基于谱聚类的聚类集成算法

周 林[1] 平西建[1] 徐 森[2] 张 涛[1]

摘 要 谱聚类是近年来出现的一类性能优越的聚类算法, 能对任意形状的数据进行聚类, 但算法对尺度参数比较敏感, 利用聚类集成良好的鲁棒性和泛化能力, 本文提出了基于谱聚类的聚类集成算法. 该算法首先利用谱聚类算法的内在特性构造多样性的聚类成员; 然后, 采用连接三元组算法计算相似度矩阵, 扩充了数据点之间的相似性信息; 最后, 对相似度矩阵使用谱聚类算法得到最终的集成结果. 为了使算法能扩展到大规模应用, 利用 Nyström 采样算法只计算随机采样数据点之间以及随机采样数据点与剩余数据点之间的相似度矩阵, 从而有效降低了算法的计算复杂度. 本文算法既利用了谱聚类算法的优越性能, 同时又避免了精确选择尺度参数的问题. 实验结果表明: 较之其他常见的聚类集成算法, 本文算法更优越、更有效, 能较好地解决数据聚类、图像分割等问题.

关键词 谱聚类, 聚类集成, 连接三元组, 图像分割

引用格式 周林, 平西建, 徐森, 张涛. 基于谱聚类的聚类集成算法. 自动化学报, 2012, **38**(8): 1335−1342

DOI 10.3724/SP.J.1004.2012.01335

Cluster Ensemble Based on Spectral Clustering

ZHOU Lin[1] PING Xi-Jian[1] XU Sen[2] ZHANG Tao[1]

Abstract Spectral clustering has become increasingly popular in recent years. It can deal with arbitrary distribution dataset, however, it is sensitive to the scaling parameter. Cluster ensemble based on spectral clustering is proposed which utilizes the good robustness and generalization ability of cluster ensemble. Multiform clustering components are generated by exploiting the property of spectral clustering, and the connected triple algorithm which can expand the similarity information among data is used to compute the affinity matrix, then the affinity matrix is used by spectral clustering algorithm to produce ensemble results. In order to make the algorithm extensible to large scale applications, only the similarity among the rand sampling data and the similarity between the random sampling data and the rest data are computed by adopting the Nyström sampling method. The proposed algorithm makes full use of the excellent performance of spectral clustering as well as avoids the selection of the accurate parameter in spectral clustering. Experiments show that compared with other common cluster ensemble techniques, the proposed algorithm is more excellent and efficient, and that it can provide a good way to solve data clustering and image segmentation problem.

Key words Spectral clustering, cluster ensemble, connected triple, image segmentation

Citation Zhou Lin, Ping Xi-Jian, Xu Sen, Zhang Tao. Cluster ensemble based on spectral clustering. *Acta Automatica Sinica*, 2012, **38**(8): 1335−1342

聚类问题一直是机器学习和模式识别领域一个比较活跃而且极具挑战性的研究方向[1]. 所谓聚类就是将数据对象分组成为多个类或簇, 使得在同一簇中的对象之间具有较高的相似度, 而不同簇中的对象差别较大. 聚类技术在机器学习、数据挖掘、模式识别以及图像分析等领域都有广泛的应用. 传统的聚类算法, 如 K-均值算法、EM (Expectation-maximization) 算法等都是建立在凸球形的样本空间上, 但当样本空间不为凸时, 算法会陷入局部最优[2]. 近年来, 谱聚类算法开始受到广泛关注[3−14]. 该算法首先根据给定的样本数据集定义一个描述数据点之间相似度的矩阵, 计算矩阵的特征值和特征向量, 然后, 选择合适的特征向量聚类不同的数据点. 从本质上来说, 谱聚类算法是通过矩阵谱分析理论导出聚类对象的新特征, 利用新的数据特征对原数据进行聚类. 与其他聚类算法相比, 谱聚类算法实现简单, 不易陷入局部最优解, 且具有识别非凸分布的聚类的能力. 但谱聚类算法自身也存在计算量较大、构造相似度矩阵时对尺度参数敏感等问题, 至今还没有有效的解决办法.

为了获得鲁棒性好且稳定的聚类性能, 人们提出了聚类集成算法. 聚类集成被认为在很多方面都能够超越单个聚类算法的性能, 如鲁棒性、稳定性和一致性估计[1]. 然而, 聚类集成要比对分类器的集成困难得多, 其中的关键问题是如何根据不同聚类成

收稿日期 2011-07-11 录用日期 2011-10-17
Manuscript received July 11, 2011; accepted October 17, 2011
国家自然科学基金 (60970142, 60903221, 61105057), 盐城工学院人才引进专项基金 (XKR2011019) 资助
Supported by National Natural Science Foundation of China (60970142, 60903221, 61105057) and Talent Introduction Special Foundation of Yancheng Institute of Technology (XKR2011019)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 解放军信息工程大学信息工程学院 郑州 450002 2. 盐城工学院信息工程学院 盐城 224000
1. Institute of Information Engineering, The Chinese Peoples′s Liberation Army Information Engineering University, Zhengzhou 450002 2. Scholl of Information Engineering, Yancheng Institute of Technology, Yancheng 224000

员得到的簇标签组合得到更好的聚类结果. Fred 等[15] 根据簇标签得到数据对象之间的互关联 (Co-association, CO) 矩阵, 避免了簇标签对应问题, 并使用层次聚类算法对互关联矩阵进行聚类得到聚类结果. Strehl 等[16] 提出了 CSPA (Cluster-based similarity partitioning algorithm) 算法, 其中, 调用了图划分算法 METIS 对相似度矩阵进行聚类, 他们还提出了 HGPA (Hypergraph partitioning algorithm) 算法和 MCLA (Meta-clustering algorithm) 算法. Zhou 等[17] 提出了多数投票法, 该方法的关键是簇标签的对应问题, 在此基础上采用多数投票来确定最终的分类. Iam-On 等[18] 提出了两种新的基于连接的相似度矩阵 CTS (Connected-triple-based similarity) 矩阵和 SRS (Simrank-based similarity) 矩阵, 并调用层次聚类法对这两个矩阵聚类得到聚类结果.

