基于本征间隙与正交特征向量的自动谱聚类

孔万增[1],孙志海[1],杨 灿[2],戴国骏[1],孙昌思核[1]
(1. 杭州电子科技大学计算机学院,浙江杭州 310018; 2. 香港科技大学电子及计算机工程学系,香港九龙)

摘 要: 针对经典谱聚类算法无法自动确定数据类个数的问题,本文提出了一种基于本征间隙与正交特征向量的自动谱聚类算法.该方法利用样本数据构建亲和度矩阵,然后进行谱分解得到相应的特征值和特征向量,对特征值从大至小依次排序,用本征间隙来刻画相邻特征值之间的差,通过第一个极大本征间隙出现的位置来自动确定类个数,最后以特征向量之间的夹角作为相似度和已获得的类个数相结合来实现数据分类.本文算法的正确性在人造数据库上得到了验证,并在 UCI 数据库上与 k-means、FCM、Jordan 算法进行了分类准确性比较实验,结果表明本文方法比其他三种方法的分类准确率更高.

关键词: 谱聚类; 亲和度矩阵; 本征间隙; 类个数; 正交特征向量
中图分类号: TP181 **文献标识码:** A **文章编号:** 0372-2112 (2010) 08-1880-06

Automatic Spectral Clustering Based on Eigengap and Orthogonal Eigenvector

KONG Wan-zeng[1], SUN Zhi-hai[1], YANG Can[2], DAI Guo-jun[1], SUN Changsihe[1]
(1. *College of Computer Science, Hangzhou Dianzi University, Hangzhou, Zhenjiang* 310018, *China*;
2. *Department of Electronic and Computer Engineering, HKUST, Kowloon, Hongkong*)

Abstract: To deal with the problem that classical spectral clustering methods can not automatically determine the number of class. A new algorithm called automatic spectral clustering(ASC) based on eigengap and orthogonal eigenvector was presented in this paper. The proposed method first constructed the affinity matrix of data, and gained series of eigenvalues and eigenvectors through spectral decomposition. Second, ordered the eigenvalues and used the first maximum eigengap to determine the number of classes. The data was classified by the class number and the angle between two eigenvectors as similarity. The effectiveness of the proposed algorithm was verified on artificial data, and was compared with k-means, FCM and Jordan algorithm on UCI database. The experiment results demonstrate that the proposed method ASC outperforms other three methods in respect of classification accuracy.

Key words: spectral clustering; affinity matrix; eigengap; class number; orthogonal eigenvector

1 引言

在模式识别研究中,聚类算法是一种进行数据分组的有效方法.目前聚类算法被广泛应用于数据挖掘、数字图像处理、信息检索等领域.比较常见的聚类算法有:模糊均值聚类[1]、层次聚类[2]等.这些聚类算法,它们在使用时一般需要指定聚类个数,聚类结果往往对所假定的聚类个数非常敏感,聚类算法鲁棒性也很差.尽管山峰聚类[3]、减法聚类[4~5]无需指定类个数能自动结束分类过程,但在面对非凸球形数据分布时,算法失效.同时,上述算法有一个共同点,就是都会产生聚类中心的结果,聚类中心又被称为数据集代表点,但是当某些数据集(如环形数据)的聚类中心缺失或不明显时,上述算法的聚类效果就很不理想.另外,k 均值、减法聚类等传统聚类算法是建立在凸球形的数据样本空间上的,若数据样本空间为非凸时,算法就会陷入局部最优,为了能在任意的数据样本空间进行正确聚类,且能收敛于全局最优解,因此有研究者提出了谱聚类(Spectral clustering)[6,7]算法.它在解决非块状、非凸球形数据的聚类问题时有着十分出色的表现.谱聚类算法的核心思想是:转换数据聚类的特征空间,在新选择的特征空间对数据运用 k 均值等方法进行聚类,聚类的结果最后映射回原数据空间.目前流行的谱聚类算法仍旧需要输入数据类个数,但在很多现实的应用场景中比如目标定位、图像分割、文本挖掘等,类个数对计算机来说还是未知的.因此,设计某种聚类算法能根据数据自身信息自动确定类别数并完成数据聚类具有重要意义.

2 谱聚类基本理论

谱聚类算法最早是从谱图划分理论演化而来.设每一个样本数据看成是图中的顶点 V,根据样本两两之间的相似度将相应顶点之间的连接边 E 赋权重 W,于是

收稿日期:2009-10-20;修回日期:2010-04-29
基金项目:国家自然科学基金(No.60803126);浙江省重大科技专项(No.2008C11108-1);浙江省国际合作项目(No.2009C14013)

就得到基于样本的相似度的无向加权图 G=(V,E,W).从图论的最优划分理论来看,类似于线性判别算法,就是使划分成的若干子图之间相似度最小,子图内部相似度最大[8].常用的划分准则有最小割集、规范割集、平均割集、比例割集以及最小最大割集等准则[9~11].图划分问题的最优解是一 *NP* 难题,一个较好的求解方法是考虑问题的连续松弛形式,于是便可将原问题转换成相似度矩阵或 *Laplacian* 矩阵的谱分解.

