文章编号: 1001-0920(2011)01-0080-05

单位超球面上的二元聚类算法

程 成, 萧蕴诗, 岳继光

(同济大学 电子与信息工程学院，上海 201804)

摘 要: 鉴于在数据学习理论中, 基于各种几何正则化的处理方法引起人们的广泛关注, 以基于独立成分分析的地震数据处理为背景, 针对数据向量的单位模长约束, 研究了单位超球面上的二元聚类问题. 通过欧氏诱导度量, 推导了单位超球面上的黎曼梯度公式, 并据此构造了求取其上二元聚类和数据平均的不动点迭代算法. 实验结果表明了其有效性和优越性.

关键词: 单位超球面；聚类；黎曼梯度；测地距离

中图分类号: TN911；O186 **文献标识码:** A

Clustering algorithm for two classes on unit Hypersphere

CHENG Cheng, *XIAO Yun-shi*, *YUE Ji-guang*

(School of Electronics and Information Engineering，Tongji University，Shanghai 201804，China．Correspondent: CHENG Cheng，E-mail：chengcheng_lcc@163.com)

Abstract: The methods of geometrical regularization in data learning theory have caused wide public concern. In the background of seismic data processing based on independent component analysis, a problem of two-class clustering on unit hypersphere is studied for the unit norm constraint. The Riemannian gradient is formulated based on the induced metric from Euclidean space, which realizes the construction of a fixed-point algorithm for two-class clustering and data averaging on unit hypersphere. Finally, the simulation result shows the effectiveness and superiority of this algorithm.

Key words: unit hypersphere；clustering；riemannian gradient；geodesic distance

1 引 言

自经典的流形学习算法局部线性嵌入 (LLE)[1]和等距映射 (ISOMAP)[2]提出以来, 基于几何正则化的数据处理方法开始引起人们的广泛关注[3]. 流形学习算法以局部度量代替数据空间的整体欧氏度量, 适应了数据空间弯曲的内蕴特征. 在流形上, 连接任意两点间的最短路径称为测地线, 而其长度称为测地距离, 基于测地距离的数据聚类算法是通常 K-means 算法在弯曲流形上的自然推广. Goh 等人[4]通过二次均方重整化对参数空间赋予了特殊的度量结构, 并据此研究了概率密度函数簇在黎曼流形上的聚类问题, 这为数据空间的概率密度估计和统计回归分析提供了一种新的方法, 并将此方法应用于图像分割, 取得了满意的结果[5]. 文献 [6] 提出了一种称为软测地核 K-means 聚类的算法, 该方法利用测地距离构造核函数, 保证了数据集在特征空间的线性可分性, 也克服了通常方法中对核函数选取的随意性.

在数据挖掘中, 相对于主成分分析 (PCA), 基于独立成分分析 (ICA) 的特征提取能够获得更有意义的结果[7-8]. 本文基于 ICA 研究了地震数据处理中的子波提取问题. 在提取过程中, 为了保证算法的稳定性, 对特征向量, 即地震子波施加的单位范数约束条件赋予了特征空间一个自然的单位超球面几何结构. 又由于 ICA 中符号的不确定性, 使得提取出的特征向量呈现出近似对称的二元分布特点. 对此, 本文通过构造黎曼度量, 研究了单位超球面上的二元聚类分析, 并提出了相应的不动点迭代算法. 在特征维数上, 单位超球面与其嵌入的欧氏空间相差 1, 因此这种度量适应了空间弯曲的结构特征. 鉴于在子波提取中数据向量近似对称的分布特点, 将聚类结果中的一类数据反号, 就可以自然地得到求取这些数据向量单位超球面平均的算法. 该方法不仅能够平滑数据中的不规则点, 还可以消除测量系统中的随机影响. 最后, 将该二元聚类分析和数据平均的方法应用于多道模拟地震

收稿日期: 2009-10-29；**修回日期:** 2010-02-02.

