参赛密码 _____

（由组委会填写）

第十二届“中关村青联杯”全国研究生数学建模竞赛

学　　校		首都师范大学
参赛队号		10028002
队员姓名	1.	袁沅祥
	2.	陈 迪
	3.	鲁世嘉

参赛密码 _____

（由组委会填写）

第十二届“中关村青联杯”全国研究生数学建模竞赛

题 目 **数据的多流形结构分析**

摘 要：

我们已经进入大数据时代, 数据的分析和处理方法成为了诸多问题成功解决的关键. 对高维数据的结构进行分析, 分离出我们感兴趣的信息, 挖掘数据的潜在使用价值, 对于生产实践具有重要意义. 由文献[1-10], 本文对问题1-4进行建模与求解, 得到结果如下:

问题1

由 SSC 算法, 我们对附件一中的 200 个 100 维数据分为两类, 一类 98 个点, 从图像来看分布在两端; 二类 102 个点, 其中第 21 和第 146 个点夹杂在二类中.

问题2

(a) 我们提出一种基于训练和拟合的算法, 将本题数据分为两类, 一类 174 点, 二类 166 点. 从图像来看, 在交叉位置有 5 个点没有分好, 具体可由程序算出.

(b) 我们将 $SMMC$ 与上一小问的算法结合起来, 把数据分为三类, 一类 30 点, 二类 30 点, 三类（平面） 240 点, 从图像来看, 有 3 个点没有分好.

(c) 也利用 $SMMC$ 算法将本题数据分成两类, 一类 200 点, 二类 200 点, 分类效果较好, 从图像来看, 没有分错的点.

(d) 不同于直线和平面, 螺旋曲线较复杂, 经多次尝试聚类, 最终被分成两类, 一类 532 点, 二类 456 点, 在交叉处分类效果较差.

问题3

(a) 将问题 2a 的算法应用到该实际问题, 分为两类, 横向 1499 点, 纵向 1336 点, 十字的中心位置为 $(124.212, 135.537)$. 从图像看， 部分二类的点被分在一类, 这是我们算法自身的缺点.

(b) 本题我们结合运动分割理论,应用 $Ncut$ 算法,将本题运动特征点轨迹分为三类,一类 67 点,二类 138 点,三类 92 点,并将分割完的数据逐帧打出.

(c) 本题我们结合多流形聚类理论,应用 $SMMC$ 算法,将 20 张人脸图片进行分类,由于图像的数据矩阵维数比较高,而且因为是实际数据,里面带有噪声,导致程序对第 19 个数据出现了误判.

问题4

(a) 我们根据文献 [6] 所推荐参数区间,对参数进行了调整.最终结果分为三类,侧面 3818 点,底面 3600 点,顶面 900 点.

(b) 我们参考文献 [10] 所提方法,估计出最优聚类类数为六类,最终结果表明,对光滑的曲线分割效果比较好.

关键字: 聚类、子空间、流形、分割、谱多流形聚类

目　录

数据的多流形结构分析

I 问题重述

1.1 背景与研究意义

二十一世纪是信息爆炸的时代，海量的数据不断产生，所谓“大数据”概念日嚣尘上，数据挖掘技术特别热门，我们迫切需要对这些大数据进行有效的分析，以抓取我们感兴趣的有用信息，挖掘数据背后的使用价值。提出分析和处理数据的新方法成为了成功解决诸多问题的关键，渐渐涌现出了大量的数据分析方法，其中几何结构分析是进行数据处理的重要基础,已经被广泛应用在数据分类、人脸识别、图象分割等计算机视觉问题中。将问题更一般化，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。

1.2 国内外研究现状

假设数据集采样于一个线性的欧氏空间，数据降维方法是一种用来挖掘数据集的低维线性子空间结构的方法。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。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] 指出，运动分割中的特征点数据具有多个混合子空间的结构，我们会常常面临要做的事情就是判断哪些特征点是属于同一子空间，将数据集分裂为若干个类别。

在单一子空间结构假设的条件下，学者们的后续讨论主要是如下两个方面。

一方面，从线性到非线性进行扩展，主要的代表性工作包括流形学习等。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；然而不足之处是，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。

另一方面，流形或子空间从一个到多个进行扩展，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]可知，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。

稀疏子空间聚类，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。

1.3 本文主要研究内容

本文中的几何结构分析问题，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，更特殊情况是分布在独立的子空间上，如问题1 和2a。我们列出本文研究的四个问题如下。

1、处理200个来自两个独立子空间里的100维的数据，利用子空间聚类算法将其分为两类，需要为结果添加类别标签。

2、利用子空间聚类和多流形聚类处理四个低维子空间的聚类问题，并用图像表示出来。(a) 两条交点不在原点且相互垂直的两条直线，需将其数据分为两类；(b) 一个平面与两条直线，它们不满足独立子空间的性质，需将其数据分为三类；(c) 两条不相交的二次曲线，需将其数据分为两类；(d) 两个相交的螺旋线，需将其数据分为两类。

3、利用子空间聚类方法解决实际问题：(a) 将从工业测量中特征提取出的十字上的点数据，用聚类的方法分成横，竖两类，来确定十字中心的问题；(b) 将动态场景比如视频中的三个不同运动的特征点轨迹数据用聚类的方法分成三类，来解决运动分割的问题；(c) 将两个人在不同光照下的人脸图像数据，用聚类的方法分成两类。(b) 和 (c) 的结果输出在$EXCEL$中。

4、利用多流形聚类方法解决实际问题：(a) 将圆台上的点数据，按照其所在的平面，即圆台的顶、底、侧面分成三类，并用图像表示出来；(b) 将机器工件的外部边缘轮廓线，按照线的类型，自定义类数，进行分类。

II 假设与符号

2.1 问题假设

(a) 数据在全局上位于或近似位于光滑的非线性流形上，局部地，每个数据点和它的近邻点位于流形的一个局部线性块上。

(b)每个数据点的局部切空间提供了非线性流形局部几何结构的优良低维线性近似。

(c)在不同流形聚类的相交区域，来自于同一个流形聚类的数据点有相似的局部切空间，而来自不同流形聚类的数据点其切空间是不同的。

(d) 数据集由内在的多个流形生成。

(e) 如果数据点i和数据点j在同一类里，那么$B_{ij}=1$。如果数据点i和数据点j不在同一类里，则$B_{ij}=0$。一般用$B_{ij}=\exp(-dist^2{}_{ij})$衡量数据点之间的相似度，其中的$dist_{ij}$表示数据点之间的距离。

(f)题目所给的数据集没有异常点，问题1和问题2是非实际问题，它们是数值模拟的数据，没有噪声。

2.2 符号说明

本文使用到的主要符号描述如表格2-1所示。

表 2-1: 本文用到的符号说明

符号与缩略语	
S_i	子空间
$\{S_i\}_{i=1}^n$	仿射子空间之并的数据集
μ_i	子空间S_i中的某一点
d_i	维数
$dist$	距离
R^D	维数为D的实数空间
Γ	置换矩阵
C	稀疏稀疏矩阵
F	稀疏奇异矩阵
Z	噪声矩阵
$diag(C)$	矩阵C的对角元素组成的向量
u	集合A中的任一点
v	集合B中的任一点
V	顶点的集合
$assoc(A,V)$	从A中的节点到图形中所有节点的总连接
$assoc(A,A)$	A内部边缘连接点的全部权重
$e(i,j)$	连接顶点i和j的边

接下页表格

(续表)

E	所有边$e(i,j)$的结合
W	相似性矩阵
$\omega(i,j)$	边$e(i,j)$的权重
$Q(i)$	顶点i的空间坐标
f	融合函数
Θ_i	局部切空间
θ	两个子空间之间的主角度
$Knn(x)$	x的k个邻近数据点
μ_m	数据的均值向量
C_m	模型的协方差
I	单位矩阵
x_i	第i个原始数据
A_i	第i个矩阵型数据
PCA	主成分分析
SCC	稀疏子空间聚类算法
$SMMC$	谱多流形聚类算法

