密度敏感的谱聚类

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)

摘 要: 谱聚类是近来出现的一种性能极具竞争力的聚类方法,它的成功很大程度依赖于相似性度量的选择. 本文通过分析这一性质并结合数据聚类特性,提出一种数据依赖的相似性度量——密度敏感的相似性度量. 该相似性度量可以有效描述数据的实际聚类分布. 将其引入谱聚类得到密度敏感的谱聚类算法. 与原有的谱聚类算法相比,新算法不仅能够处理多尺度聚类问题,而且对参数选择相对不敏感. 算法有效性分析以及实验验证了所提算法的有效性和可行性.

关键词: 聚类; 谱聚类; 距离测度; 相似性度量; 相似性矩阵
中图分类号: TP18 **文献标识码**: A **文章编号**: 0372-2112 (2007) 08-1577-05

Density-Sensitive Spectral Clustering

WANG Ling, BO Lie-feng, JIAO Li-cheng
(Institute of Intelligent Information Processing, Xidian University, Xi'an, Shaanxi 710071, China)

Abstract: Spectral clustering has become increasingly popular in recent years. Being a pairwise method, the success of spectral clustering depends heavily on the choice of similarity measure. Through analyzing the property of data clusters, a novel data-dependent similarity measure is proposed, namely density-sensitive similarity measure, which has the ability of describing the characters of data clustering compared with the traditional Euclidian metric based similarity measure. Based on the novel similarity measure, we have a density-sensitive spectral clustering algorithm. Compared with the original spectral clustering, it has the advantages of effectively dealing with the multi-scale problems and relatively not sensitive to parameter. It obtains promising results not only on artificial datasets but also on USPS handwritten digit dataset.

Key words: clustering; spectral clustering; distance metric; similarity measure; similarity matrix

1 引言

聚类问题一直是模式识别领域一个比较活跃而且极负挑战性的研究方向. 现有的基于产生式模型的聚类方法由于使用参数密度估计,不得不简化问题的模型,如假设每一类的分布是高斯分布. 这使得算法仅在具有凸形结构的数据集上有好的聚类效果,不适于具有任意复杂分布形状的聚类问题. 其他算法如基于中心的聚类方法在紧凑的具有超球形分布的数据集上具有很好的聚类性能[1],但这些算法仍不适合任意形状聚类问题.

最近,一种称为谱聚类的聚类方法开始得到关注,它的思想来源于谱图划分理论[2,3]. 谱聚类仅与数据点的数目有关,而与维数无关,因而可以避免由高维特征向量造成的奇异性问题. 谱聚类又是一种判别方法,不对数据的全局结构作假设. 尽管是一种极具竞争力的聚类方法,但其目前仍处在研究初期,算法本身存在一些亟待解决的问题. 现有的算法对分析尺度的选择非常敏感,使得尺度参数的正确选择成为算法成功的关键. 使用者必须花费大量的精力用于选取参数. 在真实世界问题中,数据通常具有多重尺度,谱聚类仍然不适合处理一些多尺度聚类问题[4]. 本文试图改进谱聚类算法以解决上述问题.

本文设计了一种简单有效的相似性度量——密度敏感的相似性度量,它可以放大不同高密度区域内数据点间距离,同时缩短同一高密度区域内数据点间距离,最终有效描述数据的实际聚类分布. 将其引入谱聚类得到密度敏感的谱聚类算法. 有效性分析和实验表明,所提算法相对于原有算法在聚类性能上有了显著提高.

2 密度敏感的谱聚类

我们观察到数据聚类具有如下两个所谓的一致性特征,这与半监督学习中数据的一致性先验假设恰好是相符的[5].

(1) 局部一致性 指的是在空间位置上相邻的数

收稿日期:2006-06-02;修回日期:2007-04-26
基金项目:国家自然科学基金(No. 60372050,60133010);国家 863 高技术研究发展计划(No. 2002AA135080);国防预研项目(No. A1420060172)

据点具有较高的相似性;

(2) 全局一致性 指的是位于同一流形上的数据点具有较高的相似性.

