基于密度调整和流形距离的近邻传播算法

夏春梦[1,2] 倪志伟[1,2] 倪丽萍[1,2] 张 霖[3]
(合肥工业大学管理学院 合肥 230009)[1]
(合肥工业大学过程优化与智能决策教育部重点实验室 合肥 230009)[2]
(北京航空航天大学自动化科学与电气工程学院 北京 100191)[3]

摘 要 针对近邻传播聚类算法在构造相似度矩阵时因对多重尺度和任意形状数据敏感而聚类效果不理想的缺陷,提出一种基于密度调整和流形距离的近邻传播算法。该算法将“领域密度”和“流形理论”的思想引入近邻传播算法,利用基于密度调整和流形的距离更好地刻画了样本空间的真实分布状况,解决了相似度矩阵不能充分表示数据之间内在关系的问题,在一定程度上提高了近邻传播聚类算法的聚类效果。通过在人工数据集和标准数据集上进行实验对比,验证了算法的有效性和优越性。

关键词 近邻传播聚类,密度调整,流形相似度,多重尺度数据集,任意形状数据集
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.035

Affinity Propagation Clustering Algorithm Based on Density Adjustment and Manifold Distance

XIA Chun-meng[1,2] NI Zhi-wei[1,2] NI Li-ping[1,2] ZHANG Lin[3]
(School of Management, Hefei University of Technology, Hefei 230009, China)[1]
(Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei University of Technology, Hefei 230009, China)[2]
(College of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China)[3]

Abstract As affinity propagation (AP) clustering is sensitive to the dataset with scaling parameter and various form while calculating the similarity matrix and the cluster result is not ideal, an affinity propagation clustering algorithm based on density adjustment and manifold distance was proposed. The algorithm introduces local density of data and manifold theory into affinity propagation clustering, and uses a way of distance measure based on manifold structure and density adjustment to describe the clusters' actual structure better, making up the similarity matrix's deficiency. At the same time, the algorithm is more efficient. Simulation experiment was done on artificial datasets and standard datasets. The result shows the effectiveness and superiority of proposed algorithm.

Keywords Affinity propagation clustering, Density adjustment, Manifold similarity, Multi-scale dataset, Various form dataset

1 引言

聚类分析是数据挖掘的主要任务之一,其目的是按照个体或样本的特征对样本分类,使同一类别内的样本具有尽可能高的同质性,使类别之间的样本具有尽可能高的异质性。随着信息技术的迅猛发展,数据聚类被广泛应用于数据挖掘、模式识别、机器学习、图像分割和生物信息处理等领域[1]。如今,聚类算法有很多种,包括传统的K-means算法[2]、K-medoids算法[3]、FCM算法[4]等。这些经典的基于划分的算法对初始聚类中心敏感,需要人为确定聚类类数;且初始点选择的随机性引起了聚类结果的不稳定,容易陷入局部最优。2007年,Fray等人在《Science》杂志上提出了近邻传播聚类算法(Affinity Propagation Clustering, AP)[5-6],解决了聚类结果对初始聚类中心敏感的问题。它根据N个数据点之间的相似度进行聚类,不需要事先指定聚类数目;由于它将所有数据点都作为潜在的聚类中心,因此不受限于初始类代表点。与传统的聚类方法相比,AP算法在处理多类、大规模数据时能得到较好的聚类结果,而且算法的运行结果相对稳定[7-8]。AP算法对相似度矩阵的对称性没有要求,适应范围更广,目前已经成功应用于图像分割[9]、目标跟踪[10]、用户社区识

到稿日期:2016-09-12 返修日期:2016-12-31 本文受国家“863”云制造主题项目(2015AA042101),国家自然科学基金重大研究计划培育项目(91546108),国家自然科学基金项目(71271071,71301041)资助。

夏春梦(1992-),女,硕士生,主要研究方向为数据挖掘,E-mail:240877306@qq.com;倪志伟(1963-),男,博士,教授,主要研究方向为数据挖掘、机器学习等,E-mail:zhwnelson@163.com;倪丽萍(1981-),女,博士,副教授,主要研究方向为分形数据挖掘等,E-mail:niliping@hfut.edu.cn(通信作者);张 霖(1966-),男,博士,教授,主要研究方向为云制造、复杂服务网络与多智能体等,E-mail:zhanglin@buaa.edu.cn。

别[11]等领域，是一种极具竞争力的聚类算法。

AP 算法比其他聚类算法具有更优越的性能，但其仍是基于中心的聚类，在紧凑的具有超球形分布的数据上具有较好的聚类性能，但并不适用于任意形状和多重尺度数据集的聚类。为此，已有不少学者对近邻传播算法进行了研究。张震等考虑到数据流的先验分布信息，结合流形理论定义了一种流形相似度的距离测度，提高了流量分类器的性能[12]。董俊等综合数据的全局与局部分布特征，设计了一种流形搜索算法，提出了一种基于可变相似性度量的近邻传播算法[13]。以上算法着重于考虑数据的流形结构，使之更符合原始数据集的真实分布状况，但是没有对多重尺度数据问题进行研究。目前，大量学者通过引入密度调整(这里的密度调整是指依据数据点周围的密度分部信息来调整相似度的计算)的思想对谱聚类相似度计算进行改进，以解决多尺度数据聚类问题。Zelnik-Manor 等提出了基于谱聚类的 Self-Tuning 算法[15]，在计算相似度的过程中加入了数据点的领域信息，使数据点周围的密度分布对它们之间的相似度产生作用，更真实地反映了数据之间的内在联系。王雅琳等进一步对添加领域信息的相似度进行改进，得到了较好的聚类结果[16]。

