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

当今社会，各式各样的数据充斥着人们生活的各方各面，对大规模数据的分析与处理在科学研究领域占据着越来越重要的地位。数据的维数之高，结构之复杂为数据的分析与处理带来了一定的困难。本文针对数据多流形结构的特点，结合已有聚类模型进行聚类分析，得到如下成果：

针对问题 1：依据该题数据采样于完全独立的两个子空间的特点以及稀疏表示的含义，本文对数据建立稀疏子空间聚类（SSC）模型进行聚类分析，得到第 41~140 个数据属于类别 1、其余编号数据属于类别 2 的结果。并利用基于主成分分析（PCA）的 K-means 聚类算法建模降维，进行模型检验。经检验，稀疏子空间聚类（SSC）模型的聚类分析结果有效。

针对问题 2：问题 2 可分为（1）线性流形聚类问题；（2）非线性流形聚类问题。根据线性、非线性的不同特点，本文对线性问题建立稀疏子空间聚类（SSC）模型进行聚类分析，对非线性问题建立谱多流形聚类（SMMC）模型进行混合流形聚类分析。有效地将 2(a)的两条交点不在原点且互相垂直的直线分为两类；将 2(b)的一个平面和两条直线，分为三类；将 2(c)的两条不相交的二次曲线分为两类；将图 2(d) 为两条相交的螺旋线分为两类。

针对问题 3：针对视觉重建中的特征提取问题，依据该问数据局部非线性

但整体线性的特点，本文建立基于 K-means 的 SSC 模型进行聚类分析，并采用 SMMC 模型进行检验，获得了可靠的聚类分析成果，有效地将 3(a)中十字上的点分成两类；针对 3(b)运动分割问题，依据该题数据高维特点，本文建立 PCA、Isomap 及 LLE 三种降维模型，与 K-means 算法相结合进行分析，并建立 SMMC 模型进行检验，将视频中一帧的特征点轨迹分成三类，得到了误差极小的聚类分析结果；针对 3(c)人脸识别问题，依据人脸图像维度高和亮度变化等因素，本文先对数据进行标准化处理，消除光照影响，再通过建立 PCA、Isomap 及 LLE 三种降维模型，使用流形学习方法，提取到不受亮度变化因素影响的人脸低维流形，与 K-means 算法相结合进行分析，最终成功将这 20 幅人脸图像分成两类，获得了有效聚类分析成果。

针对问题 4：本问可概括为非线性流形聚类问题，通过建立谱多流形聚类（SMMC）模型进行聚类分析。将 4(a) 圆台的点云，按照其所在的面分为（即圆台按照圆台的顶、底、侧面）三类，获得了有效分析成果；将 4(b)机器工件外部边缘轮廓线中不同的直线和圆弧进行了聚类，将侧面与地面分开，同时人为地进行减噪处理，所得结果与减噪前一致，并与解析几何原理聚类的结果有所误差，模型需进一步改进。

关键词：稀疏子空间；谱聚类；多流形聚类；降维模型

目　　录

1 问题背景

如今，我们已经进入了一个信息爆炸的时代，对大规模数据的分析与处理在科学研究领域占据着越来越重要的地位。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。数据的维数之高，结构之复杂给数据的分析与处理带来了一定的困难。为此，人们通常用含参数的模型来表示一组数据，子空间表示作为一种简单的参数模型被广泛地应用。

通过子空间聚类（Subspace Clustering），可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。本文中针对所给问题采用基于谱聚类（Spectral Clustering）的子空间分类方法进行求解。代表性的基于谱聚类的子空间分割方法包括稀疏表示（Sparse Representation, SR）和低秩表示（Low-Rank Representation, LRR）[5][6]。

2 模型假设

(1) 题中所给数据均无差误。

(2) 高维数据分布于多个低维子空间的并，即高维空间中的数据本质上属于某个低维子空间，能够在低维子空间中进行线性表示，即高维数据的低维化表示能够揭示数据所在的本质子空间。

(3) 数据集可通过混合线性子空间建模。

(4) 低维流形的数目和维数已知。

(5) 全局非线性流形在局部能被一系列局部线性流形很好地逼近。

(6) 主成分分析模型可有效穿过相交线性流形。

(7) 被同一个线性分析模型逼近的数据点通常具有相似的局部切空间，并且这些切空间可被局部分析建模的主子空间很好地近似。

3 符号系统

符号	含义
r	矩阵的秩
S_α	第α个子空间
L	拉普拉斯矩阵
$\|X\|_+$	矩阵X的核范数，$\|X\|_+ = \sum_i \sigma_i$，$\sigma_i$是$X$的奇异值
$\|X\|_1$	矩阵X的l_1-范数，$\|X\|_1 = \sum_i \sum_j \|X_{ij}\|$
S_t	协方差矩阵
$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(q_i - q_j\right)^2$	*Laplacian* 矩阵
W	权重矩阵
D	度矩阵
$Cut(G_1, G_2)$	损失函数
$R(L,q)$	*Rayleigh quotient*，瑞利熵
E	所有数据的均方差之和
Θ_i	局部切空间
p_{ij}	x_i与x_j之间结构相似性函数
q_{ij}	x_i与x_j之间局部相似性函数
w_{ij}	相似性权值
$Knn(x)$	x的K个近邻数据点
$\sum_x$	x点处的局部采样协方差矩阵
μ_m	数据的均值向量
ε_m	数据的噪声
$error(M)$	M个局部线性分析器逼近潜在流形的重构误差
M	局部化模型数
$\phi(Y)$	成本函数值

4 问题 1 模型的建立与求解

4.1 问题 1 重述

当子空间独立时，子空间聚类问题相对容易。附件一 1.mat 中有一组高维数据，它采样于两个独立的子空间。请将该组数据分成两类，并制作一个表格输出样本的类别标签，每行 20 个。

4.2 问题 1 分析

由题干给出的子空间聚类叙述分析可知，问题 1 的本质是将高维空间降为低维空间后，进行聚类分析。本文利用原数据采样于完全独立的两个子空间及高维的数据具有稀疏与降维相似的特点，选用稀疏子空间聚类（SSC）模型进行聚类分析，再利用主成分分析（PCA）建模降维，结合 K-means 算法进行模型检验。

4.3 模型建立与分析

4.3.1 PCA 模型

PCA（Principal Component Analysis, 主成分分析）是一种对数据进行分析的技术，最重要的应用是对原有数据进行简化，有效的找出数据中最“主要”的元素和结构，去除噪音和冗余，将原有的复杂数据降维，揭示隐藏在复杂数据背后的简单结构[9] [10]。虽然 PCA 模型本身也存在诸多的假设条件，决定它存在一定的限制，但该模型具有与简单、无参数限制及普适性特点，符合题目“一般性”要求。

PCA 以方差的大小来衡量信息量的多少，方差越大提供的信息越多，方差越小提供的信息越少。PCA 通过线性变换构造数据的低维表示，其基本思想是尽可能地保留数据的方差或信息量，丢掉信息量少的方差。

给定一组观测数据 $X=\{\mathrm{x}_i \in \mathrm{R}^D, \mathrm{i}=1,\ldots,\mathrm{N}\}$，主成分分析的目标可以表述为寻找一组相互正交的投影方向或一个列正交的线性投影矩阵 $G \in R^{D\times d}$，使得投

影后的低维嵌入表示 $y_i = G^T x_i$ 具有最大的方差。

记原始数据按列堆叠构成的矩阵为 $X = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N] \in \mathrm{R}^{D \times d}$，低维嵌入表示按列堆叠构成的矩阵为 $Y = [y_1, y_2, \ldots, y_N] \in \mathrm{R}^{D \times d}$。原始数据的样本协方差矩阵为 $S_t = \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T = XHX^T$,其中 $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$ 为样本均值，$H = I - \frac{1}{N} ee^T$ 为中心化矩阵，I 是单位矩阵，$e \in \mathrm{R}^{D \times d}$ 是元素全为 1 的列向量。进而，可求得低维嵌入表示的协方差矩阵为 $\sum_{i=1}^{N} (y_i - \bar{y})(\mathrm{y}_i - \bar{y})^T = YHY^T = G^T XHX^T G = G^T S_t G$,

其中 $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$ 为低维嵌入表示的均值。PCA 的目标函数可以表示为下列数学形式：

$$\begin{aligned} &\arg\max_{G} \quad \mathrm{tr}(\mathrm{G}^T \mathrm{S}_t \mathrm{G}) \\ &\text{S.t.} \quad\quad \mathrm{G}^T \mathrm{G} = \mathrm{I} \end{aligned} \tag{4-1}$$

该目标函数的最优解 G 可以通过对原始数据的协方差矩阵 S_t 进行谱分解或特征分解来求解，即假设 $\mathrm{S_t}$ 的谱分解为：

$$S_t = U \Lambda U^T \tag{4-2}$$

其中，$\Lambda = diag(\lambda_1, \ldots, \lambda_D)$ 是由 S_t 的特征值组成的对角矩阵，满足：$\lambda_i \geq \lambda_{i+1} (i = 1, 2, \ldots, D-1), U = [\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_D], \mathrm{u}_i (\mathrm{i} = 1, 2, \ldots, \mathrm{D})$ 为 λ_i 对应的特征向量且 $\mathrm{U}^T \mathrm{U} = \mathrm{I}$。

在 PCA 的目标函数下，通常取最优解 G 为协方差矩阵 S_t 的最大的 d 个特征值对应的特征向量，即 $G = [\mathrm{u}_1, \mathrm{u}_2, \ldots, \mathrm{u}_d]$。PCA 学习后的低维嵌入表示的中心通常在原点，即 $y_i = G^T (\mathrm{x}_i - \bar{x})$。

PCA 的一个显著特点和优势是：在不同的理解下可以有不同的解释。其中一个解释是，PCA 是最小二乘意义下的最优线性重构模型，即其目标函数的数学形式可以重述为：

$$\begin{aligned} \arg\min_{G} \quad & \sum_{i=1}^{N} \left\| (x_i - \bar{x}) - GG^T (x_i - \bar{x}) \right\|^2 \\ \text{S.t.} \quad & G^T G = I \end{aligned} \tag{4-3}$$

4.3.2 稀疏子空间聚类模型

给定一组数据设 $X = [x_1, x_2, \cdots, x_n] \in R^{D \times N}$，这组数据属于 k（k 已知或未知）个线性子空间 $\{S_i\}_{i=1}^{k}$ 的并，子空间聚类是指将这组数据分割为不同的类，在理想情况下，每一类对应一个子空间。而稀疏性是指用尽可能少的基的线性组合表示数据，使数据的线性表示中的非零系数最少。根据子空间的定义，非零系数的位置表明该数据属于由相应基组成的子空间，同时非零系数的个数也反映了数据本身的维数，因此可以通过稀疏子空间聚类分析可以反映数据的子空间特性[1]。

稀疏子空间聚类（Sparse Subspace Clustering, SSC）的基本思想是：将数据 $x_i \in S_\alpha$（S_α 为 $\{S_i\}_{i=1}^{k}$ 中的一个子空间）表示为所有其他数据的线性组合，

$$x_i = \sum_{j \neq i} Z_{ij} x_j \tag{4-4}$$

并对系数施加一定的约束使得在一定条件下对所有的 $x_j \notin S_\alpha$。对应的 $Z_{ij} = 0$。将所有数据及其表示系数按一定方式排成矩阵，则上式可写为：

$$X = XZ \tag{4-5}$$

当 x_i 和 x_j 属于不同的子空间时，则表示系数 $Z_{ij} = 0$。若已知数据的基，则在一定条件下可使系数矩阵 Z 具有块对角结构，即

$$Z = \begin{pmatrix} z_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & z_n \end{pmatrix} \tag{4-6}$$