相似度矩阵又称亲和度矩阵,常用 $\boldsymbol{A}$ 或 $\boldsymbol{W}$ 表示,该矩阵元素定义为:

$$\boldsymbol{A}_{ij}=\exp\left(-\frac{d^2(v_i,v_j)}{2\sigma^2}\right) \tag{1}$$

其中 v_i 是数据样本点,$d(v_i,v_j)$是两样本点之间的距离,一般取欧式距离 $\|v_i-v_j\|$,也可以是其他形式的距离定义,σ 是尺度参数,决定了样本点之间的衰减速度.

将相似度矩阵的每行元素相加,得到该样本点的度,体现了该点周围数据的分布状况.以所有度值为对角元素构成的对角矩阵为度矩阵,常用 $\boldsymbol{D}$ 表示,

$$\boldsymbol{D}_{ii}=\sum_j \boldsymbol{A}_{ij} \tag{2}$$

规范相似度矩阵一般形式为

$$\boldsymbol{A}_{nor}=\boldsymbol{D}^{-1/2}\,\boldsymbol{A}\,\boldsymbol{D}^{-1/2} \tag{3}$$

为了更好地理解规范化相似度矩阵 $\boldsymbol{A}_{nor}$,可根据式(3)将 $\boldsymbol{A}_{nor}$每个元素表示为:

$$\boldsymbol{A}_{nor}(i,j)=\frac{\boldsymbol{A}(i,j)}{\sqrt{\boldsymbol{D}(i,i)}\sqrt{\boldsymbol{D}(j,j)}} \tag{4}$$

在不同谱聚类算法中,谱分解方程有所不同,但分解矩阵形式上都类似式(3),有的文章如最小割集(Mcut)算法[12]就把式(3)中的亲和度矩阵 $\boldsymbol{A}$ 用 Laplacian 矩阵 $\boldsymbol{L}$ 代替. Laplacian 矩阵分为非规范 Laplacian 矩阵和规范 Laplacian 矩阵.非规范 Laplacian 矩阵表示为 $\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{A}$.规范 Laplacian 矩阵两种形式分别是

$$\boldsymbol{L}_{nor1}=\boldsymbol{D}^{-1/2}\,\boldsymbol{L}\,\boldsymbol{D}^{-1/2}=\boldsymbol{I}-\boldsymbol{D}^{-1/2}\,\boldsymbol{A}\,\boldsymbol{D}^{-1/2} \tag{5}$$

以及

$$\boldsymbol{L}_{nor2}=\boldsymbol{D}^{-1}\,\boldsymbol{L}=\boldsymbol{I}-\boldsymbol{D}^{-1}\,\boldsymbol{A} \tag{6}$$

Ng,Jordan[7]等人就是用式(5)Laplacian 规范矩阵 L_{nor1}做特征值分解,然后选取前 k 个最大特征值对应的特征向量构成$\mathbb{R}^k$ 空间,并使$\mathbb{R}^k$空间中的每一点与原数据一一对应,在$\mathbb{R}^k$空间应用 k 均值等经典聚类算法聚类. Meila[13]将相似性解释为 Markov 链中的随机游走,以式(6)中 $\boldsymbol{D}^{-1}\,\boldsymbol{A}$ 为概率转移矩阵 $\boldsymbol{P}$,同时根据随机游走对最小割集(Mcut)进行概率解释,然后对概率转移矩阵进行谱分解,其后续的步骤与 Jordan 算法基本一致.除了对谱分解矩阵构造进行研究外,国内焦李成等人对相似性度量的选择进行研究,提出了一种基于数据依赖的密度敏感的谱聚类方法处理多尺度聚类问题[14];王娜等人提出了一种主动式半监督的谱聚类算法[15],使得类内各点紧凑,类间散布.

谱聚类根据不同的准则函数有不同具体实现方法,但总的流程可以归纳为以下三个步骤:

Step1　构建能表示数据样本集关系属性的矩阵 $\boldsymbol{Z}$.

Step2　计算 $\boldsymbol{Z}$ 的前 k 个特征值与特征向量,构建新的数据特征空间.

Step3　利用 k-means 或 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　自动谱聚类算法

上述的谱聚类方法需要输入聚类个数参数 k,在谱分解后,在特征向量空间又需运用 FCM 或 k 均值等经典聚类算法.一方面在实际应用中参数 k 很难预先确定;另一方面,在转化后的新特征空间里运用 FCM 等经典算法其有效性本身值得商榷,而且二次聚类使算法显得臃肿.因此本文提出了一种自动谱聚类算法.该算法运用本征间隙来自动确定类个数,直接用正交特征向量来实现数据分类.

3.1　本征间隙估计类个数

本文拟从数据的亲和度矩阵中充分挖掘信息,提炼出数据类的个数.对于存在 k 个理想的彼此分离簇的有限数据集且数据按类依次排序,可以证明规范化亲和度矩阵的前 k 个最大特征值为 1,第 $k+1$ 个特征值则严格小于 1,二者之间的差距取决于这 k 个聚类的分布情况[16].k 个理想的彼此分离簇的数据集体现在亲和度矩阵上,就是该阵对角线上分布 k 个全 1 分块矩阵,其余位置都为 0.但在实际情况中,很难构造这样的亲和度矩阵.实际的亲和度矩阵对角线上的分块矩阵元素不全为 1,准确的说是全 1 矩阵加上一个负的扰动量;对角线分块矩阵外的元素也不全为 0,而是一个正的扰动量.根据矩阵摄动理论,将上述的第 k 个和第 $k+1$ 个特征值之间的差称之为本征间隙(Eigengap),本征间隙越大,选取的 k 个特征向量所构成的子空间就越稳定[17].对于普通分布(理想分类加扰动情形)的聚类数据,通过寻找本征间隙序列的第一个极大值便可确定原数据类的个数.