本文将领域信息引入近邻传播算法中并加以改进，提出一种基于密度调整和流形距离的近邻传播算法。通过添加数据点周围的密度参数，使得算法可以更好地处理密度不均匀的数据集；同时构造 K 近邻网络，根据流形距离进一步调整相似度矩阵，使其更符合真实的数据点的内在结构。

2 近邻传播算法

假设数据集 D 有 n 个样本$\{x_1,x_2,\cdots,x_n\}$，AP 算法首先计算每两个样本点之间的相似度，通过相似度来计算吸引度(responsibility)和归属度(availablity)，结合吸引度和归属度两方面的信息找到最优的类代表点集合(exemplars)，最终使得所有数据点到其最近的类代表点的相似度之和最大。

定义 1(相似度矩阵 S) $s(i,k)$代表数据点 x_k 作为数据点 x_i 的聚类中心的合适程度。相似度矩阵可以是对称的也可以是不对称的，即 $s(i,k)$和 $s(k,i)$可以不相等，因此 AP 算法能较好地解决非欧氏空间问题。

S 可以通过很多方法来度量，一般情况下，为方便后续计算，选择欧氏距离平方的负值来衡量，如式(1)所示：

$$s(i,k)=-\|x_i-x_k\|^2 \tag{1}$$

其中，当 $i=k$ 时 $s(i,k)$代表偏向参数 $P(i)$，$P(i)$越大，x_i 点作为聚类中心的可能性就越大。初始化算法时，所有的 $P(i)$都取相同的 P 值，即所有数据点成为聚类中心的可能性相同，P 值越大，得到的类簇个数也越多，P 通常取相似度矩阵的中值。

定义 2(吸引度矩阵 R) 吸引度 $r(i,k)$是由点 x_i 发送到候选聚类中心点 x_k 的消息，表示点 x_k 适合作为点 x_i 的聚类中心的程度。$r(i,k)$越大，表示 x_k 作为类代表点的可能性越大。

定义 3(归属度矩阵 A) 归属度 $a(i,k)$是由候选聚类中心点 x_k 发送到 x_i 的消息，表示点 x_i 选择点 x_k 作为其聚类中心的程度。$a(i,k)$越大，表示点 x_i 隶属于点 x_k 的可能性越大。

吸引度和归属度的关系分别如图 1 和图 2 所示。

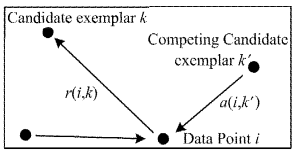

图 1 吸引度的传递

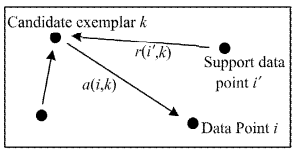

图 2 归属度的传递

AP 算法的具体实现过程如下。

步骤 1 根据式(1)计算相似度矩阵，其中 $P(i)$等于 P。

步骤 2 吸引度矩阵和归属度矩阵的交替更新过程(初始时 $a(i,k)=0$)：

$$r(i,k)=\begin{cases}s(i,k)-\max\limits_{k'\neq k}\{a(i,k')+s(i,k')\}, & i\neq k\\ P(k)-\max\limits_{k'\neq k}\{a(k,k')+s(k,k')\}, & i=k\end{cases} \tag{2}$$

$$a(i,k)=\begin{cases}\min\{0,r(k,k)+\sum\limits_{i'\notin\{i,k\}}\max\{0,r(i',k)\}\}, & i\neq k\\ \sum\limits_{i'\neq k}\max\{0,r(i',k)\}, & i=k\end{cases} \tag{3}$$

在更新 R 和 A 的同时，AP 算法为避免发生振荡引入了一个阻尼系数 $lam\in[0,1)$对 R 和 A 进行缩放，阻尼系数越大，消除振荡的效果越好，但是会减慢算法的收敛速度，通常默认为 0.5。缩放公式如下：

$$R=(1-lam)*R+lam*Rold \tag{4}$$

$$A=(1-lam)*A+lam*Rold \tag{5}$$

步骤 3 确定数据点 x_i 的聚类中心点 x_k，k 应该满足公式：

$$\arg\max(r(i,k)+a(i,k)) \tag{6}$$

其中，当 $i=k$ 时，数据点 x_i 本身就是聚类中心。

步骤 4 满足以下条件之一迭代终止：

1)超过预先设定的最大迭代次数；

2)$R+A$ 值的改变量低于某一固定阈值；

3)选择的类代表点在连续几步迭代过程中保持稳定。

AP 算法对 P 值敏感，迭代结束后可以通过改变 P 值重复运行算法得到满意的聚类结果。

3 基于密度调整和流形距离的近邻传播算法

3.1 密度调整距离

多重尺度数据集即密度不均匀的数据集，是指在各个簇中数据点密度差异较大的数据集。标准的近邻传播算法计算相似度时依据的是欧氏距离，无法完整地衡量多重尺度数据间的相似度，因此无法得到较好的聚类结果。

由文献[15-16]可知，数据点所处的领域内的密度影响着数据点之间的相似度，而且当两个数据点的领域内的密度越接近时，它们在同一类中的可能性就越大。文献[16]中通过 σ_i 表示点 x_i 到其第 T 个最近邻居点的欧氏距离，表示点的稀疏与稠密，通过权值$|\sigma_i-\sigma_j|/\sigma_{\max}$对相似度进行调整，并且得出两点之间的密度差别越大，相似度越小。

本文为了克服近邻传播算法对多重尺度数据集的聚类效果不理想的缺点，将领域密度的思想引入近邻传播算法的相似度计算中，得到了一个初始距离公式。由于近邻传播聚类的相似度一般取负值，因此在本文中距离公式取正值，并且要求两点之间的初始距离随着两点之间的欧氏距离的增大而增

