图像分割的谱聚类集成算法

贾建华[1,2]，焦李成[1]，柳炳祥[2]
(1. 西安电子科技大学智能感知与图像理解教育部重点实验室，710071，西安；
2. 景德镇陶瓷学院信息工程学院，333002，江西景德镇)

摘要：针对谱聚类算法对尺度参数敏感的问题，利用集成学习算法良好的鲁棒性和泛化能力，提出了一种无监督集成学习算法——谱聚类集成算法. 该算法先利用谱聚类的内在特性产生集成学习所需的多个聚类个体，再采用 Hungarian 算法对生成的聚类个体进行重新标记，计算每个样本点关于每一个类别所占的比例，得到一个成分向量，然后运用对数比变换将所得的成分向量映射到另一个空间，去除成分数据的不适定性，最后对映射后的数据进行聚类，从而得到最终的集成结果. 通过对 UCI 数据集和纹理图像的仿真实验表明，所提算法的聚类准确率与常用的共识函数具有一定的可比性，且运算代价较小，所需时间大约为 MCLA 算法的一半，同时避免了精确选择谱聚类算法的尺度参数.

关键词：谱聚类；集成学习；Hungarian 算法；成分数据
中图分类号：TP391.41 **文献标志码**：A **文章编号**：0253-987X(2010)06-0093-06

A Spectral Clustering Ensemble Algorithm for Image Segmentation

JIA Jianhua[1,2], JIAO Licheng[1], LIU Bingxiang[2]
(1. Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi′an 710071, China; 2. School of Information Engineering, Jingdezhen Ceramic Institute, Jingdezhen, Jiangxi 333002, China)

Abstract: An unsupervised ensemble learning algorithm, spectral clustering ensemble, is proposed to solve the sensitivity of scaling parameter of spectral clustering. The algorithm utilizes the good robustness and generalization ability of ensemble learning. The property of spectral clustering is exploited to generate the components for integrating learning, and the Hungarian algorithm is used to relabel the resulting component clusterings. Then a compositional data vector is obtained by computing the ratio of each label for each sample. The compositional data vectors are mapped into another space via log contrast transform to solve the ill-posed characteristic of compositional data. The final aggregated results are generated by clustering the mapped data. Experiments on UCI data and texture images show that the proposed algorithm is comparable with some common consensus functions in accuracy, its computational cost is about half of that of MCLA, and it avoids the selection of the accurate parameter in spectral clustering.

Keywords: spectral clustering; ensemble learning; Hungarian algorithm; compositional data

聚类问题[1]是智能科学中的核心问题之一. 传统的聚类算法如 k 均值算法等都是建立在凸球形的样本空间上. 当样本空间不为凸时，算法会陷入局部最优. 新近出现的谱聚类算法[2]克服了 k 均值算法的缺点，具有识别非凸分布聚类的能力，而且实现简单，不会陷入局部最优解，且能避免数据的过高维数

收稿日期：2009-09-23. 作者简介：贾建华(1979—)，男，博士生；焦李成(联系人)，男，教授，博士生导师. 基金项目：国家“863计划”资助项目(2008AA01Z125，2009AA12Z210)；教育部长江学者和创新团队支持计划资助项目(IRT0645).

所造成的奇异性问题，但谱聚类算法自身也存在一些如计算量较大、构造相似性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.

集成学习将不同算法或者同一算法下不同参数得到的结果进行合并，从而得到比单个聚类算法更优的结果，具有更好的泛化能力[3-4]. 值得注意的是，Bagging 和 Boosting 算法[1]等其他大多数的集成学习算法都是为监督学习而设计的，对于聚类这样的非监督学习来说，由于没有训练样本的类别标记，聚类结果之间缺乏应有的先验信息，使得非监督集成学习比一般的监督学习的集成学习设计更加困难. Strehl 等[4]提出了 3 种基于超图划分的聚类集成算法(CSPA、HPGA 和 MCLA)，这些基于超图划分的算法对平衡的数据集集成效果较好，但 CSPA 算法由于计算量的原因而不适于较大规模数据. Topchy 等[5]提出了一种基于混合模型(mixture)的聚类集成算法，但算法本身需估计的参数较多，计算量较大且容易收敛到局部最优解. Fred[6]提出了一种基于关联矩阵的聚类融合方法，但算法需较大的存储量和计算量，不适合于大规模数据如图像数据等.

