ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 3,May 2019,pp. 548－557
DOI:10. 16337/j. 1004-9037. 2019. 03. 018

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于消息传递的谱聚类算法

王丽娟[1,2]　丁世飞[1]　贾洪杰[3]

(1. 中国矿业大学计算机科学与技术学院,徐州,221116;2. 徐州工业职业技术学院信息与电气工程学院,徐州,221400;3. 江苏大学计算机与通信工程学院,镇江,212013)

摘　要: 谱聚类将数据聚类问题转化成图划分问题,通过寻找最优的子图,对数据点进行聚类。谱聚类的关键是构造合适的相似矩阵,将数据集的内在结构真实地描述出来。针对传统的谱聚类算法采用高斯核函数来构造相似矩阵时对尺度参数的选择很敏感,而且在聚类阶段需要随机确定初始的聚类中心,聚类性能也不稳定等问题,本文提出了基于消息传递的谱聚类算法。该算法采用密度自适应的相似性度量方法,可以更好地描述数据点之间的关系,然后利用近邻传播(Affinity propagation,AP)聚类中"消息传递"机制获得高质量的聚类中心,提高了谱聚类算法的性能。实验表明,新算法可以有效地处理多尺度数据集的聚类问题,其聚类性能非常稳定,聚类质量也优于传统的谱聚类算法和k-means算法。

关键词: 谱聚类;相似矩阵;消息传递;聚类稳定性

中图分类号: TP301　　**文献标志码:** A

Spectral Clustering Algorithm Based on Message Passing

Wang Lijuan[1,2], Ding Shifei[1], Jia Hongjie[3]

(1. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China;2. School of Information and Electrical Engineering, Xuzhou College of Industrial Technology, Xuzhou, 221400, China;3. School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, 212013, China)

Abstract: Spectral clustering transforms data clustering problem into a graph partitioning problem and classifies data points by finding the optimal sub-graphs. The key to spectral clustering is constructing a suitable similarity matrix, which can truly describe the intrinsic structure of the dataset. However, traditional spectral clustering algorithms adopt Gaussian kernel function to construct the similarity matrix, which results in their sensitivity of selection for scale parameter. In addition, the initial cluster centers need randomly determing at the clustering stage and the clustering performance is not stable. The paper presents an algorithm based on message passing. The algorithm uses a density adaptive similarity measure, which can well describe the relations between data points, and it can obtain high-quality cluster centers through message passing mechanism in affinity propagation (AP) clustering. Moreover, the performance of clustering is optimized by the method. Experiments show that the proposed algorithm can effectively deal with the clustering problem of multi-scale datasets. Its clustering performance is very stable, and the clustering quality is better than traditional spectral clustering algorithm and k-means algorithm.

Key words: spectral clustering; similarity matrix; message passing; clustering stability

基金项目: 国家自然科学基金(61676522,61379101)资助项目;徐州市科技发展基金(KC17132)资助项目。
收稿日期: 2018-01-08;**修订日期:** 2019-04-09

引 言

聚类分析的一般过程是根据数据之间的相似度将其划分到不同的簇集中,使同一簇中的数据相似度较大,不同簇间数据的相似度较小。传统的聚类方法,如k-means算法等,缺乏处理复杂数据结构的能力,当样本空间非凸时,算法容易陷入局部最优[1]。

近年来,谱聚类由于其良好的表现,易于实现的特点,引起了学术界的关注[2]。谱聚类能在任意形状的样本空间上聚类,且收敛于全局最优,特别适用非凸数据集[3]。谱聚类算法将数据集中的每个点都作为图的顶点,任意两点之间的相似性值作为连接两顶点的边的权值,这样就构造了一个无向加权图。然后根据某种图划分方法将图分割为若干不连通的子图,子图中包含的点集就是聚类后生成的簇[4]。

传统的图划分方法有很多种,如最小割集法、比例割集法、规范割集法和最小最大割集法等[5]。通过最大化或最小化图割方法的目标函数得到最优割值,来获得聚类结果。但对于各种图分割方法来说,求目标函数的最优解往往是NP-hard的。根据瑞利熵理论,将原问题的离散最优化问题松弛到实数域,即可在多项式时间内解决[6]。对于图的划分,可以认为某个点的一部分属于子集A,另一部分属于子集B,而不是非此即彼。一般,在聚类过程中充分利用图的拉普拉斯矩阵的特征值和特征向量所包含顶点的分类信息,就可得到良好的聚类结果[7]。由于算法是基于矩阵谱分析理论来聚类的,因此称为谱聚类。目前,谱聚类已广泛应用在计算机视觉、数据分析、图像处理、视频监控以及自动控制等领域[8-10]。

