非线性流形上的线性结构聚类挖掘

王 力[1] 吴成东[1] 陈东岳[1] 李孟歆[2] 陈 莉[2]

摘 要 针对非线性数据流形的线性结构挖掘问题, 提出一种基于 Grassmann 流形和蚁群方法的聚类算法. 为抑制噪声对线性结构探测的影响, 对含噪数据集进行算法处理最小单元提升, 利用 Grassmann 流形定义提升后单元间相似度, 同时设计了一种类测地距离作为簇连通性约束. 为提高蚁群解的线性结构挖掘质量, 提出了曲面复杂度最小方向定义, 并将其作为信息素更新的启发信息引入. 在多个数据集上的实验和分析表明, 与 K-means、Geodesic K-means 以及有限混合模型 (Finite mixture model, FMM) 等传统算法相比, 本文算法具备挖掘非线性流形上线性结构的新特性, 并且能够保证线性结构内部的连通性.

关键词 数据流形, 线性结构, Grassmann 流形, 蚁群聚类, 流形假设

引用格式 王力, 吴成东, 陈东岳, 李孟歆, 陈莉. 非线性流形上的线性结构聚类挖掘. 自动化学报, 2012, **38**(8): 1308−1320

DOI 10.3724/SP.J.1004.2012.01308

Exploring Linear Homeomorphic Clusters on Nonlinear Manifold

WANG Li[1] WU Cheng-Dong[1] CHEN Dong-Yue[1] LI Meng-Xin[2] CHEN Li[2]

Abstract This paper proposed a new clustering alogorithm based on ant colony optimization and Grassmann manifold for exploring linear homeomorphic clusters on non-linear dataset manifold. The minimum processed units of algorithm were first lifted to suppress the influence of noise, and then the similarity of unit was measured according to Grassmann manifold and a geodesic-like distance was designed for ensuring the connectivity of cluster. To improve the quality of cluster generated by ant colony clustering, the direction of minimum surface complexity was defined and introduced into the pheromone update strategy as heuristic information. Experiments and analysis on several datasets have shown the successful performance on linear homeomorphic clustering compared to traditional clustering algorithms.

Key words Data manifold, linear homeomorphic clusters, Grassmann manifold, ant colony clustering, manifold assumption

Citation Wang Li, Wu Cheng-Dong, Chen Dong-Yue, Li Meng-Xin, Chen Li. Exploring linear homeomorphic clusters on nonlinear manifold. *Acta Automatica Sinica*, 2012, **38**(8): 1308−1320

Seung 等的相关研究[1] 指出, 人类的视觉感知是以流形的方式存在, 图像流形与人脑神经网络流形之间存在对应关系. 以此结论为基础, 产生了样本分布的流形假设[2], 该假设认为, 采集所得的样本数据分布于嵌入在外围观测空间的低维流形上, 数据观测值的变化是由少量自由参数的连续变化造成的. 目前已有大量实例证实了该假设用于数据分析的可行性[2−15], 但现有算法的目的仅为获取数据的低维表示, 并未对流形建立显式表达, 而基于流形与欧氏空间的同胚性, 建立显式表达的一种可行思路是利用若干线性结构对流形进行近似[16], 但其先决条件是挖掘出非线性流形的线性结构. 由于对线性数据集的分析研究已有很多成熟技术, 因此在挖掘获得非线性流形的线性结构后, 便可利用线性分析技术对非线性流形进行分析, 从而降低原问题难度. 综上可知, 对非线性流形的线性结构挖掘问题展开研究具有重要的意义, 为此, 本文提出一种基于 Grassmann 流形的线性结构挖掘算法, 通过将原始数据集聚类为若干具有近似线性分布的简单数据簇的集合, 来表示流形的复杂非线性结构.