本文提出了一种无监督集成学习算法——谱聚类集成算法. 该算法利用谱聚类产生聚类集成所需的多个聚类个体，避免了精确选择谱聚类算法的尺度参数，再利用对成分数据的聚类方法将多个聚类结果综合起来，从而得到最终的集成结果. 所提的算法能够应用到较大规模的数据集如图像数据，和单个谱聚类算法相比，谱聚类集成算法有较好的鲁棒性和稳定性.

1 谱聚类算法和聚类集成

1.1 谱聚类算法

谱聚类算法[2]将聚类问题转化为一个无向图的多路划分问题. 数据点看成是无向图 $G(V,E)$ 中的顶点 V，加权边的集合 $E=\{S_{ij}\}$ 表示基于某一相似性度量的两点间的相似度，用 $\boldsymbol{S}$ 表示待聚类数据点之间的相似性矩阵. 图 G 中把聚类问题转变为在图 G 上的图划分问题，即将图 $G(V,E)$ 划分为 k 个互不相交的子集 $V_1,V_2,\cdots,V_k$，划分后每个子集 V_i 内部的相似程度较高，不同的集合 V_i 和 V_j 之间的相似程度较低. 本文采用规范切谱聚类算法(NJW 算法)[2]作为基学习器.

1.2 聚类集成

聚类集成首先由 Strehl 等人[4]提出. 与单一的聚类算法相比，聚类集成具有更好的鲁棒性、适用性、稳定性、并行性和扩展性[5]. 在聚类集成中，先要产生对数据集 X 的 H 个聚类成员，然后对这 H 个成员的聚类结果根据某个准则进行合并，因此聚类集成研究主要包括如何产生具有适当差异度的聚类成员，以及如何设计共识函数两个方面.

2 谱聚类集成

2.1 多样性聚类个体的产生

多样性已被证明是提高集成系统集成学习性能的关键因素. 将谱聚类算法作为个体学习器，结合谱聚类算法机理，采用多种扰动来增加聚类个体的多样性，这种策略对于谱聚类算法而言不仅容易实现，同时也部分解决了谱聚类算法存在的问题. 本文采用随机选择尺度参数[7-8]、Nyström 逼近[9]和随机初始化 k 均值算法[7-8]这 3 种方法来构造集成学习所需的谱聚类个体成员.

2.2 共识函数的设计

本文提出一种基于成分数据聚类的共识函数设计方法. 对于谱聚类算法产生的聚类个体，每个样本对应的标记是随机的，必须对所得到的每个个体的标记向量进行匹配. 最佳的匹配划分与真实划分相比应该具有最少的被错误标记的样本，本文将此问题看成最大化双向权值图问题[10]，首先构造一个相依表，其中的元素为 2 个不同聚类结果中样本所对应的标识对组合出现的次数，令 $C_r=\{1,1,2,2,2,3,3\}$ 和 $C_t=\{3,1,3,3,2,2,2\}$ 为对某数据集的 2 个划分，则其所对应的相依表如图 1 所示.

可以通过求解下列优化问题来得到最佳匹配

$$\max_{\{y_{ij}\}}\sum_{i=1}^{k}\sum_{j=1}^{k}\omega_{ij}y_{ij};\ \sum_{j=1}^{k}y_{ij}=\sum_{i=1}^{k}y_{ij}=1$$
$$y_{ij}\in\{0,1\} \tag{1}$$

式中：$\{\omega_{ij}\}$ 为相依表的元素值；y_{ij} 为指示变量值. 本文利用 Hungarian 算法[11]来求解式(1)，得到聚类结果标记之间的对应关系，如图 1 中虚线所示，即 2 个聚类结果标记 1 和 2 互换.

C_r \ C_t	1	2	3
1	1	0	1
2	0	1	2
3	0	2	0

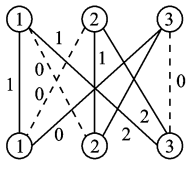


图 1 相依表和双向权值图

对于多个聚类个体,选择任意一个为参考向量.对于所有其余的非参考向量与参考向量做匹配,即求解式(1),得到所有个体的重新标记向量.

