谱聚类算法研究及其在图像分割中的应用

肖 潇 史 惠 孔凡之
(浙江传媒学院电子信息学院 杭州 310018)

摘 要 提出了一种改进谱聚类的算法。首先介绍了谱聚类和基于路径的思想;然后为了改善传统谱聚类算法对GAUSS函数中尺度参数的敏感性,提出了一种新算法,并将其扩展到半监督的情况;最后将其应用在图像分割实验中,证明了该算法的有效性。

关键词 谱聚类,尺度参数,相似度,监督,图像分割
中图法分类号 TP391 **文献标识码** A

Spectral Clustering Algorithm and its Application in Image Segmentation

XIAO Xiao SHI Hui KONG Fan-zhi
(College of Electronic Information, Zhejiang University of Media and Communications, Hangzhou 310018, China)

Abstract An improved spectral clustering algorithm was proposed in this paper. Firstly, the spectral clustering based on the idea of path was introduced. And then in order to improve the sensitivity of traditional spectral clustering on scale GAUSS function parameters, this paper put forward a new algorithm, which is extended to the semi-supervised situation. At last the algorithm was applied in the experiments of image segmentation, and the effectiveness of the algorithm was proved.

Keywords Spectral clustering, Scale parameter, Similarity, Supervision, Image segmentation

1 引言

在当今这个信息爆发的时代,图像是人类获取信息、传递信息及表达信息的重要方式,图像处理技术也因此在社会生活中也发挥着重要的作用。图像分割是图像处理到图像分析的关键,是图像理解和识别的基础,因此,图像分割手段是否有效也决定了图像处理效果是否理想。谱聚类算法是近年来机器学习领域的研究热点,与传统的聚类算法相比,它具有能够在任意形状的样本空间上聚类且收敛于全局最优解的优点,适用于解决许多实际聚类问题。近年来谱聚类算法在模式识别和图像分割中的应用研究受到了众多学者的广泛关注。

目前谱聚类算法在图像分割中的研究已有了一些进展,尽管谱聚类方法具有很多优势,并在实践中也取得了良好的效果[1],但是样本的密度和数量都影响着聚类算法的稳定性,所以仍有许多亟需研究和解决的问题。文献[2]提出了一种基于免疫谱聚类的图像分割方法,利用谱聚类的维数缩减的特性获取数据映射空间中的分布,在此基础上构造一种免疫克隆聚类算法,将其用于在映射空间中的样本聚类。文献[3]从谱聚类和权核K-均值的等价性出发,结合图像的空间一致特性,提出了一种空间约束特性的谱聚类算法。文献[4]采用特征向量组的选择性集成方法以提高谱聚类性能,其涉及基特征向量组的选取、选择性集成策略等问题,解决了谱聚类中K个最大特征值对应的特征向量不一定使聚类效果达到最好的问题。

本文将谱聚类算法应用到图像分割中,提出了一种基于改进的相似度参数估计半监督谱聚类算法,通过最小生成树优化了相似度的求取算法,并对算法的时间复杂度进行了分析。仿真实验表明,本算法不仅能够对图像进行比较准确的分割,而且能以较少的时间代价获得最优阈值。

2 基于路径的谱图聚类算法

2.1 谱聚类

谱聚类[5]是近年提出的基于谱图理论的聚类算法,它通过对样本点之间相似度矩阵的谱分析获得聚类结果。谱聚类算法没有迭代过程,所以避免了K-means聚类算法可能陷入局部极小值的情况。因此谱聚类算法一经提出就引起了学者的广泛关注。谱聚类的算法流程可以表示如下(假设有 k 个类):

Begin

Step1 对数据集构建一个相似度矩阵 $W\in M_{N\times N}$,其中如果 $i\neq j$,则 $w_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2)$;否则 $w_{ij}=0$;

Step2 定义对角矩阵 D,其中 $D_{ii}=\sum_j w_{ij}$,同时构造矩阵 $L=D^{-1/2}WD^{-1/2}$;

Step3 求矩阵 L 的 k 个最大特征值所对应的特征向量,组成矩阵 $U=[u_1,u_2,\cdots,u_k]\in M_{N\times k}$,并对矩阵 U 做行归一化,得到矩阵 $Y(Y_{ij}=U_{ij}/(\sum_j U_{ij}^2)^{1/2})$;

Step4 将 Y 的每一行视为 R^k 中间的一点,使用 K-means 对这些数据点进行聚类,最后将由数据点得到的类别标签按照顺序赋值给原始数据。

END

本文受浙江省科技厅公益项目:基于多屏互动的数字电视节目交互技术的研究和开发(2013C33G2240030)资助。
肖 潇(1980—),女,博士,讲师,主要研究方向为图像处理与模式识别。

