ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.18, No.10, October 2007, pp.2412−2422
DOI: 10.1360/jos182412

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

密度敏感的半监督谱聚类*

王 玲+, 薄列峰, 焦李成

(西安电子科技大学 智能信息处理研究所,陕西 西安 710071)

Density-Sensitive Semi-Supervised Spectral Clustering

WANG Ling+, BO Lie-Feng, JIAO Li-Cheng

(Institute of Intelligent Information Processing, Xidian University, Xi'an 710071, China)

+ Corresponding author: Phn: +86-29-88209788, E-mail: wangling@mail.xidian.edu.cn, http://see.xidian.edu.cn/graduate/wangling

Wang L, Bo LF, Jiao LC. Density-Sensitive semi-supervised spectral clustering. *Journal of Software*, 2007, 18(10):2412−2422. http://www.jos.org.cn/1000-9825/18/2412.htm

Abstract: Clustering has been traditionally viewed as an unsupervised method for data analysis. In real world application, however, some background prior knowledge can be easily obtained, such as pairwise constraints. It has been demonstrated that constraints can improve clustering performance. In this paper, the drawback of only incorporating pairwise constraints in clustering is firstly analyzed, and then an inherent prior knowledge in data sets, namely space consistency prior knowledge is exploited. The method of utilizing space consistency prior knowledge is also given. Incorporating the two types of prior knowledge into original spectral clustering, a density-sensitive semi-supervised spectral clustering algorithm (DS-SSC) is proposed. Experimental results on UCI (University of California Irvine) benchmark data, USPS (United States Postal Service) handwritten digits and text data from TREC (Text REtrieval Conference) show that the two types of prior knowledge can supplement each other in clustering process, leading to substantial performance enhancement of DS-SSC over other semi-supervised clustering methods which only incorporate pairwise constraints.

Key words: spectral clustering; semi-supervised clustering; pairwise constraints; prior knowledge

摘 要: 聚类通常被认为是一种无监督的数据分析方法,然而在实际问题中可以很容易地获得有限的样本先验信息,如样本的成对限制信息.大量研究表明,在聚类搜索过程中充分利用先验信息会显著提高聚类算法的性能.首先分析了在聚类过程中仅利用成对限制信息存在的不足,尝试探索数据集本身固有的先验信息——空间一致性先验信息,并提出利用这类先验信息的具体方法.接着,将两类先验信息同时引入经典的谱聚类算法中,提出一种密度敏感的半监督谱聚类算法(density-sensitive semi-supervised spectral clustering algorithm,简称 DS-SSC).两类先验信息在指导聚类搜索的过程中能够起到相辅相成的作用,这使得 DS-SSC 算法相对于仅利用成对限制信息的聚类算法在聚类性能上有了显著的提高.在 UCI 基准数据集、USPS 手写体数字集以及 TREC 的文本数据集上的实验结果验证了这一点.

关键词: 谱聚类;半监督聚类;成对限制;先验信息

* Supported by the National Natural Science Foundation of China under Grant Nos.60372050, 60372045 (国家自然科学基金); the National Basic Research Program of China under Grant No.2001CB309403 (国家重点基础研究发展计划(973))

Received 2006-04-12; Accepted 2006-06-30

中图法分类号: TP18 文献标识码: A

作为一种有效的数据分析方法,聚类算法已被广泛应用于计算机视觉、信息检索、数据挖掘等领域.聚类算法在执行过程中不能获得任何关于预先定义的数据项的类属信息,因而通常被看作是一种无监督学习方法.由于没有利用任何关于类属的信息,当所定义的聚类目标函数不适合数据本身时,数据聚类将变成一个病态问题.另外,聚类定义的任意性有可能产生对于实际问题没有任何意义的聚类划分.尽管对于现实世界问题要获得所有数据的类属信息需要付出相当大的代价,然而,忽视那些很容易获得的少量样本类属信息将是很大的浪费.人们已经开始尝试在一些实际问题中利用可获得的先验信息,例如在图像分割中可以获得一些区域的部分划分信息,用其来辅助整个图像的聚类[1];在视频检索中,不同的用户可以对数据库中小的子集中的图像提供注释,利用这些划分信息来对整个数据库进行聚类以改善聚类效果[2].

利用样本先验信息来改善无监督聚类算法的性能,已成为最近机器学习领域的一个研究热点,所提出的算法被统称为半监督聚类.根据使用先验信息方法的不同,已有的半监督聚类算法被分成两大类:一类是基于限制的方法,该类方法修改聚类算法本身,利用成对限制先验信息来指导聚类算法向一个较好的数据划分进行[3];另一类是基于测度的方法,这类方法首先训练相似性测度用以满足类属或限制信息,然后使用基于测度的聚类算法进行聚类[4].基于测度的方法在测度学习阶段仅利用了有限的限制信息,而将大量的无类属数据排除在外,本文试图挖掘存在于大量的无类属数据中可利用的先验信息.

谱聚类作为一种新颖的聚类方法近年来受到了模式识别领域的广泛关注.该聚类方法具有识别非凸分布聚类的能力,适用于许多实际应用领域.由于谱聚类是一种配对聚类算法,这使得在聚类过程中利用成对限制先验信息变得非常容易.本文尝试利用无类属数据内部存在的结构先验信息,并同时结合成对限制信息来改善经典谱聚类算法的聚类性能.本文首先介绍谱聚类算法,然后详细分析了从数据中可获得的聚类先验信息,认为仅使用样本层面上的成对限制先验信息[5]是不够的,应该充分利用数据本身固有的分布特性这一先验信息.受半监督学习中聚类假设[6]的启发,我们认为数据集本身可以向聚类算法提供一种所谓的空间一致性先验信息.为了利用这种先验信息,本文设计出一种基于密度敏感的距离测度的方法.该方法不仅能够同时利用用户提供的成对限制信息,而且挖掘了无类属数据中的空间一致性信息.将两类先验信息引入经典的谱聚类算法中,提出一种密度敏感的半监督谱聚类算法(density-sensitive semi-supervised spectral clustering algorithm,简称DS-SSC).在DS-SSC算法中,成对限制先验信息起到直接修改距离测度的作用,使得相似性关系随限制发生改变,而空间一致性先验信息根据数据聚类的空间分布特性自动调节数据间的距离,起到间接改变相似性关系的作用,从而避免了由用户所提供的信息含量少的限制所造成的聚类偏斜.实验表明,DS-SSC相对于仅结合成对限制信息的谱聚类算法以及对成对限制信息进行空间传播的CCL(complete connected link)算法[4]在聚类性能上有了显著的提高,并且算法在提供少量限制的情况下仍然可以取得很好的聚类效果.

