可变相似性度量的近邻传播聚类

董 俊[①] 王锁萍[①] 熊范纶[②]
[①](南京邮电大学信息网络研究所 南京 210003)
[②](中国科学院智能机械研究所 合肥 230031)

摘 要：近邻传播(AP)聚类算法面临的一个问题是不适用于多重尺度及任意空间形状的数据聚类处理。该文从数据分布特性的表征出发，提出了一种改进的近邻传播聚类算法 AP-VSM (Affinity Propagation based on Variable-Similarity Measure)。首先，综合数据的全局与局部分布特性，设计了一种数据可变相似性度量计算方法，该度量可以有效地反映数据实际聚类的分布特性；然后在传统 AP 算法框架基础上，构造出基于可变相似性度量的近邻传播聚类算法，从而拓展了传统 AP 算法的数据处理能力。仿真实验验证了新方法性能优于传统 AP 算法。

关键词：数据处理；聚类分析；近邻传播聚类；可变相似性度 b 量；流形分析
中图分类号：TP391 **文献标识码**： A **文章编号**：1009-5896(2010)03-0509-06
DOI:10.3724/SP.J.1146.2009.01066

Affinity Propagation Clustering Based on Variable-Similarity Measure

Dong Jun[①] Wang Suo-ping[①] Xiong Fan-lun[②]
[①](*Institute of Information Network*, *Nan Jing University of Posts & Telecommunations*, *Nanjing* 210003, *China*)
[②](*Institute of Intelligent Machines*, *Chinese Academy of Sciences*, *Hefei* 230031, *China*)

Abstract: Affinity Propagation (AP) clustering is not fit to deal with multi-scale data cluster as well as the arbitrary shape cluster issue. Therefore, an improved affinity propagation clustering algorithm AP-VSM (Affinity Propagation based on Variable-Similarity Measure) is proposed embarking from the token of data distribution characters. First, a kind of variable-similarity measure method is devised according of characters of global and local data distribution, which has the ability of describing the characters of data clustering effectively. Then AP-VSM clustering algorithm is proposed base on the frame of traditional AP algorithm, and this method has extended data processing capacity compared with traditional AP. The simulation results show that the new method is outperforming traditional AP algorithm.

Key words: Data processing; Cluster analysis; Affinity Propagation (AP) clustering; Variable-similarity measure; Manifold analysis

1 引言

由于经典的K均值算法对初始聚类中心的选择敏感且容易陷入局部极值，因此，需要在不同初始化下运行很多次去寻找一个最好的聚类结果。为了避免这个问题，Frey等人提出了与K均值算法同属于K中心聚类方法的近邻传播(Affinity Propagation, AP)聚类[1,2]，它是一种新的聚类算法。相比较众多经典算法，它将所有数据点都作为候选的类代表点，避免了聚类结果受限于初始类代表点的选择，同时对于相似度矩阵的对称性没有要求，并在处理多类数据时运算速度快，所以性能更好[3−5]。

由于AP算法是基于中心的聚类方法，因此它也像其它中心聚类算法一样在紧凑的具有超球形分布的数据集上具有较好的聚类性能，并不适合任意空间形状聚类问题和具有多重尺度的数据。这都为AP的进一步应用带来了困难，要克服上述问题就必须改进AP算法。

目前出现了很多改进的方法，例如建立一种软限制策略[6]，改进偏向参数[7]等，但是对于上述问题则效果并不理想。由于AP聚类算法是在数据形成的相似性矩阵基础上进行聚类的，所以这里从聚类的相似性度量着手进行改进。本文从数据流形分析的角度研究并设计了一种相似性度量，它可以改善在类代表点(或者是候选类代表点)附近的数据点以及与类代表点在同一流形上的数据点，使其数据点间的相似性度量值增大，且对于不在类代表点附近的数据点之间的相似性度量值将减小，从而提高算法的聚类精度。虽然在构建相似性矩阵的过程中时间有所增加，但是由于改善了相似性度量，相似性矩阵的结构也随之改善，使AP聚类过程中的迭代次数

2009-08-05 收到，2010-01-13 改回
国家 863 计划项目(2006AA10z249)资助课题
通信作者： 董俊 dongjun@mail.hf.ah.cn