计算规范化亲和度矩阵 $\boldsymbol{A}_{nor}$的特征值并按顺序从大到小排列 $\lambda_1\geqslant\lambda_2\geqslant\cdots\lambda_n$,计算本征间隙序列 $\{g_1,g_2\cdots,g_{n-1}\mid g_i=\lambda_i-\lambda_{i+1}\}$.在本征间隙序列中依次寻找第一个极大值,则该值对应的下标即为类个数.即类别数 k 为

$$k=\arg\min_i\{g_i-g_{j\mid j<i}>0\,\&\,g_i-g_{i+1}>0\} \tag{7}$$

有了数据类的信息,很多经典聚类算法就可以对原数据进行分类.但是如果再引入其他的聚类算法就使算

法一方面显得不精简,令一方面时间消耗也增大.因此本文在类数目估计的基础上,直接从特征向量着手实施数据分类.

3.2 基于正交特征向量的数据分类

首先我们对理想情况的亲和度矩阵进行分析.假设对象数据 $\boldsymbol{V}=\{v_1,v_2\cdots v_n\}$ 有 k 个类,数据元素按顺序属于每一类进行排列即,

$$\underbrace{\{v_1,v_2\cdots v_{n_1}\}}_{n_1}\in\boldsymbol{V}_1,\underbrace{\{v_{n_1+1},\cdots v_{n_1+n_2}\}}_{n_2}\in\boldsymbol{V}_2,\cdots\underbrace{\{v_{n_{k-1}+1},\cdots v_n\}}_{n_k}$$

$\in\boldsymbol{V}_k$ 其中 $\boldsymbol{V}=\bigcup_{i=1}^{k}\boldsymbol{V}_i,\boldsymbol{V}_i\cap\boldsymbol{V}_j=\phi(i\neq j)$.

亲和度矩阵则为

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{A}_{n_1}&0&\cdots&0\\0&\boldsymbol{A}_{n_2}&\cdots&0\\\vdots&\cdots&\ddots&\vdots\\0&0&\cdots&\boldsymbol{A}_{n_k}\end{bmatrix}\tag{8}$$

若在理想情况下,亲和度定义为下式

$$A_{ij}=\begin{cases}1, & v_i\in\boldsymbol{V}_t\wedge v_j\in\boldsymbol{V}_t\\0, & v_i\in\boldsymbol{V}_t\wedge v_j\notin\boldsymbol{V}_t\end{cases}\tag{9}$$

因此在理想情况下,式(8)中 $\boldsymbol{A}_{n_i}=\boldsymbol{E}_{n_i}\in\mathbb{R}^{n_i\times n_i}$ 为全 1 矩阵.有矩阵理论的知识可知 $\lambda_1\geqslant\lambda_2\geqslant\cdots\lambda_n$ 为 $\boldsymbol{E}_{n_i}$ 的特征值,则 $\lambda_1=n_i,\lambda_2=\lambda_3=\cdots=\lambda_{n_i}=0$,且 λ_1 对应的特征向量为 $\boldsymbol{x}_1=\boldsymbol{L}(\boldsymbol{e}),\boldsymbol{e}=[1,1,\cdots1]^{\mathrm{T}}\in\mathbb{R}^{n_i}$.理想情形下,相似度函数以式(9)为前提,使得 $\boldsymbol{A}=\mathrm{diag}(\boldsymbol{E}_{n_1},\boldsymbol{E}_{n_2}\cdots\boldsymbol{E}_{n_k})$,约束条件很强.但在实际应用中数据还有待分类,类信息未知,不能采用式(9)而采用公式(1).现在把约束条件局部弱化,把 $\boldsymbol{A}_{n_i}$ 全 1 矩阵弱化为数值$\leqslant$1(接近1)的实对称矩阵,即 $\boldsymbol{A}=\mathrm{diag}(\boldsymbol{A}_{n_1},\boldsymbol{A}_{n_2}\cdots\boldsymbol{A}_{n_k})$.则有定理[16]表明以下结论:$\lambda(\boldsymbol{A})=\bigcup_{i=1}^{k}\lambda(\boldsymbol{A}_{n_i})$,其中 $\lambda(\boldsymbol{A})$是矩阵 $\boldsymbol{A}$ 的所有特征值集合,且如果 $\boldsymbol{x}_i\in\mathbb{R}^{n_i}$是矩阵$\boldsymbol{A}_{n_i}$对应于特征值 λ 的特征向量,则

$$\underbrace{(0,0\cdots0}_{n_1},\underbrace{0,0\cdots0}_{n_2},\cdots x_i^{\mathrm{T}}\cdots\underbrace{0,0\cdots0)^{\mathrm{T}}}_{n_k}$$