对于每一个样本点,计算其关于每个标记所占的比例,得到一个长度为分类类别数的成分向量 $\boldsymbol{P}_i=(P_{i1},P_{i2},\cdots,P_{iL})$,其中 L 为分类类别数,向量 $\boldsymbol{P}_i$ 中的元素 P_{ij} 为每个类别所占的比例.很明显,$\boldsymbol{P}_i$ 满足 $\sum_{j=1}^{L}P_{ij}=1$.成分数据向量由于其总和为 1,使得其具有不适定性,因此不能直接利用传统聚类方法进行聚类.

对于 L 维成分数据向量 $\boldsymbol{P}_i=(P_{i1},P_{i2},\cdots,P_{iL})$,对其做中心化对数比变换,将 $\boldsymbol{P}_i$ 映射成另一个空间中的数据点

$$Z_i=\ln\left[\frac{P_i}{\exp\left(1/L\sum_{j=1}^{L}\ln P_{ij}\right)}\right] \tag{2}$$

此时,向量可在 L 维空间 $\mathbf{R}^L$ 中任意取值,因为对数比变换是空间 $\{(P_{i1},P_{i2},\cdots,P_{iL})\mid\sum_{j=1}^{L}P_{ij}=1,P_{ij}>0\}$ 到 $\mathbf{R}^L$ 空间的一一映射[12].由此可以看出,中心化对数比变换有效地摆脱了成分数据的约束.对所得到的新样本 Z_i,本文利用传统的模糊 C 均值算法(FCM)进行聚类,得到了最终的共识划分结果.

综上所述,可以得到谱聚类集成实现策略如下:

(1)利用 2.1 节中的方法产生谱聚类个体集合;

(2)从聚类个体集合中随机选择一个聚类个体作为参考向量,计算所有其余非参考向量与参考向量的相依表,应用 Hungarian 算法完成标记的匹配;

(3)对每个样本,计算其属于每个类别的百分比,得到每个样本的成分向量;

(4)对成分向量做中心化对数变换,去除成分数据约束;

(5)应用 FCM 算法对变换后的数据进行聚类,得到最终的聚类结果.

3　实验结果与分析

3.1　UCI 数据集的划分

为了评价本文谱聚类集成算法的性能,我们选取 UCI 机器学习数据库中的部分数据集来验证算法的有效性.表 1 给出了这些数据集的特性.对本文算法性能的评价是将聚类结果与真实已知类属信息进行匹配后,以误分率作为评价标准.

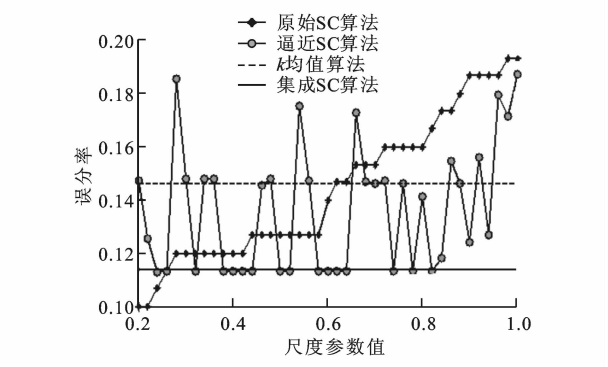
表 1　实验中所用 UCI 数据集属性

数据集	类别数	特征维数	数据量/个
Iris	3	4	150
Sonar	2	60	208
Wine	3	13	178
Breast_cancer	2	9	683
Segmentation	7	19	2 310

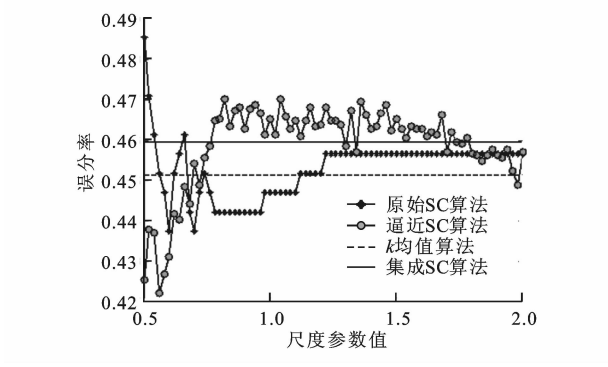
对于小样本数据,如 Iris、Sonar、Wine 和 Breast_cancer 数据集,分别采用原始谱聚类算法(SC)、基于 Nyström 的逼近谱聚类算法(SC_Nys)和本文的谱聚类集成算法(SCE)以及 k 均值算法进行聚类.单个 SC 算法和 SC_Nys 算法的尺度参数按照一定的步长在给定区间内取值.对于 Segmentation 数据,由于数据规模较大,仅采用后 3 种算法对其进行聚类.实验中,对于单个 SC_Nys 算法,从原始数据集中随机采样 100 个数据点为代表点,对于 SCE 算法随机采样 50 个数据点为代表点,尺度参数在给定的尺度参数区间随机取值,集成个体数为 30.图 2 和图 3 中所示的 SC 算法和 SC_Nys 算法的错误率是在给定某一尺度参数时 10 次运行结果的平均值.SCE 算法也是在给定的尺度参数范围内 10 次运行结果的平均值.