大为减少，所以总体时间是减少的，有时甚至比原算法少很多。有效性分析和实验表明，本文所改进的算法相对于原算法在聚类的性能上有显著提高。

本文第2节简单叙述AP算法的思想；第3节详细介绍可变相似性度量的计算方法，并将其引入到AP算法中得到可变相似性度量的近邻传播(Affinity Propagation based on Variable-Similarity Measure，AP-VSM)聚类算法，同时介绍了评价该算法的有效性分析指标；第4节实验验证可变相似性度量在AP算法中的性能，并对比原算法和其它改进算法；最后是结束语。

2 AP 算法思想

AP算法[1−5]首先将数据集的所有N个样本点都视为候选的聚类中心，为每个样本点建立与其他样本点的吸引程度的信息，即任意2个样本点x_i和x_j之间的相似度$s(i,j)=-\|x_i-x_j\|^2$被存储在$N\times N$的矩阵中；在聚类之前，每个点将被赋予一个先验值$p(i)=s(i,i)$表示数据点i被选作聚类中心的倾向性，称为偏向参数，程序开始时取它的中位值。AP算法为选出合适的聚类中心而不断迭代更新并搜索信息：对每个数据点i为j搜集信息，用$r(i,j)$来描述数据点j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j搜集信息，用$a(i,j)$来描述数据点i选择数据点j作为其类代表的适合程度。$r(i,j)$与$a(i,j)$越大，点j作为最终聚类中心的可能性就越大。每个样本点通过反复迭代，各样本点进行竞争而得到最终的聚类中心。

3 可变相似性度量的近邻传播聚类算法思想

可变相似性度量主要是根据数据在观测空间的流形分布情况来区别对待数据集中的数据点：对于全局而言，通过函数变换完成对不同流形数据点对之间的相似性度量的缩小或放大；对于局部而言，通过对空间数据流形的搜索，识别不同流形的分布形状，将不同形状的数据分布映射成 AP 算法易操作的超球形或超椭球形的凸分布形态。

3.1 可变相似性度量的流形搜索思想与算法

由于流形是从观测空间的角度分析数据，可以直观地发现数据集(特别是高维数据)分布的内在规律性，而从数据分布的局部来看，主要考虑位于同一流形的数据点要具有较高的相似度或较低的不相似度。课题研究中采用一种边缘搜索流形的策略，即通过数据点近邻的边缘扩散搜索数据空间的流形分布情况。以下是数据点之间可达和相连关系定义，并以此关系搜索位于同一流形的数据点对，以便对其相似性度量进行变换。

定义 1 若对于论域中的所有样本点可以聚类成个ω簇，设$C_i(i=1,2,\cdots,\omega)$有N_t个样本$x_p^{(i)}(p=1,2,\cdots,N_t)$。$x_q^{(j)}\left(p=1,2,\cdots,N_r\right)$是样本集$D$上与$x_p^{(i)}$不同簇内的样本点，$\|d_{x_p^{(i)},\ x_q^{(j)}}\|$是样本点$x_p^{(i)}$到$x_q^{(j)}$的欧氏距离，则类间间隙阈值$\sigma_{\text{Min-gap}}$的值被定义为是两个不同簇的样本间最短距离，表示如下：

$$\sigma_{\text{Min-gap}}=\min_{\substack{i,j=(1,2,\cdots,w)\\ p,q=(1,2,\cdots,N_r,\cdots,N_t)}}\|d_{x_p^{(i)},\ x_q^{(j)}}\|$$

定义 2 令$C_1,C_2,\cdots,C_w$是样本集中分别关于参数$\sigma_{\text{Min-gap}}$构成的一个划分，$C_i\ (i=1,2,\cdots,w)$是这个划分中的一个簇，设定噪声阈值δ(δ为小簇内的样本点的个数)，当簇内的样本点的个数$\le\delta$时(通常设定≤ 3)，判定该簇为孤立簇，即：$\{p\in D\mid\forall i,p\in C_i$且$\|C_i\|\le\delta\}$，而$p$为孤立点。