也是矩阵 $\boldsymbol{A}$ 对应于特征值λ 的特征向量.因此由矩阵 $\boldsymbol{A}$ 前 k 个特征向量构造的特征矩阵可表达为

$$\boldsymbol{X}=\begin{bmatrix}\boldsymbol{x}_1&0&\cdots&0&0\\0&\boldsymbol{x}_2&0&\cdots&0\\\vdots&0&\ddots&\vdots&\vdots\\0&\vdots&\cdots&\boldsymbol{x}_{k-1}&0\\0&0&\cdots&0&\boldsymbol{x}_k\end{bmatrix}\tag{10}$$

其中 $\boldsymbol{X}\in\mathbb{R}^{n\times k}$,令 $\boldsymbol{\alpha}_i$ 是 $\boldsymbol{X}$ 的第 i 个行向量,对应原空间的数据点 $v_i,\boldsymbol{X}=[\boldsymbol{\alpha}_1^{\mathrm{T}},\boldsymbol{\alpha}_2^{\mathrm{T}}\cdots\boldsymbol{\alpha}_n^{\mathrm{T}}]^{\mathrm{T}}$,从公式(10)可以明显看出有这样的性质:当 v_i,v_j 属于同一类时,$\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j$ 线性相关;当 v_i,v_j 不属于同一类时,$\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j$ 正交.即

$$\cos(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j)=\frac{|\boldsymbol{\alpha}_i^{\mathrm{T}}\boldsymbol{\alpha}_j|}{\boldsymbol{\alpha}_i\boldsymbol{\alpha}_j}=\begin{cases}1, & v_i\in\boldsymbol{V}_t\wedge v_j\in\boldsymbol{V}_t\\0, & v_i\in\boldsymbol{V}_t\wedge v_j\notin\boldsymbol{V}_t\end{cases}\tag{11}$$

它们类与类之间正交地分布于 k 维空间上,并且该 k 维空间形成的这 k 个聚类对应着原空间中所有点形成的 k 个聚类.由此,可通过特征矩阵的行向量之间夹角的余弦值来判断原数据的类属性.以上亲和度矩阵的约束条件虽然比理想情形有所弱化,即放松了类内点的约束,但仍维持类间点约束即类间点的相似度为零.实际应用中,亲和度矩阵可看成是公式(8)加上一对称扰动阵 $\boldsymbol{\varepsilon}\in\mathbb{R}^{n\times n}$ 即

$$\boldsymbol{A}=\mathrm{diag}(\boldsymbol{A}_{n_1},\boldsymbol{A}_{n_2}\cdots\boldsymbol{A}_{n_k})+\boldsymbol{\varepsilon}\tag{12}$$

其谱分解的特征矩阵也要增加一相应的扰动阵.文献[16]已以数值计算形式说明扰动约束在一定范围内,当特征矩阵的行向量夹角的余弦值 $\cos(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j)>0.3877$ 时,就认为数据点 v_i,v_j 就属于同一类,同时文献也表明当扰动更小时,$\cos(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j)$的阈值可适当增大,本文选取 0.3877 为阈值.

以上结论是在数据顺序属于每一类进行排列的前提下进行讨论的,但数据可能是任意排列的.对于任意排列的数据 $\boldsymbol{V}$,由线性代数的知识可知总存在一个置换 $\boldsymbol{h}$,使得 $\boldsymbol{hV}$ 按顺序属于每一类排列.同理也存在一系列置换矩阵 $\boldsymbol{H}_1,\boldsymbol{H}_2,\cdots\boldsymbol{H}_m$,使得 $\boldsymbol{H}_m\cdots\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{A}\boldsymbol{H}_1\boldsymbol{H}_2\cdots\boldsymbol{H}_m$ 成为式(12)的矩阵.因此数据是否按顺序属于每一类排列对上述结论没有影响.下面讨论数据的分类问题.本文提出了一种矩阵赋零方法来实施分类.其思想步骤如下:

Step1 构建特征矩阵行向量的相似度矩阵 $\boldsymbol{B}$,

$$B(i,j)=\begin{cases}1, & \cos(\boldsymbol{\alpha}_i,\boldsymbol{\alpha}_j)>0.3877\\0, & \text{otherwise}\end{cases};$$

Step2 从 $\boldsymbol{B}$ 阵的第 1 行开始检索数值为 1 的项,并记录它们的序号,对应序号的原数据归一类,然后在 $\boldsymbol{B}$ 阵中相应行的数值赋为 0.

Step3 检索 $\boldsymbol{B}$ 阵中下一个不全为 0 的行中数值为 1 的项,并记录它们的序号,对应的原数据归为一类,同时在 $\boldsymbol{B}$ 阵中相应行的数值赋为 0.

Step4 若 $\boldsymbol{B}$ 阵不全为 0,重复 Step3;否则算法结束.

3.3 自动谱聚类算法小结

本文与以往谱聚类算法如 Jordan 算法[7],MS 算法[13]的区别就在于不仅能自动估计类数目,而且在特征向量空间不再采用其他经典聚类算法,而采用上一节中的矩阵赋零方法直接在矩阵 $\boldsymbol{B}$ 中完成分类,现把本文的自动谱聚类算法步骤归纳如下:

Step1 利用式(1)(3)构造数据点的规范化亲和度矩阵 $\boldsymbol{A}_{nor}$.

Step2　计算规范化亲和度矩阵 $\boldsymbol{A}_{nor}$的顺序本征间隙序列,并利用式(7)确定类数目 k.

