第 34 卷第 3 期 通 信 学 报 Vol.34 No. 3
2013 年 3 月 Journal on Communications March 2013
doi:10.3969/j.issn.1000-436x.2013.03.003

基于局部密度构造相似矩阵的谱聚类算法

吴健[1]，崔志明[1]，时玉杰[1]，盛胜利[2]，龚声蓉[1]
（1. 苏州大学 智能信息处理及应用研究所，江苏 苏州 215006；2. 美国阿肯色中央大学 计算机科学系，阿肯色州 康威 72035-0001）

摘 要：依据样本数据点分布的局部和全局一致性特征，提出了一种基于局部密度构造相似矩阵的谱聚类算法。首先通过分析样本数据点的分布特性给出了局部密度定义，根据样本点的局部密度对样本点集由密到疏排序，并按照设计的连接策略构建无向图；然后以 GN 算法思想为参考，给出了一种基于边介数的权值矩阵计算方法，经过数据转换得到谱聚类相似矩阵；最后通过第一个极大本征间隙出现的位置来确定类个数，并利用经典聚类方法对特征向量空间中的数据点进行聚类。通过人工仿真数据集和 UCI 数据集进行测试，实验结果表明本文谱聚类算法具有较好的顽健性。

关键词：谱聚类；相似矩阵；局部密度；无向图构建；边介数
中图分类号：TP391 **文献标识码：**A **文章编号：**1000-436X(2013)03-0014-09

Local density-based similarity matrix construction for spectral clustering

WU Jian[1], CUI Zhi-ming[1], SHI Yu-jie[1], SHENG Sheng-li[2], GONG Sheng-rong[1]
(1. The Institute of Intelligent Information Processin lication, Soochow University, Suzhou 215006, China;
2. Department of Computer Science, University of Central Arkansas, Conway 72035-0001, USA)

Abstract: According to local and global consistency characterist points' distribution, a spectral clustering algorithm using local density-based similarity matrix construction was proposed. Firstly, by analyzing distribution characteristics of sample data points, the definition of local density was given, sorting operation on sample point set from dense to sparse according to sample points' local density was did, and undirected graph in accordance with the designed connection strategy was constructed; then, on the basis of GN algorithm's thinking, a calculation method of weight matrix using edge betweenness was given, and similarity matrix of spectral clustering via data conversion was got; lastly, the class number by appearing position of the first eigengap maximum was determined, and the classification of sample point set in eigenvector space by means of classical cluster g method was realized. By means of artificial simulative data set and UCI data set to carry out the experimental tests, show that the proposed spectral algorithm has better clustering capability.

Key words: spectral clustering; similarity matrix; local density; undirected graph building; edge betweenness

1 引言

传统的聚类分析方法受限于非凸形状的样本空间，当样本空间不凸时，传统聚类算法会陷入局部最优。为了克服样本空间形状的限制，研究者提出了谱聚类(spectral clustering)算法[1,2]。该算法不仅能够在任意形状的样本空间上聚类，而且收敛于全局最优。与传统的聚类算法相比，它能很好地解决非块状和非凸形数据的聚类问题。谱聚类的这种优良特性在图像分割[3,4]和文档聚类[5,6]等领域得到了成功的应用。

谱聚类是一种基于相似矩阵的聚类算法，它对相似矩阵进行变换得到拉普拉斯矩阵，然后对其特征向量进行聚类[1,2,7,8]。所以，相似矩阵构造

收稿日期：2012-11-07；修回日期：2013-01-27
基金项目：国家自然科学基金资助项目（61003054, 61170020, 61170124）
Foundation Item: The National Natural Science Foundation of China (61003054, 61170020, 61170124)

的好坏是谱聚类算法优劣的重要因素。传统的谱聚类算法采用欧式距离来表示样本点之间的距离，并通过高斯核变换来计算样本点之间的相似度，其仅考虑到了局部一致性，没有考虑到全局一致性。王玲等人[9]提出密度敏感的相似性度量方法，该方法采用密度敏感的距离测度描述数据的实际聚类分布，它可以放大不同高密度区域内数据点间的距离，同时缩短同一高密度区域内数据点间的距离。相对传统的相似矩阵计算方法，该方法的定义较复杂且计算复杂度较高。孔万增等人[10]采用传统谱聚类中的方法构造相似矩阵，利用本征间隙自动确定数据的聚类个数，并利用确定的类数和谱分解的特征向量之间的余弦值完成数据的聚类。文献[1~10]中的谱聚类方法都没有充分利用样本点分布特性所隐含的先验信息，不能构造很好的相似矩阵。当其面临复杂样本数据点集时，无法得到理想的聚类结果。

为了构建更符合样本数据点分布特性的相似矩阵，本文提出了一种基于局部密度构造相似矩阵的谱聚类(LDSC, local density-based spectral clustering)算法。通过人工仿真数据集和 UCI 数据集进行测试，实验结果表明，本文算法得到的相似矩阵能更好地表示数据样本点之间的相似性，算法具有较好的顽健性。

2　LDSC 算法

2.1　算法思想

样本数据点集分布具有如下 2 个一致性特征[11]。

1) 局部一致性：指的是在空间位置上相邻的数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。

本文依据样本数据点分布的局部和全局一致性特征，提出了一种基于局部密度构造相似矩阵的谱聚类算法。算法首先给出局部密度定义，对样本数据点由密到疏排序，按序依次对样本点进行连接操作，完成无向图的构建。同时，借鉴 GN 算法思想[12]，采用边介数作为样本点对间的权值，从而计算出相似矩阵。然后，通过第一个极大本征间隙出现的位置来确定类个数，并利用经典聚类方法对特征向量空间中的数据点进行聚类。

