基于监督信息特性的主动半监督谱聚类算法

王 娜,李 霞
(深圳大学信息工程学院,广东深圳 518060)

摘 要: 半监督聚类是利用少部分监督信息辅助大量未标签数据进行非监督的学习,其聚类性能的改善依赖于监督信息,因此挖掘适合半监督聚类的监督信息非常关键.提出了一种基于监督信息特性的主动学习策略,即找出同一类中距离相对较远的数据对象对和不同类中距离相对较近的数据对象对组成监督信息,并将其引入谱聚类算法,构建新颖的主动半监督谱聚类算法 ASSC(Active Semi-supervised Spectral Clustering).利用该监督信息调整谱聚类中点与点之间的距离矩阵,使类内各点紧聚,类间散布.通过对 UCI 基准数据集以及人工数据集的实验结果表明,ASSC 算法优于采用随机选取监督信息的谱聚类性能.

关键词: 谱聚类; 半监督聚类; 主动学习; 监督信息
中图分类号: TP311, TP187 **文献标识码:** A **文章编号:** 0372-2112 (2010) 01-0172-05

Active Semi-supervised Spectral Clustering Based on Pairwise Constraints

WANG Na, LI Xia
(*College of Information Engineering, Shenzhen University, Shenzhen, Guangdong* 518060, *China*)

Abstract: Semi-supervised clustering uses a small amount of supervised data such as pairwise constraints to aid unsupervised learning. The improved clustering performance depends heavily on the choice of constraints. This makes it important to explore the appropriate pairwise constraints for semi-supervised clustering. This paper presents a method for actively selecting informative pairwise constraints, which corresponds to pick up data pairs far apart in the same cluster and those close in different clusters. An active semi-supervised spectral clustering (ASSC) is then developed by utilizing the selected pairwise constraints to adjust the distance matrix in spectral clustering. As a result, the intra-cluster distance is decreased and the inter-cluster distance is increased. Experimental results on UCI benchmark data sets and artificial data set show that these informative pariwise constraints lead to substantial performance enhancement over the random selective pairwise constraints spectral clustering.

Key words: spectral clustering, semi-supervised clustering, active learning, pairwise constraints

1 引言

近年来,随着自然语言分析、网络与电信数据分析、Web 信息的有效获取、图像与视频数据分析等巨大应用驱动,聚类已成为模式识别、决策支持、机器学习、图像分割等领域中最重要的数据分析方式之一[1].传统的聚类算法主要是 K 均值算法、EM 算法以及在它们基础上的改进算法.这些算法都是建立在凸样本空间分布上,当样本空间不为凸时,算法会陷入"局部"最优.由于谱聚类具有识别非凸分布聚类的能力,能在任意形状的样本空间上聚类,且收敛于全局最优解,因此谱聚类是近来出现的一种性能极具竞争力的聚类方法[2,3].

聚类是通过抽取数据的"潜在"结构,根据相似性将数据样本分入不同的集合.聚类过程中通常没有类别标签信息,是一种无监督的学习.随着研究的深入,主观因素的重要性逐渐为人们所认识."对于不同的应用,其相应的聚类结果应该不同".因此如何把用户倾向结合入聚类过程成为近年研究的一个热点问题.用户倾向实际上就是背景知识,也称监督信息,利用监督信息来改善无监督聚类算法的性能,称之为半监督聚类.根据使用监督信息的方法不同,目前,半监督聚类算法大致分为三类:一类是基于限制的方法.该类算法通过修改聚类目标函数[4],或在聚类过程中遵循监督信息限制条件[5,6]等引导聚类过程向一个较好的数据划分进行;另一类是基于距离测度函数学习的方法.该类算法通过对监督信息学习,改变聚类算法中的距离测度函数,得到新的适合数据聚类的新度量[7~9],如利用监督信息基于图的最短路径方法而得到的欧式距离;利用监督信息基于凸的优化方法而得到的马氏距离等;第三类就是集成上述两种思想的聚类算法,如 Bilenko 等人提出的 MPCK

收稿日期:2008-07-21,修回日期:2009-07-20
基金项目:国家自然科学基金(No.60902069,60772148),广东省自然科学基金(No.9151806001000025);深圳市科技计划(No.200613);深圳大学科研项目(No.200632)

– means 算法等[10].