大，随着密度差的增大而增大。得到的计算公式如式(7)所示：

$$A_{ij}=-\exp[-\frac{d^2(x_i,x_j)}{\bar{\sigma}^2}(1+\frac{|\sigma_i-\sigma_j|}{\sigma_{\max}})]+1 \tag{7}$$

其中，σ_i 表示点 x_i 到其第 T 个最近邻居点的欧氏距离；$\sigma_{\max}=\{\max|\sigma_i-\sigma_j|,i=1,\cdots,n;j=1,\cdots,n\}$；$\bar{\sigma}=\frac{1}{n}\sum_{i=1}^{n}\sigma_i$。

两点之间的相似度即为距离的负值。在利用式(7)计算相似度时，$\bar{\sigma}$ 起到减小噪声的作用。从式(7)可以看出，初始距离公式得到的结果是一个在[0,1)范围内上升的曲线，并且可以保证初始距离随着欧氏距离的增大而增大，随着密度差的减小而减小。

3.2　流形距离

AP 算法在处理任意形状的数据集时效果并不理想，本文从数据空间的流形分布出发，用流形搜索的方法区分出不同形状的簇。

流形思想源自于“流形理论”，考虑了数据集的全局一致性，从样本空间的整体分布来发现样本的内在规律[17]。很多文献将流形思想应用于数据分析中，以解决复杂的非线性流形结构的数据分析问题，通过流形学习的方法对数据进行降维，降低了问题的复杂性[18-19]。在复杂形状的数据集中，欧氏距离不能反映数据集的全局一致性，如图 3 所示，a 和 c 在同一流形分布中，a 与 c 的欧氏距离大于 a 与 b 的欧氏距离，若 a,b,c 的领域密度相差不大，则根据式(7)计算的 a 与 b 的相似度将大于 a 与 c 的相似度。

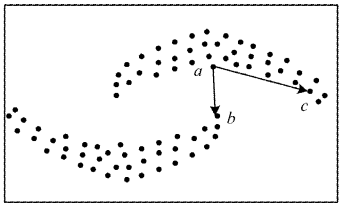

图 3　基于欧氏距离的样本空间分布示意图

为了搜索出不同形状的簇，本文选用 K 近邻法搜索流形，建立 K 近邻网络图，即将一个点与其距离最近的 K 个点之间用一条边连接起来，最终得到的结果是：处于同一个簇中的点之间相互连接，处于不同簇中的点无法连接在一起。由此可以区分出不同形状的不同簇结构。值得注意的是，本文在构造 K 近邻网络时选用的是上文提到的初始距离作为 K 近邻的距离测度，即构造 K 近邻网络的前提是基于密度敏感的距离测度。

为了更好地区分簇内结构和簇与簇之间的结构，分别对数据点与 K 近邻点相连的边和数据点与非 K 近邻点相连的边赋权，在密度敏感距离的基础上有效地增大了 K 近邻点与非 K 近邻点之间的差异。如果点 x_j 是点 x_i 的 K 近邻点，则它们之间的权重变化趋势为[0,1)上缓慢上升的曲线；如果点 x_j 是点 x_i 的非 K 近邻点，则它们之间的权重变化趋势为[0,+∞]上快速上升的曲线。赋权公式如式(8)所示：

$$A_{ij}=\begin{cases}-\exp[-\dfrac{d^2(x_i,x_j)}{\bar{\sigma}^2}(1+\dfrac{|\sigma_i-\sigma_j|}{\sigma_{\max}})]+1, & x_j\in K_i\\ \exp[\dfrac{d^2(x_i,x_j)}{\bar{\sigma}^2}(1+\dfrac{|\sigma_i-\sigma_j|}{\sigma_{\max}})]-1, & x_j\notin K_i\end{cases} \tag{8}$$

其中，K_i 代表点 x_i 的 K 近邻点集合。式(8)满足权重随着欧氏距离的增大而增大，随着密度差的减小而减小。通过式(8)缩短了簇内部的距离，增大了簇与簇之间的距离。

依据以上过程构建得到一个赋权无向图，再根据此无向图搜索最短路径：

$$S(x_i,x_j)=-\min_{q\in Q_{ij}}\sum_{k=1}^{|q|-1}A(q_k,q_{k+1}) \tag{9}$$

其中，Q_{ij} 代表连接 x_i 和 x_j 的所有路径的集合，q 代表连接 x_i 和 x_j 的一条路径，$|q|$ 代表路径上的节点个数，q_k 代表路径上的第 k 个点，$A(q_k,q_{k+1})$ 代表依据式(8)计算的路径上点 q_k 到点 q_{k+1} 的距离。

本文最终计算得到的距离空间是在[0,+∞]区间内的离散值，因此相似度空间是[−∞,0]。最终的相似度计算结合了领域密度和流形思想，更加真实地反映了数据集的内在分布，添加的尺度参数根据欧氏距离自适应调整，使得算法对尺度参数不敏感，并且可以根据本文的流形搜索算法区分出不同形状的簇，在处理密度不均匀和任意形状的复杂数据集聚类时具有更大的优越性。

3.3　算法的流程

本文将密度调整距离与流形距离相结合，得到了基于密度调整和流形距离的近邻传播算法(DAMS-AP)，具体如算法 1 所示。

算法 1　基于密度调整和流形距离的近邻传播算法(DAMS-AP)