III 问题分析

3.1 问题1分析

首先利用MATLAB（当前使用版本R2013a）加载1.mat文件，发现它是200个100维的数据点，如图3-1所示，其中一列为一个数据；然后调用imshow函数进行显示，容易看到图像具有“流形”，像缓缓行进的水流一样，仔细观察可以看到它有两层（尺寸较大的一层我们称之为底层，背景层），我们猜想这正是题目要求我们分两类的原因。但是，有八个“竖道”很明显也表现出来了，底层有两个，上层有六个，在作出没有异常点的假设下，可以猜测这种情况与数据的排序有关。

图3-1的灰度值整体偏暗，说明数据集中只有在少数维度下的值比较大，其余值相对很小。

调用pdist 函数计算距离矩阵如图3-2 所示，这是一个对称矩阵，一条对角线呈现黑色，这是因为数据点到自己的距离为0，也能明显注意到有水平方向和竖直方向各有八条亮度较大的“白条”，这表明“竖道”到其余任意点的距离都比较大，可能是“中心”。从整体上看图3-2，呈现较黑（深色）与较灰（浅色）两种颜色，按这两种颜色可对数据进行粗略分类。

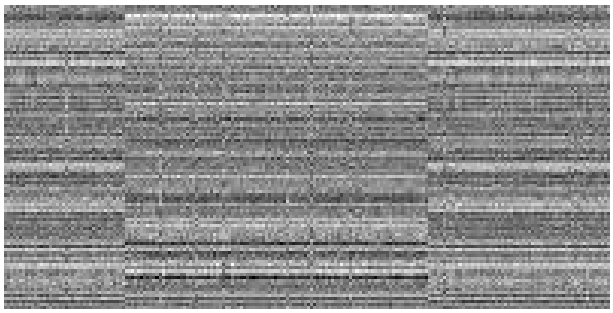

图 3-1: 原始数据

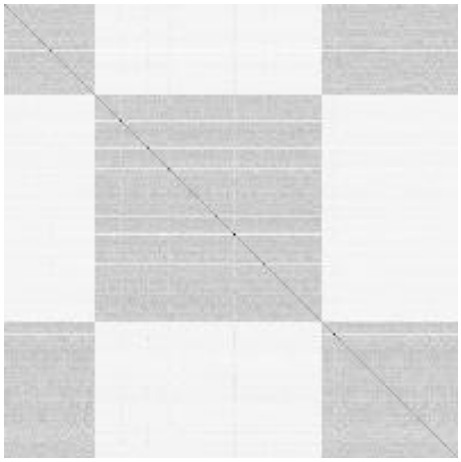

图 3-2: 距离矩阵

本题中的数据集维数高达100，传统聚类算法不能有效将其分类。许多文献也指出，传统聚类算法并不适用于高维分类数据集，因为传统的距离测量方式不再适用于高维环境下的分类数据，在高维数据空间上，对象之间的距离亦变得几乎等同。[9]在高维数据集中，有很多聚类结果在全维属性空间中并不相似，但在某些属性子集空间中，却具有很高的相似性．包括前述算法的全维聚类算法都不能发现这种隐含在属性子集上的聚类，因此，研究者提出并开展了子空间聚类算法的相关研究。

3.2 问题2分析

分析本问题中的四个数据文件可知，a,c,d是二维点列，题目要求各分为两类，b 是三维点列，题目要求分为三类。我们首先采用传统的聚类方法，得到如图3-3 所示的结果，显然聚类算法已经失效。我们再根据题目所给文献的算法，试着重新对数据集进行分类。题目明确说这是子空间聚类和多流形问题。图1(a) 为两条交点不在原点且互相垂直的两条直线，而直线是一维的，两条相交直线没有除交点之外的共同元素，这是一个满足独立子空间关系的分类问题；图1(b) 为一个平面和两条直线，平面是二维的，直线是一维的，这是一个不满足独立子空间的关系的例子。图1(c)为两条不相交的二次曲线，二次曲线上的点没有像直线或平面具有良好的线性性质，图1(d)为两条相交的螺旋线，情况更为复杂。

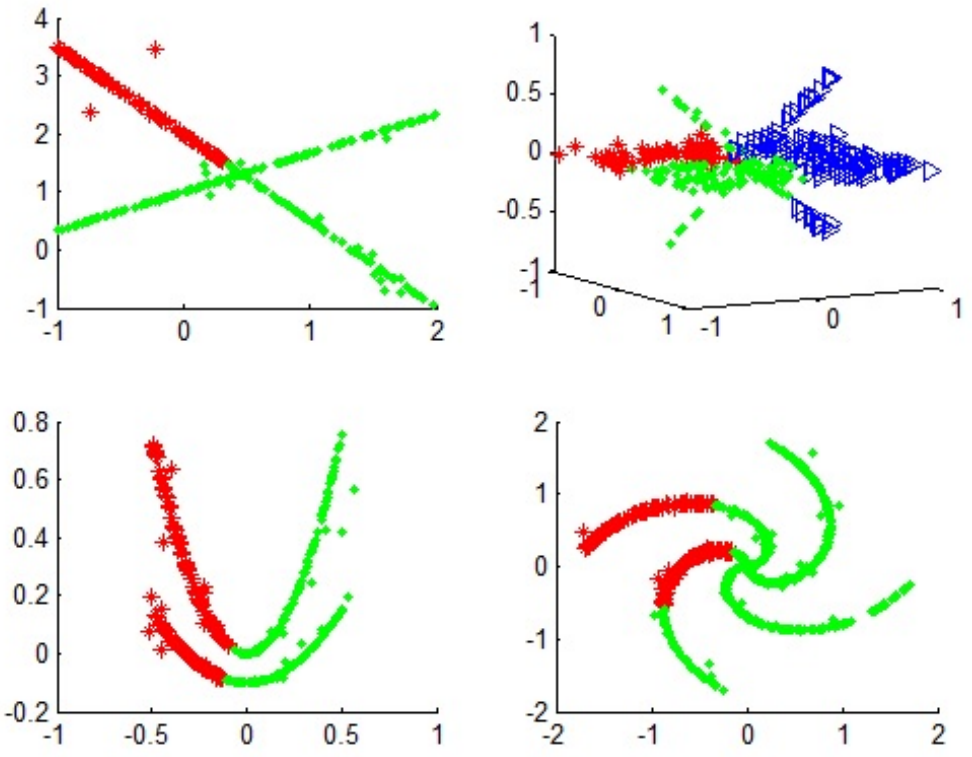

图 3-3: 传统方法聚类结果

3.3 问题3分析

本题a为视觉重建中的特征提取问题，待解决的实际问题是将一个十字形“十字架”数据集按照“横”和“竖”两个方向分为两类。本题a与问题2中的a特别类似。因为是实际问题，难免会引入误差，从图也可看出这个十字架中间是有点弯的，而且“横”和“竖”两个方向上的点的数量也比2题a中的数据要多出数倍。所以，虽然这题也用和2 题a中的方法，但是在算法上应作改进，比如要对数据矩阵进行特别的处理，来达到聚类的效果。

本题b是运动分割中的特征点轨迹分类问题，待解决的实际问题是将被采集到的数据矩阵中的三个运动特征点轨迹分成三类。这里我们考虑用文献[7]或者文献[3]中的方法进行聚类。

本题c是对两个人在不同光照条件下的20幅面部图像进行分类，分成两类。数据矩阵中的每一列是由人脸图像矩阵拉直而成。初步观察，数据虽然量不大，但是维数很高。直接用数据矩阵进行聚类，效果不会很好。所以我们考虑将这个图像处理问题转化为高维子空间聚类问题。

3.4 问题4分析

本题是实际应用中的多流形聚类问题。

本题a要将圆台上的点云按照它所在的面，即顶面、底面、侧面来分成三类。这道题感觉与第2 题的b类似，都是三维空间的并且不是独立的，所以打算用类似的算法进行聚

类。不过这题的数据量比较大，需要进行一下处理。

本题b要将机器工件外部边缘轮廓进行分类，但这道题并没有明确给出类数，我们考虑先用文献[10]中提到的方法来估计类数，再进行聚类分析。

IV 建立模型与求解

4.1 模型一

子空间聚类

设$\{x_j \in \mathbb{R}^D\}_{j=1}^N$是取自$n\,(n \geq 1)$个线性或仿射子空间$\{S_i\}_{i=1}^n$之并的数据集，子空间的维数$d_i = \dim(S_i), 0 < d_i < D, i = 1, \cdots n$.子空间能表示为

$$S_i = \{x \in \mathbb{R}^D : x = \mu_i + U_i y\}, i = 1, \cdots, n, \tag{4-1}$$