从图 2 可以看出,k 均值算法的聚类结果多数都与数据集本身的特性和初始迭代样本点有关,有很大的随机性,所以在某些数据集(如 Sonar)上的聚类结果要稍优于 SCE 算法.因为采用了逼近策略,所以 SC_Nys 算法所得的结果在同一尺度参数下绝大多数要劣于 SC 算法,但集成后的结果明显要优于 SC_Nys 算法,在很大的尺度参数范围内也要优于 SC 算法.从图 3a 可以看出,SC_Nys 算法的结果受尺度参数的影响很大,即使是同一个尺度参数,10 次独立实验的结果也存在不小的差异.图 3b 为 SC_Nys 算法在每个尺度参数 10 次运行的结果的平均错误率和集成 SC 算法 10 次结果的平均值和标准差,参数区间为[1,5]、[5,10]、[1,10].SCE 算法在尺度参数区间[1,5]上的性能要优于在区间[1,10]和[5,10]上的性能,这是由于个体谱聚类在尺度参数区间[1,5]上所得的单个谱聚类的性能要优于其在区间[5,10]、[1,10]上的性能.这也说明,在集成学习中,除多样性外,个体精度同样非常重要.

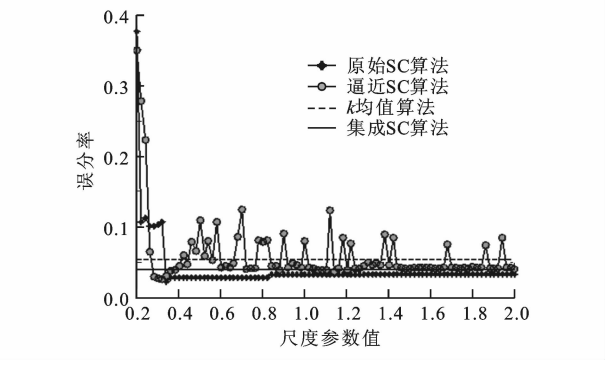
表 2 给出了各种算法在不同数据集上的运行时间,计算机配置为:双核 1.86 GHz Pentium IV 处理器,2GB内存,使用Matlab7.01语言编程.从表2



