

数据的多流形结构分析

我们已经进入了一个信息爆炸的时代,海量的数据不断产生,迫切需要对这些大数据进行有效的分析,以至数据的分析和处理方法成为了诸多问题成功解决的关键,涌现出了大量的数据分析方法。几何结构分析是进行数据处理的重要基础,已经被广泛应用在人脸识别、手写体数字识别、图像分类、等模式识别和数据分类问题,以及图象分割、运动分割等计算机视觉问题(人脸识别、图像分类、运动分割等实例见下文)中。更一般地,对于高维数据的相关性分析、聚类分析等基本问题,结构分析也格外重要。

文献[1]指出一个人在不同光照下的人脸图像可以被一个低维子空间近似,由此产生大量的数据降维方法被用来挖掘数据集的低维线性子空间结构,这类方法假设数据集采样于一个线性的欧氏空间。但是,在实际问题中很多数据具备更加复杂的结构。例如,文献[2]中指出,运动分割(motion segmentation)中的特征点数据具有多个混合子空间的结构,判断哪些特征点属于同一子空间是这个问题能否有效解决的关键。

针对单一子空间结构假设的后续讨论主要是两个方面,首先是从线性到非线性的扩展,主要的代表性工作包括流形(流形是局部具有欧氏空间性质的空间,欧氏空间就是流形最简单的实例)学习等。流形学习于2000年在著名杂志Science上被首次提出,之后逐渐成为了研究热点。基于数据均匀采样于一个高维欧氏空间中的低维流形的假设,流形学习试图学习出高维数据样本空间中嵌入的低维子流形,并求出相应的嵌入映射。流形学习的出现,很好地解决了具有非线性结构的样本集的特征提取问题。然而流形学习方法通常计算复杂度较大,对噪声和算法参数都比较敏感,并且存在所谓的样本溢出问题,例如,当增加新的样本点时,不能快速地提取新特征。

其次是流形或子空间从一个到多个的扩展,即假设数据集采样于多个欧氏空间的混合。子空间聚类(又称为子空间分割,假设数据分布于若干个低维子空间的并)是将数据按某种方式分类到其所属的子空间的过程。通过子空间聚类,可以将来自同一子空间中的数据归为一类,由同类数据又可以提取对应子空间的相关性质。根据综述[2],子空间聚类的求解方法有代数方法、迭代方法、统计学方法和基于谱聚类的方法。其中基于谱聚类的方法在近几年较为流行,这类方法

首先定义一个关于样本点相互关系的图，然后利用 Normalized Cut[3]等谱聚类方法（其输入是一个反应样本关系的相似度矩阵，矩阵的第 i 行 j 列的数值越大说明第 i 个样本和第 j 个样本的关系越密切，如果能将同类样本的相似度构造的较大，不同类的较小，这类方法一般都能得到正确的分类结果）得到分割结果。代表性的基于谱聚类的子空间分割方法包括低秩表示[4]和稀疏表示[5]等，下面对这两种方法的做个简单介绍。

稀疏子空间聚类：

稀疏子空间聚类方法，是对子空间表示系数进行稀疏约束的一类子空间聚类方法。子空间聚类的最终结果是将同一子空间的数据归为一类。在子空间相互独立的情况下，属于某一子空间的数据只由这个子空间的基的线性组合生成，而在其他子空间中的表示系数为零。这样高维数据的表示系数就具有稀疏的特性。同一子空间中的数据，因为都仅在这一子空间中有非零的表示系数，表现为相同的稀疏特性，通过对表示系数稀疏约束的求解，突出了数据表示系数的这种稀疏特性，进而为数据的正确聚类提供支持。

低秩子空间聚类

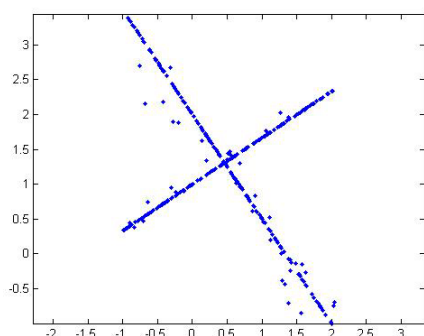
通过对子空间表示系数矩阵的研究，有些学者在求解子空间表示系数矩阵时，引入核范数（一个矩阵的核范数是指矩阵的所有奇异值的加和）约束，希望通过系数矩阵的低秩要求得到更好的数据的子空间表示。文章[4]给出了低秩表示模型的闭解且理论上保证了当子空间独立且数据采样充分的情况时，低秩表示可以得到块对角的解。这个结论基本保证了低秩表示方法在解决独立子空间分割问题的有效性。

