代表的程度, 也为数据点 i 从候选类代表点 k 搜集证据 $A(i,k)$ (称为点 i 对点 k 的 availability 或归属度) 来描述数据点 i 选择数据点 k 作为其类代表的适合程度. 证据越强 (即 $R(i,k)$ 与 $A(i,k)$ 越大), 点 k 作为最终聚类中心的可能性就越大. AP 算法通过一个迭代循环不断进行证据的搜集和传递 (亦称为消息传递) 以产生 m 个高质量的类代表和对应的聚类, 同时聚类的能量函数也得到了最小化. 将各数据点分配给最近的类代表所属的类, 则找到的 m 个聚类即是聚类结果.

AP 算法中有两个重要参数: 置于相似度矩阵 S 对角线的偏向参数 p 和迭代中针对 R 与 A 更新的阻尼因子 l_{am}. 偏向参数 $p(i)$(通常是负数) 表示数据点 i 被选作聚类中心的倾向性, 并对哪些类代表会作为最终的聚类中心产生重要影响. 依据文献 [1] 中吸引度 R 和归属度 A 的计算公式: $R(i, k) =$

收稿日期 2007-8-24 收修改稿日期 2007-10-25
Received August 24, 2007; in revised form October 25, 2007
国家自然科学基金 (60574039, 60371044) 资助
Supported by National Natural Science Foundation of China (60574039, 60371044)
1. 西安电子科技大学计算机学院 西安 710071 2. 中国计量学院 杭州 310018
1. School of Computer Science and Technology, Xidian University, Xi′an 710071 2. China Jiliang University, Hangzhou 310018
DOI: 10.1360/aas-007-1242

$S(i,k)-\max\{A(i,j)+S(i,j)\}$ ($j\in\{1,2,\cdots,N\}$, 但 $j\neq k$), $A(i,k)=\min\{0, R(k,k)+\sum_j\{\max(0,R(j,k))\}\}$ ($j\in\{1,2,\cdots,N\}$, 但 $j\neq i$ 且 $j\neq k$), 可知参数 p 出现在 $R(k,k)=p(k)-\max\{A(k,j)+S(k,j)\}$ 中. 这样, 当 $p(k)$ 较大使得 $R(k,k)$ 较大时, $A(i,k)$ 也较大, 从而类代表 k 作为最终聚类中心的可能性较大; 同样, 当越多的 $p(i)$ 较大时, 越多的类代表倾向于成为最终的聚类中心. 因此, 增大或减小 p 可以增加或减少 AP 输出的聚类数目, 且文献[1]推荐在无先验知识时将所有的 $p(i)$ 设定为 p_m(S 中元素的中值). 然而, 在许多情况下 p_m 不能使 AP 算法产生最优的聚类结果, 这是由于 p_m 的设定并不是依据数据集本身的聚类结构. 此外, p 与 AP 输出的聚类数目之间没有一一对应关系, 这使得很难用聚类有效性方法[3] 来寻找最优聚类结果 (最优类数). 因此, 使用 AP 算法时如何找出最优的聚类结果是尚未解决的问题.

在 AP 算法的每一步循环迭代 i 中, 吸引度 R_i 和归属度 A_i 要与上一步的 R_{i-1} 与 A_{i-1} 进行加权更新: $R_i=(1-l_{am})\times R_i+l_{am}\times R_{i-1}$, $A_i=(1-l_{am})\times A_i+l_{am}\times A_{i-1}$ (其中 $l_{am}\in[0,1]$, 默认值为 0.5[1]), 这体现了阻尼因子 l_{am} 的作用. l_{am} 的另一个作用是改进收敛性: 当 AP 算法发生震荡 (迭代过程中产生的类数不断发生摆动) 不能收敛时, 增大 l_{am} 可消除震荡[1]. 发生震荡时, 需手动增大 l_{am} 并重新运行算法, 直到算法收敛; 另一种做法是直接将 l_{am} 设定为接近 1 来避免震荡, 但 R 与 A 更新很慢使得算法运行缓慢. 因此, 当震荡发生时, 如何自动消除震荡是一个需要解决的重要问题.

为了解决这两个问题, 本文提出了自适应仿射传播聚类 (Adaptive affinity propagation clustering, adAP) 方法, 包括扫描偏向参数空间来搜索聚类个数空间以寻找最优聚类结果 (称为自适应扫描)、调整阻尼因子来消除震荡 (称为自适应阻尼) 以及降低 p 值以逃离震荡 (称为自适应逃离). 本文方法的 Matlab 源程序可由文献 [4] 获得.