定义 3 (边界判则)若点$x_p^{(i)}$为簇C_i中的样本点，$x_p^{(i)}$的邻域$\sigma_i\ (\sigma_i>\sigma_{\text{Min-gap}})$内包含至少一个点属于$C_i$且至少一个点不属于$C_i$，则$x_p^{(i)}$为簇$C_i$的一个临界样本(或称为簇$C_i$的一个边界点)。簇$C_i$中所有临界样本点构成簇$C_i$的边界。

这里，根据$\sigma_{\text{Min-gap}}$的定义，在临界样本点$x_p^{(i)}$下以参数为$\sigma_{\text{Min-gap}}$邻域内不可能有其它簇样本点。

定义 4[9] 若样本点$x_p^{(i)}$是簇C_i中的样本点，则在以$x_p^{(i)}$为圆心，半径为$\sigma_{\text{Min-gap}}$邻域内的所有样本点称为$x_p^{(i)}$的近邻样本，即$\|d_{x_p^{(i)},x_q^{(j)}}\|<\sigma$。并设定近邻样本与样本点$x_p^{(i)}$之间关于$\sigma_{\text{Min-gap}}$直接可达。

定义 5 若样本点$x_q^{(i)}$是$x_p^{(i)}$在参数为$\sigma_{\text{Min-gap}}$下的近邻样本，而$x_r^{(i)}$是$x_q^{(i)}$在参数为$\sigma_{\text{Min-gap}}$下的近邻样本，则$x_r^{(i)}$称为$x_p^{(i)}$在参数为$\sigma_{\text{Min-gap}}$下的近邻样本$x_q^{(i)}$的扩散。

图 1 中的$x_p^{(i)}$，$x_q^{(i)}$都是属于同一个簇，$x_p^{(i)}$通过$x_q^{(i)}$扩散到$x_r^{(i)}$，也就是通过$x_q^{(i)}$把$x_r^{(i)}$聚类到与自己相同的簇中，可以看出$\|d_{x_px_q}\|$，$\|d_{x_qx_r}\|$都小于$\sigma_{\text{Min-gap}}$，而$\|d_{x_px_r}\|<2\sigma_{\text{Min-gap}}$。

定义 6[8] 令$D=\{x_1,x_2,\cdots,x_n\}$是一个样本总数为n的样本集，x_i是样本集D中的一个样本点，且$\|d_{x_i,x_f}\|>\sigma_{\text{Min-gap}}$，则集合$F=P\{x_f\mid x_f\in D,$且$x_f$

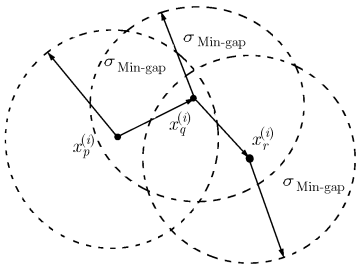

图 1 样本点在参数为$\sigma_{\text{Min-gap}}$的近邻扩散

是从x_i关于$\sigma_{\text{Min-gap}}$可达} 是一个关于阈值 $\sigma_{\text{Min-gap}}$ 的簇。这里的簇不一定是自然簇。

定理 1　若样本集上有样本点 a 和样本点 b 在样本集 D 上属同一簇 C，且 C 上的样本数大于 $n+2$，则必然能在该样本集上找到这样一个样本点序列 $a,x_1,x_2,\cdots,x_i,x_j,\cdots x_n,b$，使得样本点序列中任意相邻两点 x_i,x_j 之间的距离 $d_{ij}<\sigma_{\text{Min-gap}}$。

证明　若 $a,x_1,x_2,\cdots,x_i,x_j,\cdots x_n,b$ 中有相邻两点 x_i,x_j，且 $\| d_{x_i,x_j} \| \geq \sigma$，则 x_i,x_j 之间在经有限次扩散迭代后不可达，即 x_i,x_j 不可能属于同一个簇。除非在 x_i,x_j 之间还有别的样本点，这样与 x_i,x_j 之间是相互相邻的两点矛盾。　　证毕

同理，如果样本集上的两点 a,b 是属于同一簇 C，且 C 上的样本数大于 $n+2$，则在样本集 D 上必然能找到一个样本序列 $a,x_1,x_2,\cdots,x_i,x_j,\cdots x_n,b$，使得其序列上的点满足从样本点 a 到样本点 b 点点可达。如果 a,b 之间点点可达，称为 a,b 相连。