稀疏子空间聚类的基本框架如图 4-1 所示：

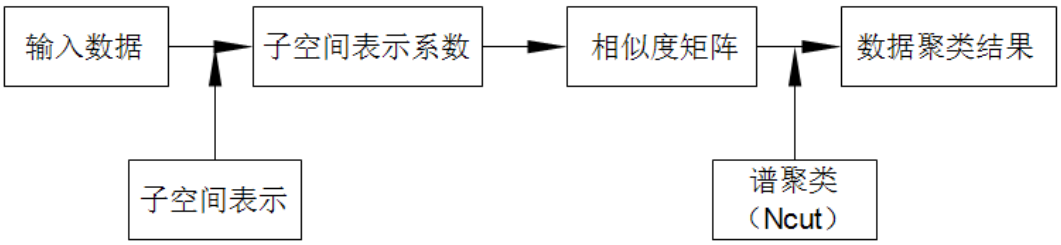


图 4-1 稀疏子空间聚类的基本框架

4.3.3 谱聚类

谱聚类（Spectral Clustering, SC）是一种基于图论的聚类方法——将带权无向图划分为两个或两个以上的最优子图，使子图内部尽量相似，而子图间距离尽量距离较远，以达到常见的聚类的目的[2]。其中的最优是指最优目标函数不同，可以是割边最小分割——如图 4-2 的 Smallest Cut（如后文的 Minimum Cut），也可以是分割规模差不多且割边最小的分割——如图 4-2 的 Best Cut（如后文的 Normalized Cut）。

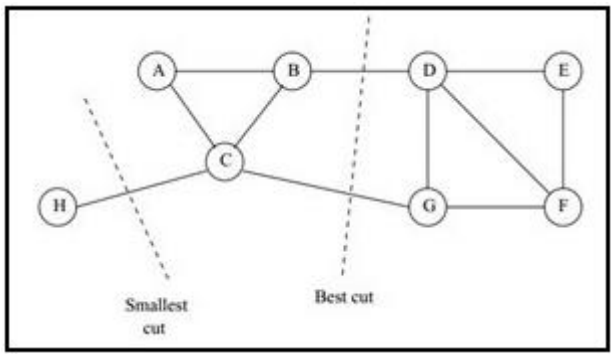


图 4-2 谱聚类无向图划分——Smallest Cut 和 Best Cut

这样，谱聚类能够识别任意形状的样本空间且收敛于全局最优解，其基本思想是利用样本数据的相似矩阵（拉普拉斯矩阵）进行特征分解后得到的特征向量进行聚类。谱聚类可以理解为：降维过程+其他聚类方法，最终对 $m \times n$ 矩阵的行向量聚类。

(1) 几项定义：

① $G(V,E)$ 表示无向图，其中：$V=\{v_1,v_2,\ldots,v_n\}$ 表示点集，E 表示边集。$G_1,G_2,\ldots,G_n$ 为 G 被划分成的子图。

② w_{ij} 表示 v_i 与 v_j 之间的关系，称为权重，对于无向图，$w_{ij}=w_{ji}>0$ 并

且。

③ 损失函数：

$$Cut(G_1,G_2)=\sum_{i\in G_1,j\in G_2}w_{ij} \tag{4-7}$$

为划分时子图被“截断”边的权重和。

④ $q=(q_1,q_2,\ldots,q_n)$ 为 n 维向量，用来表示划分方案，以下以二分为例叙述，

$$q_i=\begin{cases}c_1,i\in G_1\\c_2,i\in G_2\end{cases} \tag{4-8}$$

$$Cut(G_1,G_2)=\sum_{i\in G_1,j\in G_2}w_{ij}=\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(q_i-q_j\right)^2}{2\left(c_1-c_2\right)^2} \tag{4-9}$$

其中：*Laplacian* 矩阵：

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(q_i-q_j\right)^2&=\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(q_i^{\,2}-2q_iq_j+q_j^{\,2}\right)\\&=-\sum_{i=1}^{n}\sum_{j=1}^{n}2w_{ij}q_iq_j+\sum_{i=1}^{n}2q_i^{\,2}\sum_{j=1}^{n}w_{ij}\\&=2q^T(D-W)q\end{aligned} \tag{4-10}$$

D（度矩阵）为对角矩阵：

$$D_{ii}=\sum_{j=1}^{n}w_{ij} \tag{4-11}$$

⑤ 拉普拉斯矩阵，拉普拉斯矩阵 $L=D-W$，其中 W 为权重矩阵，则：

$$\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(q_i-q_j\right)^2=2q^TLq \tag{4-12}$$

$$q^TLq=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left(q_i-q_j\right)^2\geq 0 \tag{4-13}$$

可知，L 为半正定矩阵，所有特征值非负，最小特征值为 0，且对应的特征向量为单位向量。此时，损失函数：

$$Cut(G_1,G_2)=\sum_{i\in G_1,j\in G_2}w_{ij}=\frac{q^TLq}{\left(c_1-c_2\right)^2} \tag{4-14}$$

通过以上转换，可以看到图的划分问题转化为 q^TLq 的条件最小值问题。

(2) 最优化方法

① Minimum Cut 法（以下简称 Mcut）

令 $q_i = \begin{cases} c, i \in G_1 \\ -c, i \in G_2 \end{cases}$，s.t. $q^T q = \sum_{i=1}^{n} q_i^2 = nc^2$，求 $\min(q^T L q)$。

瑞利熵（*Rayleigh quotient*）：

$$R(L,q) = \frac{q^T L q}{q^T q} \tag{4-15}$$

根据瑞利熵性质：$R(L,q)$ 的最小值，次小值，……，最大值，分别在 q 为 L 的最小特征值，次小特征值，……，最大特征值对应的特征向量时取得。

② Ratio Cut 法（以下简称 Rcut）

$$RCut(G_1,G_2) = Cut(G_1,G_2) \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \tag{4-16}$$

其中：n_1、n_2 为划分到子空间（子图 1，子图 2）中的元素个数。

$$\begin{aligned} RCut(G_1,G_2) &= \sum_{i \in G_1, j \in G_2} w_{ij} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) = \sum_{i \in G_1, j \in G_2} w_{ij} \left(\frac{(n_1+n_2)^2}{n_1 n_2 n} \right) \\ &= \sum_{i \in G_1, j \in G_2} w_{ij} \left(\sqrt{\frac{n_1^2}{n_1 n_2 n}} + \sqrt{\frac{n_2^2}{n_1 n_2 n}} \right)^2 \\ &= \sum_{i \in G_1, j \in G_2} w_{ij} \left(\sqrt{\frac{n_1}{n n_2}} + \sqrt{\frac{n_2}{n_1 n}} \right)^2 \end{aligned} \tag{4-17}$$

令 $q_i = \begin{cases} \sqrt{\frac{n_1}{n n_2}}, i \in G_1 \\ -\sqrt{\frac{n_2}{n n_1}}, i \in G_2 \end{cases}$，$RCut(G_1,G_2) = \sum_{i \in G_1, j \in G_2}^{n} w_{ij} \left(q_i - q_j \right)^2 = q^T L q$

s.t. $q^T q = \sum_{i=1}^{n} q_i^2 = \sum_{i \in G_1}^{n} q_i^2 + \sum_{i \in G_2}^{n} q_i^2 = n_1 \frac{n_2}{n_1 n} + n_2 \frac{n_1}{n_2 n} = 1$，求 $\min(q^T L q)$。

瑞利熵：

$$R(L,q) = \frac{q^T L q}{q^T q}$$

③ Normalized Cut 方法（以下简称 Ncut）

$$NCut(G_1,G_2)=Cut(G_1,G_2)\times\left(\frac{1}{d_1}+\frac{1}{d_2}\right) \tag{4-18}$$

其中：d_1、d_2为划分到子空间（子图 1，子图 2）中的权重和。

令 $q_i=\begin{cases}\sqrt{\dfrac{d_1}{nd_2}}, i\in G_1\\ -\sqrt{\dfrac{d_2}{dn_1}}, i\in G_2\end{cases}$，$NCut(G_1,G_2)=\sum_{i\in G_1,j\in G_2}^{n} w_{ij}\left(q_i-q_j\right)^2=q^TLq$

s.t. $q^TDq=\sum_{i\in G_1}^{n}q_i^2\sum_{j=1}^{n}w_{ij}+\sum_{i\in G_2}^{n}q_i^2\sum_{j=1}^{n}w_{ij}=\frac{d_2}{d_1d}d_1+\frac{d_1}{d_2d}d_2=1$，求 $\min(q^TLq)$。

广义瑞利熵：

$$R(L,q)=\frac{q^TLq}{q^TDq} \tag{4-19}$$

$$\begin{gathered}Lq=\lambda Dq\\ Lq=\lambda D^{\frac{1}{2}}D^{\frac{1}{2}}q\\ D^{-\frac{1}{2}}LD^{-\frac{1}{2}}D^{\frac{1}{2}}q=\lambda D^{\frac{1}{2}}q\\ L'=D^{-\frac{1}{2}}LD^{-\frac{1}{2}},q'=D^{\frac{1}{2}}q\end{gathered} \tag{4-20}$$

其中：λ为L的广义特征值，L'为规范拉普拉斯矩阵，对角元素全为 1。

(3) 以上 3 种方法中，Mcut 和 Rcut 只考虑了顶点数或权重和其中之一个要求，Ncut 法的目标是同时考虑最小化 cut 边和划分平衡两个要求，而计算的L'相比计算L要稍具优势。

(4) 计算步骤：

① *Un-normalized Spectral Clustering* 步骤：

输入：样本及类别数 K。

A) 根据样本建立权重矩阵W；

B) 根据W，计算度矩阵D，进而计算拉普拉斯矩阵L；

C) 计算L的特征值及特征向量$V_e=\left(v_{e_1},v_{e_2},\ldots,v_{e_k}\right)$，并对$V_k$的行向量进行聚类，得到$K$个 Cluster。

② *Normalized Spectral Clustering* 步骤：

输入：样本及类别数 K。

A) 根据样本建立权重矩阵 W；

B) 计算拉普拉斯矩阵 L 及 $L' = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$；

C) 计算 L' 的特征值及特征向量 $V_e = \left(v_{e_1}, v_{e_2}, \ldots, v_{e_k}\right)$；

D) 取出前 k 小特征值对应的特征向量 $V_k = \left(v_{e_1}, v_{e_2}, \ldots, v_{e_k}\right)$ 并对 V_k 的行向量进行聚类，得到 K 个聚类。

(5) 可见不管是 L、L' 都与 W 联系特别大。如果将 W 看成一个高维向量空间，也能在一定程度上反映 item 之间的关系。将 W 直接 K-means 聚类，得到的结果也能反映 V 的聚类特性，而谱聚类的引入 L 和 L' 是使得 G 的分割具有物理意义。

而且，如果 W 的 item（即 n）足够大，导致难以计算出它的 K-means，可用 PCA 降维（仍为最大的特征值与向量）。

上述对将 W 当成向量空间矩阵，直观地符合我们的认知，但缺乏理论基础；而 L (L'等)的引入，使得计算具有理论基础，其前 k 个特征向量，也等价于对 L (L'等)的降维。