谱聚类算法尽管在实践中取得了很好的效果,但是作为一种新型的聚类方法,仍处于发展阶段,还有很多问题值得进一步深入研究。例如,传统的谱聚类对初始值敏感,而且无法有效处理多尺度的聚类问题[11]。为了处理多重尺度的数据集,Zelnik-Manor等[12]提出了自适应的谱聚类。它不再指定统一的参数σ,而是根据每个点自身的邻域信息,为每个点x_i计算一个自适应的参数σ_i,其中σ_i为点x_i到其第p个近邻的欧式距离,该相似度度量称为自适应的高斯核函数。由于考虑了每个点邻域的数据分布,自调节的谱聚类能够有效分离出稀疏背景簇中包含的紧密簇。刘馨月等[13]通过局部密度获得数据中隐含的簇结构特征,再与自调节的高斯核函数结合,提出了一种基于共享近邻的自适应相似度的谱聚类算法。陶新民等[14]提出了一种在流形结构数据点间相似度计算方法,提高算法的聚类性能。为了解决传统谱聚类算法对尺度参数和聚类中心初始化敏感的问题,本文提出一种基于消息传递的谱聚类算法(Spectral clustering algorithm based on message passing,MPSC)。

1 谱聚类算法的初始化敏感分析

谱聚类算法是将数据聚类问题转化为图的最优分割问题。最小化或最大化图割方法的目标函数,均为NP离散最优化问题。幸运的是,谱方法可以为该最优化问题提供一种多项式时间内的宽松解[15]。这里的“宽松”指的是将离散最优化问题宽松到实数域,然后利用某种启发式方法将其重新转换为离散解。图分割的本质可以归结为矩阵的迹最小化或最大化问题,而完成该最小化或最大化的任务需要依靠谱聚类算法。

通常任何谱聚类算法都由3个部分组成:预处理、谱表示和聚类[16]。先构造相似图来描述数据集;然后建立相关的拉普拉斯矩阵,计算拉普拉斯矩阵的特征值和特征向量,基于一个或多个向量,把每个数据点映射到一个低维的代表点;最后,基于新的代表点,将数据点划分成两个或多个类。

在划分数据集或图时,有两个基本的方法:递归2-way划分法和k-way划分法[17]。递归2-way划分法是以一个层次的方式递归地调用2-way划分算法,当把图划分成两部分之后,再对子图应用相同的过程,直到聚类数目满足要求或不再符合递归条件为止;k-way划分法首先按照某种策略,将拉普拉斯矩阵的含有聚类信息的特征向量挑选出来,然后直接利用这些特征向量对数据集进行k-way划分。k-way划

分的一个典型的谱聚类算法是由文献[18]提出的NJW算法。下面是汪中等[19]就谱聚类算法中初始值敏感进行的分析。

命题1 设数据集$\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,以不同顺序输入得到的相似矩阵为$\boldsymbol{W}_1,\boldsymbol{W}_2$,对角矩阵为$\boldsymbol{D}_1,\boldsymbol{D}_2$,拉普拉斯矩阵为$\boldsymbol{L}_1,\boldsymbol{L}_2$,生成矩阵为$\boldsymbol{Y}_1,\boldsymbol{Y}_2$,则$\boldsymbol{W}_1$和$\boldsymbol{W}_2$相似,$\boldsymbol{D}_1$和$\boldsymbol{D}_2$相似,$\boldsymbol{L}_1$和$\boldsymbol{L}_2$相似,$\boldsymbol{Y}_1$和$\boldsymbol{Y}_2$相似。

证明:设数据集以$x_1,x_2,\cdots,x_n$的顺序输入得到矩阵$\boldsymbol{W}_1$,以$x_n,x_{n-1},\cdots,x_2,x_1$的顺序输入得到矩阵$\boldsymbol{W}_2$,矩阵$\boldsymbol{W}_1$经过若干次初等变换可得到矩阵$\boldsymbol{W}_2$,可证矩阵$\boldsymbol{W}_1$和$\boldsymbol{W}_2$相似。同理推出以任意顺序输入得到的矩阵$\boldsymbol{W}_2$均与$\boldsymbol{W}_1$相似。由于矩阵$\boldsymbol{D}_1$为对角矩阵,其主对角元素为相似矩阵$\boldsymbol{W}_1$的相应各行元素之和,故矩阵$\boldsymbol{D}_1$和$\boldsymbol{D}_2$相似。又因$\boldsymbol{L}_1=\boldsymbol{D}_1^{-1/2}\boldsymbol{W}_1\boldsymbol{D}_1^{-1/2}$,$\boldsymbol{L}_2=\boldsymbol{D}_2^{-1/2}\boldsymbol{W}_2\boldsymbol{D}_2^{-1/2}$,且$\boldsymbol{D}_1,\boldsymbol{D}_2$为对角矩阵,由矩阵$\boldsymbol{W}_1$和$\boldsymbol{W}_2$相似、$\boldsymbol{L}_1$和$\boldsymbol{L}_2$相似,其对应的特征向量$\boldsymbol{x}_1$经过若干次初等变换可以得到特征向量$\boldsymbol{x}_2$,可证生成矩阵$\boldsymbol{Y}_1$和$\boldsymbol{Y}_2$亦相似。故结论成立。