同时，这里不需要核函数及样本点密度的支持，以适应高维稀疏样本的情况。

流形搜索算法：

算法1　流形搜索算法如表1所示。

表1 流形搜索算法

输入：n 个数据点 $\{x_i \mid i=1,2,\cdots,n\}$，$\sigma$，$\mu$

输出：$C_1,C_2,\cdots,C_w$

(1) C_{temp} 为临时簇，sum_{temp} 为临时簇中样本总数，k 为聚类的类标号，i 为 C_{temp} 中的样本点号，n 为样本总数；μ 为孤立簇判定阈值；

(2)初始化类间间隙阈值 σ；

(3)while $\mu \neq 0$，任取其中一个样本点 x_i 到一个类 C_1 中，并与剩下的其它样本点比较，搜索近邻样本，计算距离，若 $d<\sigma$ 则放到类 C_{temp} 中；sum_{temp} ++；if $C_{1\_\text{sum}} \geq \mu$，end do；

(4)在 C_{temp} 中任取一样本 B_i，放到类 C_1 中，计算与除类 C_1、C_{temp} 中的所有样本与 B_i 的距离，

若 $d>\sigma$，则 i++，

　　若 $d<\sigma$，放入 C_{temp} 类中，sum_{temp} ++；

则 i++，

$\mu=\mu-1$，

$i=\text{sum}_{\text{temp}}$，$k$++；清除 C_{temp} 退出；

(5)while $\mu \neq 0$，再任取剩下一样本点放入 C_k 中，并与除 C_1 剩下的其它样本点比较，若 $d<\sigma$ 则放到类 C_{temp} 中；sum_{temp} ++；$\mu=\mu-1$；

(6)去除孤立簇。

在表 1 中，$C_1,C_2,\cdots,C_w$ 只是数据集基于参数 $\sigma_{\text{Min-gap}}$ 的一个划分。另外，当 $\mu=1$ 时，该算法只有一次循环，当 $\mu>1$ 时，即孤立簇内的样本点大于1，算法需要判定孤立簇内的样本点是否存在间接联通的情况。例如，孤立簇内的样本点数为 2，则还要判定这两个样本点是否还有其它近邻，若有则不能判定该簇为孤立簇。以上算法的平均时间复杂度：$O(n^2)$，平均空间复杂度：$O(n)$。

3.2 可变相似性度量的全局尺度度量

为了增大同一流形区域内的数据点之间的相似性度量，同时缩小不同流形区域内的数据点之间的相似性度量，AP聚类算法中的相似性度量可以从两个方面来考虑：

其一，从数据分布空间的全局来看，靠近中心或有可能成为聚类中心的数据点及其近邻的其它数据点比较集中，并有可能成为聚类中的同一簇，所以要尽量缩短其距离，而距离中心点较远的数据点尽量扩大它们之间的距离。据此，考虑指数函数 $a^x\ (-\infty<x<+\infty)$，函数曲线呈凹形递增。当 $a>1$，$1>x>0$时，a^x 的值域为 $(1,a)$，值域变化的幅值有限，但是当 $x>1$时函数的值域变化相当大，函数曲线也突然变得很陡。由此我们利用指数函数的这个特性设计可变距离度量，写成

$$\text{Dist}_{ij}(x_i,x_j)=\theta^{\left(D(x_i,x_j)/\tau_{ij}\right)}-1 \tag{1}$$

这里 $D(x_i,x_j)=\dfrac{d(x_i,x_j)}{\max(d(x_{i'},x_{j'}))}$，是归一化后的欧式距离，$d(x_i,x_j)$ 为数据点之间的欧氏距离，$\max(d(x_{i'},x_{j'}))$ 是所有数据点之间最大的距离；τ_{ij} 是为了调整 $D(x_i,x_j)$ 的幅值；$\theta\ (\theta>1)$为尺度伸缩参数。该类似方法在其它文献也有用到，例如：王玲等人在文献[8]中，使用该思想作为定义一种线段的长度，但是没有设定 τ 幅值限制参数。