2 自适应仿射传播聚类

本节首先设计自动消除震荡的方法和扫描偏向参数空间来搜索聚类个数空间的自适应方法, 然后讨论采用聚类有效性方法从算法产生的一系列聚类结果中找出最优的聚类结果. 下文中提及的偏向参数 p, 除特别指明外, 均指置于矩阵 S 对角线的 $p(i)$, 且给 p 赋值时, $p(i)$ 均取相同的值.

adAP 的目标是在震荡发生时既能消除震荡又能保持算法快速. 虽然将 l_{am} 增大到接近 1 消除震荡的可能性更大, 但 l_{am} 越大, R 更新越慢, 算法需要越多的迭代循环以达到 $l_{am}=0.5$ 时的更新效果, 故在检测到震荡后逐步增大 l_{am} 并同时查看效果是更好的选择. 依此思路, 自适应调整阻尼因子技术设计如下步骤: 1) AP 算法进行一次循环, 检测是否发生震荡; 2) 若有震荡发生, 以一个步幅 (例如 0.05) 增大 l_{am}, 否则转步骤 1); 3) 继续 w 次循环 (目的是等待 w 次循环再查看效果); 4) 重复以上步骤直到算法达到停止条件.

震荡的检测是自适应阻尼技术的关键, 但要描述震荡的特征很困难, 于是定义易于描述的非震荡特征, 即在算法的迭代过程中产生的类代表数目不断下降或保持不变 (这也是算法走向收敛的特征). 为了记录在连续的迭代循环过程中出现非震荡特征的次数, 特设计一个移动监视窗 $K_b(j)(j=1,2,\cdots,w, w$ 为窗宽) 进行记录 (可连续记录 w 次迭代循环), 例如在第 i 次迭代循环当非震荡特征出现时, $K_b(i)=1$; 否则 $K_b(i)=0$. 震荡发生与否的判定准则设计如下: 若 K_b 中 1 的数目小于窗宽的三分之二, 则震荡发生. 这种准则是一种容忍设计, 考虑了偶尔出现的少量震荡情况以及算法刚开始阶段的不稳定情况.

当自适应阻尼技术效果不佳 (例如 $l_{am}=0.85$ 或更大, 但震荡仍存在) 时, 需要启用自适应逃离震荡技术. 很大的 l_{am} 仍不能抑制震荡说明在给定的 p 下震荡是 “固执的”, 因而替代方法是离开这个 p 值以摆脱震荡. 这种逃离方法可行的原因在于 adAP 中的 p 值是可变的 (参见下文), 这与 AP 工作在一个固定的 p 下是不同的. 自适应逃离技术设计如下: 当震荡发生且 $l_{am}\geq 0.85$ 时, 逐步降低 p 值, 直到震荡消失. 这项工作只需要添加在自适应阻尼技术的步骤 2) 中: 若有震荡发生, 以一个步幅 (例如 0.05) 增大 l_{am}; 若 $l_{am}\geq 0.85$, 以某个步幅 p_{step} 降低 p, 否则转步骤 1). 自适应逃离是对自适应阻尼技术的补充, 两者同时使用有利于尽早消除震荡. 依据实验效果, 监视窗宽 $w=40$ 是兼顾算法速度的合适选择 (w 太小不能实现容忍设计的目的; 而 w 太大会增加算法的迭代次数). (这两项技术的伪代码可参见文献 [4]).

AP 输出的聚类数目依赖于输入的 p, 但对给定的数据集, p 取何值能产生最优的聚类结果却是未知的. 聚类有效性技术 (通常使用聚类有效性指标)[3] 是评价聚类结果质量的有效方法, 它需要一个聚类算法产生一系列具有不同类数的聚类结果, 然后由聚类有效性指标从中找出最优的聚类结果. 考虑到 p 与输出的聚类个数之间没有一一对应关系, 故设计扫描偏向参数空间的方法来搜索聚类个数空间, 以获得一系列具有不同类数的聚类结果.