有些实际问题的数据并不符合混合子空间结构的假设，例如图 3（a）中一个圆台的点云，圆台的顶，底和侧面分别采样于不同流形。所以假设数据的结构为混合多流形更具有一般性。由于混合流形不全是子空间的情况，数据往往具有更复杂的结构，分析这种数据具有更大的挑战性。基于谱聚类的方法仍然是处理该类问题的流行方法如文献[6]。虽然这类数据本身无法使用相互表示的方式，但是数据的特征可相互线性表示且表示系数具有稀疏性或低秩性的特点。由此一些学者通过提取数据的特征将低秩表示模型扩展用于处理图像分割[7]、图像的显著性检测[8]等问题。

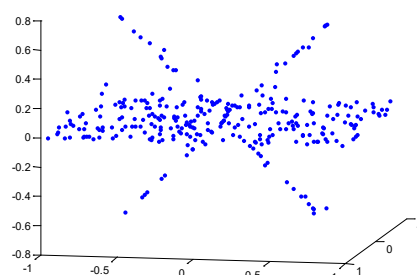
本几何结构分析问题中假设数据分布在多个维数不等的流形上，其特殊情况是数据分布在多个线性子空间上。请按照文献中的方法或以文献中的方法为基础创新新的方法完成以下问题，创新部分一定要讲清思路，要具有一般性（例如不仅适应低维数据也适应高维数据）。回答方式：第 1 题，第 3 题的 b 与 c 请制作一个表格输出样本的类别标签，每行 20 个，其余题目请将分类结果画出：

1. 当子空间独立时，子空间聚类问题相对容易。附件一中 1.mat 中有一组高维数据（.mat 所存矩阵的每列为一个数据点，以下各题均如此），它采样于两个独立的子空间。请将该组数据分成两类。

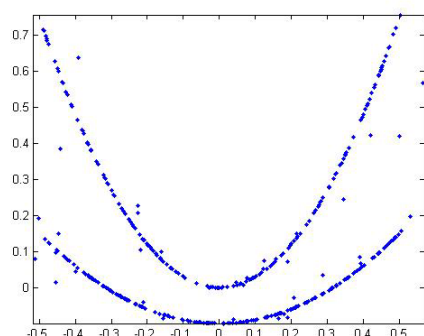
2. 请处理附件二中四个低维空间中的子空间聚类问题和多流形聚类问题，如图 1 所示。图 1(a) 为两条交点不在原点且互相垂直的两条直线，请将其分为两类；图 1(b) 为一个平面和两条直线，这是一个不满足独立子空间的关系的例子，请将其分为三类。图 1(c) 为两条不相交的二次曲线，请将其分为两类。图 1(d) 为两条相交的螺旋线，请将其分为两类。



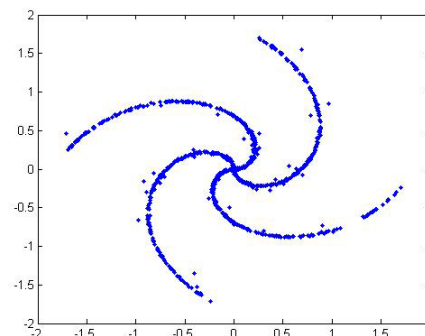
(a)



(b)



(c)



(d)

图 1

3. 请解决以下三个实际应用中的子空间聚类问题，数据见附件三

(a) 受实际条件的制约，在工业测量中往往需要非接触测量的方式，视觉重建是一类重要的非接触测量方法。特征提取是视觉重建的一个关键环节，如图 2 (a) 所示，其中十字便是特征提取环节中处理得到的，十字上的点的位置信息已经提取出来，为了确定十字的中心位置，一个可行的方法是先将十字中的点按照“横”和“竖”分两类。请使用适当的方法将图 2 (a) 中十字上的点分成两类。

(b) 运动分割是将视频中有着不同运动的物体分开，是动态场景的理解和重构中是不可缺少的一步。基于特征点轨迹的方法是重要的一类运动分割方法，该方法首先利用标准的追踪方法提取视频中不同运动物体的特征点轨迹，之后把场景中不同运动对应的不同特征点轨迹分割出来。已经有文献指出同一运动的特征点轨迹在同一个线性流形上。图 2 (b) 显示了视频中的一帧，有三个不同运动的特征点轨迹被提取出来保存在了 3b.mat 文件中，请使用适当方法将这些特征点轨迹分成三类。