2.2　无向图构建

2.2.1　现有构图方法分析

目前，用来构造无向图的方法有 e 阈值法、k 近邻法和互为 k 近邻法。e 阈值法虽然简便，但是由于样本点分布的多样性，e 的选取比较困难，很难选择一个合适的 e 以得到既连通又稀疏的图。较之更好且常用的是 k 近邻方法，k 容易选取且能得到一个稀疏图。但是 k 近邻法把每一个数据点看成同等重要的点，随机对点进行 k 近邻连线，不仅计算量大而且可能导致不同类数据点间的互连，从而将不同类数据点归为同类。互为 k 近邻方法[8]虽能保证数据点间互连的对称性，但其可能使同类样本之间也无法得到连通图。笔者以图 1 所示样本数据集为例进行分析。

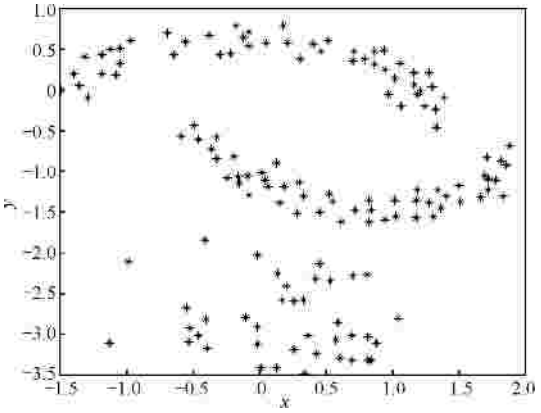
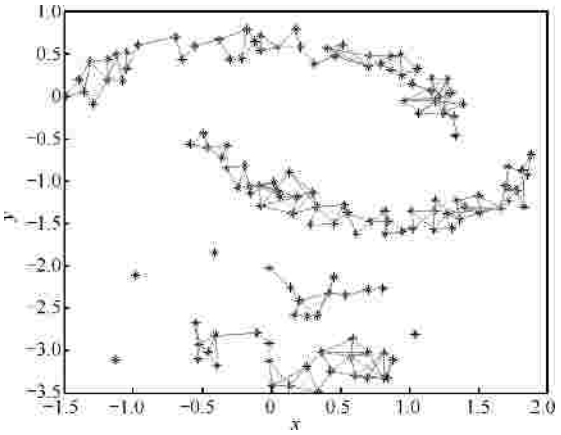


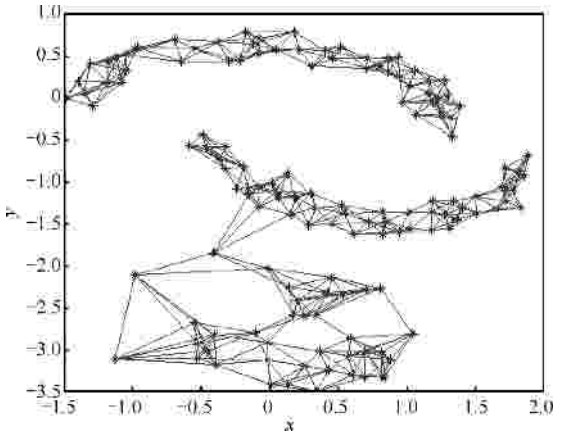
图 1　原始样本数据集

构建无向图，即根据一定的策略给样本空间中的数据点连线，最终得到样本数据集对应的无向图。图 1 为原始数据集，分为上月、下月和最下面的不规则分布 3 类。

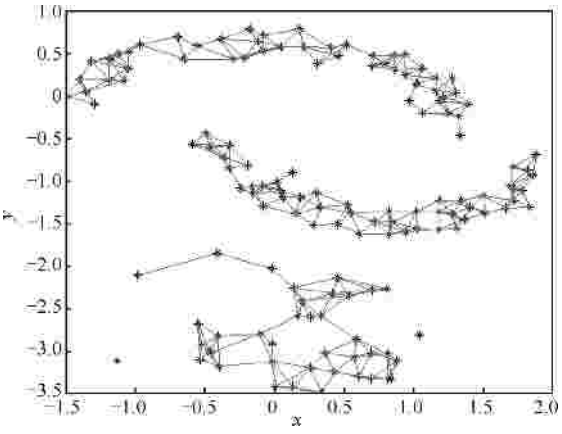
图 2 中给出了利用 3 种现有方法构建无向图的结果，从图 2 中可以看出现有无向图构建方法的局限性。图 2(a)的 e 阈值构图法中，阈值为 0.3，从图 2(a)中很容易发现其存在的问题：同类样本数据点间不连通，且存在较多孤立点。图 2(b)的 k 近邻方法虽然已形成连通图，但也存在明显的缺陷：不同类样本数据点互连，即图 2(b)中的下月型数据集和最下面的不规则分布数据集中有互连现象，如果采用该方法，则必然会对谱聚类结果产生较大的影响。图 2(c)的互为 k 近邻方法中，上月、下月和不规则分布间没有互连，但与 e 阈值构图法一样，同类样本之间不能构成连通图，则其谱聚类结果会将同类样本归为不同类。



(a) e 阈值 (e=0.3)



(b) k 近邻方法 (k=6)



(c) 互为 k 近邻方法 (k=6)

图 2 传统构图方法的构图结果

2.2.2 局部密度定义

针对上节 3 种方法中存在的问题，期望设计出一种新的构图方法，能够使得各类数据集有效分开且同类样本点连通，形成独立的连通子图。本文充分考虑样本点集的局部密度，首先对样本点按照局部密度进行排序，然后依照一定的连接策略完成无向图的构建。为了叙述方便，先给出局部密度的定义。