由命题1可知,以不同的顺序输入数据点,得到的相似矩阵$\boldsymbol{W}$和最终的生成矩阵$\boldsymbol{Y}$都是相似的,所以谱聚类算法对聚类中心的初始化敏感,本质上是因为在最后聚类阶段使用了k-means算法。为了解决这个问题,将AP聚类中的“消息传递”机制引入到谱聚类中,用来确定聚类中心,以改善谱聚类算法的性能。

2 AP聚类算法

近邻传播(Affinity propagation,AP)聚类是文献[20]提出的一种新的聚类算法。AP算法不需要事先指定初始聚类中心。实验表明,AP算法具有很高的效率,例如,对数千个手写的邮政编码的图片,AP算法只花费5 min就可以找出能准确解释各种笔迹类型的少量图片,而k-means算法要达到同样的精度需要耗费500万年。

AP算法以数据点之间的相似关系矩阵$\boldsymbol{S}$为基础,将$\boldsymbol{S}$的对角线上的数值$\boldsymbol{S}(k,k)$作为点k能否成为聚类中心的评判标准,这个值称作偏向参数,用$p(k)$表示。如若$p(k)$越小,则点k成为聚类中心的可能性就越小。然后通过在数据点之间传递消息,经过数次迭代来寻找最优的类代表点集合,使网络能量函数达到最小[21]。即有

$$E(c)=-\sum_{i=1}^{n}S(i,c_i) \tag{1}$$

式中:n表示数据的个数;c_i表示点i所在类的聚类中心点;$S(i,c_i)$表示点i与聚类中心点c_i之间的相似度。

在AP算法中数据点之间相互传播的信息有两种:一种信息被称为“吸引度”(Responsibility),简写为R;另一种信息被称为“归属度”(Availability),简写为A。

$R(i,k)$是数据点i向其备选代表点k发出的信息,定义为

$$R(i,k)=S(i,k)-\mathrm{Max}\{A(i,j)+S(i,j)\}\qquad j\neq k \tag{2}$$

它反映了在考虑数据点i的其他备选代表点的情况下,数据点k点适合作为i点的代表点的累积证据。

$A(i,k)$是备选代表点k向数据点i发出的信息,定义为

$$A(i,k)=\mathrm{Min}\{0,R(k,k)+\sum_{j=1}^{n}\mathrm{Max}\{0,R(j,k)\}\}\qquad j\neq i\text{且}j\neq k \tag{3}$$

它反映了在考虑了其他数据点是否选择点k作为自己的代表点的情况下,数据点i选择点k作为其代表点的累积证据。

$R(i,k)$与$A(i,k)$的和越大,则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[22]。这两种信息中包含了不同的竞争机制,在任何一个阶段根据这些信息可以判断出哪些数据点是代表点,以及每个数据点被划分到哪个数据类。

另外,为了平衡前后两次迭代的吸引度和归属度,AP算法引入了阻尼系数(Damping factor,即damp)来调整吸引度$R(i,k)$和归属度$A(i,k)$,从而得到本次迭代最终的吸引度和归属度

$$R_T = R_T \times (1-\text{damp}) + R_{T-1} \times \text{damp} \tag{4}$$

$$A_T = A_T \times (1-\text{damp}) + A_{T-1} \times \text{damp} \tag{5}$$

式中:R_{T-1},A_{T-1}表示第$T-1$次迭代的吸引度和归属度;R_T,A_T表示第T次迭代的吸引度和归属度;阻尼系数damp的取值范围是[0.5,1)。

AP算法聚类的过程如图1所示,从图中可以看到n个样本点$x_1,x_2,\cdots,x_n$,初始时都被当作潜在的聚类中心。AP算法在迭代过程中,根据相似关系矩阵$\boldsymbol{S}(i,i)$不断更新每个点的吸引度和归属度值,并且在网络节点之间不断地传递这两个信息量,直到产生m个高质量的聚类中心为止(满足停止条件)。

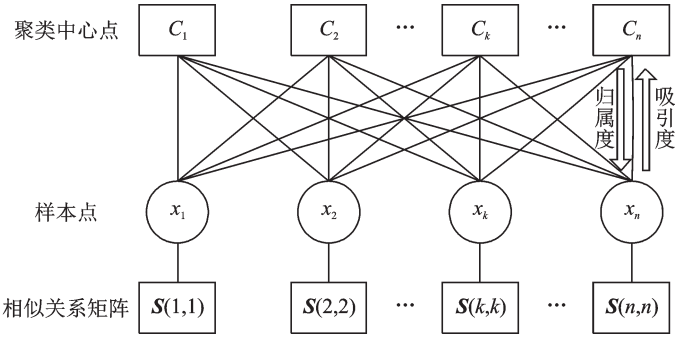


图1 AP算法聚类过程

Fig.1 Clustering process of AP algorithm

3 本文算法

谱聚类的关键是选择一个恰当的距离度量方法,将数据集的内在结构真实地描述出来。相同类中的数据点应该具有较高的相似性,并且要保持空间一致性。因此相似矩阵的构造非常重要,其优劣会在很大程度上影响谱聚类的性能[23]。度量数据点相似性的方法常用高斯核函数,但是在高斯核函数中,尺度参数σ通常是固定的,这样两个数据点之间的相似度只取决于它们的欧氏距离,而对其周围的点没有适应性[24]。当处理复杂数据集时,简单的基于欧氏距离的相似性无法准确地反映数据的分布情况,会显著降低谱聚类的性能,导致较差的聚类结果。