Step3　利用 $\boldsymbol{A}_{nor}$阵前 k 个最大特征值对应的特征向量构建特征空间,计算其行向量的相似度矩阵 $\boldsymbol{B}$.

Step4　利用上节的矩阵赋零方法对数据进行分类,算法结束.

4　实验及分析

为了验证自动谱聚类算法的正确性以及其有效的分类性能.本文分别在人造数据以及 UCI 标准数据中进行了相关实验.

4.1　算法验证

4.1.1　理想分类数据实验

本文首先在人造理想分类数据上进行试验.理想分类数据由计算机随机产生,共 210 个样本,分为七类,如图 1 所示.

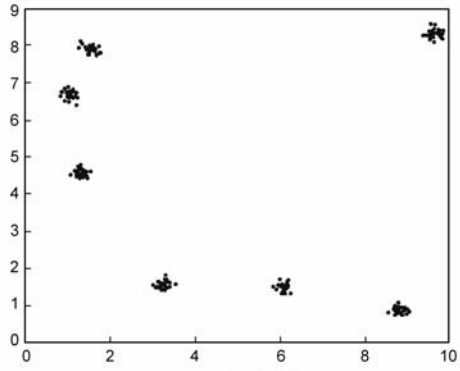
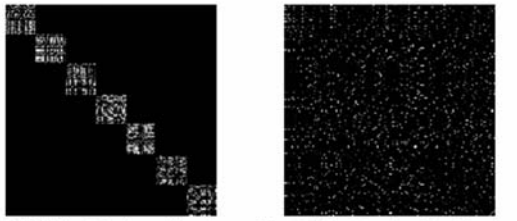


图1　人造7类数据

实验时数据按顺序属于每一类进行排列.由于数据是类顺序排列,因此它的规范化亲和度矩阵 $\boldsymbol{A}_{nor}$呈明显的对角分块矩阵,如图 2(a)所示;若数据不是按类顺序而随机排列,则 $\boldsymbol{A}_{nor}$阵看上去杂乱无章如图 2(b),但不影响算法的聚类结果.图 3 显示了规范化亲和度矩阵的从大到小排列的特征值及其本征向量序列.由于每个数据类被理想的分开,因此 $\boldsymbol{A}_{nor}$阵的前 7 个最大特征值为 1,在本征间隙序列中横坐标 7 对应的位置是该序列的第一个极大值,如图 3 所示.从图 3 中发现当数据理想分类时,本征间隙序列的第一个极大值恰好也是该序列的最大值.图 4 显示了 $\boldsymbol{A}_{nor}$阵的 7 个最大特征值对应的特征向量.图 4 的这七个特征向量互相正交,是



(*a*) Affinity Matrix with order　(b) Affinity Matrix with disorder

图2　亲和度矩阵

一组正交基,这也就说明由前 k 个特征向量构建的特征空间中能进行最佳分类的本质原因.如果亲和度是按式(9)定义,则 $\boldsymbol{A}_{nor}=\mathrm{diag}(\boldsymbol{E}_{n_1},\boldsymbol{E}_{n_2}\cdots\boldsymbol{E}_{n_k})$,即每个子类矩阵的值都相同为 1,则它的特征向量的非零部分则是一条直线.而在本文实验中,亲和度按式(1)定义,实验中数据类之间分离的较好,因此 $\boldsymbol{A}_{nor}=\mathrm{diag}(\boldsymbol{A}_{n_1},\boldsymbol{A}_{n_2}\cdots\boldsymbol{A}_{n_k})$,即每个子类矩阵的值都不同,但接近 1,所以其特征向量的非零部分是波浪线.采用本文的自动谱聚类算法对该合成的七类数据有很好的分类性能,结果如图 5 所示,以不同的形状代表一类.针对数据随机排列,如文中 2.2 节所述,自动谱聚类算法对其聚类结果与按顺序属于每一类排列的数据聚类结果完全一致.

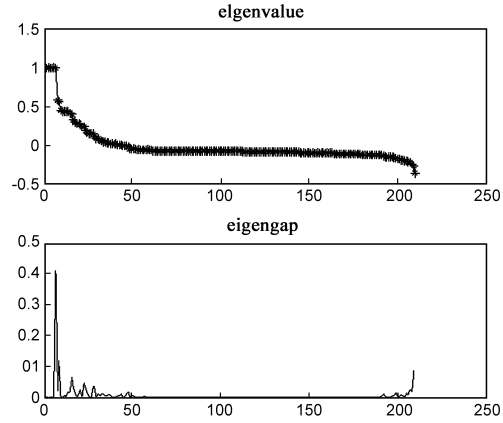


图3　特征值及其本征间隙序列

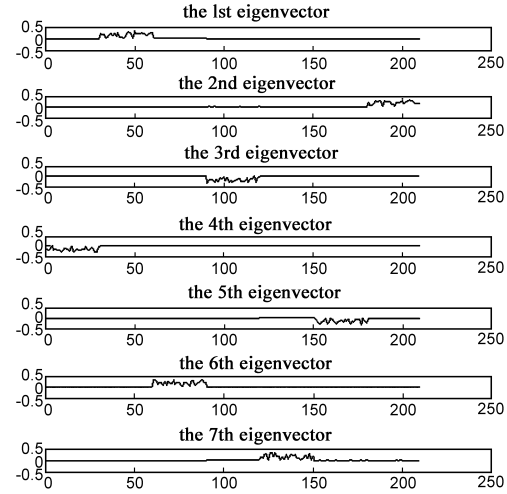


图4　规范化亲和度矩阵的前7个特征向量

4.1.2　非线性数据实验

4.1.1 节利用理想分类数据验证了自动谱聚类算法的正确性.这一节中,我们将对一些复杂数据进行分类.传统聚类算法往往使用了类中心的概念,但有些数据类中心不明显,如图 6 所示的三环数据,传统算法 FCM 的聚类效果就十分不理想,如图 6 所示.采用公式(1)作为亲和度,按类顺序排列以及随机排列的亲和度