定义 1 如果存在一个点与其 k 近邻点的距离之和比其他样本点都小，则该点所处的局部区域在整个样本点集中越稠密。因此，笔者用样本点与其 k 近邻点距离之和的大小表示该点所处区域的稠密程度。记数据点 v_i 与其 k 近邻点的距离之和为 D_i，用 D_i 表示数据点 v_i 的局部密度，D_i 定义为

$$D_i = d_{i1} + d_{i2} + \cdots + d_{ij} + \cdots + d_{ik} \tag{1}$$

其中，$d_{i1} \le d_{i2} \le \cdots \le d_{ij} \le \cdots \le d_{ik}$，$d_{ij}$ 表示 v_j 和 v_i 之间的距离。

由定义 1 可知，D_i 越小，则表明 v_i 附近数据点越集中；D_i 越大，则表明 v_i 附近数据点越稀疏。

2.2.3 无向图构建步骤

采用局部密度思想，本文提出了基于局部密度的无向图构建方法，局部密度越大的数据点能够与其 k 邻近点相连的机会越大。本文提出的方法可以避免 e 阈值法、k 近邻方法和互为 k 近邻方法在构建无向图时存在的问题。

构图算法步骤如下。

Step1 求每一个点 V_i 到其 k 邻近点的距离之和 $D_i(i=1,\cdots,n)$。

Step2 对 $D_i(i=1,\cdots,n)$ 从小到大排序，选取最小的 $D_{(n)}$ 对应的点 V_i，并记 $num=1$，$D_{(n)}$ 定义如下（下标 n 表示未进行 k 邻近点连线操作的点的个数）

$$D_{(n)} = \arg\min\{D_i, i=1,2,\cdots,n\} \tag{2}$$

Step3 对 V_i 进行操作，即与 k 邻近点连线。对应邻接矩阵 $\boldsymbol{P}$ 中，初始 $\boldsymbol{P}_{\text{initial}}$ 中值都为 -1，如果点 V_i 和 V_j 相连，则置 $p_{ij}=1$ 且 $p_{ji}=1$；若两点不连，则置 $p_{ij}=0$ 且 $p_{ji}=0$；顶点自身一直为 -1。定义 $\boldsymbol{P}_{\text{initial}}$ 为

$$\boldsymbol{P}_{\text{initial}} = \begin{pmatrix} -1 & -1 & \cdots & -1 \\ -1 & & & -1 \\ \vdots & & \ddots & \vdots \\ -1 & -1 & \cdots & -1 \end{pmatrix} \tag{3}$$

则矩阵 $\boldsymbol{P}$ 可表示为

$$\boldsymbol{P} = \begin{pmatrix} -1 & p_{12} & \cdots & p_{1n} \\ p_{21} & & & p_{2n} \\ \vdots & & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & -1 \end{pmatrix} \tag{4}$$

Step4 选取剩余 $\{D_x, x=1,2,\cdots,n-num\}$ 中的

最小值 $D_{(n-num)}$（下标 $n-num$ 表示未进行 k 邻近点连线操作的点的个数），其对应点为 V_x，$D_{(n-num)}$ 定义为

$$D_{(n-num)} = \arg\min\{D_x, x = 1, 2, \text{L}, n-num\} \quad (5)$$

统计出 V_x 一行对应邻接矩阵中 1 的个数 m，然后将与 $k-m$ 个邻近点连线。可知 V_x 的 k 邻近点集 $\{V_{xl}, l = 1, 2, \text{L}, k\}$，$V_{xl}$ 表示距离点 V_x 第 l 近的点，初始化 $l = 1$，$count = 0$。

Step5　当 $l \le k$ 时，进行如下判断。

1) 如果 V_{xl} 已与 V_x 连接，$l = l+1$，重新执行 Step 5；如果 V_{xl} 的度已饱和（即度为 k），则点 V_x 不与点 V_{xl} 相连，即邻接矩阵中置 0，$l = l+1$；否则，连接两点，对应邻接矩阵中置 1，且 $count = count+1$，$l = l+1$。

2) 如果 $count > k-m$，转 Step 6，否则重新执行 Step 5。

Step6　当 $num < n$ 时，$num = num+1$，重复执行 Step 4 和 Step 5；否则，程序结束。

利用本文构图方法所得构图结果如图 3 所示。

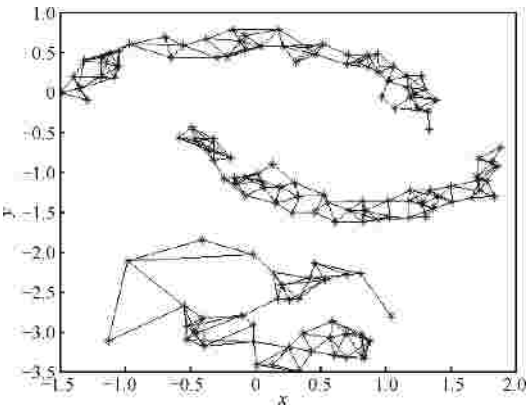


图 3　改进 KNN 构图结果(k=6)

从图 3 中可以看出，3 类样本集有效分开且内部连通，得到 3 个独立的连通子图，达到了预期目标。基于局部密度构图，可以保证空间位置上相邻的点互连，从而同属一类。Step 5 中的改进可以避免不同类数据点间互连。

2.3　相似矩阵构造

现实世界中的很多系统都以网络形式存在，具有同簇节点相互连接密集、异簇节点相互连接稀疏的特点。复杂网络聚类方法可归纳为 2 类：基于优化的方法和启发式方法[13]。基于图分割理论的谱聚类是一种基于优化的聚类方法。启发式方法将复杂网络聚类问题转化为预定义启发式规则的设计问题，被广为引用的 GN 算法的启发式规则是：簇间连接的边介数应大于簇内连接的边介数。边介数概念最早由 Girvan 和 Newman 提出，被用作评估复杂网络关键边的重要度指标。本文借鉴 GN 算法思想，提出了一种新颖的基于边介数度量的权值矩阵计算方法。