4.3.4 K-means 算法

K-means（K 均值）是一种基于划分的方法，该算法的优点是简单易行，时间复杂度为 $O(n)$，适用于处理大规模数据[7]。其核心思想是把 n 个数据对象划分为 k 个聚类，使每个聚类中的数据点到该聚类中心的平方和最小[8]，算法处理过程：

输入：聚类个数 k，包含 n 个数据对象的数据集。

输出：k 个聚类。

(1) 从 n 个数据对象中任意选取 k 个对象作为初始的聚类中心；

(2) 分别计算每个对象到各个聚类中心的距离，把对象分配到距离最近的聚类中；

(3) 所有对象分配完成后，重新计算 k 个聚类的中心；

(4) 与前一次计算得到的 k 个聚类中心比较，如果聚类中心发生变化，转(2)，否则转(5)；

(5) 输出聚类结果。

K-means 算法工作流程图如下：

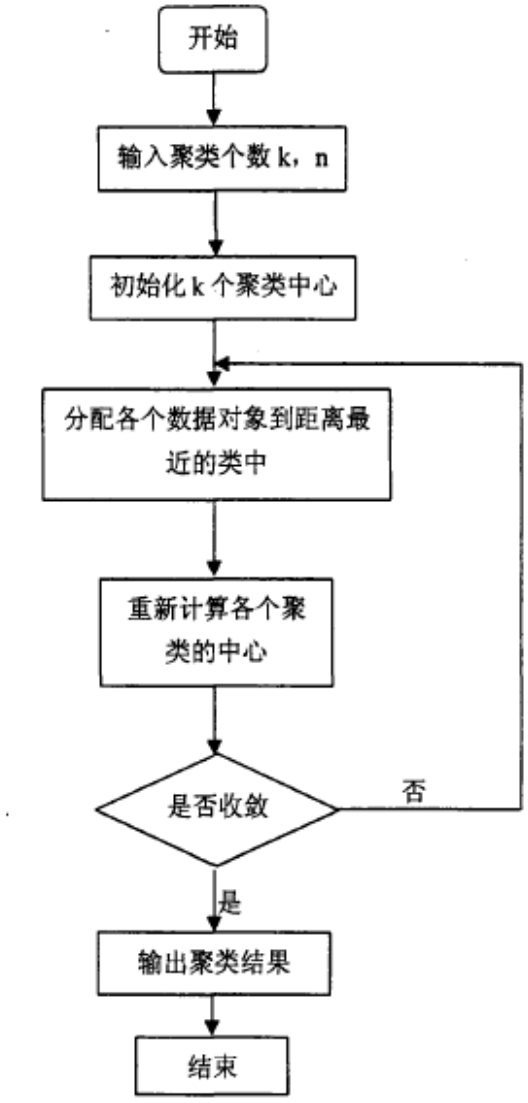


图 4-3 K-means 方法流程图

首先从 n 个数据对象中任意选择 k 个对象作为初始聚类中心，而对于所剩下的其它对象，则根据他们与这些聚类中心的相似度（距离），分别将他们分配给与其最相似的（聚类中心所代表的）聚类。然后再计算每个所新聚类的聚类中心（该聚类中所有对象的均值）。不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数，具体定义如下：

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \tag{4-21}$$

其中：E 为数据库中所有对象的均方差之和；

p 为代表对象的空间中的一个点；

m_i 为聚类 c_i 的均值(p 和 m_i 均是多维的)。

公式(4-21)所示聚类标准旨在使所获得的 k 个聚类具有以下特点：各聚类本身尽可能紧凑，而各聚类间尽可能的分开。K-means 算法是相对可伸缩的和高效率的，因为它的计算复杂度为 $O(nkt)$，其中 n 为对象个数，k 为聚类个数，t 为迭代次数，通常有 $t \le n$，$k \le n$，因此它的复杂度通常也用 $O(n)$ 表示。

4.3.5 问题 1 求解

将 200 个数据按列从小到大每行排 20 个（即第一行对应的第 1~20 列），求解结果用数字 1、2 分别作为各类的类别标签，对结果进行列表表示。

(1) 解法一

利用 SSC 模型，在 Matlab 中引入凸优化函数 CVX（CVX 能使目标函数快速求到全局最优解），建立相似度矩阵，从而可求得邻接矩阵，接着分别运用三种谱聚类切割方法（Mcut、Rcut、Ncut），再利用 K-means 算法进行聚类分析。流程图如下：

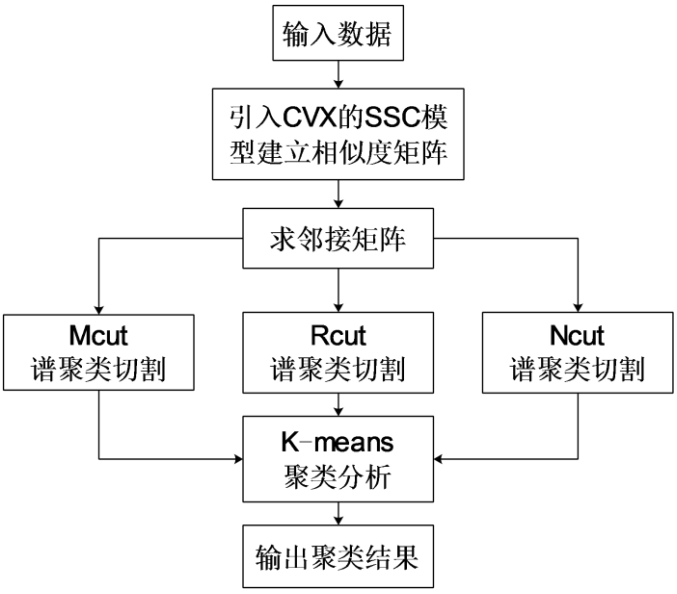


图 4-4 SSC 解法流程图

SSC 的三种切割方法结果一致，如下表：

表 4-1 SSC 模型按样本类别标签求解结果

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
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
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

(2) 解法二

先利用 PCA 模型进行主成分提取，各主成分的 Latent 值如下图所示：

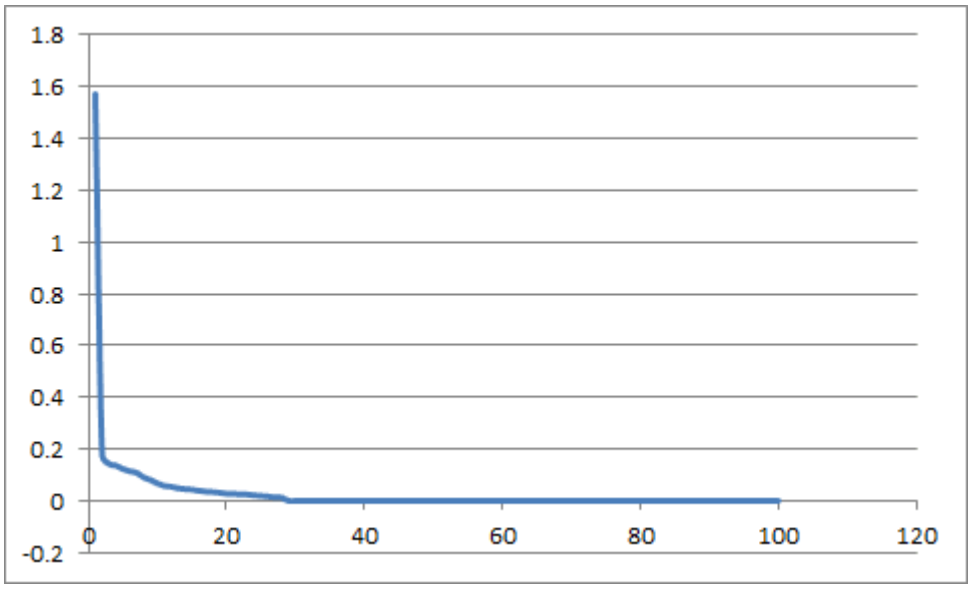


图 4-5 PCA 主成分提取 Latent 值结果

由上图可知，从原有的 200 个主成分中提取前 28 个，这 28 个主成分能够代表原数据的 99.9% 以上，因而可以用这前 28 个主成分来表示 1.mat 中的所有数据，再利用 K-means 算法进行聚类分析。

其结果如下：

表 4-2 PCA 模型按样本类别标签求解结果

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

4.4 结果分析

可以看到，两种解法结果基本一致，区别在于解法一将第 21 个和第 146 个的数据归入了第 2 类中，而解法二是将该两个编号的数据归入到了第 1 类中。原因是，从向量角度分析，通过求解极大无关组向量[3][4]，可得到解法二中的两个子空间的基向量类别标签分别如下：

表 4-3 基向量编号

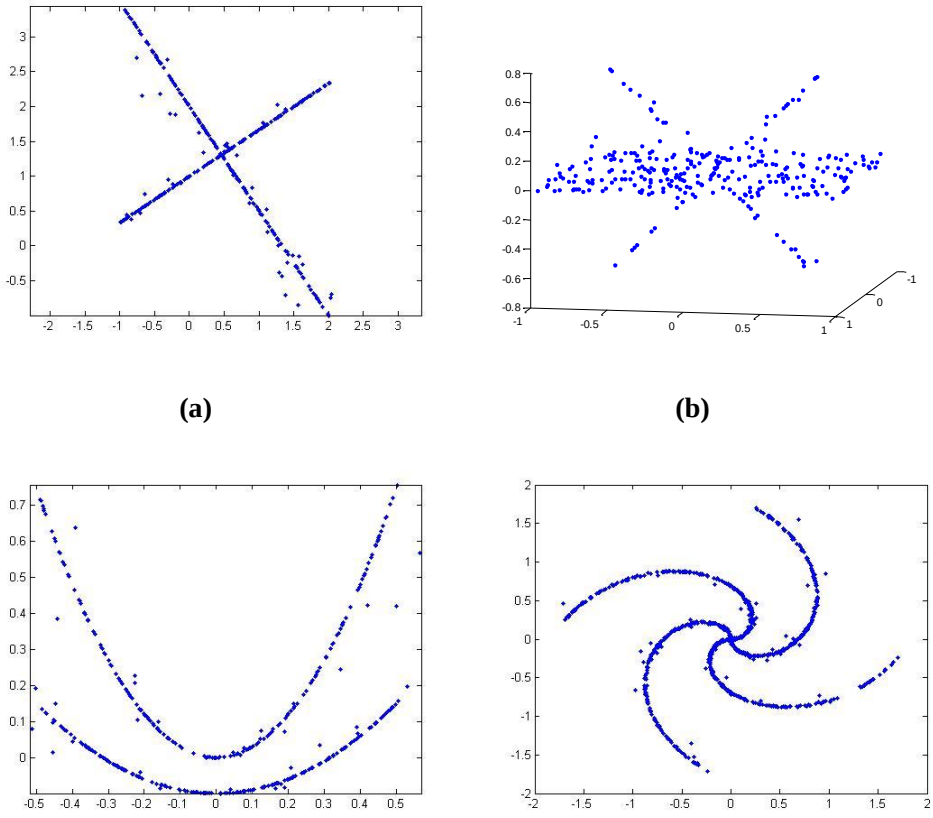
类别 1	1	2	3	4	5	6	7	8	9	10	21	146				
类别 2	41	42	43	44	45	46	47	48	49	50	52	64	73	94	102	115

