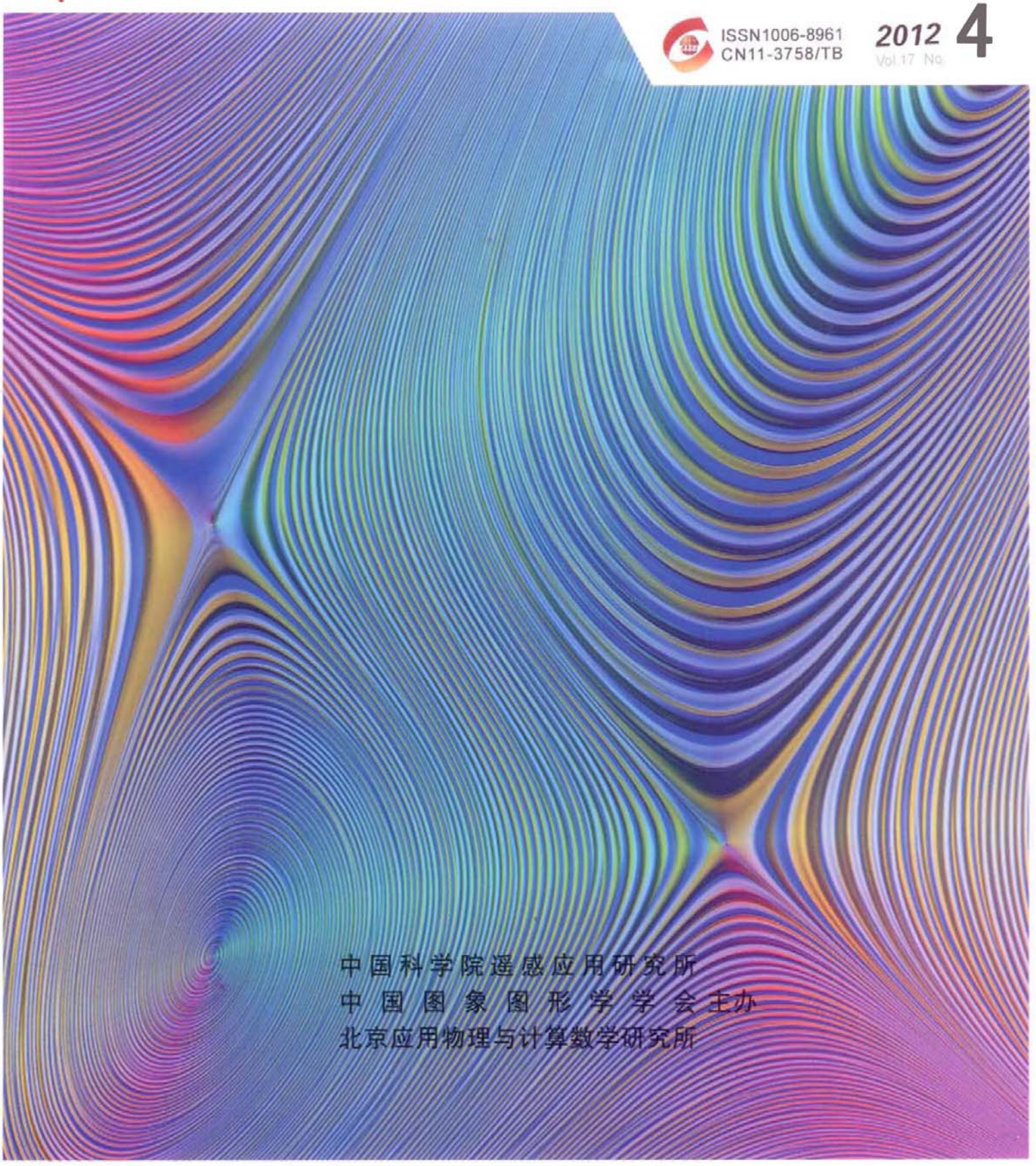
Journal of Image and Graphics

中国图象图形学报

ISSN1006-8961
CN11-3758/TB
2012 Vol.17 No. 4



中国科学院遥感应用研究所
中国图象图形学学会主办
北京应用物理与计算数学研究所

中国精品科技期刊

中国图象图形学报

Zhongguo Tuxiang Tuxing Xuebao

2012 年 4 月　第 17 卷　第 4 期(总第 192 期)

目　次

中国图象图形学报

刊名题字:宋 健　　月刊(1996 年创刊)　　第 17 卷 第 4 期　　2012 年 4 月 16 日出版

主管单位 中国科学院
主　　办 中国科学院遥感应用研究所
中国图象图形学学会
北京应用物理与计算数学研究所
主　　编 李小文
编辑出版 《中国图象图形学报》编辑出版委员会
北京 9718 信箱　邮编 100101
电子信箱:jig@ irsa. ac. cn
电话:010-68407995　010-82614429
网　址:www. cjig. cn
印刷装订 北京北林印刷厂
广告经营许可证 京朝工商广字第 0346 号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
国外发行 中国国际图书贸易总公司
(中国国际书店)
(北京 399 信箱　邮编 100044)

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing Application, CAS China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics
Chief editor LI Xiaowen
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
(P. O. Box 9718, Beijing 100101, China)
E-mail: jig@ irsa. ac. cn
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Foreign China International Book Trading Corporation
(P. O. Box 399, Beijing 100044, China)
Printed by Beijing Beilin Printing House