为了保持算法快速, 偏向参数空间的扫描设计在算法的迭代循环过程中, 但不改变 AP 算法的核心内容: 算法从初始给定的 p 出发, 循环过程的每

次迭代更新吸引度 R 和归属度 A (但相似度矩阵 S 固定不变); 若循环过程收敛到某个类数 K, 以步幅 p_{step} 逐步减小 p (即改变 S 对角线上的 $p(i)$) 并重复同样的循环过程 (p 扫描的实施), 以获得不同的 K(p 扫描的目的). 为避免可能的重复计算, 在 p 扫描的实施中设计每次减小 p (即减小 $S(i, i) = p(i)$, 而 S 的其他元素未变) 后以当前的 $R(i, j)$ 与 $A(i, j)$ 值作为起点, 继续计算 $R(i, k)$ 与 $A(i, k)$. 自适应扫描技术设计如下: 1) 确定一个较大的 p 启动算法; 2) AP 算法进行一次循环, 产生 K 个类代表; 3) 检测 K 个类代表是否收敛 (收敛条件是满足预先设定的连续不变次数 v); 4) 若收敛, 转步骤 5); 否则, 转步骤 2); 5) 若连续 $delay$ 次循环均收敛到 K(等待 $delay$ 次循环是使算法更加可靠), 以某个步幅 p_{step} 减小 p, 否则转步骤 2); 6) 转步骤 2).

下降步幅的选择是自适应扫描技术的关键. 依据笔者的经验, 可将下降步幅设定为 $p_{step} = 0.01p_m$. 这个扫描间隔是一种折中设计, 它考虑了 $|p_{step}|$ 太小时算法运行缓慢 (需要更多的迭代次数来完成扫描) 以及 $|p_{step}|$ 太大时反映数据集固有聚类结构的类数有可能被错过. 然而, 固定的步幅对大类数和小类数的不同情况很难同时都适合. 因此, 针对较大的类数对 p_{step} 比小类数敏感 (敏感度 $K/|p_{step}|$) 的情况, 设计下降步幅的自适应调整技术如下: 下降系数 $q = 0.1\sqrt{K+50}$, $p_{step} = 0.01p_m/q$. 这样, 算法在产生 K 个类代表时可动态调整 q, 以实现在 K 较大时下降步幅小一些而在 K 较小时下降步幅大一些的目的. 为了检查收敛条件是否满足, 再设计一个与自适应阻尼技术中类似的监视窗 B 以记录 K 个类代表连续不变的次数, 并设置窗宽 v=40 (40 加上考虑算法更加可靠而设计的 $delay = 10$ 次延时, 则与 AP 算法中默认的收敛次数 50[1] 相同).

扫描区间的确定也是重要的, 因为总希望扫描区间小一些以节省运算时间. 偏向参数空间为 $[-\infty, 0]$, 其对应的聚类数目空间为 $[1, N]$. 对 N 个数据点进行聚类, 通常认为其最优聚类个数的上限为 N 的平方根是合理的[5]. 这样, 需要确定较大的初始 $p < 0$ 作为向 $-\infty$ 扫描的起点. 根据实验观察, 当初始 $p = p_m/2$ 时, 算法第一次收敛到的类数 K_1 基本能达到或超过 $\sqrt{N}$ (K_1 还依赖于数据集的固有聚类结构), 而且 AP 算法搜索过的类数远大于 $\sqrt{N}$ (因算法起始时将每个数据点都视为类代表), 故可将起始值设定为 $p = p_m/2$. 最小类数 2 决定了 p 的扫描下限, 即减小 p 直到获得类数 $K = 2$. 为了不让最大循环次数 $maxits$ 影响算法是否到达 $K = 2$, 不妨设置较大的 $maxits = 50\,000$.

最后, 还需设计扫描 p 的加速技术以节省运算时间. 由于 p 与类数是非一一对应关系, 某些聚类数目会对应比较广的 p 值范围, 此时需要很多次迭代使 p 值有较大的下降才能使聚类数目发生变化. 出现这种情况时, 可加大 p 的下降步幅以尽快获得更小的聚类数目. p 下降的加速技术设计如下: 1) AP 算法进行一次迭代, 检测类代表数目是否收敛到 K, 若是转步骤 2), 否则令 $b = 0$, 重复步骤 1); 2) 继续迭代 $delay = 10$ 次, 检查类代表数目是否收敛到 K; 若是则计数 $b = b + 1$, 否则转步骤 1); 3) 令 $p = p + b{\times}p_{step}$, 转步骤 2). 这样, 当类代表数目收敛时每迭代 $delay$ 次循环就使 p 下降 $b{\times}p_{step}$, 即 p 的下降幅度逐步加速直到 K 下降为止 (自适应扫描的伪代码参见文献 [4]).