图 1 示出了谱聚类算法的聚类效果，左图是原始样本集，共有 3 个封闭的圆形，每一个圆形(不同颜色、不同形状)可以看做是一类数据样本，共有 600 个样本点，每一类都有包含 200 个样本点，右图是聚类结果。可以看出，聚类的效果是比较理想的。

图 1 传统谱聚类算法效果

文献[6]进行了多次的实验，比较充分地说明了谱聚类算法的有效性。但是也有不同的情况，比如对于图 2 的样本集，谱聚类算法就无法得出正确的聚类结果。图 2 中左图是原始样本，其是由 3 条螺旋线组成的，每一条螺旋线(不同颜色、不同形状)代表一类样本，共有 63 个样本点，每一类 21 个，右图是聚类结果图，可以看出聚类效果不理想。

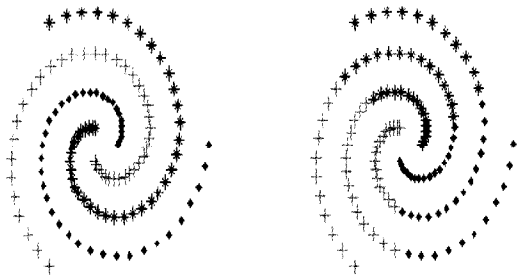

图 2 传统谱聚类算法效果

上述实验说明了谱聚类算法也存在一定的局限性。很多学者同时提出了谱聚类算法的一些改进算法[7,8]，其中之一就是利用基于路径的样本点之间的相似度来代替采用欧氏距离和 Gauss 核计算得到的相似度，也就是基于路径的谱聚类算法。

2.2 基于路径的谱聚类

谱聚类中样本点间的相似度定义如下：

$$w_{ij}=\begin{cases}\exp\left[-\dfrac{\|x_i-x_j\|^2}{2\sigma^2}\right], & i\neq j\\ 0, & i=j\end{cases} \tag{1}$$

基于路径的相似度[9,10]定义如下：如果样本点构成一个连通图 G，令 P_{ij} 表示图上顶点 i 到顶点 j 的所有路径，对每一条路径 $p\in P_{ij}$，x_i 和 x_j 之间的相似度为路径上的最小权重边的值，也就是路径上相似度最小的两个相邻样本点的相似度，考虑到所有 $p\in P_{ij}$，那么 x_i 和 x_j 之间基于路径的相似度是：

$$w_{ij}=\max_{p\in P_{ij}}\{\min_{1\leqslant h\leqslant|p|}w_{p[h]p[h+1]}\} \tag{2}$$

其中，$p[h]$表示路径上第 h 个顶点。重新定义了样本点之间的相似度后，仍然使用上述例子来观察基于路径的谱聚类算法的有效性。图 3 所示为聚类效果，左图为原始数据，右图为聚类结果。可见基于路径的谱聚类算法对于图 2 中样本点的聚类效果也是理想的。

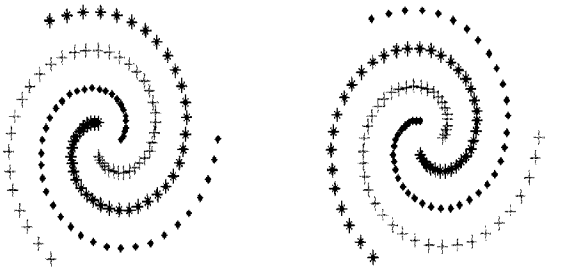

图 3 基于路径的谱聚类算法聚类效果

但是，需要指出的是，基于路径的谱聚类算法除了相似度有所改进，其他操作与谱聚类算法并无不同，仍然存在一个尺度参数 σ，这个 σ 的选取对聚类结果有很大影响，在上面实验中，通过反复尝试，我们取 $\sigma=3$，当选择其他参数时效果都不是很理想。

为此，我们考虑是否可以采用更鲁棒的方法来确定数据间相似度，文献[11]中采用二次规划方法求取相邻样本点间相似度的方法是比较有效的。

3 改进的相似度参数估计谱聚类算法

3.1 相似度矩阵

对任何数据集，样本点的邻域都可以看作是线性的，假设每一样本点在其邻域内都可以由其邻近点的线性组合来近似重构，如果限定这些重构权值的取值范围，令其必须大于或等于零，那么这个权值就可以看作是数据点和它近邻点之间的一个相似度。