事实上，作为数据点间的相似性度量是不一定需要满足欧式距离的三角不等式的，并且AP算法也不需要相似性度量的值小于1，所以这里取其负值作为数据点之间的不相似性度量，即 $S_{ij}(x_i,x_j)=-\text{Dist}_{ij}(x_i,x_j)$。这种全局相似性度量设计使与类代表点较近的数据点的相似性测度变化缓慢，而与待定类代表点较远的数据点相似性测度变化加大，从而更有利于总体数据的聚类。

3.3 可变相似性度量的局部尺度度量

经过上面的流形搜索，我们可以得到关于数据集中数据的流形分布情况，改进数据局部流形中的各数据点之间测度问题。通过对局部数据点之间的相似性度量变换，将任意流形的数据集转换为超椭球或超球形状的。这样可以对每个流形进行局部处理，使得到相似度矩阵更合理，更能反映数据集中数据分布的实际情况，当然也容易被AP等聚类算法识别，提高算法的准确性。

假定：数据点 x_p 和 x_q 是关于 ε 相连的，则数据点 x_p 和 x_q 是关于 ε 在数据集 D 同一流形上，$x_p \neq x_q$。则定义 x_p 和 x_q 的距离度量为

$$\mathrm{Dist}(x_p,x_q)=\theta^{\left(\frac{D(x_p,x_q)}{\tau_{pq}\cdot\bar{d}(x_p,x_q)/\varepsilon}\right)}-1\big|_{p\neq q}$$

$$\approx\theta^{\left(\frac{\varepsilon}{\tau_{pq}\cdot\max(d(x_p,x_q))}\right)}-1,\ x_i,x_j\text{关于}\varepsilon\text{相连} \quad (2)$$

其中 $\bar{d}(x_p,x_q)/\varepsilon$ 为 x_p 和 x_q 路径的曲线距离/ε。这样可以将同一流形的数据点通过变换转换到以 x_p 为中心的近似超球上来。如图2所示：

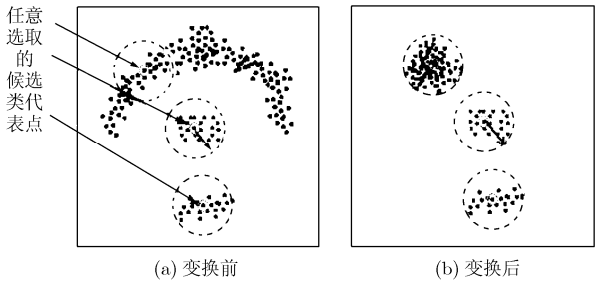

图2 任意形状中位于同一流形数据点之间度量变换前后，其数据点之间的关系

这里不去掉 θ 尺度伸缩参数，因为当 ε 设定较小时，所形成的类数较多，可能出现同一流形的数据在聚类后不在同一簇，这时 θ 尺度伸缩参数可以合并相近类形成聚类最终的簇。需要指出的是：按表达式(2)“$\approx$”前的表达式进行流形转换，其结果是超椭球；按“$\approx$”后的表达式转换，结果是超球；在当所涉及数据有重叠或重叠严重时将逐步减小 ε，这样可以将数据集中的数据以最近邻的方式进行优化。

由于AP算法是不需要进行度量的归一化的，所以对于所有数据点，当 $x_i \neq x_j$，其不相似性度量为

$$S_{ij}(x_i,x_j)=\begin{cases}-\theta^{\left(D(x_i,x_j)/\tau_{ij}\right)}+1, & x_i,x_j\text{关于}\varepsilon\text{不可达}\\ -\theta^{\left(\frac{\varepsilon}{\tau_{ij}\cdot\max(d(x_i,x_j))}\right)}+1, & x_i,x_j\text{关于}\varepsilon\text{相连}\end{cases} \quad (3)$$

3.4 基于可变相似性度量的 AP 算法(AP-VSM)