传统的欧氏距离仅能反映聚类结构的局部一致性特征,而不能反映全局一致性特征. 用一个简单例子说明该问题. 如图 1(a) 所示,期望的是在某一测度下点 a 和点 c 更接近. 在欧氏距离测度下,点 a 更接近于点 b,这便没有反映聚类的全局一致性. 对于复杂问题,简单的基于欧氏距离的相似性度量不能完全反映聚类结构.

从数据的空间分布情况观察到,同一聚类内的数据趋向于分布在一个密度较高的区域,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. 我们可以根据数据的这一局部密度特征来设计一类数据依赖的相似性度量.

将数据点看作是一个加权无向图 $G=(V, E)$ 的顶点 V,边集合 $E=\{W_{ij}\}$ 表示数据点间的相似度. 期望设计这样的相似性度量:如果两个数据点可以由一条穿过高密度区域的路径相连接,则这两点间将被赋予较高的相似度,否则将被赋予较低的相似度. 有实验表明,利用这一相似性度量的思想可以显著改善半监督学习分类精度[6]. 用图 1(a) 来解释这一思想,也就是说要设计这样的相似性度量,使得点 a 和点 c 之间的距离比点 a 和点 b 之间的距离短. 最终达到的目的是:放大那些穿过低密度区域的路径长度,而同时缩短那些没有穿过低密度区域的路径长度.

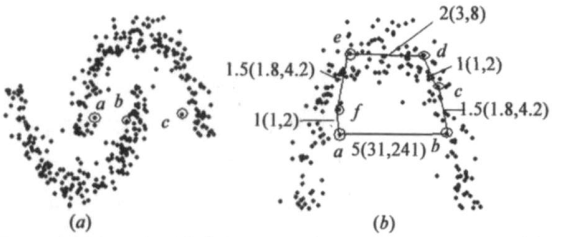


图 1 (a)寻找距离测度使得点 a 更接近于点 c;(b) 全局一致性距离满足:$\overline{af}+\overline{fe}+\overline{ed}+\overline{dc}+\overline{cb}<\overline{ab}$ (括号外的数字为欧氏距离,括号内的数字为 ρ 取 2 和 3 时计算的线段长度)

为达到该目的,首先需要设计一个密度可调节的线段长度. 观察到这样一个性质: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路径不一定最短. 如图 1(b) 所示,为了满足聚类的全局一致性,必须使得穿过高密度区域用较短边相连的路径长度比穿过低密度区域直接相连的两点间距离来得短,即图 1(b) 中 $\overline{af}+\overline{fe}+\overline{ed}+\overline{dc}+\overline{cb}<\overline{ab}$.

定义 1 密度可调节的线段长度:

$$L(\mathrm{x}, \mathrm{y}) = \rho^{dist(\mathrm{x}, \mathrm{y})} - 1 \qquad (1)$$

其中 $dist(\mathrm{x}, \mathrm{y})$ 为求数据点 x 和 y 间的欧氏距离,$\rho>1$ 称为伸缩因子.

这样定义的线段长度满足上面的性质,可以用来描述聚类的全局一致性. 我们通过调节伸缩因子 ρ 来放大或缩短两点间线段长度.

基于密度可调节的线段长度,进一步定义密度敏感的距离测度:

定义 2 将数据点集 $V=\{\mathrm{x}\}_{i=1}^{n}$ 看作是图 $G=(V, E)$ 的顶点. 令 $\rho=\{p_1 \to p_2 \to \cdots \to p_l\} \in V^l$ 表示图上一顶点数为 $l=:|\rho|$ 的连接点 p_1 和 p_l 的路径,其中边 $(p_k, p_{k+1}) \in E, 1 \le k < l$. 令 $P_{ij}, 1 \le i, j \le n$,表示连接数据点对 $\mathrm{x}_i, \mathrm{x}_j$ 的所有路径的集合,则数据点对 $\mathrm{x}_i, \mathrm{x}_j$ 间的密度敏感的距离定义为:

$$D_{ij} = \min_{\rho \in P_{ij}} \sum_{k=1}^{l-1} L(p_k, p_{k+1}) \qquad (2)$$

其中 $L(\cdot, \cdot)$ 表示两点间密度可调节的线段长度.