所以，目标就是最小化下式：

$$\varepsilon=\sum_i\|x_i-\sum_{j:x_j\in N(x_i)}w_{ij}x_j\|^2 \tag{3}$$

其中，规定 $\sum_{j:x_j\in N(x_i)}w_{ij}=1, w_{ij}\geqslant 0$。

显然如果 x_i, x_j 越相似，则 w_{ij} 越接近 1；反之则越接近 0，因此 w_{ij} 确实能作为 x_i, x_j 之间相似度的一个量度。需要注意的是，w_{ij} 和 w_{ji} 一般是不相等的。

根据式(3)有如下的推导：

$$\begin{aligned}\varepsilon&=\sum_i\|x_i-\sum_{j:x_j\in N(x_i)}w_{ij}x_j\|^2\\&=\|\sum_{j:x_j\in N(x_i)}w_{ij}(x_i-x_j)\|^2\\&=\sum_{j:x_j\in N(x_i)}w_{ij}w_{ik}(x_i-x_j)^{\mathrm{T}}(x_i-x_k)\\&=\sum_{j:x_j\in N(x_i)}w_{ij}G^i_{jk}w_{ik}\end{aligned}$$

其中，G^i_{jk} 表示 Gram 矩阵 $(G^i)_{j,k}=(x_i-x_j)^{\mathrm{T}}(x_i-x_k)$ 的第 (i,j) 个元素。因此求解重构权值可以通过求解下面 N 个二次规划的问题来得到：

$$\begin{aligned}&\min_{w_{ij}}\sum_{j,k:x_j,x_k\in N(x_i)}w_{ij}G^i_{jk}w_{ik}\\&\text{s. t. }\sum_j w_{ij}=1, w_{ij}\geqslant 0\end{aligned} \tag{4}$$

求出上式解后，由于 w_{ij} 和 w_{ji} 一般不相等，因此再令 $w_{ij}=(w_{ij}+w_{ji})/2$，这样可以得到一个稀疏矩阵 W，即有

$$(W)_{i,j}=\begin{cases}w_{ij}, & x_j\in N(x_i)\text{或 }x_i\in N(x_j)\\0, & \text{其他}\end{cases} \tag{5}$$

这个矩阵就可以看作是样本点之间的相似度关系矩阵。

通过这个矩阵就可以在图上最短路径算法中得到样本点间的基于路径的相似度。

3.2 相似度计算优化算法

将通过上述计算得到的相似度矩阵构造一个图，通过图上最短路径算法可以计算出样本点间基于路径的相似度。考虑到最短路径算法的时间复杂度是相当高的，如样本点个数为 N，那么在构造图上运行最短路径算法的时间复杂度为 $O(N^3)$，这对于处理大规模的数据来说代价是相当高的。所以我们采用一个策略，即先在定义的图上构造一个最小生成树(Minimum Spanning Tree，MST)，要注意的是这个最小生成树是通过选取相对相似度大的边构造而成的，算法的时间复杂度为 $O(|E|\log N)$，$|E|$ 为图中边的个数，所以时间复杂度为 $O(N^2\log N)$。获得最小生成树后，可以利用 Fibonacci 堆的 Dijkstra 算法得到基于路径的相似度，这个算法的复杂度也为 $O(N^2\log N)$，从而在一定程度上节省了算法运行的时间。

通过上述算法构造最小生成树，图 4、图 5 示出了构造结果。

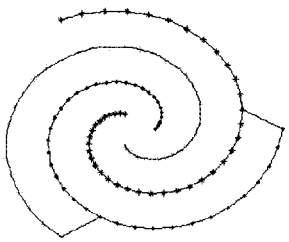

图 4 数据集 1 最小生成树

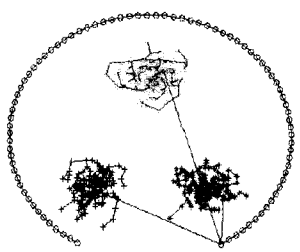

图 5 数据集 2 最小生成树

图 5 中出现了 3 条长线段，这是因为我们将不相邻的样本点的相似度全都设为 0，构造 MST 时要使得所有数据连通，因此对于相对相似度为 0 的边，算法随机选择了几条，从而导致这样的结果。