式4-1中$\mu_i \in \mathbb{R}^D$是子空间S_i中的某一点，对于线性子空间，取$\mu_i = 0$，$U_i \in \mathbb{R}^{D \times d_i}$是$S_i$的一组基，$y_i \in \mathbb{R}^D$是$x$的低维稀疏表示。子空间聚类的目的是寻找$n$个子空间，它们的维数为$\{d_i\}_{i=1}^n$，基底为$\{U_i\}_{i=1}^n$，以及其中某个点$\{\mu_i\}_{i=1}^n$，使得分割出来的数据点分别属于自己的子空间。

当子空间的个数为1时，这个问题等价于寻找一个向量$\mu_i \in \mathbb{R}^D$，一组基底，线性表示$Y = [y1, \cdots, y_N] \in \mathbb{R}^{d \times N}$，以及维数$d$。这个问题就是大家熟知的$PCA$。文献[2]指出，当$n > 1$时子空间聚类变得更难。首先，是数据分割和模型估计之间的耦合强。具体而言，如果数据的分割是已知的，可以很容易地适应一个单一的子空间，每个组的点使用标准的主成分分析。显然，如果子空间参数已知，我们很容易发现属于每一子空间的数据点是哪些。实际上，数据的分割和子空间参量都是不知道的，我们需要同时解决两个问题。

数据在子空间内的分布一般是未知的。如果每个子空间内的数据分布在一个集群中心，不同子空间的聚类中心是相距甚远的，子空间聚类问题简化到简单并已经充分研究的中心聚类问题上。然而，如果在子空间中的数据点的分布是任意的，子空间的聚类问题并不能用中心聚类的方法解决。此外，当许多点靠近交叉点或存在较多的子空间时，问题就变得更加困难了。

子空间的位置和方向可以是任意的。当子空间是不相交的或独立的子空间聚类问题，可以更容易地解决，这一点我们在后面的文章中进一步说明。然而，当子空间的相关性较强时，子空间聚类问题变得更加困难。(如果两个子空间的交点是原点，那么线性子空间是不相交的。如果子空间合并的维数与各自子空间维数的总和是相等的，则说明维线性子空间是独立的。独立的子空间是无交集的但反过来并非总是如此。如果维仿射子空间是独立不相交的，那么在齐次坐标系中的相对应的线性子空间。)

数据可能会损坏由噪声、缺少条目和离群值。虽然在一个单一的子空间的情况下对其稳健性有一定的研究，但是在多个子空间的情况仍不是很清楚的。

模型的选择。在经典的主成分分析中，唯一的参数就是子空间维度，其中可以通过寻找最小尺寸的子空间，找到适合的数据和一个给定的精度。在多维子空间的情况下，我

们可以把数据分成那个不同的一维子空间，例如：将一个数据点视为一个子空间，或者分成一个D维的子空间。很显然，这两种方法并不好。所以我们面临的挑战是如何找到一个倾向小维度和子空间数量较少的模型的评价标准。

高维数据聚类里比较流行的是谱聚类算法。谱聚类算法的关键是构造相似度矩阵$B \in R^{N\times N}$，用来衡量数据点i和数据点j之间相似度。在理想情况下，如果数据点i和数据点j在同一类里，那么$B_{ij}=1$。如果数据点i 和数据点j不在同一类里，则$B_{ij}=0$。一般用$B_{ij}=\exp(-dist^2{}_{ij})$衡量数据点之间的相似度，其中的$dist_{ij}$表示数据点之间的距离。由相似度矩阵计算出拉普拉斯矩阵，再计算拉普拉斯矩阵$n(n \ll N)$个特征向量，最后使用$k-means$对由特征向量构成的矩阵进行聚类完成数据的划分。

稀疏最优化模型I

基于谱聚类的子空间聚类关键是如何构建相似度矩阵。当两个数据点非常接近，可能两个数据点并不是在同一个子空间里。相反，两个数据点距离非常远，却很有可能在同一个子空间里。在这里不能用传统的方法即数据点之间的距离衡量。而近年来，随着稀疏学习的流行，相似矩阵构造有基于稀疏表示和基于低秩描述的。

位于线性或仿射子空间集合的高维数据可以稀疏地被同一个子空间的点线性或者仿射表示。本文通过文献[5] 中稀疏表示技巧获得高维数据的稀疏表示。

设有N个D维数据$\{y_i\}_{i=1}^N$，处于R^D空间的n个线性子空间$\{S_i\}_{i=1}^n$中，子空间的维度分别为$\{d_i\}_{i=1}^n$，定义一个矩阵Y 为：

$$Y=[y_1\cdots y_N]=[Y_1\cdots Y_n]\Gamma \tag{4-2}$$

其中，$Y\in R^{M\times N}, Y_1\in R^{M\times N_l}$是一个秩为$dl$的矩阵，表示第$l$个子空间数据组成的矩阵。$\Gamma$为未知的置换矩阵。子空间聚类目的就是获得$Y_l\in R^{M\times N_l}$ 矩阵。

对于每个数据点都可以被一些除它以外的数据点表示。为了获得每个数据点的最稀疏的表示，选择最小化其l_0范数对其进行凸松弛处理。稀疏最优化模型为：

$$\min \|C\|_1, s.t. Y=YC, diag(C)=0 \tag{4-3}$$

其中，$C=[c_1,c_2,\cdots,c_N]\in R^{N\times N}$是一个矩阵，每一列对应每个数据点的稀疏表示$diag(C)\in R^N$ 是矩阵C的对角元素组成的向量。

因此，稀疏代表的获得转化为一个凸优化的问题。注意，该稀疏最优化模型只对独立的子空间和不相交的子空间有效。另外，相对于一些分类算法，本文算法不需要提前获知子空间的个数和维数。

我们利用SSC算法对问题一的数据集进行聚类，得到结果如图4-4所示，样本的类别标签如表格4-5所示，横坐标为点索引，纵坐标为归类id，很清晰地看到有两个2类中的点，穿插在1类的索引中，这两个点的索引是21 和146，在$EXCEL$中的坐标为$(A,2)$，$(F,8)$。

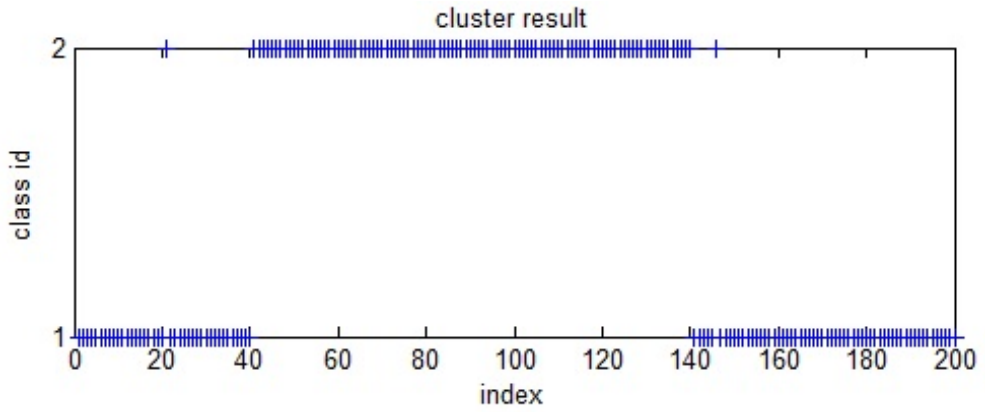

图 4-4: 传统方法聚类结果

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
3	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
4	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
5	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
6	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
7	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
8	1	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

图 4-5: problem 1 类别标签

题目中给出的点采样于两个独立的子空间，利用稀疏子空间聚类算法(SSC) 进行求解，从运行结果中我们可以看出，98 个点属于1类，有102 个点属于2 类。

4.2 模型二

4.2.1 模型$2-a$

随机训练拟合算法

针对数据集，$X_i, i=1,\cdots,N$是来源于二维平面内两条直线上的情况，比如问题2(a)，我们提出随机训练拟合算法（Random training and fitting algorithm）。我们已经知道，任何一条直线可以表示为$l: kx+b=(k,b)\begin{pmatrix} x \\ 1 \end{pmatrix}=M\begin{pmatrix} x \\ 1 \end{pmatrix}=y$。$RTFA$算法的步骤如下:

步骤1. 迭代计数器置零，给定最大迭代次数$\max_it$；给定非负常量C，通常取$C\leq 1$；给定计数器上限U。

步骤2.如果$iterator \geq \max_it$（$\max_it$是最大迭代次数），跳出循环，终止算法；否则继续，并将计数器归零，即$Counter=0$。任取两个不同的点$X_1=(x_1,y_1)$，$X_2=(x_2,y_2)$，假设$X_1\in l, X_2\in l$，可解得系数k,b（如果不能解出系数，则重复取点操作），并将X_1,X_2归为A类。

步骤3.遍历数据集，令$y = M\begin{pmatrix} x_i \\ 1 \end{pmatrix}$，$\Delta = \left|\frac{y_i - y}{y}\right|$。进行判断，若$\Delta \leq C$，则将$X_i$归为$A$类，并采用最小二乘算法校正系数$k, b$；否则，进行第4步。

步骤4.若B类个数$N_B = n$（n是一个不小于2的整数），同上采用最小二乘算法校正另外一条直线的系数$(k', b') = M'$。否则，令$y' = M'\begin{pmatrix} x_i \\ 1 \end{pmatrix}$，$\Delta' = \left|\frac{y_i - y'}{y'}\right|$。进行判断，若$\Delta' \leq C$，则将$X_i$归为$B$类，否则计算器$Counter + 1$。当$Counter > U$时，转步骤2，否则转步骤3 继续遍历数据，直到完成。

步骤5.如果$Counter = 0$，算法结束，循环退出；否则转步骤2。

本算法的详细实现过程见$MATLAB$代码。

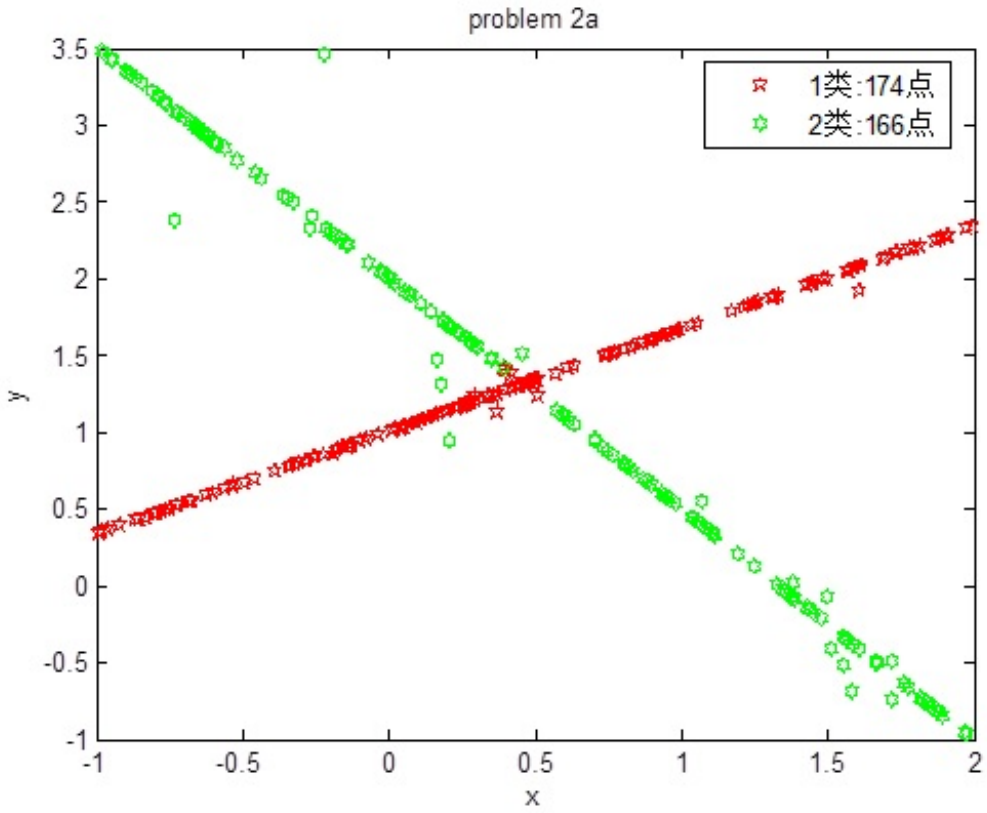

图 4-6: problem 2a

本题指明数据点可以按照两条互相垂直的直线进行聚类，对于数据集中任一点，不属于此类，定属于彼类，因为前文已经假设没有异常点存在，其实上述算法中的$Counter$就是用来统计异常点个数的，当其超过上限U时可能是初值不好，程序开始重新选取初始点。

我们利用随机训练和拟合算法最终将本题数据分成两类，如图4-6所示。其中，五角星表示1 类，共174 点，六角星表示2 类，共166 点。从图像看出，绝大部分点分类正确，某些点即使偏离2 类很远，也能正确被归类，然而同样距离的某些靠近交叉位置的2类点却归类失败了，可见算法还有待改进。

4.2.2 模型$2-b$

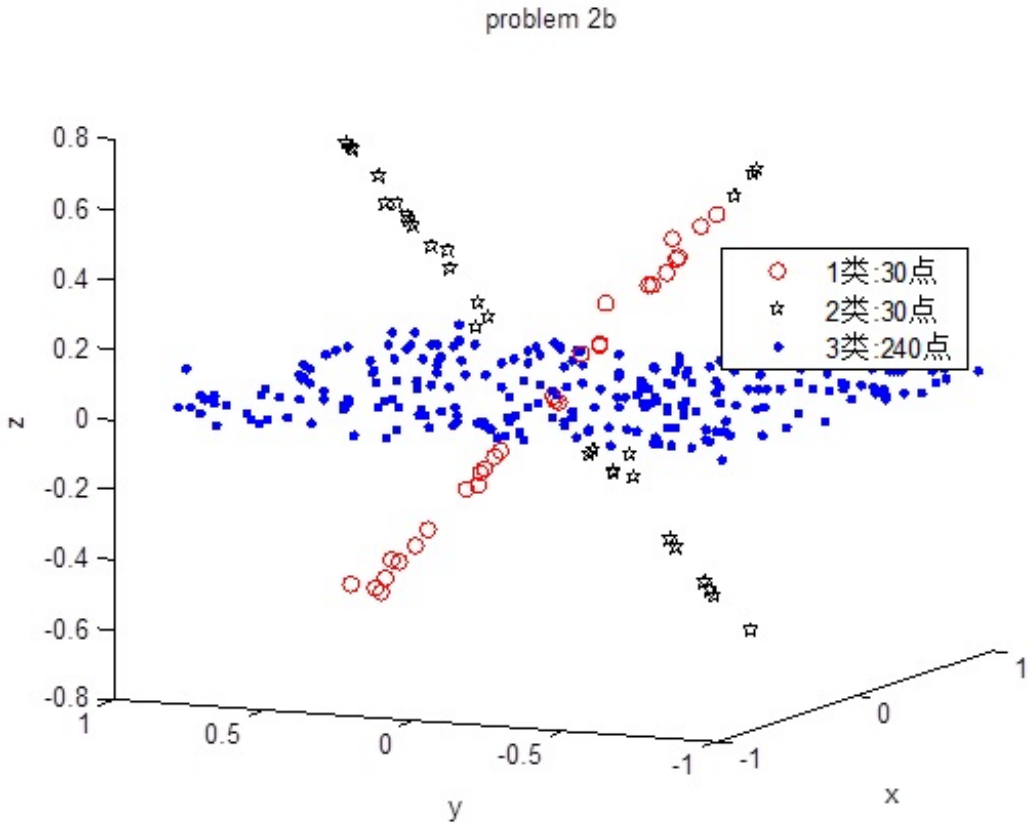

图 4-7: problem 2b

本题与上一问的不同之处在于，数据维数变成3，并由题目可知这是一个不满足独立子空间关系的分类问题。我们结合$SMMC$与随机训练和拟合算法将其分为三类，如图4-7所示，红圈表示的1 类有30 点，黑色五角星表示的2 类也有30 点，它们代表两条直线，蓝点表示的3 类有240点，代表的是平面。从整体上看分类效果比较好，但是有个别1类中的点被分到了2 类，如右上角的几个点。