矩阵如图 8 所示.由于数据是环形分布,不是块状分布,再加上欧式距离的因素,在类顺序排列的亲和度矩阵上呈现出一条狭细的对角块.从图 9 本征间隙序列来看,第一个极大值点出现在横坐标为 3 的位置,但该位置仅是极大值,远不是最大值.这也从实验角度验证了由第一个极大值确定聚类个数的原因.另外实验展示了环形数据的前三特征向量,特征向量的每个行向量表示为三维空间的一个点.从图 10 可知,三维空间中的点明显呈 3 条线状分布,这说明同类的点是线性相关,但由于 $\boldsymbol{A}_{n_i}$ 阵不是严格意义的 $\boldsymbol{E}_{n_i}$,使得不同类的行向量也不能严格正交.最后采用本文自动谱聚类算法后,三环数据聚类准确无误,结果如图 7 所示.

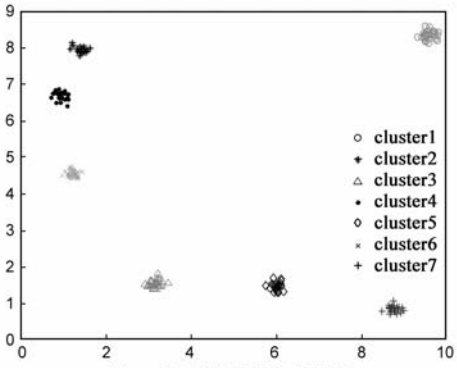


图5 自动谱聚类分类结果

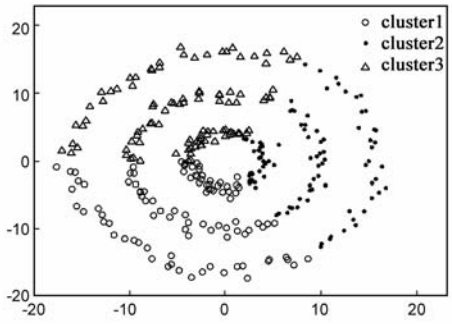


图6 FCM 3环数据聚类结果

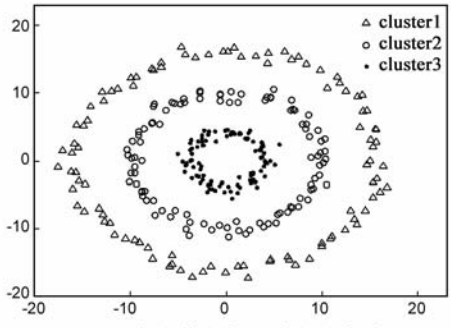
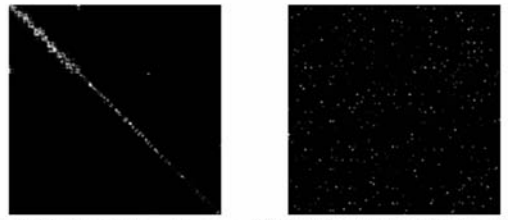


图7 自动谱聚类3环数据聚类结果

4.2 UCI 数据实验

本文在国际通用数据库 UCI 数据库(专门用于测试分类、聚类算法)中的 Satimage、Iris、Ionesphere 及 New-thyroid 四个数据集进行了四种聚类算法的比较实验,其


(*a*) Affinity Matrix with order (*b*) Affinity Matrix with disorder
图8 亲和度矩阵

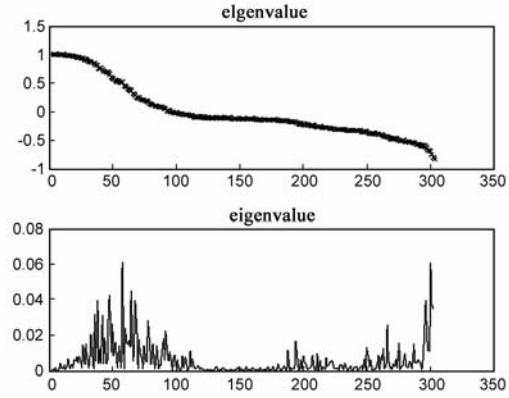


图9 特征值及本征间隙序列

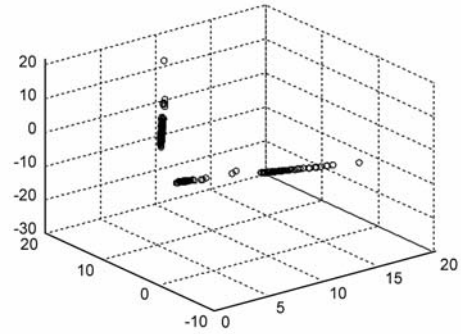


图10 特征向量3维空间分布

中 Satimage 原数据为 6 类共 4435 个数据样本点,这 6 个类的数据个数分别为(1072,479,961,415,470,1038),本文按照比例在各个类中分别随机抽取(107,48,96,42,47,104)共 444 个样本点.本文自动谱聚类算法(ASC)与 k-means,FCM 以及谱聚类中的经典算法 Jordan 算法进行了实验比较.实验之前首先对 UCI 四个数据集进行归一化处理,以便于 Jordan 算法和本文算法在亲和度函数中选取统一的尺度参数 σ.实验在计算机内存 1.5G、CPU 2.2GHz 的硬件条件以及 Matlab7.0 环境下运行.