本文提出了一种基于谱聚类的聚类集成算法. 该算法先利用谱聚类算法构造多样性的聚类成员, 并采用连接三元组方法 (Connected triple algorithm) 和 Nyström 采样方法得到随机采样数据点之间以及随机采样数据点与剩余数据点之间的相似度矩阵, 然后, 对相似度矩阵使用谱聚类算法得到聚类结果. 该算法既利用了谱聚类能对任意形状的数据进行聚类, 且算法不易陷入局部最优解, 能够得到比其他聚类算法更优越的结果的优点; 又避免了谱聚类算法对尺度参数敏感的问题.

1 谱聚类和聚类集成

1.1 谱聚类

谱聚类的思想来源于谱图划分[1], 它将数据聚类问题看成是一个无向图的多路划分问题. 数据点看成是一个无向图 $G(V,W)$ 的顶点 V, 边权重的集合 $W=\{S_{ij}\}$ 表示基于某一相似性度量计算的两点间的相似度, S 表示待聚类数据点间的相似度矩阵, 将其看做是该无向图的邻接矩阵, 它包含了聚类所需的所有信息. 然后, 定义一个划分准则, 最优化这一准则, 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.

虽然谱聚类算法具有坚实的理论基础, 并且已经在很多领域获得了成功应用, 但是仍存在以下两个主要缺点: 1) 对尺度参数比较敏感, 这使得相似度矩阵 S 构造困难; 2) 需要求解矩阵的特征值分解问题, 对于大规模应用, 其计算量和存储量太大让人难以接受.

1.2 聚类集成

通常可以将聚类集成问题表述如下: 令 $X=\{x_1,x_2,\cdots,x_N\}$ 表示由 N 个数据点组成的集合, $\Pi=\{\pi_1,\pi_2,\cdots,\pi_M\}$ 表示对 X 进行 M 次聚类得到的聚类结果的集合, 其中, $\pi_i=\{\boldsymbol{C}_1^i,\boldsymbol{C}_2^i,\cdots,\boldsymbol{C}_{k_i}^i\}$ 为第 i 次聚类得到的簇集合, k_i 表示第 i 个簇集合中簇的个数. 对数据点 $x\in X$, $\boldsymbol{C}(x)$ 表示其簇标签, 在第 i 个簇集合中, 如果 $x\in\boldsymbol{C}_j^i$, 则有 $\boldsymbol{C}(x)=j$. 聚类集成就是对集合 Π 进行合并, 得到数据集 X 最终的聚类结果. 聚类集成研究主要包括两个方面: 1) 构造聚类成员; 2) 设计共识函数.

多样性已被证明是提高聚类集成性能的关键因素. 与分类器集成类似, 对于一个给定任务, 具有多样性的聚类成员可以通过多种方式构造: 选择不同的算法[16]、对一个算法选择不同的初值[19]、选择不同的对象子集[20−21]、选择不同的特征子集投影到数据子空间[22] 等.

在得到聚类成员后, 学者们设计了各种共识函数来得到最终的聚类结果. 总体来说, 主要有以下三类共识函数: 1) 基于特征的方法[20−23]; 2) 基于图的方法[16−24]; 3) 基于数据点间相似度的方法[15]. 第一种方法由每个聚类成员提供各个数据点的聚类标签, 使用这些标签作为新的特征来得到最后的聚类结果. 这种方法需要估计的参数较多, 计算量较大而且容易陷入局部最优解. 第二种方法将聚类集成问题表示成超图的形式, 并调用图划分算法. 由于图划分算法为了避免得到平凡解而加了平衡性约束, 因此, 得到的每个簇的大小大致相等, 而且该方法的计算量较大. 最后一种方法建立数据点之间的相似度矩阵 — 互关联矩阵, 然后, 使用基于相似度聚类的方法来得到聚类结果. 该方法显式或隐式地给最终的簇强加了某种结构, 如单连接算法仅考虑最近邻信息, 它侧重于发现簇间的连通性, 因此, 容易产生具有链式结构的簇, 另外, 该方法的计算量和存储量都较大.

互相关矩阵简单、易于构造, 但它只能得到少部分数据点之间的相似度, 其他数据点之间的相似度都为 0. 为了得到更多数据点之间的相似度信息, 文献 [18] 提出了一种新的基于连接的相似度矩阵构造方法: 连接三元组算法. 该算法可表述如下: 连接三元组 $\bigwedge=(V_\wedge,W_\wedge)$ 是 G 的一个子图, 包含三个顶点 $V_\wedge=(A_1,A_2,A_3)\subset V$ 和两条边 $W_\Lambda=\{e_{A_1,A_2},e_{A_1,A_3}\}\subset W$, 连接其他两个顶点的顶点称为这个三元组的中心. 其基本思想是: 如果两个节点都与第三个节点有连接, 则认为这两个节点之间存在相似性.

该方法在聚类问题上的应用如图 1 所示, 圆形顶点代表数据点, 方形顶点代表簇集合中的簇, 如果数据点 $x_i\in\boldsymbol{C}_j^k$, 则它们之间存在一条边. 对于簇集合 π_2 和 π_3 来说, 由于数据点 x_1 和 x_2 分别都属于簇 $\boldsymbol{C}_1^2$ 和簇 $\boldsymbol{C}_1^3$, 因此认为它们是相似的. 而对于簇集合 π_1, 点 x_1 和 x_2 属于不同的簇, 但如果这些簇是相似的, 则它们之间也存在相似性. 根据连接三元组方法, 由于簇 $\boldsymbol{C}_1^1$ 和簇 $\boldsymbol{C}_2^1$ 具有两个连接三元组, 且簇 $\boldsymbol{C}_1^2$ 和簇 $\boldsymbol{C}_1^3$ 分别是这两个连接三元组的中心,

因此, 簇 $\boldsymbol{C}_1^1$ 和簇 $\boldsymbol{C}_2^1$ 是相似的, 从而对于簇集合 π_1 来说, x_1 和 x_2 也是相似的. 可见该方法扩充了数据点之间的相似性信息, 有利于具有复杂结构的数据聚类.