边介数[14]定义如下：图中任意两点的最短路径经过这条边的数目；如果不止一条最短路径，则在这些路径间等分边介数值。边 e 的边介数 B^e 计算式为

$$B^e = \sum_{u \ne v} b^e(u, v) \quad (6)$$

其中，$b^e(u,v) = \dfrac{s_{uv}(e)}{s_{uv}}$，$s_{uv}$ 表示点 u 到点 v 的最短路径的数目，$s_{uv}(e)$ 表示经过边 e 的点 u 到点 v 的最短路径数目。

样本数据点对应的无向图构建完成后，需要给图中的边赋予权值。若样本数据点有多类，则构建出的无向图中会有多个独立连通子图，子图内直接相连的点对的边介数可以直接求解，但是在子图内部和子图之间存在点对间无直接边相连的情况，求解这些点对间的“边介数”至关重要，即如何给这些样本点对赋予权值，需要作深入的研究。

通过分析样本数据集的分布特性，笔者研究得出样本点分布具有如下性质。

性质 1　点间传递相似性。

已知点 V_a 和点 V_b 具有较高的相似性，V_b 和 V_c 具有较高的相似性，则 V_a 和 V_c 也具有较高的相似性。如下式所示

$$S(V_a, V_b)、S(V_b, V_c) \rightarrow S(V_a, V_c) \quad (7)$$

性质 2　点间阻断性。

若不能直接也不能通过传递相似性得出两点相似，则两点不具有相似性。

结合边介数的定义和上述两条性质，给无向图中任意两点赋予权重，即可得到无向图对应的权值矩阵。以图 4 为例具体说明如何给任意两点赋权值。

任意点对间权重赋值的步骤如下。

1) 首先计算图中每一条边的边介数，这个边介数作为该边所连接的两点间的权值，即 $w_{ab} = B^{ab}$，w_{ab} 表示点 a 和点 b 的权值，显然 $w_{ab} = w_{ba}$。

2) 由性质 1 中的点间传递相似性可知，虽然点 a 和 c 没有边直接相连，但其却有较高的相似性。同理可知，点 a 和点 g 也具有较大的相似性。为解

决此类问题，本文给出如下定义。

$$w_{uv}=\frac{\sum B^{u\to v}}{sum(e_{u\to v})} \tag{8}$$

其中，w_{uv} 表示任意两点 u 和 v 之间的权值，$sum(e_{u\to v})$ 表示点 u 到点 v 的最短路径上边的条数，$\sum B^{u\to v}$ 表示点 u 到点 v 的最短路径上各边的边介数之和。经分析可知，有边直接相连的两点间的权值亦可通过这个方法计算，用式(8)作为独立连通子图内任意两点间权值的计算公式。显然，$w_{uv}=w_{vu}$。

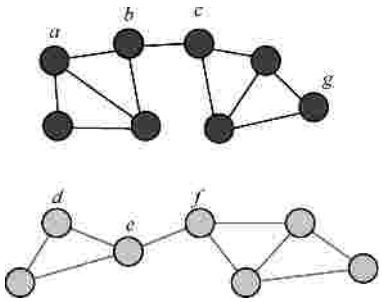


图 4 权值计算示例

3) 前两点解决的是独立连通子图内任意两点间权重的赋值问题，不能应用于独立子图间的样本数据点，如图 4 中的点 a 和点 d，因为这两点间不存在路径。但由性质 2 中的点间阻断性可知，点 a 和点 d 不具有相似性，据此作如下规定，若两点间无路径，则其权值为一较大的正数值 M（或为无穷大）。此规定亦适用于无向图中的孤立点。

经过上述 3 个步骤，可以计算得出数据样本集中任意点对间的权值，从而得到了无向图对应的权值矩阵。权值矩阵中 2 个数据点之间的数值越小，表示 2 个数据点越相似。谱聚类算法在聚类过程中需要对权值矩阵经过一定的变换得到相似矩阵。相似矩阵中的数值越大表示 2 个数据点越相似，属于同一类的可能性越大。反之，2 个点属于同一类的可能性越小。这与权值矩阵中的数值表示含义相反。本文采用倒数的方式计算相似矩阵，由权值矩阵到相似矩阵的计算公式为

$$S_{ij}=\frac{1}{W_{ij}+1}, S_{ij}=S_{ji} \tag{9}$$

如果在权值矩阵中 2 个点之间的权值是无穷大，则在相似矩阵中就设置为 0。由于权值矩阵中主对角线上的数值为 0，则在相似矩阵中主对角线上的数值应该为 1。

2.4 算法实现

本文针对谱聚类算法中的相似矩阵构造进行了研究。首先利用上节中给出的方法对样本数据点集进行无向图构建，然后利用边介数思想计算权值矩阵，经过数据变换得到相似矩阵，最后利用经典聚类方法对特征向量空间中的数据点进行聚类。具体步骤如下。

输入：n 个数据点 $x_i(i=1,\text{L},n)$。

输出：数据点集的划分结果 C_1,L,C_k。

1) 利用本文提出的改进 KNN 算法对输入的数据点构造相似图 G。

2) 对构造的相似图 G 进行边介数计算，按照 2.3 节所述方法求取权值矩阵 $\boldsymbol{W}$，并采用式(9)计算相似矩阵 $\boldsymbol{S}$。

3) 构造 Laplacian 矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$，其中，$\boldsymbol{D}$ 为对角度矩阵 $D_{ii}=\sum_{j=1}^{n}S_{ij}$。