获得最小生成树后，很容易就可以得到样本间的基于路径的相似度，然后通过谱聚类算法就能得到最后聚类结果。上述两个数据集的聚类结果如图 6 所示。

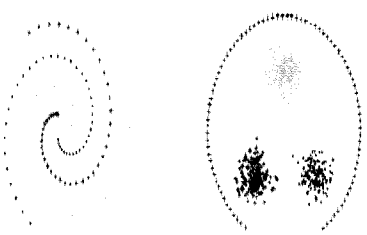

图 6 数据集 1 和数据集 2 的聚类效果

3.3 基于路径相似度的半监督谱聚类算法

当某些数据必须属于同一类或必定不属于同一类，这种情况实际上属于一个半监督学习的问题。我们可以将上述算法扩展到半监督的情况下。

假设成对同类约束集为 M，成对不同类约束集 C。我们重新定义数据点间相似度矩阵，即可得

$$(W_{i,j})=\begin{cases} w_{ij}, & x_j\in N(x_i)\text{或}\ x_i\in N(x_j)\text{且}\ x_i,x_j\notin M\cup C \\ 0, & x_j\notin N(x_i)\text{或}\ x_i\notin N(x_j)\text{且}\ x_i,x_j\notin M\cup C \\ \max(w_{ij}), & x_i,x_j\in M \\ \min(w_{ij}), & x_i,x_j\in C \end{cases} \tag{6}$$

4 实验分析

4.1 MNIST 数据实验

采用 MNIST 手写数字库来进行实验，MNIST 手写数字库包含了 0～9 10 个手写数字的图片样本，每张图片被正规化成 28×28 大小的灰度图。选取数字“5”和“6”两个数字各 50 个样本，通过基于路径相似度谱聚类可以获得它们在二维空间的分布描述，再在这个二维空间中进行 K-means 聚类，效果如图 7 所示。

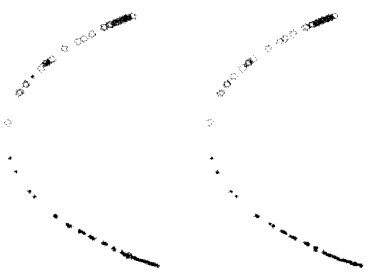

图 7 数字“5”和数字“6”的二维分布和聚类结果

再分别选择数字(1,3)、(8,5)、(1,7)、(8,9) 4 个数字对进行聚类实验，每个数字随机选取 100 个。对比算法采用 K-means、谱聚类和基于路径的谱聚类算法，对于谱聚类和基于路径的谱聚类算法，选取不同的尺度参数做 20 次实验，聚类结果为 20 次实验的最优结果，而对于本文所提的算法，在邻域大小上，同样选取 10 个不同的值，选取最佳聚类结果作为最终结果。然后整个实验再重复做 20 次，结果取平均值，结果如图 8 所示。

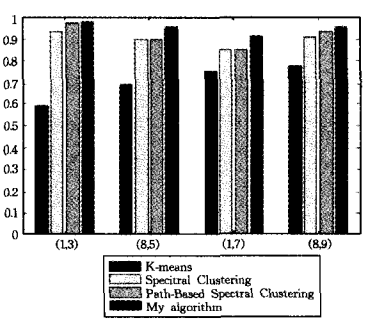

图 8 MNIST 聚类效果比较

(下转第 230 页)

[2] 吴迪，胡钢，倪刚. 无线传感器网络安全路由协议的研究[J]. 传感技术学报，2008，21(7)：1195-1201
[3] 郭书城，卢昱，许定根. 基于分簇无线传感器网络的路由算法研究[J]. 通信学报. 2010，31(8A)：63-69
[4] 孙利民，李建中，陈渝，等. 无线传感器网络[M]. 北京：清华大学出版社，2005：94-96
[5] Heinzelman W，Chandrakasan A，Balakrishnan H. Energ y-Efficient Communication Protocol for Wireless[C]// Proceedings of the Hawaii Conference on System Sciences Microsensor Networks. 2000，1：3005-3014
[6] 刘睿琼，齐小刚，孙正海. 基于节点位置的无线传感器网络分簇路由协议[J]. 电子科技，2013，26(3)：130-133
[7] 李亚男，徐夫田，陈金鑫. 基于 LEACH 的 WSNs 分簇优化策略[J]. 传感技术学报，2014，27(5)：671-674
[8] Lindsey S，Raghavendra C. PEGASIS：Power-efficient gathering in sensor information systems[J]. IEEE Aerospace Conference Proceedings，2002，3(10)：1125-1130
[9] 黄飞，金心宇，张昱，等. 基于 GASA 的能耗均衡 WSN 路由协议[J]. 传感技术学报，2009，22(4)：586-592
[10] 张震，闫连山，潘炜. 基于 LEACH 和 PEGASIS 的簇头成链可靠路由协议研究[J]. 传感技术报，2010，23(8)：1173-1178
[11] 乔钢柱. 基于无线传感器网络的煤矿安全综合监控系统设计与关键技术研究[D]. 兰州：兰州理工大学，2012：77-83
[12] 王伟. 长距离带状无线传感器网络路由协议设计[J]. 计算机工程，2014，40(3)：132-136
[13] 胡钢，谢冬梅，吴元忠. 无线传感器网络路由协议 LEACH 的研究与改进[J]. 传感技术学报，2007，20(6)：1391-1396
[14] 李鉴，石鑫，刘贺平. 煤矿巷道无线传感器网络非均匀分簇数据传送机制[J]. 地球科学-中国地质大学学报，2013，38(1)：195-200