第 21 个和第 146 个向量是与其他所有向量都线性无关，换句话说，第 21 个和第 146 个向量本身可以构成独立的子空间，归在类别 1 或类别 2 中均可，因此，所得结果是有效的。

5 问题 2 模型的建立与求解

5.1 问题 2 重述

已知附件二中含有四个低维空间中的子空间聚类问题和多流形聚类问题的原始数据，4 个低维空间如图 5-1 所示。图 5-1(a)为两条交点不在原点且互相垂直的两条直线，要求将其分为两类；图 5-1(b)为一个平面和两条直线，不满足独立子空间的关系，要求将其分为三类；图 5-1 (c)为两条不相交的二次曲线，要求将其分为两类；图 5-1 (d)为两条相交的螺旋线，要求将其分为两类。并将结果分类画出。



(a) **(b)**

(c) **(d)**

图 5-1 四个低维子空间

5.2 问题 2 分析

问题 2 的四幅原图由图 5-1 及题干所示可分为两类：一类是图 5-1(a)与图 5-1 (b)，属于线性流形聚类问题；另一类为图 5-1 (c)与图 5-1 (d)，属于非线性

流形聚类问题。对这四幅图时进行聚类分析时，根据线性、非线性的不同特点，选取合适的谱聚类方法进行建模，并进一步验证聚类结果。本文中，针对图 5-1 (a)与图 5-1 (b)的线性问题，采用 SSC 方法来进行聚类（方法同问题 1 的解法一），并利用谱多流形聚类（SMMC）模型对聚类结果进行检验。针对图 5-1 (c)与图 5-1 (d)的非线性问题，采用 SMMC 方法来实现混合流形聚类分析。

5.3 模型建立与分析

5.3.1 谱多流形聚类模型

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[9] [10]。

(1) 相似性矩阵

相似性矩阵构造基于下述事实：① 尽管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性块上；② 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；③ 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。因此，可利用数据点所内含的局部几何结构信息来辅助构造更合适的相似性矩阵 W 。

只有当下面的两个条件同时满足时，才能够断定两个数据点是来自同一个流形聚类的：它们相互靠近同时具有相似的局部切空间。因此，在构造相似性矩阵时，既要考虑数据点之间的欧氏距离关系 $q_{ij}=q(\|x_i-x_j\|)$（称为 Local Similarity, 局部相似性），又要考虑数据点局部切空间之间的相似性 p_{ij}（称为 Structural Similarity, 结构相似性）。这两个相似性融合在一起来决定最后的相似性权值：

$$w_{ij}=f(p_{ij},q_{ij}) \tag{5-1}$$

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氏距离的一个单调递减函数同时是局部切空间之间相似性的单调递增函数。

SMMC 方法中所采用的函数 p ， q 和 f 的具体形式如下：

假设数据点 $x_i(i=1,\ldots,N)$ 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：

$$p_{ij}=p(\Theta_i,\Theta_j)=(\prod_{l=1}^{d}\cos(\theta_l))^o \tag{5-2}$$

在(5-2)中， $o\in N^+$ 是一个可调节参数。 $0\le\theta_1\le\ldots\le\theta_d\le\pi/2$ 是两个切空间 Θ_i,Θ_j 之间的主角度，递归地定义为：

$$\cos(\theta_1)=\max_{\substack{u_1\in\Theta_i,v_1\in\Theta_j\\ \|u_1\|=\|v_1\|=1}} u_1^T v_1 \tag{5-3}$$

$$\cos(\theta_l)=\max_{\substack{u_l\in\Theta_i,v_l\in\Theta_j\\ \|u_l\|=\|v_l\|=1}} u_l^T v_l, l=2,\ldots,d \tag{5-4}$$

其中： $u_l^T u_i=0, v_l^T v_i=0, i=1,\ldots,l-1$ 。

数据点 x_i 和 x_j 之间的局部相似性简单地定义为：

$$q_{ij}=\begin{cases}1 & if\ x_i\in Knn(x_j)\ or\ x_j\in Knn(x_i)\\ 0 & otherwise\end{cases} \tag{5-5}$$

其中： $Knn(x)$ 代表 x 的 K 个近邻数据点。换句话说，局部相似性要求在构造近邻图时采用 K -近邻图，而不能采用完全图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：

$$w_{ij}=p_{ij}q_{ij}=\begin{cases}(\prod_{l=1}^{d}\cos(\theta_l))^o & if\ x_i\in Knn(x_j)\ or\ x_j\in Knn(x_i)\\ 0 & otherwise\end{cases} \tag{5-6}$$

容易验证，式（5-6）中定义的相似性权值具有所期望的性质，即来自不同聚类或流形的数据点之间有相对低的权值。其原因是：① 当来自不同流形的两个数据点相互远离时，根据（5-6）其相似性权值为 0；② 当来自不同流形的两个数据点靠近流形的相交区域时，它们有不同的局部切空间结构，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，可以期望得到更好的性能。

上述过程中是建于假设每个数据点的局部切空间是已知的情况，还需要有效地近似或估计这些局部切空间。

(2) 局部切空间

传统上，每个数据点的局部切空间可以通过给定样本点附近的局部近邻点来估计。具体地说，给定样本点 x 和它在欧氏空间度量下的 n 个近邻点 $N(x)=\left\{x^1,\ldots,x^n\right\}$，$x$ 附近的局部几何信息内含在该点处的局部采样协方差矩阵 $\sum_x$ 中：

$$\sum_x = \frac{1}{n}\sum_{i=1}^{n}(x^i-\mu_x)(x^i-\mu_x)^T \tag{5-7}$$

其中：$\mu_x = \frac{1}{n}\sum_{i=1}^{n}x^i$ 。

样本点 x 处的局部切空间 Θ_x 由 $\sum_x$ 的最大 d 个奇异值对应的左奇异向量给出。即，假设 $\sum_x$ 的 SVD 为：

$$\sum_x = \begin{pmatrix} U_d & \tilde{U}_d \end{pmatrix}\begin{pmatrix} \sum_d & 0 \\ 0 & \tilde{\sum}_d \end{pmatrix}\begin{pmatrix} V_d & \tilde{V}_d \end{pmatrix}^T \tag{5-8}$$

其中：$\begin{pmatrix} U_d & \tilde{U}_d \end{pmatrix} \in R^{D\times D}$ 是正交矩阵并且 $U_d \in R^{D\times d}$，则有：

$$\Theta_x = span(U_d) \tag{5-9}$$

但当两个数据点 x 和 y 非常靠近时，即使他们来自于不同的流形，根据 (5-9) 估计出的局部切空间 Θ_x 和 Θ_y 也非常相似。其原因是，在这种情况下 x 和 y 的基于欧氏距离度量的局部近邻 N(x) 和 N(y) 会严重地交叠在一起，从而导致了相似的局部协方差矩阵 $\sum_x$ 和 $\sum_y$。因此，将用一个快速有效的方法来逼近每个数据点附近的局部切空间。

基本思想基于如下事实：① 全局非线性流形在局部能被一系列局部线性流

形很好的逼近；② 主成分分析器可以有效地穿过相交线性流形；③ 被同一个线性分析器逼近的数据点通常具有相似的局部切空间并且这些切空间可以被局部分析器的主子空间很好地近似。因此，通过训练一系列局部线性分析器来逼近潜在的流形，估计每个给定数据点的局部切空间为其相应局部分析器的主子空间。

通过训练 M 个混合概率主成分分析器（Mixture of Probabilistic Principal Component Analyzers，MPPCA）来估计局部切空间，其中每个分析器由模型参数 $\theta_m = \left(\mu_m, V_m, {\sigma_m}^2\right)(m = 1,\ldots,M)$ 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。M 是用于逼近所有潜在的线性或非线性流形的局部线性子模型的个数。在第 m 个分析器模型下，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：

$$x = V_m y + \mu_m + \varepsilon_m \tag{5-10}$$

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 $y \sim N(0, I)$ 和 $\varepsilon_m \sim N(0, \sigma^2 I)$。在此模型下，$x$ 的边际分布为：

$$p(x \mid m) = \left(2\pi\right)^{-D/2} |C_m|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu_m)^T {C_m}^{-1}(x-\mu_m)\right\} \tag{5-11}$$

其中：模型协方差为：

$$C_m = {\sigma_m}^2 I + V_m {V_m}^T \tag{5-12}$$

模型参数 μ_m，V_m，${\sigma_m}^2$ 可以通过利用 EM 算法最大化观测数据 $X = \left(x_i, i = 1,\ldots N\right)$ 的对数似然来得到：

$$L = \sum_{i=1}^{n} \ln\left\{\sum_{m=1}^{M} \pi_m p(x_i \mid m)\right\} \tag{5-13}$$

其中：π_m 是混合比例，满足条件 $\pi_m \geq 0$ 和 $\sum_{m=1}^{M} \pi_m = 1$ 。EM 算法的过程为：

E-step：利用当前模型参数 $\theta_m = \left(\mu_m, V_m, {\sigma_m}^2\right)$ 计算：

$$R_{im}=\frac{\pi_m p(x_i\mid m)}{\sum_{m=1}^{M}\pi_m p(x_i\mid m)} \tag{5-14}$$

$$\pi_m^{new}=\frac{1}{N}\sum_{i=1}^{N}R_{im} \tag{5-15}$$

$$\mu_m^{new}=\frac{\sum_{i=1}^{N}R_{im}x_i}{\sum_{i=1}^{N}R_{im}} \tag{5-16}$$

M-step：重新估计参数V_m和$\sigma_m{}^2$为：

$$V_m^{new}=S_mV_m(\sigma_m^2I+T_m^{-1}V_m^TS_mV_m)^{-1} \tag{5-17}$$

$$(\sigma_m^2)^{new}=\frac{1}{d}tr[S_m-S_mV_mT_m^{-1}(V_m^{new})^T] \tag{5-18}$$

其中：

$$S_m=\frac{1}{\pi_m^{new}}\sum_{i=1}^{n}R_{im}(x_i-\mu_m^{new})(x_i-\mu_m^{new})^T \tag{5-19}$$

$$Tm=\sigma_m^2I+V_m^TV_m \tag{5-20}$$

采用 K-means 来初始化上述 EM 过程，最后，样本点x_i根据下述关系分组到第j个局部分析器：

$$p(x_i\mid j)=\max_m p(x_i\mid m) \tag{5-21}$$

同时其局部切空间由下式给出：

$$\Theta_x=span(V_j) \tag{5-22}$$

利用M个局部线性分析器逼近潜在流形的重构误差为：

$$error(M)=\sum_{j=1}^{M}\sum_{l=1}^{N_j}(x_l^j-\mu_j)^T(I-V_jV_j{}^T)(x_l^j-\mu_j) \tag{5-23}$$

其中：$x_l^j,l=1,\ldots,N_j$是分组到第j个局部分析器的N_j个数据点（$\sum_{j=1}^{M}N_j=N$）。

(3) SMMC 算法流程

通过(2)的方法估计出每个数据点的局部切空间后，即可计算得到相似性矩

阵 W，随后通过谱方法即可得到聚类结果，谱多流形聚类（SMMC）方法分组混合结构数据的基本过程如下：

输入：原始数据集 X，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。

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 根据式 (5-22) 确定每个点的局部切空间；

③ 利用式 (5-16) 计算两个局部切空间之间的结构相似性；

④ 利用式 (5-20)计算相似性矩阵 $W \in R^{D\times D}$，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 $u_1,\ldots,u_k$；