Yang等[25]提出了基于密度敏感的相似性度量方法,该方法可以处理多尺度的聚类问题,还相对对参数选择不敏感。实验证明该算法能有效地描述数据的实际聚类分布。

定义一个密度可调节的线段长度

$$L(x_i,x_j) = (\mathrm{e}^{\rho d(x_i,x_j)} - 1)^{1/\rho} \tag{6}$$

式中:$d(x_i,x_j)$为数据点x_i和x_j间的欧式距离,$\rho>1$称为伸缩因子。

数据点x_i和x_j间的密度敏感的距离定义为

$$D_{ij} = \text{Min} \sum_{k=1}^{l-1} L(p_k,p_{k+1}) \qquad p \in P_{ij} \tag{7}$$

式中:$p=\{\vec{p}_1,\vec{p}_2,\cdots,\vec{p}_l\}$表示从$p_i$到$p_l$的路径,其长度为$l=|p|$;$P_{ij}$表示连接数据点$x_i$和$x_j$的所有路径的集合。

密度敏感的相似性度量可表示为

$$w_{ij} = \frac{1}{D_{ij}+1} \tag{8}$$

与高斯核函数相比,该相似性度量不需要引入核函数,可以在距离测度上直接计算相似度。密度敏感的距离可以度量沿着流形上的最短路径,使得位于同一高密度区域内的两点可用许多较短的边相连,而位于不同高密度区域内的两点要用穿过低密度区域的较长边相连,最终达到这一目的:使不同高密度区域的数据点间距离放大,同一高密度区域内的数据点间距离缩小。因此,这一距离度量是数据依赖的,且可以反映数据的局部密度特征,即所谓的密度敏感。

在上述相似性度量方法的基础上,将“消息传递”机制引入到谱聚类中,提出了一种基于消息传递的谱聚类算法(Spectral clustering algorithm based on message passing,MPSC),其基本思想是:首先使用基于密度敏感的相似性度量方法计算相似性,构造相似性矩阵和拉普拉斯矩阵;然后选择拉氏矩阵的前k个最大特征向量,构造特征空间并将原始数据集中的点映射到$\mathbf{R}^k$空间中;最后在$\mathbf{R}^k$空间中,用AP聚类方法将数据点划分成k个类。MPSC算法的详细步骤如下。

输入:数据集$X=\{x_i|i=1,\cdots,n\}$,聚类数目k。

输出:k个划分好的类。

Step 1 根据式(8),计算数据点之间的相似性值,建立基于密度的相似性矩阵$\boldsymbol{W}\in\mathbf{R}^{n\times n}$。

Step 2 建立图的度矩阵$\boldsymbol{D}\in\mathbf{R}^{n\times n}$,$\boldsymbol{D}$是一个对角矩阵:对角线上的元素为$d_i$,$d_i=\sum_{j=1}^{n}w_{ij}$称为顶点$i$的度,而对角线外的元素值为0。

Step 3 根据相似性矩阵$\boldsymbol{W}$和度矩阵$\boldsymbol{D}$,构造拉普拉斯矩阵$\boldsymbol{L}$:$\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$。

Step 4 计算矩阵$\boldsymbol{L}$的前k个最大特征值所对应的特征向量$\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k$(重复特征值取其相互正交的特征向量),然后将这些特征向量纵向排列,形成矩阵$\boldsymbol{U}$,$\boldsymbol{U}=[u_1\vdots\cdots\vdots u_k]\in\mathbf{R}^{n\times k}$。

Step 5 规范化矩阵$\boldsymbol{U}$的每一行,将行向量转变成单位向量,得到矩阵$\boldsymbol{Y}$:$y_{ij}=u_{ij}\Big/\left[\sum_{j=1}^{n}u_{ij}^2\right]^{1/2}$。

Step 6 将矩阵$\boldsymbol{Y}$的每一行看作空间$\mathbf{R}^k$中的一个点,利用AP聚类方法将这些点划分成k类。

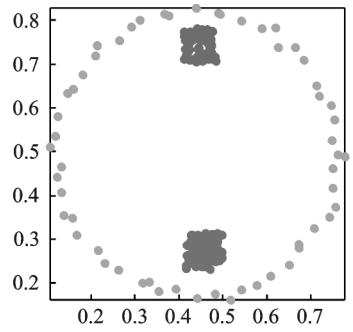
Step 7 如果矩阵$\boldsymbol{Y}$的第i行被分配到第j类,就将原始的数据点x_i划分到第j类。