现在 adAP 方法通过搜索类数空间能够输出一系列具有不同聚类数目的聚类结果, 于是可以采用聚类有效性方法来评价聚类结果的质量, 并且通常采用聚类有效性指标来评价聚类算法产生的哪个聚类结果是最优的. 在众多有效性指标中, Silhouette 指标以其对明显的聚类结构具有良好的评价能力而被广泛应用. Silhouette 指标反映了聚类结构的类内紧密性和类间可分性, 既可用于估计最优的聚类数目, 也可应用于评价聚类质量. 因此, 这里以 Silhouette 指标为例来求解最优聚类结果.

设一个具有 n 个样本 (或数据点) 的数据集被划分为 K 个聚类 $C_i (i = 1, 2, \cdots, K)$, $a(t)$ 为聚类 C_j 中的样本 t 与 C_j 内所有其他样本的平均不相似度或距离, $d(t, C_i)$ 为 C_j 的样本 t 到另一个类 C_i 的所有样本的平均不相似度或距离, 则 $b(t) = \min\{d(t, C_i)\}$, 其中 $i = 1, 2, \cdots, K$ 且 $i \neq j$. 于是, 一个样本 t 的 Silhouette 指标为

$$S_{il}(t) = \frac{b(t) - a(t)}{\max\{a(t), b(t)\}} \tag{1}$$

由 $S_{il}(t)$ 容易计算一个聚类 C_i 的所有样本的平均 S_{il} 值 $S_{av}(C_i)$, 它反映了类 C_i 的紧密性 (例如类内平均距离) 和可分性 (例如最小类间距离); 而一个数据集的所有样本的平均 S_{il} 值 $S_{av}(C)$ 则可以反映聚类结果的质量. 对于一系列聚类结果的 Silhouette 指标值, 值越大表示聚类质量越好, 最大值对应的类数是最优的聚类个数, 对应的聚类结果也是最优的[3]. 聚类结果的 Silhouette 值超过 0.5 说明各个聚类能明显地分开 (好的分类), 低于 0.5 表明一些聚类有重叠的情况, 而 0.2 以下是缺乏实质的聚类结构[6]. 由于 $S_{av}(C)$ 反映的是所有聚类可分性的平均情况, 故在类数很多时应考虑最靠近的两类的可分性, 特别是聚类结果的 $S_{av}(C)$ 较小应优先考虑, 即从 K 个聚类的 $S_{av}(C_i)$ 中找出最小值 $S_{\min}(K) = \min\{S_{av}(C_i)\}$, 再由一系列聚类结果的 $\{S_{\min}(K)\}(K = 2, 3, \cdots, K_{\max})$ 中找出最大值对应的 K 作为最优聚类个数.

3 实验结果

本节对 adAP 算法和 AP 算法的聚类性能进行实验比较, 以检验 adAP 算法能否通过自适应的 p 扫描技术并结合聚类有效性指标找出正确或更优的聚类结果, 以及能否在震荡发生时自动消除震荡或逃离震荡, 从而给出正确或更优的聚类结果. 两种算法均采用相同的阻尼因子初始值 $l_{am} = 0.5$ (但 Travelroute 实验中 $l_{am} = 0.8$[1]), 而使用固定 p 值的 AP 算法设置 $p = p_m$ 以及 $maxits = 2000$. 对 Document 和 Travelroute 实验, 两种算法均采用依据先验知识获得的固定 p 值[1].

设一个数据集表示为 m 行 (样本/数据点) d 列 (维数) 的矩阵 $X = \{x_i\}$, 其每个样本 x_i 是 d 维的. 对一般数据采用欧式距离作为 x_i 和 x_j 之间的相似性测度, 而基因表达数据则采用普遍使用的 Pearson 相关系数作为相似性测度, 即 x_i 和 x_j 之间的线性相关系数 $P(i, j)$[3]. 为了避免负数可能引起的计算混乱, 将 $P(i, j) \in [-1, 1]$ 进行转换: $P(i, j) = 1 - (1 + P(i, j))/2$, 从而 Pearson 系数转换为正的 Pearson 距离 $P(i, j) \in [0, 1]$(值越大表示两个样本相距越远). 这样, 两个样本间的相似度为 $S(i, j) = -P(i, j)$.