输入：n 个数据点$\{x_1,x_2,\cdots,x_n\}$，K，T

输出：数据点的 C 个划分

Step1　根据式(7)计算每个点与其他点的初始密度调整距离；

Step2　利用初始密度调整距离选取 KNN 网络的每个点的最近邻的 K 个值；

Step3　利用式(8)对 KNN 网络赋权，得到最终的赋权矩阵 W；

Step4　根据式(9)对矩阵 W 求最短路径，并求出最终的相似度矩阵 S；

Step5　在达到迭代停止条件之前，根据式(2)－式(5)对相似度矩阵进行迭代更新；

Step6　根据式(6)选出最优的类代表点；

Step7　将数据点分配到最近的聚类中心。

调整 P 值，重新运行算法，直至得到满意的聚类结果。

4　实验结果及分析

本文的实验环境为：处理器 Intel(R)Core(TM)i3-2330M 2.2GHz，内存为 2GB，硬盘为 750GB，操作系统为 Windows7，编程语言为 Matlab2012a。

4.1　对比算法及参数设置

4.1.1　对比算法描述

为了验证本文所提算法的有效性，选用 AP，STI-AP[12]，DA-AP，IASCBDA[16]作为对比实验算法。

AP 算法选用传统的 AP 算法，实验代码来源于文献[5]；本文的 STI-AP 算法除去了半监督的信息，计算流形距离依据没有半监督信息更改的欧氏距离进行判断，主要是为了将本文算法与文献[12]中的基于流形相似度改进的近邻传播聚类算法进行对比；DA-AP 算法是基于密度敏感距离的 AP 算法，密度敏感距离由式(8)中非 K 近邻点的距离计算得到，其中数据点周边的密度差异越小，欧氏距离越小，相似度越大；

IASCBDA 算法是基于密度敏感相似度的谱聚类算法，密度敏感相似度由文献[16]中的相似度计算公式计算得到。本文的 DAMS-AP 算法是基于密度敏感和流形距离的 AP 算法，距离计算公式可由式(7)－式(9)计算得到。

4.1.2 算法参数设置

在近邻传播算法中，偏向参数 P 的选择影响着聚类质量。通常，P 值越大聚类数目越多，P 值越小聚类数目越少，但是类数和 P 值并非一一对应的关系，某些聚类数目对应的 P 值范围比较大，此时需要经过多次迭代使 P 值有较大的下降才能使聚类数目发生变化[14]。传统的近邻传播算法中 P 值初始设置为相似度矩阵的中位数，设 P_0 为相似度矩阵的中位数。AP 中还有一个参数为阻尼系数，其可以消除震荡，阻尼系数越大，消除震荡的效果越好，但会降低算法的收敛速度，一般将其设置为 0.5。

AP 算法：在 AP 算法中有 3 个 P 值可以利用，即相似度最大值 $S_{\max}$、相似度中位数 S_{median} 和相似度最小值 $S_{\min}$。一般情况下，P 值的合理范围为 $[S_{\min}, S_{\max}]$，但 AP 算法趋向于产生多于真实聚类数目的类，因此本文选用的 P 值范围是 $[S_{\min}, S_{median}]$，更新规则为 $P = S_{median} - \beta * (S_{median} - S_{\min})$，其中 β 值选取 0.1，0.2，…，0.9，1。阻尼系数 $lam=0.5$。

STI-AP 算法：距离公式的伸缩因子 $\theta_1 = 1/2$，$\theta_2 = 2$，阻尼系数为 0.5。基于前文的算法描述，STI-AP 算法最终的相似度范围是 $[-\infty, 0]$，因此 P 值的合理范围是 $[-\infty, 0]$。本文的 P 值按 P_0 倍数的形式下降(例如 $P_0, 2P_0, 3P_0, \cdots, 9P_0, 10P_0$)，也可以根据数据集相似度矩阵的差异适当减小或增大 P 值的下降速度。

DA-AP 算法：该算法主要计算密度敏感距离，计算过程中 $T=5$，由于 P 值的合理范围也是 $[-\infty, 0]$，因此 P 值的变换同样使用 STI-AP 算法中的倍数下降方法。

IASCBDA 算法：该算法中的领域密度度量参数 $K=7$，谱聚类选用 NJW 算法。

DAMS-AP 算法：本文算法中密度调整距离的参数 $T=5$，KNN 网络的 $K=5$，阻尼系数 $lam=0.5$，基于本文的相似度计算公式，算法最终的相似度范围是 $[-\infty, 0]$，因此 P 值的合理范围是 $[-\infty, 0]$，本文中 P 值的选取方式同 STI-AP 算法。

本文先后采用人工数据集和标准数据集对算法进行验证。

4.2 人工数据集验证

选用 4 个人工数据集(见图 4(a)－图 4(d))对算法进行验证。图 4(a)由两个密度不一样的簇组成，其中一个簇紧密，另一个簇稀疏；图 4(b)由 3 个流形形状组成，流形之间的距离较近；图 4(c)由不同形状的 3 个簇组成，即一个圈里包含两个正方形；图 4(d)包含两个不同形状的簇，两个簇的密度不同。使用本文所提算法以及对比算法对这 4 个数据集进行聚类，不同的类簇已在图中用不同的形状表示。

