**文章编号:** 1001-0920 (2014) 03-0396-07 **DOI:** 10.13195/j.kzyjc.2012.1783

# 球极平面逆投影迭代谱聚类算法

何 力, 曲仕茹

(西北工业大学 自动化学院，西安 710129)

**摘 要:** 提出一种相似矩阵迭代修正并聚类算法, 分为偏振定理的谱分离数据和球极平面逆投影的几何分离数据两步. 首先将数据谱分解, 得到低维距离矩阵; 然后投影到双随机矩阵, 隐式进行一次球极平面逆投影, 几何对称分离数据; 最后解算投影后坐标, 得到新相似矩阵. 实验在人工合成数据和自然数据上进行, 结果表明所提出算法修正了数据的相似度, 并获得了正确的聚类个数, 对尺度参数变化有较强的鲁棒性, 聚类性能比修正前有较大提升.

**关键词:** 球极平面逆投影；谱聚类；偏振定理；模式识别

**中图分类号:** TP391.4 **文献标志码:** A

# Iterative spectral clustering by inverse stereographic projection

*HE Li*, *QU Shi-ru*

(School of Automation，Northwestern Polytechnical University，Xi'an 710129，China. Correspondent：HE Li, E-mail：heli9903@mail.nwpu.edu.cn)

**Abstract:** A method of iterative spectral clustering based on the inverse stereographic projection is proposed. The proposed method contains two steps for data clustering, one by the polarization theorem as the spectral clustering, the other by the inverse stereographic projection. Firstly, the affinity matrix of input data is eigen-decomposed, leading to the embedding of data in a low-dimensional space. The Euclidean distance matrix of the embedded data is then projected to its nearest doubly stochastic matrix. This approach is shown as a critical step to implicitly call the inverse stereographic projection that maps data into a hyper sphere. The last step is to solve the center and the scale factor of the hyper sphere. Experiments on the challenging synthetic data and the Iris and Wine data sets demonstrate the successful use of the proposed method in modifying the affinity matrix, and the modified affinity matrix can obtain better clustering results than the original one.

**Key words:** inverse stereographic projection；spectral clustering；polarization theorem；pattern recognition

## 0 引 言

聚类算法是模式识别领域的研究热点. 基于数据凸球形分布假设的传统聚类方法, 如 $k$-均值算法、EM 算法等, 当凸性难以保证时, 这些算法可能落入局部最优, 导致聚类效果下降. 为此, 学术界不断提出更多的聚类算法, 其中, 以规范割 (NCut)[1]为代表的谱聚类算法, 被深入研究并广泛应用于数据挖掘[2]、文本识别[3]等. 谱聚类算法的性能对于聚类个数和核方程参数 (如常见的高斯核的方差) 较为敏感, 这要求用户拥有对数据更多的先验知识. 越来越多的研究致力于赋以算法更强的自适应能力, 减少对用户的依赖.

谱聚类算法的理论基础是图谱分割理论, 通过特征分解获得数据类别划分的全局最优解. Dhillon 等[4]证明了谱聚类与核 $k$-均值聚类代价函数的统一性, 并将谱聚类问题利用核 $k$-均值算法加以求解, 解决了大数据计算问题. Alzate 等[5]证明了多种谱聚类算法可以在加权核主成分分析框架下描述, 并采用最小二乘支持向量机解决多类划分和 out-of-sample 问题. 这类算法要求类别个数为已知, 但实际中往往难以获得对该值的先验知识, 影响算法的适用范围. 另一方面, 对聚类使用的特征向量进行选择或修正也是研究热点. Xiang 等[6]研究了相似矩阵的单一特征向量对数据类别划分的影响, 以数据相关性为指标, 选择区分能力最优的特征向量, 其寻优过程通过期望最大化算法对模型迭代运算实现, 该方法建立在混合高斯模型上, 对于特征向量的非高斯性有待研究. 在文献 [6] 的基

**收稿日期:** 2012-11-29；**修回日期:** 2013-03-04.

**基金项目:** 教育部博士点基金项目(20096102110027)；航天科技创新基金项目(CASC201104)；航空科学基金项目(2012ZC53043).

**作者简介:** 何力(1983−), 男, 博士生, 从事谱聚类、目标检测的研究；曲仕茹(1963−), 女, 教授, 博士生导师, 从事模式识别、目标检测与跟踪、智能交通等研究.

础上, Zhao 等[7]采用熵指标取代相关性度量重要性, 选取最优的特征向量聚类数据. Rebagliati 等[8]进一步通过选择多于类别数的特征向量, 实现有冗余的特征向量聚类. 除了迭代选取特征向量的方法, 还可以迭代修正相似矩阵, 实现谱聚类. Lin 等[9]采用数学上快速求解矩阵主特征值的思想, 将相似矩阵幂次迭代后, 寻找若干主特征向量, 并利用其线性组合聚类. Brand 等[10]提出了利用偏振定理迭代谱分解数据. 这类谱聚类算法对原始输入的相似矩阵比较敏感[11], 对于变化的参数 (如高斯核参数 $\sigma$), 聚类的稳定性有待提高.