基金项目: 国家自然科学基金项目(40872090).

作者简介: 程成(1980−), 男, 博士生, 从事盲源分离、机器学习等研究；萧蕴诗(1946−), 男, 教授, 博士生导师, 从事信号处理、智能控制理论等研究.

数据的子波提取, 经比较发现其结果明显优于基于欧氏度量的 K-means 聚类所得的结果.

2　问题描述

假设欧氏空间 $\boldsymbol{R}^p$ 中一数据点集记为 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}$, K-means 聚类即是将这 N 个数据点分别划归到 K 个不同的类别中, 其中类别数 K 是预先已知的, 而每个类别分别以聚类中心 $\boldsymbol{c}_k$ 标记. 任意数据点 $\boldsymbol{x}_n$ 与各聚类中心 $\boldsymbol{c}_k$ 间距离的大小作为其划归相应类别的依据. 在 $\boldsymbol{R}^p$ 中, K-means 聚类问题可以描述为如下优化问题:

$$\arg\min_{r_{nk},\boldsymbol{c}_k} J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\boldsymbol{x}_n - \boldsymbol{c}_k\|^2, \tag{1}$$

其中

$$r_{nk} = \begin{cases} 1, \ k = \arg\min\limits_j \|\boldsymbol{x}_n - \boldsymbol{c}_j\|^2; \\ 0, \ \text{otherwise}. \end{cases}$$

r_{nk} 是判断数据点类别属性的二元指标, 目标函数 J 的建立依赖于 $\boldsymbol{R}^p$ 中由 2- 范数结构诱导的直线距离.

流形学习的结果对数据点的度量结构具有明显的依赖关系, 因此对于特殊的数据流形, 必须考虑与其相应的几何与度量结构. 假设 d 就是满足这种结构的二元泛函, 则相应的聚类问题可以描述为

$$\arg\min_{r_{nk},\boldsymbol{c}_k} J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} d(\boldsymbol{x}_n, \boldsymbol{c}_k), \tag{2}$$

其中

$$r_{nk} = \begin{cases} 1, \ k = \arg\min\limits_j d(\boldsymbol{x}_n, \boldsymbol{c}_k); \\ 0, \ \text{otherwise}. \end{cases}$$

当 $K = 2$, 且所有参数向量具有单位模长时, 问题 (2) 就是本文所要研究的单位超球面上的二元聚类问题.

以下通过研究单位超球面上的几何结构, 建立一种自然的度量关系, 并且用于数据二元聚类和求取平均的研究.

3　单位超球面上的几何结构

3.1　数学定义

欧氏空间 $\boldsymbol{R}^p$ 中的单位超球面可以表示为

$$\boldsymbol{S}^{p-1} = \{\boldsymbol{x} \in \boldsymbol{R}^p | \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} = 1\}.$$

利用 $\boldsymbol{R}^p$ 中的标准内积, 任意 $\boldsymbol{x} \in \boldsymbol{S}^{p-1}$ 处的仿射空间可以正交分解为切空间和法空间的直和, 分别为

$$T_{\boldsymbol{x}}\boldsymbol{S}^{p-1} = \{\boldsymbol{v} \in \boldsymbol{R}^p | \boldsymbol{v}^{\mathrm{T}}\boldsymbol{x} = 0\}, \ N_{\boldsymbol{x}}\boldsymbol{S}^{p-1} = \{\alpha\boldsymbol{x} | \alpha \in \boldsymbol{R}\}.$$

作为 $\boldsymbol{R}^p$ 中的一个嵌入子流形, 可以对 $\boldsymbol{x}$ 处的切空间 $T_{\boldsymbol{x}}\boldsymbol{S}^{p-1}$ 赋予一个诱导度量, 使其局部欧氏化. 该诱导度量为

$$\langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle_{\boldsymbol{x}}^{\boldsymbol{S}^{p-1}} = \langle \boldsymbol{v}_1, \boldsymbol{v}_2 \rangle_{\boldsymbol{x}}^{\boldsymbol{R}^p} = \boldsymbol{v}_1^{\mathrm{T}}\boldsymbol{v}_2. \tag{3}$$