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

输出：原始数据对应的聚类结果。

(4) 参数影响

SMMC 方法中游三个可调节参数，即局部化模型数 M，近邻点数 K 和调节参数 o。需要考查这些参数的设置对 SMMC 方法聚类性能的影响，进而得出参数设置的一些推荐原则。

① SMMC 的性能更多地依赖于局部化模型数 M 的值，随着局部化模型数增加，平均重构误差减小。这意味着对潜在流形局部线性块的近似越来越好，从而对每个数据点局部切空间的估计也越来越可靠，进而使得 SMMC 具有更好的性能，因为其性能依赖于局部切空间的正确估计。

② 当近邻点数 K 既不太大也不太小时，SMMC 的性能在一个很大的参数选取范围内都是稳健的。其原因在于，当 K 值太小时会出现很多不连通的子聚类，而当它太大时局部限制会逐渐丧失。

③ 当调节参数 o 足够大时，SMMC 的性能很好。其原因在于，o 越大，来自不同流形的数据之间的可分离性越好，因为对 $x<1$ 而言当 o 变大时 x^o 趋向于 0。

④ 估计局部切空间的时间和局部化模型数 M 近似成线性关系，这与理论分析是一致的。SMMC 的总运行时间独立于局部化模型数，因为 SMMC 的计算复杂度主要由计算相似性矩阵和执行谱方法进行聚类分析组成，它们是独立于 M 的。

5.3.2 问题 2 求解

针对图 5-1 (a)与图 5-1 (b)的线性问题，采用 SSC 方法来进行聚类（方法同问题 1 的解法一），并利用 SMMC 模型对聚类结果进行检验。针对图 5-1 (c)与图 5-1 (d)的非线性问题，采用 SMMC 方法来实现混合流形聚类分析。

SMMC 模型流程图如下：

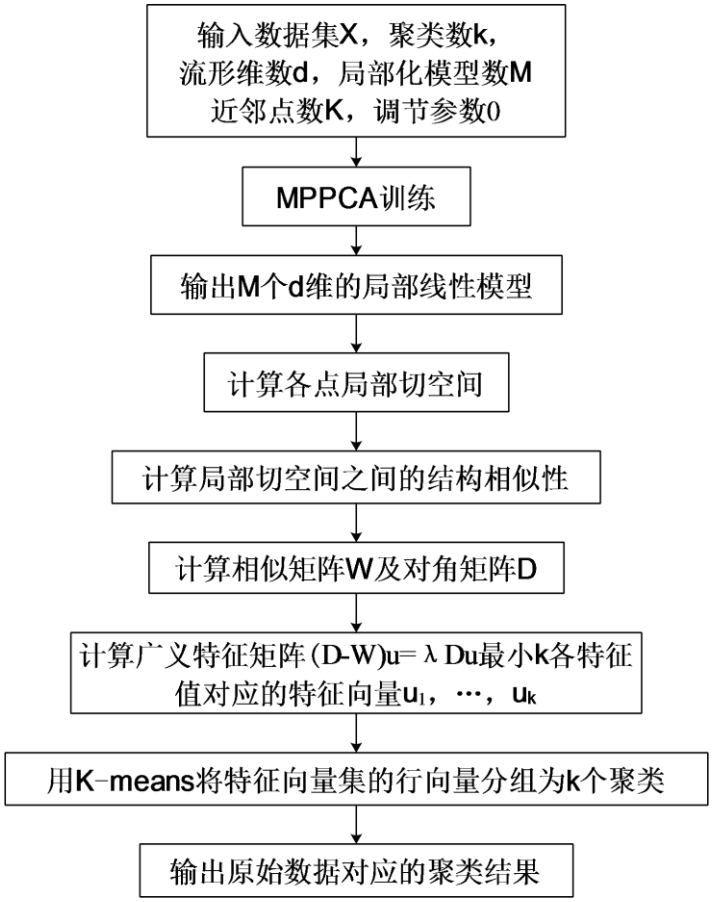
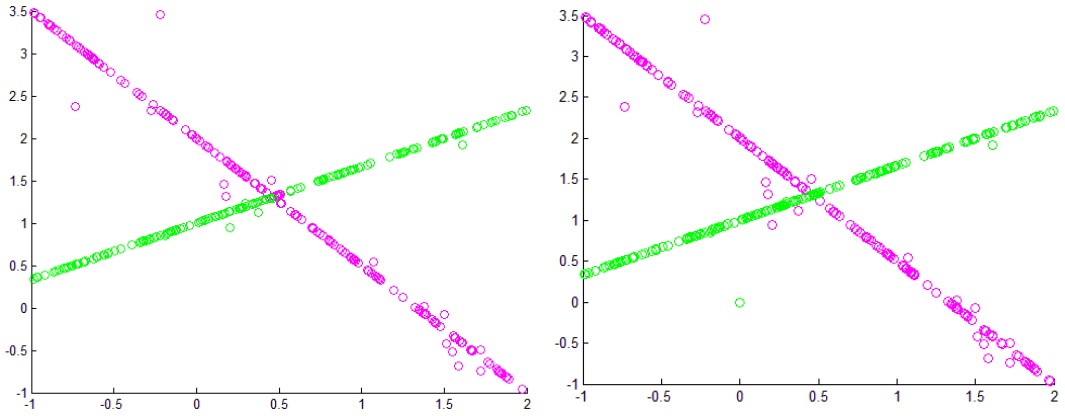


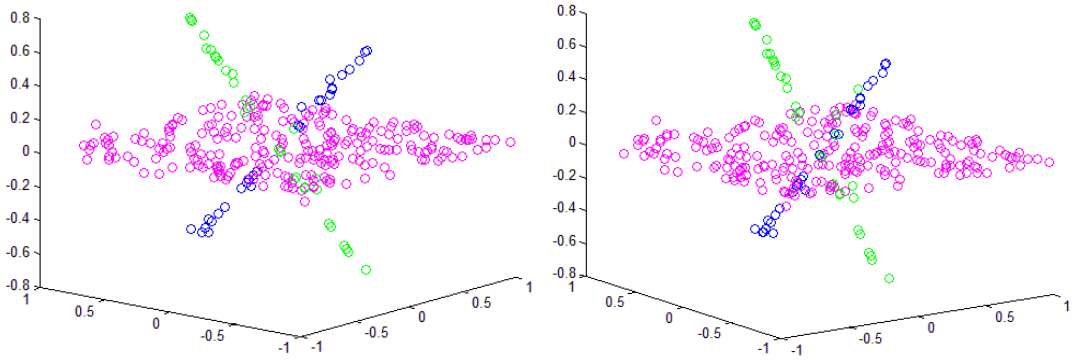
图 5-2 SMMC 模型流程图

SSC 的三种切割方法结果一致，计算结果如下：



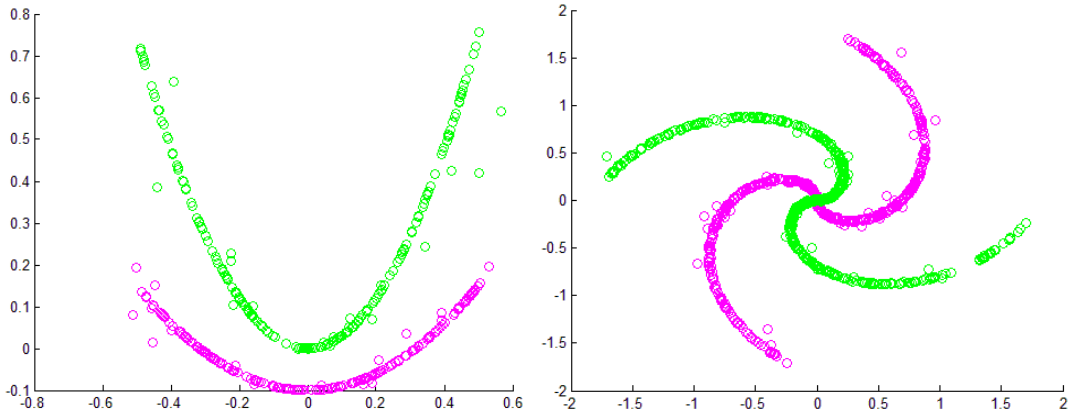
(1) SSC **(2) SMMC**

图 5-3 题 2(a)计算结果



(1) SSC **(2) SMMC**

图 5-4 题 2(b)计算结果



(1) 2(c) **(2) 2(d)**

图 5-5 题 2(c)(d)计算结果

5.4 结果分析

由图 5-3 与图 5-4 所示结果可以看出，对于线性流形问题，SSC 方法与 SMMC 方法所得出的结果基本一致，相较于 SSC 法，SMMC 法运行速度更快，但对参数的选择有一定的要求。

由图 5-5 所示结果，对于非线性流形问题，SMMC 方法所进行的聚类分析结果与观察所得结论基本一致。

SMMC 求解上述四个小题的相关参数选择如下：

表 5-1 SMMC 法的相关参数

参数 / 问题	k	d	M	K	o
2(a)	2	1	4	100	8
2(b)	3	2	9	200	8
2(c)	2	1	40	12	8
2(d)	2	1	60	20	15

为使 SMMC 模型运行稳定，基于上述分析，参数选取的一些推荐设置原则如下：

建议设置 $M=\lceil N/10d \rceil$， $K=2\lceil \log(N) \rceil$， $o=8$。当所有潜在流形都为线性时，可以采用一个较小的 M，例如 $M=3k$。此外，在下述参数范围内寻找最优参数较为合适： $M\in\left[\lceil N/(10d) \rceil,\lceil N/(2d) \rceil\right]$， $K\in\left[\lceil \log(N) \rceil,3\lceil \log(N) \rceil\right]$，$o\in[4,12]$。但对一般的数据集而言，这些参数的最优选取仍依赖于数据的分布和噪声水平等因素。

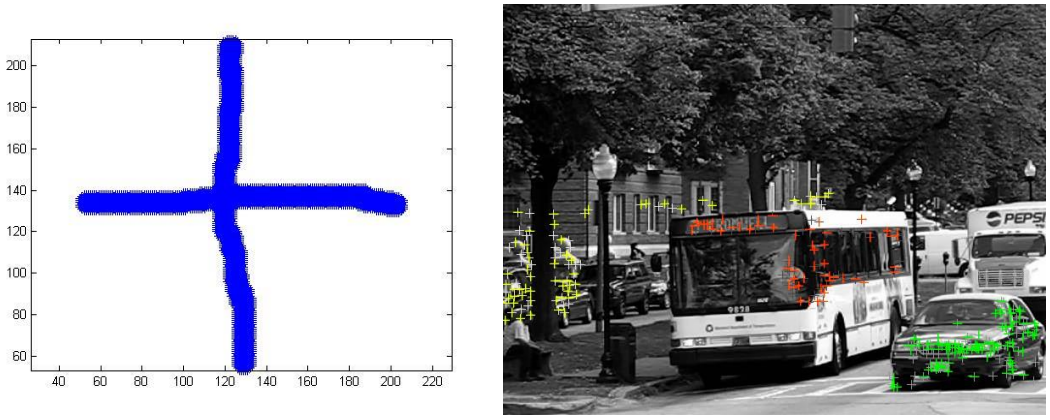
6 问题 3 模型的建立与求解

6.1 问题 3 重述

有以下三个实际应用中的子空间聚类问题，原始数据见附件三：

(1) 视觉重建是一类重要的非接触测量方法，而特征提取是视觉重建的一个关键环节，图 6-1(a)所示的 3a.mat 中数据组成的十字便是特征提取环节中处理得到的，十字上点的位置信息已经提取出来，为确定十字中心位置，将 6-1(a)中十字上的点按照 “横”和“竖”分成两类，并将分类结果画出。

