

文章编号:1007-6735(2014)01-0049-04

一种新的层次谱聚类算法

杨晓慧, 王莉莉, 李登峰

(河南大学 数学与信息科学学院, 开封 475004)

摘要: 提出一种新的聚类算法——层次谱聚类算法. 该算法在传统二分的 SM 谱聚类过程中嵌入了层次聚类算法, 目的是为了提高谱聚类的聚类正确率, 同时又利用谱聚类纠正了层次聚类过程中所得到的歪斜划分. 实验结果表明: 提出的层次谱聚类算法的聚类正确率比层次聚类算法、谱聚类算法的聚类正确率都要高, 同时又纠正了层次聚类过程中的歪斜划分.

关键词: 层次聚类; 谱聚类; 层次谱聚类
中图分类号: TP 301.6 **文献标志码:** A

New Hierarchical Spectral Clustering Algorithm

YANG Xiao-hui, WANG Li-li, LI Deng-feng

(College of Mathematics and Information Sciences, Henan University, Kaifeng 475004, China)

Abstract: A new clustering algorithm—hierarchical spectral clustering algorithm was proposed. In the algorithm the hierarchical clustering was embedded in traditional dichotomy SM spectral clustering, in order to raise the accuracy of spectral clustering and check skewed divisions in the process of hierarchical clustering. The experimental results show that the hierarchical spectral clustering is superior to spectral clustering or hierarchical clustering on clustering accuracy and checking skewed divisions in the process of hierarchical clustering.

Key words: hierarchical clustering; spectral clustering; hierarchical spectral clustering

聚类算法在数据分析和模式识别领域都扮演着重要的角色, 其目的是将相似对象聚为一类. 在目前计算机视觉的研究中存在的困难是如何有效地提高聚类算法的性能. 作为一种有效的数据分析方法, 聚类算法在计算机视觉、信息检索及数据挖掘等领域都有广泛的应用. 聚类搜索策略是当前研究的一个热点. 1998 年, Iyengar 等^[1]利用聚类算法已达到对大型数据库进行有效的访问. 2002 年, Saux 等^[2]提

出利用图像聚类可以从更好的角度帮助用户快速的从大型数据库中找到所要找的图像. 2003 年, Käster 等^[3]提出了一种利用图像分割技术实现图像聚类的方法. 在聚类算法的研究中, 代表性的常用聚类算法有: LBG 算法^[4-6]、K-means 算法^[7-11]、谱聚类算法^[12-15]和层次聚类算法^[10, 16-18]. 层次聚类算法可以提高聚类精确度, 而谱聚类算法能够尽可能的平衡分割, 而这正是所要测试的两个图像库所需

收稿日期: 2012-10-03
基金项目: 国家自然科学基金资助项目(61071189), 河南省青年骨干教师资助项目(2013GGJS-027), 河南省教育厅科技攻关重点资助项目(14A120009)
第一作者: 杨晓慧(1978-), 女, 副教授. 研究方向: 小波分析、图像处理和模式识别. E-mail: xhyang@henu.edu.cn
通信作者: 李登峰(1964-), 男, 教授. 研究方向: 调和分析、小波分析及其应用. E-mail: wavelet2009@126.com

要的.于是提出了一种层次谱聚类算法,它融合了两种聚类算法的优点,并且抑制了两者的缺点.实验表明,层次谱聚类算法在聚类精确度上优于谱聚类算法,相对于层次聚类算法又大大减少了运算时间.

文中介绍了层次谱聚类算法的具体实施过程及两种聚类评价标准.并选取 Wang 图像库为实验图像库,对比试验结果.得出了层次谱聚类算法的聚类正确率高于层次聚类算法、谱聚类算法的结论.