半监督聚类对聚类性能的改善依赖于监督信息,Davidson 等人已经证明寻找满足所有监督信息的聚类解是一个 NP 完备问题[11],监督信息越多,半监督聚类算法的复杂性越高,但聚类性能不一定更好,因此挖掘适合半监督聚类的监督信息非常关键. Basu 提出适用于 K 均值算法的 Farthest-first traversal 策略[12],该算法选择离当前集合最远的 k 个数据点来初始化 K 均值算法,提高了半监督聚类性能;Wagstaff 提出适用于类间边界不明显甚至类间有重叠的谱特征值策略[13],该策略只适用于两类聚类.在谱聚类与半监督结合方面,焦李成等提出了一种密度敏感的半监督谱聚类算法[14],该算法通过引入空间一致性先验信息构造密度敏感距离,将其应用到谱聚类算法中,提高了聚类性能.

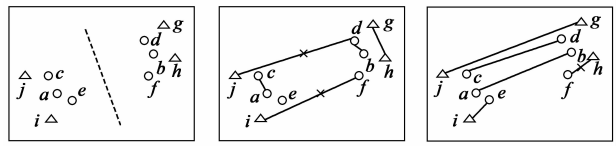
本文根据监督信息的信息含量提出一种主动学习策略,挖掘含有丰富聚类信息的监督信息,并将其应用到谱聚类算法中,调整点与点之间的距离矩阵,使类内各点紧密分布,类间距离尽量拉大,形成主动半监督谱聚类算法,提高聚类性能.

2　主动学习策略

在半监督聚类中,监督信息分为两种,一种是标记有类属信息的数据对象,称之为有标签数据,另一种是对聚类数据对象的一些限制,称之为成对约束监督信息:如某两个数据对象 P 和 Q 应该划归一类,标记为 ML(P, Q),或某两个数据对象应分属不同集合,用 CL(P, Q)表示.在实际应用中,如交谈中的说话人识别[15]、GPS 数据中的道路检测[7],获得数据对象的类属信息比较困难,数据对象之间的关系则较容易获取.此外,有标签数据根据类属信息很容易转换成成对约束监督信息,反之则不然.因此本文探讨的是对成对约束监督信息的主动挖掘.

在半监督聚类中,如果用户提供的监督信息信息含量较少,或提供的监督信息是聚类算法本身就能发现的,则这些监督信息难以对聚类算法起到积极的指导作用,聚类性能提高有限甚至可能下降[15].因此,在半监督聚类中要尽可能获得含有丰富信息量的监督信息,即要挖掘聚类算法本身不能发现的数据对象关系.图 1 示意性给出具有丰富信息量的监督信息的含义.设有两类数据对象 $\{a, b, \cdots, j\}$,分别用△,○表示,若采用经典的划分聚类算法来说,就是寻找 2 个中心,再将每个数据对象分入与其最近中心所代表的类中,为了使各数据对象与其所在类中心距离的平方和较小,很显然会得到图 1(a)所示的结果.若提供图 1(b)所示的监督信息,即 ML = {[g, h][d, b][a, c]},CL = {[d, j][i, f]}则不能改善聚类的性能,因为聚类算法本身就能发现这种监督信息.而如图 1(c)所示的监督信息 ML = {[d, c][g, j][a, b]},CL = {[h, f][i, e]},则含有较丰富的信息量,因为这些监督信息反映了聚类数据的结构,是划分聚类算法本身不能发现的.

本文依据信息量这一特性,提出了一种挖掘成对约束监督信息的主动学习策略.在基于划分的聚类过程中,距离远的数据对象会被认为不相似,从而分入不同类中.相反,距离近的数据对象会被认为具有较强的相似性,从而分入同一类中.因此本文提出基于丰富信息特性的主动学习,其基本思想是找出同一类中距离远的数据对象对,标记为 ML,不同类中距离近的数据对象对,标记为 CL.定义变量 P 为当前的监督信息数目,L 为当前集合 S 中的类别数目,具体策略如下:



(a) 高信息量的监督信息 (b) 低信息量的监督信息 (c) 划分聚类结果示例

图1　监督信息特性示例

算法 1　基于丰富监督信息特性的主动学习策略

(1)初始化监督信息限制数目 N 和聚类数目 K;初始化集合 $S = \{\lambda_i\}_{i=1}^{K}$ 为空;定义数据点 x 到集合 S 的距离 $d(x, S) = \min(d_{y \in S}(x, y))$;

(2)随机在数据集中选一点做为初始点,加入到子集合 λ_i 中;初始化 $L = 1$, $P = 0$;

(3)当 $P < N$,选择离集合 S 距离最远的点 x,即 $x \mid d(x, S) = \max_{i=1,\cdots,K}(\min(d_{y \in \lambda_i}(x, y)))$