4) 计算矩阵 $\boldsymbol{L}$ 的特征值，并对特征值进行从大到小的排序，找到第一个极大本征间隙出现的位置，记为 k，k 即为聚类类别数。

5) 计算 k 个最大特征值对应的特征向量 v_1,v_2,L,v_k，构造矩阵 $\boldsymbol{V}=[v_1,v_2,\text{L},v_k]$，对矩阵 $\boldsymbol{V}$ 中的每一行进行单位化处理，得到矩阵 $\boldsymbol{Y}$，即

$$Y_{ij}=\frac{V_{ij}}{\left(\sum_j V_{ij}^2\right)^{\frac{1}{2}}} \tag{10}$$

6) 把矩阵 $\boldsymbol{Y}$ 的每一行看成 k 维空间中的点，利用传统的聚类算法（如 K-means 算法）将其聚成 k 类。

7) 如果 $\boldsymbol{Y}$ 的第 i 行属于第 j 类，则将原数据点 x_i 也划分到第 j 类。算法结束。

3 实验结果与分析

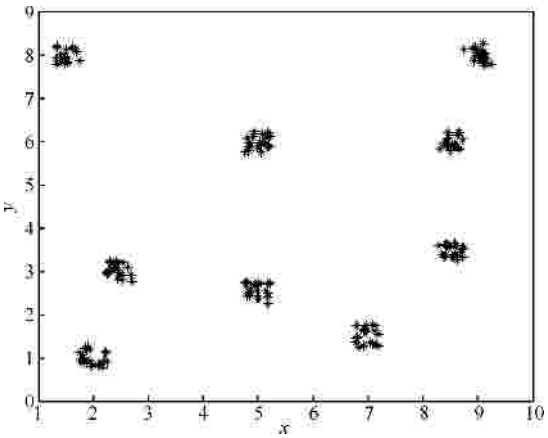
为了验证本文谱聚类算法的分类性能，本文进行了 2 组实验：采用人工仿真数据集，分为理想分类数据集和非线性数据集，重点测试聚类算法的一般性和特殊性；选用具有真实数据含义的 UCI 数据集测试聚类算法的实际应用效果。实验程序采用 Dell PC 机上的 MATLAB 2010a 实现，机器配置如下：Pentium(R) Dual-Core CPU E5300@2.60 GHz，

4GB 内存，Windows 7 操作系统。

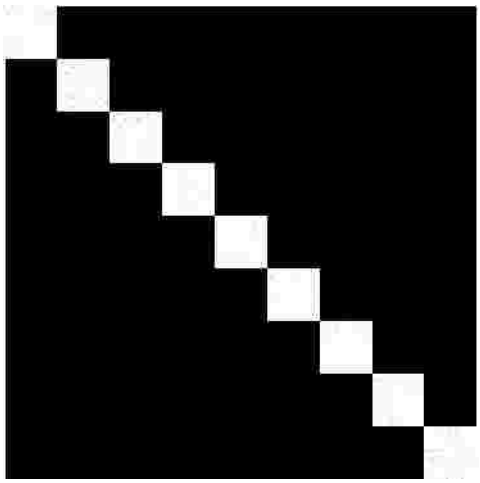
实验 1　人工数据集测试

1) 理想分类数据实验

为了验证本文算法的一般性，首先在理想分类数据集上进行实验，理想数据集由计算机随机产生，共 180 个样本点，分为 9 类，如图 5(a)所示。为了使实验结果更直观、明显，本文在实验时对数据按类顺序排列，由于样本数据点按照类顺序进行排列，因此它对应的相似矩阵 S 在主对角线上呈现出 9 个明显的分块，如图 5(b)所示。



(a) 样本点分布



(b) 相似性矩阵

图 5　人造 9 类数据

文献[15]指出，当谱聚类算法的相似性矩阵是块对角矩阵时，该算法可以找到完全正确的聚类结果。由此可知，本文谱聚类算法完全能够解决一般性的问题。本文谱聚类算法对图 5(a)所示的 9 类数据样本集有很好的分类性能，聚类的结果如图 6 所示。

2) 非线性数据实验

通过上一节对理想分类数据的实验分析，验证了本文谱聚类算法的有效性。在本节中，笔者将对一些复杂数据进行聚类实验分析。传统聚类算法，如 K-means 算法、FCM 算法，基于欧式距离来描述样本数据点之间的相似性。但是，样本数据点之间的欧式距离较小并不意味着它们即属于同一类，如图 7 所示的数据样本集。由于数据集是交叉的月牙型分布，不是块状分布，如采用传统聚类算法对其进行聚类分析，其聚类效果会很不理想。