1 层次谱聚类

根据层次是自底向上还是自顶向下形成,可以将层次聚类算法分为合并型层次聚类算法和分裂型层次聚类算法两种.两者的不同之处在于合并型算法初始时,每一个成员都组成一个单独的簇,在以后的迭代过程中,再将那些相邻的簇合并为一个簇,直到所有的成员组成一个簇为止.而分裂型算法则在初始时将所有元组归于同一簇,然后将上层的簇重复的分裂为两个下层簇,直到每一个元组都组成一个单独的簇为止.本文聚类的初始状态是将每一幅图像视为一类,所以文中采用了合并型层次聚类算法.谱聚类来源于谱图划分准则,是一种受欢迎的功能强大的计算方法.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 $G(V, E)$ (如图 1) 的顶点 V , 边权重的集合 $E = \{W_{ij}\}$, 表示基于某一相似性度量计算的两点间的相似性, W 表示待聚类数据点间的相似性矩阵, 将其看做是该无向图的邻接矩阵, 它包含了聚类所需要的所有信息.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.本文采用的谱聚类算法是由 Shi 和 Malik 提出的 SM 算法^[12]. 它作为一个启发式算法, 目标在于最小化由同一作者提出的规范切准则(normalized cut, NCut)^[12].

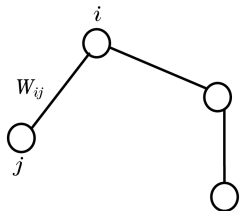


图 1 无向图 $G(V, E)$

Fig.1 Undirective graph $G(V, E)$

充分融合层次聚类算法较高的聚类精确度的优点和谱聚类算法尽可能平衡分割的优点,将两者结

合并提出了一种层次谱聚类算法.具体步骤如下:

步骤 1 用 SM 聚类算法将整个图像库分为 S_1 和 S_2 两类.

步骤 2 比较 S_1 和 S_2 哪一类所包含的节点多,不失一般性,假设 S_2 包含的节点多于 S_1 ,对 S_2 施行层次聚类.当所要合并的两类之间的距离大于等于阈值 T 时,层次聚类终止(T 为图像库中任意两幅图之间距离的均值).

步骤 3 计算对 S_2 施行层次聚类所得类的类中心,并对其施行 SM 聚类.

步骤 4 重复步骤 2、步骤 3,直到得出所需的类数为止.

2 聚类的评价标准

2.1 聚类正确率

聚类正确率是一种常用的聚类算法评价标准,它将聚类结果和已知的真实类属信息进行匹配后,得出聚类的正确率.其计算式为

$$Accuracy = \frac{a + b}{n(n-1)/2} \quad (1)$$

式中, n 为图像库中所包含的图像个数; a 为两幅图像在已知的真实类属信息中属于同一类,而聚类所得的结果中他们也属于同一类的对数; b 为两幅图像在已知的真实类属信息中不属于同一类,而聚类所得的结果中它们也不属于同一类的对数.

2.2 聚类的纯度

聚类所得的结果在一类中可能会包含属于不同语义的对象.聚类的纯度是指一类中主语义类所占的百分比.假设对于某一类 C_j 中有 n 幅图像属于 c 个语义(在试验中, $c \leq 10$),那么该类的纯度计算式为

$$p(C_j) = \frac{1}{|C_j|} \max_{k=1, \dots, c} |C_{j,k}| \quad (2)$$

式中, $|C_j|$ 为 C_j 类中所包含的图像个数; $|C_{j,k}|$ 为 C_j 类中所包含的语义 k 的图像总数.

3 实验结果及分析

文中所有程序的运行环境为 matlab R2010a on Dual-core Intel (R) Pentium (R) CPU P6000 @ 1.87 GHZ, 512M memory, operating system: windows7. 对 Corel Database 的子图像库 Wang Database 进行试验. Wang Database 共包含 10 类,每类

中包含 100 幅图像. 这 10 类分别是 Africa people and Villages, Beach, Buildings, Buses, Dinosaurs, Elephants, Flowers, Horses, Mountains and Glaciers 和 Food. 图 2 给出了每类中的一个代表图像.



