一种基于粗糙集理论的谱聚类算法

郑 吉 苗夺谦 王睿智 钟才明

(同济大学计算机科学与技术系 上海 201804)

摘 要 谱聚类算法利用特征向量构造简化的数据空间,在降低数据维数的同时,使得数据在子空间中的分布结构更加明显。现有谱聚类算法的聚类结果多为精确集,而真实数据集中重叠现象广泛存在。基于粗糙集理论提出了一种新的谱聚类算法,其主要思想是对谱聚类算法进行粗糙集扩展,使得聚类结果成为具有下近似和上近似定义的、类与类之间存在重叠区域的结构。实验表明,该算法与现有的谱聚类算法相比,稳定性和准确率都有一定的提高。

关键词 粗糙集,谱聚类,k均值聚类

中图法分类号 TP301.6 **文献标识码** A

Rough-set Based Spectral Clustering

ZHENG Ji MIAO Duo-qian WANG Rui-zhi ZHONG Cai-ming

(Department of Computer Science and Technology, Tongji University, Shanghai 201804, China)

Abstract The spectral clustering algorithm constructs a simplified data space making the use of the eigenvectors that not only reduces the dimension of data but also gives clearer distribution of data in the subspace. The results of most existing spectral clustering algorithm are precise sets while widespread 'overlapping' exists in real data sets. This paper proposed a new spectral clustering algorithm which is based on the rough set theory. The main idea is to extend spectral clustering with rough set theory to obtain the results with lower-and-upper-approximation definition and between-cluster-overlapped structure. Experiment results indicate that the proposed algorithm outperforms the existing spectral clustering algorithms in both stability and accuracy.

Keywords Rough set, Spectral clustering, K-means clustering

聚类是数据挖掘、模式识别等研究方向的重要研究内容之一。机器学习中的聚类算法应用于图像分割和机器视觉。图像处理中聚类算法应用于数据压缩、信息检索。聚类的另一主要应用是数据挖掘(多关系数据挖掘)、时空数据库应用(GIS等)、序列和异类数据分析等。此外,聚类还应用于统计科学[1]。

近年来所提出的谱聚类是一种较为实用的聚类方法[2]。根据图的谱分割原理,谱聚类在进行聚类时首先以待聚类的对象集为顶点集构造带权图,然后通过分析一个与图相关的矩阵的特征向量和特征值来得到聚类结果。由于图的边权可以结合待聚类对象的各种特征,因此谱聚类方法简单,可以处理复杂的数据类型,目前已经出现了许多谱聚类模型和算法,如Ratio-cut[3],Normalized-cut[4]及Min-max-cut[5]等。

粗糙k均值算法是P. Lingras提出的一种新型聚类算法[6]。通过对经典k均值算法进行粗糙集扩展,使得聚类的结果成为一种具有下近似、上近似两个层次的结构。

本文将粗糙k均值的概念引入谱聚类算法中,提出了一种具有两个层次的基于粗糙集理论的谱聚类算法。

1 谱聚类算法

谱聚类算法的思想来源于谱图划分理论。它将聚类问题看成是一个无向图的多路划分问题。定义一个图划分判据,如Shi&Malik提出的一个有效的图划分判据——Normalized-cut判据[4],最优化这一判据,使得同一类内的点具有较高的相似性,而不同类之间的点具有较低的相似性。由于图划分问题的组合本质,求图划分判据的最优解是一个NP难问题。一个很好的求解方法是考虑问题的连续放松形式,这样便可将原问题转换成求图的Laplacian矩阵的谱分解,因此,将这类方法统称为谱聚类,可以认为谱方法是对图划分判据的逼近[7]。谱聚类也可以利用类似于PCA子空间方法中的嵌入思想来解释。该方法同时使用矩阵的多个特征向量,利用这些特征向量构造一个简化了的数据空间,在该空间中数据的分布结构更加明显。

谱聚类算法本质上是利用邻接矩阵的特征向量进行聚类。已知对象集合(数据点)的相似矩阵,可以通过全连通法构造邻接矩阵。本文中我们用高斯相似函数$s(x_i,x_j)=\exp$

到稿日期:2008-06-04 本文受国家自然科学基金(60475019, 60775036),教育部博士点专项基金(20060247039)资助。

郑 吉(1983—),男,硕士研究生,主要研究方向为粗糙集、聚类分析等,E-mail:uowenz@gmail.com;苗夺谦(1964—),男,教授,博士生导师,主要研究方向为粗糙集、主曲线和粒度计算等;王睿智(1968—),女,博士研究生,主要研究方向为聚类分析、粗糙集、粒度计算等;钟才明(1970—),男,博士研究生,主要研究方向为数据挖掘、机器学习等。