3.2　黎曼梯度

设 f 是 $\boldsymbol{S}^{p-1}$ 上的一个光滑函数. 作为嵌入子流形, f 也可以看作是 $\boldsymbol{R}^p$ 中的函数. 根据微积分理论, f 在 $\boldsymbol{R}^p$ 中的梯度向量 ∂f 指向函数增长最快的方向. 由于 $\boldsymbol{S}^{p-1}$ 是一个约束空间, 可以忽略 f 在法向上的变化而只考虑其在切空间中的梯度分量, 这就引出了黎曼梯度的概念.

根据诱导度量 (3), f 所对应的黎曼梯度 $\nabla f \in T_{\boldsymbol{x}}\boldsymbol{S}^{p-1}$ 定义为满足如下约束的切向量:

$$\langle \boldsymbol{v}, \nabla f \rangle_{\boldsymbol{x}}^{\boldsymbol{S}^{p-1}} = \langle \boldsymbol{v}, \partial f \rangle_{\boldsymbol{x}}^{\boldsymbol{R}^p} = \boldsymbol{v}^{\mathrm{T}}\partial f. \tag{4}$$

其中: ∂f 是 $\boldsymbol{R}^p$ 中的常规梯度, $\boldsymbol{v}$ 是 $T_{\boldsymbol{x}}\boldsymbol{S}^{p-1}$ 中任意一个切向量.

由于 $T_{\boldsymbol{x}}\boldsymbol{S}^{p-1}$ 被赋予与 $\boldsymbol{R}^p$ 相同的内积形式, 由式 (4) 可得 $\langle \boldsymbol{v}, \nabla f - \partial f \rangle = 0$, 即存在 $\alpha \in \boldsymbol{R}$ 使得

$$\nabla f - \partial f = \alpha\boldsymbol{x}. \tag{5}$$

进一步计算可知 $\alpha = -\langle \boldsymbol{x}, \partial f \rangle = -\boldsymbol{x}^{\mathrm{T}}\partial f$. 将该结果带入式 (5) 即可得到 f 在 $\boldsymbol{x}$ 点处的黎曼梯度公式

$$\nabla f = (\boldsymbol{I}_p - \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}})\partial f, \tag{6}$$

其中 $\boldsymbol{I}_p$ 代表 p 阶单位矩阵.

根据诱导度量, 可以证明式 (6) 中的黎曼梯度 ∇f 就是常规梯度 ∂f 在切空间 $T_{\boldsymbol{x}}\boldsymbol{S}^{p-1}$ 中的投影向量. 以球面 $\boldsymbol{S}^2$ 为例, 图 1 给出了各类梯度的图示说明, 其中 $\nabla^{\perp} f = \partial f - \nabla f$ 是法向梯度.

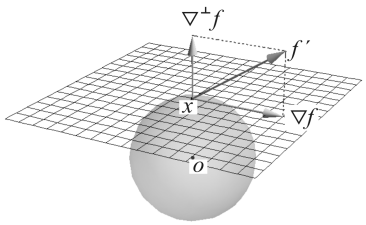


图 1　二维球面上的各种梯度向量

3.3　单位超球面上的测地距离

由文献 [9] 可知, 被赋予诱导度量后, $\boldsymbol{S}^{p-1}$ 上的测地线就是其表面的大圆弧. 通过将测地线看成质点在 $\boldsymbol{S}^{p-1}$ 上运动的轨迹可以给出其动力学解释. 假设质点以 $\boldsymbol{x}$ 为起点, 以 $\boldsymbol{v} \in T_{\boldsymbol{x}}\boldsymbol{S}^{p-1}$ 为初始速度运动, 则其运动轨迹就是由这两个向量所决定的单位圆周, 测地线方程为