图 2 Wang database 中每类的代表图像

Fig.2 Example images of each category in Wang database

文献[21]中提出的 MPEG-7 边缘直方图特征能够保留传统直方图的强度,并包含有图像的边缘连通性和区域块边缘模式的连续性信息,因此基于图像的 MPEG-7 边缘直方图特征进行聚类. 两幅图像之间的相似性用两幅图像的直方图 A, B 之间的距离 $D(A, B)$ 来度量

$$D(A, B) = \sum_{i=1}^n |A(i) - B(i)| \quad (3)$$

式中, $A(i), B(i)$ 分别为图像的直方图 A, B 的第 i 个直方条的度量值.

表 1 和表 2 是分别对不同特征用 LBG 和 K-Means 算法所得的结果^[22]和本文实验结果的对比. 表中(1)代表不变特征直方图, (2)代表不变特征关系直方图, (3)代表 Tamura 特征直方图. 对比结果表明层次聚类算法所得的聚类正确率是最高的, 这主要得益于在谱聚类算法的过程中运用了层次聚类算法, 而层次聚类算法可以提高聚类的正确率.

表 1 不同特征用 LBG 算法^[22]和本文结果的比较

Tab.1 Comparison between the results by LBG algorithm^[22] and our results for different features

算法	(1) LBG	(2) LBG	(3) LBG	(1)(2) LBG
正确率/%	46	22	25	43
算法	(1)(3) LBG	(1)(2)(3) LBG	层次谱聚类	
正确率/%	48	57	81	

表 3 和图 3 及图 4 为层次聚类、谱聚类、层次谱聚类 3 种聚类算法所得结果的比较, 所采用的图像特征和相似性度量与文献[21]中相同.

表 3 是 3 种聚类算法所得结果的聚类正确率和计算时间的比较, 从表 3 可以看出层次谱聚类的聚类正确率比层次聚类、谱聚类的聚类正确率都高. 这再次证明了在谱聚类过程中采用层次聚类有助于提高聚类的正确率. 而层次谱聚类所用的时间比层次聚类所用的时间少, 却远远多于谱聚类所用的时间,

表 2 不同特征用 K-Means 算法^[22]和本文结果的比较

Tab.2 Comparison between the results by K-Means algorithm^[22] and our results for different features

算法	(1) K-均值	(2) K-均值	(3) K-均值	(1)(2) K-均值
正确率/%	47	24	27	53
算法	(1)(3) K-均值	(1)(2)(3) K-均值	层次谱聚类	
正确率/%	54	58	81	

这是因为层次聚类的计算复杂度较高, 在谱聚类过程中运用层次聚类虽然提高了聚类正确率, 却大大延长了运算时间.

表 3 3 种聚类算法的聚类正确率和计算时间

Tab.3 Accuracy and computing time consumption of the three clustering algorithms

算法	层次聚类	谱聚类	层次谱聚类
正确率/%	17	79	81
计算时间/s	40089	44	17020

图 3 是层次聚类、谱聚类和层次谱聚类所得结果的每类中的图像个数(按从多到少排列)比较, 横坐标表示图像类, 纵坐标表示每类中包含图像的个数. 从图 3 的结果可以看出层次聚类在聚类过程中得到了歪斜划分, 而谱聚类和层次谱聚类所得的聚类结果每类中所含的图像个数相对平均, 这是由于谱聚类有尽可能平衡分割的特点, 而这是 Wang Database 所需要的.

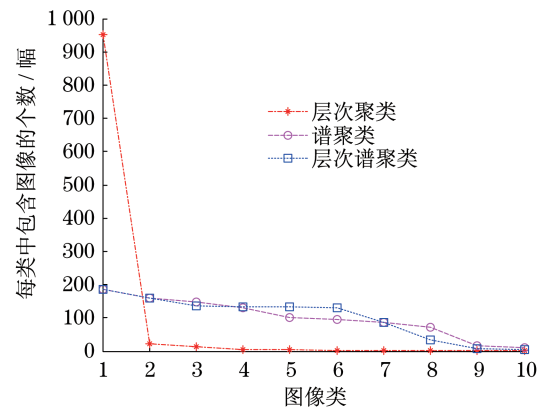


图 3 W 层次聚类、谱聚类和层次谱聚类所得的每类中的图像个数比较

Fig.3 Comparison of the number of images in each cluster of hierarchical clustering, spectral clustering and hierarchical spectral clustering