$(-\|x_i-x_j\|^2/2\sigma^2)$为原始数据集$X=\{x_1,\cdots,x_n\}$建立相似矩阵，其中$\sigma$是事先给出的参数，用来控制数据点之间的距离宽度。Ng等人在2002年提出一种谱聚类算法[8]，该算法选取矩阵的前k个最小的特征值所对应的特征向量，从而在$\mathbb{R}^k$空间中构成与原数据一一对应的表述，并在$\mathbb{R}^k$空间中进行聚类。该算法表述如下。

算法1

输入：数据集$X=\{x_1,\cdots,x_n\}$，聚类个数k，参数σ。

输出：聚类$A_1,\cdots,A_k$。

第1步　计算相似矩阵$S=(s_{ij})_{i,j=1,\cdots,n}$，$s_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2)(i\neq j,s_{ii}=0)$，并根据相似矩阵构造邻接矩阵$W\in\mathbb{R}^{n\times n}$。

第2步　计算归一化Laplacian矩阵$L_{sym}=D^{-1/2}LD^{-1/2}=I-D^{-1/2}WD^{-1/2}$，其中$D$为对角阵，$d_{ii}=\sum_{j=1}^{n}\omega_{ij}$，$L=D-W$。

第3步　计算L_{sym}的最小的k个特征值对应的特征向量$v_1,\cdots,v_k$，以每个特征向量为一列构造矩阵$V\in\mathbb{R}^{n\times k}$。

第4步　对矩阵V进行行归一化处理得到矩阵$U\in\mathbb{R}^{n\times k}$，$u_{ij}=v_{ij}/(\sum_k v_{ik}^2)^{1/2}$。

第5步　令$y_i\in\mathbb{R}^k$对应矩阵U的第i个行向量$(i=1,\cdots,n)$。

第6步　用k均值算法将数据点集$(y_i)_{i=1,\cdots,n}$聚成k类$C_1,\cdots,C_k$。

第7步　得出聚类结果$A_1,\cdots,A_k$，其中$A_i=\{j\mid y_j\in C_i\}$。

谱聚类算法是一种配对聚类方法，算法仅与数据对象的数目有关，而与维数无关，因而可以避免由特征向量的过高维数所造成的奇异性问题。谱聚类算法又是一种判别方法，不用对数据的全局结构作假设。谱聚类算法成功的原因在于：通过特征分解，可以获得聚类判据在放松了的连续域中的全局最优解。谱聚类相对于其他聚类方法具有明显的优势，它具有识别非凸分布聚类的能力，非常适合于许多实际问题，而且执行起来比较容易。该方法已成功应用于语音识别、图像和视频分割等领域。

现有的谱聚类算法多基于k均值算法，由于k均值算法本身基于精确集，使得谱聚类结果是基于划分的，相互不存在重叠区域的精确集合。k均值聚类不具备识别聚类之间重叠区域的能力。另外，k均值算法处理数据集中孤立点的能力较弱，造成了聚类结果的类内平均距离增大，影响了谱聚类算法的准确率。

本文将现有的谱聚类算法进行了粗糙集扩展。扩展后的算法，其聚类结果基于粗糙集概念，孤立点并不会被分配到某个精确的聚类(下近似)中，而是分配到类间重叠区域(多个类的上近似)中，有效地降低了聚类结果的类内平均距离，较好地解决了上文提到的问题。

2　基于粗糙集理论的谱聚类算法

2.1　k均值算法的粗糙集扩展

近年来，软计算方法如模糊集、神经网络、粗糙集等等被用于解决数据挖掘中的问题。2004年，P. Lingras等人基于粗糙集理论和k均值算法提出了粗糙聚类算法。在P. Lingras的粗糙聚类算法中，每个聚类包括一个下近似和一个上近似。同一类的下近似是上近似的子集。一个类的下近似的成员一定属于这个类，不能属于其他的类。一个类的上近似的成员可能属于这个类，它们属于哪个类是不确定的，所以它们至少还是另外一个或多个类的成员。