$$g(s) = \cos(s)\boldsymbol{x} + \sin(s)\boldsymbol{v}. \tag{7}$$

其中: $\boldsymbol{x}$ 和 $\boldsymbol{v}$ 均是单位向量; s 是弧长参数, 由于圆周具有单位半径, s 也是弧度参数. 此处限定质点具有单位速度并非本质的约束, 因为任何曲线通过弧长参数化都具有单位速度, 并且经过参数替换, 弧长的大小是不会发生改变的[9]. 进一步计算可知, 该测地线方程可以看成是质点在 $\boldsymbol{S}^{p-1}$ 上做匀速圆周运动的轨迹.

以下考虑 $\boldsymbol{S}^{p-1}$ 上任意两点 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 间的测地距离 $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$. 设测地线 $g(s)$ 满足 $g(0) = \boldsymbol{x}_1, g(S) = \boldsymbol{x}_2$.

由于曲线已经被弧长参数化, 连接该两点间的测地距离即为 $d(\boldsymbol{x}_1,\boldsymbol{x}_2)=S$. 代入边界条件及单位模长约束可得 $\cos(S)=\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2$. 由于大圆的优弧与劣弧都满足测地线方程, 显然应该选取劣弧的长度作为测地距离, 而反三角余弦函数能够自动取值于劣弧范围. 经计算, $\boldsymbol{S}^{p-1}$ 上任意两点间的测地距离为

$$d(\boldsymbol{x}_1,\boldsymbol{x}_2)=\arccos(\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2).$$

可以发现以上结果同 $\boldsymbol{S}^1$ 的情形是一致的.

4 单位超球面上的二元聚类分析

根据式 (2) 对于广义聚类问题的描述, 选取 $\boldsymbol{S}^{p-1}$ 上的测地距离来衡量两个数据点间的相似程度, 故此时二元聚类问题可以表示为

$$\arg\min_{r_{nk},\,\boldsymbol{c}_k} J=\sum_{n=1}^{N}\sum_{k=1}^{2} r_{nk}\arccos(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_k). \tag{8}$$

其中

$$r_{nk}=\begin{cases}1,\ k=\arg\min\limits_{j}\arccos(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_j);\\ 0,\ \text{otherwise.}\end{cases}$$

$$\boldsymbol{x}_n,\boldsymbol{c}_k\in\boldsymbol{S}^{p-1}.$$

对于优化问题 (8), 可以通过两步迭代的方法进行求解. 首先注意到 r_{nk} 是一个二元变量, 通过比较的方法确定, 而 $\boldsymbol{c}_k$ 可以通过连续优化求解. 假设当前迭代中 r_{nk} 已确定, 转而考虑优化问题

$$\arg\min_{\boldsymbol{c}_k} J=\sum_{n=1}^{N}\sum_{k=1}^{2} r_{nk}\arccos(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_k). \tag{9}$$

由约束优化理论可知, 在最优点处目标函数的常规梯度应平行于数据变量, 即此刻黎曼梯度为零. 计算可得目标函数 J 关于 $\boldsymbol{c}_k$ 的常规梯度向量为

$$\partial_{\boldsymbol{c}_k}J=-\sum_{n=1}^{N}\frac{r_{nk}\boldsymbol{x}_n}{\sqrt{1-(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_k)^2}}. \tag{10}$$

代入黎曼梯度公式 (6) 可得

$$\nabla_{\boldsymbol{c}_k}J=-\sum_{n=1}^{N}\frac{r_{nk}(\boldsymbol{I}_p-\boldsymbol{c}_k\boldsymbol{c}_k^{\mathrm{T}})\boldsymbol{x}_n}{\sqrt{1-(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_k)^2}}. \tag{11}$$

将式 (11) 置零并整理可得

$$\boldsymbol{c}_k\sum_{n=1}^{N}\frac{r_{nk}\boldsymbol{c}_k^{\mathrm{T}}\boldsymbol{x}_n}{\sqrt{1-(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_k)^2}}=\sum_{n=1}^{N}\frac{r_{nk}\boldsymbol{x}_n}{\sqrt{1-(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_k)^2}}. \tag{12}$$