1 谱聚类算法

谱聚类算法的思想来源于谱图划分理论.它将聚类问题看成是一个无向图的多路划分问题.定义一个图划分判据,如Shi和Malik提出的一个有效的图划分判据——规范切判据[7],最优化这一判据,使得同一类内的点具有较高的相似性,而不同类之间的点具有较低的相似性.由于图划分问题的组合本质,求图划分判据的最优解是一个NP难问题.一个很好的求解方法是考虑问题的连续放松形式,这样便可将原问题转换成求图的Laplacian矩阵的谱分解问题,因此,将这类方法统称为谱聚类,可以认为谱方法是对图划分判据的逼近[8].谱聚类也可以利用类似于PCA(principle component analysis)子空间方法中的嵌入思想来解释.该方法同时使用矩阵的多个特征向量,利用这些特征向量构造一个简化了的数据空间,在该空间中数据的分布结构更加明显.代表性算法有Ng等人提出的SC(spectral clustering)算法[9].Meila和Xu指出,SC算法和一种基于图划分判据的方法——多路规范切算法(multiway normalized cut,简称MNCut)[10]的不同之处仅在于所使用的谱映射不同,并且当相似性矩阵是理想矩阵时,它们是等价的.

谱聚类算法是一种配对聚类方法,算法仅与数据点的数目有关,而与维数无关,因而可以避免由特征向量的过高维数所造成的奇异性问题.谱聚类算法又是一种判别方法,不用对数据的全局结构作假设.谱方法成功的原因在于:通过特征分解,可以获得聚类判据在放松了的连续域中的全局最优解.谱聚类相对于其他聚类方法具有明显的优势,它具有识别非凸分布聚类的能力,非常适合于许多实际问题,而且执行起来比较容易.该方法已成功应用于语音识别、图像和视频分割等领域.

由于谱聚类算法是一种配对聚类算法,这使得在算法中利用用户提供的成对限制信息变得非常容易,已有研究者提出利用成对限制信息的半监督谱聚类算法[11].下面我们将考虑如何在谱聚类算法中充分利用数据集合中可获得的数据聚类先验信息,期望进一步提高算法的聚类性能.

2 可获得的数据聚类先验信息

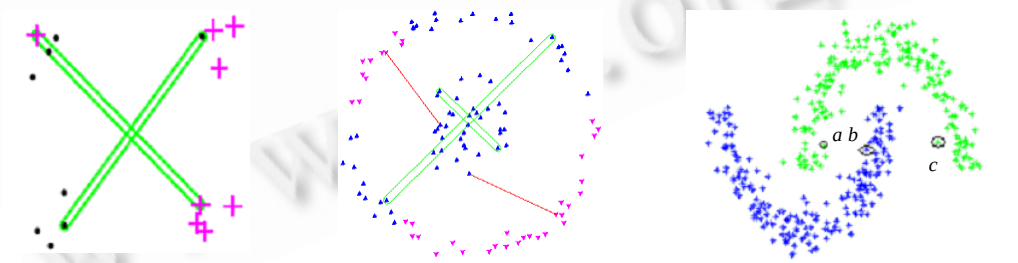
2.1 成对限制先验信息

对于许多聚类应用领域,如交谈中的说话人识别[12]、GPS数据中的道路检测问题[4],考虑以成对点限制形式出现的监督信息而不是样本类属信息会比较实际一些.这是由于,对于用户来说,要确定样本类属会比较困难,而获得一些关于样本点是否可以或不能位于同一类的限制信息将会比较容易.另外,基于限制的先验信息比类属信息更为一般化,我们可以从类属信息获得等价的成对限制信息,反之则不然.Wagstaff等人最早在文献[5]中引入两种类型的成对点限制,即*must-link*和*cannot-link*来辅助聚类搜索.*must-link*限制规定两个样本必须在同一聚类中;*cannot-link*限制规定两个样本不能在同一聚类中.

成对限制先验信息给出的仅仅是样本层面上的限制,如果单纯地只利用这类限制信息容易导致聚类算法产生如图 1(a)所示的单个奇异点的聚类偏斜[4].Klein等人通过研究*must-link*限制在样本上具备的一组二值传递关系:

$$\begin{cases}(x_i,x_j)\in must\text{-}link \;\&\; \&(x_j,x_k)\in must\text{-}link \Rightarrow (x_i,x_k)\in must\text{-}link\\ (x_i,x_j)\in must\text{-}link \;\&\; \&(x_j,x_k)\in cannot\text{-}link \Rightarrow (x_i,x_k)\in cannot\text{-}link\end{cases},$$

将样本层面上的限制进行空间传播.该方法首先使用求最短路的方法施加*must-link*限制,然后在改变了的测度空间中仅施加样本层面上的*cannot-link*限制,从而实现上述传递闭包关系,最后利用完全链接层次聚类算法将*cannot-link*限制进行空间传播,最终达到利用了所有成对限制信息所造成的空间影响.这里,我们将Klein的方法称为CCL算法[4].