图 4(见下页)给出 3 种聚类算法所得结果中每类的纯度比较, 横坐标表示图像类, 纵坐标表示图像类中图像的纯度. 从图 4 的结果可以看出层次谱聚

类所得的结果除了第1类(Africa people and Villages)和谱聚类所得的结果一致外,其它9类均优于谱聚类的结果.而层次聚类算法得到的是歪斜划分,10类中有5类都只含有1幅图像(只包含1幅图像的类纯度当然是100%),因此对于本文的实验图像库其纯度结果不具有参考价值.这同时说明在谱聚类过程中运用层次聚类可以提高聚类的正确率.

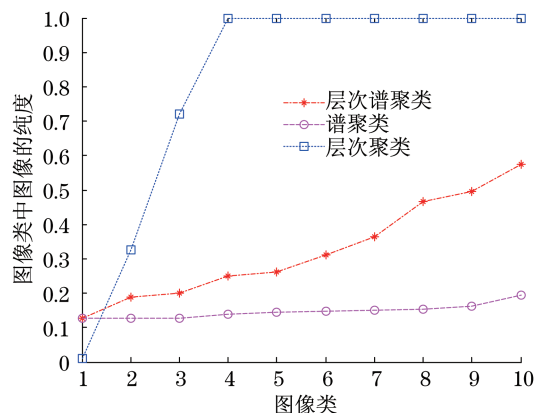


图4 3种聚类算法每类的纯度比较

Fig.4 Comparison of cluster purity of the three clustering algorithms

3 结论

充分考虑了各种聚类算法的优缺点,将层次聚类和谱聚类结合在一起,提出的层次谱聚类算法吸收了层次聚类算法和谱聚类算法的优点.实验结果表明,层次谱聚类算法既避免了聚类过程中的歪斜划分,又比谱聚类算法提高了聚类正确率,同时又比层次聚类算法减少了运算时间.但其运算时间仍然较长,且其聚类正确率有待于进一步提高.希望以后能够寻找到更优的聚类算法.

参考文献:

- [1] Iyengar G, Lippman A B. Clustering images using relative entropy for efficient retrieval [C] // International Workshop on Very Low Bitrate Video Coding, Urbana, 1998.
- [2] Saux B L, Boujemaa N. Unsupervised robust clustering for image database categorization [C] // Proceeding International Conference Pattern Recognition Quebec, Canada; IEEE, 2002; 259 - 262.
- [3] Käster T, Wendt V, Sagerer G. Comparing clustering methods for database categorization in image retrieval [J]. Pattern Recognition, 2003, 2781; 228 - 235.
- [4] Linder Y, Buzo A, Gray R M. An algorithm for vector quantization design [J]. Proceeding IEEE Transaction Communications Society, 1980, 28(1); 84 - 95.
- [5] Gersho A, Gray R M. Vector Quantization and Signal Compression [M]. Boston: Kluwer Academic, 1991.
- [6] Kekre H, Sarode T, Bharadi V, et al. Iris recognition using vector quantization [C] // International Conference Signal Acquisition and Processing, Bangalore: IEEE, 2010; 58 - 62.
- [7] Bradley P, Fayyad U. Refining initial points for K-means clustering [C] // Proceeding International Conference on Machine Learning, San Francisco: Morgan kaufmann publishers Inc, 1998; 91 - 99.
- [8] Liu H, Yu X H. Application research of K-means clustering algorithm in image retrieval system [C] // Proceeding of the Second Symposium International Computer Science and Computation Technology, Huangshan, 2009; 274 - 277.
- [9] Yang Y, Xu D, Nie F P, et al. Image clustering using local discriminate models and global integration [J]. IEEE Transactions on Image Processing, 2010, 19 (10); 2761 - 2773.
- [10] Chen T S, Tsai T H, Chen Y T, et al. A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray [C] // Proceeding of 2005 International Symposium on Intelligent Signal Processing and Communication System, Hong Kong; IEEE, 2005; 405 - 408.
- [11] Honda K, Notsu A, Ichihashi H. Fuzzy PCA-guided robust K-means clustering [J]. IEEE Transaction Fuzzy Systems, 2010, 18(1); 67 - 79.
- [12] Shi J, Malik J. Normalized cuts and image segmentation [J]. IEEE Transaction Pattern Analysis and Machine Intelligence, 2000, 22(8); 888 - 905.
- [13] Xu L, Li W, Schuurmans D. Fast normalized cut with linear constraints [C] // IEEE Conference on Computer Vision and Pattern Recognition, Florida: IEEE, 2009; 2866 - 2873.
- [14] Li Z, Liu J, Tang X. Constrained clustering via spectral regularization [C] // IEEE Conference on Computer Vision and Pattern Recognition, Miami: IEEE, 2009; 421 - 428.
- [15] Ning H, Xu W, Chi Y, et al. Incremental spectral clustering by efficiently updating the eigen-system [J]. Pattern Recognition, 2010, 43(1); 113 - 127.
- [16] Bandyopadhyay S. An automatic shape independent clustering technique [J]. Pattern Recognition, 2004, 37 (1); 33 - 45.