ISSN 1006-8961　CN11-3758/TB　CODE ZTTXFZ　国内邮发代号:82-831　国外发行代号:M1406　国内定价:45.00 元

第 18 届中国遥感大会征文通知

“第 18 届中国遥感大会”将于 2012 年 10 月 19 日 - 23 日在武汉召开。本届会议由中国遥感委员会主办,中国测绘学会摄影测量与遥感专业委员会和武汉大学承办。会议将围绕“遥感—全方位的社会服务”这一宗旨,以遥感学界院士与知名专家的特邀报告,分会场专题技术交流与技术讲座,重点项目研讨汇报、技术展览,新技术与新产品发布,专业委员会理事会等多种形式开展,同时举行“第 7 届中国青年遥感辩论会”和“第 2 届全国高分辨率遥感数据处理与应用研讨会”。

会议将全方位地展示遥感(RS)、全球定位系统(GPS)、地理信息系统(GIS)等方面的最新成果,为专家、学者和政府主管部门搭建联系纽带,为研发和用户提供技术交流平台,共同促进遥感科技的发展、遥感产业化的推进和大遥感体系的建立。

本届会议围绕大会主题将就遥感新理论、技术、方法和应用进行征文,范围包含但不限于以下方面:

1)国家遥感中长期发展战略、国际遥感前沿与进展;

2)航天、航空、低空、地面遥感技术及系统;

3)光学、红外、高光谱及激光遥感技术;

4)主、被动微波及雷达遥感技术;

5)数字摄影测量与制图;

6)高分辨率遥感数据处理与应用;

7)地理空间数据处理技术与方法;

8)地理国情监测(土地、农业、林业、矿产、环境、地质及水资源等);

9)海洋、气象与全球变化;

10)遥感、地理信息系统与导航定位系统(3S)集成与应用;

11)智慧城市与数字地球;

12)深空探测与行星测绘;

13)教育、培训与社会公共事业。

征文采用在线方式投稿;

投稿要求:论文内容不涉密,且未在国内外学术刊物或正式学术会议上发表过;被录用的全文将收入大会论文集(送 ISTP 检索),并精选 70 ~ 90 篇口头报告论文编辑出版英文 SPIE 会议文集;大会将评选青年优秀论文(参加口头报告),论文将直接进入英文 SPIE 会议文集。

论文摘要截止日期为 2012 年 5 月 15 日,全文截稿日期为 2012 年 6 月 15 日。

会议相关信息,请查阅会议网址:http://rsgis.whu.edu.cn/18ccrs/index.html

“第 18 届中国遥感大会”组委会

第 33 届亚洲遥感会议征文通知

“第 33 届亚洲遥感会议”将由泰国地理信息和空间技术发展局(GISTDA)、科技部(MOST)和亚洲遥感协会(AARS)联合主办,于 2012 年 11 月 26 - 30 日,在泰国芭堤雅市宗滴恩酒店举行。这是亚洲遥感协会每年一届的系列学术会议。本届大会征文包括传感器与平台、算法和图像处理、GIS 与 Web GIS、全球导航卫星系统、灾害、自然资源、环境科学、教育和宣传、健康科学、制图、其他等方面。

会议重要日期:

论文摘要提交截止:2012 年 5 月 15 日;

论文接收通知:2012 年 7 月 1 日;

论文全文提交截止:2012 年 9 月 30 日;

网上注册截止:2012 年 10 月 26 日;

会议召开日期:2012 年 11 月 26 - 30 日。

会议还将组织学生专场和技术展览,其他信息请访问会议网站:http://acrs2012.gistda.or.th

与往年一样,中国遥感委员会仍将鼓励中国遥感科研人员和企事业单位参加会议,并组团参加学术交流和会议展览。

中国遥感委员会

Journal of Image and Graphics

(Monthly, Started in 1996)

Vol. 17 No. 4 April 2012

Contents

中图法分类号: TP391.9 **文献标志码**: A **文章编号**: 1006-8961(2012)04-0523-07
论文引用格式:孔万增,孙昌思核,张建海,胡三清,杨灿. 近邻自适应局部尺度的谱聚类算法[J]. 中国图象图形学报,2012,17(4):523-529

近邻自适应局部尺度的谱聚类算法

孔万增[1],孙昌思核[1],张建海[1],胡三清[1],杨灿[2]
1. 杭州电子科技大学计算机学院,杭州 310018;2. 香港科技大学电子及计算机工程学系,中国香港

摘 要: 针对尺度参数选取对使用高斯核函数的传统谱聚类算法性能的影响,提出一种以近邻自适应局部尺度代替全局统一尺度的新谱聚类算法。该算法在数据聚类一致性特征的基础上,首先强调局部尺度的灵活性,即每个样本数据对应一个尺度参数,克服了传统方法中所有样本对应单一全局尺度参数的局限性,更好地刻画数据集的本征结构。其次注重参数选取的便捷性,即通过对样本周围 N 个近邻计算加权距离和作为局部尺度的值,从而实现了尺度参数的自动选取。从理论和实验两个角度阐述该算法不仅对离群点有一定的抑制作用,而且能对尺度分布不同的数据类进行准确聚类。最后,在人工数据集和 UCI 数据集上验证了该算法的有效性。