(a) An instance-level inductive bias results in single outliers
(a) 样本层面上限制造成的单个奇异点聚类偏斜

(b) With some constraints, CCL can not recognize the two circles
(b) 在某些限制条件下 CCL 无法识别两个圆问题

(c) Looking for a distance metric according to which point *a* is much closer to point *c*
(c) 寻找距离测度使得点 *a* 更接近于点 *c*

Fig.1
图 1

由于这里我们所考虑的限制信息是由用户任意提供的,这样就有可能产生一些信息含量相对少的限制,这些限制信息对于聚类算法不能起到积极的指导作用,反而有可能误导聚类.由于Klein所采用的传播限制的方法利用完全链接层次聚类算法对 *cannot-link* 限制进行空间传播,该算法有可能造成 *cannot-link* 限制的错误传播,

从而误导聚类搜索,造成聚类性能的下降.另外,在有限的限制条件下,该算法对限制的传播作用也非常有限,因而此时该方法无法识别如图 1(b)所示的两个圆问题(图中两点连线代表 *cannot-link* 限制,两点画圈代表 *must-link* 限制).如果仅利用样本层面的限制信息来指导聚类,则或者可以得到有限的聚类性能的提高,或者仅起到误导聚类的反作用.我们将在后面的仿真实验中进一步验证这一点.

2.2 空间一致性先验信息

根据上一节的分析,成对限制先验信息所能提供给聚类算法的指导作用是有限的,必须利用一些方法来扩充限制的数目,而传播限制的方法有时会对聚类起误导的反作用.因此,充分提供的有用的先验信息对于半监督聚类来说至关重要.为了能够在聚类过程中克服成对限制信息所造成的不良影响,进一步提高半监督聚类算法的性能,本节我们将探索利用一种存在于数据集合本身的可获得的聚类先验信息.

除了用户可提供的有限的成对限制信息以外,我们考虑能否从数据集合本身获得一些对于聚类有用的先验信息.根据半监督学习中数据的聚类一致性先验假设[6],数据聚类具有下面的两个一致性特征:

- 局部一致性——是指在空间位置上相邻的数据点具有较高的相似性;
- 全局一致性——是指位于同一流形上的数据点具有较高的相似性.

我们用如图1(c)所示的人工数据来说明聚类的两个一致性特征.所希望的是位于同一聚类中的点具有较高的相似性,如点 *a* 和点 *c* 具有较高的相似性,即在某一测度下点 *a* 和 *c* 更接近,然而在欧氏距离测度下,点 *a* 更接近于点 *b*,这样便没有反映出聚类的全局一致性.因此,对于复杂的实际应用问题,简单的基于欧氏距离测度的相似性度量不能完全反映数据聚类复杂的空间分布特性.

下面的问题是,我们如何利用数据的空间一致性先验假设来辅助聚类.从数据的空间分布情况可以观察到,同一聚类内的数据趋向于分布在一个密度比较高的区域,而在不同聚类之间存在一个数据分布相对稀疏的低密度区域.我们考虑引入一个空间一致性距离测度,该距离测度既能够描述数据聚类的局部一致性特征,又能够描述数据聚类的全局一致性特征,从而体现了聚类的空间分布特性.然而需要注意的是,满足聚类全局一致性的距离并不一定满足欧氏测度下的三角不等式.也就是说,满足聚类全局一致性的距离能够使得两点间直接相连的路径长度不一定最短.为了达到这一目的,我们首先需要定义一个所谓的密度可调节的线段长度.

定义 1. 密度可调节的线段长度: $L(x_i,x_j)=\rho^{dist(x_i,x_j)}-1$,其中,$dist(x_i,x_j)$为求数据点$x_i$和$x_j$之间的欧氏距离,$\rho>1$ 称为伸缩因子.

显然,这样定义的线段长度可以满足上面的性质,从而可以用来描述聚类的全局一致性.另外,我们还可以通过调节伸缩因子ρ来放大或缩短两点间线段长度.根据密度可调节的线段长度,我们需要进一步定义一个新的距离测度,称为密度敏感的距离测度.这里,将数据点看作是一个加权无向图G的顶点V,边集合$E=\{W_{ij}\}$表示的是在每一对数据点间定义的相似度,密度敏感的距离测度通过寻找图G中两点间的最短路径来度量两点间的距离.

定义 2. 将数据点看作是图$G=(V,E)$的顶点,令$p\in V^l$表示图上一个长度为$l=:|p|$的连接点p_1和$p_{|p|}$的路径,其中,边$(p_k,p_{k+1})\in E,1\le k\le|p|$.令$P_{i,j}$表示连接数据点$x_i,x_j$的所有路径的集合.密度敏感的距离测度定义为

$$D_{ij}=\min_{p\in P_{i,j}}\sum_{k=1}^{|p|-1}L(p_k,p_{k+1}),$$

其中,$L(a,b)$表示求两点间密度可调节的线段长度.D_{ij}满足测度的 4 个条件,即对称性$D_{ij}=D_{ji}$;非负性$D_{ij}\ge0$;三角不等式$D_{ij}\le D_{ik}+D_{kj}$(对于任意的x_i,x_j,x_k);自反性$D_{ij}=0$,当且仅当$x_i=x_j$.

密度敏感的距离测度可以度量沿着流形上的最短路径,这使得位于同一高密度区域内的两点可以用许多较短的边相连接,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的边相连接,最终达到了这一目的:放大位于不同高密度区域的数据点间的距离,而缩短位于同一高密度区域内的数据点间的距离.因此,这一距离测度是数据依赖的,而且可以反映出数据的局部密度特征,即所谓的密度敏感.

综上所述,利用密度敏感的距离测度可以充分挖掘无类属数据内部存在的空间一致性先验信息.然而,如果我们仅在聚类算法中利用数据的空间一致性先验信息,那么本质上得到的仍然是一个完全无监督的聚类算法.

对于某些数据集,也许仅能提供有限的空间分布信息,在这种情况下,用户提供的有限的成对限制信息对于聚类搜索所起到的指导作用就变得至关重要.下面我们考虑如何将用户提供的成对限制先验信息和数据的空间一致性先验信息同时引入谱聚类算法中,期望进一步改善算法的聚类性能.