定理 1 对于数据点 $\mathrm{x}_i, \mathrm{x}_j, \mathrm{x}_k \in V, 1 \le i, j, k \le n$,密度敏感的距离满足测度的四个性质:

(1) 非负性 $D_{ij} \ge 0$; (2) 自反性 $D_{ij}=0$ 当且仅当 $\mathrm{x}_i = \mathrm{x}_j$; (3) 对称性 $D_{ij}=D_{ji}$; (4) 三角不等式 $D_{ij} \le D_{ik} + D_{kj}$.

证明:欧氏距离满足非负性、对称性和自反性. 显然,根据定义 1,密度可调节的线段长度也满足非负性、对称性和自反性. 由于密度敏感的距离为对密度可调节的线段长度的线性求和,从而也满足这三个性质.

根据密度敏感的距离定义,$\forall \mathrm{x}_i, \mathrm{x}_j, \mathrm{x}_k \in V$,经过数据点 x_k 连接 x_i 和 x_j 的路经的最短长度为 $D_{ik}+D_{kj}$. 由于 D_{ij} 为连接 x_i 和 x_j 的所有路径中的最短路径,从而 $D_{ij} \le D_{ik}+D_{kj}$. 定理 1 证明完毕.

算法 1 密度敏感的谱聚类算法(DSSC)

输入:n 个数据点 $\{\mathrm{x}_i\}_{i=1}^{n}$,聚类数目 k

输出:数据点集的划分 $C_1, \cdots, C_k$

(1) 根据密度敏感的相似性度量构造相似性矩阵 $\mathrm{W} \in R^{n \times n}$,其中

$$\mathrm{W}_{ij} = \frac{1}{\min\limits_{\rho \in P_{ij}} \sum\limits_{k=1}^{l-1} (\rho^{dist(p_k, p_{k+1})} - 1)}, \quad \mathrm{W}_{ii}=0;$$

(2) 构造 Laplacian 矩阵 $\mathrm{P}=\mathrm{D}^{-1/2}\mathrm{W}\mathrm{D}^{-1/2}$,其中 D 为对角度矩阵

$$\mathrm{D}_{ii} = \sum_{j=1}^{n} W_{ij};$$

(3) 求 P 的 k 个最大特征值对应的特征向量 $\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_k$,构造矩阵 $\mathrm{V}=[\mathrm{v}_1, \mathrm{v}_2, \cdots, \mathrm{v}_k] \in R^{n \times k}$,其中 v_i 为列向量;

(4) 规范化 V 的行向量,得到矩阵 Y,其中 $\mathrm{Y}_{ij} = \mathrm{V}_{ij} / \left(\sum_j \mathrm{V}_{ij}^2\right)^{1/2}$;

(5) 将的 Y 每一行看成是 R^k 空间内的一点,使用 K 均值或其他算法将其聚为 k 类;

(6) 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.

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:放大位于不同高密度区域的数据点间距离,而缩短位于同一高密度区域内的数据点间距离. 因此,这一距离测度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. 图 1(*b*) 说明伸缩因子如何受密度的控制. 显然,利用式(1) 计算的线段长度可以起到相对缩小同一流形上数据间距离的作用,ρ 的增大使得这一作用更加显著,从而达到对数据密度的敏感.

基于密度敏感的距离测度可以很容易地设计一个新颖的相似性度量,称为密度敏感的相似性度量:

$$W_{ij} = \frac{1}{\min\limits_{\rho \in P_{ij}} \sum\limits_{k=1}^{l-1} (\rho^{dist(p_k, p_{k+1})} - 1) + 1} \tag{3}$$

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. 需要指出的是,密度敏感的相似性度量仅涉及一个自由参数 ρ,而文献[7]的方法是通过高斯核函数获得相似性度量,从而引入两个自由参数,增加了参数选择的困难度. 根据上述分析,将密度敏感的相似性度量引入谱聚类算法中,得到密度敏感的谱聚类算法.