本文构造了一种新的迭代谱聚类算法, 以增强聚类算法对数据相似矩阵的鲁棒性, 并给出了该算法的几何解释. 具体而言, 迭代分为 3 步, 首先根据偏振定理, 采用数据相似矩阵的低秩近似代替原矩阵, 获得数据的低维嵌入, 即传统的谱分离方法; 然后计算低维数据的距离矩阵, 并映射到与其最接近的双随机矩阵, 该过程实际上是对数据进行一次受几何对称条件约束的球极平面逆投影, 再次分离数据; 最后, 解算数据所在超球的中心和半径参数, 将双随机阵转化为相似阵, 供下次迭代使用. 通过实验在合成数据集上验证了所提出方法寻找正确聚类个数的能力, 并通过多组 $\sigma$ 输入参数实验, 检验了算法对 $\sigma$ 的鲁棒性. 在 Iris 数据集和 Wine 数据集上与多种聚类算法进行比较, 验证了所提出算法的性能.

本文所提出算法继承文献 [10] 的框架, 也采用谱分离数据并投影到双随机矩阵的思路 (与基于随机游走解释的投影到随机矩阵的方法有所区别). 与文献 [10] 采用对角线归零的相似矩阵投影不同, 本文利用数据的欧氏距离矩阵计算双随机矩阵, 并解算投影后超球上数据的坐标, 从而使谱分离和几何分离操作能够循环迭代. 本文的主要贡献在于指出上述投影过程隐式地将低维数据按球极平面逆投影原则映射到高维超球上, 并利用该性质对数据按几何对称原则进行分离.

## 1　球极平面逆投影迭代谱聚类算法流程

首先给出所提算法的流程, 详细解释将在后面逐渐展开. 与一般习惯不同, 本文涉及到的欧氏距离矩阵 $E$ 为距离平方矩阵 (在不产生歧义的前提下, 仍简称为距离矩阵), 元素 $e_{ij}=||P_i-P_j||^2$. 其中: $P$ 为欧氏空间中数据的坐标. 相似矩阵 $A$ 采用高斯核, 其元素 $a_{ij}=\exp(-||P_i-P_j||^2/\sigma^2)$.

算法的输入为数据坐标 $P$, 输出为修正后近似理想的相似矩阵. 具体步骤如下.

Step 1: 计算 $P$ 的欧氏距离矩阵 $E$.

Step 2: 将距离矩阵 $E$ 核映射到相似矩阵 $A$, 核函数选高斯核, 尺度参数为 $\sigma$ (初次迭代后使用 $\sigma_S$).

Step 3: 对相似矩阵 $A$ 作特征分解, $A=V\Lambda V^{\mathrm{T}}$, 取特征向量矩阵前 $d$ 行 ($d$ 随迭代严格递减) 作为数据低维坐标 $X_d$, 计算 $X_d$ 的距离矩阵 $E_d$.

Step 4: 将 $E_d$ 投影到其最近的双随机矩阵 $B_d$.

Step 5: 由 $B_d$ 解算数据所在超球的球心 $z$ 和半径 $r$, 经球极平面逆投影, 得到数据在超球上的坐标 $X_{d+1}$, 进而得到其距离矩阵.

Step 6: 重复 Step 2 ∼ Step 5, 直至 Step 2 中相似矩阵的谱间隙大于阈值或 $d<1$.

算法大致可分为两部分, 即应用偏振定理对数据进行谱分离 (Step 2 和 Step 3) 和利用球极平面逆投影对数据几何分离 (Step 4 和 Step 5).

## 2　谱分离数据

利用图谱理论进行数据分离并聚类是谱聚类算法的基础. 本文算法中, 首先对相似矩阵 $A$ 特征分解, 即 $A=V\Lambda V^{\mathrm{T}}$, 其中特征值所在对角矩阵 $\Lambda$ 按降序排列. 记 $X\triangleq\Lambda^{1/2}V^{\mathrm{T}}$, 并记 $X$ 前 $d$ 行为 $X_d$, 当以 $X_d$ 重构相似矩阵, 即 $A_d\triangleq X_d^{\mathrm{T}}X_d$ 时, $A_d$ 是原相似阵 $A$ 最优的秩 $d$ 近似 ($F$ 范数意义下). 在文献 [10] 中, "谱分解——截断——重构矩阵"的过程已深入研究, 并证明, 随着 $d$ 不断减小, $X_d$ 中两两列向量夹角的余弦平方之和为严格递增 (偏振定理). 直观而言, 随着 $d$ 减小, 原本纠缠在一起的数据 (夹角 $\pi/2$ 附近, 余弦接近 0) 经以上步骤分离或聚拢 (夹角 $\in\{0,\pi\}$, 余弦接近 1), 从而实现聚类、分离数据. 另一方面, 在单位圆上, 由于有 $\mathrm{corr}(x_i,x_j)=x_i^{\mathrm{T}}x_j=\cos\theta_{ij}$, 即数据相关性与向量夹角余弦值相等, 余弦平方和的增加事实上也意味着总相关性的增加. 以上过程实际上是对数据进行一次多维尺度映射 (MDS)[12], $X_d$ 便是原始数据在低维空间中的最优嵌入. 这种利用谱分离对数据进行划分的方法在很多领域得以应用[1,4-10], 本文也将此作为首步数据分离操作, 并将 $X_d$ 作为数据在 $d$ 维空间的坐标.