MPSC算法继承了AP聚类的优点,它在初始时将所有的数据点都看作候选聚类中心,在不断的迭代过程中,选择某个数据点作为中心或每个数据点通过“消息传递”来竞争成为聚类中心,最终获得若干个优化的聚类中心。为了克服传统划分算法中随机选择参数对整个数据聚类结果的影响,引入消息传递机制优化传统谱聚类算法中初始化敏感的问题,得到更稳定的聚类结果。

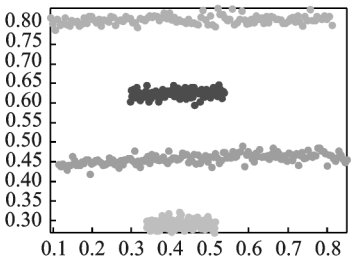
4 实验分析

4.1 仿真数据集聚类

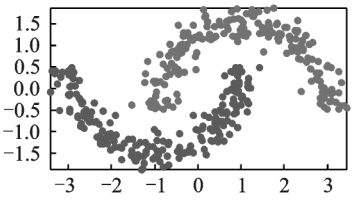
文献[26]给出了一些有“挑战性”的人工数据集,例如:Blobs and circle,Four lines,Two moons,Two circles,Two spirals和Three circles。用MPSC算法分别对这些数据集进行聚类,得到的聚类结果如图2所示(伸缩因子$\rho=2$,偏好参数p取数据集全体样本相似度的中位数)。从图2中可以看出,MPSC算法可以有效识别不同尺度的流形数据结构,得到令人满意的聚类结果。这是因为该算法采用密度敏感的相似性度量方法,通过拉普拉斯变换,将原始数据点映射到谱空间,使得相同类内的数据点更加紧凑,而不同类间的数据点更加分离。在2002年,Ng, Jordan和Weiss提出NJW算法(算法以3位作者名字的首字母缩写命名),图3给出了NJW算法(分别取$\sigma=0.1$和$\sigma=0.2$)在Blobs and circle,Two moons和Two spirals人工数据集上的聚类结果,可见NJW算法对尺度参数σ的取值比较敏感,当数据结构复杂时,无法得到准确的聚类结果。相比之下,MPSC算法在聚类过程中可以保持数据的全局一致性,从而



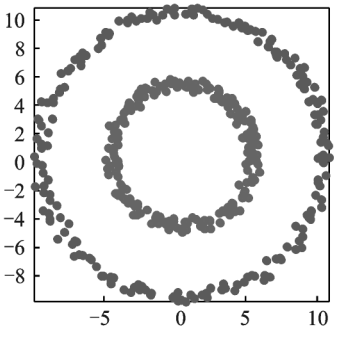
(a) MPSC算法在Blobs and circle的聚类结果
(a) Clustering results of MPSC algorithm on Blobs and circle



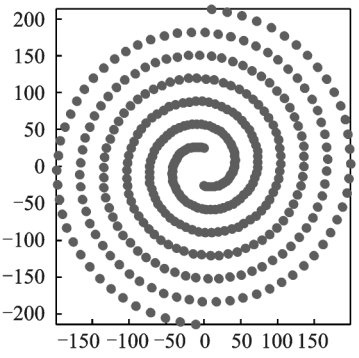
(b) MPSC算法在Four lines的聚类结果
(b) Clustering results of MPSC algorithm on Four lines



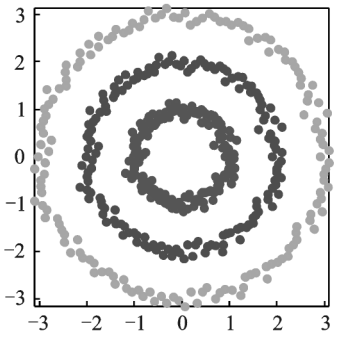
(c) MPSC算法在Two moons的聚类结果
(c) Clustering results of MPSC algorithm on Two moons



(d) MPSC算法在Two circles的聚类结果
(d) Clustering results of MPSC algorithm on Two circles



(e) MPSC算法在Two spirals的聚类结果
(e) Clustering results of MPSC algorithm on Two spirals



(f) MPSC算法在Three circles的聚类结果
(f) Clustering results of MPSC algorithm on Three circles

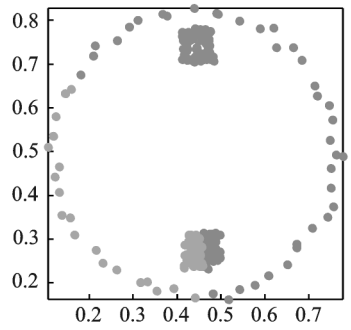
图2 MPSC算法在人工数据集上的聚类结果

Fig.2 Clustering results of MPSC algorithm on artificial data sets

克服了高斯核函数无法处理多尺度数据集的缺点。图2,3中横纵坐标分别为人工数据集的第1列和第2列。

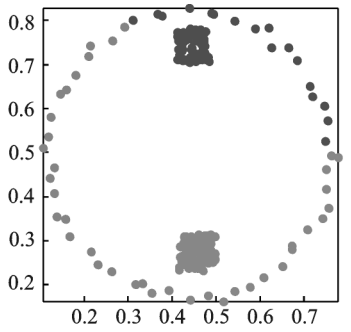
4.2 真实数据集聚类

为了进一步验证MPSC算法的有效性,从UCI机器学习数据库中选取了3个真实的数据集,它们的数据特征如表1所示。如果将聚类得到的结果和真实的划分情况进行比较,对于数据集中的每对数据点,存在着下面4种可能:(1) SS:属于同一类的两个数据点在聚类时也被分到相同的类中;(2) SD:属于同一类的两个数据点在聚类时却被分到不同的类中;(3) DS:不属于同一类的两个数据点在聚类时却被分