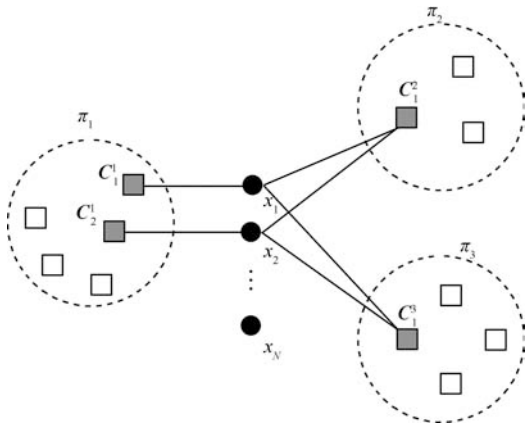


图 1　连接三元组聚类集成示意图

Fig. 1　A graphical representation of a cluster ensemble

2　基于谱聚类的聚类集成

2.1　聚类成员的构造

本文将单个谱聚类算法作为聚类成员. 谱聚类算法对尺度参数非常敏感, 不同的参数将产生完全不同的结果, 这种缺点对于构造有效的集成学习系统非常有用. 在预先给定的参数范围内, 为每个聚类成员随机选择一个尺度参数, 从而既避免了尺度参数的选择问题, 又产生了具有差异的聚类成员. 另外, K-均值聚类算法对于初始化非常敏感, 在谱聚类中, 用于特征分量聚类的 K-均值算法采用随机初始化, 这有助于多样性聚类成员的产生. 对于大规模数据处理来说, 直接使用原始的谱聚类算法计算量太大, 因此, 使用基于 Nyström 采样方法的谱聚类, 可以有效降低计算量. 另外, 由于 Nyström 采样每次抽取的样本点是不同的, 因此, 聚类结果也是存在差异的, 这恰恰有利于聚类集成中聚类成员多样性的构成. 总结起来, 不同尺度参数、基于不同初始化的 K-均值算法以及 Nyström 采样方法的使用一起来构造多样性的聚类成员.

2.2　聚类集成

在构造好聚类成员后, 鉴于已有聚类集成算法存在的不足和谱聚类算法的优点, 本文引入谱聚类算法思想来解决聚类集成问题. 从谱聚类的角度来看, 聚类集成为其提供数据点之间有意义的相似度矩阵, 聚类成员可以在设定的参数范围内随机选择参数, 避免了参数设置问题; 从聚类集成的角度来看, 谱聚类对簇的形状不做强的假设, 且算法不易陷入局部最优解, 能够得到比其他聚类算法更优越的结果. 聚类集成的核心问题之一是如何根据这些由聚类成员得到的簇标签构造数据点之间的相似度矩阵. 本文选用能得到数据点之间更多相似性信息的连接三元组算法[18] 来构造数据点之间的 CTS 矩阵.

连接簇 $\boldsymbol{C}_i$ 和簇 $\boldsymbol{C}_j$ 的边的权重 W_{ij} 由这两个簇共同包含的数据点个数得到, 如下式:

$$W_{ij} = \frac{|X_i \cap X_j|}{|X_i \cup X_j|} \tag{1}$$

其中, X_i 为属于簇 $\boldsymbol{C}_i$ 的数据点的集合. 邻接点为簇 $\boldsymbol{C}_k$ 的两个簇 $\boldsymbol{C}_i, \boldsymbol{C}_j$ 之间连接三元组的个数为

$$WCT_{ij}^k = \min(w_{ik}, w_{jk}) \tag{2}$$

簇 $\boldsymbol{C}_i, \boldsymbol{C}_j$ 之间的相似度 $Sim^{WCT}(i,j)$ 计算如下:

$$Sim^{WCT}(i,j) = \frac{\sum_{k=1}^{q} WCT_{ij}^k}{WCT_{\max}} \tag{3}$$

其中, $1 \le q < \infty$.

对任意的簇集合 $\pi_m \in \prod$, $m = 1, \cdots, M$, 数据点 $x_i, x_j \in X$ 之间的相似度 $S_m(x_i, x_j)$ 计算如下:

$$S_m(x_i, x_j) = \tag{4}$$
$$\begin{cases} 1, & \boldsymbol{C}(x_i) = \boldsymbol{C}(x_j) \\ Sim^{WCT}(\boldsymbol{C}(x_i), \boldsymbol{C}(x_j)) \cdot DC, & \text{否则} \end{cases}$$

其中, DC 为衰减因子, 即认为两个不相同事物相似的置信水平. CTS 矩阵 S 中的元素 $S(x_i, x_j)$ 计算如下:

$$S(x_i, x_j) = \frac{1}{M} \sum_{m=1}^{M} S_m(x_i, x_j) \tag{5}$$

S 即为数据点之间的相似度矩阵. 相对于传统的互关联矩阵来说, CTS 矩阵考虑了当两个数据点不属于同一个簇时它们之间的相似度, 扩充了数据点之间的相似性信息, 有利于具有复杂结构的数据聚类. 但 CTS 矩阵的计算量和存储量太大, 无法直接应用于大规模数据处理. 以图像分割为例, 一幅大小为 256 像素 × 256 像素的图像的样本数目为 65 536, 矩阵大小为 65 536 × 65 536, 在内存为 1 GB 的计算上使用 Matlab 2007 Ra 来生成这么大的矩阵时, 程序报错 “Out of memory”, 即内存溢出, 无法对其做矩阵分解.