3 结合先验信息的谱聚类算法

3.1 密度敏感的半监督谱聚类

大量研究表明,充分利用领域先验知识可以显著地提高聚类性能.通过上一节的分析我们可以看出,成对限制信息对于聚类算法所起的指导作用是有限的,应该同时考虑利用数据集本身所固有的空间一致性先验信息来辅助聚类搜索,这样便尽可能充分地利用了领域先验知识.由于谱聚类算法是一种配对聚类方法,因而相似性度量的选择与算法在实际问题中的性能有着直接的关系.利用好的相似性度量不仅能够获得好的聚类性能,而且可以克服谱聚类算法对尺度参数选择较为敏感这一缺陷.本文试图充分利用数据自身所能提供的先验信息来指导相似性度量的选择.我们提出同时利用用户提供的成对限制先验信息以及数据空间一致性先验信息来调节数据间的相似性关系,期望获得更加接近于理想矩阵的相似性矩阵,从而利用这一相似性矩阵的谱聚类算法获得较为理想的聚类效果,所提算法称为密度敏感的半监督谱聚类算法(DS-SSC).

首先,我们根据第 2.2 节中提出的利用数据空间一致性先验信息的方法来设计一个数据自适应的相似性度量.基于密度敏感的距离测度,可以很容易地设计出一个新颖的相似性度量,即密度敏感的相似性度量.

$$S_{ij}=\frac{1}{\min\limits_{p\in P_{i,j}}\sum\limits_{k=1}^{|p|-1}(\rho^{dist(x_i,x_j)}-1)+1} \tag{1}$$

与谱聚类算法中经常使用的 Gauss 核函数作为相似性度量相比,该相似性度量不需要引入核函数,可以直接在距离测度上计算相似性关系.此外,该相似性度量能够根据数据空间分布特性自动调节相似性关系,从而可以更加准确地反映数据间的相似性关系,使得相似性矩阵更加接近于理想矩阵.下面给出 DS-SSC 算法的执行步骤.

算法 1. 密度敏感的半监督谱聚类算法(DS-SSC).

1. $\forall x_i,x_j\in X$,计算两点间欧式距离 $D_{ij}=\left(\left\|x_i-x_j\right\|_2^2\right)^{\frac{1}{2}}$;
2. 添加成对限制信息: $\begin{cases}D_{ij},D_{ji}=0, & \text{if } (x_i,x_j)\in must\text{-}link\\ D_{ij},D_{ji}=\infty, & \text{if } (x_i,x_j)\in cannot\text{-}link\end{cases}$;
3. $\forall x_i,x_j\in X$,计算两点间密度敏感的距离: $D_{ij}=\min\limits_{p\in P_{i,j}}\sum\limits_{k=1}^{|p|-1}(\rho^{dist(x_i,x_j)}-1)$;
4. 根据密度敏感的相似性度量构造相似性矩阵$S\in R^{n\times n}$,其中, $S_{ij}=\frac{1}{D_{ij}+1}$,$S_{ii}=0$;
5. 构造Laplacian矩阵$P=L^{-1/2}SL^{-1/2}$,其中L为对角度矩阵 $L_{ii}=\sum\limits_{j=1}^{n}S_{ij}$;
6. 求P的k个最大特征值所对应的特征向量$v_1,v_2,\ldots,v_k$(重复特征值选择正交的特征向量),构造矩阵$V=[v_1,v_2,\ldots,v_k]\in R^{n\times k}$,其中,$v_i$为列向量;
7. 规范化 V 的行向量,得到矩阵 Y,其中, $Y_{ij}=V_{ij}/(\sum_j V_{ij}^2)^{\frac{1}{2}}$;
8. 将Y的每一行看成是R^k空间内的一点,使用K均值将其聚为k类;
9. 如果Y的第i行属于第j类,则将原数据点x_i也划分到第j类.

DS-SSC 算法首先通过直接修改距离矩阵的方法来施加成对限制先验信息.然后,利用密度敏感的距离测度将有限的限制信息进行空间传播.由于密度敏感的距离测度可以发现数据内部固有的空间分布信息,这种传

播方法是根据数据聚类的空间分布特性来传播限制的影响,因而可以有效地避免由于提供了信息含量少的限制所造成的对聚类的误导作用.算法接着根据密度敏感的相似性度量计算各个数据点间的相似性关系,这一相似性关系同时反映了成对限制先验信息和空间一致性先验信息对相似性关系的影响.对最终得到的相似性矩阵进行谱特征分解,由于该相似性矩阵更加接近于理想矩阵,数据集合在映射空间便可以形成较容易识别的聚类簇,最后使用 K 均值方法便可得到理想的聚类效果.

在 DS-SSC 算法中,成对限制先验信息起到直接修改距离测度的作用,使得相似性关系随限制发生改变,而空间一致性先验信息根据数据聚类的空间分布特性自动调节数据间的距离,起到间接修改相似性关系的作用,而且能够同时传播成对限制的影响作用,并且不断根据数据分布特性修正限制可能引起的偏斜.在整个聚类过程中,两类先验信息相辅相成,最大限度地发挥了其对聚类搜索的指导作用.

3.2 空间一致性先验有效性分析

本节我们利用一个定量指标来分析空间一致性先验信息指导聚类的有效性.在一般情况下,相似性矩阵可以看作是块对角相似性矩阵 S 经过扰动得到的扰动矩阵 $\hat{S}=S+E$ .Ng 等人利用矩阵扰动理论分析了一般情况下谱聚类算法取得成功的条件:

引理 1[9]. 如果 $\delta>(2+\sqrt{2})\varepsilon$,则存在k个正交向量$r_1,r_2,\ldots,r_k$(如果$i=j$, $r_i^T r_j=1$,否则 $r_i^T r_j=0$),使得由扰动矩阵 $\hat{S}$ 得到的矩阵Y的行向量满足

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\| y_j^{(i)}-r_i\|_2^2\le 4C(4+2\sqrt{k})^2\frac{\varepsilon^2}{(\delta-\sqrt{2}\varepsilon)^2} \tag{2}$$

其中,$\delta,\varepsilon,C>0$ 为依赖于扰动矩阵 $\hat{S}$ 的常数.