4.2.3 模型$2-c$

流形聚类

流形聚类的目的是把输入流形数据集分为若干个类别，使得每个类别中的数据点都来自单一、简单、低维嵌入流形。首先假设低维流形的数目和维数是已知的。

如果假定数据集由多个内在流形生成，那么我们可以給出流形聚类的形式化描述：

定义：设$X = \{x_1, x_2, \cdots, x_n\}$是$m$维空间的高维数据集合，流形聚类就是将数据集$X$划分为$K$个互不相交的子集$X = X_1 \cup X_2 \cup \cdots \cup X_K$，且$X_i \cap X_j = \emptyset (1 \leq i, j \leq K, i \neq j)$，并指定不同的类别标签。

令$Y = Y_1 \cup Y_2 \cup \cdots \cup Y_K$，其中低维流形子集$Y_i$是高维数据子集$X_i$对应的低维嵌入，聚类后每个低维子集$Y_i$ 都是简单、单一的低维流形结构。流形聚类是一类特殊的聚类问

题，主要考虑数据间的流形结构，按照数据潜在的流形结构进行聚类，使得同一种流形结构的数据在同类别中。

谱多流形聚类

谱多流形聚类方法($Spectral\ Multi-Manifold\ Clustering$简记为$SMMC$) 来实现混合流形聚类。它的基本思想是：从相似性矩阵的角度出发，充分利用流形采样点所内含的自然的局部几何结构信息来辅助构造更合适的相似性矩阵并进而发现正确的流形聚类。

(1)相似性矩阵

在构造相似性矩阵时，既要考虑数据点之间的欧式距离关系$q_{ij}=q(\|x_i-x_j\|)$，又要考虑数据点局部切空间的相似性p_{ij}，这两个相似性融合在一起来决定最后的相似性权值：

$$w_{ij}=f(p_{ij},q_{ij}), \tag{4-4}$$

其中f是一个合适的融合函数。

为了使得构造出的相似性矩阵具有前面分析中所期望的性质，f应该是关于数据点间欧式距离的一个单调递减函数，同时使局部切空间之间相似性的单调递增函数。

下面我们给出$SMMC$方法中所采用的函数p,q,和f的具体形式。

假设数据点$x_i(i=1,2,\cdots,N)$处的局部切空间为Θ_i，则两个数据点x_i和x_j的局部切空间之间的结构相似性可以定义为：

$$p_{ij}=p(\Theta_i,\Theta_j)=(\prod_{l=1}^{d}\cos(\theta_l))^o \tag{4-5}$$

其中，$o\in N^+$是一个可调节参数。$0\le\theta_1\le\theta_2\le\cdots\le\theta_d\le\pi/2$是两个切空间$\Theta_i$和$\Theta_j$之间的主角度，递归地定义为：

$$\cos(\theta_1)=\max_{\substack{u_l\in\Theta_i,v_l\in\Theta_j\\ \|u_l\|=\|v_l\|=1}}u_l^Tv_l,l=2,\cdots,d, \tag{4-6}$$

其中，

$$u_l^Tu_i=0,v_l^Tv_i=0,i=1,\cdots,l-1\cdot \tag{4-7}$$

数据点x_i和x_j之间的局部相似性简单地定义为：

$$q_{ij}=\begin{cases}1 & if\ x_i\in Knn(x_j)\ or\ x_j\in Knn(x_i),\\ 0 & otherwise,\end{cases} \tag{4-8}$$

其中$Knn(x)$代表x的K个近邻数据点。话句话说，局部相似性要求我们在构造近邻图时采用$K-$近邻图，而不能采用完全将所有的数据点都通过边连接起来。最后函数f将这两个函数p和q简单的乘在一起得到相似性权值：

$$\omega_{ij}=p_{ij}q_{ij}=\begin{cases}(\prod_{l=1}^{d}\cos(\theta_l))^{o} & if\ x_i\in Knn(x_j)\ or\ x_j\in Knn(x_i),\\ 0 & otherwise.\end{cases} \tag{4-9}$$

因此，当调节参数o足够大时，也可以使得相似性权值相对低。因此，当谱方法应用于上述定义的相似性矩阵W时，可以得到很好的性质。

(2)局部切空间

我们通过训练M个混合概率主成分分析器来估计局部切空间，其中每个分析器由模型参数$\theta_m=\{\mu_m,V_m,\sigma_m^2\}(m=1,\cdots,M)$刻画，其中$\mu_m\in R^D,V_m\in R^{D\times d}$，而$\sigma_m^2$是一个标量。在$m$个分类器模型下，一个$D$维的观测数据向量$x$通过下式对应一个相应的$d$维潜在向量$y$：

$$x=V_m y+\mu_m+\varepsilon_m \tag{4-10}$$

其中μ_m是数据的均值向量，潜在变量y和噪声ε_m分别是高斯分布y 服从$N(0,I)$和ε_m服从$N(0,\sigma_m^2 I)$。模型的协方差为：$C_m=\sigma_m^2 I+V_m V_m^T$。模型参数$\mu_m,V_m,\sigma_m^2$可以通过$EM$算法得到[10]。最后采用$K-means$来初始化$EM$过程，即可得到分类器。

谱多流形聚类算法

算法$SpectralMulti-Manifold\ Clustering(SMMC)$

输入：原始数据集Ψ，聚类数k，流形维数d，局部化模型数M，近邻点数K，调节参数o.

算法过程：

1 :利用$MPPCA$训练M个d维的局部线性模型来近似潜在的流形数据；

2 :确定每个点得局部切空间；

3 :计算两个局部切空间之间的结构相似性；

4 :计算相似性矩阵$W\in R^{N\times N}$,并计算对角矩阵D,其中$d_{ii}=\sum_j w_{ij}$；

5 :计算广义特征矩阵$(D-W)u=\lambda Du$最小k个特征值对应的特征向量$u_1,\cdots,u_k$；

6 :利用$K-means$将$U=[u_1,u_2,\cdots.u_k]\in R^{N\times k}$的行向量分组为$k$个聚类。

输出：原始数据对应的聚类结果。

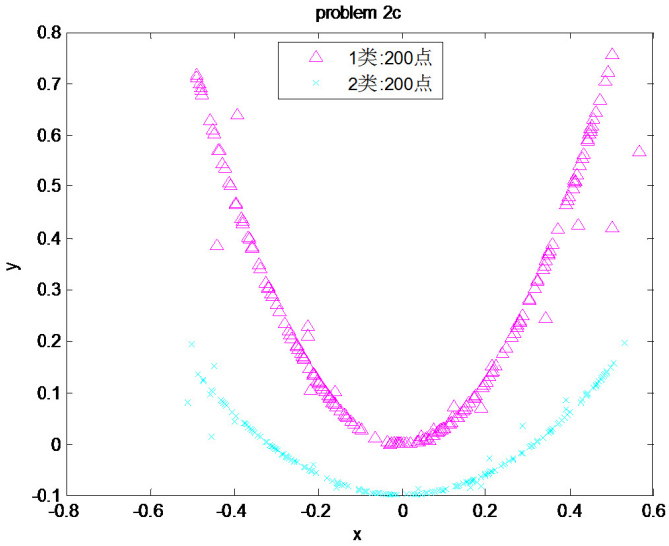

图 4-8: problem 2c

我们利用$SMMC$算法对数据进行处理，得到如图4-8所示。我们成功的将两条抛物线分成两类，第1类有200个点，第2 类有200个点。这里我们使用的参数是$(2, 1, 2, 12, 8)$，它们分别代表（类数，主成分子空间的维数，混合模型的中心数，临近点个数，亲和度）。此算法在处理没有交叉的数据时，分类结果比较好。