3　算法有效性分析

首先给出谱聚类涉及的几个重要概念[8]:

谱映射:如果 n 维向量集合 $v^1, v^2, \cdots, v^k$ 相对于一个聚类 $\Delta = \{C_1, \cdots, C_k\}$ 是分段常数向量,那么谱映射 $i \rightarrow (v_i^1, v_i^2, \cdots, v_i^k)$ 将每一个类 $C_k \in \Delta$ 映射到 R^k 中的唯一点. 所谓 n 维向量 v 对于聚类 Δ 是分段常数向量是指:如果 i, j 在同一个类,那么 $v_i = v_j$.

理想矩阵:如果每个类 $C_k \in \Delta$ 经过谱映射以后简化为单个点,则所使用的相似性矩阵 W 对于算法是理想的.

块对角矩阵:对于一个聚类 $\Delta = \{C_1, \cdots, C_k\}$,当数据点 i, j 属于不同类时 $W_{ij} = 0$,则称矩阵 W 为对应于聚类 Δ 的块对角矩阵.

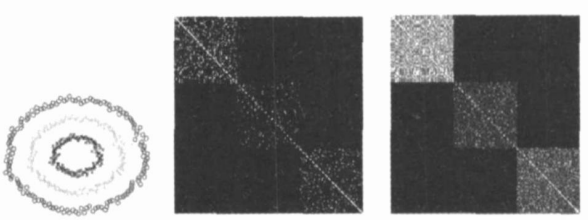
块随机矩阵:随机矩阵 P 称为相对于一个聚类 $\Delta = \{C_1, \cdots, C_k\}$ 是块随机的,当且仅当对于所有的 $s, s' = 1, \cdots, k$,求和 $P_{is} = \sum\limits_{j \in C_s} P_{ij}$ 对于所有的 $i \in C_s$ 都是相等的,并且矩阵 $\bar{P} = [\bar{P}_{ss'}]$(其中 $\bar{P}_{ss'} = \sum\limits_{j \in C_{s'}} P_{ij}, i \in C_s$) 是非奇异的.

从前两个概念可以看出,谱映射本身具有产生理想聚类的功能. 如果相似性矩阵是理想的,则在 R^k 中聚类映射后的数据点将变得相当容易. 理想相似性矩阵代表了谱聚类的理想情况. 很容易证明块对角相似性矩阵对于所有谱聚类算法都是理想矩阵. 然而实际情况中相似性矩阵一般不是块对角矩阵,那么怎样判断相似性矩阵是否为理想矩阵呢?可以将数据点间的相似性关系看作是一个 Markov 随机游走中的边流,谱聚类中规范化的相似矩阵 P 即为该 Markov 链的转移概率矩阵[9]. 引理 1 给出了一个矩阵为理想矩阵的充要条件.

引理 1[8]　相似性矩阵 W 对于聚类 $\Delta = \{C_1, \cdots, C_k\}$ 为理想矩阵,当且仅当规范化的相似性矩阵 P 对于聚类 $\Delta = \{C_1, \cdots, C_k\}$ 是块随机矩阵.

根据引理 1,如果 P 对应的 Markov 链可以聚集成一个具有状态空间 $\Delta = \{C_1, \cdots, C_k\}$ 和转移概率矩阵 $\bar{P}$ 的 Markov 链,则相似性矩阵对于算法来说是理想的. 也就是说,规范化的相似性矩阵 P 经过一定步数的随机走动可以很明显地显示出块状分布结构.

当相似性矩阵是块对角矩阵这一理想情形时,谱聚类算法可以找到完全正确的聚类. 谱聚类算法成功的关键是选择合适的相似性度量,使得产生的相似性矩阵具有明显的块分布. 图 2 显示基于高斯核函数和密度敏感的相似性度量计算出的相似性矩阵. 对矩阵按照样本类属重新排序后可明显看出,密度敏感的相似性度量得到的相似性矩阵的块状效应更明显. 这说明使用该相似性度量可以缩小不同类数据点间相似度,同时增大同一类内数据点间相似度. 这一特性使得规范化的相似性矩阵 P 能够经过很少步的随机走动显示出明显的块结构. 也就是说,由密度敏感的相似性度量计算的相似性矩阵更接近于理想矩阵. 因此,相对于原有的谱聚类算法 DSSC 能够更好地识别数据本质的聚类结构.