本文提出利用 Nyström 采样方法[25] 来对 CTS 矩阵进行分解, 可以有效解决 CTS 矩阵计算量和存储量大的问题, 使得算法能扩展到大规模应用. 假设随机采样的数据点为 n 个, 记为 $x_{s(i)}$, 剩下的数据点为 $N-n$ 个, 记为 $x_{r(k)}$. 对任意的簇集合 $\pi_m \in \prod$, $m = 1, \cdots, M$, 用 $s_m^{ss}(i,j)$ 表示第 i 个和第 j 个随机采样的数据点之间的相似度, $s_m^{sr}(i,k)$ 表示第 i 个随机采样的数据点和第 k 个剩余数据点之间的相似度, S_{ss} 表示随机采样数据点之间的相似度矩阵, S_{sr} 表示随机采样数据与剩余数据点之间的相似度矩阵, 则有:

$$s_m^{ss}(i,j)=\begin{cases}1, & \boldsymbol{C}(x_{s(i)})=\boldsymbol{C}(x_{s(j)})\\ Sim^{WCT}(\boldsymbol{C}(x_{s(i)}),\boldsymbol{C}(x_{s(j)}))\cdot DC, & \text{否则}\end{cases},i,j=1,\cdots,n \tag{6}$$

$$s_m^{sr}(i,k)=\begin{cases}1, & \boldsymbol{C}(x_{s(i)})=\boldsymbol{C}(x_{r(k)})\\ Sim^{WCT}(\boldsymbol{C}(x_{s(i)}),\boldsymbol{C}(x_{r(k)}))\cdot DC, & \text{否则}\end{cases},i=1,\cdots,n;\ k=1,\cdots,N-n \tag{7}$$

$$S_{ss}(i,j)=\frac{1}{M}\sum_{m=1}^{M}s_m^{ss}(i,j) \tag{8}$$

$$S_{sr}(i,k)=\frac{1}{M}\sum_{m=1}^{M}s_m^{sr}(i,k) \tag{9}$$

则 CTS 矩阵 S 可以分解为

$$S=\begin{bmatrix}S_{ss} & S_{sr}\\ S_{sr}^{\mathrm{T}} & Q\end{bmatrix} \tag{10}$$

对 S_{ss} 做特征值分解:

$$S_{ss}=U\Lambda U^{\mathrm{T}} \tag{11}$$

$$\Lambda=\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix} \tag{12}$$

如果所有的特征值 $\lambda_i,\ i=1,\cdots,n$ 都大于零, 则 S_{ss} 为正定矩阵, 反之则不是.

如果 S_{ss} 为正定矩阵, 定义:

$$P=S_{ss}+S_{ss}^{-\frac{1}{2}}S_{sr}S_{sr}^{\mathrm{T}}S_{ss}^{-\frac{1}{2}} \tag{13}$$

其中 $S_{ss}^{1/2}$ 为矩阵 S_{ss} 满足对称正定的平方根, 将 P 对角化得到:

$$P=U_P\Lambda_PU_P^{\mathrm{T}} \tag{14}$$

定义矩阵 Y 为

$$Y=\begin{bmatrix}S_{ss}\\ S_{sr}^{\mathrm{T}}\end{bmatrix}S_{ss}^{-\frac{1}{2}}U_P\Lambda_P^{-\frac{1}{2}} \tag{15}$$

则有 $\hat{S}=Y\Lambda_PY^{\mathrm{T}},\ Y^{\mathrm{T}}Y=I$, 其中 $\hat{S}$ 表示对 S 的逼近.

如果 S_{ss} 为非正定矩阵, 令:

$$\bar{U}_P^{\mathrm{T}}=U_P^{\mathrm{T}}\Lambda_P^{-1}U_P^{\mathrm{T}}S_{sr} \tag{16}$$

定义 $Z=\bar{U}_P\Lambda^{1/2}$, 对 $Z^{\mathrm{T}}Z$ 进行对角分解得:

$$Z^{\mathrm{T}}Z=F\Lambda_FF^{\mathrm{T}} \tag{17}$$

定义矩阵 Y 为

$$Y=ZF\Lambda_F^{-\frac{1}{2}} \tag{18}$$

则 $\hat{S}=Y\Lambda_FY^{\mathrm{T}},\ Y^{\mathrm{T}}Y=I$.

在得到矩阵 S 的特征向量 Y 后, 即可使用谱聚类算法进行聚类. 综上所述, 基于谱聚类的聚类集成算法的基本流程如下:

输入. N 个数据点组成的集合 X, 随机采样数据点个数 n, 类别数目 K, 聚类成员的数目 M, 基于 Nyström 方法的谱聚类算法 A, 其中, K-均值算法中采用随机初始化.

1) 集成系统中簇集合的产生

for $i=1:M$

σ_i: 在 $[\sigma_{\min},\sigma_{\max}]$ 之间随机选择一个尺度参数

$\pi_i=A(\sigma_i)$

end

2) 利用集合 $\prod=\{\pi_1,\pi_2,\cdots,\pi_M\}$ 构造 CTS 矩阵, 使用 Nyström 采样方法计算矩阵 S 的特征向量 Y;

3) 将 Y 的前 K 个最大特征值对应的特征向量形成矩阵 $H\in\mathbf{R}^{N\times K}$, 设 $\boldsymbol{g}_i\in\mathbf{R}^K$ 为对应于 H 的第 i 行的列向量, 使用 K-均值算法把 $G=\{\boldsymbol{g}_i|i=1,\cdots,N\}$ 聚为 K 个簇 $\boldsymbol{C}_1,\cdots,\boldsymbol{C}_K$.

输出. 数据簇 $D_1,\cdots,D_K$, 其中, $D_i=\{x_j|\boldsymbol{g}_j\in\boldsymbol{C}_i,x_j\in X\},\ 1\le i\le K$.