算法 2 可变相似性度量的 AP 算法(AP-VSM)如表 2 所示。

表 2 AP 算法

输入：$S(x_i,x_j)$，$r(i,j)=0$，$a(i,j)=0$，p_f，maxits，λ，conv，ε，θ，τ_{ij}

输出：划分后的各聚类族以及 SiI，FMI 等各类指标

(1)初始化 $r(i,j)=0$，$a(i,j)=0$，p_f，maxits，λ，conv，ε，θ，τ_{ij}

(2)建立不相似性矩阵，其中：

$$S_{ij}(x_i,x_j)=\begin{cases}-\theta^{\left(D(x_i,x_j)/\tau_{ij}\right)}+1, x_i, & x_j\text{关于}\varepsilon\text{不可达}\\ -\theta^{\left(\frac{\varepsilon}{\tau_{ij}\cdot\max(d(x_i,x_j))}\right)}+1, & x_i,x_j\text{关于}\varepsilon\text{相连}\end{cases}$$

(3)利用 AP 原理聚类，AP 中的 $r(i,j)$ 与 $a(i,j)$ 的按如下规则更新：

$$r(i,j)\leftarrow\lambda r(i,j)+(1-\lambda)\{s(i,j)-\max_{j'\neq j}[a(i,j')+s(i,j')]\};$$

$$a(i,j)\leftarrow\lambda a(i,j)+(1-\lambda)\min_{i\neq j}\{0,r(j,j)+\sum_{i'\neq i,i'\neq j}\max[0,r(i',j)]\};$$

$$a(j,j)\leftarrow\sum_{i'\neq i,i'\neq j}\max[0,r(i',j)]。$$

(4)查看 AP 算法是否收敛，若收敛则计算 SiI[7]，FMI[9,10]等各类指标并寻找最优解。

在表 2 中 maxits 为 AP 最大迭代次数；λ 为阻尼系数，用它来防止聚类过程中因不收敛而产生的振荡；conv 是收敛迭代系数。由于 AP 聚类的偏向参数 $p(i)$ 在聚类过程中占有非常重要的地位，通常 AP 算法的初始值设置成为其中值，但是在实际运行中并不能产生最优解，所以这里设置了 $p(i)$ 的因子参数 p_f，以调节 $p(i)$ 的初始值。

有效性分析指标 由于聚类结果的好坏优劣一直都没有统一的标准，我们只能通过各种指标来近似估算聚类质量并寻找出聚类的最优解和判定算法的有效性。对于上述算法中我们利用一个内部指标——SiI指标(Silhouette Index)[7]来具体评价聚类的质量和算法的有效性。对于 n 个样本 k 个聚类 $C_i(i=1,2,\cdots,k)$，它的 SiI 指标 $\mathrm{SiI}(t)=\dfrac{\min(\bar{d}(t,C_i))-\bar{d}(t,C_j)}{\max(a(t),b(t))}$，其中 $\bar{d}(t,C_j)$ 为聚类 C_j 中的样本 t 与 C_j 内所有其他样本的平均相似度(或距离)，$\bar{d}(t,C_i)$ 为 C_j 的样本 t 到另一个类 C_i 的所有样本的平均相似度(或距离)。记 $\mathrm{SiI}_{av}(C_i)$ 为一个聚类 C_i 的所有样本的 $\mathrm{SiI}(t)$ 平均值，它反映了类 C_i 的紧密性和可分性，而一个数据集的所有样本的 $\mathrm{SiI}(t)$ 平均值 $\mathrm{SiI}_{av}(C)$ 则可以反映聚类结果的质量，一般是值越大越好，这里用SiI指标来判定并寻找聚类算法中的最优解。

本算法的计算复杂度主要是在于构建相似度矩阵和使用 AP 算法进行聚类的时间，而构建相似度矩阵的时间是 $O(n^2)$。对于 AP 算法的时间复杂度主要是由其迭代次数决定。所以全部的时间复杂度不低于 $O(n^2)$ 这个量级，但是最大不高于 AP 聚类算法最大迭代次数所消费的时间(通常也不会达到，除非不收敛)，值得注意的是收敛迭代系数 conv 如果设定较大，也会大幅增加 AP 的聚类时间，实验中设定为 50。但是由于使用了上述对于相似性矩阵的优