根据引理 1,当满足一定条件时,矩阵 Y 的行向量在 k 个位于 k 维球上的点附近形成紧的聚类.不等式(2)左边可以看作是谱聚类算法在特征空间中获得聚类结果与真实聚类间的偏斜.这是由于如果矩阵 Y 是由理想矩阵得到的,那么,它的行向量在 k 维映射空间便形成上述 k 个相互正交的点;而当 Y 是由扰动矩阵得到的,它的行向量便在这k个相互正交的点附近形成扰动.因此,这个量越小,则表明聚类结果越接近于真实聚类.然而,由于理想矩阵一般是未知的,这一偏斜程度通常无法直接计算得到.由于我们在后处理方法中采用的是 K 均值算法,所以很自然地考虑利用 K 均值算法的目标函数值来逼近这一偏斜程度,即

$$dist(\Delta)=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\| y_j^{(i)}-c_i\|^2 \tag{3}$$

其中, $c_i=\frac{1}{|C_i|}\left(\sum_{j\in C_i}y_j^{(i)}\right)$ 为聚类划分Δ中的每一类C_i的中心.因此,我们可以利用该指标来定量分析密度敏感的距离测度所挖掘的空间一致性先验信息对聚类指导的有效性.

以USPS手写体数字为例.这里取没有利用先验信息的SC算法[9]在该数据集上的最优聚类结果所对应的模型参数,然后求SC算法和没有利用成对限制信息的DS-SSC算法在该参数设置下的K均值目标函数值.从表 1 可以看出,在SC算法取得最优聚类的情况下,没有利用成对限制信息的DS-SSC算法可以获得更小的目标函数值,这使得后续的K均值算法可以较为容易地发现聚类,从而验证了空间一致性先验信息指导聚类的有效性.

Table 1 Target function value of K-means on USPS data sets

表 1 USPS 数据集上的 K 均值目标函数值

Digits	Target function value	
	DS-SSC	SC
0,8	56.34	188.86
3,5,8	155.21	167.39
1,2,3,4	162.01	194.15

3.3 相关工作比较

Kamvar等人通过直接修改相似性矩阵的方法将成对限制信息引入谱聚类算法中,提出了受限的谱聚类算法[11].Kamvar使用相似性矩阵$N=(W+d_{\max}I-D)/d_{\max}$,其中,$d_{\max}$为度矩阵$D$中的最大元素,$I$为单位矩阵.Xu等人认为,当数据集合中包含奇异点时,矩阵N中的非零对角元素容易造成单元素聚类.为了克服这一缺陷,在他们所提出的CSC(constrained spectral clustering)算法[13]中采用相似性矩阵$P=D^{-1}W$.

与 CSC 不同,DS-SSC 首先根据成对限制信息修改距离矩阵,然后利用密度敏感的距离测度将限制信息进行空间传播,该方法不仅充分利用了有限的限制信息,而且在传播限制的同时,通过密度敏感的距离测度引入数据的空间一致性结构信息,间接修改了相似性矩阵,使得最终得到的相似性矩阵更接近于反映数据间正确相似性关系的理想矩阵.同时,空间一致性结构信息能够间接修正由信息含量少的成对限制所造成的对聚类的误导,这一点从仿真实验中可以得到证实.

Klein等人提出受限的完全链接层次聚类算法[4](CCL),该算法的思想已在第 2.1 节中加以阐述.由于链接层次聚类算法本身的局限性,使得CCL对用户提供的限制信息极其敏感,加之其采用的传播方法,更进一步加重了限制信息对聚类搜索所起的误导的反作用,最终导致错误划分.

4 实 验

为了验证同时将两类先验信息引入谱聚类算法相对于在谱聚类中仅结合成对限制信息或空间一致性信息所能获得的聚类性能上的进一步提高,下面我们分别在一组UCI基准数据集[14]、USPS手写体数据集以及一组TREC文本数据集上进行仿真实验.

4.1 实验方法与评价准则

本文所有实验中限制的数目取自 0~200 之间.对于每一个给定的限制数目进行 100 次实验,输出平均结果.由于所选限制的不同对于聚类算法的性能有着很大的影响,为了实验的公平性,我们采用如下方法随机产生限制:对于同一个限制数目要产生 100 组不同的限制,为此首先采用不同的种子初始化随机数发生器,种子数目从 1 递增到 100,这样便可以得到同一限制数目下不同的限制集合.另外,DS-SSC 和 CSC 算法中还涉及到参数选择问题.对于所有实验,我们采用的是网格搜索的方法来自动选择参数,算法汇报在最优参数上的聚类性能.

聚类算法性能的评价一直是一个具有挑战性的问题.这里将聚类划分看作是样本之间的一种关系,即每一对样本要么被划分在同一类,要么在不同类.对于一个具有n个样本的数据集,存在$n(n-1)/2$ 个样本对,也就是说,对于一个聚类划分存在$n(n-1)/2$ 个成对决策.Wagstaff[8]根据算法所获得的正确决策数来评价聚类算法的性能,即使用所谓的Rand指标: $RI=\dfrac{\#CD}{n(n-1)/2}$,其中,$\#Cn$表示由算法所获得的正确决策对数.为了反映限制信息对算法性能的影响以及方便与已有算法进行比较,我们仅在不是由限制所确定的决策上评价聚类性能,因此采用修正Rand指标[6]: $CRI=\dfrac{\#CD-\#Cn}{n(n-1)/2-\#Cn}$,其中,$\#Cn$表示所有限制的数目,包括根据用户提供的限制直接传播所得的所有限制.本文实验中提到的聚类精确度即指CRI值.