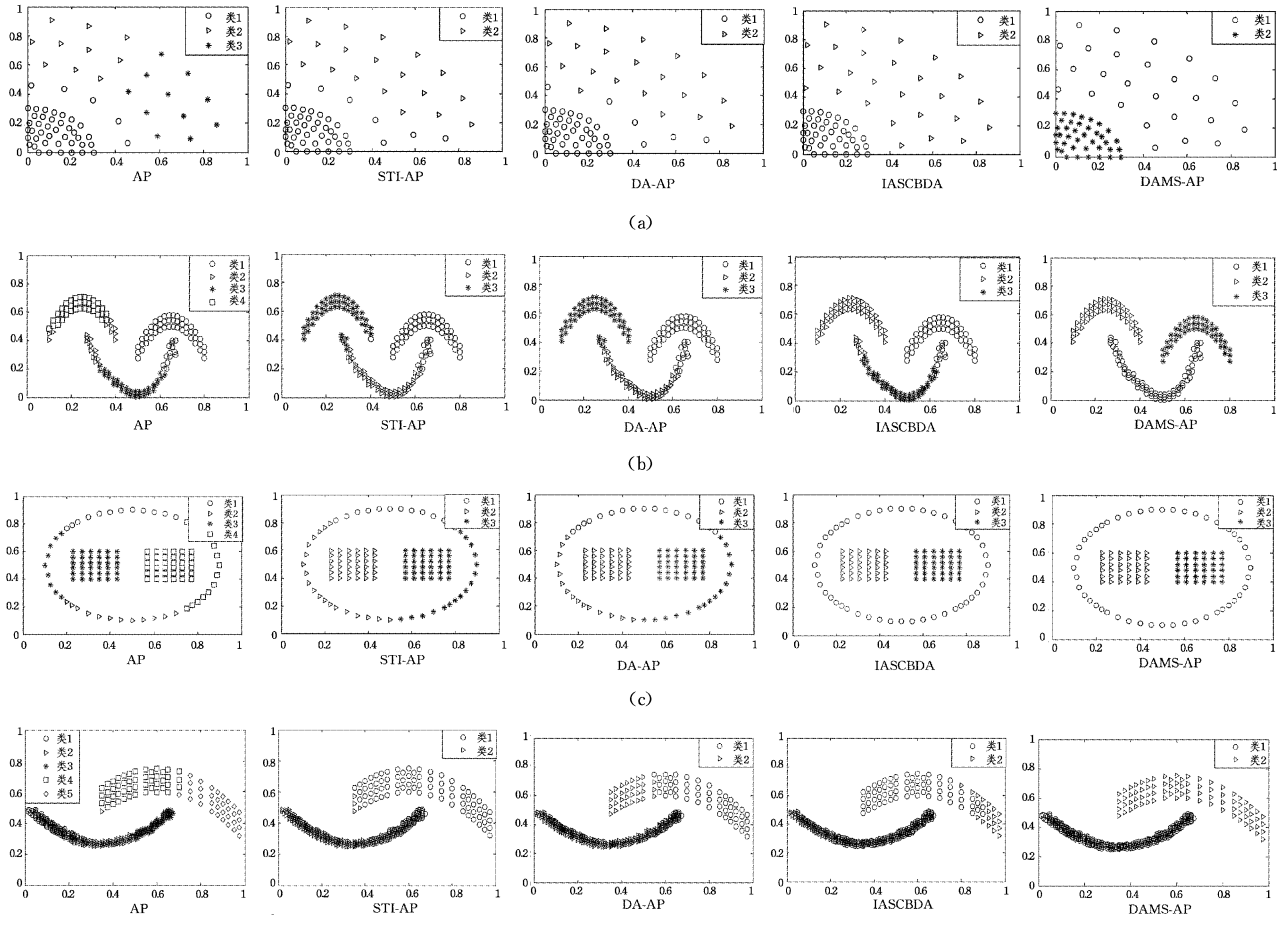

图 4 5 种算法的聚类结果

从图 4 中可以看出，按照前文的算法步骤，AP 算法在处理 4 个人工数据集时均无法得到准确的聚类个数，聚类效果较差；STI-AP 算法和 DA-AP 算法在处理这 4 个人工数据集时可以得到准确的聚类个数，虽然得到的聚类结果相对较差，但是 STI-AP 算法和 DA-AP 算法的性能相对于 AP 算法略有提高；IASCBDA 处理图 4(a)和图 4(c)的数据集时能有效地分出各个类簇，但处理图 4(b)和图 4(d)的数据集时效果不佳；而本文提出的算法对这 4 个数据集都能很好地聚出与正确结果相同的类簇，得到较理想的聚类效果。

表 1 列出了 5 种算法聚类 4 个数据集后出现的误分点个数。可以看出，AP 算法对 4 个人工数据集的误分率分别是 23%，22%，30%，78%，聚类结果较差；STI-AP 相对于 AP 算法的聚类效果稍好，对 4 个数据集的误分率分别为 11%，13%，29%，31%；DA-AP 算法对图 4(a)和图 4(d)的数据集的误分率为 6%和 21%，相对于前两种算法有一定提高，对于图 4(b)和图 4(c)数据集的误分数与 STI-AP 类似；IASCBDA 算法对图 4(a)和图 4(c) 的误分率都是 0，效果很好，它对图 4(b)和图 4(d)的误分率分别是 11%和 45%，误分率较大，相对前 3 种算法提升不大；本文算法 DAMS-AP 在最好的情况下对 4 个数据集能达到很好的聚类效果。由人工数据集的聚类结果可以看出，本文提出的基于密度调整和流形距离的近邻传播算法具有一定的优越性。

表 1　误分点数

数据集（点数）	误分点数				
	AP	STI-AP	DA-AP	IASCBDA	DAMS-AP
(a)(65)	15	7	4	0	0
(b)(169)	37	22	20	19	0
(c)(122)	36	35	31	0	0
(d)(344)	229	105	73	154	0

4.3　标准数据集验证

4.3.1　评价指标

F-measure 指标是常用的聚类算法评价指标，反映了聚类算法的准确程度。F-measure 的取值范围是[0，1]，F-measure的值越大表示算法越准确。F-measure 组合了“准确率”和“召回率”两个指标，准确率、召回率、F-measure 的公式如下。