2.3 计算复杂度分析和比较

在后面的聚类集成对比实验中, 本文都是先使用谱聚类算法对数据集合进行聚类得到簇集合, 然后, 使用不同的聚类集成算法对这些簇集合进行集成得到最后的聚类结果, 因此, 只比较聚类集成阶段的计算复杂度. 在得到簇集合后, 本文算法第 1) 步计算矩阵 S_{ss} 和 S_{sr}, 复杂度分别为 $\mathrm{O}(n^2M)$ 和 $\mathrm{O}(n(N-n)M)$. 第 2) 步, 如果 S_{ss} 为正定矩阵, 对 S_{ss} 进行矩阵分解, 复杂度为 $\mathrm{O}(n^3)$; 求解矩阵 P 的复杂度为 $\mathrm{O}(n^2(N-n))$, 其矩阵分解的复杂度为 $\mathrm{O}(n^3)$; 求解矩阵 Y 的复杂度为 $\mathrm{O}(n^2N)$. 如果 S_{ss} 为非正定矩阵, 求解矩阵 Z 的复杂度为 $\mathrm{O}(n^2N)$; 对 $Z^{\mathrm{T}}Z$ 进行对角分解的复杂度为 $\mathrm{O}(n^3)$; 求解矩阵 Y 的复杂度为 $\mathrm{O}(n^2N)$. 第 3) 步, 使用 K-均值算法进行聚类, 其中复杂度为 $\mathrm{O}(nK^2T_1)$, T_1 为 K-均值算法的迭代次数. 可见, 本文算法的计算复杂度为 $\mathrm{O}(n^2N)$. 而 CSPA、HGPA 和 MCLA 算法的计算复杂度分别为 $\mathrm{O}(KMN^2)$, $\mathrm{O}(KMN)$ 和 $\mathrm{O}(K^2M^2N)$. 文献 [15] 中计算 CO 矩阵的计算量为 $\mathrm{O}(MN^2)$, 调用层次聚类算法 SL (Single-linkage) 的计算复杂度为 $\mathrm{O}(N^2)$, 调用 AL (Average-linkage) 和 CL (Complete-linkage) 算法的计算复杂度为 $\mathrm{O}(N^2\lg N)$. 文献 [18] 中计算 CTS

矩阵的计算量为 $O(MN^2+K^2T_2)$, 其中, T_2 表示直接连接的簇的平均个数, 调用三种层次聚类算法的计算量与文献 [15] 相同. 由此可见, CSPA 算法、基于 CO 矩阵的聚类集成算法和基于 CTS 矩阵的聚类集成算法的计算复杂度为数据点个数 N 的二次方, 而本文算法的计算复杂度与 HGPA、MCLA 算法相当, 都为 N 的一次方.

3 实验结果与分析

3.1 UCI 数据集聚类

选用 UCI (University of California-Irvine) 数据库中 4 个真实的数据集验证算法的有效性, 表 1 给出了这些数据集的特性. 采用聚类结果与已知真实类属信息匹配后的误分率和 FM (Fowlkes-mallows) 作为评价标准. 其中, FM 表示测量聚类结果与真实类属信息一致性的 Fowlkes-mallows 指标, 其值处于 0 和 1 之间, 且值越大表示一致性越好[26].

表 1 实验所用 UCI 数据集的属性

Table 1 The attribute of the selected UCI datasets

数据集	类别数	特征维数	数据量
Iris	3	4	150
Sonar	2	60	208
Segmentation	7	19	2 310
Pen digits	10	16	10 992

在前 3 个数据集上, 分别采用 6 类聚类算法进行聚类: 1) K-均值算法; 2) 谱聚类 (Spectral clustering, SC) 算法; 3) 基于图方法的谱聚类集成算法: CSPA、HGPA 和 MCLA; 4) 基于 CO 矩阵的聚类集成算法[15]: CO-SL、CO-AL 和 CO-CL; 5) 基于 CTS 矩阵的聚类集成算法[18]: CTS-SL、CTS-AL 和 CTS-CL; 6) 本文算法. 在 Pen digits 数据集上, 由于样本数较大, CSPA 算法、基于 CO 矩阵的聚类集成算法和基于 CTS 矩阵的聚类集成算法所需存储空间和计算量太大, 会出现 "Out of memory", HGPA 算法效果较差, 因此, 集成算法只采用 MCLA 算法和本文算法. 由于 K-均值算法对初始化比较敏感, 对每个数据集独立运行 50 次, 取最好的一个结果. 在谱聚类算法和本文算法中, 都使用 K-均值算法对相似度矩阵的前 K 个最大特征值对应的特征向量进行聚类, 对特征向量独立运行 50 次, 取所有样本点到聚类中心距离的平方和中最小的作为结果. 对于谱聚类算法, 用每个参数运行一次, 取最好的一个结果. 在聚类集成算法中, 每个个体聚类成员从原始数据集中随机采样 100 个数据点作为代表点, 集成系统包括 25 个聚类成员, 其中, 尺度参数在区间 $[1:0.1:5]$ 中随机取值. 表 2 为独立运行 30 次各种谱聚类集成算法的平均值. 在聚类前将数据归一化到 $[0,1]$ 范围内.

从表 2 和表 3 可以看出: 1) 在上述 4 个数据集上, 本文算法的性能均高于 SC 算法, 说明基于谱聚类的聚类集成算法有效提高了谱聚类算法的性能; 2) 除了在 Iris 数据集上本文算法的性能与其他聚类集成算法相当外, 在其他 3 个数据集上, 本文算法的性能均高于其他聚类集成算法; 3) 在调用相同的层次聚类算法 (SL、AL 和 CL) 时, 采用 CTS 矩阵算法的性能等于或高于采用 CO 矩阵算法的性能, 这验证了本文采用连接三元组算法 CTS 矩阵是恰当的.

表 2 不同算法的误分率比较 (%)

Table 2 The error rates obtained by different algorithms (%)

	K-均值	SC	CSPA	HGPA	MCLA	CO-SL/AL/CL	CTS-SL/AL/CL	本文算法
Iris	11.33	10.72	14.87	36.53	10.00	**10.00/10.00/10.00**	**10.00/10.00/10.00**	**10.00**
Sonar	44.71	44.53	43.87	49.51	43.26	47.87/43.64/44.87	46.95/43.51/44.20	**41.95**
Segmentation	33.46	22.08	21.19	51.17	20.87	44.02/20.84/29.30	43.57/20.74/23.42	**19.22**
Pen digits	33.29	25.27	—	—	24.71	—	—	**23.13**

表 3 不同算法的 FM 比较

Table 3 The FM obtained by different algorithms

	K-均值	SC	CSPA	HGPA	MCLA	CO-SL/AL/CL	CTS-SL/AL/CL	本文算法
Iris	0.766	0.819	0.766	0.586	0.830	**0.830/0.830/0.830**	**0.830/0.830/0.830**	**0.830**
Sonar	0.502	0.506	0.516	0.469	0.528	0.476/0.521/0.498	0.482/0.525/0.511	**0.539**
Segmentation	0.575	0.648	0.667	0.437	0.680	0.512/0.681/0.609	0.526/0.687/0.661	**0.715**
Pen digits	0.587	0.625	—	—	0.631	—	—	**0.644**