4.2 基准数据集

表 2 给出了实验数据集的数据特征.图 2 分别绘出了 DS-SSC,CSC,CCL 这 3 种方法在 4 个数据集上的聚类精确度.从图中我们可以观察到:

- 当不提供成对限制信息时,即在图中的初始点上,除了在 Sonar 数据集上 3 种算法性能相当以外,在其他 3 个数据集上两种谱聚类算法的性能均优于 CCL 算法.这说明谱聚类算法本身在聚类性能上相对于完全链接层次聚类算法具有一定的优势.注意,这时的 DS-SSC 算法仅利用了空间一致性先验信息,我们可以观察到,除了在 Ionosphere 数据集上其性能明显优于不结合任何先验信息的谱聚类算法外,在其他 3 个数据集上两者性能相当.这说明,单独使用空间一致性先验信息对聚类的指导作用不明显.

- 当仅提供成对限制信息时,在 Iris 数据集上,CCL 算法出现了明显的性能波动,而且仅在限制数少的情况下算法性能有所改善;而随着限制数目的增多,算法性能反而下降.这说明对于有聚类重叠的数据集,成对限制对于完全链接层次聚类算法起到了误导聚类搜索的反作用.相反地,DS-SSC 和 CSC 在 Iris 上取得了很好的聚类效果,而且 DS-SSC 在少量限制数目的情况下,性能优于 CSC.对于另外 3 个数据集,我们观察到,CSC 随着限制数目的增多其性能没有显著的提高.当限制数目比较多时,在 Glass 和 Sonar 数据集上,由于没有采用限制传播机制,其性能反而不如 CCL.这说明,对于谱聚类算法仅采用修改相似性矩阵的方法来结合成对限制信息是远远不够的,必须采取一定的传播限制的方式.
- 当仅结合空间一致性先验信息时,DS-SSC 仅在 Ionosphere 数据集上的性能优于其他两类算法.随着限制数目的增多,其他两类算法的性能优于仅结合空间一致性先验信息的 DS-SSC 算法.这说明,没有用户提供的成对限制先验信息,完全依靠数据自身所能提供的先验信息对聚类起到的指导作用是有限的.
- 当同时结合两类先验信息时,在所有 4 个数据集上,DS-SSC 算法相对于 CSC 和 CCL 在聚类精确度上有了显著的提高.由图 2 所示的 4 幅图中我们可以看出,随着限制信息的逐步加入,DS-SSC 算法的性能曲线逐步攀升,这说明,成对限制信息与空间一致性信息在指导聚类搜索上起到了相辅相成的作用,两者共同使得谱聚类算法获得了性能上的大幅提高.空间一致性先验信息充分挖掘了数据本身包含的聚类结构信息,并且起到传播成对限制信息的作用.成对限制信息在这里起到了辅助搜索聚类结构的作用,这样,随着限制数目的增多,聚类性能得到了显著提高.此外,同时结合两类先验信息克服了仅结合成对限制信息所造成的对聚类的误导.根据上述分析,两类先验信息的结合不是简单的结合,而是一种有机的结合,两者相互补充,相互作用,共同指导聚类搜索.