关键词:局部尺度;谱聚类;近邻自适应;全局尺度

Spectral clustering based on neighboring adaptive local scale

Kong Wanzeng[1], Sun Changsihe[1], Zhang Jianhai[1], Hu Sanqing[1], Yang Can[2]
1. *College of Computer Science, Hangzhou Dianzi University, Hangzhou* 310018, *China*;
2. *Department of Electronic and Computer Engineering, HKUST, Clear Water Bay, Hongkong, China*

Abstract: Considering the performance of traditional spectral clustering using Gaussian kernels, a new spectral clustering based on neighboring adaptive local scale is presented in this paper. Based on clustering consistency characteristics, the proposed method first emphasizes the flexibility of the local scale, which means each sample has a corresponding scale parameter. Furthermore, it overcomes the limitations of traditional methods in all samples with the same global scale parameter. Hence, it can depict the intrinsic structure of data sets better. Second, it stresses the convenience of parameter selection. It can determine the value of a local scale for one sample by computing the sum of weighted distances of N neighbors. Therefore, it can determine the scale parameter automatically. This paper illustrates the proposed algorithm not only has inhibition for certain outliers but is able to cluster the data sets with different scales. Finally, experiments on both, artificial data and UCI data sets, show that the proposed method is effective.

Key words: local scale; spectral clustering; neighboring adaptive; global scale

0 引 言

谱聚类算法[1-2]最早从谱图划分理论演化而来,在谱聚类出现以前,传统聚类算法(如 k-means, FCM 等)在进行聚类时,将由各种特征(如距离等)构造而成的相似矩阵,使用矩阵的具体元素当作两个样本间的相似度来对待,只是将相似度作为算法的输入数据,并没有考虑相似度矩阵自身对于聚类结果的影响。另外,k-means 等算法本身存在样本需要服从高斯分布的隐性假设,谱聚类算法本质上不需要样本服从某种分布,它直接分析相似度矩阵本

收稿日期:2011-03-16;**修回日期**:2011-09-29
基金项目:国家自然科学基金项目(61102028,61070127);浙江省国际合作重大项目(2009C14013)
第一作者简介:孔万增(1980—),男,副教授,2008 年在浙江大学控制理论与控制工程专业获博士学位,主要研究方向为模式识别与人机交互。E-mail: kongwanzeng@hdu.edu.cn

身,通过求其拉普拉斯矩阵的特征向量问题,来达到聚类的目的。因此,很大程度上避免了引入样本空间分布假设带来的局限性,理论上能在任意形状的样本空间上进行聚类。谱聚类这一优良特性使其已被成功应用到图像分割[3-4]、数据挖掘[5-6]、信息检索[7-8]等领域。但传统谱聚类算法有两个关键技术参数需要事先人为确定,即聚类个数 k 和尺度参数 σ,孔万增等人[9]针对谱聚类如何自动估计类个数进行了研究,但即便在类个数确定的前提下,聚类效果很大程度上还依赖于距离尺度参数的选择,若盲目选取,不但随机性强,且其要花费大量的时间。针对密度分布不同的数据,王玲等人[10]提出一种密度敏感的相似性度量方式,能对密度分布不同的数据进行分类,但这种密度敏感的距离定义和计算相对比较复杂。本文在分析了数据聚类一致性特征的基础上,提出一种近邻自适应局部尺度代替传统谱聚类算法中的全局统一尺度参数,该方法简化了尺度参数的选取,降低算法的人为随机性,而且对数据的分布类型与分布密度均不需要任何限制,同时提高聚类的准确性。最后通过人工数据集和UCI数据库上的实验验证了本文算法性能的优越性。

1 谱聚类及其尺度参数影响

1.1 谱聚类算法

本文的分析是建立在经典谱聚类算法(NJW算法[11])的基础上,首先简单描述NJW算法。算法输入:待聚类数据集 $\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_N\}$,尺度参数 σ 和聚类个数 k;输出:数据聚类划分。具体流程如下:

1)计算相似度矩阵 $\boldsymbol{A}\in\mathbf{R}^{n\times n}$,其中

$$A_{ij}=\begin{cases}\exp\left(-\dfrac{\|\boldsymbol{s}_i-\boldsymbol{s}_j\|^2}{2\sigma^2}\right) & i\neq j\\ 0 & i=j\end{cases} \tag{1}$$

2)构建规范化Laplacian矩阵

$$\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$$

式中,$\boldsymbol{D}$ 为度矩阵,定义 $D_{ii}=\sum_{j=1}^{n}A_{ij}$;

3)计算Laplacian矩阵的前 k 个最大特征值所对应的特征向量 $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k$(必要时需作正交化处理),构造矩阵 $\boldsymbol{X}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_k]\in\mathbf{R}^{n\times k}$;

4)将矩阵 $\boldsymbol{X}$ 的行向量转变为单位向量,得到矩阵 $\boldsymbol{Y}$,即

$$\boldsymbol{Y}_{ij}=\frac{\boldsymbol{X}_{ij}}{\sqrt{\sum_j\boldsymbol{X}_{ij}^2}}$$

5)将矩阵 $\boldsymbol{Y}$ 的每一行看做是 $\mathbf{R}^k$ 空间中的一个点,对其使用 k 均值算法或任意其他经典算法,得到 k 个聚类,并将数据点 $\boldsymbol{y}_i$ 划分到聚类 $\mathbf{C}$ 中,当且仅当 $\boldsymbol{Y}$ 的第 i 行被划分到聚类 $\mathbf{C}$ 中。