P. Lingras用到了粗糙集理论的以下主要性质：

性质1　一个数据对象属于至多一个类的下近似。

性质2　若一个数据对象属于一个类的下近似，那么该数据对象也属于这个类的上近似。

性质3　若一个数据对象不属于任何一个类的下近似，那么该数据对象一定属于至少两个类的上近似。

因此，严格地说，粗糙k均值算法是一种带有下近似和上近似定义的双层次的聚类算法。

在P. Lingras提出的粗糙k均值算法中，计算聚类均值向量的公式为

$$m_i=\begin{cases}w_l\sum\limits_{x_n\in\underline{C_i}}\dfrac{x_n}{|\underline{C_i}|}+w_b\sum\limits_{x_n\in\overline{C_i}-\underline{C_i}}\dfrac{x_n}{|\overline{C_i}-\underline{C_i}|} & ,\overline{C_i}-\underline{C_i}\neq\emptyset\\ w_l\sum\limits_{x_n\in\underline{C_i}}\dfrac{x_n}{|\underline{C_i}|} & ,\text{其它}\end{cases}\tag{1}$$

其中$\underline{C_i}$表示第i个聚类的下近似，$\overline{C_i}$表示第i个聚类的上近似，$i=1,\cdots,k$。w_l表示计算均值向量时，各个类的下近似中各元素具有的权值；w_b表示各个类的边界区域(即上近似与下近似的差集)中各元素具有的权值。

用$d(x_n,m_h)$表示对象x_n与各均值向量的距离中最小的一个。在确定类的边界区域时，构造集合T

$$T=\{t\mid d(x_n,m_t)-d(x_n,m_h)\leqslant\varepsilon\wedge h\neq t\}\tag{2}$$

ε是一个阈值。ε越大，类的边界区域就越大。若$T\neq\emptyset$，则$x_n\in\overline{C_t},\forall t\in T$。否则，$x_n\in\underline{C_h}$。

最初的粗糙k均值算法存在一些潜在问题。首先，两个权值w_l,w_b的选取存在随意性，当$\underline{C_i}=\overline{C_i}$时，须特别指定权值$w_l=1$才能得到与经典$k$均值算法一致的结果。其次，当$\underline{C_i}=\emptyset$时式(1)中分母为0，均值向量无法计算。再次，构造集合T时使用阈值ε判别两个距离值的绝对差值，这使得ε的选取与具体数据对象之间存在一定依赖性。

G. Peters对粗糙k均值算法进行了严格的分析，提出了一种改进算法[9]。该算法较好地解决了上述问题。

2.2　基于粗糙集理论的谱聚类算法

我们提出一种基于粗糙集理论的谱聚类算法(RSC算法)。该算法通过矩阵的谱分解，将原始数据集一一映射到$\mathbb{R}^k$子空间上，再使用改进的粗糙k均值算法得出聚类结果。相比现有的谱聚类算法，粗糙谱聚类算法得出的结果含有类间重叠区域，能够更好地描述原始数据集的内在类别特征。该算法相对最初的粗糙k均值算法，在输入参数以及处理细节方面进行了一些调整，避免了最初算法存在的一些问题。该算法具体步骤如下。

算法2

输入：数据集$X=\{x_1,\cdots,x_n\}$，参数w_l,w_u,ζ。其中w_l表示计算均值向量时，各个类的下近似中各元素具有的权值；w_u表示各个类的上近似中各元素具有的权值；ζ是一个阈值，用于控制边界区域的大小。

输出：聚类$\underline{A_1},\cdots,\underline{A_k},\overline{A_1},\cdots,\overline{A_k}$。其中$\underline{A_i}$表示第$i$个聚类的下近似，$\overline{A_i}$表示第$i$个聚类的上近似，$i=1,\cdots,k$。

第0步　执行算法1的第1步到第5步，求得数据对象

集 $Y=\{y_1,\cdots,y_n\}$。

第 1 步　将数据对象 $y_i(i=1,\cdots,n)$随机地分配到一个聚类的下近似中。由于下近似是上近似的子集，这些数据对象也一定属于相应类的上近似(所有数据对象分配后，若发现某一类的下近似为空集，则重新进行这一步)。

第 2 步　使用下述公式计算均值向量。

$$m_i=w_l\sum_{y_n\in\underline{C_i}}\frac{y_n}{|\underline{C_i}|}+w_u\sum_{y_n\in\overline{C_i}}\frac{y_n}{|C_i|}\tag{3}$$