表 1 显示了 UCI 四个实验数据集的基本属性以及四种算法在每个数据集上随机 10 次实验的平均分类准确率与耗时情况.由于 Jordan 算法与本文算法原理一致,均属于谱聚类方法,实验中发现尺度参数的选择与两算法的分类性能好、差体现较为一致,因此表 1 中参数 σ 是两种算法分类性能俱佳时的取值.Satimage 和 New-thyroid 数据维数较高,分别为 34 和 36 维,Iris、Ionesphere 数据维数较低.一般来说,维数高、类别多的数据

对准确分类是有一定的难度.在 36 维的 Satimage 数据集上,表 1 结果表明无论是 Jordan 算法还是本文的自动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.表 1 同时表明同一类型的算法,其实验性能也比较接近:k-means 与 FCM 同属样本最小均方差的方法,分类准确率比较接近;Jordan 算法和本文算法同属谱方法,分类准确率也较为接近.Jordan 算法和本文算法的差别在于前者需要事先确定类个数,后者能自动确定;前者在特征空间采用了 k 均值等分类算法,后者则运用特征向量的夹角作为相似度来直接分类.UCI 数据库实验结果显示本文算法在四个数据库上均要比 Jordan、FCM、k-means 算法优越.在计算时间花费上,k-means 与 FCM 均比谱聚类算法要快的多,原因是谱聚类在计算相似度矩阵和矩阵特征值分解耗时较多,而本文算法比 Jordan 算法耗时略多,主要由于自动确定类个数计算耗时所产生.实验也表明谱聚类在样本规模大、数据维数高的情况下,存在计算困难的问题.

表 1　不同聚类算法在 UCI 数据库的实验结果

数据集	数据属性			算法性能(分类准确率/算法耗时)				
	样本数	维数	固有类数	k-means(%/s)	FCM(%/s)	JORDAN(%/s) 预先指定	ASC(%/s) 自动确定	参数 σ
Satimage	444	36	6	74.26/0.0581	77.45/0.1314	86.04/3.774	86.49/4.3225	0.13
Iris	150	4	3	79.27/0.0095	82.60/0.0083	90.33/0.1611	92.00/0.1712	0.16
Ionesphere	351	34	2	71.23/0.0125	66.21/0.0262	72.08/1.4977	72.08/1.5947	0.20
New-thyroid	215	5	3	86.93/0.0115	85.26/0.0271	83.72/0.4415	89.30/0.4709	0.13

5　结论

本文在 Jordan 等经典谱聚类算法的基础上做了两方面改进.首先对分类数据建立亲和度矩阵并进行谱分解,利用本征间隙自动确定数据的类个数;其次利用确定的类个数和谱分解的特征向量之间的余弦值完成数据的分类.本文方法本质把谱分解矩阵看成是理想分类情况下的亲和度矩阵的对称扰动情形,文章理论分析了该方法的原理,同时在人造数据和非线性数据上对本文算法原理正确性进行了验证,最后在 UCI 数据库上与 k-means、FCM、Jordan 算法进行了分类准确性比较实验,结果表明该方法的有效性.但本文方法在数据分布扰动较大以及数据维数过高时会出现类估计错误和分类准确率低的现象,同时对大规模样本数据集,本文方法跟其他谱聚类算法一样存在计算困难的挑战.如何实现高维数据和大规模数据集的自动谱聚类是我们下一步的研究工作.

作者简介:

孔万增　男,讲师,中国计算机学会会员,美国 ACM 会员.1980 年 4 月出生于浙江杭州.2003 年和 2008 年在浙江大学获工学学士和工学博士学位.现为杭州电子科技大学计算机应用技术研究所教师,主要从事机器学习、认知计算等方面的研究工作.

E-mail:kongwanzeng@hdu.edu.cn

戴国骏　男,教授.中国计算机学会高级会员、IEEE 高级会员.1965 年 6 月出生于浙江湖州.1991 年和 1998 年在浙江大学获工学学士和工学博士学位.现为杭州电子科技大学计算机学院副院长,计算机应用技术研究所所长,主要从事认知计算、无线传感网络,计算机体系结构等方面的研究工作.

参考文献:

[1] J C Bezdek, R Ehrlich, W Full. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2 – 3): 191 – 203.

[2] A K Jain, M N Murty, P J Flynn. Data clustering: a review[J]. ACM Comput. Surveys. 1999, 31(3) 264 – 323.

[3] R Yager, D FileV. Approximate clustering via the mountain method[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1994, 24(8): 1279 – 1284.