3.2 图像分割

3.2.1 合成纹理图像分割

利用谱聚类集成算法对合成图像进行分割, 图 2 (a) 为一幅 300 像素 × 300 像素的原始合成图像, 包含 4 类纹理. 采用基于非下采样 Contourlet 变化的纹理特征提取方法[27] 进行纹理特征提取, 将每幅图像分解为 3 层, 每一个像素由一个 16 维的区域特征向量表示, 区域窗口大小为 9 像素 × 9 像素. 在聚类前对特征向量进行归一化处理. 由于计算量的问题, 为了比较只采用 K-均值算法、单个基于 Nyström 方法的谱聚类算法、MCLA 算法和本文算法对图像进行分割, 其结果如图 2 所示. 其中, Nyström 方法随机采样 100 个像素点作为代表点. 集成学习算法中包括 25 个聚类成员, 每个成员的参数 σ_ν 和 σ_c 分别在 [0.3, 1.0] 和 [30, 80] 之间随机取值, 其中, σ_ν 和 σ_c 是表示特征相似性信息和空间邻近信息重要程度的尺度参数. 表 4 列出了实际分割结果相对于理想分割的误分率和 FM.

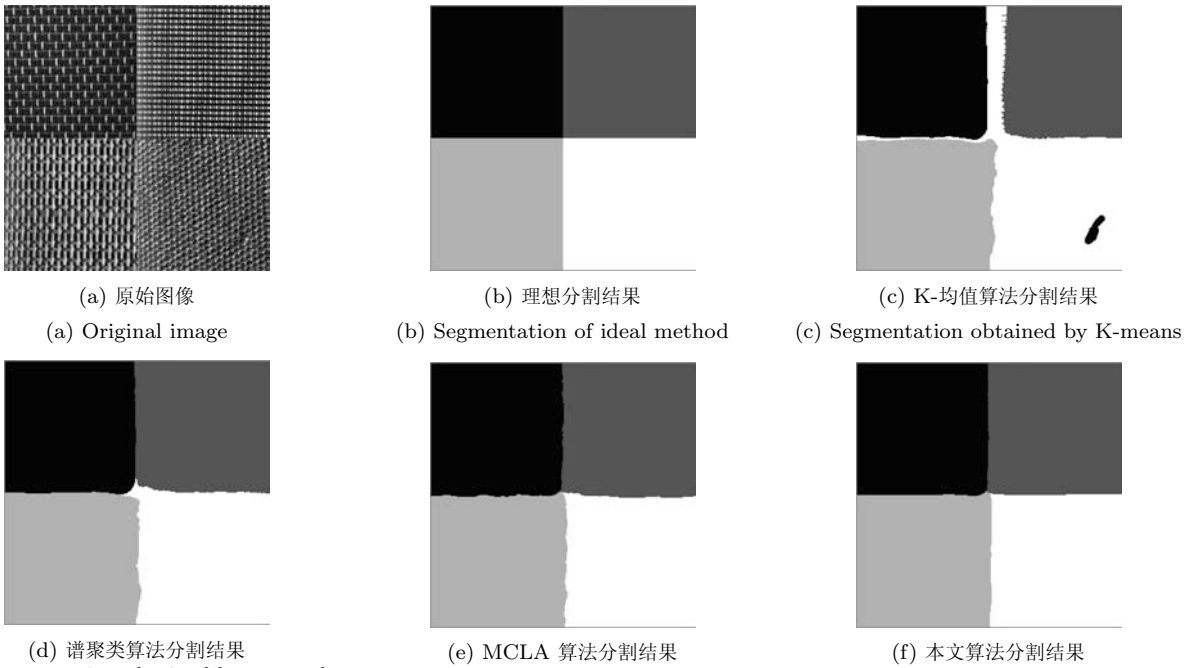
从图 2 和表 4 可以看出, 谱聚类算法的分割结果优于 K-均值算法的结果, 这是由于谱聚类算法具有识别非凸分布聚类的能力, 更适合于复杂数据的划分问题. 然而谱聚类算法中的尺度参数却是个很难确定的值, 谱聚类集成算法避免了尺度参数的选择, MCLA 算法和本文算法均获得了较好的分割结果. 本文算法在视觉上与 MCLA 算法相差不大, 分类性能 (误分率和 FM) 有所提高.

表 4 不同算法的误分率/FM 比较
Table 4 The error rates/FM obtained by different algorithms

算法	误分率 (%)	FM
K-均值算法	11.38	0.783
谱聚类算法	3.95	0.926
MCLA 算法	2.67	0.954
本文算法	1.46	0.978

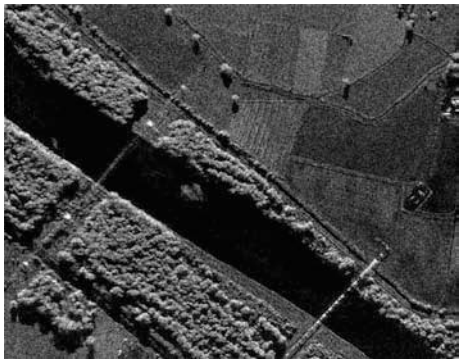
3.2.2 合成孔径雷达图像分割

将本文算法应用于合成孔径雷达 (Synthetic aperturo radar, SAR) 图像分割, 对图像提取基于 4 个方向灰度共生矩阵的 12 维特征和基于 3 层小波分解的 10 维小波能量特征, 构成 22 维特征向量. 图 3 (a) 是一幅 536 像素 × 418 像素的 Ku 波段 SAR 图像, 其中, 包含三类地物: 河流、植被和平原地带. 集成学习算法中包括 25 个聚类成员, 每个成员的尺度参数 σ_ν 和 σ_c 分别在 $[0.1, 0.6]$ 和 [20, 80] 之间随机取值. 为了比较, 采用单个基于 Nyström 方法的谱聚类算法、MCLA 算法和本文提出的基于谱聚类的聚类集成算法对 SAR 图像进行分割. 其中, Nyström 方法随机采样 100 个像素点作为代表点.