(3.1)当 $L < K$ 时,询问 x 是否属于当前某子集合 $\{\lambda_i\}_{i=1}^{K}$,若属于,则随机抽取子集合 λ_i 中的一点 y,构建成对约束监督信息 ML(x, y),并将其加入子集合 λ_i, $P = P + 1$;否则,从每个子集合中各抽取一点 y_i,构建 L 个成对约束监督信息 CL(x, y_i),并创建新的子集合 λ_{L+1}, $P = P + L$, $L = L + 1$;

(3.2)当 $L = K$ 时,计算 x 到 K 个子集合中心的距离,并按该距离从小到大对子集合排序,依次询问 x 是否属于当前某子集合 $\{\lambda_i\}_{i=1}^{K_0}$,若不属于,随机抽取当前子集合中一点 y,构建成对约束监督信息 CL(x, y), $P = P + K$;直到询问到 x 属于某一个子集合 λ_{K_0+1},随机抽取 λ_{K_0+1} 集合中的一点 y,构建成对约束监督信息 ML(x, y), $P = P + 1$,并将其加入子集合 λ_{K_0+1} 中,返回(3);

(4)退出;

从主动学习策略中可以看出,当 $L < K$ 时,选择离当前集合 S 距离最远的点意味着利用较少的成对约束监督信息找出代表 K 个聚类类别的数据点.由于

$x \mid d(x,s) = \max_{i=1,\cdots,k}(\min(d_{y\in\lambda_i}(x,y)))$,这时挖掘的ML成对约束监督信息肯定是同一个子集合中距离远的两个点,CL是不同子集合中距离近的数据点.同样,$L=K$ 表明 K 个聚类已经形成,寻找离当前集合 S 距离最远的点意味着该点是对当前 K 个类最有奇异性的,即最难判定该点属于哪个子集合,利用 x 到 K 个子集合中心从小到大的距离排序,构建出的ML和CL成对约束监督信息保证了同一类中是距离远的数据对象对而不同类中是距离近的数据对象对,因此通过这种主动学习策略构建的监督信息是具有丰富信息量的.

3 基于主动学习策略的半监督谱聚类算法(ASSC)

在谱聚类中,属于同一类别的两个点之间的距离应该尽可能小,反之,二者之间的距离越大越好.本文提出的主动半监督谱聚类算法,就是利用主动学习获得的成对约束监督信息,来改变点与点之间的距离矩阵,由此使类内各点尽可能紧密分布,类间各点尽可能地彼此分离,从而最终获得好的聚类结果.具体算法如下:

算法 2 主动半监督谱聚类算法(ASSC)

(1)计算两点间欧式距离 $D_{ij}=(\|x_i-x_j\|^2)^{1/2}$;

(2)根据成对约束监督信息修改距离矩阵:

$$\begin{cases} D_{ij}, D_{ij}=0, \text{if}(x_i,x_j)\in ML \\ D_{ij}, D_{ij}=\infty, \text{if}(x_i,x_j)\in CL \end{cases}$$

(3)构造矩阵 $\boldsymbol{S}=1/D_{ij}$,若 $D_{ij}=0$,则另 $S_{ij}=\max(\boldsymbol{S})$,$S_{ii}=0$;

(4)构造矩阵 $\boldsymbol{P}=\boldsymbol{L}^{-1/2}\boldsymbol{S}\boldsymbol{L}^{1/2}$,其中 $\boldsymbol{L}$ 为对角矩阵 $L_{ii}=\sum_{j=1}^{n}S_{ij}$;

(5)求 $\boldsymbol{P}$ 的 $\boldsymbol{K}$ 最大特征值所对应的特征向量 $v_1, v_2,\cdots,v_K$,构造矩阵 $\boldsymbol{V}=[v_1,v_2,\cdots,v_K]$;

(6)规范化 $\boldsymbol{V}$ 的行向量,得到矩阵 $\boldsymbol{Y}$,其中 $Y_{ij}=\dfrac{V_{ij}}{\left(\sum_j V_{ij}^2\right)^{1/2}}$

(7)将 $\boldsymbol{Y}$ 的每一行看成是 R^K 空间内的一点,使用核模糊聚类KFCM将其聚为 K 类;

从以上算法可以看出,ASSC算法与传统谱聚类算法有三点不同:

(1)与谱聚类算法经常使用的Gauss核函数作为相似性度量不同,该相似性直接在欧式距离测度上计算.谱聚类算法性能对Gauss核中的尺度参数非常敏感,只有在尺度参数很小的区间内才能识别聚类,这样就造成了对于不同的聚类数据,参数的选择非常困难.直接采用欧式距离构造相似性关系避免了该问题,但同时也弱化了算法的非线性聚类能力.