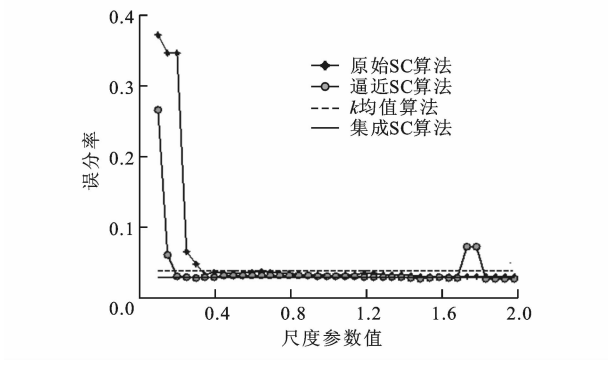
(a)Iris 数据集



(b)Sonar 数据集

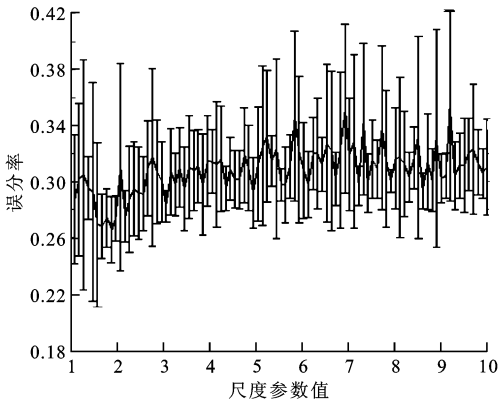


(c)Wine 数据集

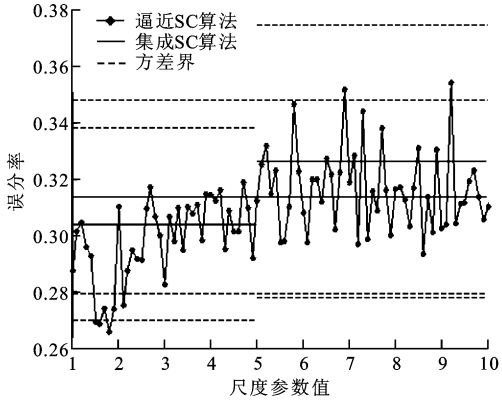


(d)Breast_cancer 数据集

图 2 小规模 UCI 数据集的聚类结果比较



(a)SC_Nys 算法 10 次实验的平均错误率及方差



(b)逼近 SC 算法和集成 SC 算法 10 次实验的平均错误率和方差

图 3 较大规模数据集(Segmentation 数据)的聚类结果

表 2 不同共识算法在 UCI 数据集上的运算时间

数据集	个体数	运算时间/s				
		CSPA 算法	HGPA 算法	MCLA 算法	Mixture 算法	本文算法
Iris	30	0.234	0.187	0.172	3.844	0.062
Sonar	30	0.328	0.203	0.265	6.484	0.094
Wine	30	0.266	0.234	0.359	9.930	0.094
Breast_cancer	30	1.484	0.219	0.187	10.750	0.078
Segmentaion	30	7.954	1.141	0.844	130.797	0.531

可看出,本文的算法所需的运算代价最小,这对于处理大规模数据如图像数据而言是比较有意义的.

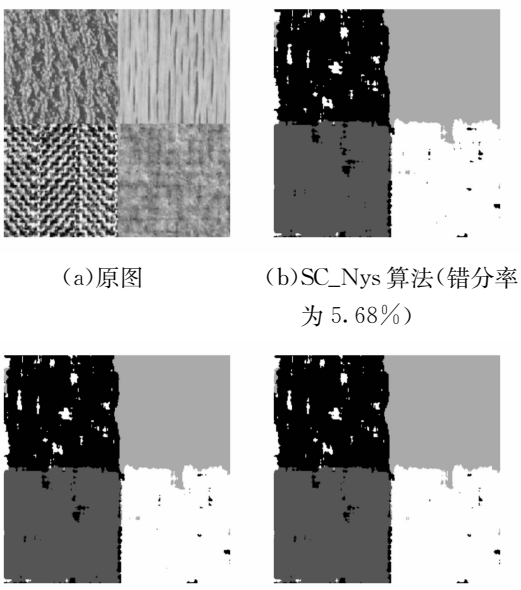
3.2　纹理图像分割

无监督图像分割可以看成是一个聚类问题,以下利用谱聚类集成对纹理图像进行分割来评价本文算法的性能.在下面的实验中,对图像每个像素点提取非下采样小波能量特征,采用 DB3 小波,分解 3 层,总共 10 维特征,每个像素点选取的变换窗口大小为 15×15 像素,边界像素采用镜像延拓.在聚类前对特征向量进行对应分量归一化,尺度参数在所给定的参数区间随机选取,集成学习器包含 30 个谱聚类个体.

由于单个 SC 算法和 CSPA 算法所需存储空间和计算量太大,HGPA 算法在图像分割集成中效果较差,基于混合模型的集成方法不仅计算量太大且易于收敛到局部最优,所以本文只采用 SC_Nys 算法、MCLA 集成算法以及本文提出的 SCE 算法对图像进行分割,Nyström 逼近采样 100 个像素点作为代表点.

3.2.1　合成纹理图像分割　图 4a 是一幅 4 类的合成纹理图像,采用前面所定义的特征,尺度参数区间为[1,5].图 4 给出了 SC_Nys 算法(30 个个体谱聚类的最优结果)、基于 MCLA 的谱聚类集成算法和本文算法的分割结果.

由图 4 可知,本文算法在视觉效果上和分割精度上与 MCLA 算法相差不大,误分率有所降低.