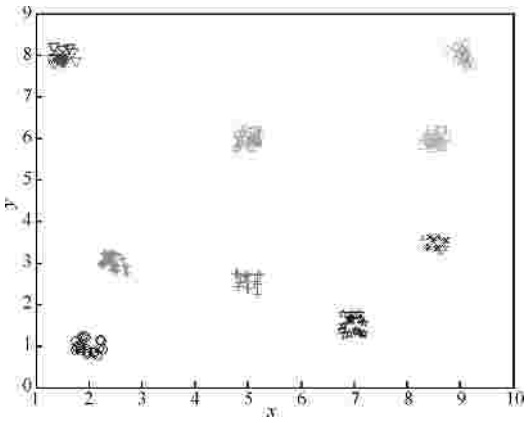


图 6　本文谱聚类分类结果

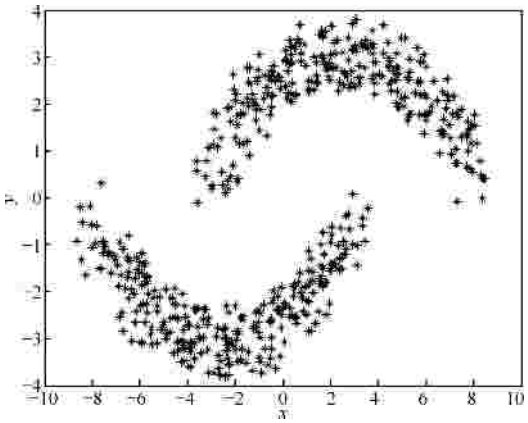
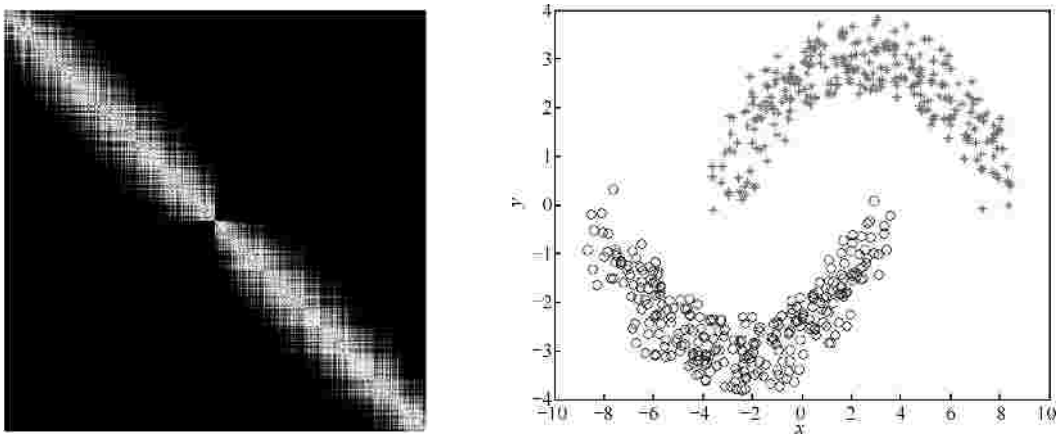


图 7　双月牙数据

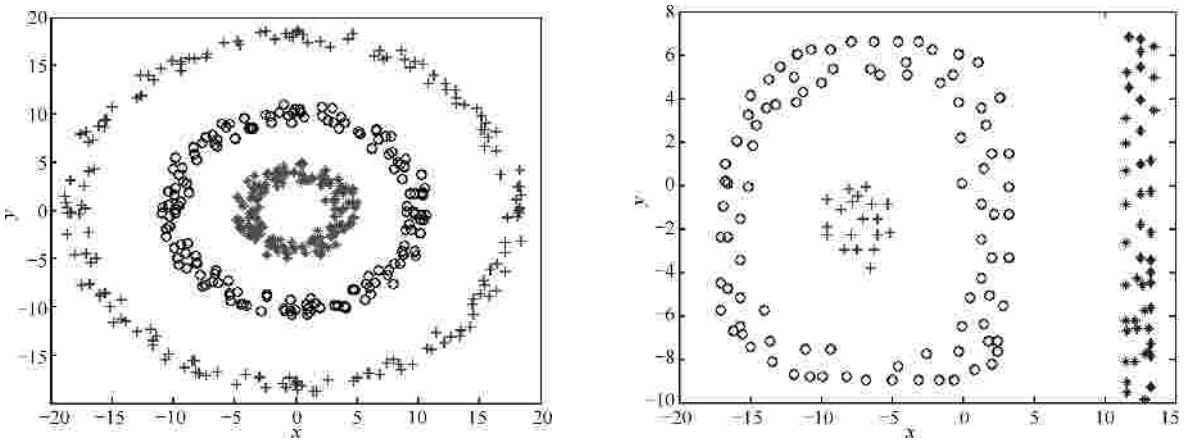
采用本文谱聚类算法，在类顺序排列的相似矩阵上呈现出一条狭长的对角块，如图 8(a)所示。在双月型的数据集中属于同一类的样本点在物理位置上有可能比较远，样本点之间的最短路径需要经过很多条边相连，引起相似矩阵中同类样本点间的相似度有所差异，所以该图并不是严格意义上的块对角分布，但还是可以从图 8(a)中清晰地看出样本数据点集分成了 2 类。双月型数据聚类结果准确无误，聚类结果如图 8(b)所示。

笔者从一些挑战性问题中选择了 3 个较为困难的数据集。本文以文献[10]提出的 ASC 算法作为比较算法。图 9(a)~图 9(c)是分别采用本文算法在这 3 个数据集上的聚类结果，对于这 3 个问题，本文算法可以成功得到聚类结果。图 10 是 ASC 算法在这 3 个数据集上的聚类结果，从图 10 中可以看出，图 10(b)和图 10(c)发生了聚类错误，不能对线球形和波浪线数据集进行正确聚类。