假设算法运行的实验结果为 k 类：$C'=\{c_1',c_2',\cdots,c_k'\}$，真实数据集分为 t 类：$C=\{c_1,c_2,c_3,\cdots,c_t\}$，则准确率为：

$$Precision(c_i',c_j)=\frac{|c_i'\bigcap c_j|}{|c_i'|} \tag{10}$$

召回率为：

$$Recall(c_i',c_j)=\frac{|c_i'\bigcap c_j|}{|c_j|} \tag{11}$$

F-measure 为：

$$F(C',C)=\frac{1}{N}\sum_{j=1}^{t}|c_j|\times\max_{1\leqslant i\leqslant k}\left(\frac{2\times Precision(c_i',c_j)\times Recall(c_i',c_j)}{Precision(c_i',c_j)+Recall(c_i',c_j)}\right) \tag{12}$$

化简后的 F-measure 为：

$$F(C',C)=\frac{1}{N}\sum_{j=1}^{t}|c_j|\times\max_{1\leqslant i\leqslant k}\left(\frac{2\times|c_i'\bigcap c_j|}{|c_i'|+|c_j|}\right) \tag{13}$$

除 F-measure 指标外，本文还采用平均 NMI 值[20]共同作为聚类评价标准。平均 NMI 数值越接近 1 代表聚类结果越好，越接近 0 代表聚类结果越差。计算公式如式(14)所示：

$$NMI=\frac{I(C',C)}{\sqrt{H(C')\times H(C)}} \tag{14}$$

其中，$I(C',C)=\sum_{c'\in C',c\in C}p(c',c)\log\frac{p(c',c)}{p(c')p(c)}$，$H(C)=\sum_{j=1}^{t}p(c_j)I(c_j)$。

4.3.2　实验数据说明

实验中使用的标准数据集包括：wine，glass，balance-scale，ionosphere，segment。数据集的基本情况如表 2 所列。

表 2　标准数据集的基本情况

数据名称	样本个数	非标签属性个数	聚类数目
wine	178	13	3
glass	214	9	6
balance	625	4	3
ionosphere	351	34	2
segment	2310	18	7

4.3.3　实验结果及分析

将本文算法和 4 个对比算法在 5 个标准数据集上进行聚类，采用 F-measure 指标和 NMI 指标来衡量聚类效果的好坏。对每个数据集重复实验 20 次，取均值作为最终实验结果。聚类结果以及聚类类数如表 3—表 5 所列。

表 3　F-measure 指标对比

数据集	AP	STI-AP	DA-AP	IASCBDA	DAMS-AP
wine	0.6910	0.6910	0.6236	0.5787	0.6910
glass	0.4907	0.5000	0.5000	0.4767	0.5280
balance	0.1504	0.4336	0.4304	0.4928	0.5552
ionosphere	0.6211	0.6866	0.7179	0.7493	0.7949
segment	0.4795	0.4228	0.3987	0.1446	0.6472

表 4　NMI 指标对比

数据集	AP	STI-AP	DA-AP	IASCBDA	DAMS-AP
wine	0.4315	0.4360	0.4095	0.4176	0.4238
glass	0.3995	0.3926	0.4432	0.4002	0.3749
balance	0.1553	0.1870	0.1537	0.1183	0.1873
ionosphere	0.2696	0.2368	0.3590	0.3458	0.4049
segment	0.4789	0.5375	0.4758	0.0956	0.6438

表 5　算法的聚类类数比较

数据集	真实聚类	AP	STI-AP	DA-AP	DAMS-AP
wine	3	3	3	3	3
glass	6	6	6	8	6
balance	3	16	4	4	3
ionosphere	2	5	2	2	2
segment	7	7	13	13	7

从表 3 和表 4 中可以看出，STI-AP 算法的聚类效果相较于 AP 算法的聚类效果有一定的提高，STI-AP 对 wine，glass 和 balance 数据集的聚类效果比 AP 算法的聚类效果好；DA-AP 算法对 glass 和 ionosphere 数据集的聚类效果较好；IAS-CBDA 算法对 balance 和 segment 数据集的聚类效果较差；对比其他算法，本文提出的 DAMS-AP 算法对 balance，ionosphere 和 segment 数据集的聚类效果较好，对 wine 和 glass 数据集的聚类效果相对较差，但差异不明显。由此可见，本文所提算法与 4 种对比算法相比有一定的优越性。

表 5 列出了各个算法对数据集的聚类数目的对比，其中 IASCBDA 算法是基于谱聚类的算法，需要提前给定聚类数

目，在此不做比较。由表 5 可见，AP 算法对 balance 和 ionosphere 数据集的聚类数目在多次实验中都是偏大的，无法得到准确的聚类数目；STI-AP 和 DA-AP 在聚类 balance 和 segment 数据集时无法得到与真实类数相等的聚类类数；DA-AP 对于数据集 glass 也无法得到与真实类数相等的类数；本文算法根据 P 值的调整可以得到与原数据集类簇数目相等的簇。本文通过结合两种聚类评价指标值综合评估验证了所提算法的有效性。

4.4 算法分析

在本文的对比算法中，STI-AP 没有利用密度信息，聚类时主要依据欧氏距离进行指数函数变换来刻画两个样本之间的距离，然后再进行流形搜索，可能造成样本点的错分，如图 4(a)和图 4(d)中靠近高密度区域但是属于低密度簇的点更容易被错分。DA-AP 算法只考虑了密度敏感距离，对于复杂结构的数据容易分出更多的类。IASCBDA 算法对密度不均匀的数据集进行聚类时有一定的优势，但是对于任意形状的数据集还有所欠缺，因此对更复杂的数据集的聚类效果相对较差。对于不同数据集，IASCBDA 首先分别确定聚类数目，如果是未知类数的数据集，输入不同的聚类数目则会对算法结果造成很大的影响，而且对相似度矩阵对称性的要求也使得该算法的应用范围不如本文提出的算法。