[4] S Chiu. Fuzzy model identification based on cluster estimation [J]. Journal of Intelligent and Fuzzy Systems, 1994, 2(3): 267 – 278.

[5] W Kong, S Zhu. Multi-face detection based on downsampling and modified subtractive clustering for color images[J]. Journal of Zhejiang University SCIENCE A. 2007, 8(1): 72 – 78.

[6] N Cristianini, J S Taylor, J S Kandola. Spectral kernel methods for clustering[C]. In: NIPS, Cambridge, MIT Press, 2001: 649 – 655.

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

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

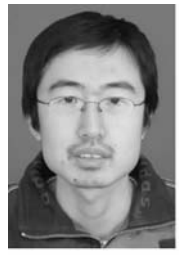
[9] 蔡晓妍,戴冠中,杨黎斌.谱聚类算法综述[J].计算机科学,2008,35(7):14 – 18.

(下转第 1891 页)

[17] Harb A M, Harb B A. Controlling chaos in Josephson junction using nonlinear backstepping controller[J]. IEEE Trans Appl Supercond, 2006, 16(4): 1988 – 1998.
[18] 高金峰. 非线性电路与混沌[M]. 北京: 科学出版社, 2006. 25 – 28, 130 – 131.
[19] Clark A. Hamilton. Josephson voltage standards[J]. Review of Scientific Instruments, 2000, 71(10): 3611 – 3623.
[20] S P Benz. Superconductor-normal-superconductor junctions for programmable voltage standards[J]. Appl Phys Lett, 1995, 67(18): 2714 – 2716.
[21] 王争, 岳宏卫, 周铁戈, 赵新杰, 何明, 谢清连, 方兰, 阎少林. $SrTiO_3$ 基片上 Tl-2212 双晶约瑟夫森结的动态特性及噪声影响研究[J]. 物理学报, 2009, 58(10): 7216 – 7221.
Wang Zheng, Yue Hong-wei, Zhou Tie-ge, et al. Dynamic characteristics of Tl-2212 bicrystal Josephson junctions on $SrTiO_3$ substrates and the effect of noise on it[J]. Acta Physica Sinica, 2009, 58(10): 7216 – 7221. (in Chinese)
[22] S J Berkowitz, W J Skocpol, P M Mankiewich, et al. Thermal noise in high temperature superconducting normal superconducting step-edge Josephson junctions[J]. J Appl Phys, 1994, 76(2): 1337 – 1339.
[23] Giovanni Filatrella, Boris A. Malomed, Sergio Pagano. Noise-induced dephasing of an ac-driven Josephson junction[J]. Physical Review E, 65, 05116.
[24] Ambegaokar V, Baratoff A. Voltage due to thermal noise in the dc Josephson effect[J]. Phys Rev Lett, 1969, 22(25): 1364 – 1366.
[25] K K Likharev, V K Semenov. RSFQ logic/memory family: A new Josephson-junction technology for sub-Terahertz-clock-frequency digital system[J]. IEEE Transactions on Applied Superconductivity, 1991, 7(2): 3176 – 3180.

作者简介:

樊　彬　男, 1985 年 7 月生于天津, 硕士, 研究方向为超导电子学.
E-mail: nankaifanbin@hotmail.com



周铁戈　男, 1980 年 2 月生于内蒙古, 博士, 副教授, 研究方向为超导电子技术、非线性电路和混沌保密通信等.
E-mail: zhoutg@nankai.edu.cn

(上接第 1885 页)

CAI Xiao-yan, DAI Guan-zhong, YANG Li-bin. Survey on Spectral Clustering Algorithms[J]. Computer Science, 2008, 35(7): 14 – 18.
[10] M Filipponea, F Camastrab, F Masullia, S Rovetta. A survey of kernel and spectral methods for clustering[J]. Pattern Recognition, 2008, 41(1): 176 – 190.
[11] U Luxburg. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395 – 416.
[12] C Ding, X F He, H Y Zha, et al. A min-max cut algorithm for graph partitioning and data clustering[A]. The 2001 IEEE International Conference on Data mining[C]. San Jose, USA, 2001: 107 – 114.
[13] M Meila, J Shi. Learning segmentation by random walks[C]. In: NIPS, Cambridge, MIT Press, 2000: 873 – 879.
[14] 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, Jiao Li-cheng. Density sensitive spectral clustering[J]. Acta Electronica Sinica, 2007, 35(8): 1577 – 1581. (in Chinese)
[15] 王娜, 李霞. 基于监督信息特性的主动半监督谱聚类算法[J]. 电子学报, 2010, 38 (1): 172 – 176.
Wang Na, Li Xia. Active semi-supervised spectral clustering based on pairwise constraint[J]. Acta Electronica Sinica, 2010, 38(1): 172 – 176. (in Chinese)
[16] 田铮, 李小斌, 句彦伟. 谱聚类的扰动分析[J]. 中国科学, 2007, 37(4): 527 – 543.
Tian Zheng, Li Xiao-bin, Ju Yan-wei. Spectral clustering based on matrix perturbation theory[J]. Science in China. 2007, 37(4): 527 – 543. (in Chinese)
[17] 孙继广. 矩阵扰动分析[M]. 北京: 科学出版社, 2001. 146 – 160.
Sun Ju-guang. Matrix Perturbation Analysis[M]. Beijing: Science Press, 2001. 146 – 160. (in Chinese)