(2)为了克服此缺点,对 $\boldsymbol{Y}$ 使用核模糊聚类KFCM聚类.对于不同的数据聚类问题,由于此时 $\boldsymbol{Y}$ 矩阵都是一个归一化矩阵,因此所有的核尺度参数都可以选择为1,这样既弥补了欧式距离度量非线性聚类能力差的弱点,又避免了参数的选择困难问题,后面的实验也有效验证了该算法.

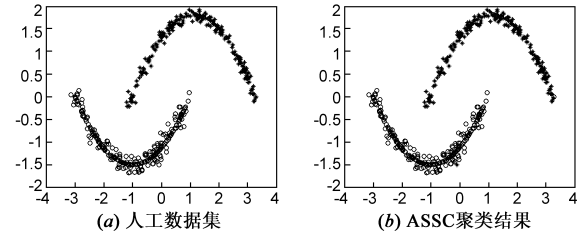
(3)通过直接修改距离矩阵的方法来施加成对约束监督信息.谱聚类算法聚类性能依赖于相似性矩阵,因此相似性矩阵应尽可能准确地反映数据间的相似性关系.通过主动学习得到的成对约束监督信息都是聚类数据相似性关系不明显,甚至通过相似性计算会出错的数据点对,因此通过强制修改距离矩阵而构造相似关系,使其随监督信息发生改变,从而更能发现数据内部固有的空间分布信息,发挥对聚类搜索的指导作用.

4 实验结果与分析

为了验证本文提出的主动半监督聚类算法,分别在一组取自真实世界的UCI基准数据Iris,Wine和Inosphere数据集和人工数据集上[17]进行仿真实验,并与基于随机选取成对约束监督信息的半监督谱聚类算法RSSC、核模糊聚类算法KFCM进行了对比研究.表1给出了UCI数据集的数据特征,人工数据集如图2(a)所示.在本文所有实验中,成对约束监督信息限数目介于0~200之间.对于以上三种算法,在同一实验条件下,采用100次实验的平均结果作为输出.

表 1 UCI数据集的数据特征

	Iris	Wine	Ionosphere
样本数	150	178	351
维数	4	14	34
类数	3	3	2



(*a*) 人工数据集 (*b*) ASSC聚类结果

图2 人工数据集实验结果

本文采用常用的聚类精度衡量不同算法的性能,假设已知聚类划分为 $C^{true}=\{c_1^{true},c_2^{true},\cdots,c_K^{true}\}$,算法获得的聚类划分为 $C=\{c_1,c_2,\cdots,c_K\}$,$\forall i,j\in[1,2,\cdots,K]$用 $Iden(i,j)$表示已知聚类 c_i^{true} 和算法划分聚类 c_j 之间相同的数据点个数,聚类精度(Ratio of correct categorized)定义为:

$$R_{cc}=\frac{1}{n}\sum_{i=1}^{K}\sum_{j=1,i=j}^{K}Iden(i,j)$$

其中 n 为数据点个数.KFCM是基于核方法的聚类算

法,这里采用 Gauss 核,不同的核参数会引起聚类性能很大波动,图 3 显示了核参数对 KFCM 聚类结果的影响,可见在 KFCM 中,对于不同数据,选取合适的核参数是很困难的.在与本文提出的算法比较时,采用的都是 KFCM 较好的聚类结果,即在三个 UCI 数据集上对应的核参数取值分别为 $Q=100,10,10$.在 ASSC 算法中,所有核参数均为 $Q=1$.图 2(b)绘出了 ASSC 算法对人工数据集的聚类结果.图 4(a ~ c)分别绘出了 KFCM、RSSC、ASSC 三种方法在三个 UCI 数据集上的聚类精度.表 2 给出了主动半监督谱聚类和随机半监督谱聚类在相同成对约束监督信息数目下的谱特征值.