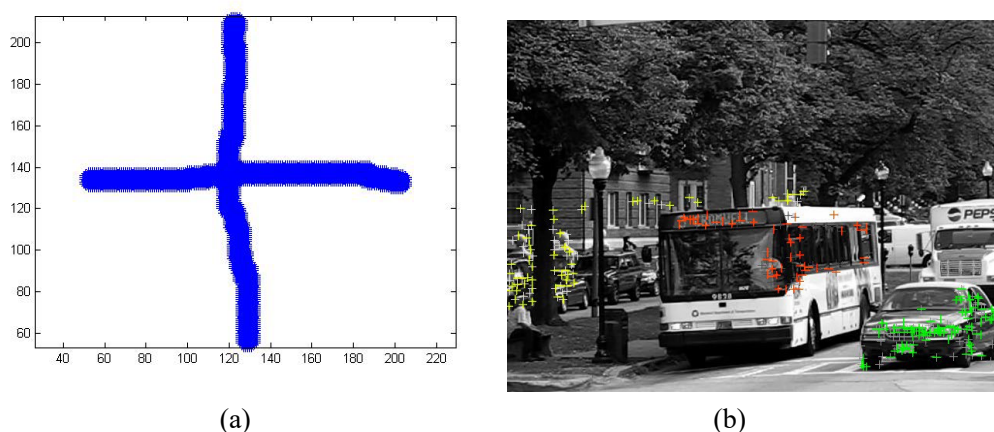


图 2

(c) 3c.mat 中的数据为两个人在不同光照下的人脸图像共 20 幅 (X 变量的每一列为拉成向量的一幅人脸图像)，请将这 20 幅图像分成两类。

4. 请作答如下两个实际应用中的多流形聚类问题

图 3(a) 分别显示了圆台的点云，请将点按照其所在的面分开 (即圆台按照圆台的顶、底、侧面分成三类)。

图 3(b) 是机器工件外部边缘轮廓的图像，请将轮廓线中不同的直线和圆弧分类，类数自定。

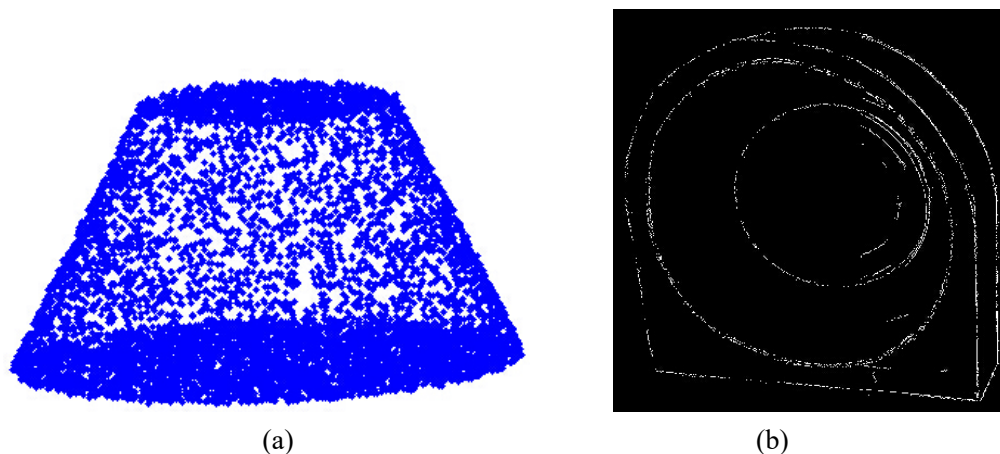


图 3

参考文献

- [1] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2):218–233, 2003.
- [2] R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.
- [3] J. Shi and J. Malik, “Normalized cuts and image segmentation,” *IEEE Transactions Pattern Analysis Machine Intelligence*, 22(8):888–905, 2000.
- [4] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.
- [5] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.
- [6] Y. Wang, Y. Jiang, Y. Wu, and Z. Zhou. Spectral clustering on multiple manifolds. *IEEE Transactions on Neural Networks*, 22(7):1149–1161, 2011.
- [7] B. Cheng, G. Liu, J. Wang, Z. Huang, and S. Yan, Multi-task low rank affinity pursuit for image segmentation, *ICCV*, 2011.
- [8] C. Lang, G. Liu, J. Yu, and S. Yan, Saliency detection by multitask sparsity pursuit, *IEEE Transactions on Image Processing*, 21(3): 1327–1338, 2012.