(*a*) Gauss 核函数　　(*b*) 密度敏感的相似性度量

图 2　三个圆问题的相似性矩阵(按每类样本顺序重新排序得到,这一顺序对算法是未知的)

DSSC 算法的计算复杂度由求最短路的计算量所决定. 本文采用 Dijkstra 最短路算法[10],该算法的计算复杂度为 $O(n^3)$. 原有的谱聚类算法需要对整个核矩阵进行特征分解,这一运算的计算复杂度同样为 $O(n^3)$. 因此,DSSC 算法的计算复杂度与原有谱聚类算法的计算复杂度在同一个数量级上,仅相差一个常数因子.

4　实验

4.1　人工数据集

文献[11]给出一些"挑战性"问题. 从中挑选较为困

难的,并给出另一些困难数据集. 图 3 给出 DSSC 在八个数据集上的聚类结果. 对于所有问题,DSSC 都可成功识别聚类. 这里需要强调的是,DSSC 能够在一个很宽的参数范围内获得最优聚类,而 SC 只能在一个较小的参数范围内取得最优聚类,即 SC 对尺度参数比较敏感. 以两个圆问题为例,对于在区间 $[1, e^{29}]$ 上的任意伸缩因子 ρ,DSSC 都可识别聚类;而 SC 中的核参数 σ 必须在区间 $[0.1443, 1.6013]$ 中选择才能识别聚类,很小的扰动都会引起错分. 图 4 显示核参数 σ 对 SC 聚类结果的影响. DSSC 在一定程度上克服了谱聚类算法对尺度参数选择敏感的缺陷. 此外,SC 不能识别一些多尺度聚类问题如"blobs and circle",而 DSSC 却可成功识别. 这是由于 SC 所使用的相似性度量(高斯核函数)本身缺陷所致,该相似性度量不适合这类问题.

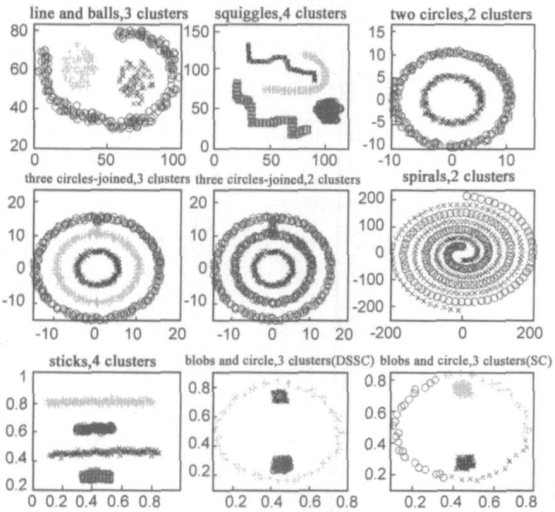


图 3 DSSC 成功识别八个人工"挑战性"聚类问题(类属用不同的符号和颜色标出),而 SC 算法无法识别 blobs and circle

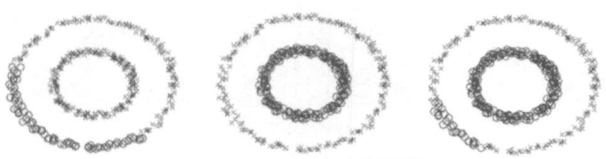


图 4 SC 算法对核参数选择的敏感性

4.2 手写体数据集

USPS 数据集由 16×16 维灰度图像构成,其中包含 7291 个训练样本,2007 个测试样本. 取全部测试样本作为聚类数据集,分别执行四组实验,包括识别两组较难的数字 0、8 和 3、5、8;两组较容易的数字 1、2、3、4 和 0、2、4、6、7. 四组实验分别在以下参数区间上执行:$[2^{-8}, 2^{-7.9}, \cdots, 2^{-5}]$,$[2^{-9}, 2^{-8.9}, \cdots, 2^{-5}]$,$[2^{-10}, 2^{-9.9}, \cdots, 2^{-5.5}]$,$[2^{-7}, 2^{-6.95}, \cdots, 2^{-5}]$. 为避免后处理步骤中 K 均值陷入局部最优,在每一个候选参数上运行 K 均值 100 次取最优结果. K 均值最大迭代步数设为 500,停止阈值设为 10^{-5}.