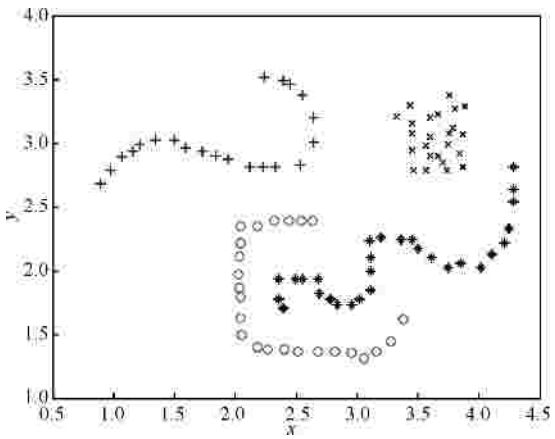


(a) 相似性矩阵　　(b) 聚类结果

图 8　本文谱聚类算法聚类分析

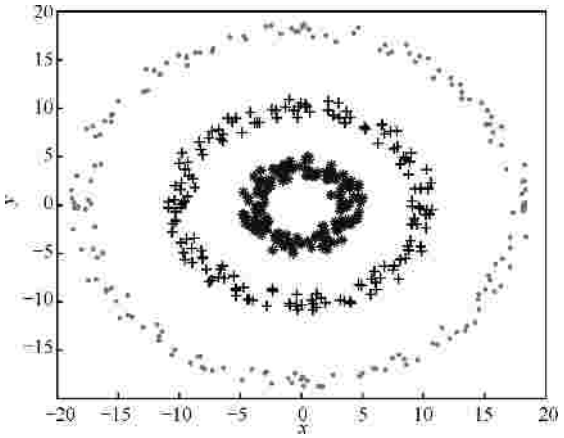


(a) 同心圆，3 类　　(b) 线球形，3 类

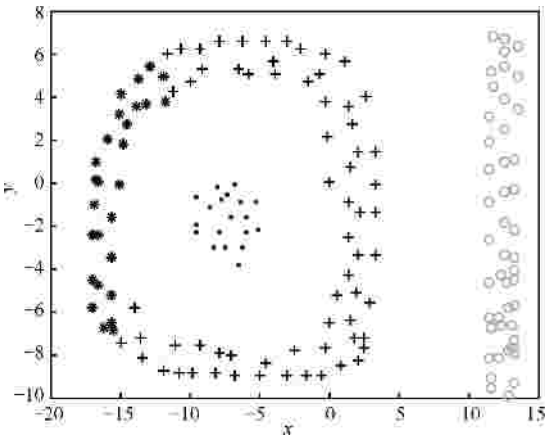


(c) 波浪线，4 类

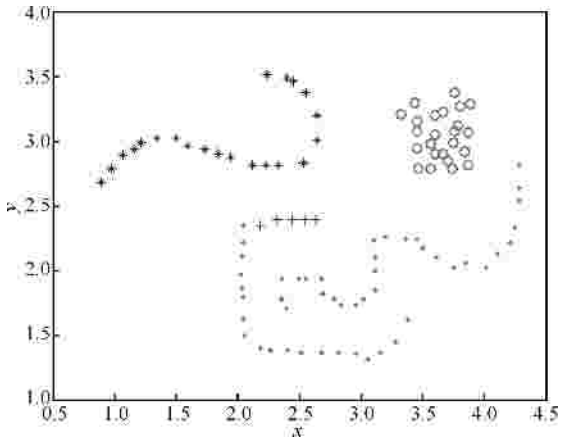
图 9　本文算法针对 3 种人工挑战性问题的聚类结果



(a) 同心圆，3 类



(b) 线球形，3 类



(c) 波浪线，4 类

图 10　文献[10]算法针对 3 种人工挑战性问题的聚类结果

实验 2　UCI 数据集测试

实验采用的数据集选自国际通用数据库 UCI 基准数据集中的 Satimage 数据集、Iris 数据集、Ionosphere 数据集以及 New-thyroid 数据集进行本文算法和 ASC 算法的实验比较。具体真实数据集的信息如表 1 所示，显示了 UCI 4 个实验数据集的基本属性。

表 1　UCI 数据集基本信息

样本数据集	维数	样本数	固有类数
Satimage	36	444	6
Iris	4	150	3
Ionosphere	34	351	2
New-thyroid	5	215	3

为了比较不同算法的性能，笔者使用的评价指标是聚类正确率[9]和时间开销。假设已知聚类划分为 $D^{\text{true}}=\{C_1^{\text{true}},C_2^{\text{true}},\L,C_{k_{\text{true}}}^{\text{true}}\}$，算法获得的聚类划分为 $D=\{C_1,C_2,\L,C_k\}$。$\forall i\in[1,\L,k_{\text{true}}]$，$j\in[1,\L,k]$，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，则聚类错误率定义为

$$CE(D,D^{\text{true}})=\frac{1}{n}\sum_{i=1}^{k_{\text{true}}}\sum_{\substack{j=1\\ i\neq j}}^{k}Confusion(i,j) \qquad (11)$$

其中，n 为数据点的个数。聚类正确率的定义很容易根据式(11)得到。

$$CT(D,D^{\text{true}})=1-CE(D,D^{\text{true}}) \qquad (12)$$

表 2 是 2 种算法在每个数据集上的最优聚类误差和平均运行时间结果。本文算法中的参数是 k，而 ASC 算法中的主要参数是 s，2 种算法中笔者都选择最优的聚类结果进行实验比较分析。文献[10]中给出了 Iris、Ionoshpere 和 New-thyroid 数据集在得到最优结果时的 s 值及其分类准确率，这里笔者使用同样的参数值进行实验。由于 Satimage 数据集中数据较多，文献[10]在 6 类中随机选取 444 个样本数据进行实验，本文也随机选择 444 个样本数据，但由于数据不尽相同，所以这里选取的最优 s 参数和文献[10]中给出的有所不同。

表 2　UCI 数据集上的最优聚类误差和平均运行时间

样本数据集	算法					
	LDSC			ASC		
	CT/%	时间/s	K	CT/%	时间/s	S
Satimage	80.41	2.386 8	5	77.93	2.315 3	0.07
Iris	93.33	0.214 9	8	92.00	0.140 5	0.16
Ionosphere	90.88	1.092 6	8	72.08	1.042 4	0.20
New-thyroid	89.76	0.387 3	5	89.30	0.323 5	0.13

有研究表明，大多数聚类算法仅在少量数据形成的低维特征空间中拥有较好的聚类结果[16]。从表 2 中可以看出，由于 Satimage 和 Ionoshpere 数据集的维数