(2) 运动分割是将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或缺的一步。基于特征点轨迹的方法，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，之后将场景中不同运动对应的不同特征点轨迹分割出来。同一运动的特征点轨迹在同一个线性流形上。图 6-1(b) 所示的 3b.mat 中数据显示了视频中的一帧， 3b.mat 中有三个不同运动的特征点轨迹，将这些特征点轨迹分成三类，并制作一个表格输出样本的类别标签，每行 20 个。



(a) **(b)**

图 6-1 实际应用中的子空间聚类

(3) 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），将这 20 幅图像分成两类，并制作一个表格输出样本的类别标签，每行 20 个。

6.2 问题 3 分析

6.2.1 视觉重建中的特征提取

图 6-1(a)为由 2835 个数据点组成的二维图像，虽然从局部来看为非线性流形问题，但从整体上来看而呈线性。针对此类问题，可以运用 K-means 算法将图形分成 k 个核心，以 k 个核心代表所有数据点，达到消除局部非线性特征的目的。

首先用 K-means 将 2835 个数据分为 200 组，形成 200 个聚类核心，然后利用 SSC 模型的三种切割方法进行聚类分析。

在利用 K-means 和 SSC 模型结合的方法分离出“横”“竖”两类后，再利用 SMMC 方法进行模型检验。

6.2.2 运动分割

图 6-1(b)中展示了 3b.mat 中运动轨迹的一帧，由于数据的维数较高，需要进行降维处理，本题的处理方法为分别通过 PCA、Isomap 及 LLE 得出各自的降维结果，利用 K-means 算法进行聚类，即采用基于 PCA、Isomap 及 LLE 的 K-means 算法解决运动分割问题，再利用 SMMC 模型进行检验。

6.2.3 人脸识别

人脸资料属于高维度数据，而高维度数据往往难以叙述与计算，需用一个较低维度的非线性流形（Non-linear Manifold）进行模拟，即需要进行降维处理再加以聚类。问题 3(c)中所涉及的人脸识别问题，由于人脸识别应以人脸特征作为主要考量对象，故应首先消除光照强度影响。本题中对 3c.mat 中每列 2016 个数据利用 matlab 中 Z-score 命令采用标准化处理的方法消除光照影响，接着分别通过 PCA、Isomap 及 LLE 得出各自的降维结果，利用 K-means 算法进行聚类，即标准化数据消除光照影响后，采用基于 PCA、Isomap 及 LLE 的 K-means 算法解决人脸识别问题。计划与 3(b)方法相同再利用 SMMC 进行模型检验，由于 SMMC 模型需利用统计学知识（6.3.3 SMMC 模型建立中有详细叙述），而 3c.mat 中的人脸数据仅有 20 个人脸数据，样本数量偏少，数据缺乏代表性，导致利用 SMMC 方法时所得聚类结果稳定性不高，因而在处理 3c.mat 数据进行

人脸识别时本文未采用 SMMC 模型进行检验。

6.3 模型建立与分析

6.3.1 PCA 模型

PCA 算法（线性算法）用于人脸识别的基本原理是利用 *K-L* 变换（特征向量变换或主分量法）抽取人脸的主要成分，构成特征脸空间，识别时将测试图像投影到此空间，得到一组投影系数，通过与各个人脸图像比较进行识别。这种方法使得压缩前后的均方误差最小，且变换后的低维空间有很好的分辨能力。

3c.mat 数据中共有 2016 个主成分，利用 PCA 提取前五的 5 个主成分，占可表示出原始数据的 98%以上，因而只需要选取前 5 个主成分，既可表示出原始数据的特征。

6.3.2 Isomap 模型

Isomap（非线性算法）是以 MDS(Multidimensional Scaling)为计算工具，在计算高维流形上数据点间距离时，不是用传统的欧式距离，而是采用微分几何中的测地线距离（或称为曲线距离），并且可以用实际输入数据估计其最小路径逼近测地线距离。

(1) Isomap 算法有如下优点：

③ 求解过程依赖于线性代数的特征值和特征向量问题，保证了结果的稳健性和全局最优性；

④ 能通过剩余方差判定隐含的低维嵌入的本质维数；

⑤ 计算过程中只需要确定唯一的一个参数（近邻参数 k 或邻域半径 e）。

(2) 计算步骤如下：

① Isomap 的输入是许多高维度的数据，并把它们当作一个图，只要两个顶点相邻，就会有一个边连接，判定相邻的方法可用 K-Nearest Neighbors（最邻近规则分类）或是用直接距离再取临界值。

② 利用弗洛伊德算法（Floyd’s Algorithm）算出每个顶点之间的最短距离。

③ 将②中所得结果作为 MMDS 的输入，就可得到一个坐标轴，利用该坐标轴描述出的数据就是一个低维度的流形。

Isomap 算法虽简单，但解决了 PCA 或其他线性方法在非线性流形上所遇到的问题，透过邻居（Neighbor）的定义，加强了各个数据之间的连结性，而不是只以绝对距离作为衡量。

6.3.3 LLE 模型

LLE 算法（非线性算法）是针对非线性数据的一种降维方法，是将高维流型用剪刀剪成很多的小块，每一小块可以用平面代替，然后在低维中重新拼合出来， 且能保持各点之间的拓扑关系。将整个问题最后被转化为两个二次规划问题。

(1) LLE 算法的基本步骤：

① 寻找每个样本点的 k 个近邻点；

② 由每个样本点的近邻点计算出该样本点的局部重建权值矩阵；

③ 由该样本点的局部重建权值矩阵和其近邻点计算出该样本点的输出值。

(2) 算法分析：

算法的第一步是计算出每个样本点 X_i 的 k 个近邻点，把相对于所求样本点距离最近的 k 个样本点规定为所求样本点的 k 个近邻点。k 是一个预先给定值。距离的计算既可采用欧氏距离也可采用 Dijkstra 距离。Dijkstra 距离是一种测地距离，能够保持样本点之间的曲面特性。

LLE 算法的第二步是计算出样本点的局部重建权值矩阵。定义一个成本函数(Cost Function)，如式(6-1)所示，以测量重建误差：

$$\varepsilon(W)=\sum_{i}\left|X_i-\sum_{j}W_{ij}\right| \tag{6-1}$$

即全部样本点和他们的重建之间的距离平方和。W_{ij} 表示第 j 个数据点到第 i 个重建点之间的权重。为计算权重 W_{ij}，设置两个限制条件而使成本函数取最小值：

① 每个数据点 X_i 仅从它的邻居那里被重建，如果 X_j 不属于 X_i 的邻居的集合，则 $W_{ij}=0$；

② 矩阵中每行的权重和为 1：即 $\sum_{j}W_{ij}=1$。

为使重建误差最小化，W_{ij}服从一种重要的对称性，即对所有特定数据点来说，它们和邻居点之间经过旋转、重排、转换等变换后，之间的对称性是不变的。由此可见重建权重能够描述每个邻居本质的几何特性。因此可以认为原始数据空间内的局部几何特征同在流形局部块上的几何特征是完全等效的。

LLE 算法的最后一步是将所有的样本点X_i映射到在流形中表示内部全局坐标的低维向量Y_j上。映射条件满足如下成本函数，

$$\phi(Y)=\sum_i\left|Y_j-\sum_j W_{ij}Y_j\right|^2 \tag{6-2}$$

其中：$\phi(Y)$为成本函数值；Y_j为X_i的输出向量；

Y_j是 Y_i的k个近邻点，且满足两个条件，即：

$$\sum Y_i=0(i=1,2,\ldots,N) \tag{6-3}$$

$$\sum Y_iY_i^T=I(i=1,2,\ldots,N) \tag{6-4}$$

其中：I是$m\times m$的单位矩阵。

要使成本函数值达到最小，则取Y_j为M（流形）的最小 m 个非零特征值所对应的特征向量。在处理过程中，将M的特征值从小到大排列，第一个特征值接近于零，则舍去第一个特征值。通常取第 2~m+1 之间的特征值所对应的特征向量作为输出结果。

LLE 主要是一种局部方法，它试图保持数据的局部几何特征，就本质上而言，它是将流形上的近邻点映射到低维空间的近邻点，而 Isomap 是一种全局方法，它试图保持整个数据的几何特征，将流形上的近邻点映射到低维空间的近邻点，将流形上的远点映射到低维空间的远点。

6.3.4 问题 3 求解

(1) 问题 3(a)求解

问题 3(a)的解法流程图如下：

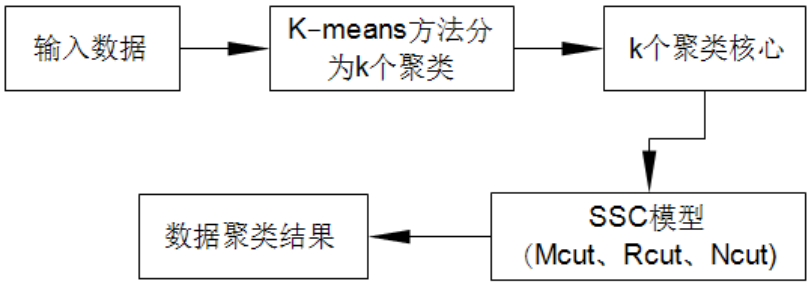
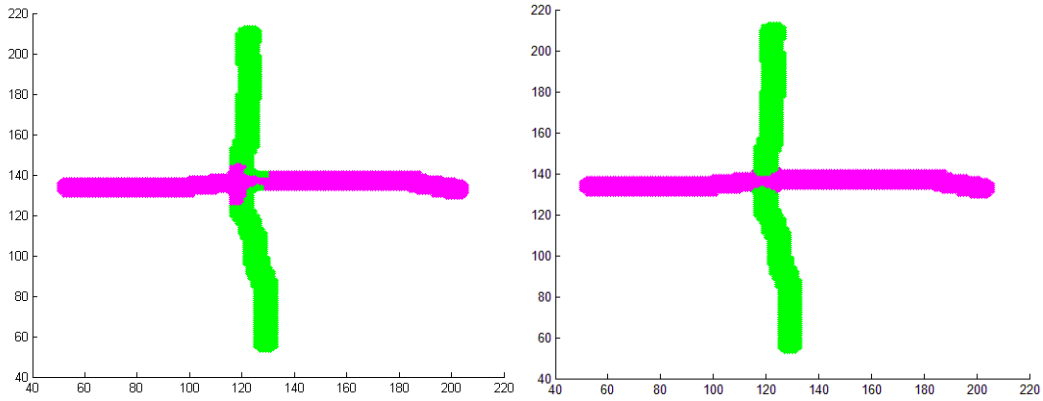


图 6-2 问题 3(a)解法流程图

SSC 的三种切割方法结果一致，计算所得结果为：



(1) 基于 K-means 的 SSC 模型 **(2) SMMC**

图 6-3 题 3(a)计算结果

SMMC 的参数分别为 k=2，d=1，M=16，K=150，o=8。

(2) 问题 3(b)求解

问题 3 (b)将 3b.mat 中的数据分别利用 PCA、Isomap、LLE 和 SMMC 模型进行聚类计算，将三个运动轨迹分别记为 1、2、3 三类，Isomap 模型流程图如图 6-4 所示：