根据矩阵摄动理论, 谱分离得到的低秩相似矩阵 $A_d$ 可用来判断数据聚类是否显著. 由 Davis-Kahan 定理[13]可知, 对于理想数据, 即同类数据相似度为 1, 异类为 0 的数据, 记其归一化拉普拉斯矩阵为 $L$, 扰动后的矩阵为 $\tilde{L}=L+H$, $H$ 为扰动阵. 记 $L$ 的特征值 $\gamma_1\leqslant\gamma_2\leqslant\cdots\leqslant\gamma_n$($n$ 为数据个数) 对应特征向量 $u_i(i=1,2,\cdots,n)$, $\tilde{L}$ 的特征值 $\tilde{\gamma}_1\leqslant\tilde{\gamma}_2\leqslant\cdots\leqslant\tilde{\gamma}_n$ 对应特征向量 $\tilde{u_j}(j=1,2,\cdots,n)$. 对于指定正数 $l$, 记谱间隙 $\delta_l=\gamma_{l+1}-\gamma_l$, 由 $u_1,u_2,\cdots,u_l$ 张成的空间为 $U=\mathrm{span}\{u_1,u_2,\cdots,u_l\}$, $\tilde{u}_1,\tilde{u}_2,\cdots,\tilde{u}_l$ 张成的空间为 $\tilde{U}$, 则 $U$ 与 $\tilde{U}$ 的距离上限为

$$d(U,\tilde{U}) \leqslant ||H||/\delta_l, \tag{1}$$

其中 $||\cdot||$ 为 $F$ 范数. 式 (1) 表明, 谱间隙 $\delta_l$ 越大, 扰动后特征向量张成的子空间与理想数据的子空间的距离上限越小, 更易于聚类. 本文使用 $A_d$ 的归一化拉普拉斯矩阵 $L_{A_d}$ 计算谱间隙. 将 $L_{A_d}$ 的特征值升序排列, 检查相邻两特征值之差, 若接近 1 则认为数据聚类显著, 停止迭代. 由于 $L_{A_d}$ 的特征值在 $[0,1]$ 内, 谱间隙接近 1 表明间隙两边特征值分别接近 0 或 1, 与理想数据类似. 小于 2 的类别数对于聚类而言无意义, 因此本文的谱间隙由 $\delta_2 = \gamma_3 - \gamma_2$ 开始, 并不计算 $\delta_1$.

## 3 几何分离数据

对于 $d$ 维空间中数据 $X_d$ 的距离矩阵 $E_d$, 采用文献 [14] 的快速迭代方法, 将 $E_d$ 投影到其最近的双随机矩阵 $B_d$, 即 $B_d = C_d E_d C_d$, 其中 $C_d$ 为对角阵. 本节将说明, 该过程实际上是隐式地进行一次球极平面逆投影, 从而进一步分离数据.

### 3.1 球极平面逆投影分离数据

球极平面投影是地理学常用的投影方法之一, 该投影从球北极点出发, 引一条过球面点 $\alpha$ 的直线, 并与某个与赤道平行的平面相交于点 $A$, 则点 $A$ 是点 $\alpha$ 的球极平面投影, 相反的过程称为球极平面逆投影.

球极平面投影与球面上点的欧氏距离矩阵及其对应的双随机矩阵有密切关系[15]. 根据文献 [15] 的定理 7, 对于 $d$ 维空间中若干点 $X_d$ 的欧氏距离矩阵 $E_d$, 记其最近的双随机矩阵为 $B_d$, 则在 $d+1$ 维空间中存在一个球 $S_{d+1}(z,r)$ (球心 $[z,r]$, $z$ 为 $d\times 1$ 向量, $r$ 为半径, 标量), 记 $X_d$ 在这个球上的球极平面逆投影点 $X_{d+1}$ 的距离矩阵为 $E_{d+1}$, 同时满足: 1) $E_{d+1}$ 与 $B_d$ 相差一个标量因子, 即 $E_{d+1} = kB_d$, $k$ 为标量; 2) 球上逆投影点 $X_{d+1}$ 的质心为球心, 即 $[z,r] = \overline{X}_{d+1}$.

上述定理中, 球心为球面逆投影点质心这一性质, 可视为对数据进行一次几何分离. 对于类别数为 2 且元素个数相同的理想数据, 球面上逆投影点对称分布于球心两侧, 类间夹角最大 ($\pi$), 类内最小 (0); 含噪的数据, 夹角会因噪音而扰动; 数据个数不同的类, 夹角会在理想情况的基础上相应扩散, 但整体的分离性仍得保存. 球面逆投影后, 两类数据间的夹角较大, 易于分割. 可见, 投影到双随机矩阵的过程依据几何对称要求分离数据, 本文利用该性质, 在谱分离数据后, 再次进行数据的几何分离.

由于超球参数 $[z,r]$ 未知, $B_d = C_d E_d C_d$ 的运算可以视为隐式地应用逆投影. 为了形成循环迭代, 需要将距离矩阵转为相似矩阵, 即需要解算超球的参数.