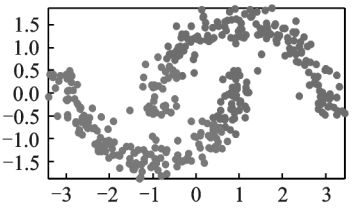
(a) NJW算法在Blobs and circle上的聚类结果($\sigma=0.1$)

(a) Clustering results of NJW algorithm on Blobs and circle ($\sigma=0.1$)



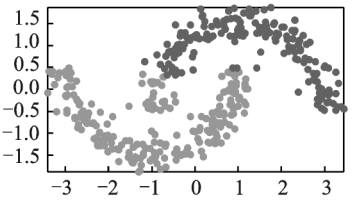
(b) NJW算法在Blobs and circle上的聚类结果($\sigma=0.2$)

(b) Clustering results of NJW algorithm on Blobs and circle ($\sigma=0.2$)



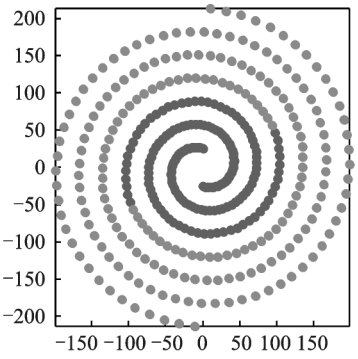
(c) NJW算法在Two moons上的聚类结果($\sigma=0.1$)

(c) Clustering results of NJW algorithm on Two moons ($\sigma=0.1$)



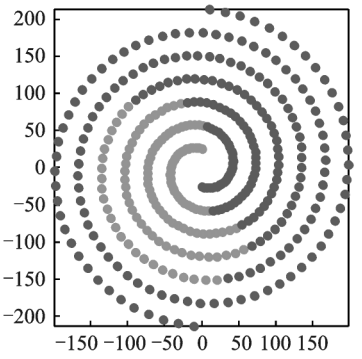
(d) NJW算法在Two moons上的聚类结果($\sigma=0.2$)

(d) Clustering results of NJW algorithm on Two moons ($\sigma=0.2$)



(e) NJW算法在Two spirals上的聚类结果($\sigma=0.1$)

(e) Clustering results of NJW algorithm on Two spirals ($\sigma=0.1$)



(f) NJW算法在Two spirals上的聚类结果($\sigma=0.2$)

(f) Clustering results of NJW algorithm on Two spirals ($\sigma=0.2$)

图3 NJW算法在人工数据集上的聚类结果

Fig.3 Clustering results of NJW algorithm on artificial data sets

到相同的类中;(4) DD:不属于同一类的两个数据点在聚类时也被分到不同的类中。

ARI(Adjusted rand index)指标就是根据数据点对的这4种关系来评价聚类结果的,它可以定量描述聚类的质量,客观反映聚类算法的优劣,是一种常用的聚类评价准则[27]。

设满足SS,SD,DS,DD关系的数据点对的数目分别是a,b,c,d,则ARI的计算公式为

表1 真实数据集的数据特征

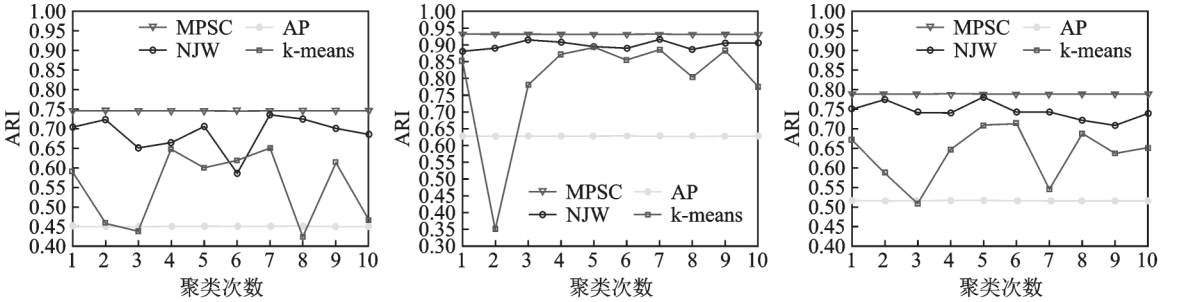
Tab. 1 Data characteristics of real data sets