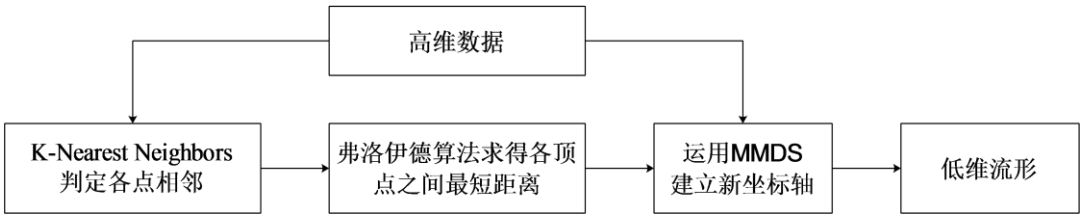


图 6-4 Isomap 模型流程图

LLE 模型流程图如图 6-5 所示：

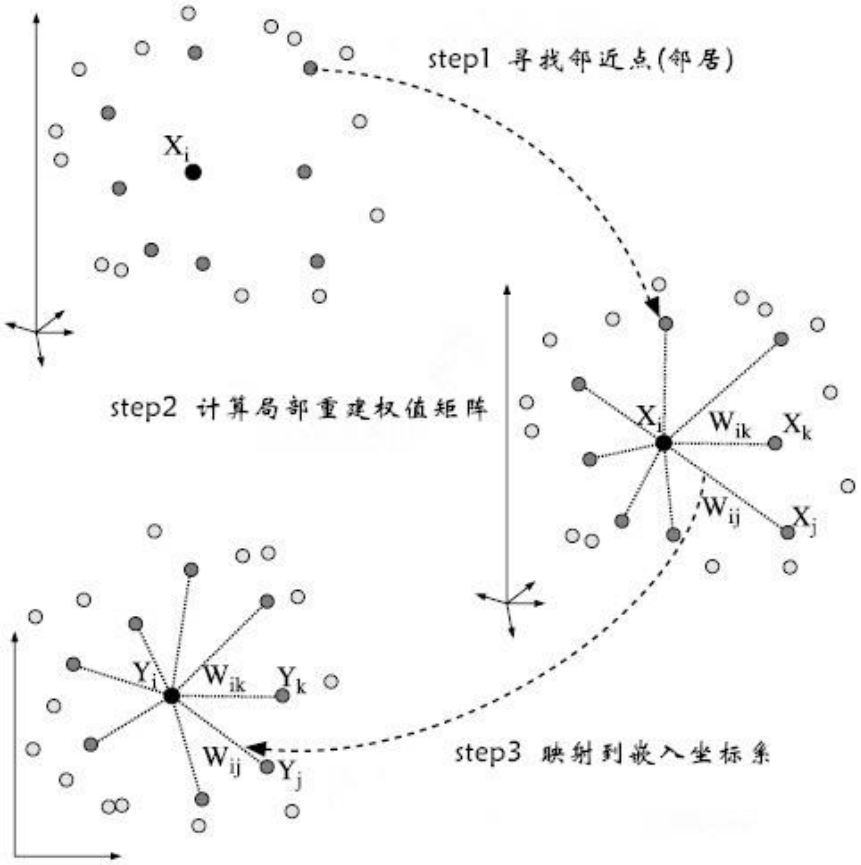


图 6-5　LLE 算法步骤图

将 297 个数据按列从小到大每行排 20 个（即第一行对应的第 1~20 列），求解结果用数字 1、2、3 分别作为各类的类别标签，对结果进行列表表示。

计算结果如下：

① 用 PCA 模型降维后 K-means 法聚类结果：

表 6-1 PCA 模型聚类结果

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	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	2	3	2	3	3	3	3	2	3	2	3
3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	3	3
3	3	3	2	3	3	2	2	3	2	3	2	3	3	3	3	3	3	3	3
3	2	3	3	3	3	2	3	3	3	3	3	3	3	3	3	2			

② 用 Isomap 模型降维后 K-means 法聚类结果：

表 6-2 Isomap 模型聚类结果

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	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	2	3	2	3	3	3	3	2	3	2	3
3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	3	3
3	3	3	2	3	3	2	2	3	2	3	2	3	3	3	3	3	3	3	3
3	2	3	3	3	3	2	3	3	3	3	3	3	3	3	3	2			

③ 用 LLE 模型降维后 K-means 法聚类结果：

表 6-3　LLE 模型聚类结果

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	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	2	3	2	3	3	3	3	2	3	2	3
2	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	3	3
3	3	3	2	3	3	2	2	3	2	3	2	3	3	2	2	3	3	3	3
2	2	3	3	3	3	2	3	3	3	2	3	3	3	3	3	2			

④ SMMC 模型聚类结果（参数分别为 k=3，d=4，M=20，K=15，o=12）：

表 6-4　SMMC 模型聚类结果

1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2
2	1	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2
1	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	1
2	2	2	2	2	1	2	2	2	2	2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	2	3	2	3	3	3	3	2	3	2	3
3	3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	3	3
3	3	3	1	3	3	2	2	3	2	3	2	3	3	2	2	3	3	3	3
3	1	3	3	3	3	2	3	3	3	3	3	3	3	3	3	2			

(3) 问题 3(c)求解

将 20 个数据按列从小到大每行排 20 个（即第一行对应的第 1~20 列），求解结果用数字 1、2 分别作为各类的类别标签，对结果进行列表表示。

方法同 3(b)，分别利用 PCA、Isomap 和 LLE 模型进行聚类计算，所得结果如下（三种方法所得结果一致）：

表 6-5　3(c)聚类结果

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

6.4 结果分析

问题 3(a)采用的 SSC 与 SMMC 进行聚类的结果大体一致。

问题 3(b)分别采用了 PCA、Isomap、LLE 和 SMMC，后三类模型的聚类结果与 PCA 相比误差分别为 0，1.7%（5 个样本不同）与 3.7%（11 个样本不同）。

问题 3(c)分别采用了 PCA、Isomap 和 LLE 模型，所得结果完全一致。从所画出的人脸模型灰度图（图 6-6）可以看出，聚类结果合理有效。

图 6-6 人脸模型灰度图

由此可得，上述模型均合理，具有一般性。

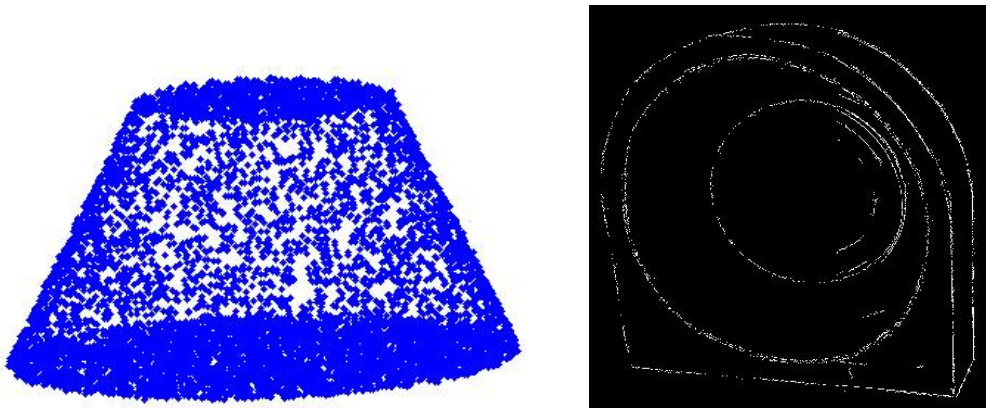
7 问题 4 模型的建立与求解

7.1 问题 4 重述

附件 4 中所给数据反映了如下两个实际应用中的多流形聚类问题。

图 7-1(a)显示了圆台的点云，将点按照圆台的顶、底、侧面分成三类，并画出分类结果。

图 7-1(b)是机器工件外部边缘轮廓图像，将轮廓线中不同的直线和圆弧分类，自定类数，并画出分类结果。



(a) **(b)**

图 7-1 实际应用中的多流行聚类问题

7.2 问题 4 分析

针对图 7-1 (a)云台的点云与图 7-1(b)机器工件外部轮廓图像的非线性问题，采用与问题 2(c)与问题 2(d)相同 SMMC 模型来实现混合流形聚类。

7.3 问题 4 求解

SMMC 模型的基本思想是：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。SMMC 模型流程图如下：