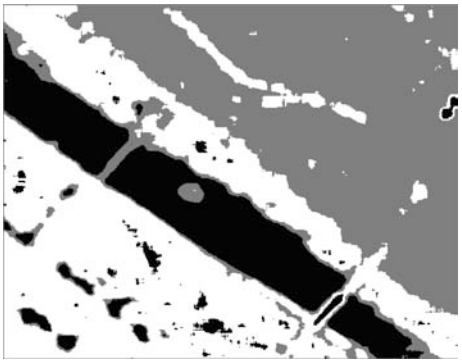
(a) 原始图像 (a) Original image
(b) 理想分割结果 (b) Segmentation of ideal method
(c) K-均值算法分割结果 (c) Segmentation obtained by K-means
(d) 谱聚类算法分割结果 (d) Segmentation obtained by spectral clustering
(e) MCLA 算法分割结果 (e) Segmentation obtained by MCLA
(f) 本文算法分割结果 (f) Segmentation obtained by our method

图 2 合成纹理图像分割
Fig. 2 Segmentation results of synthesized texture image



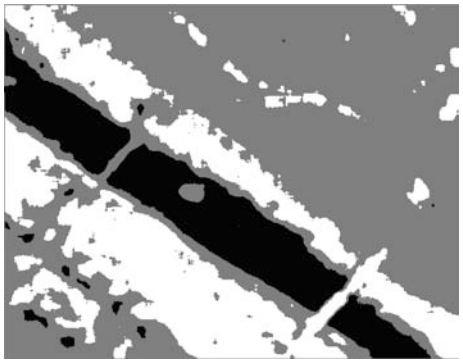
(a) 原始图像

(a) Original image



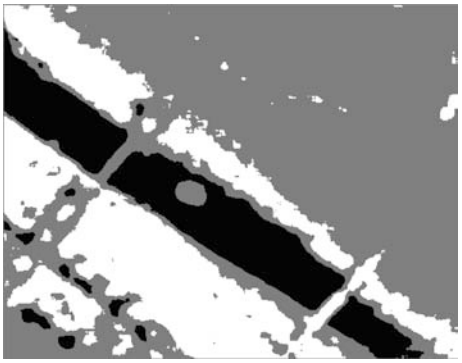
(b) 谱聚类算法分割结果

(b) Segmentation obtained by spectral clustering



(c) MCLA 算法分割结果

(c) Segmentation obtained by MCLA



(d) 本文算法分割结果

(d) Segmentation obtained by our method

图 3 SAR 图像分割

Fig. 3 Segmentation results of SAR image

图 3 (b) 是谱聚类算法分割结果, 从图中可以看出存在错分区域, 尤其是在图的左下角, 错分较为严重. MCLA 算法的分割结果 (图 3 (c)) 好于谱聚类算法, 但对于河流和植被的边界区域以及右上方的植被区域, 区分不是很好. 图 3 (d) 为本文算法分割结果, 可见本文算法对河流、植被和平坦区域都能正确分割.

4 结论

本文提出了基于谱聚类的聚类集成算法, 将谱聚类作为基聚类成员来构造聚类集成系统, 使用连接三元组算法计算相似度矩阵, 并对连接三元组算法进行改进, 有效降低了算法的计算复杂度, 最后, 对相似度矩阵使用谱聚类算法得到聚类结果. 与现有的聚类集成算法相比, 本文降低了计算复杂度, 提高了聚类效果, 可以有效解决数据聚类、图像分割等问题. 另外, 本文提出的 "谱 + 谱" 的方法既利用了谱聚类算法的优越性能, 同时又避免了谱聚类算法对尺度参数敏感的问题, 同样可以用于解决其他应用领域的聚类问题, 且本文方法适用于大规模应用.

References

1 Jiao Li-Cheng, Zhang Xiang-Rong, Hou Biao, Wang Shuang, Liu Fang. *Intelligent SAR Image Processing and Interpretation*. Beijing: Science Press, 2008. 398−435
(焦李成, 张向荣, 侯彪, 王爽, 刘芳. 智能 SAR 图像处理及解译. 北京: 科学出版社, 2008. 398−435)

2 Cai Xiao-Yan, Dai Guan-Zhong, Yang Li-Bin. Survey on spectral clustering algorithms. *Computer Science*, 2008, **35**(7): 14−18
(蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述. 计算机科学, 2008, **35**(7): 14−18)

3 Wu Rui, Huang Jian-Hua, Tang Xiang-Long, Liu Jia-Feng. Method of text image binarization processing using histogram and spectral clustering. *Journal of Electronics and Information Technology*, 2009, **31**(10): 2460−2464
(吴锐, 黄剑华, 唐降龙, 刘家锋. 基于灰度直方图和谱聚类的文本图像二值化方法. 电子与信息学报, 2009, **31**(10): 2460−2464)

4 Wang Na, Li Xia. Active semi-supervised spectral clustering based on pairwise constraints. *Acta Electronica Sinica*, 2010, **38**(1): 172−176
(王娜, 李霞. 基于监督信息特性的主动半监督谱聚类算法. 电子学报, 2010, **38**(1): 172−176)

5 Jia Jian-Hua, Jiao Li-Cheng. Image segmentation by spectral clustering algorithm with spatial coherence constraints. *Journal of Infrared and Millimeter Waves*, 2010, **29**(1): 69−75
(贾建华, 焦李成. 空间一致性约束谱聚类算法用于图像分割. 红外和毫米波学报, 2010, **29**(1): 69−75)

6 Ersahin K, Cumming I G, Ward R K. Segmentation and classification of polarimetric SAR data using spectral graph partitioning. *IEEE Transactions on Geoscience and Remote Sensing*, 2010, **48**(1): 164−167