其中 $w_l+w_u=1,i=1,\cdots,k$。

第 3 步　将数据对象 $y_i(i=1,\cdots,n)$分配到各个上、下近似中。

(i) 把最能代表每个聚类的数据对象分配到这个聚类的下近似和上近似中。

(a) 找到所有 k 个聚类和所有 n 个数据对象的最小距离，假设这一距离是聚类 h 和数据对象 l 之间的距离，将数据对象 l 分配到类 h 中：

$$d(y_l,m_h)=\min_{n,i}d(y_n,m_i)\Rightarrow y_l\in\underline{C_i}\wedge y_l\in\overline{C_i}\tag{4}$$

其中 $i=1,\cdots,k$。

(b) 将 y_l 和 m_h 排除考虑范围。如果存在没有被分配到数据对象的类，转到(a)，否则转到(ii)。

(ii) 对余下的每一个数据对象 $y_m'(m=1,\cdots,M,M=N-K)$，找到与它距离最近的均值 m_h

$$d_{m,h}^{\min}=d(y_m',m_h)=\min_{i=1,\cdots,k}d(y_m',m_i)\tag{5}$$

将 y'_m 分配到类 h 的上近似中。

(iii) 使用下述公式定义的相对距离，找到其它与数据对象 y'_m 距离较近的类 m_t。其中 ζ 是事先给出的相对阈值，$i=1,\cdots,k$

$$T'=\left\{t:\frac{d(y_m',m_i)}{d(y_m',m_h)}\leqslant\zeta\wedge h\neq i\right\}\tag{6}$$

如果 $T'\neq\varnothing(y_m'$和至少一个除 m_h 之外的类 m_t 距离也很近)，则 $y_m'\in\overline{C_t},\forall t\in T'$，否则 $y_m'\in\underline{C_h}$。

第 4 步　检查算法是否收敛。如果算法未收敛，转到第 1 步，否则结束。

第 5 步　得到聚类结果$\underline{A_1},\cdots,\underline{A_k},\overline{A_1},\cdots,\overline{A_k}$其中$\underline{A_i}=\{j\mid y_j\in\underline{C_i}\},\overline{A_i}=\{j\mid y_j\in\overline{C_i}\}$。

分析上述算法可以发现，两个权值 w_l,w_u 的选取满足 $w_l+w_u=1$，使得计算出的均值向量是聚类上、下近似均值向量的线性相加，当$\underline{C_i}=\overline{C_i}$时，无论权值如何选取，算法总是能够得到与经典 k 均值算法相一致的结果，总的来说使得权值的选取难度降低。第 3 步的第(i)子步保证了每个类的下近似至少有一个元素，从而避免了聚类下近似为空集造成式(1)中分母为 0 的问题。第(iii)子步构造集合 T 时使用相对阈值 ζ 判别两个距离值的相对比值，ζ 的选取与具体的数据对象无关，只与它们的比值有关。综上所述，算法较好地解决了上一节中提出的粗糙 k 均值算法的问题。同时，RSC 继承了现有谱聚类算法的优良特性，包括谱聚类算法的数据维数无关性、数据全局结构无关性等等，这些优点都是 k 均值算法以及粗糙 k 均值算法所不具有的。

3　实验与分析

为了评估 RSC 算法我们首先使用一组人工数据集考察算法的有效性和稳定性。然后使用 3 组 UCI 基准数据集[10]考察算法的有效性和准确率。

3.1　主要评价指标

本文中我们用两种评价指标评价 RSC 以及相关算法的性能：Davies-Bouldin 指标和 Rand 指标。

(1) Davies-Bouldin 指标

Davies-Bouldin 指标(简称 DB 指标)定义为类内平均距离与类间平均间距的比值。我们用欧式距离作为距离量度标准，采用沃德法描述对象间的连接关系。令$\{x_1,\cdots,x_{|C_i|}\}$代表类 C_i 内的元素，则 C_i 内元素的类内平均距离表示为 $S(C_i)=\sum_{x_i\in C_i}\|x_i-m_i\|^2/|C_i|$，其中 m_i 代表类 C_i 的均值向量，$i=1,\cdots,k$。类 C_i,C_j 间的类间平均间距表示为 $d(C_i,C_j)=\sum_{x_i\in C_i,x_j\in C_j}\|x_i-x_j\|/(|C_i||C_j|)$，其中 $i,j=1,\cdots,k,i\neq j$。则 DB 定义为下式：

$$DB=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left\{\frac{S(C_i)+S(C_j)}{d(C_i,C_j)}\right\}\tag{7}$$