### 3.2 超球参数求解

双随机矩阵 $B_d$ 是球面上点的距离矩阵 (相差一个因子), 需要将其转化为相似矩阵, 从而降维并谱分离数据. 一个直观的思路是直接将距离矩阵转化为相似矩阵, 即相似矩阵元素 $a_{ij} = \exp(-b_{ij}/\sigma_S^2)$. 其中: $\sigma_S$ 为某个尺度参数, $b_{ij}$ 为 $B_d$ 中对应元素. 但是由于 $b_{ij} = k^{-1}||X_{d+1}(i) - X_{d+1}(j)||^2$, 其中 $X_{d+1}(i)$ 为第 $i$ 个数据在 $d+1$ 维空间中的坐标, 有

$$a_{ij} = [\exp(-||X_{d+1}(i) - X_{d+1}(j)||^2/\sigma_S^2)]^{k^{-1}},$$

即相似矩阵元素 $a_{ij}$ 被指数缩放了常数因子 $k^{-1}$, 造成失真. 为此, 本文在得到 $B_d = C_d E_d C_d$ 后, 利用文献 [15] 的定理 4、定理 6 和引理 3, 进一步解算超球参数 $[z,r]$ 和常数因子 $k$, 以得到数据在球面上真实坐标 $X_{d+1}$.

在 $d+1$ 空间的 $x_{d+1} = 0$ 平面上, 首先将坐标原点平移到 $z$, 使超球的球心为 $[0,r]$ (此时 $z$ 和 $r$ 仍未知). 由文献 [15] 的定理 6 易得

$$||X_d(i) - z||^2 = \frac{4r^2}{k^{1/2}C_{ii}} - 4r^2, \tag{2}$$

其中 $C_{ii}$ 为 $C$ 对角线上第 $i$ 个元素. 令 $r_B = (1/2n)^{1/2}$, 数据个数 $n$ 已知, 由文献 [15] 的引理 3 和定理 4 有 $r = k^{1/2} r_B$, 代入式 (2) 得到

$$||X_d(i) - z||^2 = \frac{4k^{1/2}r_B^2}{C_{ii}} - 4kr_B^2. \tag{3}$$

式 (3) 为多元四次方程组, 含有 $n$ 个方程和 $d+1$ 个未知量 ($z$ 的 $d$ 个未知量和标量 $k$, 一般 $n \gg d$). 显然, 方程组的精确解不易获得, 因此采用 dogleg 信赖域方法求数值解, 其核心是采用 dogleg[16] 沿分段线性路径搜索近似解. 得到 $[z,r]$ 和 $k$ 后, 通过球极平面逆投影可立即得到 $X_{d+1}$, 计算 $X_{d+1}$ 的距离矩阵, 重复本文算法, 直至满足条件退出.

本文再次使用高斯核将 $X_{d+1}$ 的距离矩阵转化为相似矩阵. 与原始输入数据不同, 此时的数据分布于超球上, 可以利用这一先验知识构造高斯核参数 $\sigma_S$. 参数由超球半径给出, $\sigma_S = \pi r/t$, 其中 $t$ 是人为指定的常数因子. 通过简单计算可得, 若希望第 $i$ 个数据与第 $j$ 个数据之间相似度为 $a_{ij}$, 则其夹角相应为 $\theta_{ij} = \arccos(1 + \pi^2 \log a_{ij}/2t^2)$. 此时, 相似度仅与夹角有关, 从而避免了估计尺度参数, 降低了用户对数据缩放尺度先验知识的要求.

## 4 实验分析

为了验证本文算法的有效性, 在多组数据上进行实验. 首先在人工合成数据上运行本文算法, 给定初始相似矩阵后, 实验将验证所提出算法是否终止于正确的类别个数和聚类效果. 为了验证本文算法对高斯核 $\sigma$ 参数的敏感性, 在一组不同 $\sigma$ 上检验聚类结果, 最后在自然数据上验证聚类效果, 并与其他几种聚类方法进行比较. 需要指出的是, 聚类算法大多对

高斯核 $\sigma$ 参数选取敏感[11], 本节不予讨论. 较为常用的 $\sigma$ 参数选取方法有两种: 1) 指定一个全局通用的常量 $\sigma$, 这需要用户对数据有较强的先验知识; 2) 文献[11]的方法, 记数据 $i$ 到其第 $K$ 近邻的距离为 $\sigma_i$, 则数据 $i$ 与数据 $j$ 互相的高斯核参数为 $\sigma_{ij}^2 = \sigma_i\sigma_j$, 额外的时间开销和 $K$ 值的选取是这类方法的问题. 本文实验采用全局通用 $\sigma$, 其值由用户给出, 高斯核参数 $\sigma_S$ 由 $t$ 指定, 实验中取 $t = 4$, 这是一个相对宽松的设定, 此时相似度大于 0.9 的数据夹角在 14.6°(角度制) 以内.

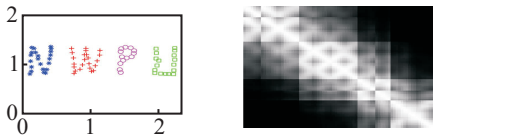
谱间隙阈值是本文算法的关键参数, 虽然理论上该值为 1, 但考虑到实际数据的扰动, 其取值往往远小于 1. 一般而言, 谱间隙达到 0.3 时, 相似矩阵的聚类性比较显著, 若不加说明, 则实验所取的值为 0.3. 算法另一个退出条件由 $d$ 控制, $d$ 指定了数据所在维数, 因此最小取 1, 而谱分离要求 $d$ 每次迭代严格递减. 若 $d = 1$ 时谱间隙仍未达到退出阈值, 则算法结束, 且没有找到符合条件的聚类. 一般而言, 可以通过改变高斯核 $\sigma$ 参数的选取, 再次迭代寻找聚类. $d$ 的初值并无严格要求, 用户可估计类别个数上限并上浮若干作为初值, 若不加说明, 实验取 $d = 6$ 为初值.