以上给出了求取 $\boldsymbol{c}_k$ 的不动点迭代格式. 注意到式 (12) 左端仅是 $\boldsymbol{c}_k$ 与一个比例因子的乘积, 因此在迭代过程中可以仅考虑该比例因子的符号而忽略其数值大小, 最后通过范数归一化达到更新 $\boldsymbol{c}_k$ 的目的. 由式 (12) 也可以发现, 正是黎曼梯度的引入才使得不动点迭代成为可能.

总结以上优化过程可以得到如下单位超球面 $\boldsymbol{S}^{p-1}$ 上的二元聚类算法:

Step 1: 选取初始聚类中心 $\boldsymbol{c}_k\in\boldsymbol{S}^{p-1}$, $k=1,2$.

Step 2: 确定每个数据点的类别属性指标

$$r_{nk}=\begin{cases}1,\ k=\arg\min\limits_{j}\arccos(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_j);\\ 0,\ \text{otherwise.}\end{cases}$$

Step 3: 更新聚类中心

$$\boldsymbol{c}_k\leftarrow\frac{\sum\limits_{n=1}^{N}\frac{r_{nk}\boldsymbol{x}_n}{\sqrt{1-(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_k)^2}}}{\operatorname{sign}\Big(\sum\limits_{n=1}^{N}\frac{r_{nk}\boldsymbol{c}_k^{\mathrm{T}}\boldsymbol{x}_n}{\sqrt{1-(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c}_k)^2}}\Big)}.$$

Step 4: 对 $\boldsymbol{c}_k$ 进行单位化 $\boldsymbol{c}_k\leftarrow\boldsymbol{c}_k/\|\boldsymbol{c}_k\|$.

Step 5: 如果目标函数 J 收敛则停止迭代; 否则返回 Step 2 继续以上过程.

当限定数据类别数 $K=1$ 时, 对于每个数据点 $\boldsymbol{x}_n$ 有 $r_n\equiv 1$, 这就得到如下求取数据点平均的算法:

Step 1: 选取初始向量 $\boldsymbol{c}\in\boldsymbol{S}^{p-1}$.

Step 2: 更新数据中心

$$\boldsymbol{c}\leftarrow\frac{\sum\limits_{n=1}^{N}\frac{\boldsymbol{x}_n}{\sqrt{1-(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c})^2}}}{\operatorname{sign}\Big(\sum\limits_{n=1}^{N}\frac{\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_n}{\sqrt{1-(\boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{c})^2}}\Big)}.$$

Step 3: 对向量单位化 $\boldsymbol{c}\leftarrow\boldsymbol{c}/\|\boldsymbol{c}\|$.

Step 4: 如果向量 $\boldsymbol{c}$ 收敛则停止; 否则返回Step 2 继续以上过程.

下面对上述算法做 4 点说明:

1) 在欧氏空间中实施通常的 K-means 聚类或求取算术平均然后向 $\boldsymbol{S}^{p-1}$ 投影可能被认为是更加直接的方法. 其实不然, 以求取平均为例, 如果数据点在超球面上呈现中心对称分布, 则通常的算术平均将给出退化的解. 此外, 对于更一般的非线性流形凸性条件往往难以得到满足, 因此投影算子不再是适定的.