在粗糙聚类中，类 $C_i(i=1,\cdots,k)$被扩展为$\underline{C_i}$和$\overline{C_i}$，类似于算法 2 中计算类的均值向量的方法，我们把 DB 的定义扩展到粗糙集理论中，则有：

$$S_r(C_i)=w_l\frac{\sum_{x_i\in\underline{C_i}}\|x_i-m_i\|^2}{|\underline{C_i}|}+w_u\frac{\sum_{x_j\in\overline{C_i}}\|x_j-m_i\|^2}{|\overline{C_i}|}\tag{8}$$

其中 $w_l+w_u=1,i=1,\cdots,k$。

分析式(8)，当$\overline{C_i}=\underline{C_i}=C_i$ 时，

$$S_r(C_i)=w_l\sum_{x_i\in C_i}\|x_i-m_i\|^2/|C_i|+w_u\sum_{x_j\in C_i}\|x_j-m_i\|^2/|C_i|=(w_l+w_u)\sum_{x_i\in C_i}\|x_i-m_i\|^2/|C_i|$$

根据新的类内平均距离定义我们给出 DB 的粗糙集扩展形式：

$$DB=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left\{\frac{S_r(C_i)+S_r(C_j)}{d(C_i,C_j)}\right\}\tag{9}$$

根据聚类的定义，同一类中的两个数据对象之间应具有较小的距离，而不同类的两个数据对象之间应有较大的距离。故 DB 的值越小，聚类效果越好。

(2) Rand 指标

在已知分类结果的条件下，最常用的评价聚类准确率的指标是 Rand 指标，该指标视聚类结果为对每一成对点是否来自同一类所作出的判断。Rand 指标定义为：

$$Rand=\frac{\text{正确的判别对数}}{\text{总的判别对数}}\times100\%\tag{10}$$

3.2　人工数据实验

随机生成一组数据，数据集中包含两个分散开的服从高斯分布的点集，并包含了一些干扰点。数据对象的维数为 2，对象个数为 100，聚类个数为 2。

考虑下近似的定义，即确定属于某一类的数据对象集合，这就决定了，一般情况下，下近似对于计算类均值向量的影响较大，所以应取 $w_l>w_u$。这里我们分别取 $w_l=0.9,0.8,0.7,0.6,(w_u=1-w_l)$进行 4 组实验。对于阈值 ζ 的选取，它决定了类的边界区域(即上近似和下近似的差集)的大小，当 $\zeta=1$ 时，下近似与上近似相等，退化到经典 k 均值算法，所以应取 $\zeta>1$。考虑实验的实际情况，在 4 组实验中，每组取等间隔的 50 个阈值 $\zeta,\zeta\in(1,2]$。粗糙 k 均值算法(RKM)和粗糙谱聚类算法(RSC)结果的 DB 指标值随阈值 ζ 的变化情况

如图 1 所示。

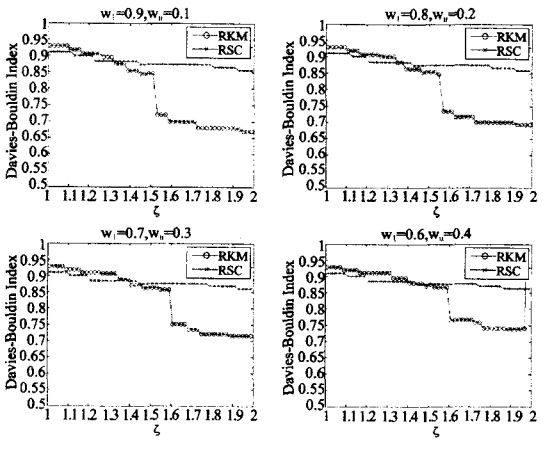

图 1　RKM 和 RSC 算法 DB 指标对比

从图中我们可以观察到，RKM 算法的 DB 指标值在阈值超过一定范围时会发生跃变或振荡现象。相比之下，RSC 算法呈平稳下降趋势，不会发生跃变或振荡。这说明 RSC 算法受阈值影响的敏感程度低于 RKM 算法，体现了较好的稳定性。另外，在 $\zeta\in(1,1.3]$范围内，RSC 算法得出结果的 DB 指标值普遍低于 RKM 算法，这表明在这一范围内选取阈值，得到较好聚类效果的可能较大，RSC 算法有效地降低了聚类结果的 DB 指标值。

3.3　UCI 基准数据集实验

我们取 UCI 机器学习数据集中 Wine，Iris 和 Ionosphere 3 个数据集进行实验，数据集的基本信息如表 1 所列。

表 1　UCI 数据集基本信息

数据集	对象个数	维数	类数
Wine	178	13	3
Iris	150	4	3
Ionosphere	351	34	2

对于每个数据集，我们分别用 k 均值算法(KM)、粗糙 k 均值算法(RKM)、谱聚类算法(SC)、粗糙谱聚类算法(RSC)进行聚类。采用 20 次随机实验取平均值的方法统计实验结果。各数据集上各算法 DB 指标值和 Rand 指标值比较如图 2、图 3 所示。