Dataset	Sample number	Attribute number	Class number
Iris	150	4	3
Wine	178	13	3
Zoo	101	16	7

$$\mathrm{ARI}=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{9}$$

式中ARI∈[0,1],如ARI的值越小,也就表明聚类的结果越差。

在实验中,以ARI指标作为衡量标准,分别在Iris,Wine,Zoo数据集上,对比了MPSC算法、NJW算法、AP算法和k-means算法的聚类性能。实验结果如图4所示。



(a) 4种算法在Iris数据集上的实验结果 (b) 4种算法在Wine数据集上的实验结果 (c) 4种算法在Zoo数据集上的实验结果

(a) Experimental results of four algorithms on Iris (b) Experimental results of four algorithms on Wine (c) Experimental results of four algorithms on Zoo

图4 4种算法在真实数据集上的实验结果

Fig.4 Experimental results of four algorithms on real data sets

在图4中,NJW算法和k-means算法的曲线都有明显的波动,说明它们对聚类中心的初始化比较敏感。这两种算法都需要随机地确定初始聚类中心,当聚类中心的选择不合适时,就会产生较差的聚类结果。AP算法和MPSC算法的曲线非常平稳,没有出现任何波动,说明AP算法和MPSC算法的性能很稳定,但AP算法表现不理想,本文提出的MPSC算法将AP算法中"消息传递"机制引入,以确定聚类中心,有效地解决了传统谱聚类算法对聚类中心初始化敏感的问题。而且从图4中也可以看出,MPSC算法的ARI指标也要明显优于NJW算法、AP算法和k-means算法,因此MPSC算法可以得到比较理想的聚类结果。表2给出了MPSC算法、NJW算法、AP算法和k-means算法10次聚类的平均准确率和ARI指标。

表2 4种算法的平均准确率和ARI指标

Tab. 2 Average accuracy and ARI index of four algorithms

Algorithm	Accuracy			ARI		
	Iris	Wine	Zoo	Iris	Wine	Zoo
MPSC	0.906 7	0.977 5	0.842 6	0.758 3	0.931 0	0.788 5
NJW	0.853 4	0.966 9	0.816 0	0.685 0	0.898 6	0.744 1
AP	0.681 7	0.620 3	0.754 9	0.451 3	0.627 4	0.515 8
k-means	0.727 3	0.896 7	0.715 9	0.551 6	0.794 3	0.635 9

5 结束语

本文分析了传统的谱聚类算法初始化敏感的原因,提出了一种基于消息传递的谱聚类算法。该算法引入了"消息传递"机制,通过在数据点之间不断传递"吸引度"和"归属度"信息,可以获得高质量的聚类中心。而且该算法使用了密度敏感的相似性度量方法,这样在度量数据点之间的相似性时,可以更好地描述数据的分布情况,保持数据的全局一致性。为了验证MPSC算法的有效性,在人工数据集

和真实数据集上分别进行了仿真实验。实验结果表明,MPSC算法不仅可以有效识别数据全局的和局部的分布特征,而且对聚类中心的初始化不再敏感,其聚类准确率和稳定性都明显好于传统的谱聚类算法和k-means算法。

不过MPSC算法的计算复杂度较高,在处理大数据集时,会花费较长的时间,如何降低算法的时间复杂度,提高聚类的效率,还有待进一步研究。现实的聚类问题中,除了含有大量无标记的数据,有时也会含有一些标记数据,因此将半监督学习与MPSC算法相结合,利用少量的标记信息来指导聚类的过程,也是一个有价值的研究方向。另外,MPSC算法也有着广阔的应用前景,可以用来处理图像分割、语音分离和文字识别等实际问题。

References:

[1] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.
Sun Jigui, Liu Jie, Zhao Lianyu. Clustering algorithms research[J]. Journal of Software, 2008, 19(1): 48-61.

[2] Nascimento M C V, Andre C, De Carvalho A. Spectral methods for graph clustering—A survey[J]. European Journal of Operational Research, 2011, 211(2): 221-231.

[3] Ding S F, Jia H J, Zhang L W, et al. Research of semi-supervised spectral clustering algorithm based on pairwise constraints [J]. Neural Computing & Applications, 2014,24(1): 211-219.

[4] Chen W F, Feng G C. Spectral clustering: A semi-supervised approach[J]. Neurocomputing, 2012, 77(1): 229-242.

[5] Fan N, Pardalos P M. Multi-way clustering and biclustering by the ratio cut and normalized cut in graphs[J]. Journal of Combinatorial Optimization, 2012, 23(2): 224-251.

[6] Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416.

[7] 李建元, 周脚根, 关佶红, 等 谱图聚类算法研究进展[J]. 智能系统学报, 2011, 6(5): 405-414.
Li Jianyuan, Zhou Jiaogen, Guan Jihong, et al. A survey of clustering algorithms based on spectra of graphs[J]. CAAI Transactions on Intelligent Systems, 2011, 6(5): 405-414.