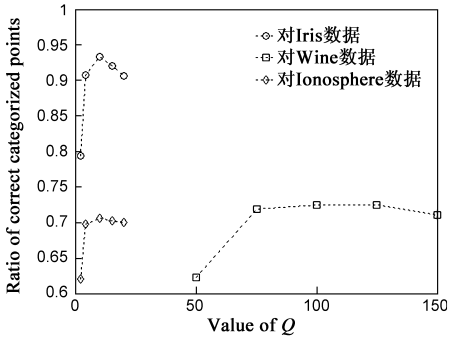


图3　Q值对聚类结果的影响

从图 4(a ~ c)可以看出,ASSC 在所有数据集上的聚类精度都明显高于随机 RSSC,表 2 也相应反映了这种聚类性能的提高.如在 200 对成对约束监督信息下,对于 Wine 数据来说,ASSC 的聚类精度是 80.40%,而 RSSC 是 72.82%.在表 2 中,ASSC 的三个特征值为(1,0.783,0.564),而 RSSC 的三个特征值分别为(1,0.756,0.589),可以看出 ASSC 中类内部点之间比 RSSC 分布地更紧密,并且 ASSC 中第二和第三特征值的差距也大于 RSSC,使得类间分离程度高于 RSSC,更利于后面的核聚类,因此 ASSC 聚类性能优于 RSSC.同时,从表 2 还可以观察到随着成对监督信息数目的增加,第二特征值会随之增加,且特征值间的差距也会加大,如在 Wine 数据中,对于 ASSC 算法在监督信息为 100 对时,特征值为(1,0.764,0.593),而监督信息增加到 200 对时,特征值变为(1,0.783,0.564),这就充分证明了成对约束监督信息有效影响了谱聚类性能,反映了图 4(a ~ c)中随着监督信息的增加,两种谱聚类性能曲线向上的形状.此外,ASSC 聚类算法只有在较少监督信息下,聚类性能低于 KFCM,这是由于较少的监督信息不能完全反映聚类结构,但随着监督信息限制数目的增加,聚类性能优于 KFCM,并且 ASSC 克服了 KFCM 核参数选取困难问题.

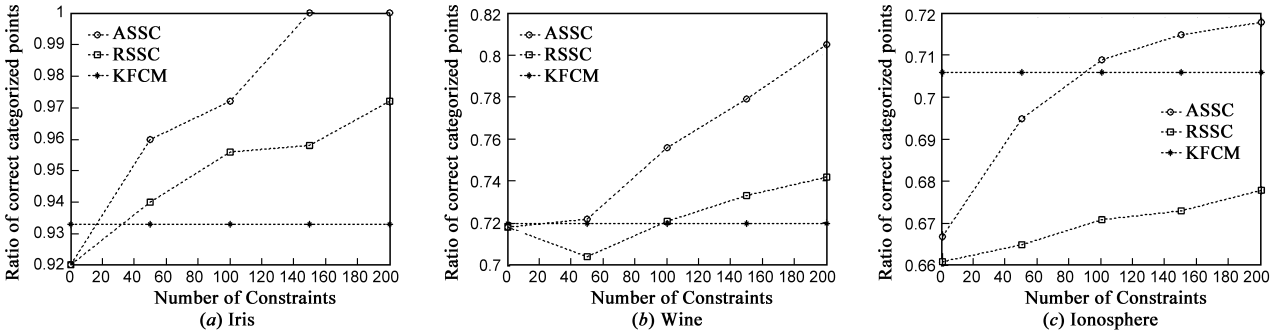


图4　三种算法在基准数据集上的对比实验结果

表 2　两种谱聚类算法的特征值比较

聚类算法	Wine		Iris		Ionosphere	
	200	100	200	100	200	100
ASSC	(1,0.783,0.564)	(1,0.764,0.593)	(1,0.712,0.409)	(1,0.697,0.393)	(1,0.3986)	(1,0.3935)
RSSC	(1,0.756,0.589)	(1,0.736,0.594)	(1,0.692,0.368)	(1,0.687,0.362)	(1,0.3634)	(1,0.3630)

5　结论

本文提出了一种基于监督信息特性的主动半监督谱聚类算法,通过主动学习找出同一类中距离相对较远的数据对象对,不同类中距离相对较近的数据对象对,分别组成监督信息 ML 和 CL,并利用此类含有丰富聚类信息的监督信息来调整谱聚类算法中点与点之间的距离矩阵,从而使得同一聚类内各点之间分布的更为紧密,而不同聚类之间彼此分开.基于被调整距离矩阵所形成的谱特征矩阵,进行核模糊聚类.实验结果表明该算法聚类性能优于基于随机选取监督信息的半监督谱聚类,且克服了核参数敏感的问题.未来工作包括两方面,一方面探索如何在较少的监督信息下扩展,以期使用更少的监督信息,就能获得很好的聚类结果;另一方面,探索正约束信息 ML 和负约束信息 CL 对于聚类性能的影响作用,以提高半监督聚类性能.