4.2.4 模型$2-d$

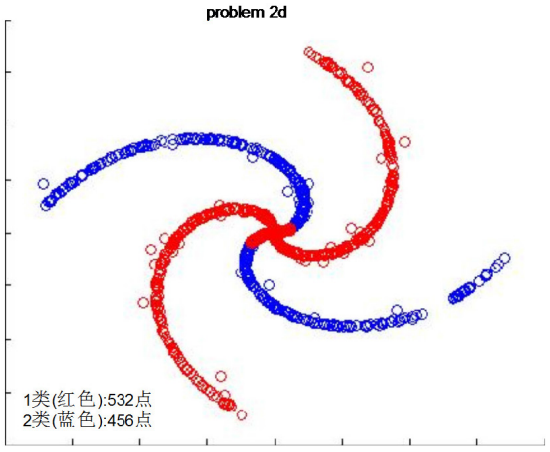

图 4-9: problem 2d

我们利用$SMMC$算法对数据进行处理，并用一个三重循环对参数按照文献中提供的范围进行了参数调整，从而确定参数是$(2, 1, 2, 10, 30)$，它们分别代表（类数，主成分子空间的维数，混合模型的中心数，临近点个数，亲和度）。但这个算法不是最优的，因为这个算法处理有交叉的数据不太稳定。我们将两条螺旋线大致分为两类，其中一类包含532个

点，另一类包含456个点。从图4-9 中我们可以看出，分类的结果稍显粗糙，在交叉处，2类（蓝色）中的一些点被分在了1 类（红色）中。

4.3 模型三

4.3.1 模型$3-a$

稀疏最优化模型II

与模型一中的仿真问题不同，在实际问题中，数据点通常混合着稀疏的奇异值和噪声。另外，数据常常处于仿射子空间的并而不是线性子空间。为解决这些问题，稀疏最优化模型转化为：

$$\min \|C\|_1 + \lambda e\|F\|_1 + 1/2\lambda z \|Z\|_F^2 \tag{4-11}$$

约束条件为：$Y = YC + F + Z$，$1^T C = 1^T$，$diag(C) = 0$。其中，C 为稀疏系数矩阵，F为稀疏奇异值矩阵，Z为噪声矩阵，系数$\lambda e = \alpha_e/\mu_e > 0$，$\lambda z = \alpha_z/\mu_z > 0$，$\mu_e \triangleq \min\limits_i \max\limits_{j\neq i} \left|{y_i}^T y_i\right|$，$\mu z \triangleq \min\limits_i \max\limits_{j\neq i} \|y_j\|_1$。

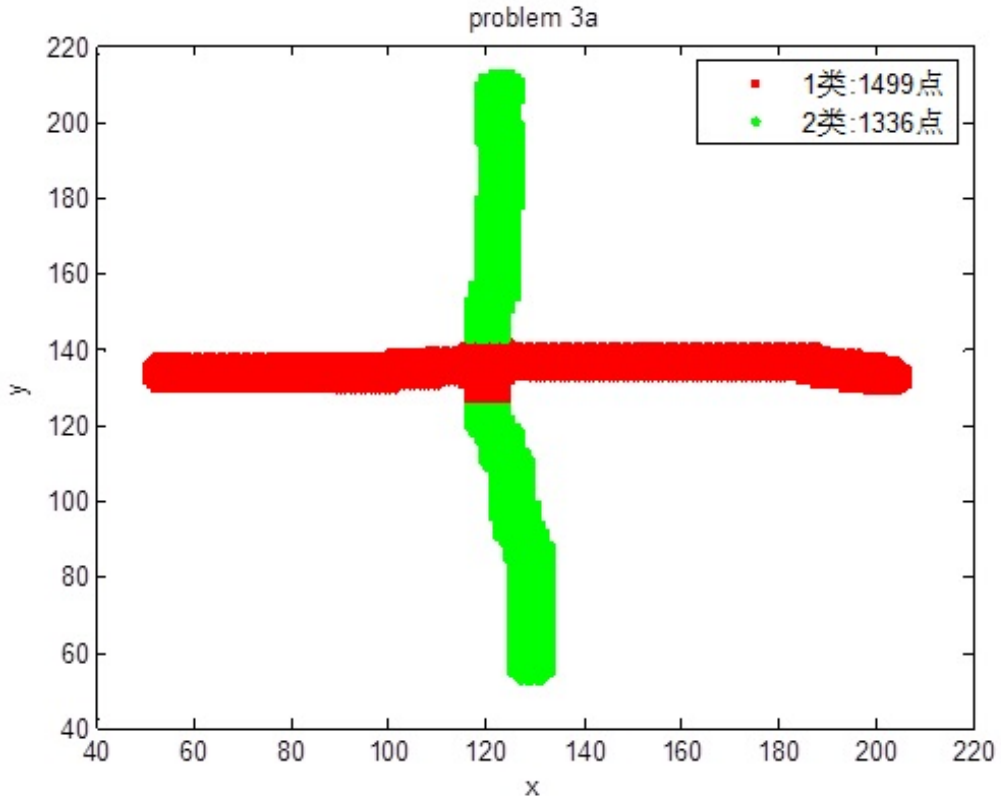

图 4-10: problem 3a

将第2 问a 中方法用到实际当中，正确地将图2(a) 分成两类，1 类（横向）1499 点，2 类1336 点，十字的中心位置大约为$(124.212, 135.537)$。我们发现，即使原本应该在直线上的数据变得“弯”了一些，算法也能得出较好的结果。从图看出，竖线比横线似乎更弯曲。

4.3.2 模型$3-b$

运动分割

运动分割的方法与图的理论分组制定最有关。任意特征空间的点集可以表示为加权无向图$G=(V,E)$，特征空间的点是图的节点，每对双节点之间形成一个边缘。每个节点的权重，$w(i,j)$是i 和j两个节点之间的功能相似性。在分组中，我们将定点集分割成不相交的点集$V_1,V_2,\cdots,V_m$，在某种程度上，在V_i点集中的顶点相似程度较高，不同点集V_iV_j的顶点的相似程度低。

图形$G=(V,E)$可以分割成两个不相交的集合，$A,B,A\cup B=V,A\cup B=\emptyset$，只需连接两个组合的边缘。这两部分的相似度可以通过已经被移除边缘的权重计算，在图像理论中，称为切割：

$$cut(A,B)=\sum_{u\in A,v\in B}w(u,v) \tag{4-12}$$

对一个图形的最优分割是将切割值降到最低。

为了避免切割成小集合中出现异常，我们利用解除关联规范化切割($Ncut$)：

$$Ncut(A,B)=\frac{cut(A,B)}{assoc(A,V)}+\frac{cut(A,B)}{assoc(B,V)} \tag{4-13}$$

其中，$assoc(A,V)=\sum_{u\in A,t\in V}w(u,t)$是从$A$中的节点到图形中所有节点的总连接，同理也可以得到$assoc(B,V)$的定义。对于给定的一个分割，我们定义一个组内规范化关联的测量值：

$$Nassoc(A,B)=\frac{assoc(A,A)}{assoc(A,V)}+\frac{assoc(B,B)}{assoc(B,V)} \tag{4-14}$$

其中，$assoc(A,A)$和$assoc(B,B)$分别是A和B内部边缘连接点的全部权重。

一个区分是否关联的一个重要特征是他们是自然相关的：

$$\begin{aligned}Ncut(A,B)=&\frac{cut(A,B)}{assoc(A,V)}+\frac{cut(A,B)}{assoc(B,V)}\\=&\frac{assoc(A,V)-assoc(A,A)}{assoc(A,V)}+\frac{assoc(B,V)-assoc(B,B)}{assoc(B,V)}\\=&2-(\frac{assoc(A,A)}{assoc(A,V)}+\frac{assoc(B,B)}{assoc(B,V)})\\=&2-Nassoc(A,B)\end{aligned} \tag{4-15}$$

因此，我们在分组算法中寻求两个分组准则，使组内非关联最小化同时使组内关联最大化。

规范化分割聚类算法

规范化分割聚类算法($normalized\,cut,Ncut$) 是一种图像分割方法，该方法的基本思想是把图像中的每个像素点（标量场的每个标量数据）看作图中的一个顶点，以图像像