传统的 NCut 采用 $k$-均值法聚类特征向量, 初始中心选择和空类处理对结果影响较大. 为了直观比较相似矩阵优劣, 本文采用文献 [5] 的方法对修正后的相似矩阵进行聚类. 输入为迭代后相似矩阵 $A_d$ 和算法得到的类别个数 $l$, 输出为聚类结果. 对于 $A_d$ 的归一化拉普拉斯矩阵 $L_{A_d}$, 首先二值化 $L_{A_d}$ 最小的 $l$ 个特征向量, 并以特征向量的每一行表示一个数据; 然后选取所有数据中出现次数最多的 $l$ 个 ($l$ 为类别个数) 二值组合为各类中心, 其他数据, 按照到各中心的 Hamming 距离最近原则, 归属各类. 直观而言, 该方法是变形的最近邻法, 出现次数最多保证了所在类的代表性, 而二值化大幅减少了数据可能的取值 (共 $2^l$ 种), 并使数据间距的度量离散化.

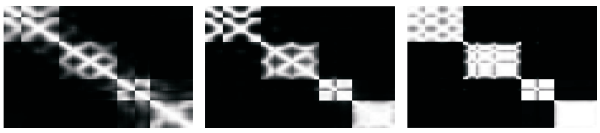
### 4.1　迭代聚类实验

本实验在一组人工合成数据集上检验所提出算法的性能. 图 1 显示了一组数据实验过程中, 相似矩阵迭代变化的情况. 图 1(a) 为输入数据, 不同色点对应不同类别, 且数据按照类别顺序排列 (该数据理想的相似矩阵为对角分块矩阵, 块内为 1, 块间为 0). 图 1(b) 为 $\sigma = 0.5$ 时原始相似矩阵. 图 1(c) ~图 1(e) 显示了迭代过程中 $d = 6, 5, 4$ 下的相似矩阵, 当 $d = 4$ 时, 谱间隙为 0.579, 达到退出条件. 可以看到, 经本文算法迭代后相似矩阵的聚类性比原始矩阵有较大提高.