表 1 列出了实验中采用的 12 个数据集的特征, 其中前 8 个数据集已知类数和正确类标. 这些数据集的聚类结构特征包括: 相距远的和靠近的完全分离的聚类、轻微重叠的聚类、聚类是紧密的和松散的聚类. 表中前 4 个是模拟数据集, 其他是真实的数据集 (其中 Yeast 与 NCI60 是基因表达数据, Exons 是 75 067 个 exons 数据的子集, 即前 3499 个样本与最后一个样本).

表 2 是两种算法的聚类结果, 其中 “adAP 错误率” 表示 adAP 算法的聚类结果 (类标) 与正确类标相比的错误率; “adAP 消震荡” 栏中 “yes” 表示有震荡发生而 adAP 自动消除了它们; “adAP 时间” 与 “AP 时间” 分别表示两种算法的 Matlab 程序在同一 PC 机 (Intel CPU 3.60GHz 2GB) 上的运行时间 (秒). FM 表示测量聚类结果与正确类标一致性的 Fowlkes-Mallows 指标[3], 其值处于 0 与 1 之间且越大表示一致性越好, 例如当聚类结果与正确类标完全一致时, $FM = 1$. 当聚类结果的类数与正确类数不同时, 可用 FM 指标评价聚类结果的质量. 最后 4 个数据集没有提供正确类标, 故无错误率与 FM 值; 对 Exons, FM 栏中的值表示识别出 exons 的正确率.

从表 2 中可以看出, 对于前 8 个数据集, adAP 算法均给出了正确的聚类个数, 而 AP 算法均未产生正确的聚类个数; 并且在 FM 指标上 adAP 均高于 AP 表明 adAP 的聚类质量更好; 由于 AP 不能自动消除震荡, 故在数据集 22k10far 和 Ionosphere

表 1 数据集的特征

Table 1 Features of data sets

数据集	聚类结构特征	类数	样本数	维数	来源
3k2lap	重叠, 松散	3	300	2	文献 [7]
5k8close	靠近, 松散	5	1000	8	文献 [8]
14k10close	靠近, 松散	14	480	10	文献 [7]
22k10far	相距远, 紧密	22	790	10	文献 [7]
Ionosphere	重叠, 松散	2	351	4	文献 [9]
Wine	重叠, 松散	3	178	3	文献 [9]
Yeast	相距远, 松散	4	208	79	文献 [10]
NCI60	重叠, 松散	8	58	20	文献 [11]
FaceImage	重叠	100	900	50×50	文献 [1]
Document	/	4	125	/	文献 [1]
Travelroute	/	7	456	3	文献 [1]
Exons	/	/	3500	12	文献 [1]

表 2 adAP 与 AP 的聚类结果

Table 2 Clustering results of adAP and AP

数据集	已知类数	adAP 类数	adAP 错误率 (%)	adAP FM	adAP 时间 (s)	adAP 消震荡	AP 类数	AP FM	AP 时间 (s)
3k2lap	3	3	7.7	0.85	144.0	/	16	0.39	2.1
5k8close	5	5	0	1.00	1851.0	/	17	0.56	31.2
14k10close	14	14	0	1.00	275.5	/	15	0.97	6.0
22k10far	22	22	0	1.00	1125.9	yes	168	0.80	307.8
Ionosphere	2	2	17.4	0.75	445.3	yes	28	0.43	56.8
Wine	3	3	10.7	0.80	34.5	/	11	0.46	0.5
Yeast	4	4	3.4	0.97	54.7	/	11	0.66	0.9
NCI60	8	8	/	0.56	12.9	/	9	0.48	0.1
FaceImage	100	102	/	/	3701.2	/	103	/	14.5
Document	4	4	/	/	0.3	/	4	/	0.2
Travelroute	7	7	/	/	24.7	/	7	/	20.0
Exons	/	102	/	32.8%	83073.7	/	37	22.4%	996.0

上的聚类质量比较差. 对于 Document 数据, 两种算法均找出了同样的 4 个聚类中心作为代表句子 (聚类任务); 对于 Travelroute 数据, 两种算法均找出了同样的 7 个城市 (即聚类中心) 作为最适合的航空枢纽; 对于 Exons 数据, 需找出聚类结果中非 exons 的聚类 (事先已知最后一个样本不是 exon), 则不属于此类的为识别出的 exons, 由表 2 可知, adAP 识别 exons 的正确率高于 AP; 对于 FaceImage 数据, 由于人脸原本就比较相似, 再对 100 个人脸图像进行变化而产生 900 幅图像, 其可分性并不好 (100 个聚类的 Silhouette 值 0.207 很低说明了这一点). 因此, 两种方法的聚类结果接近 100 类已是相当好了, 而 adAP 的 102 类更接近正确的类数. 这些结果表明 adAP 算法能够找出正确或更好的聚类结果和自动消除震荡, 也验证了 adAP 算法中采用的自适应扫描、自适应阻尼和自适应逃离技术是十分有效的.