1.2 尺度参数取值的影响

从NJW算法可以看出,有一个参数影响着最终的聚类效果:相似函数中的尺度参数 σ。尺度参数的选取与相似度函数的最终值紧密相关,因此在相似度函数确定的条件下(一般取高斯核函数),尺度参数的选取等价于相似度矩阵的选取,而相似度矩阵又是谱聚类算法的基础,直接影响了谱聚类算法的聚类结果。

本节将首先分析尺度参数 σ 对聚类结果的影响。实验数据集由300个样本点组成的三环数据如图1(a)所示,不同类用不同符号和颜色表示。对三环数据集选用不同 σ 进行谱聚类,聚类结果如图1所示。图1中,当 $\sigma=1.2$ 时,数据集得到正确分类结果;当 $\sigma=2$ 和 $\sigma=3.9$ 时,数据集得到的是错误的分类结果,且分类结果又大不相同。通过大量实验,我们得到数据集的最佳尺度参数 σ 取值范围为0.5~1.2。可以看到,尺度参数 σ 的最佳取值范围很有限,有时候甚至很难在实值空间找到合适的值。

2 近邻自适应局部尺度谱聚类算法

由上面的分析可知,谱聚类算法不仅很难给尺度 σ 设定一个固定值,且其聚类性能很大程度上依赖于 σ 值的选取。

2.1 局部尺度

通过观察可以发现,数据聚类有两个所谓的一致性特征[12]。

1)局部一致性 在空间位置上相邻的数据点具有更高的相似性。

2)全局一致性 具有同一结构(同类)的数据点具有更高的相似性。

特别是密度分布不同的类,更体现了同类数据的全局一致性和不同类数据的局部一致性。如图2有两个密度分布相差很大的类 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$,其中 $\boldsymbol{s}_t$ 和 $\boldsymbol{s}_q$ 属于稀疏类 $\boldsymbol{C}_1$,$\boldsymbol{s}_p$ 属于稠密类 $\boldsymbol{C}_2$,为了方便说明,不妨设 $\|\boldsymbol{s}_t-\boldsymbol{s}_p\|=\|\boldsymbol{s}_t-\boldsymbol{s}_q\|$。若使用全局统一尺度参数 σ,可得

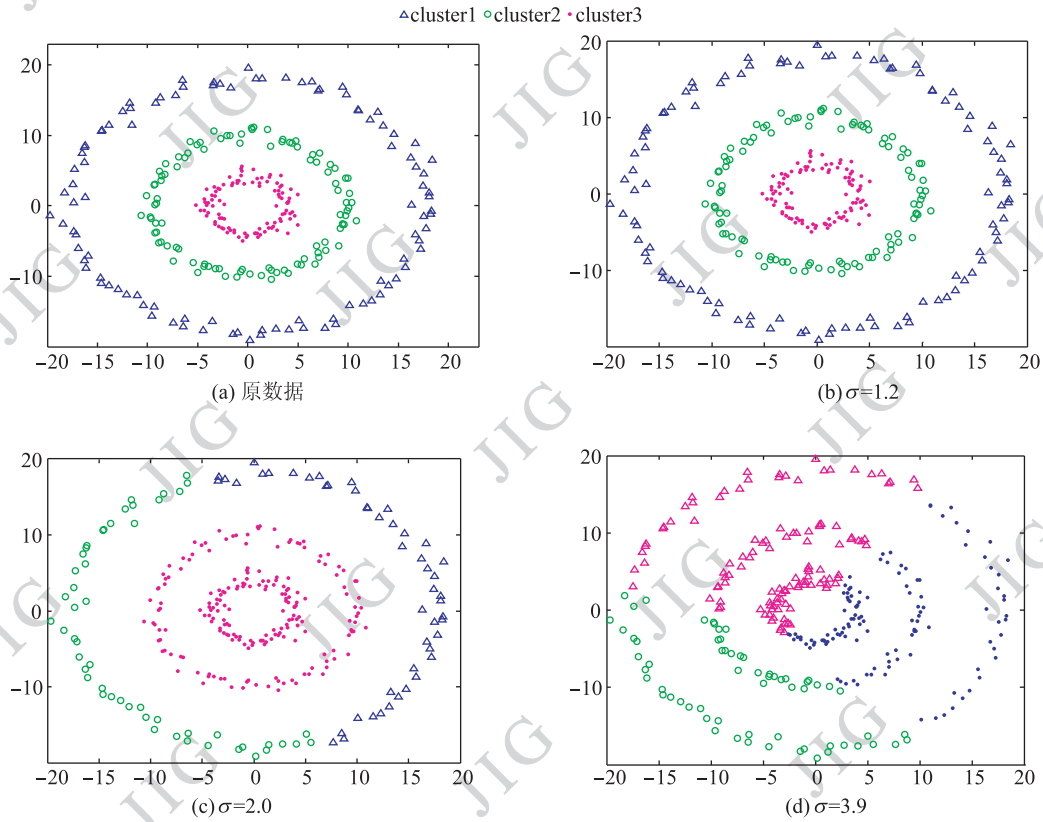


图 1　实验数据及不同 σ 取值的聚类效果

Fig. 1　Experiment data and classification results with differentσ

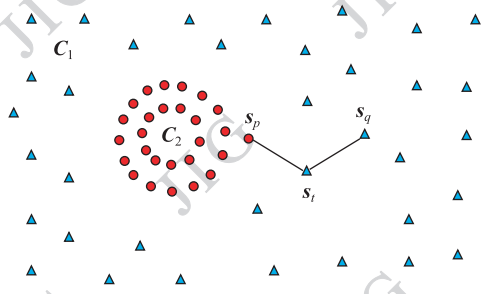


图 2　加权近邻自适应尺度的原理图

Fig. 2　Illustration of weighted neighboring adaptive scale

$$\boldsymbol{A}_{tp} = \exp\left(-\frac{\|\boldsymbol{s}_t - \boldsymbol{s}_p\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\|\boldsymbol{s}_t - \boldsymbol{s}_q\|^2}{2\sigma^2}\right) = \boldsymbol{A}_{tq} \tag{2}$$