素点的颜色信息、空间信息和纹理信息来定义其相似度作为顶点间的权重，来构造一个带权重的无向图，在图上建立一个稀疏的关系矩阵，并利用该矩阵的谱信息对图或图像进行划分。具体步骤如下：

步骤1：对含有n个像素的图像，构造一个带权重的无向图$G = (V; E; W)$ 。V 是图中顶点的集合，E是图中连接顶点i和j的边$e(i; j)$的集合；W是一个n阶对称矩阵，矩阵中的元素$w(i; j)$表示边$e(i; j)$上的权值，它表示顶点i 和j之间的相关性。

步骤2：解广义特征值方程：$(D - W)x = \lambda Dx$。这里只需要计算最小的几个特征值及其对应的特征向量；D是对角矩阵，对角线上的元素$d(i, j)$等于权重矩阵W第i行上所有元素的和。

步骤3：利用已经得到的谱的信息（第二小的特征值对应的特征向量）将图一份为二。

步骤4：判断分割后的图像是否还需要再次分割，如需要分割，再对分割后的两部分图像分别重复上述过程进行分割。

n阶权重矩阵W中元素$\omega(i, j)$表示顶点i 和j 之间的相关性；权$\omega(i, j)$ 越大，顶点i和j的相关性越强，分在同一个子集中的可能性越大。任一顶点与自己的相关性最大。权$\omega(i, j)$的取值中应包含图像的重要特征信息，可以用高斯核定义权：

$$\omega(i, j) = \exp(-\|Q(i) - Q(j)\|/2\sigma^2) \tag{4-16}$$

其中$Q(i)$是顶点i的空间坐标或其它的特征信息。由于图像一般只是局部相关的，为方便计算，当顶点i和j间距离大于$r(r \ll n)$ 时，$\omega(i, j)$取值为0。因此，最终得到的权重矩阵W 是一个稀疏的对称矩阵。

规范化分割方法的主要特点有：(1)图像在大多数情况下仅仅是局部相关的，所以最后得到的矩阵形都是稀疏矩阵。(2) 通常只需要几个最小的特征向量就可以对图像进行很好分割，不必求出矩阵的全部特征向量。(3) 分割对特征向量的精确度要求不高。大多数情况下，只需要其分量的正、负号就可以分割图像。

2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	3	1	3	1	1	1	1	3	1	3	1
1	1	1	1	1	1	1	1	1	1	3	3	1	1	1	1	1	1	1	1
1	1	1	3	1	1	3	3	1	3	1	3	1	1	3	3	1	1	1	1
1	3	1	1	1	1	3	1	1	1	1	1	1	1	1	1	3	0	0	0

图 4-12: problem 3b 类别标签

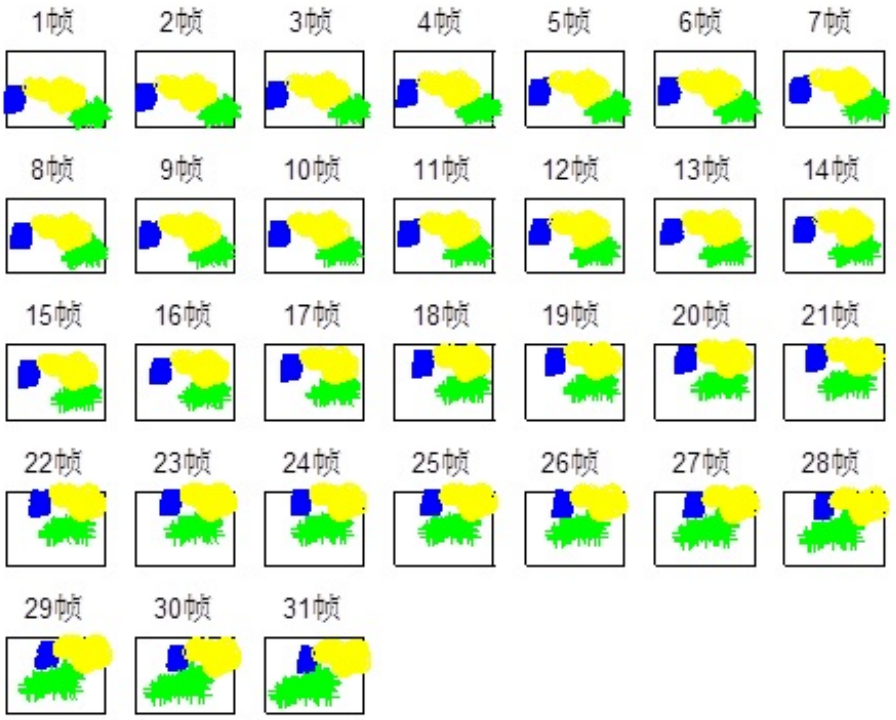

图 4-11: problem 3b

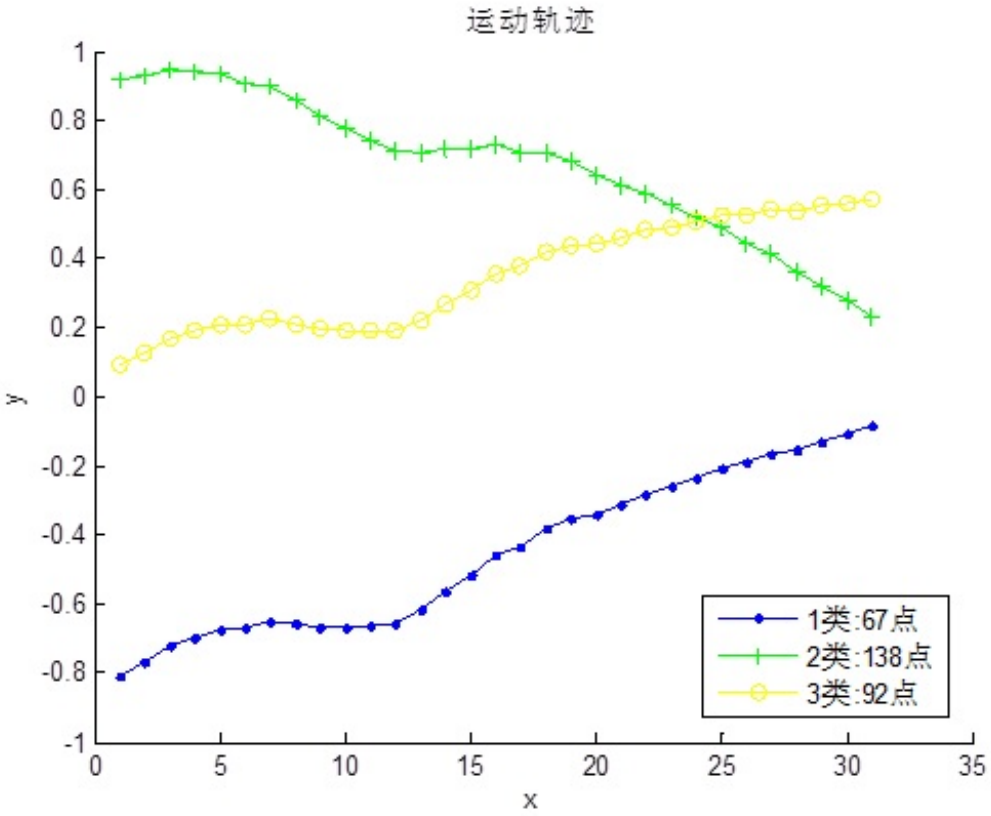

图 4-13: problem 3b 运动轨迹

我们利用$Ncut$算法，从数据矩阵中，提取出来三个不同运动的特征点轨迹，并将它们逐帧显示出来。特征点的分类如图4-11所示，详细结果见图4-12。图像中物体运动轨迹如图4-13 所示，一类67 个点，表示物体一，二类138 个点，表示物体二，三类92个点，表示物体三。物体一和物体三的运动方向基本一致，但与物体二反向。

4.3.3 模型$3-c$

图 4-14: problem 3c

1	1	1	1	1	2	2	2	2	2	1	1	1	1	1	2	2	2	1	2

图 4-15: problem 3c 类别标签

我们利用$SMMC$算法对数据进行处理。虽然这组数据的数据量并不是很大，但是维数很高，所以在运行时间上比较长。

从运行结果上看（如图4-14），只有一张人像错判，因此我们有理由认为这个算法适合人脸识别问题。关于错判图像，究其原因可能是由于光线太暗，使得其中的特征点变得模糊，从而导致错判。分类的标签如4-15所示，倒数第2个就是归类出错的图像。