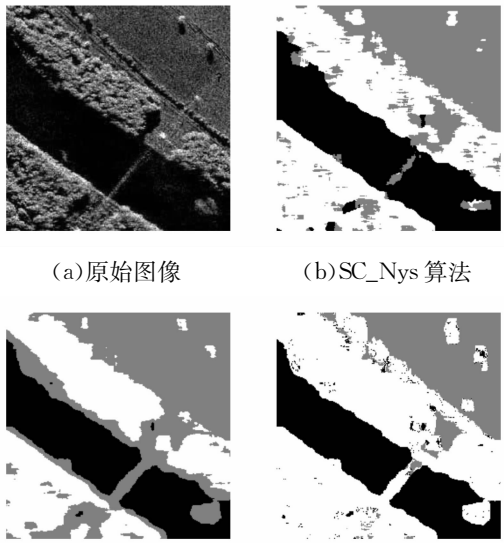


(a)原图　(b)SC_Nys 算法(错分率为 5.68%)

(c)MCLA 算法(错分率为 5.32%)　(d)本文算法(错分率为 4.94%)

图 4　采用 3 种算法对合成纹理图像的分割结果

3.2.2　SAR 图像分割　将本文算法应用于合成孔径雷达(SAR)图像分割,实验参数同上.图 5a 为一 Ku 波段 SAR 图像,包含 3 类地物:河流、植被和平原地带,大小为 256×256 像素,谱聚类个体的尺度区间为[1,5].分别采用 SC_Nys 算法、基于 MCLA 的谱聚类集成算法和本文算法的分割结果见图 5.



(a)原始图像　(b)SC_Nys 算法

(c)MCLA 集成算法　(d)本文算法

图 5　SAR 图像的分割结果

从视觉效果看,对于河流和植被的边界区域,本文的算法能正确地区分,而基于 MCLA 的谱聚类集成算法对此边界的区分不是很好.从计算代价来看,表 3 给出了 2 种算法的计算时间,本文算法具有较小的计算代价,运算时间同样大约为 MCLA 算法运算时间的一半.

表 3　不同图像的共识算法运算时间

图像	运算时间/s	
	MCLA 算法	本文算法
图 4a	10.969	5.750
图 5a	8.906	4.423

4　结　论

本文提出了一种无监督聚类集成的学习算法,利用谱聚类算法的优点,将谱聚类作为基学习器来构造集成学习系统,最后运用一种基于成分数据聚类的共识函数将谱聚类个体的结果综合起来.所设计的共识函数与现存的几种共识函数,特别是与常用的 MCLA 相比,在分类性能差别不大的情况下,运算时间约减少了一半,且在图像分割中取得了较为满意的分割结果.

参考文献：

[1] DUDA R O, HART P E, STORK D G. Pattern classification [M]. New York, USA: Wiley Interscience Publication, 2001.

[2] NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm [C]// Proceedings of Advance of Neural Information Processing System. Cambridge, MA, USA: MIT Press, 2002: 849-856.

[3] DIETTERICH T G. Machine-learning research: four current directions [J]. AI Magazine, 1997, 18(4): 97-136.

[4] STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions [J]. The Journal of Machine Learning Research, 2002, 3(12): 583-617.

[5] TOPCHY A, JAIN A K. Clustering ensembles: models of consensus and weak partitions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881.

[6] FRED A L N, JAIN A K. Combining multiple clusterings using evidence accumulation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850.

[7] ZHANG Xiangrong, JIAO Licheng, LIU Fang, et al. Spectral clustering ensemble applied to SAR image segmentation [J]. IEEE Transactions on Geoscience and Remote Sensing, 2008, 46(7): 2126-2136.

[8] LOURENCO A, FRED A L N. Comparison of combination methods using spectral clustering ensembles [C]// Proceedings of the 4th International Workshop on Pattern Recognition in Information Systems. Setubal, Portugal: INSTICC Press, 2004: 222-233.

[9] FOWLKES C, BELONGIE S, CHUNG F, et al. Spectral grouping using the Nyström method [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(2): 214-225.

[10] TOPCHY A P, LAW M H C, JAIN A K, et al. Analysis of consensus partition in cluster ensemble [C]// Proceedings of the 4th IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE Press, 2004: 225-232.

[11] MUKRES J. Algorithms for the assignment and transportation problems [J]. Journal of the Society for Industrial and Applied Mathematics, 1957, 5(1): 32-38.

[12] AITCHISON J. The statistical analysis of compositional data [J]. Journal of the Royal Statistical Society, Series B: Methodological, 1982, 44(2): 137-177.

（编辑 刘杨 苗凌）