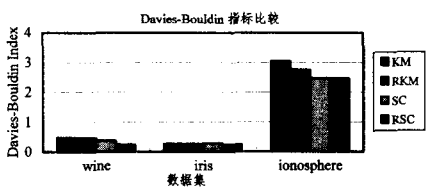

图 2　KM，RKM，SC，RSC 算法 DB 指标对比

DB 指标方面。对于 Wine 数据集，KM 和 RKM 的结果基本持平，SC 算法已经比较明显地降低了聚类结果的 DB 指标值，RSC 算法较 SC 算法又有一定程度的降低。说明在该数据集上谱聚类较传统 k 均值及粗糙 k 均值算法，DB 指标值降低较为明显。对于 Iris 数据集，KM 和 RKM 继续保持基本持平，SC 算法的 DB 指标值略微升高，不过经过粗糙集扩展，DB 指标值又有一定程度的降低。可以看出 Iris 经谱聚类算法降维之后 DB 指标值降低不明显，但经粗糙集扩展后，DB 值有了一定程度的降低。对于 Ionosphere 数据集，RKM 算法较 KM 算法 DB 值明显降低，SC 算法相对 RKM 算法又有了一次比较明显的降低。RSC 和 SC 相比有一定程度的降低。

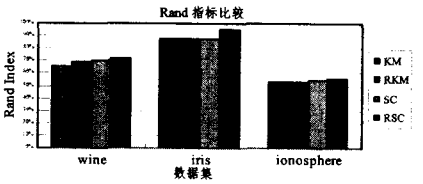

图 3　KM，RKM，SC，RSC 算法 Rand 指标对比

Rand 指标方面。对于 Wine 数据集，KM，RKM，SC，RSC 的 Rand 值依次升高。对于 Iris 数据集，KM，RKM，SC 的 Rand 值基本持平，RSC 有明显提升。对于 Ionosphere 数据集，KM 和 RKM 保持持平，SC 和 RSC 各有一定程度的提升。

综上所述，在 3 个基准测试集上，RSC 算法在 DB 指标和 Rand 指标上都有较好的表现。

结束语　本文将现有的谱聚类算法进行了扩展，提出了一种基于粗糙集理论的谱聚类算法。实验表明，该算法在稳定性和结果的准确率上都有超出现有算法的表现。本文中使用了高斯相似度函数构造数据对象的相似度矩阵，今后我们将进一步通过实验讨论相似矩阵的构造方法问题。谱聚类算法与对象的维数无关，适用于处理高维数据，经过粗糙集扩展后的谱聚类算法的结果具有层次性，可应用于 Web 文本数据挖掘、图像检索等领域。

参 考 文 献

[1] 孙吉贵，刘杰，赵连宇. 聚类算法研究 [J]. 软件学报，2008，19(1)：48-61

[2] Bach R，Jordan M I. Learning spectral clustering [R]. UCB/CSD-03-1249. University of California at Berkeley，2003

[3] Hagen L，Kahng A B. New spectral methods for ratio cut partitioning and clustering [J]. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems，1992，11(9)：1074-1085

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

[5] Ding C H Q，He X，Zha H，et al. A min-max cut algorithm for graph partitioning and data clustering [C]// Cercone N，Lin T Y，Wu X，eds. ICDM 2001. Los Alamitos，California：IEEE Computer Society，2001：107-114

[6] Lingras P，West C. Interval set clustering of web users with rough k-means [J]. Journal of Intelligence Information Systems，2004，23(1)：5-16

[7] Gu M，Zha H，Ding C，et al. Spectral relaxation models and structure analysis for k-way graph clustering and bi-clustering [R]. CSE-01-007. Penn State University，2001

[8] Ng A Y，Jordan M I，Weiss Y. On spectral clustering：Analysis and an algorithm [C]// Dietterich T G，Becker S，Ghahramani Z，eds. Advances in Neural Information Processing Systems 14. Cambridge，MA：MIT Press，2002：849-856

[9] Peters G. Some refinements of rough k-means clustering [J]. Pattern Recognition，2006，39：1481-1491

[10] UC Irvine Machine Learning Repository [DB/OL]. URL：http://archive.ics.uci.edu/ml/