图 2 为原始相似矩阵、NCut 在原始相似矩阵聚类结果、迭代终止时相似矩阵和本文算法聚类的结

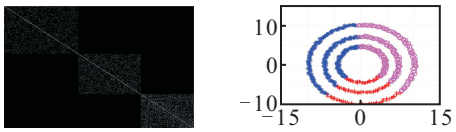


(a) 输入数据　(b) $\sigma$=0.5相似矩阵

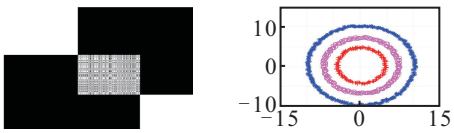


(c) 秩 6 相似矩阵　(d) 秩 5 相似矩阵　(e) 秩 4 相似矩阵

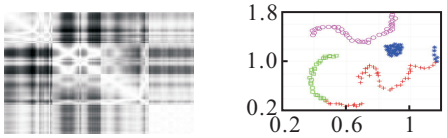
**图 1　相似矩阵迭代变化**



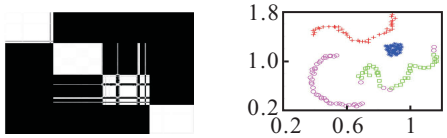
(a) 原始相似矩阵 1　(b) 原始分割结果 1



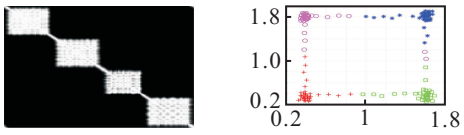
(c) 修正后相似矩阵 1　(d) 修正后分割结果 1



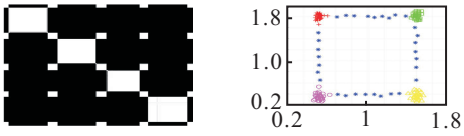
(e) 原始相似矩阵 2　(f) 原始分割结果 2



(g) 修正后相似矩阵 2　(h) 修正后分割结果 2



(i) 原始相似矩阵 3　(j) 原始分割结果 3



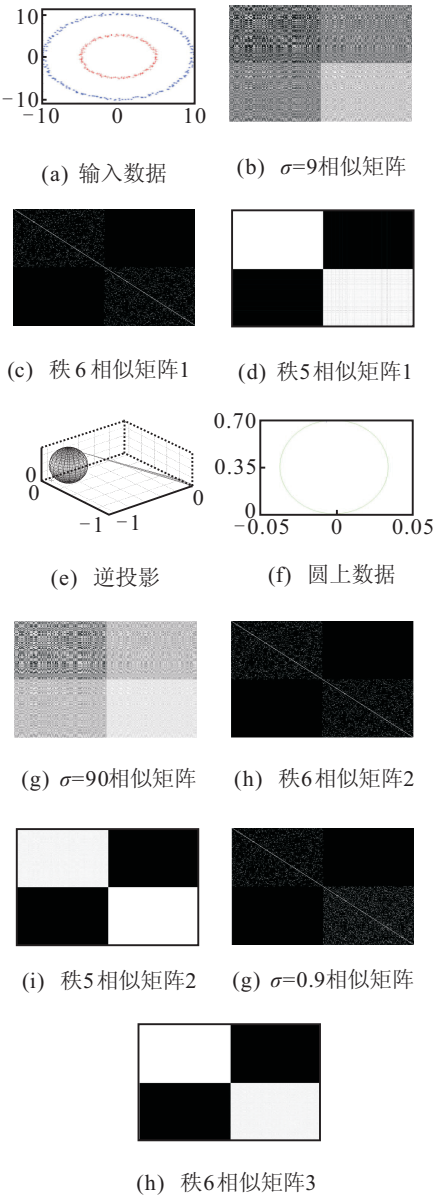
(k) 修正后相似矩阵 3　(l) 修正后分割结果 3

**图 2　迭代聚类结果**

果. NCut 在原始相似矩阵聚类时, 其聚类个数参数人为给出. 经过迭代, 相似矩阵聚合性更加明显, 聚类结果比未迭代有较大提高. 对于第 1 组三圈同心圆数据, 迭代前后分别得到这类数据最为常见的两种聚类结果之一. 一般而言, $\sigma$ 决定了聚类更看重数据的全局连通性 (原始聚类结果) 或局部连通性 (迭代聚类结果). 图 2 中后一组实验展示了较为有趣的结果, 迭代后聚类将正方形四边上的数据归为新类, 而不是像原始聚类一样, 将边与顶点数据归为一类.

### 4.2 不同高斯核 $\sigma$ 参数实验

为了验证本文方法在不同高斯核 $\sigma$ 参数下的聚类鲁棒性, 本实验在一组同心圆数据上, 分别令 $\sigma$ 为 0.9, 9 和 90, 得到原始相似矩阵, 并应用本文算法进行迭代聚类. 输入数据如图 3(a) 所示, 图 3(b), 图 3(g) 和图 3(j) 分别是 $\sigma$ 为 9, 90, 0.9 时的原始相似矩阵. 本实验中, 首次迭代将数据降到秩 6 的空间, 即 $d = 6$, 分别得到图 3(c), 图 3(h)和图 3(k) 的相似矩阵. 首次迭代后, $\sigma = 90$ 一组实验获得 0.987 的谱间隙, 完成迭代; 次轮迭代后另两组实验谱间隙分别达到 0.888 和 0.985, 较迭代前聚类程度增加显著, 如图 3(d) 和图 3(i) 所示. 在 $\sigma = 9$ 实验中, 继续迭代实验, 当 $d = 2$ 时的球极平面逆投影如图 3(e) 所示, 进而得到 $d = 1$ 时的低维数据如图 3(f) 所示. 此时, 数据对称分布在圆两侧, 其谱间隙为 1(对应相似矩阵图没有给出). 本文算法在不同 $\sigma$ 下均找到正确类别个数, 聚类结果见图 3(a).




(a) 输入数据 (b) $\sigma$=9相似矩阵

(c) 秩 6 相似矩阵1 (d) 秩5相似矩阵1

(e) 逆投影 (f) 圆上数据

(g) $\sigma$=90相似矩阵 (h) 秩6相似矩阵2

(i) 秩5相似矩阵2 (g) $\sigma$=0.9相似矩阵

(h) 秩6相似矩阵3

**图 3 变化的高斯核参数聚类结果**

### 4.3 自然数据聚类实验

为了验证算法在自然数据集上的迭代聚类能力, 在 UCI 数据库的 Iris 数据集和 Wine 数据集[17]上运行所提出算法, 并与其他几种聚类方法进行比较. Iris 数据集样本容量 150, 属性 4 个. 为了比较迭代前后聚类效果, 采用 Adjusted Rand Index (ARI)[5]对聚类结果与真实类别标注进行比较. ARI 值在 [0,1] 内变化, 为 1 时表明两类标注完全一致. 由于是与真实标注计算 ARI, 实验中 ARI 越高, 性能越优. 传统谱聚类大多采用 $k$-均值算法对特征向量作最后的聚类, 随机的初值选取和空类时随机选取新中心, 为这些方法引入了随机性. 当采用本节开头所述聚类方法后, 无论迭代环节或是聚类环节, 均无随机因素在内, 结果随输入惟一确定.

表 1 对比了不同 $\sigma$ 下 NCut、模糊核 $c$ 均值聚类 (KFCM)[18]、NJW[19]和本文算法在 Iris 数据集上的 ARI. 由于 NCut、KFCM 算法和 NJW 算法聚类结果均具有一定随机性, 3 类算法的 ARI 为 10 次实验的均值. 实验中, 前 3 种算法的类别个数由人为输入, 本文算法自适应确定类别个数. 表 1 的最后一列显示了迭代终止时谱间隙的大小.