本文算法的主要任务是计算相似度矩阵和 AP 算法的迭代过程，AP 算法构建相似度矩阵的时间复杂度是 $O(n^2)$，更新支持度和吸引度的时间复杂度是 $O(n^2\log n)$；本算法中构建 KNN 网络并赋权的时间复杂度为 $O(n^2)$，利用 Dijkstra 算法搜索最短路径时使用邻接表的方法，其时间复杂度为 $O(n^2)$，因此本文算法的时间复杂度为 $O(n^2\log n)$。本算法在处理大规模数据集时往往会花费相对较长的时间，但通过综合利用密度尺度信息和数据内在流形结构，充分考虑数据集的全局和局部信息，使得聚类时能得到更好的结果。

结束语 针对近邻传播算法在任意形状和多重尺度的复杂结构样本中无法得到较好的聚类结果，本文提出了一种基于密度调整和流形距离的近邻传播算法。通过引入密度敏感参数，构造两种差异化的密度敏感距离，并根据 KNN 网络增大了处于不同流形中的样本之间的距离，更加真实地描述了样本之间的内在联系。本文算法与 4 种算法进行了对比，对 4 个充分考虑到尺度变换和形状变换的人工数据集和 5 个标准数据集进行聚类时都得到了较好的聚类效果，验证了算法的有效性。接下来的研究工作是在对大规模数据进行聚类时减少算法的运行时间以及将算法用于文本聚类问题中。

参 考 文 献

[1] JAIN A K. Data clustering: 50 years beyond K-means [J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[2] MACQUEEN J. Some Methods for Classification and Analysis of Mulitivariate Observations [C] // Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press, 1967: 281-297.

[3] PARK H S, JUN C H. A simple and fast algorithm for K-medoids clustering[J]. Expert System with Applications, 2009, 36(2): 3336-3341.

[4] LIU Y, HOU T, LIU F. Improving fuzzy c-means method for unbalanced dataset[J]. Electronics Letters, 2015, 23(51): 1880-1882.

[5] FREY B J, DUECK D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[6] FREY B J, DUECK D. Response to comment on "Clustering by passing messages between data points"[J]. Science, 2008, 319(5864): 726.

[7] XIAO Y, YU J. Semi-Supervised Clustering Based on Affinity Propagation Algorithm[J]. Journal of Software, 2008, 19(11): 2803-2813. (in Chinese)
肖宇，于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报，2008，19(11)：2803-2813.

[8] NI Z W, JING T T, NI L P. Affinity Propagation Hierarchical Optimization Algorithm[J]. Computer Science, 2015, 42(3): 195-200. (in Chinese)
倪志伟，荆婷婷，倪丽萍. 一种近邻传播的层次优化算法[J]. 计算机科学，2015，42(3)：195-200.

[9] LIU L, JIN S H, JIAO L C, et al. Manifold Affinity Propagation Clustering for PolSAR Image Classification[J]. Journal of Signal Processing, 2016, 32(2): 135-141. (in Chinese)
刘璐，靳少辉，焦李成，等. 采用流形近邻传播聚类的极化 SAR 图像分类[J]. 信号处理，2016，32(2)：135-141.

[10] ZHANG T, WU R B. Multiple extended target tracking using AP clustering[J]. Control and Decision, 2016, 31(4): 764-768. (in Chinese)
章涛，吴仁彪. 近邻传播观测聚类的多扩展目标跟踪算法[J]. 控制与决策，2016，31(4)：764-768.

[11] GUO K, GUO W Z, QIU Q R, et al. Community detection algorithm based on local affinity propagation and user profile[J]. Journal on Communications, 2015, 36(2): 72-83. (in Chinese)
郭昆，郭文忠，邱启荣，等. 基于局部近邻传播及用户特征的社区识别算法[J]. 通信学报，2015，36(2)：72-83.

[12] ZHANG Z, WANG B Q, LI X T, et al. Semi-supervised Traffic Identification Based on Affinity Propagation[J]. Acta Automatica Sinica, 2013, 39(7): 1100-1109. (in Chinese)
张震，汪斌强，李向涛，等. 基于近邻传播学习的半监督流量分类方法[J]. 自动化学报，2013，39(7)：1100-1109.

[13] DONG J, WANG S P, XIONG F L. Affinity Propagation Clustering Based on Variable-Similarity Measure[J]. Journal of Electronics and Information Technology, 2010, 32(3): 509-514. (in Chinese)
董俊，王锁萍，熊范纶. 可变相似性度量的近邻传播聚类[J]. 电子与信息学报，2010，32(3)：509-514.

[14] WANG K J, ZHANG J Y, LI D, et al. Adaptive Affinity Propagation Clustering[J]. Acta Automatica Sinica, 2007, 33(12): 1242-1246. (in Chinese)
王开军，张军英，李丹，等. 自适应仿射传播聚类[J]. 自动化学报，2007，33(12)：1242-1246.

[15] ZELNIK-MANOR L, PERONA P. Self-tuning spectral clustering[J]. Advances in Neural Information Processing Systems, 2005, 17: 1601-1608.

[16] WANG Y L, CHEN B, WANG X L, et al. Improved adaptive spectral clustering algorithm based on density adjustment[J]. Control and Decision, 2014, 29(9): 1683-1687. (in Chinese)

(下转第 215 页)

行了分析,建立了 FTF 信号仿真优化模型,构建了多目标优化算法结合微观交通仿真的 FTF 模型实施框架。利用采集的交通数据对 3 个交叉口组成的干线进行实例验证,实验结果表明,在过饱和状态下以最小化车均时延为单一目标的信号优化会导致次路车辆排队长度过大;对相序进行优化能进一步减小时延;本文方法能有效控制车辆排队长度,均衡车辆分布,提升路网效率,所得信号配时方案可用于离线控制或在线优化的基础方案。目前本方案仿真优化耗时较长,为进一步提升信号优化方法的实用性,下一步将研究仿真环境的校准方法并实现并行优化。

参 考 文 献

[1] YAN L Q, ZHAO Z H, LI P F, et al. Review of Traffic Signal Control Methods under Over-saturated Conditions[J]. Journal of Traffic and Transportation Engineering, 2013, 13(4): 116.