4 结论

自适应 AP 算法通过对偏向参数空间的自适应扫描来搜索整个聚类个数空间, 并依据聚类有效性技术寻找符合数据聚类结构的最优聚类结果. 在自适应 AP 算法中, 设计自适应阻尼技术自动消除发生的震荡, 还设计了自适应逃离技术在阻尼技术效果不佳时逃离震荡. 依靠这些自适应技术, 自适应 AP 算法在聚类质量和震荡消除方面都优于或不低于原 AP 算法. 此外, 对于聚类结构复杂的数据集 (例如重叠的聚类), 是否有更合适的有效性指标或方法与自适应 AP 算法结合还需要进一步的研究.

References

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

2 Kelly K. Affinity program slashes computing times [Online], available: http://www.news.utoronto.ca/bin6/070215-2952.asp, October 25, 2007

3 Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 2002, **3**(7): 1∼21

4 Wang K J. Supplement of adaptive affinity propagation clustering [Online], available: http://www.mathworks.com/matlabcentral/fileexchange/loadAuthor.do?objectType=author&objectId=1095267, October 25, 2007

5 Yu Jian, Cheng Qian-Sheng. The upper bound of the optimal number of clusters in fuzzy clustering. *Science in China*, 2002, **32**(2): 119∼125
(于剑, 程乾生. 模糊聚类方法中的最佳聚类数的存在范围. 中国科学, 2002, **32**(2): 119∼125)

6 Velamuru P K, Renaut R A, Guo H B, Chen K W. Robust clustering of positron emission tomography data. In: Joint Interface CSNA. USA: 2005

7 Dembélé D, Kastner P. Fuzzy C-means method for clustering microarray data. *Bioinformatics*, 2003, **19**(8): 973∼980

8 Strehl A. Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining [Ph. D. dissertation], The University of Texas at Austin, 2002

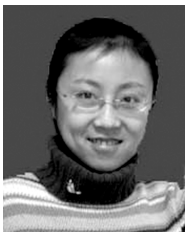
9 Blake C L, Merz C J. UCI repository of machine learning databases (University of California) [Online], available: http://mlearn.ics.uci.edu/MLRepository.html, September 27, 2007

10 Ben H A, Guyon I, Elisseeff A. A stability based method for discovering structure in clustered data. In: Proceedings of the 7th Pacific Symposium on Biocomputing. Hawaii, USA: 2002. 6∼17

11 Ross D T, Scherf U, Eisen M B, Perou C M, Rees C, Spellman P. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 2000, **24**(3): 227∼235

王开军 西安电子科技大学计算机学院博士研究生. 主要研究方向为人工智能、模式识别以及生物信息学. 本文通信作者. E-mail: sunice9@yahoo.com
(**WANG Kai-Jun** Ph. D. candidate at School of Computer Science and Technology, Xidian University. His research interest covers artificial intelligence, pattern recognition, and bioinformatics. Corresponding author of this paper.)

张军英 西安电子科技大学计算机学院教授. 主要研究方向为人工神经网络、图象处理、模式识别、优化、智能信息处理与生物信息学.
E-mail: jyzhang@mail.xidian.edu.cn
(**ZHANG Jun-Ying** Professor at School of Computer Science and Technology, Xidian University. Her research interest covers neural networks, image processing, pattern recognition, optimization, intelligent information processing, and bioinformatics.)



李 丹 西安电子科技大学计算机学院硕士研究生. 主要研究方向为人工智能和模式识别. E-mail: dli20129@126.com
(**LI Dan** Master student at School of Computer Science and Technology, Xidian University. Her research interest covers artificial intelligence and pattern recognition.)

张新娜 中国计量学院教师. 主要研究方向为自动控制.
E-mail: sinnar@cjlu.edu.cn
(**ZHANG Xin-Na** Lecturer at China Jiliang University. Her research interest covers control science and engineering.)

郭 涛 西安电子科技大学计算机学院博士研究生. 主要研究方向为自动控制.
E-mail: myit_02@126.com
(**GUO Tao** Ph. D. candidate at School of Computer Science and Technology, Xidian University. His research interest covers control science and engineering.)