(上接第 208 页)

4.2 图像分割实验

最后，将算法用于图像分割实验中。图像分割[12,13]就是将图像中各个物体按其自然属性区分开来。由于一般的彩色图像内容比较丰富，采用无监督的像素值聚类算法对图像进行分割的效果不是很理想，一般情况下需要人为提供一些先验信息来确定图中哪一块像素区属于这个物体而不属于另一个物体。这里利用本文提出的改进的谱聚类算法来进行图像分割。图 9(a)和图 9(c)是原始图像，图 9(b)和图 9(d)是图像分割结果。

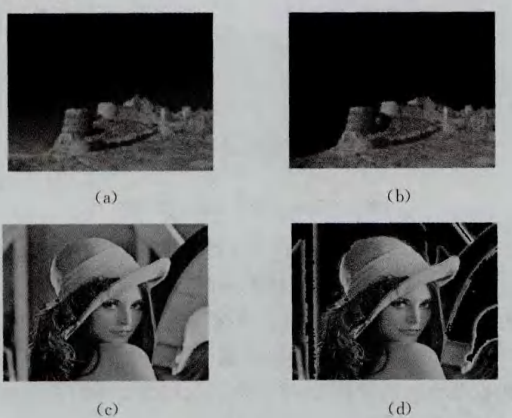

图 9 基于改进的谱聚类的图像分割效果

从图像分割实验可以看出，本算法在大样本聚类中确实是有效的。

结束语 本文提出的算法对谱聚类算法的改进主要体现在对数据点间的相似度的定义上，传统的基于路径的谱聚类算法主要是采用 Gauss 函数定义相似度，Gauss 函数中涉及了尺度参数，尺度参数选择的好坏对聚类结果的好坏影响很大。但是目前又没有确定这个尺度参数的规则，这一点给传统算法带来了一定的局限。因此我们通过二次规划的方法来鲁棒地确定数样本之间的相似度，避免了尺度参数选择的不准确性的问题，同时为了降低了计算路径相似度的时间复杂度，构造了最小生成树。最后通过 MNIST 手写数字库和图像分割实验验证了算法的有效性。

参 考 文 献

[1] 金慧珍，赵辽英，刘博. 一种基于谱聚类的灰度图像分割法[J]. 计算机系统应用，2009，4(4)：74-76
[2] 张向荣，骞晓雪，焦李成. 基于免疫谱聚类的图像分割[J]. 软件学报，2010(9)：2196-2205
[3] 贾建华，焦李成. 空间一致性约束谱聚类算法用于图像分割[J]. 红外与毫米波学报，2010(1)：69-75
[4] 王兴良，王立宏，武栓虎. 谱聚类中选取特征向量的动态选择性集成方法[J]. 人工智能与模式识别，2014，27(5)：452-462
[5] Andrew Y N，Jordan M，Weiss Y. On Spectral Clustering Analysis and an algorithm[C]// NIPS. 2002
[6] Chang H，Yeung D Y. Robust path-based spectral clustering [J]. Pattern Recognition，2008，41：191-203
[7] 潘晓英，刘芳，焦李成. 密度敏感的多智能体进化聚类算法[J]. 软件学报，2010，21(10)：2420-2431
[8] 孔万增，孙昌思，张建海，等. 近邻自适应局部尺度的谱聚类算法[J]. 中国图象图形学报，2012，17(4)：523-529
[9] Fischer B，Buhmann J M. Bagging for Path-Based Clustering [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence，2003，25(11)：1411-1415
[10] Fischer B，Buhmann J M. Path-Based Clustering for Grouping Smooth Curves and Texture Segmentation[J]. IEEE Transaction on Pattern Analysis and Machine Intelligence，2003，25(4)：513-518
[11] Wang F，Zhang C. Label Propagation through Linear Neighborhoods[C]// International Conference on Machine Learning. Pittsburgh，2006
[12] Ley J D F，van Dam A，Feiner S K，et al. Computer graphics：principles and practice[M]. New York：A ddison-Weslev Publishing，1998
[13] 章毓晋. 图像处理和分析[M]. 北京：清华大学出版社，1999