参考文献:

[1] 罗敏,王丽娜,张焕国.基于无监督聚类的入侵检测方法[J].电子学报,2003,31(11):1713 - 1716.

LUO Min ,WANG Li-na ,ZHANG Huan-guo. An unsupervised clustering Based intrusion detection method [J]. Acta Electronica Sinica,2003,31(11):1713 – 1716.(in Chinese)

[2] S Kamvar,D Klein,C Manning. Spectral learning[A]. In Proc. of IJCAI'03[C]. Mexico: Morgan Kaufmann Publishers, 2003.561 – 566.

[3] 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 Lie-feng ,J IAO Li-cheng. Density-sensitive spectral clustering[J]. Acta Electronica Sinica, 2007, 35(8): 1577 – 1581.(in Chinese)

[4] A Demiriz, K Bennett, M Embrechts. Semi-supervised clustering using genetic algorithms[A]. In Proc. of the Intelligent Engineering Systems Through Artificial Neural Networks [C]. New York: ASME Press, 1999. 809 – 814.

[5] K Wagstaff, C Cardie, S Rogers, S Schroedl. Constrained k-means clustering with background knowledge[A]. In Proc. of ICML'01 [C]. San Francisco: Morgan Kaufmann Publishers, 2001.577 – 584.

[6] S Basu, M Bilenko, R J Mooney. A probabilistic framework for semi-supervised clustering[A]. In Proc. of ICML'01[C], San Francisco: Morgan Kaufmann Publishers, 2001.577 – 584.

[7] E Xing, A Ng, M Jordon, S Russell. Distance metric learning with application to clustering with side-information[A]. Advances in Neural Information Processing System [C]. Cambridge : MIT Press, 2003.505 – 512.

[8] D Klein, S Kamvar, C Manning. From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering[A]. In Proc. of ICML'02[C]. Sydney: Morgan Kaufmann Publishers, 2002.307 – 314.

[9] N Wang, X Li. Kernel parameters optimization for semi-supervised fuzzy clustering with pairwise constraints[J]. Chinese Journal of Electronics, 2008, 17(2): 297 – 300.

[10] Bilenko M, Basu S, Mooney R. Integrating constraints and metric learning in semi-supervised clustering[A]. In Proc. of ICML'04[C]. Banff: ACM Press, 2004. 81 – 88.

[11] Davidson I, Ravi s. The complexity of Non-hierarchical clustering with instance and cluster level constraints[J]. Data Mining Knowledge Discovery, 2007, 14(1): 25 – 61.

[12] Basu S, Banerjee A. Active semi-supervision for pairwise constrained clustering[A]. In Proc. of the 4th SIAM International Conference on Data Mining[C]. SIAM, 2004. 333 – 344.

[13] Qianjun Xu, DesJardins M, Wagstaff K. Active constrained clustering by examining spectral eigenvectors[A]. In Proc. of Discovery Science [C]. Heidelberg: Springer, 2005. 294 – 307.

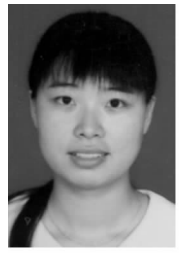
[14] 王玲,薄列峰,焦李成.密度敏感的半监督谱聚类[J].软件学报,2007,18(10):2412 – 2422.
WANG Ling ,BO Liefeng ,JIAO Licheng. Density-sensitive semi-supervised spectral clustering[J]. Journal of Software, 2007,18(10):2412 – 2422.(in Chinese)

[15] BarHillel A, Hertz T, Shental N, Weinshall D. Learning distance functions using equivalence relations[A]. In Proc. of ICML'03[C], Washington DC: AAAI Press, 2003. 11 – 18.

[16] Davidson I ,Wagstaff K, Basu S. Measuring constraint-set utility for partitional clustering algorithms[A]. In Proc. of PKDD[C], Heidelberg: Springer, 2006. 115 – 126.

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

作者简介:



王 娜 女,1977 年出生于河北保定.工学博士,深圳大学信息工程学院副教授,主要从事机器学习、图像处理和模式识别等方面的研究工作.

E-mail: wangna@szu.edu.cn

李 霞 女,1968 年出生于四川乐山.工学博士,深圳大学信息工程学院教授,博导,主要从事智能优化、图像处理和模式识别等方面的研究工作.