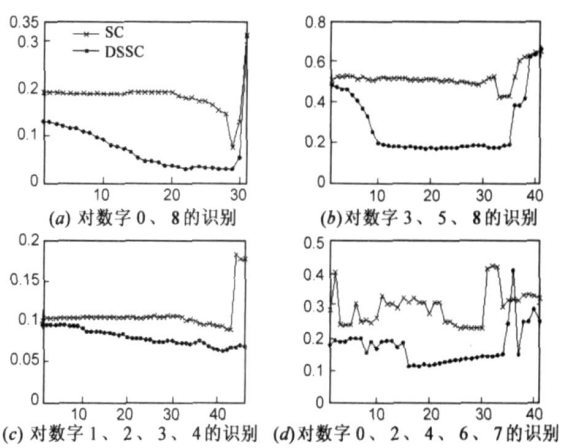
为比较不同算法的性能,我们利用了 USPS 数据集的类属信息(注意在真实聚类问题中类属信息不可获得). 假设已知聚类划分为 $\Delta^{true} = \{C_1^{true}, C_2^{true}, \cdots, C_{k_{true}}^{true}\}$,算法获得的聚类划分为 $\Delta = \{C_1, C_2, \cdots, C_k\}$. $\forall i \in [1, \cdots, k_{true}], j \in [1, \cdots, k]$,用 $Confusion(i, j)$ 表示已知聚类 C_i^{true} 和算法划分的聚类 C_j 之间相同的数据点个数,则聚类误差定义如下:

$$CE(\Delta, \Delta^{true}) = \frac{1}{n} \sum_{i=1}^{k_{true}} \sum_{\substack{j=1 \\ i \neq j}}^{k} Confusion(i, j) \qquad (4)$$

其中 n 为数据点个数. 这里存在一个重新编号的问题,例如实际聚类中的第 1 类有可能被算法指派成第 3 类。为了克服这一问题,需要对算法产生的聚类划分在所有可能的 $C_{k_{true}}^{k}$ 个序号组合上计算聚类误差,取其中最小值即为算法的聚类误差。

表 1 USPS 数据集上的最优聚类误差和平均运行时间

数字集	DSSC		SC	
	CE	Time(秒)	CE	Time(秒)
0,8	0.025	3.468	0.059	2.722
3,5,8	0.163	3.088	0.325	2.537
1,2,3,4	0.064	8.861	0.091	5.964
0,2,4,6,7	0.093	22.267	0.228	15.531



(a) 对数字 0、8 的识别 (b) 对数字 3、5、8 的识别

(c) 对数字 1、2、3、4 的识别 (d) 对数字 0、2、4、6、7 的识别

图 5 USPS 数据集上的性能对比. 横轴表示候选参数,纵轴表示聚类误差

表 1 给出算法在四个数据集上取得的最优聚类误差和对应参数段上的平均运行时间(实验执行在具有 1.86GHz 处理器,2G 内存和 Windows XP 操作系统的个人计算机上). DSSC 的性能均明显优于 SC,平均运行时间与 SC 是可比的. 图 5 给出算法在上述参数区间上的性能对比. 可以很明显地看出,DSSC 在所有数据集上的聚类误差都明显低于 SC,并在比较宽的参数范围内取得最优聚类,而 SC 仅在极少数参数上获得最优聚类. SC 对最优参数的选择极为敏感,本文算法克服了该问题. DSSC 在较为困难的数据集 0、8 和 3、5、8 上相对于

SC 具有明显的优势.这说明密度敏感的相似性度量能够很好地描绘数据间复杂的相似性关系.