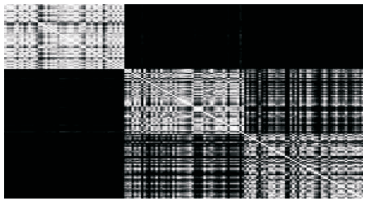
**表 1 Iris 数据迭代聚类 ARI 对比表**

| | NCut | KFCM | NJW | 本文算法 | 谱间隙 |
|---|---|---|---|---|---|
| $\sigma = 0.8$ | 0.677 2 | 0.706 0 | 0.703 2 | 0.771 1 | 0.425 6 |
| $\sigma = 8$ | 0.786 1 | 0.729 4 | 0.703 7 | 0.834 1 | 0.481 7 |
| $\sigma = 0.08$ | 0.563 8 | 0.017 4 | 0.558 9 | NaN | 0.999 9 |

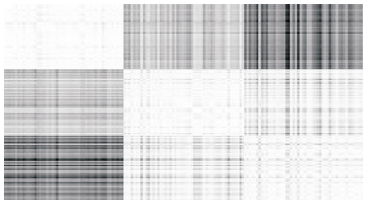
在 $\sigma = 0.8$ 和 $\sigma = 8$ 两组实验中, 本文算法成功获得了正确的类别个数; 采用迭代后相似矩阵聚类数据, 也获得最高的 ARI. 当 $\sigma = 0.08$ 时, 所提出算法的相似矩阵在首次迭代后其元素多数接近 1, 从而前两个特征值之间的谱间隙为 1, $\delta_1 = 1$, 剩余的谱间隙为 0, $\delta_i = 0$, $i = 2, 3, \cdots$, 即只有 1 个类被检出, 随着迭代进行直至 $d = 1$, 该情况没有得到改善. 由于迭代过程中谱间隙始终没有超过阈值, 算法自动判断迭代失败, 此时更改 $\sigma$ 重新迭代. 当 $\sigma = 0.08$ 时, 另外 3 种聚类算法结果也不理想. 可见, 所提出算法在某种程

度上具有一定“纠错”能力, 即 $\sigma$ 选取不当时, 算法可以自动判断迭代失败, 拒绝聚类.

表 2 对比了 Wine 数据集上不同 $\sigma$ 时各聚类算法的 ARI. Wine 数据集样本容量 178, 属性 13 个. 本文算法在 3 组实验中均得到了正确的类别个数; 利用迭代后相似矩阵聚类, 本文算法在 3 组实验中获得两组最优的 ARI.

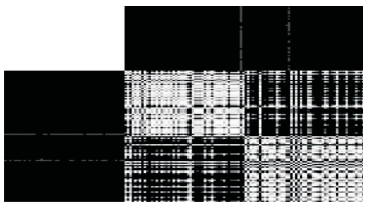


(a) $\sigma$=0.8原始相似矩阵



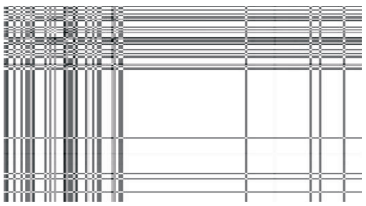
(b) $\sigma$=8原始相似矩阵

(c) $\sigma$=0.08原始相似矩阵



(d) $\sigma$=0.8修正后相似矩阵

(e) $\sigma$=8修正后相似矩阵



(f) $\sigma$=0.08修正后相似矩阵

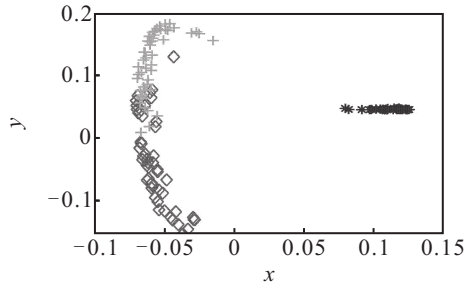
**图 4　Iris 数据迭代聚类相似矩阵**

**表 2　Wine 数据迭代聚类 ARI 对比表**

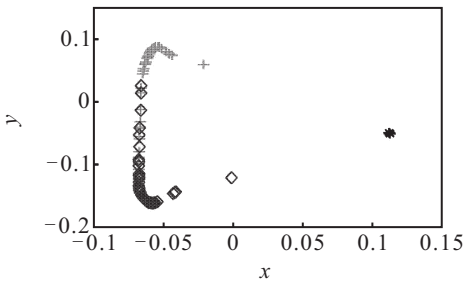
| | NCut | KFCM | NJW | 本文算法 | 谱间隙 |
|---|---|---|---|---|---|
| $\sigma=0.4$ | 0.829 8 | 0.638 1 | 0.798 9 | 0.866 6 | 0.329 4 |
| $\sigma=0.7$ | 0.830 1 | 0.783 4 | 0.831 1 | 0.831 9 | 0.341 7 |
| $\sigma=1$ | 0.829 8 | 0.768 8 | 0.846 4 | 0.831 8 | 0.567 2 |

图 4 为 Iris 数据在不同 $\sigma$ 时迭代前后相似矩阵的变化. 图 4 中, 每列从左至右对应的 $\sigma$ 为 0.8, 8, 0.08, 第 1 行是原始相似矩阵, 第 2 行是迭代终止时的相似矩阵. 图 4 最后一组实验, 算法没有找到正确聚类.