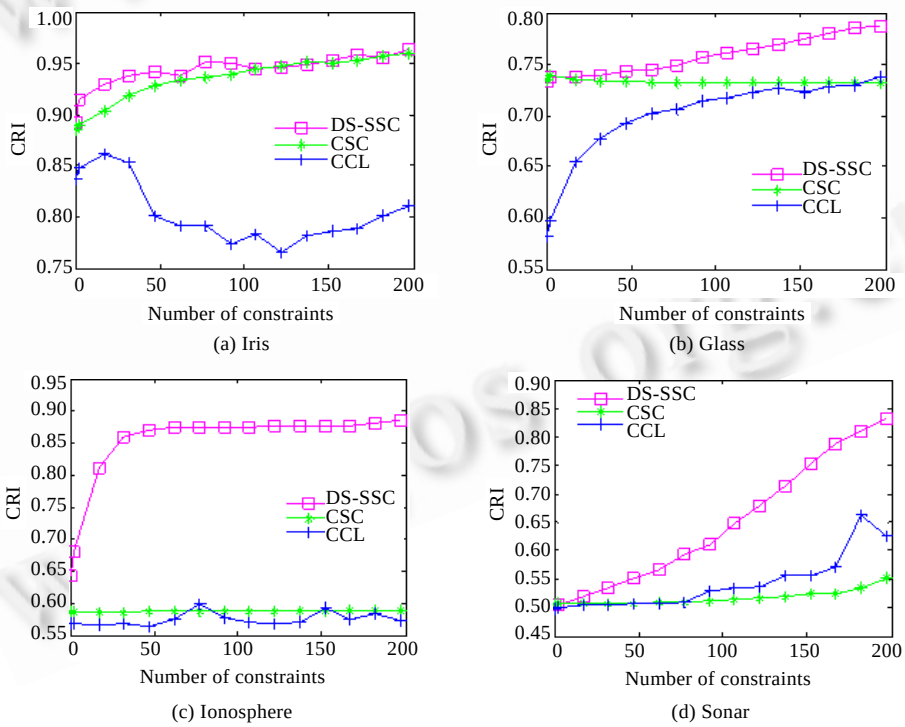


Fig.2 Comparisons of DS-SSC, CSC and CCL on UCI benchmark data sets

图 2 基准数据集上的对比实验结果

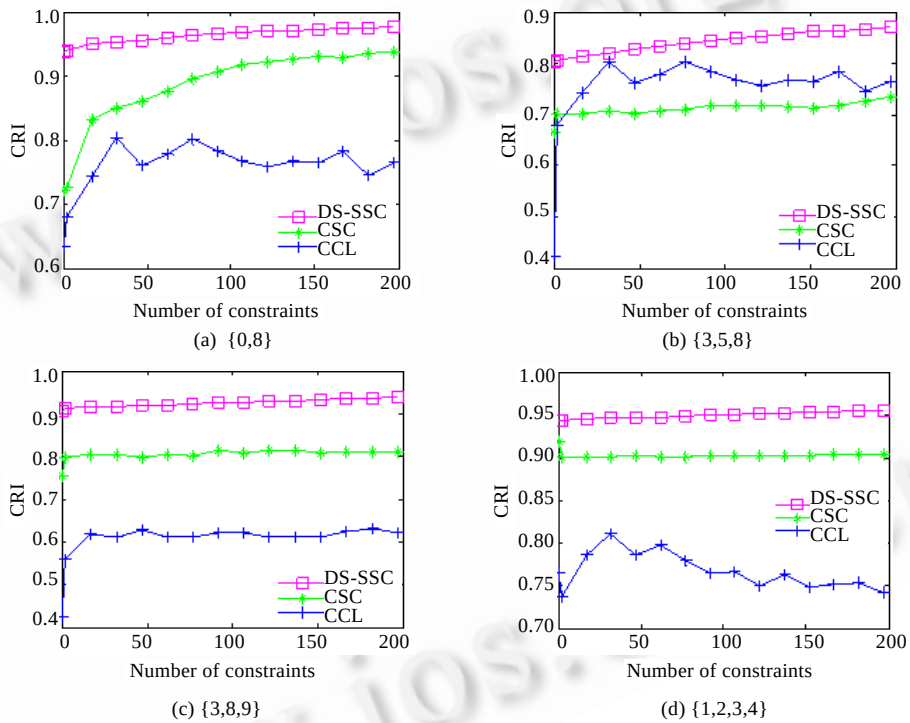
Table 2 Information on benchmark data sets

表 2 基准数据集的数据特征

Data set	No. of instances	No. of attributes	No. of classes
Iris	150	4	3
Glass	214	9	6
Ionosphere	351	34	2
Sonar	208	60	2

4.3 USPS手写体数字集

USPS 数据集是由 9 298 个 16×16 维灰度图像构成,其中包含 7 291 个训练样本,2 007 个测试样本.实验取全部测试样本作为聚类数据集,从中挑选 3 组较难识别和 1 组相对容易识别的数字集合进行识别,即数字{0,8},{3,5,8},{3,8,9},{1,2,3,4}.图 3 分别绘出了 3 种方法对 4 组数字集合的识别结果.



(a) {0,8} (b) {3,5,8}

(c) {3,8,9} (d) {1,2,3,4}

Fig.3 Comparisons of DS-SSC, CSC and CCL on USPS data sets

图 3 USPS 手写体数字集上的对比实验结果

从对比图中我们可以得出以下结论:

- DS-SSC 算法在所有 4 个数字集合上的聚类性能均明显优于 CSC 和 CCL 算法.值得一提的是,当不提供成对限制信息而仅利用空间一致性先验信息时(即在图中 DS-SSC 算法的初始点处),DS-SSC 算法在所有数字集合上的聚类性能仍然明显优于 CCL 和 CSC 算法.这说明,在该类问题上,空间一致性信息对于聚类起着关键的指导作用,DS-SSC 算法充分利用了这一信息,使得谱聚类算法的聚类性能得到了显著提高.当逐步加入成对限制信息后,我们观察到算法的聚类性能得到了逐步的提高.这说明,同时利用成对限制信息和空间一致性结构信息不仅能够充分利用有限的限制信息,而且挖掘了数据本身的聚类结构信息,起到更进一步提升算法聚类性能的作用.另外,对于较容易识别的{1,2,3,4}数字集,随着限制的增多,DS-SSC 算法的识别性能没有得到显著的提升.这说明对于较容易识别的问题,DS-SSC 仅利用空

间一致性先验信息便可充分发现聚类结构,而限制信息在这里仅能起到有限的提升聚类性能的作用.

- 当成对限制的数目逐渐增多时,CSC 算法相对于完全无监督谱聚类算法仅在数字{0,8}上表现出性能的显著提高,而在其他数字集合上其性能仅有很少的提升.在数字集合{3,5,8}上,CCL 算法的性能超过了CSC;在{1,2,3,4}数字集上,CSC 算法的性能甚至低于没有结合限制信息的谱聚类算法.这说明,对于较为复杂的聚类分布,对相似性矩阵修改的方法对于限制信息的利用作用是有限的,而且当类属增多时甚至起到了误导聚类的反作用.CCL 中的限制传播方法在这里起到了一定的作用,然而在 4 组实验中我们仍然观察到这种传播限制的方法的不稳定性.相反地,对于 DS-SSC 算法,空间一致性信息的利用起到了间接传播限制的作用,从而随着限制的加入,算法的聚类性能得到了逐步提高.

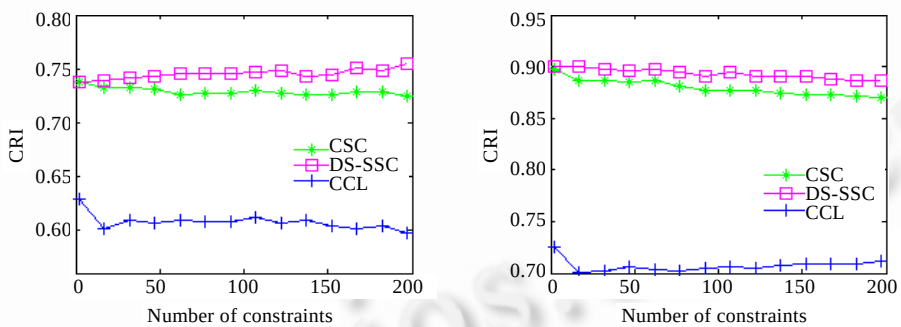
4.4 文本数据集

此外,我们还在TREC的文本数据集*tr*23 和*tr*11 上进行了实验.首先对实验所采用的数据集进行预处理,去除停用词,并进行词干抽取,然后去除那些在少于 3 个文本中出现的词后,进行词频统计.我们使用*ti*×*idf*来衡量文本中每个词的权重.表 3 给出了所用数据集的详细信息,其中,n_d表示文本数,n_w表示字数,k表示类数,Balance表示最小类规模和最大类规模之比.图 4 分别绘出了 3 种方法在文本数据集上的性能比较.从对比图中可以看出,对于*tr*11,数据集成对限制信息的加入并没有起到指导聚类的作用.这是由于该数据集各类间的规模极度不平衡所造成的.相反地,由于结合了空间一致性先验信息,DS-SSC算法获得了优于其他两类算法的聚类性能.