[2] LERTWORAWANICH P. A Simple Adaptive Signal Control Algorithm For Isolated Intersections Using Time-space Diagrams[C]// 2010 13th International IEEE Conference on Intelligent Transportation Systems (ITSC). Funchal, IEEE, 2010: 273-278.

[3] YE B L, WU W M, HANG Y S. A Green Wave Band based Method for Urban Arterial Signal Control[C]// 2014 IEEE 11th International Conference on Networking, Sensing and Control (ICNSC). Miami, FL, 2014: 126-131.

[4] LIU Y, CHANG G L. An arterial signal optimization model for intersections experiencing queue spillback and lane blockage[J]. Transportation Research Part C, 2011, 19(1): 130-144.

[5] YE B L, WU W M, MAO W J. A Two-Way Arterial Signal Coordination Method With Queueing Process Considered[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(6): 3440-3451.

[6] HAJBABAIE A, BENEKOHAL R B. A Program for Simultaneous Network Signal Timing Optimization and Traffic Assignment[J]. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(5): 2573-2586.

[7] STEVANOVIC A, MARTIN P T, STEVANOVIC J. VisSim-Based Genetic Algorithm Optimization of Signal Timings[J]. Journal of the Transportation Research Board, 2007, 2035 (2035): 59-68.

[8] LI Y, GUO X C, TAO S R, et al. NSGA II based traffic signal control optimization algorithm for over-saturated intersection group[J]. Journal of Southeast University (English Edition), 2013, 29(2): 211-216.

[9] GAO Y F, HU H, HAN H, et al. Multi-objective Optimization and Simulation for Urban Road Intersection Group Traffic Signal Control[J]. China Journal of Highway and Transport, 2012, 25(6): 129-135. (in Chinese)
高云峰,胡华,韩皓,等. 城市道路交叉口群信号协调控制多目标优化与仿真[J]. 中国公路学报,2012,25(6):129-135.

[10] CHEN S Y. Real-time Traffic Signal Control for Over-saturated Networks[D]. Lubbock, Texas: Texas Tech University, 2007.

[11] YANG X G, ZHAO J, MA W J, et al. Review on Calculation Method for Signalized Intersection Capacity[J]. China Journal of Highway and Transport, 2014, 27(5): 148-157. (in Chinese)
杨晓光,赵靖,马万经,等. 信号控制交叉口通行能力计算方法研究综述[J]. 中国公路学报,2014,27(5):148-157.

[12] KALYANMOY D, AGRAWAL S, PRATAB A, et al. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II [J]. IEEE Transactions on Evolutionary Computation, 2002, 6(2): 182-197.

[13] HUANG C, HU D M, YU X. An Improved NSGA-II Algorithm Based on Vector Space Model[J]. Journal of Chinese Computer Systems, 2015, 36(2): 391-396. (in Chinese)
黄超,胡德敏,余星. 一种基于向量空间模型的 NSGA-II 改进算法[J]. 小型微型计算机系统,2015,36(2):391-396.

[14] SUN Y J, SHEN G Z. Improved NSGA-II Multi-objective Genetic Algorithm Based on Hybridization-encouraged Mechanism [J]. Chinese Journal of Aeronautics, 2008, 21(6): 540-549.

[15] HUANG M, RAO M L, LI M. Research of Lane-level Basic Road Network Model for Simulation and Its Application[J]. Journal of System Simulation, 2014, 26(3): 657-681. (in Chinese)
黄敏,饶明雷,李敏. 面向仿真的车道级基础路网模型及其应用[J]. 系统仿真学报,2014,26(3):657-681.

[16] WANG D H, JIN S. Review and Outlook of Modeling of Car Following Behavior[J]. China Journal of Highway and Transport, 2012, 25(1): 116-127. (in Chinese)
王殿海,金盛. 车辆跟驰行为建模的回顾与展望[J]. 中国公路学报,2012,25(1):116-127.

[17] YANG L H, ZHANG X Q, GONG J K, et al. The Research of Car-Following Model Based on Real-Time Maximum Deceleration[J]. Mathematical Problems in Engineering, 2015, 2015: 1-9.

(上接第 192 页)

王雅琳,陈斌,王晓丽,等. 基于密度调整的改进自适应谱聚类算法[J]. 控制与决策,2014,29(9):1683-1687.

[17] SEUNG H S, LEE D D. The manifold ways of perception[J]. Science, 2000, 290(5500): 2263-2269.

[18] LIU S L, YAN D Q. A new global embedding algorithm[J]. Acta Automatica Sinica, 2011, 37(7): 828-835. (in Chinese)
刘胜蓝,闫德勤. 一种新的全局嵌入降维算法[J]. 自动化学报,2011,37(7):828-835.

[19] ZHANG S W, LEI Y K. Modified locally linear discriminant embedding for plant leaf recognition[J]. Neurocomputing, 2011, 74 (14/15): 2284-2290.

[20] LI Y M, NI L P, FANG Q H, et al. Research of Text Data Streams Clustering Algorithm Based on Affinity Propagation [J]. Computer Science, 2016, 43(5): 223-229. (in Chinese)
李一鸣,倪丽萍,方清华,等. 基于近邻传播的文本数据流聚类算法研究[J]. 计算机科学,2016,43(5):223-229.