4.4 模型四

4.4.1 模型$4-a$

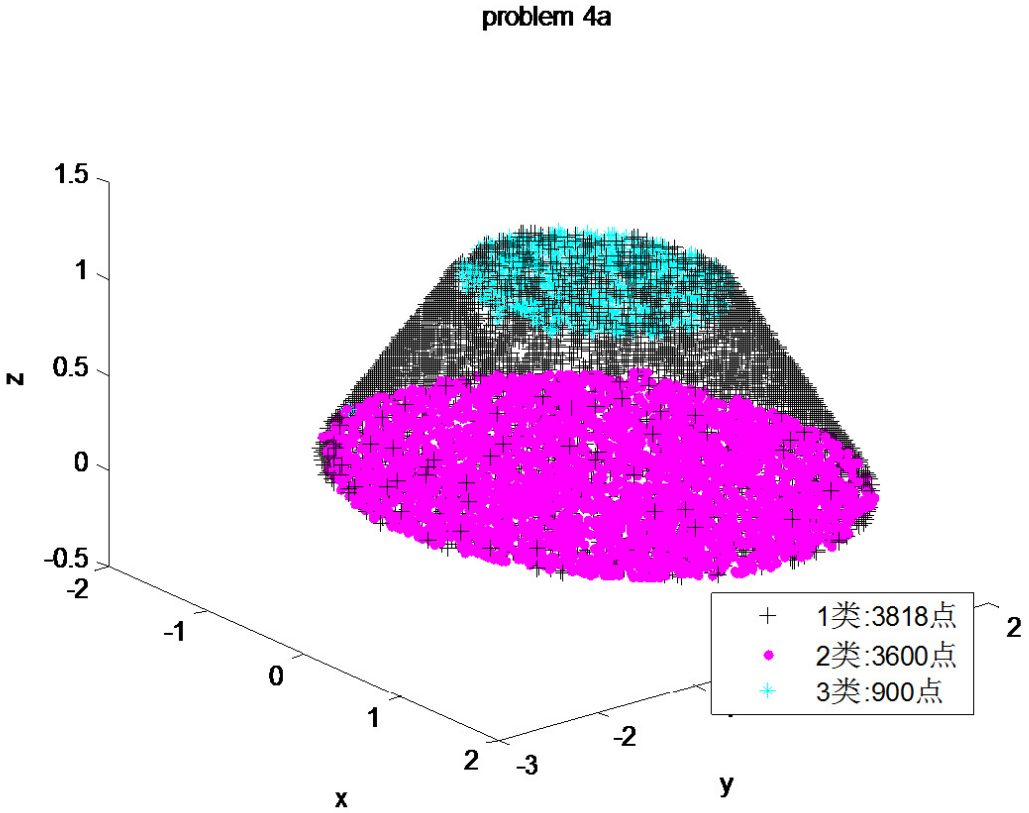

图 4-16: problem 4a

我们尝试过先把上下两个底面分离出来，也想过把三维数据投影到二维平面进行聚类，但是没有成功。采用$SMMC$方法，选取参数(3, 2, 200, 20, 11)，运行程序得到结果如下。本题程序比较耗内存，在8G内存、12核心Xeon CPU的戴尔塔式服务器Precision T5600上运行一次大约需要4分钟。

1 类：3818 点，图中黑色侧面;

2 类：3600 点，图中红色底面;

3 类：900 点，图中浅蓝色顶面。

侧面、顶面和底面被成功分开，但是仍有个别（当前结果为8个）属于顶面的点被分在底面。然而并没有错分在侧面，这是因为两个平面的结构比较相近。本文算法的参数比较难调，对参数敏感，有时相同的参数未必出同样的结果。

4.4.2 模型$4-b$

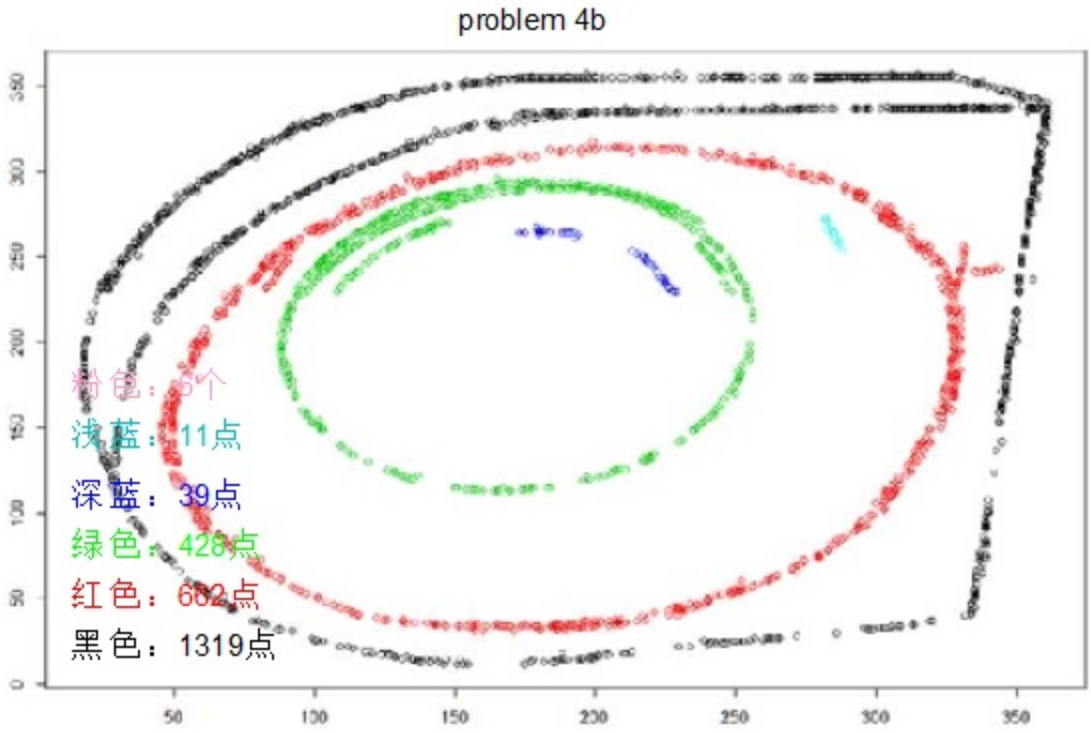

图 4-17: problem 4b

我们利用$SMMC$ 算法对机器外部轮廓数据进行聚类分析，由于本题并没有给出类数，所以我们利用参考文献[10]中给出的算法，从而估计出该机器外部轮廓数据分为六类，见图4-17。分别为第一类有6 个点，第二类有11 个点，第三类有39 个点，第四类有428个点，第五类有662个点，第六类为1319个点。

V 评价与推广

本文解决问题的过程中，大部分都使用了$SMMC$算法。这个算法具有很多优点，但是它还比较粗糙、值得继续完善和改进，例如：这个算法对整个数据集合中每个类别都需要通过谱分析来获取不同的分解成分，这使求解数据在维数高、样本量大时既占用较大的内存，也十分耗时；在最后做$k-means$聚类时，由于初始点的选取，导致结果不是十分稳定。而且这个算法在碰到有交点的情况时处理的结果也不甚理想。因此需要从例如原始矩阵的处理等方面来改进这个算法，最终得到更快速有效的算法来进行求解。除此之我们还使用了SSC和$Ncut$算法，这两个算法相比于传统的聚类方法有了一定的改进，但是都具有的各自的局限性。其中SSC算法只适用于子空间彼此相互独立的情况下，对子空间不独立的情况效果并不理想。而$Ncut$ 算法在运动分割问题上表现突出，在其他问题上的效果并不显著。

本文所有相关代码等文件，包括TEX 论文在github开源。

VI 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.

[9] 单世民,王新艳,张宪超.高维分类属性的子空间聚类算法[J]. 小型微型计算机系统,2009,10:2016-2021.

[10] 王勇,基于流形学习的分类与聚类方法及其应用研究,2011-04.