2) 根据 $\boldsymbol{S}^{p-1}$ 的紧致性, 目标函数 J 在 $\boldsymbol{S}^{p-1}$ 上是有界的. 由于在极大值点和极小值点处具有相同的不动点迭代形式, 初始向量的选取将会影响算法实施的结果. 对此可以考虑不动点迭代和梯度下降相结合的学习模式. 在算法实施的初期, 利用梯度下降法将聚类中心移动到极小值点附近; 然后利用不动点迭代求取最终的极小值点. 关于梯度下降法可以利用如下 3 种迭代形式:

$$\boldsymbol{c}_k\leftarrow\boldsymbol{c}_k-\lambda\partial_{\boldsymbol{c}_k}J,\ \boldsymbol{c}_k\leftarrow\boldsymbol{c}_k/\|\boldsymbol{c}_k\|; \tag{13}$$

$$\boldsymbol{c}_k\leftarrow\boldsymbol{c}_k-\lambda\nabla_{\boldsymbol{c}_k}J,\ \boldsymbol{c}_k\leftarrow\boldsymbol{c}_k/\|\boldsymbol{c}_k\|; \tag{14}$$

$$\boldsymbol{c}_k\leftarrow g(\lambda). \tag{15}$$

其中

$$g(t)=\cos(\|\nabla_{\boldsymbol{c}_k}J\|t)\boldsymbol{c}_k-\frac{\sin(\|\nabla_{\boldsymbol{c}_k}J\|t)}{\|\nabla_{\boldsymbol{c}_k}J\|}\nabla_{\boldsymbol{c}_k}J$$

是以 $\boldsymbol{c}_k$ 为起点, $-\nabla_{\boldsymbol{c}_k}J$ 为初始速度的测地线方程, λ

是学习步长. 式(13)和(14)分别是常规梯度流和黎曼梯度流的离散迭代格式, 显然格式(14)要优于格式(13). 而式(15)是测地流的离散迭代格式, 测地流技术直接在约束空间中进行优化, 可以省去归一化的步骤, 并具有明显的几何意义.

3) 上述降梯度和不动点相结合的优化策略也保证了算法的局部收敛性, 因为在每次迭代中目标函数的取值是下降的. 文献[10]对这种两步迭代算法的收敛性有更详尽的论述. 此外, 由于目标函数是非凸的, 只能给出局部极小值, 通过反复实验或结合一些智能算法可以解决这个问题.

4) 以上的二元聚类算法来源于ICA的实际应用背景, 但是可以发现, 上述算法可以直接推广到$K > 2$时的数据聚类问题.

5　仿真实验与分析

首先考虑超球面二元聚类算法和通常K-means聚类算法的性能比较. 在单位超球面上, 数据的概率分布是不可穷尽的, 以下仅考虑最具代表性的von Mises分布情形. von Mises分布是高斯分布在单位超球面投影的结果. 此外, 在K-means聚类中, 由于数据的单位范数约束, 还须对其聚类结果进行单位化.

在实验中, 首先随机产生两个原始的中心单位向量; 然后将一组具有相同维数和特定标准差的高斯噪声向量组叠加到中心向量上, 并进行范数归一化. 其中, 两个中心向量所叠加的噪声数据个数也是随机的. 随后分别利用这两种方法进行聚类分析, 且考虑其聚类中心与对应的原始中心向量的相关性差值作为评判这两种方法优劣的标准. 由于有两个聚类中心, 可以考虑相关性差值的平均值, 而相关性差值由球面聚类相关系数减去K-means聚类相关系数得到. 表1显示了在不同数据维数和噪声标准差下相关性差值的计算结果. 其中, 每次实验的球面样本点均取为50, 而表中每个数据是重复100次实验所得结果的平均值. 从表1可以看出, 所有数字都是正值, 这说明球面聚类的效果要优于K-means聚类的效果, 且这种优势在低维数据和高水平噪声下越发明显. 尽管表中的每个数据本身是个随机量, 但是如下统计性规律总是存在的, 即除去个别点外, 相关性差值随数据维数的增大和噪声水平的减小而减小.