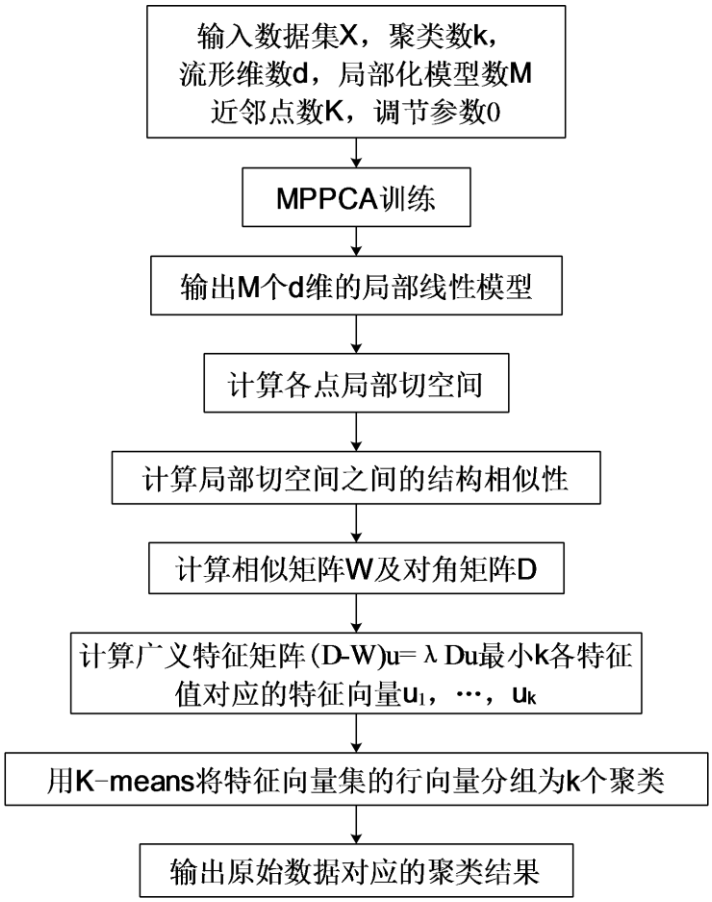


图 7-2 SMMC 模型流程图

7.3.1 问题 4(a)求解

问题 4(a)运用 SMMC 模型聚类结果如下（SMMC 的参数分别为 k=3，d=2，M=120，K=20，o=15）：

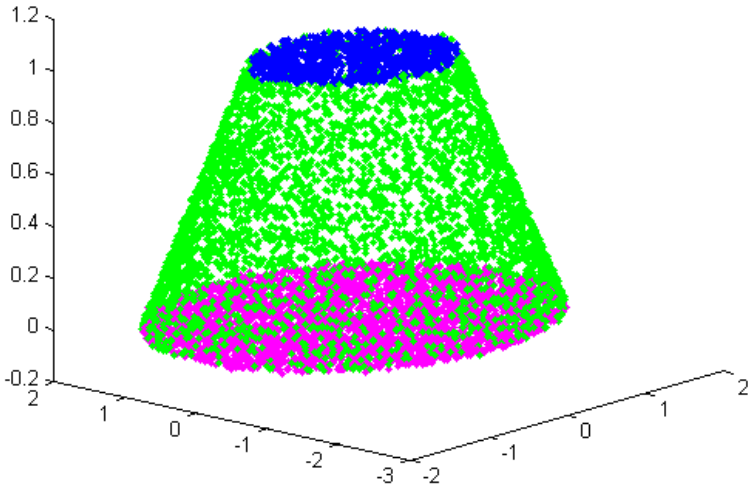


图 7-3　4(a)聚类结果

7.3.2 问题 4(b)求解

问题 4(b)首先运用 SMMC 模型聚类结果如下（SMMC 的参数分别为 k=4，d=1，M=100，K=24，o=4）：

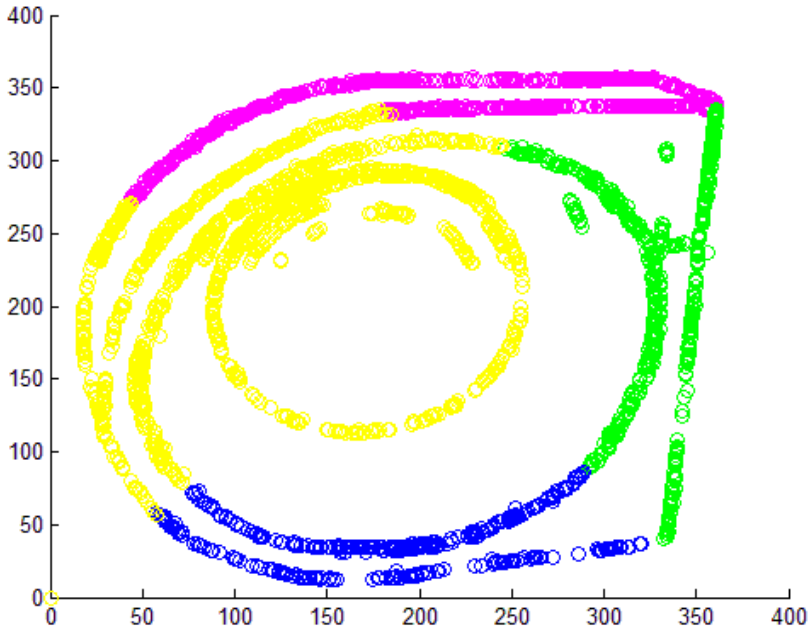


图 7-4　4(b)-减噪前 SMMC 聚类结果

直观上可见图 7-4 的聚类效果并不理想，分析可能原因是受噪声影响，故本文运用 K-means 算法人为地消除了一些噪声点后，减噪后的运行结果如下：

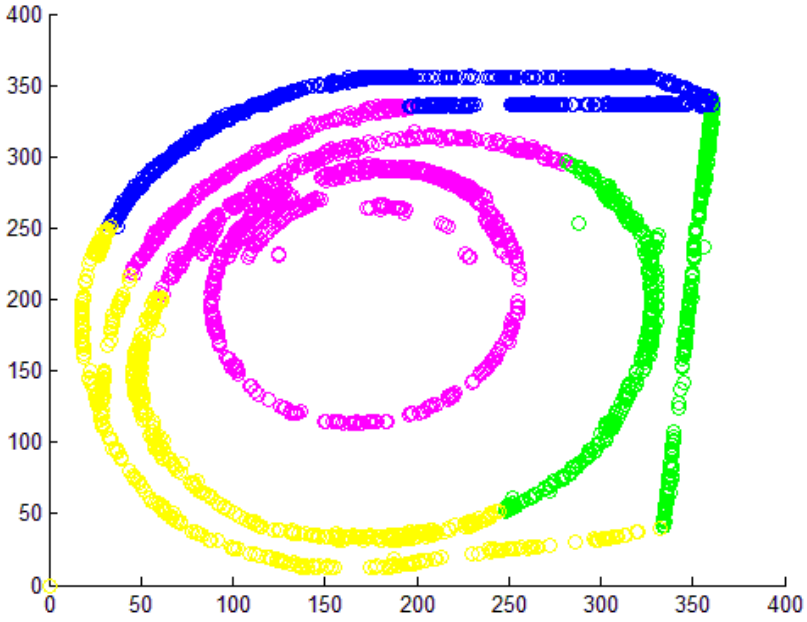


图 7-5　4(b) -减噪后 SMMC 聚类结果

利用解析几何原理人为地进行聚类，得到直观上应得到的聚类结果，解析几何原理所处理的结果如下所示：

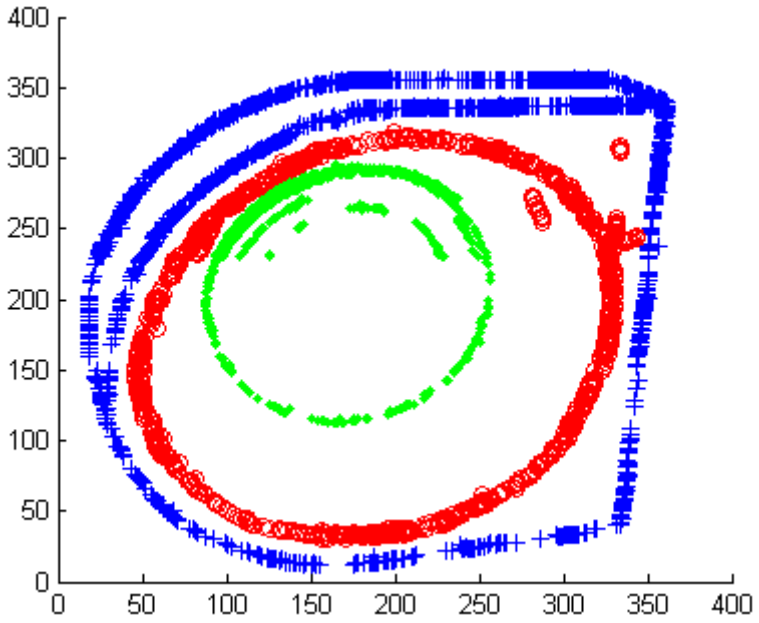


图 7-6　4(b) –解析几何原理聚类结果

7.4 结果分析

问题 4(a)的结果从直观上将点按照其所在的面（即按照圆台的顶、底、侧面）分成了三类，符合题意要求，模型合理，且具有一般性。

问题 4(b)的 SMMC 模型结果将机器工件外部轮廓按侧面平面与底面圆弧面分开，达到了一定的聚类效果，但局部区域不同的非线性流形，导致聚类效果不理想，与解析几何原理直观上所得结果有一定误差。

分析其主要原因为：(1) 同一流形上的点相似度太低；(2) 不同流形上的点距离太近，且局部切空间过多地重叠。故而需要进行深度学习，将 SMMC 模型进一步改进优化。

8 模型评价与推广

8.1 模型评价

由于时间仓促，本文在模型建立和数据处理上，还存在一些问题。为此，本文有如下几点可进行优化改进：

(1) SMMC 模型在运行时需要对参数进行事先调试，从而使其运行稳定，这是其不足之处。

(2) 针对问题 3(b) 的运动分割，PCA 模型、Isomap 模型、LL 模型、SMMC 模型这几种模型间还存在一定误差，需改进模型，使误差进一步减小。

(3) 针对问题 3(c) 的人脸识别，由于数据样本数较小，代表性较差，对 SMMC 模型的稳定性影响较大，聚类效果不太理想，未能采用 SMMC 模型进行求解校验。

(4) 针对问题 4(b) 的工件外部边缘轮廓图像聚类，由于数据噪声过大且缺乏明确的流形结构特征，SMMC 模型只能做大致的聚类分析，还需对模型进一步优化改进。

8.2 模型推广

本文建立的与多流形数据结构相关的人脸识别模型、图像分类模型、运动分割模型和计算机视觉重建模型等，对高维数据的相关性分析、聚类分析和结构分析的研究领域提供了解决思路，有一定的参考价值。

9 结论

9.1 结果

问题 1、问题 2(a)2(b)与问题 3(a)均属于线性流形聚类问题，用 SSC 模型都能得到很好的聚类结果；问题 2(c)、2(d)、4(a)、4(b)均属于非线性流形聚类问题，用 SMMC 得到了有效结果；同时，问题 3(b)、3(c)属于高维度流形降维聚类问题，用 PCA、Isomap、LLE 三种模型降维后进行聚类，并进行比较，相互间误差较小，得到了很好的结果。

具体结果如下：

问题 1：得到第 41~140 个数据属于类别 1、其余编号数据属于类别 2 的结果。

问题 2：有效地将 2(a)的两条交点不在原点且互相垂直的直线分为两类；将 2(b)的一个平面和两条直线，分为三类；将 2(c)的两条不相交的二次曲线分为两类；将图 2(d) 为两条相交的螺旋线分为两类。

问题 3：有效地将 3(a)中十字上的点分成两类；针对 3(b)运动分割问题，有效地将视频中一帧的特征点轨迹分成三类；针对 3(c)人脸识别问题，成功将请将这 20 幅人脸图像分成两类。

问题 4：有效地将 4(a) 圆台的点云，按照其所在的面分为（即圆台按照圆台的顶、底、侧面）三类；将 4(b)机器工件外部边缘轮廓线中不同的直线和圆弧进行了聚类，分成了四类。

9.2 模型优缺点

SSC 模型能有效地解决线性流形聚类问题，且适用性强；

SMMC 模型不仅能够有效地解决线性流形聚类问题，而且对于非线性流形聚类问题也能够进行很好的处理，且计算速度较快，运行时间短。但 SMMC 模型需要事先对所需的参数进行估计、调试，才能保证程序的稳定性。且对样本数有一定要求。

PCA、Isomap、LLE 三种降维模型都能高效地对高维流形进行降维处理，且降维后的聚类结果误差较小，适用性较强。

10 参考文献

[1] 王卫卫，李小平，冯象初，王斯琪. 稀疏子空间聚类综述[J]. 自动化学报，2015，08:1373-1384.

[2] 陈黎飞. 高维数据的聚类方法研究与应用[D].厦门大学，2008.

[3] 张肇炽. 关于用初等变换求向量组的极大无关组[J]. 高等数学研究，2003，04:18-21.

[4] 陈新宁. 论极大线性无关组[J]. 甘肃联合大学学报(自然科学版)，2009，S1:30-31+36.

[5] 郭鑫. 图像理解中的稀疏与低秩[D].北京邮电大学,2014.

[6] 夏建明,杨俊安. 基于稀疏流形聚类嵌入模型和 L_1 范数正则化的标签错误检测[J]. 控制与决策,2014,06:1103-1108.

[7] 冯超. K-means 聚类算法的研究[D].大连理工大学,2007.

[8] 冯晓蒲,张铁峰. 四种聚类方法之比较[J]. 微型机与应用,2010,16:1-3.

[9] 王勇. 基于流形学习的分类与聚类方法及其应用研究[D].国防科学技术大学,2011.

[10] Y. Wang, Y. Jiang, Y. Wu, and Z. Zhou. Spectral clustering on multiple manifolds. IEEE Transactions on Neural Networks, 22(7):1149–1161, 2011

11 附录

在本文中，主要利用 Matlab 进行建模与算法实现，模型的选取涉及 PCA 模型、SSC 模型、SMMC 模型、Isomap 模型和 LLE 模型等，算法以 K-means 为基础。由于计算机程序和代码数量多，所使用的命令和编写的计算机源程序不在论文中一一录入，将整理成附件上传。