[8] Tung F, Wong A, Clausi D A. Enabling scalable spectral clustering for image segmentation[J]. Pattern Recognition, 2010, 43 (12): 4069-4076.

[9] Zhang D P, Chen F Y, Peng H L. Detecting group-level crowd using spectral clustering analysis on particle trajectories[J]. Information Technology Journal, 2013, 12(1): 174-179.

[10] Ding L, Gonzalez-Longatt F M, Wall P, et al. Two-step spectral clustering controlled islanding algorithm[J]. IEEE Transactions on Power Systems, 2013, 28(1): 75-84.

[11] 刘馨月, 李静伟, 于红, 等. 基于共享近邻的自适应谱聚类[J]. 小型微型计算机系统, 2011, 32(9): 1876-1880.
Liu Xinyue, Li Jingwei, Yu Hong, et al. Adaptive spectral clustering based on shared nearest neighbors[J]. Journal of Chinese Computer System, 2011, 32(9): 1876-1880.

[12] Zelnik-Manor L, Perona P. Self-tuning spectral clustering[C]//Proceeding of NIPS. Vancouver, Canada: Neural Information Processing Systems Foundation, 2005: 1601-1608.

[13] Liu X Y, Li J W, Yu H, et al. Adaptive spectral clustering based on shared nearest neighbors[J]. Journal of Chinese Computer Systems, 2011, 32(9): 1876-1880.

[14] 陶新民, 宋少宇, 曹盼东, 等. 一种基于流形距离核的谱聚类算法[J]. 信息与控制, 2012, 41(3): 307-313.
Tao Xinmin, Song Shaoyu, Cao Pandong, et al. A spectral clustering algorithm based on manifold distance kernel[J]. Information and Control, 2012, 41(3): 307-313.

[15] Ding S F, Qi B J, Jia H J, et al. Research of semi-supervised spectral clustering based on constraints expansion[J]. Neural Computing and Applications, 2013, 22(1): 405-410.

[16] Hamad D, Biela P. Introduction to spectral clustering[C]//Proceedings of the International Conference on Information and Communication Technologies from Theory to Applications-ICTTA'08. Damascus, Syria: IEEE Computer Society, 2008: 634-639.

[17] 蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述[J]. 计算机科学, 2008, 35(7): 14-18.

Cai Xiaoyan, Dai Guanzhong, Yang Libin. Survey on spectral clustering algorithms[J]. Computer Science, 2008,35(7): 14-18.

[18] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[J]. Advances in Neural Information Processing Systems, 2002, 14: 849-856.

[19] 汪中, 刘贵全, 陈恩红. 基于模糊K-harmonic means的谱聚类算法[J]. 智能系统学报, 2009, 4(2): 95-99.
Wang Zhong, Liu Guiquan, Chen Enhong. A spectral clustering algorithm based on fuzzy K-harmonicmeans[J]. CAAI Transactions on Intelligent Systems, 2009,4(2): 95-99.

[20] Frey B J, Dueek D. Clustering by passing messages between data points[J]. Sincence, 2007, 315(5814): 972-976.

[21] 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.
Dong Jun, Wang Suoping, Xiong Fanlun. Affinity propagation clustering based on variable-similarity measure[J]. Journal of Electronics &. Information Technology, 2010,32(3): 509-514.

[22] Zhang X L, Wang W, Nørvag K, et al. K-AP: Generating specified k clusters by efficient affinity propagation[C]//Proceedings 2010 10th IEEE International Conference on Data Mining (ICDM 2010). Sydney, Australia: IEEE, 2010: 1187-1192.

[23] Zhang X C, Li J W, Yu H. Local density adaptive similarity measurement for spectral clustering[J]. Pattern Recognition Letters, 2011, 32(2): 352-358.

[24] Wang Y, Jiang Y, Wu Y, et al. Spectral clustering on multiple manifolds[J]. IEEE Transactions on Neural Networks, 2011, 22(7): 1149-1161.

[25] Yang P, Zhu Q S, Huang B. Spectral clustering with density sensitive similarity function[J]. Knowledge-Based Systems, 2011, 24(5): 621-628.

[26] 王玲, 薄列峰, 焦李成. 密度敏感的谱聚类[J]. 电子学报, 2007, 35(8): 1577-1581.
Wang Ling, Bo Liefeng, Jiao Licheng. Density-sensitive spectral clustering[J]. Acta Electronica Sinica, 2007,35(8): 1577-1581.

[27] Xie J Y, Jiang S, Xie W X, et al. An efficient global k-means clustering algorithm[J]. Journal of Computers, 2011, 6(2): 271-279.

作者简介:

王丽娟(1981-),女,博士研究生,研究方向:机器学习、数据挖掘、聚类分析等,E-mail:donglittle@126.com。

丁世飞(1963-),男,教授、博士生导师,研究方向:人工智能、机器学习、模式识别、数据挖掘等,E-mail:dingsf@cumt.edu.cn。

贾洪杰(1988-),男,博士,研究领域:机器学习、数据挖掘,谱聚类等。

(编辑:刘彦东)