(下转第59页)

- [24] Brown A B, Dabberdt W F. Modeling ventilation and dispersion for covered roadways[J]. Journal of Wind Engineering and Industrial Aerodynamics, 2003, 91(5):593-608.
- [25] Chen T Y, Lee Y T, Hsu C C. Investigations of piston-effect and jet fan-effect in model vehicle tunnels[J]. Journal of Wind Engineering and Industrial Aerodynamics, 1998, 73(2):99-110.
- [26] Ke M T, Cheng T C, Wang W P. Numerical simulation for optimizing the design of subway environmental control system[J]. Building and Environment, 2002, 37(11):1139-1152.
- [27] Maele K V, Merci B. Application of RANS and LES field simulations to predict the critical ventilation velocity in longitudinally ventilated horizontal tunnels[J]. Fire Safety Journal, 2008, 43(8):598-609.
- [28] Betta V, Cascetta F, Musto M, et al. Fluid dynamic performances of traditional and alternative jet fans in tunnel longitudinal ventilation systems[J]. Tunnelling and Underground Space Technology, 2010, 25(4):415-422.
- [29] 王福军. 计算流体力学分析 - CFD 软件原理与应用 [D]. 北京:清华大学出版社, 2006.
- [30] Wu K H, Xu H T. Comparison of 3D CFD simulation approach for aerodynamics effects in high speed railway tunnel system [C] // 15th International Symposium on Aerodynamics, Ventilation & Fire in Tunnels. Barcelona, 2013:433-441.
- [31] Wu K H, Lin Y F, Pin Y F, et al. Efficient numerical simulation of ventilation system design of road tunnels [C] // Proceedings of the ASME 2012 International Mechanical Engineering Congress & Exposition, Houston, 2012.
- [32] Daly B B. Woods Practical Guide to Fan Engineering [M]. Colchestre: Woods of Colchester Ltd, 1992.
- (编辑:金 虹)

.....

(上接第 52 页)

- [17] Maqbool O, Babri H. Hierarchical clustering for software architecture recovery[J]. IEEE Transactions on Software Engineering, 2007, 33(11):759-780.
- [18] Cilibrasi R, Vitanyi P. A fast quartet tree heuristic for hierarchical clustering[J]. Pattern Recognition, 2011, 44(3):662-677.
- [19] Jain A K, Dubes R C. Algorithms for clustering data [M]. Englewood Cliff: Prentice-Hall, 1988.
- [20] Saporta G, Youness G. Comparing two partitions: some proposals and experiments [C] // Compstat Berlin: Physical-Verlag HD, 2002:243-248.
- [21] 康勤. 基于 MPEG-7 边缘直方图描述符的图像检索算法[J]. 西南大学学报, 2008, 30(5):149-153.
- [22] Deselaers T, Ney H, Keysers D. Features for image retrieval [D]. Aachen: RWTH Aachen University, 2003:77-79.
- (编辑:金 虹)