化方法,仿真实验表明:AP 聚类的迭代次数将减少,聚类时间也将减少,特殊情况下甚至不到原来 AP 聚类时间的 1/5。

4　实验及其分析

本节对结合多种尺度对AP算法的聚类性能进行实验比较, 以检验聚类算法的有效性并结合聚类有效性指标找出正确或更优的聚类结果。

仿真实验是在Matlab 2006a仿真平台上进行的,硬件环境为CPU PM 1.6G 双核,主存为1 G。

AP聚类算法中设置初始值:$P(i)$设定为中值,对最大迭代次数为maxits=3000,收敛迭代系数conv=50[1]。

这里我们使用Fowlkes and Mallows指标(Fowlkes-Mallows Index,FMI)[9,10]去判定算法的正确率。计算方法为:$\mathrm{FMI}=F(C,C')=\dfrac{C_{11}}{\sqrt{((C_{11}+C_{10})(C_{11}+C_{01}))}}=\sqrt{(W(C,C')W'(C,C'))}$,其中$C$,$C'$是$k$个聚类中两个不同的类,$C_{11}$表示在$C$,$C'$上的同一类数据对的数量;$C_{01}$表示在$C'$上但不在$C$上的同一类数据对的数量,$C_{10}$表示在$C$上但不在$C'$上的同一类数据对的数量,$C_{00}$表示不在$C$,$C'$上的同一类数据对的数量,这里:$C_{00}+C_{01}+C_{10}+C_{11}=n(n-1)/2$。通常FMI值越大,正确率越好。

这里使用的原 AP 聚类 Matlab 源程序,来源于 Toronto 大学的 Probabilistic and Statistical Inference Group 主页 http://www.psi.toronto.edu/affinitypropagation /,而 6 个数据集均来源于 UCI(http://archive.ics.uci.edu/ml/)数据集,表 3 详细说明其样本数和维数以及真实类数。

从表 4 的仿真结果我们可以看出:(1)在消费时间上,AP-VSM 时间比原 AP 聚类时间要短,这并不是说建立不相似性矩阵的时间也比原来缩短,事实上建立不相似矩阵的时间比原来增加了,但是由于改善了不相似性矩阵使得 AP 的迭代次数比以前减少,导致总体时间也因此减少,例如:数据集 Ionosphere 的总体时间缩短了近 90%;其中数据集 Wine 与 Ionosphere 采用文献[7] adAP 的数据(减维后),并与 adAP 的实验结果进行比较,精度要好于 adAP。(2)在聚类的类数上,AP-VSM 聚类比原聚类更能接近数据集反映的聚类类数,这主要是新的相似度度量更能体现数据空间中的数据分布特性,所得到的不相似性矩阵能使 AP 在聚类过程中更容易获取数据的具体分布。

表 3 数据特征

数据集	样本数	维数	真实类数
Wine	178	3	3
Ionosphere	351	4	2
Parkinsons	197	21	2
Ozone_eighthr	2534	72	2
vehicle	92	18	4
vowel	990	10	11

图 3 说明在聚类的精度上,AP-VSM 聚类比原聚类有显著提高,特别是对高维数据的聚类更是如此,例如 Ozone_eighthr 数据集。这体现了当采用新的相似性度量后得到的不相似性矩阵更接近块随机矩阵,更能体现数据的本质聚类结构,因而使得 AP 的聚类性能得到大幅提升。当数据集中的各类数

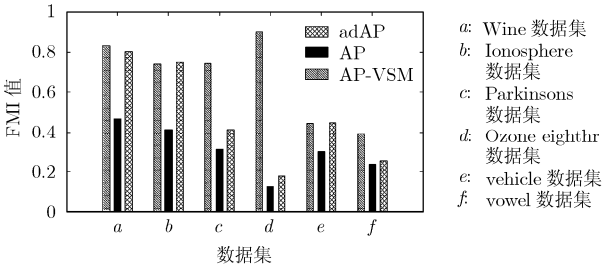

图 3 各数据集 FMI 值

表 4 AP-VSM 与 AP 的聚类结果

数据集	原类数	AP-VSM 类数	AP-VSM 建立不相似矩阵时间(s)	AP-VSM 聚类时间(s)	AP 类数	AP 建立不相似矩阵时间(s)	AP 聚类时间(s)	adAP 类数	adAP 聚类时间(s)
Wine	3	3	0.150447	0.496753	11	0.125891	0.532106	3	34.5
Ionosphere	2	2	0.576548	3.556293	42	0.569718	30.73168	2	445.3
Parkinsons	2	2	0.184654	0.644299	14	0.159041	0.71436	8	301.965
Ozone_eighthr	2	2	31.2434	216.8689	81	27.8362	231.2448	10	97175.4
vehicle	4	4	0.0552643	0.2884	6	0.0536792	0.1344	2	57.6916
vowel	11	11	4.58353	23.1464	78	4.13147	26.9266	47	20341.7