即相同距离的点相似度也相同,使得数据点 $\boldsymbol{s}_p$ 和 $\boldsymbol{s}_q$ 更偏向于同一类中,这显然与事实不符。

为了能更好的体现聚类结构的局部一致性特征,也为了能够快速准确的设置尺度参数值,本文提出为每个数据点 $\boldsymbol{s}_i$ 计算一个属于该样本的加权近邻局部尺度参数

$$\sigma_i = \sum_{j=1}^{N} W_{ij}\|\boldsymbol{s}_i - \boldsymbol{s}_j\| \tag{3}$$

式中,W_{ij}为数据点 $\boldsymbol{s}_i$ 和其 N 个邻近点之间的权系数(N 值一般取 3～7),称 σ_i 为 $\boldsymbol{s}_i$ 的近邻自适应局部尺度。而文献[13－14]只考虑样本的第 N 个邻近(N 值一般为 7),没有考虑其周围样本的分布特性,易受离群数据点的影响。本文尺度综合了样本周围 N 个样本的分布,因此,在聚类时更具鲁棒性。由此,数据点 $\boldsymbol{s}_i$ 和 $\boldsymbol{s}_j$($i \neq j$)相似度定义为

$$A'_{ij} = \exp\left(-\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{\sqrt{2}\sigma_i} \cdot \frac{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|}{\sqrt{2}\sigma_j}\right) \tag{4}$$

若 $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ 为欧氏距离,则 $\|\boldsymbol{s}_i - \boldsymbol{s}_j\| = \|\boldsymbol{s}_j - \boldsymbol{s}_i\|$,由此可得

$$A'_{ij} = \exp\left(-\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|^2}{2\sigma_i\sigma_j}\right) \tag{5}$$

如此,尺度参数 σ_i 将会随着数据邻近的分布而变化,能体现聚类数据的局部一致性特性,此时由图 2 易知 $\sigma_q > \sigma_p$,所以

$$A'_{tp} = \exp\left(-\frac{\|\boldsymbol{s}_t - \boldsymbol{s}_p\|^2}{2\sigma_t\sigma_p}\right) = \exp\left(-\frac{\|\boldsymbol{s}_t - \boldsymbol{s}_q\|^2}{2\sigma_t\sigma_p}\right) <$$

$$\exp\left(-\frac{\|\boldsymbol{s}_t-\boldsymbol{s}_q\|^2}{2\sigma_t\sigma_q}\right)=A'_{tq}$$

可见,加权近邻自适应适度使得类内点相似度值更大,类间相似度值更小,进而更好地挖掘各自类的本征结构。

2.2 权值的确定

权值的引入,使算法对离群点有一定的鲁棒性,同时使加权近邻自适应尺度谱聚类算法可以通过调整这些权值以优化聚类效果。本文权值的设定方法是:

1)选取数据点的 N 个邻近点;

2)根据下式来确定权值

$$W_{ij}=\frac{\frac{1}{d_{ij}}}{\sum_{k=1}^{N}\frac{1}{d_{jk}}} \tag{6}$$

式中,N 为邻近点个数(一般选取3~7),d_{ij} 为两样本之间的欧氏距离。该权值能抑制某样本周围分布较远的点对该样本尺度参数值的影响。特别地,当样本为严格均匀分布时,$W_{ij}=\frac{1}{N}$。

2.3 算法小结

结合上述如何确定尺度参数值及其原理分析,本文设计了近邻自适应局部尺度谱聚类算法,算法输入:待聚类数据集 $\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_N\}$ 和聚类个数 k;输出:数据聚类划分。描述如下:

1)根据式(3)计算每个数据点的加权近邻自适应尺度参数;

2)根据式(4)构造相似度矩阵 $\boldsymbol{A}'$,并构造规范化相似度矩阵 $\boldsymbol{A}'_{\text{nor}}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}'\boldsymbol{D}^{-\frac{1}{2}}$,其中 $\boldsymbol{D}$ 是对角化矩阵,$D_{ij}=\sum_j A'_{ij}$;