分别为 36 和 34，ASC 算法在这 2 个数据集上的聚类正确率较低，而在维数较低的 Iris 和 New-thyroid 数据集上的聚类正确率相对较高。本文算法由于能够更好地表示数据样本点之间的相似性，在这 4 个数据集上的聚类效果均比 ASC 算法要好，尤其是维数较高的 Ionoshpere 数据集的聚类比 ASC 算法更优。

在平均运行时间上，本文算法要比 ASC 算法运行时间略长，原因在于本文算法包含求解最短路径的环节，而 ASC 算法没有这个过程。LDSC 算法的计算复杂度由求最短路径的计算量所决定，本文采用 Floyd 最短路径算法，该算法的计算复杂度为 $O(n^3)$。因此，LDSC 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数量级上。

4 结束语

相似矩阵构造是谱聚类中的瓶颈问题。本文提出了一种基于局部密度构造相似矩阵的谱聚类算法，算法首先对样本点按照局部密度由密到疏进行排序，并按照设计的连接策略构建无向图；然后基于边介数构造无向图的权值矩阵，从而得到相似矩阵；最后利用经典聚类方法对特征向量空间中的数据点进行聚类。实验结果表明，本文算法得到的相似矩阵能更好地表示数据样本点之间的相似性，算法具有较好的顽健性。但是，谱聚类算法存在着面对高维数据集时聚类正确率降低的共同问题，下一步研究重心将放在提高高维数据集聚类正确率上。

参考文献：

[1] CRISTIANINI N, SHAWE-TAYLOR J, KANDOLA J. Spectral kernel methods for clustering[A]. Proceedings of the 13th Advances in Neural Information Processing Systems(NIPS 2001)[C]. Vancouver, British Columbia, Canada, 2001. 649-655.

[2] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm[A]. Proceedings of the 14th Advances in Neural Information Processing Systems(NIPS 2002)[C]. Vancouver, British Columbia, Canada, 2002. 849-856.

[3] 邓晓政, 焦李成, 卢山. 基于非负矩阵分解的谱聚类集成SAR图像分割[J]. 电子学报, 2011, 39(12):2905-2909.
DENG X Z, JIAO L C, LU S. Spectral clustering ensemble applied to SAR image segmentation using nonnegative matrix factorization[J]. Acta Electronica Sinica, 2011, 39(12):2905-2909.

[4] DUCOURNAU A, BRETTO A, RITAL S, *et al*. A reductive approach to hypergraph clustering: an application to image segmentation[J]. Pattern Recognition, 2012, 45(7):2788-2803.

[5] 徐森, 卢志茂, 顾国昌. 使用谱聚类算法解决文本聚类集成问题[J]. 通信学报, 2010, 31(6):58-66.
XU S, LU Z M, GU G C. Spectral clustering algorithms for document cluster ensemble problem[J]. Journal on Communications, 2010, 31(6): 58-66.

[6] ZHANG T P, TANG Y Y, FANG B, *et al*. Document clustering in correlation similarity measure space[J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(6):1002-1013.

[7] LUXBURG U V. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4):395-416.

[8] 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 J Y, ZHOU J G, GUAN J H, *et al*. A survey of clustering algorithms based on spectra of graphs[J]. CAAI Transactions on Intelligent Systems, 2011, 6(5):405-414.

[9] 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 L, BO L F, JIAO L C. Density-sensitive spectral clustering[J]. Acta Electronica Sinica, 2007, 35(8):1577-1581.

[10] 孔万增, 孙志海, 杨灿等. 基于本征间隙与正交特征向量的自动谱聚类[J]. 电子学报, 2010, 38(8):1880-1891.
KONG W Z, SUN Z H, YANG C, *et al*. Automatic spectral clustering based on eigengap and orthogonal eigenvector[J]. Acta Electronica Sinica, 2010, 38(8):1880-1891.

[11] ZHOU D, BOUSQUET O, LAL T N, *et al*. Learning with local and global consistency[A]. Proceedings of the 16th Advances in Neural Information Processing Systems(NIPS 2004)[C]. Vancouver, British Columbia, Canada, 2004. 321-328.

[12] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. The National Academy of Science, 2002, 9(12):7821-7826.

[13] 杨博, 刘大有, 刘际明等. 复杂网络聚类方法[J]. 软件学报, 2009, 20(1):54-66.
YANG B, LIU D Y, LIU J M, *et al*. Complex network clustering algorithms[J]. Journal of Software, 2009, 20(1):54-66.

[14] PINNEY J W, WESTHEAD D R. Betweenness-based Decomposition Methods for Social and Biological Networks[M]. Interdisciplinary Statistics and Bioinformatics, Leeds University Press, 2007.

[15] MEILA M, XU L. Multiway Cuts and Spectral Clustering[R]. University of Washington, 2003.

[16] 刘铭, 王晓龙, 刘远超. 基于语义的高维数据聚类技术[J]. 电子学报, 2009, 37(5):925-929.
LIU M, WANG X L, LIU Y C. Clustering technology for high dimensional data based on semantics[J]. Acta Electronica Sinica, 2009, 37(5):925-929.

作者简介：

吴健（1979-），男，江苏南通人，博士，苏州大学讲师，主要研究方向为图像与视频处理、模式识别和图像检索。

崔志明（1961-），男，上海人，苏州大学教授、博士生导师，主要研究方向为智能信息处理和数据挖掘。

时玉杰（1988-），女，河南周口人，苏州大学硕士生，主要研究方向为图像处理和模式识别。

盛胜利（1969-），男，安徽马鞍山人，美国阿肯色中央大学博士、助理教授，主要研究方向为数据挖掘和机器学习。

龚声蓉（1966-），男，湖北天门人，苏州大学教授、博士生导师，主要研究方向为图像与视频处理、模式识别和智能信息处理。