注:聚类前对 Ozone_eighthr 中的不全数据进行补零处理。

据之间重叠严重时，AP-VSM 也比原 AP 精度要高，例如第 5 个数据集 vehicle 和第 6 个数据集 vowel。也可以看出，合适的相似性度量对于聚类算法中聚类性能的改变是很大的。

综上所述，由于选用不同的相似性度量，AP-VSM 比原 AP 精度要高，也比改进算法 adAP 精度要高一些，且总体时间消费比原 AP 和 adAP 少，故总体上 AP-VSM 比以负欧氏距离作为相似性测度的原 AP 算法效果更好，性能更优越。

5 结束语

本文对数据在空间中的局部分布和总体分布两个方面的流形分布特性进行了研究，提出了基于一种可变相似性度量的近邻传播聚类算法。它采用可变度量来改善数据的总体分布特性，并利用近邻边缘搜索思想搜索局部流形以及非线性映射，从数据分布空间的局部流形方面去改进位于同一流形的数据点之间的相似性度量，通过这样变换、映射得到的数据点之间的相似性矩阵能更好地反映和体现数据在空间中的复杂分布(特别是高维数据)。因此，在此相似性度量上进行 AP 聚类可以明显改善聚类的迭代次数，提高聚类的效率和正确率，拓展 AP 算法处理多种数据的能力。仿真实验表明：它比以负欧氏距离作为相似性测度的原 AP 算法效果更好，性能更优越。以后我们将研究如何改善 AP 聚类的精度和缩短聚类的时间，为 AP 算法的扩展应用提供研究基础。

参 考 文 献

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

[2] Givoni I E and Frey B J. A binary variable model for affinity propagation. *Neural Computation*, 2009, 21(6): 1589–1600.

[3] Jia Sen, Qian Yun-tao, and Ji Zhen. Band selection for hyperspectral imagery using affinity. Propagation. Proceedings of the 2008 Digital Image Computing: Techniques and Applications, Canberra, ACT, 1–3.12.2008: 137–141.

[4] Gang Li, Lei Guo, and Liu Tian-ming, *et al*.. Grouping of brain MR images via affinity propagation. IEEE International Symposium on Circuits and Systems, 2009 (ISCAS 2009) Taipei, Taiwan, 5.24. 2009: 2425–2428.

[5] Dueck D, Frey B J, and Jojic N, *et al*.. Constructing treatment portfolios using affinity propagation[C]. Proceedings of 12th Annual International Conference, RECOMB 2008. Singapore. 3.30–4.2, 2008: 360–371.

[6] Leone M, Sumedha, and Weigt M. Clustering by soft-constraint affinity propagation: applications to gene-expression data. *Bioinformatics*, 2007, 23(20): 2708–2715.

[7] 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.
Wang Kai-jun, Zhang Jun-ying, and Li Dan. Adaptive affinity propagation clustering. *Acta Automatica Sinica*, 2007, 33(12): 1242–1246.

[8] 王玲, 薄列峰, 焦李成. 密度敏感的半监督谱聚类. 软件学报, 2007, 18(10): 2412–2422.
Wang L, Bo L F, and Jiao L C. Density-Sensitive semi-supervised spectral clustering. *Journal of Software*, 2007, 18(10): 2412–2422.

[9] Alexander Hinneburg and Daniel A Keim. A general approach to clustering in large databases with noise. *Knowledge and Information Systems*, 2003, 5(4): 387–415.

[10] Little M A, McSharry P E, Hunter E J, and Lorraine O. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 2009, 56(4): 1015–1022.

董 俊： 男，1973 年生，博士，研究方向为现代网络和多媒体技术、人工智能.

王锁萍： 男，1946 年生，教授，博士生导师，研究方向为现代网络与多媒体技术.

熊范纶： 男，1940 年生，研究员，博士生导师，研究方向为人工智能与农业信息工程.