3)利用规范化相似度矩阵前 k 个最大特征值对应的特征向量构建特征空间;利用 FCM 等经典聚类方法对特征向量空间中的数据点进行聚类,聚类结果映射回原数据空间。

3 实验结果和分析

为了验证近邻自适应局部尺度谱聚类算法理论分析的正确性及其有效的分类性能,本文分别在人造数据以及 UCI 标准数据集中进行了相关实验。

3.1 算法验证

为了验证理论分析的正确性,对复杂人工数据集进行实验,选用的相似度函数为

$$A'_{ij}=\begin{cases}\exp\left(-\dfrac{\|\boldsymbol{s}_i-\boldsymbol{s}_j\|^2}{2\sigma_i\sigma_j}\right) & i\neq j\\ 0 & i=j\end{cases}$$

$$i,j=1,2,\cdots,n$$

尺度参数为加权近邻自适应尺度参数,$\sigma_i=\sum_{j=1}^{N}W_{ij}d_{ij}$。

密度分布不同的数据集如图3所示,该数据集具有明显的局部一致性,4个块状类形象地刻画了在空间位置上相连的数据点具有更高的相似性,同时它的第5个类,又体现出该数据集具有相同结构的数据点具有更高的相似性的全局一致性。对于此数据集,即使通过大量重复实验的方法,也不一定能找到有效全局统一尺度参数 σ 使聚类效果达到最佳,只能采用近邻自适应的局部尺度谱聚类才能得到正确的分类。

图4(b)(c)分别显示了基于全局统一尺度参数计算出的相似度矩阵和近邻自适应局部尺度计算出的相似度矩阵(其中数据样本每类顺序排列)。通过对理想情况下相似度矩阵的分析可知,相似度矩阵是块对角矩阵这一理想情形时,谱聚类算法容易得到正确的聚类。显然,由近邻自适应局部尺度计算得到的相似度矩阵减少了数据集块状外的干扰,从而使其对角块状效应更加明显,这说明本文方法缩小了不同类数据点间的相似度,同时又增大了同一类内数据点间的相似度,能够更好地识别数据本质的聚类结构。

3.2 人工数据集实验

文献[8]给出一些“挑战性”问题,从中挑选较为困难的,并给出另外一些困难数据集。图5给出标准谱聚类(NJW 算法)、Self-tunning 谱聚类和自适应局部尺度谱聚类3种算法在5个数据集上的聚类结果。实验中 NJW 算法的 σ 值按文献[3]给出的方法选取,为数据点之间距离变化范围的10%~20%(记做 $d=\max(\boldsymbol{L})-\min(\boldsymbol{L})$,$L_{ij}=\|s_i-s_j\|$)。为不失一般性,这里 $\sigma=0.1\ d$,在 Self-tuning 谱聚类算法和自适应局部尺度谱聚类算法中,选取 $N=7$。图5(a)为 NJW 算法的聚类结果,全部出现了错误划分,这是因为尺度参数选取不当的原因;图5(b)是 Self-tunning 算法的实验结果,对数据集1和2的

聚类是正确的,而对其他 3 个数据集或多或少都出现了错误划分,再一次验证了仅选择近邻第 N 个点来计算尺度参数的 Self-tunning 算法的局限性,以此方法来计算尺度参数不够鲁棒;图 5(c)是自适应局部尺度谱聚类的实验结果,无论是对密度分布不均的数据集还是对非块状数据集(条形、弧形、圈形等)均取得了正确的划分,实验证明了近邻自适应局部尺度谱聚类的强大分类功能。