在 0、2、4、6 上观察到两个算法均出现性能不稳定现象,SC 表现得尤为明显.这是由于谱聚类算法存在一个在特征空间对数据后处理的问题.目前采用简单的 K 均值后处理方法.由于 K 均值对初始聚类中心选择极为敏感,造成算法产生不稳定的聚类结果.观察到本文算法存在一个参数段,在该参数段内算法表现出良好的稳定性.这归功于参数段内计算的相似性矩阵更接近于理想矩阵.

5　结论

本文研究了数据聚类的两个一致性特征,提出一个新颖的相似性度量,得到密度敏感的谱聚类算法.该算法不仅能在较大的分析尺度范围内识别"挑战性"问题,而且可以识别多尺度问题,在 USPS 数据集上的性能明显优于原有谱聚类算法,且克服了对尺度参数敏感的缺陷,这表明该算法更适合实际问题.未来工作包括将密度敏感的相似性度量嵌入到更一般的聚类方法中,期望提高聚类性能.该思想已成功应用于 K 均值算法[12].

参考文献:

[1] 李洁,高新波,焦李成.基于特征加权的模糊聚类新算法[J].电子学报,2006,34(1):89 - 92.
Li Jie, Gao Xinbo, Jiao Licheng. A new feature weighted fuzzy clustering algorithm[J]. Acta Electronica Sinica, 2006, 34(1): 89 - 92. (in Chinese)

[2] Fiedler M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(98): 298 - 305.

[3] Shi J, Malik J. Normalized cuts and image segmentation[J]. IEEE Trans on PAMI, 2000, 22(8): 888 - 905.

[4] Zelnik-Manor L, Perona P. Self-tuning spectral clustering[A]. Advances in Neural Information Processing Systems (NIPS17)[C]. Cambridge, MA: MIT Press, 2005. 1601 - 1608.

[5] Zhou D, Bousquet O, Lal T N, et al. Learning with Local and Global Consistency[A]. Advances in Neural Information Processing Systems (NIPS16)[C]. Cambridge, MA: MIT Press, 2004. 321 - 328.

[6] Blum A, Chawla S. Learning from labeled and unlabeled data using graph mincuts[A]. Proceedings of the Eighteenth International Conference on Machine Learning (ICML18)[C]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2001. 19 - 26.

[7] Chapelle O, Zien A. Semi-supervised classification by low density separation[A]. Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics[C]. Barbados: Society for Artificial Intelligence and Statistics, 2005. 57 - 64.

[8] Meila M, Xu L. Multiway cuts and spectral clustering[R]. University of Washington, 2003.

[9] Meila M, Shi J. A random walks view of spectral segmentation[A]. Proceedings of International Workshop on AI and Statistics[C]. Florida, USA: Society for Artificial Intelligence and Statistics, 2001.

[10] Dijkstra E W. A note on two problems in connection with graphs[J]. Numerical Mathematics, 1959, 1: 269 - 271.

[11] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[A]. Advances in Neural Information Processing Systems (NIPS14)[C]. Cambridge, MA: MIT Press, 2002. 894 - 856.

[12] Wang L, Bo L F, Jiao L C. A modified K-Means Clustering with a density-sensitive distance matric[A]. Lecture Notes in Artificial Intelligent 4062 (RSKT'06)[C]. Heidelberg: Springer Press, 2006. 544 - 551.

作者简介:

王　玲　女,1978 年生于陕西西安,西安电子科技大学博士研究生.研究方向为模式识别、统计机器学习和图像处理.
E-mail: wliiip@163.com

薄列峰　男,1978 年生于陕西西安,西安电子科技大学博士研究生.研究方向为核机器学习、流形学习、神经网络和计算机视觉.
E-mail: blf0218@163.com

焦李成　男,1959 年生于陕西白水,工学博士,IEEE 高级会员,现为西安电子科技大学教授、博士生导师.主要从事非线性科学和智能信息处理等领域的研究.