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

8 Lauer F, Schnorr C. Spectral clustering of linear subspaces for motion segmentation. In: Proceedings of the 12th IEEE International Conference of Computer Vision. Kyoto, Japan: IEEE, 2009. 678−685

9 Xu Sen, Lu Zhi-Mao, Gu Guo-Chang. Two spectral algorithms for ensembling document clusters. *Acta Automatica Sinica*, 2009, **35**(7): 997−1002
(徐森, 卢志茂, 顾国昌. 解决文本聚类集成问题的两个谱算法. 自动化学报, 2009, **35**(7): 997−1002)

10 Cheng Y, Tong Q. Spectral clustering on manifolds with statistical and geometrical similarity. *Lecture Notes in Computer Science*. Berlin: Springer, 2010. 422−429

11 Zhang Y L, Zhuang J, Wang S A. Fusion of manifold learning and spectral clustering algorithm with application to fault diagnosis. In: Proceedings of the 2nd International Conference on Machine Learning and Computing. Bangalore, India: IEEE, 2010. 155−160

12 Zhao F, Jiao L C, Liu H Q, Gao X B, Gong M G. Spectral clustering with eigenvector selection based on entropy ranking. *Neurocomputing*, 2010, **73**(10−12): 1704−1717

13 Zhang Xiang-Rong, Qian Xiao-Xue, Jiao Li-Cheng. Immune spectral clustering algorithm for image segmentation. *Journal of Software*, 2010, **21**(9): 2196−2205
(张向荣, 骞晓雪, 焦李成. 基于免疫谱聚类的图像分割. 软件学报, 2010, **21**(9): 2196−2205)

14 Xu Hai-Xia, Tian Zheng, Ding Ming-Tao. Multiscale segmentation for SAR image based on spectral clustering and mixture model. *Journal of Image and Graphics*, 2010, **15**(3): 450−454
(徐海霞, 田铮, 丁明涛. 基于谱聚类与混合模型的 SAR 图像多尺度分割. 中国图象图形学报, 2010, **15**(3): 450−454)

15 Fred A L, Jain A K. Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, **27**(6): 835−850

16 Strehl A, Ghosh J. Cluster ensembles — a knowledge reuse framework for combining partitions. *Journal of Machine Learning Research*, 2002, **3**: 583−617

17 Zhou Z H, Tang W. Clusterer ensemble. *Knowledge-Based Systems*, 2006, **19**(1): 77−83

18 Iam-On N, Boongoen T, Garrett S. Refining pairwise similarity matrix for cluster ensemble problem with cluster relations. In: Proceedings of the 11th International Conference on Discovery Science. Budapest, Hungary: Springer, 2008. 222−233

19 Ayad H, Basir O, Kamel M. A probabilistic model using information theoretic measures for cluster ensemble. In: Proceedings of the 5th International Workshop on Multiple Classifier Systems. Cagliari, Italy: Springer, 2004. 144−153

20 Topchy A, Jain A K, Punch W. A mixture model for clustering ensembles. In: Proceedings of the 4th SIAM International Conference on Data Mining. Florida, USA: SIAM, 2004. 379−390

21 Tang Wei, Zhou Zhi-Hua. Bagging-based selective clusterer ensemble. *Journal of Software*, 2005, **16**(4): 496−502
(唐伟, 周志华. 基于 Bagging 的选择性聚类集成. 软件学报, 2005, **16**(4): 496−502)

22 Fern X Z, Brodley C E. Random projection for high dimensional data clustering: a cluster ensemble approach. In: Proceedings of the 20th International Conference on Machine Learning. Washington D. C., USA: AAAI Press, 2003. 186−193

23 Nguyen N, Caruana R. Consensus clusterings. In: Proceedings of the 7th IEEE International Conference on Data Mining. Omaha, USA: IEEE, 2007. 607−612

24 Fern X Z, Brodley C E. Solving cluster ensemble problems by bipartite graph partitioning. In: Proceedings of the 21st International Conference on Machine Learning. Banff, Canada: ACM, 2004. 1−8

25 Fowlkes C, Belongie S, Chung F, Mailk J. Spectral grouping using the Nyström method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, **26**(2): 214−225

26 Wang Kai-Jun, Zhang Jun-Ying, Li Dan, Zhang Xin-Na, Guo Tao. Adaptive affinity propagation clustering. *Acta Automatica Sinica*, 2007, **33**(12): 1242−1246
(王开军, 张军英, 李丹, 张新娜, 郭涛. 自适应仿射传播聚类. 自动化学报, 2007, **33**(12): 1242−1246)

27 Cunha A L, Zhou J P, Do N M. The nonsubsampled contourlet transform: theory, design, and application. *IEEE Transactions on Image Processing*, 2006, **15**(10): 3089−3101

周 林 解放军信息工程大学信息工程学院博士研究生. 主要研究方向为图像处理和模式识别. 本文通信作者.
E-mail: zhoulin8382@163.com
(**ZHOU Lin** Ph. D. candidate at the Institute of Information Engineering, The Chinese People′s Liberation Army Information Engineering University. His research interest covers image processing and pattern recognition. Corresponding author of this paper.)

平西建 解放军信息工程大学信息工程学院教授. 主要研究方向为图像处理, 计算机视觉, 信息隐藏.
E-mail: pingxj@126.com
(**PING Xi-Jian** Professor at the Institute of Information Engineering, The Chinese People′s Liberation Army Information Engineering University. His research interest covers image processing, computer vision, and information hiding.)

徐 森 盐城工学院信息工程学院副教授. 主要研究方向为机器学习, 人工智能, 文本挖掘.
E-mail: xusen@hrbeu.edu.cn
(**XU Sen** Associate professor at the School of Information Engineering, Yancheng Institute of Technology. His research interest covers machine learning, artificial intelligence, and document mining.)

张 涛 解放军信息工程大学信息工程学院副教授. 主要研究方向为图像处理, 模式识别, 信息隐藏.
E-mail: ZhangT_77@163.com
(**ZHANG Tao** Associate professor at the Institute of Information Engineering, The Chinese People′s Liberation Army Information Engineering University. His research interest covers image processing, pattern recognition, and information hiding.)