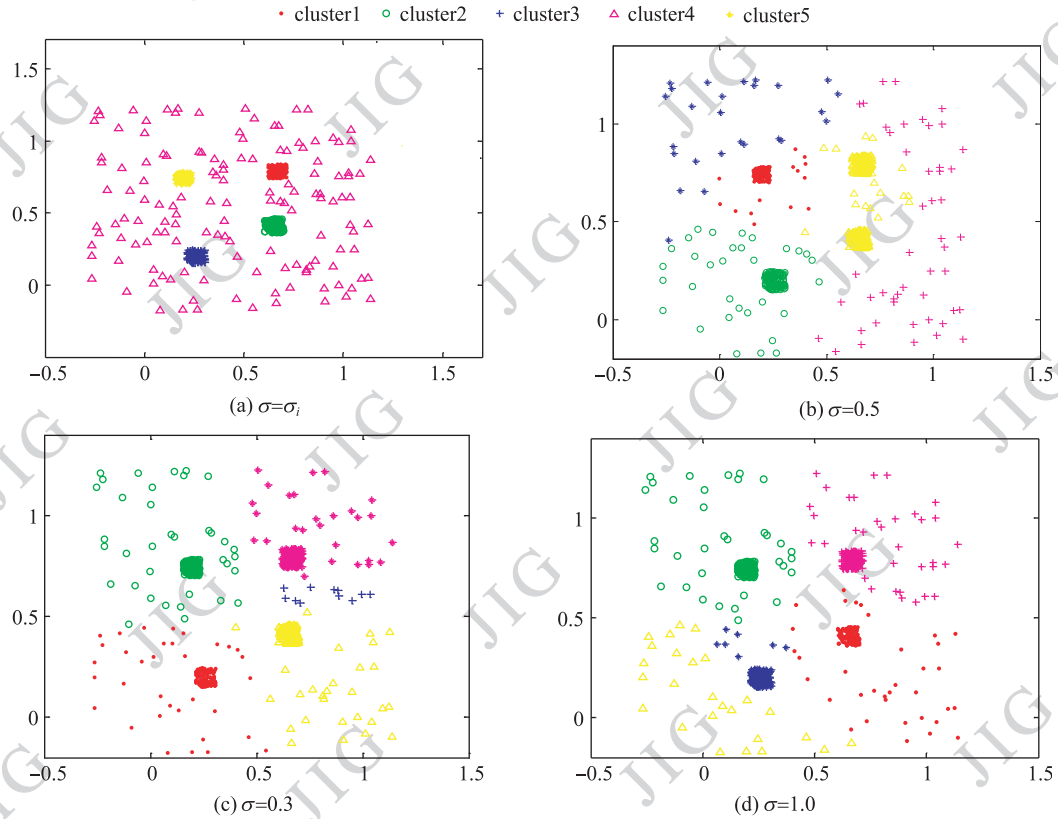


图 3　σ 的取值对谱聚类算法的影响

Fig. 3　Effect on classification result with different σ

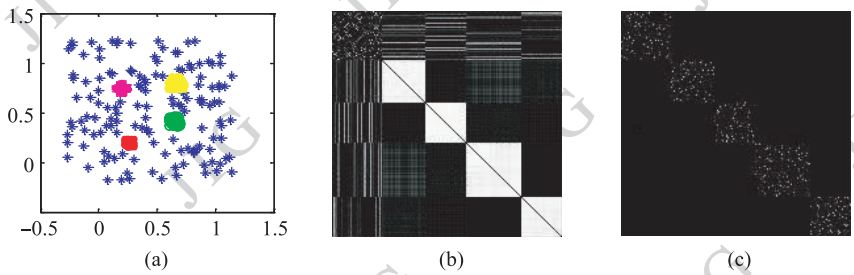


图 4　密度不均匀数据集及其相似度矩阵

Fig. 4　A data set with different scale and its affinity matrix

3.3　UCI 数据实验

为了验证本文算法的实际应用能力,在国际通用数据库 UCI 数据库(专门用于测试分类、聚类算法)中的 Iris、Hea 和 Acd 3 个数据集上进行了 5 种不同聚类算法的比较实验。其中 Iris 数据有 150 个数据点,4 个特征,分为 3 类;Hea 数据有 270 个数据点,13 个特征,分为 2 类;Acd 数据有 690 个数据点,14 个特征,也分为 2 类,如表 1 所示。

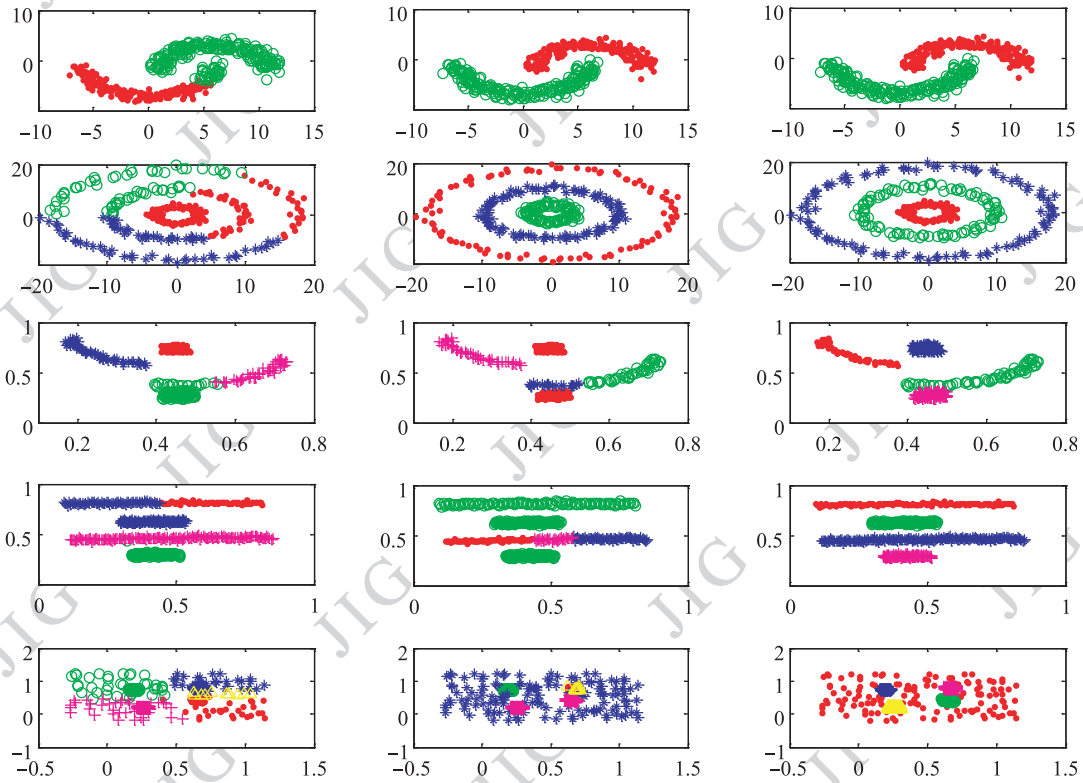


图 5 3 种算法在 5 个数据集上的聚类结果

Fig. 5 Results on the five clustering sets of the three clustering algorithms

表 1 实验中用到的 UCI 数据集

Table 1 UCI data set used in experiments

数据集	数据属性		
	数据点	维数	类数
Iris	150	4	3
Hea	270	13	2
Acd	690	14	2

表 2 则给出了 k-means、FCM、NJW 和 Self-tuning 4 种聚类算法的分类准确率，由于 k-means、FCM 算法的结果都受到初始类中心的影响，每次运行结果可能会不一样，因此表 2 中结果为运行 10 次后取算法准确率平均值及其方差。实验中 NJW 算法的 σ 值与上节一样按文献[3]给出的方法选取，不失一般性，这里在 $\sigma=0.1d$ 和 $\sigma=0.2d$ 两种情况下分别进行实验。在 Self-tuning 谱聚类算法中，则根据文献中说明，选取 $N=7$，为不失一般性，这里也选取了 $N=7$ 便于比较。