表 1　相关性差值　　10^{-4}

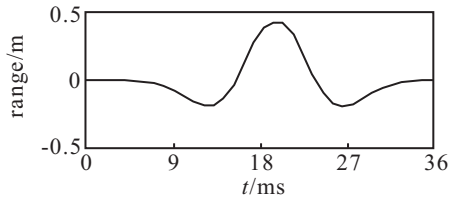
噪声标准差	维数						
	5	10	20	30	40	50	60
0.1	2.56	2.47	1.82	1.80	1.65	1.29	1.22
0.3	17.1	15.4	11.2	7.09	4.57	2.01	2.04
0.5	35.5	29.0	18.7	10.2	8.12	5.98	4.23

以下考虑该算法在地球物理勘探领域的应用. 依据线性系统假设, 人工地震记录可以看作是地层反射系数序列与地震子波卷积的结果. 由于反射系数与地震子波均未知, 地震反卷积基本上是一个盲过程. 最近, 文献[11-12]提出了一种称为带状独立成分分析(BICA)反卷积的方法. 该方法将单道反卷积问题转化为多通道ICA问题; 然后利用ICA中的算法对反卷积问题实施求解. 相对于经典的最小熵和Bussgang型反卷积方法, BICA反卷积收敛更快并具有较强的鲁棒性.

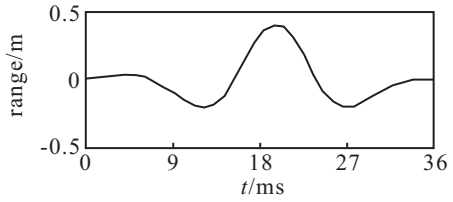
利用ICA的并行处理结构可以分解出与子波长度数值相等的多道候选子波, 为此, 文献[11]提出了一种基于最小二乘的选取方法, 但是这种方法对于子波的时移相当敏感. 此外, 在人工地震勘探中一般认为地震子波作为系统输入在一定的空间范围内具有相同或相近的形式, 因此可以利用邻近道间的相关信息来克服子波提取中时移的影响. 为了使学习过程稳定, 假设所有子波均具有归一化能量, 这样就自然赋予了子波单位超球面的几何结构. 由于ICA符号的不确定性, 提取出的子波向量在超球面上呈现出近似对称的二元分布结构. 可以考虑在单位超球面上对多道子波实施二元聚类, 然后将其中一类数据和另一类数据的反号合并, 并求取这一类新数据的球面平均, 以此消除测量系统和单道数据处理中的随机影响. 子波提取不是本文关注的重点, 相关细节见文献[11-12].

以下采用采样数为400点且满足拉普拉斯分布的序列模拟地层反射系数, 而地震子波采用零相位雷克子波模拟[11], 且采样数为36点, 两者卷积后就得到模拟地震数据. 经过BICA, 可以对提取出的多道子波实施球面二元聚类和求取平均. 图2显示了对25道模拟地震数据实验的结果. 其中, 图1(a)是卷积用的36点零相位雷克子波, 图1(b)是对基于邻近道相关所提取出的25道子波实施超球面二元聚类和平均的结果, 图1(c)是对这25道子波实施欧氏聚类和平均的结果, 图1(d)是对基于文献[11]中最小二乘法所提取出的25道子波实施欧氏聚类和平均的结果. 由图2可见, 利用单位超球面上聚类和平均的方法提取出的地震子波明显优于直接用欧氏聚类和算术平均所得的结果, 且图1(c)中的子波在时间轴上被拉伸, 因而主频降低, 利用其实施反卷积势必将降低信号的分辨率, 图1(d)的结果显然是不能被接受的.

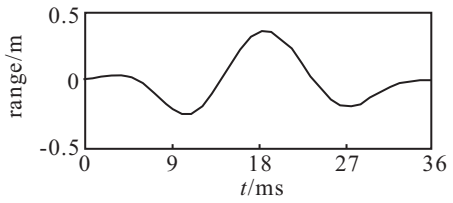
改变地震数据的道数, 进行子波提取可以得到相似的结果. 图3显示了道数分别为5, 10, 15, 20, 25和30时, 用这两种方法提取出的子波与原始雷克子波间相关系数的比较, 黑色标志为超球面聚类及平均, 灰色标志为欧氏聚类及平均. 由于时移影响, 这些相