图 5 为 Iris 数据在 $\sigma = 0.8$ 时, 迭代前后归一化拉普拉斯矩阵第 2 特征向量和第 3 特征向量对比图. 迭代前特征向量的类内方差 $S_w = 0.387\,0$, 类间方差 $S_b = 0.032\,2$, 比值为 $S_b/S_w = 0.083\,2$; 迭代后类内方差 $S'_w = 0.337\,7$, 类间方差 $S'_b = 0.033\,2$, 比值为 $S'_b/S'_w = 0.098\,3$. 迭代后特征向量比迭代前类内数据更加聚拢, 类间分离更远, 易于聚类.




(a) 原始相似矩阵特征向量




(b) 迭代后特征向量

**图 5　Iris 数据迭代前后特征向量对比**

## 5　结　论

本文针对多数聚类算法对于数据的相似矩阵较为敏感的问题, 在分析偏振定理和球极平面逆投影几何性质的基础上, 构造了一种迭代修正相似矩阵的算法. 所提出方法分为谱分离和几何分离两步, 对输入数据迭代分离, 改善聚类结果; 利用谱间隙自适应确定聚类个数, 降低了用户对数据先验知识的要求. 该方法在文献 [10] 算法的基础上, 指出距离矩阵投影到双随机矩阵实质上是进行了一次球极平面逆投影, 并利用该逆投影的几何对称性, 对数据进一步分离. 为了获得逆投影后数据的距离矩阵, 给出了球面坐标求解方法. 通过引入球极平面逆投影, 数据的相似矩阵在迭代后展现出更强的聚类性, 提高了聚类正确

率. 在人工合成数据集和自然数据集上验证了所提出算法寻找合适的聚类个数的能力和对聚类结果的改善, 实验结果表明修正后相似矩阵的聚类性能得到提升. 后续工作将围绕大数据集合上算法验证和算法简化展开, 后者主要针对本文算法球极平面参数求解时, 多元高次方程组求解问题加以研究. 此外, 另一条可行的路线是放宽 $B_d$ 计算相似矩阵时精度要求, 使用 $B_d$ 直接估算相似阵.

## 参考文献(References)


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

[2] 姜园, 张朝阳, 仇佩亮, 等. 用于数据挖掘的聚类算法[J]. 电子与信息学报, 2005, 27(4): 655-662.
(Jiang Y, Zhang Z Y, Qiu P L, et al. Clustering algorithms used in data mining[J]. J of Electronics and Information Technology, 2005, 27(4): 655-662.)

[3] 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, 131(6): 58-66.
(Xu S, Lu Z M, Gu G C. Spectral clustering algorithms for document cluster ensemble problem[J]. J on Communications, 2010, 131(6): 58-66.)

[4] Dhillon I S, Guan Y, Kulis B. Kernel $k$-means: Spectral clustering and normalized cuts[C]. Proc of the 10th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM Press, 2004: 551-556.

[5] Alzate C, Suykens J A K. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(2): 335-347.

[6] Xiang T, Gong S. Spectral clustering with eigenvector selection[J]. Pattern Recognition, 2008, 41(3): 1012-1029.

[7] Zhao F, Jiao L, Liu H, et al. Spectral clustering with eigenvector selection based on entropy ranking[J]. Neurocomputing, 2010, 73(10): 1704-1717.

[8] Rebagliati N, Verri A. Spectral clustering with more than $K$ eigenvectors[J]. Neurocomputing, 2011, 74(9): 1391-1401.

[9] Lin F, Cohen W W. Power iteration clustering[C]. Proc of the 27th Int Conf on Machine Learning. Haifa: IMLS, 2010: 655-662.

[10] Brand M, Huang K. A unifying theorem for spectral embedding and clustering[C]. Proc of the 9th Int Workshop on Artificial Intelligence and Statistics. Key West: SAIS, 2003.

[11] Perona P, Zelnik-Manor L. Self-tuning spectral clustering[C]. Advances in Neural Information Processing Systems. Vancouver: MIT Press, 2004: 1601-1608.

[12] 张学工. 模式识别[M]. 北京: 清华大学出版社, 2010: 173-176.
(Zhang X G. Pattern recognition[M]. Beijing: Tsinghua University Press, 2010: 173-176.)

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

[14] Sinkhorn R. A relationship between arbitrary positive matrices and doubly stochastic matrices[J]. The Annals of Mathematical Statistics, 1964, 35(2): 876-879.

[15] Johnson C R, Masson R D, Trosset M W. On the diagonal scaling of Euclidean distance matrices to doubly stochastic matrices[J]. Linear Algebra and Its Applications, 2005, 397(1): 253-264.

[16] 贾春霞. 凸约束的非线性方程系统的仿射内点信赖域法[D]. 上海: 上海师范大学数理信息学院, 2008.
(Jia C X. Affine scaling interior trust region methods for solving convex constrained nonlinear systems[D]. Shanghai: Mathematics and Science College, Shanghai Normal University, 2008.)

[17] UCI data set. Iris, wine data set[DB/OL]. (1988-7-1)[2011-11-28]. http://archive.ics.uci.edu/ml/datasets.html.

[18] 高翠芳. 模糊聚类新算法及应用研究[D]. 无锡: 江南大学数字媒体学院, 2011.
(Gao C F. Novel fuzzy clustering algorithms and applications[D]. Wuxi: School of Digital Media, Jiangnan University, 2011.)

[19] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[C]. Advances in Neural Information Processing Systems. Vancouver: MIT Press, 2001: 849-856.


(责任编辑: 郑晓蕾)