Table 3 Information on text data sets

表 3 文本数据集特征

Data	n_d	N_w	k	Balance
*tr*23	204	5,832	6	0.066
*tr*11	414	6,429	9	0.046



(a) *tr*23 (b) *tr*11

Fig.4 Comparisons of DS-SSC, CSC and CCL on text data sets from TREC

图 4 TREC 文本数据集上的对比实验结果

5 结束语

在许多实际应用问题中,成对限制先验信息可以很容易地获得.我们认为,仅利用这一先验信息对于提高聚类算法的性能是远远不够的,而且当用户提供了信息含量少的限制时,对于聚类搜索反而起到了误导的反作用.本文尝试探索一种存在于数据集合内部的先验信息——空间一致性先验信息,并提出了利用该类信息的具体方法.将两类先验信息同时引入经典的谱聚类算法中,提出了一种密度敏感的半监督谱聚类算法(DS-SSC).该算法通过密度敏感的相似性度量所获得的相似性关系,同时反映了限制信息和空间一致性先验信息对相似性关系的影响,使得最终得到的相似性矩阵更加接近于理想矩阵.该矩阵经过特征分解后将更加有助于在映射空间的后续聚类划分.仿真实验结果充分验证了上述结论.

本文所讨论的成对限制信息是由用户任意提供的.从实验中可以看出,这样提供的限制信息对于聚类算法不一定会起到积极的指导作用.如何主动提供给聚类算法信息含量丰富的成对限制,已成为半监督聚类的一个研究热点.我们的下一步工作包括,如何为谱聚类算法主动地提供信息含量丰富的限制信息,使其仅利用有限的限制信息便可获得聚类性能的提高.

References:

[1] Yu SX, Shi J. Segmentation given partial grouping constraints. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2004, 26(2):173−183.

[2] Hertz T, Shental N, Bar-Hillel A, Weinshall D. Enhancing image and video retrieval: Learning via equivalence constraint. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. Madison: IEEE Computer Society, 2003. 668−674.

[3] Wagstaff K, Cardie C, Rogers S, Schroedl S. Constrained *K*-means clustering with background knowledge. In: Brodley CE, Danyluk AP, eds. Proc. of the 18th Int'l Conf. on Machine Learning. Williamstown: Morgan Kaufmann Publishers, 2001. 577−584.

[4] Klein D, Kamvar SD, Manning CD. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In: Sammut C, Hoffmann AG, eds. Proc. of the 19th Int'l Conf. on Machine Learning. Sydney: Morgan Kaufmann Publishers, 2002. 307−314.

[5] Wagstaff K, Cardie C. Clustering with instance-level constraints. In: Langley P, ed. Proc. of the 17th Int'l Conf. on Machine Learning. Morgan Kaufmann Publishers, 2000. 1103−1110.

[6] Zhou D, Bousquet O, Lal TN, Weston J, Schölkopf B. Learning with local and global consistency. In: Thrun S, Saul L, Schölkopf B, eds. Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2004. 321−328.

[7] Shi J, Malik J. Normalized cuts and image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000,22(8): 888−905.

[8] Gu M, Zha H, Ding C, He X, Simon H. Spectral relaxation models and structure analysis for *k*-way graph clustering and bi-clustering. Technical Report, CSE-01-007, Penn State University, 2001.

[9] Ng AY, Jordan MI, Weiss Y. On spectral clustering: Analysis and an algorithm. In: Dietterich TG, Becker S, Ghahramani Z, eds. Advances in Neural Information Processing Systems (NIPS) 14. Cambridge: MIT Press, 2002. 894−856.

[10] Meila M, Xu L. Multiway cuts and spectral clustering. Technical Report, 442, University of Washington, 2004.

[11] Kamvar SD, Klein D, Manning CD. Spectral learning. In: Gottlob G, Walsh T, eds. Proc. of the 18th Int'l Joint Conf. on Artificial Intelligence. Acapulco: Morgan Kaufmann Publishers, 2003. 561−566.

[12] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning distance functions using equivalence relations. In: Fawcett T, Mishra N, eds. Proc. of the 20th Int'l Conf. on Machine Learning. Washington: AAAI Press, 2003. 11−18.

[13] Xu QJ, DesJardins M, Wagstaff K. Constrained spectral clustering under a local proximity structure assumption. In: Russell I, Markov Z, eds. Proc. of the 18th Int'l Conf. of the Florida Artificial Intelligence Research Society, Clearwater Beach. Florida: AAAI Press, 2005. http://www.litech.org/~wkiri/Papers/xu-short-flairs05.pdf

[14] Blake C, Keogh E, Merz CJ. UCI repository of machine learning databases. Irvine: University of California, 1998. http://www.ics.uci.edu/~mlearn/MLRepository.html

王玲(1978－),女,陕西西安人,博士生,主要研究领域为统计机器学习,模式识别.

薄列峰(1978－),男,博士生,主要研究领域为核机器学习,流形学习,神经网络,计算机视觉.

焦李成(1959－),男,教授,博士生导师,CCF高级会员,主要研究领域为非线性科学,智能信息处理,大规模并行处理.