表 2 的结果表明，NJW 算法、Self-tuning 算法和本文的近邻自适应局部尺度谱聚类算法在该数据集上分类准确率普遍比传统的 k-means 和 FCM 算法要高。NJW 与 Self-tuning 算法和本文算法的差别在于前者需要事先人为选取尺度参数值，后两者为自动确定；虽然前者选取也存在找到某一 σ 使得它在该数据集上的分类准确性高于 Self-tuning 和本文算法的可能性，但这需要花费大量的时间和精力，且不一定能找到。而后两者区别在于 Self-tuning 算法自动确定尺度参数值时只考虑数据点的第 N 个邻近点。本文算法不仅考虑了前 N 个邻近点，并对前 N 个邻近点设置了权值，使其更符合数据聚类的局部一致性特性。表 2 同时表明，本文算法性能优于 Self-tuning 算法和其他分类算法，最直接的例证便是：从分类准确类的方差来看，本文算法的稳定性最好。在对每个数据库的 10 次分类中方差均为 0，尤其对 Iris 数据库，其他所有的算法包括 NJW 和 Self-tuning 算法均有很大的波动，但本文算法非常稳定，原因就在于本文算法在选取尺度参数时考虑了周围 N 个样本点并加权处理，有效的抑制了离群点的干扰，使得到的尺度参数在算法中更加鲁棒。

表 2　5 种不同聚类算法在 UCI 数据集上的比较

Table 2　Performance compare of five clustering algorithms on UCI data　/%

数据集	k-means	FCM	NJW		Self-tuning	本文算法
			$\sigma=0.1d$	$\sigma=0.2d$	$N=7$	$N=7$
Iris	75.8 ±17.5	70.4 ±10.2	87.0 ±9.4	85.6 ±9.7	88.2 ±7.8	90.7 ±0
Hea	59.3 ±0	58.9 ±0	61.1 ±0	60.0 ±0	61.4 ±0	61.9 ±0
Acd	55.94 ±0	56.1 ±3.4	56.23 ±0	57.10 ±0	66.38 ±0.2	66.52 ±0

4　结　论

首先分析传统谱聚类算法选取尺度参数对聚类结果的影响,然后在充分分析数据聚类一致性特征的基础上,提出一种以近邻自适应局部尺度代替全局统一尺度的新谱聚类算法。由于近邻自适应尺度为 N 个近邻的加权距离比第 N 个近邻更稳定,使算法对离群点有一定的抑制作用,同时,邻近个数 N 的取值一般是 3～7 的整数,也比选取全局尺度参数简便。最后,人工数据集实验结果显示,本文方法对复杂分布数据有很强自适应能力,能刻画数据集的自身内在结构,并准确分类;UCI 数据集实验结果表明,本文算法较其他谱聚类算法和经典聚类的分类准确率高且算法运行结果更稳定。下一步的工作主要是开展谱聚类算法在高维大数据集上的快速性及准确性研究。

参考文献(References)

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

[2] Filipponea M, Camastrab F, Masullia F, et al. A survey of kernel and spectral methods for clustering [J]. Pattern Recognition, 2008, 41(1):176-190.

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

[4] Wang C, Li W, Ding L, et al. Image segmentation using spectral clustering[C]. // Proceedings of the 17th IEEE International Conference on Tools with Aritificial Intelligence, 2005:677-678.

[5] Dhillon I S, Guan Y, Kulis B. Weighted graph cuts without eigenvectors: A multilevel approach[J]. IEEE Fransactions on Pattern Anaeysis and Machine Intelligence. 2007, 29(11): 1944-1957.

[6] Sarkar S, Soundararajan P. Supervised learning of large perceptual organization: Graph spectral partitioning and learning automata [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(5):504-525.

[7] Wang C X, Wang J C. Application of spectral clustering in image retrieval[J]. Journal of Computer Research and Development, 2009,19(1):207-210. [王春雪,王继成. 谱聚类在图像检索中的应用[J]. 计算机技术与发展, 2009,19(1):207-210.]

[8] Cristianini N, Taylor J S, Kandola J S. Spectral kernel methods for clustering [C]// Proceedings of the Neural Information Processing Systems. Cambridge, MA: MIT Press, 2002:649-655.

[9] 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):1980-1985. [孔万增,孙志海,杨灿,等,基于本征间隙和正交特征向量的谱聚类算法[J]. 电子学报. 2010, 38(8):1980-1985.]

[10] Wang L, Bo L F, Jiao L C. Density-Sensitive Spectral Clustering [J]. Acta Electronica Sinica, 2007,35(8):1577-1581. [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.]

[11] Nga Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm [C]// Proceedings of the 14th Advances in Neural Information Processing Systems (NIPS). Cambridge, MA: MIT Press, 2002: 849-856.

[12] Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency [C]// Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2004: 321-328.

[13] Lihi Z M, Pietro P. Self-tuning spectral clustering [C] // Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, 2005: 1601-1608.

[14] Gu R J, Ye B, Xu W B. An improved spectral clustering algorithm[J]. Journal of Computer Research and Development, 2007,44(z2): 145-149. [谷瑞军,叶宾,须文波. 一种改进的谱聚类算法[J]. 计算机研究与发展. 2007,44(z2):145-149.]