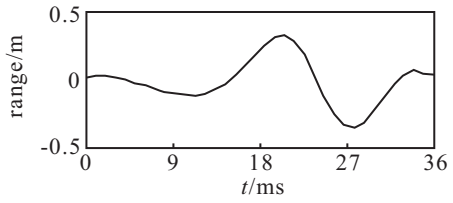
(a) 雷克子波



(b) 球面聚类及平均结果



(c) 欧氏聚类及平均结果1



(d) 欧氏聚类及平均结果2

图 2 各种子波的比较

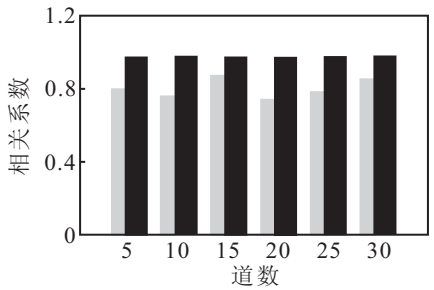


图 3 相关系数的比较

关系数是依据子波振幅谱计算而得的, 且图中的数值结果是进行50次实验的平均值. 由图3可以看出, 基于超球面聚类和平均的子波提取方法具有明显的优势, 在提取过程中几乎完全恢复了子波的振幅谱信息.

6 结 论

本文利用黎曼度量研究了单位超球面上二元聚类和求取数据平均的方法, 然后将其应用于地震子波提取并取得了较通常方法更为满意的结果. 而且, 这种方法能够直接推广到单位超球面上任何聚类分析或数据学习算法上. 本文只研究了 von Mises 分布下这两种聚类分析的比较, 对于一般情况下的性能比较还值得进一步研究. 此外, 由于单位超球面是一个常曲率空间, 它同欧氏空间的特征维数仅相差 1, 其弯曲程度并不是很明显, 而在欧氏空间中更低维嵌入子流形上的机器学习理论是一个值得关注和研究的方向.

参考文献(References)

[1] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(22): 2323-2326.

[2] Tenenbaum J B, Silva de V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(22): 2319-2323.

[3] Amari S, Fiori S. Geometrical methods in neural networks and learning[J]. Neurocomputing, 2005, 67(8): 1-7.

[4] Goh A, Vidal R. Unsupervised Riemannian clustering of probability density functions[C]. European Conf on Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Antwerp, 2008: 377-392.

[5] Goh A, Vidal R. Clustering and dimensionality reduction on riemannian manifolds[C]. IEEE Conf on Computer Vision and Pattern Recognition. Anchorage, 2008: 1-7.

[6] Kim J, Shim K H, Choi S. Soft geodesic kernel K-means[C]. IEEE Int Conf on Acoustics, Speech and Signal Processing. Honolulu, 2007, 2: 429-432.

[7] Hyvarinen A, Karhunen J, Oja E. Independent component analysis[M]. New York: Wiley, 2001.

[8] Wang F S, Li H W, Li R. Data mining with independent component analysis[C]. The 6th World Congress on Intelligent Control and Automation. Dalian, 2006, 2: 6043-6047.

[9] Neill B O. Elementary differential geometry[M]. New York: Academic, 1966.

[10] MacQueen J B. Some methods for classification and analysis of multivariate observations[C]. Proc of 5th Berkerly Symposium on Mathematical Statistics and Probability. Berkerly, 1967, 1: 281-297.

[11] Kaplan S T, Ulrych T J. Blind deconvolution and ICA with a banded mixing matrix[C]. The 4th Int Symposium on ICA and Blind Signal Separation. Nara, 2003: 223-228.

[12] Aws A Q, Woo W L, Dlay S S. Blind seismic deconvolution of single channel using instantaneous independent component analysis[C]. The 6th Int Symposium on Communication Systems Networks and Digital Signal Processing. Graz, 2008: 142-146.