由流形假设, 采样所得样本可假设分布于嵌入在观测空间的流形上, 称观测空间的维度为数据的观测维度, 流形的维度为数据的内蕴维度, 并且当样本在观测空间中呈非线性分布时, 可称样本采样自非线性流形. 为方便分析, 且不失一般性, 样本可进一步假设为分布在 C^{∞} 类流形上, C^{∞} 类流形是处处光滑的流形, 任意局部邻域均与同维度欧氏空间局部同胚, 即流形的任意邻域与同维度欧氏空间之间存在连续的双射. 由于在实际应用中不需要局部邻域任意小, 邻域可以在近似保持同胚映射的基础上

收稿日期 2011-09-20 录用日期 2011-12-19
Manuscript received September 20, 2011; accepted December 19, 2011
国家自然科学基金 (61005032), 辽宁省自然科学基金 (20102062), 沈阳市科学计划项目 (F10-147-9-00), 中央高校基本科研业务费项目 (N100604018) 资助
Supported by National Natural Science Foundation of China (61005032), Natural Science Foundation of Liaoning Province (20102062), Science and Technology Planning Project of Shenyang (F10-147-9-00), and the Fundamental Research Funds for the Central Universities (N100604018)
本文责任编委 刘成林
Recommended by Associate Editor LIU Cheng-Lin
1. 东北大学信息科学与工程学院 沈阳 110819 2. 沈阳建筑大学信息与控制工程学院 沈阳 110168
1. School of Information Science and Engineering, Northeastern University, Shenyang 110819 2. Information and Control Engineering Faculty, Shenyang Jianzhu University, Shenyang 110168

扩大成局部区域, 因此可以认为样本集所在的一“片” C^∞ 类流形是由若干“片”与之同维的欧氏空间拼接而成. 基于此, 非线性流形上的线性结构挖掘问题可总结为: 给定采集自非线性流形 M 上的数据集 $X=\{\boldsymbol{x}_i|1\le i\le n,\boldsymbol{x}_i\in\mathbf{R}^d\}$, M 的维度为 r, 且有 $r<d$, 挖掘算法对其完成分簇, 从而以数据簇的形式获得观测空间 $\mathbf{R}^d$ 中具有线性特性的部分, 并且簇内应保持数据点的聚集性. 对于非线性部分, 则将其划分为若干具有近似线性特性的簇. 如图 1 所示, 设数据集 X 位于维度为 1 的流形上, 观测空间为 $\mathbf{R}^2$, 则上述挖掘问题即为通过分簇获得线性结构簇 A, 并且将数据集非线性部分拆分为若干近似线性结构簇 $B\sim F$.

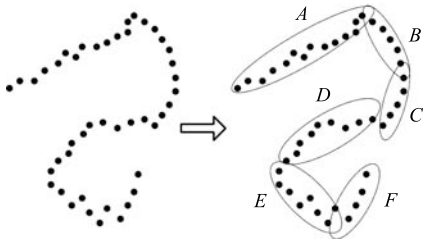


图 1　一维流形上线性结构挖掘问题示意图

Fig. 1　Linear structure mining on 1D manifold

本文算法的整体流程描述如图 2 所示.

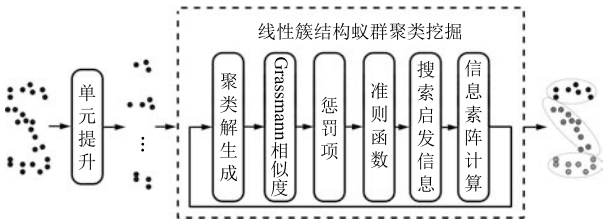


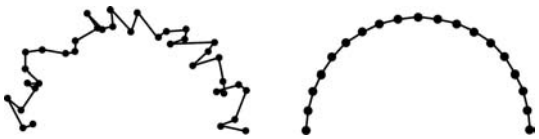
图 2　算法整体流程

Fig. 2　Algorithm procedure

算法首先将所处理的最小单元从单个样本点提升为样本子集, 以抑制噪声对线性结构探测的影响, 利用 Grassmann 流形定义最小单元间的线性分布相似度, 并用于控制最小单元的生成大小. 在此基础上设计了线性结构聚类准则函数, 并通过定义子集类测地距离为惩罚项, 以保证最终每个线性簇的单连通性. 之后基于 Shelokar 蚁群模型[17] 对聚类准则函数进行最优化求解, 并提出最小曲面复杂度方向估计, 将此信息作为启发式信息融入 Shelokar 蚁群信息素更新策略, 从而获得样本集最佳的线性簇结构划分. 通过与 K-means、Geodesic K-means[18] 以及有限混合模型 (Finite mixture model, FMM)[19] 等聚类算法的对比实验, 考察了本文算法在挖掘线性簇结构方面的优势, 并通过人脸流形上的聚类实验, 验证了本文算法在揭示流形变化方面的新特性.

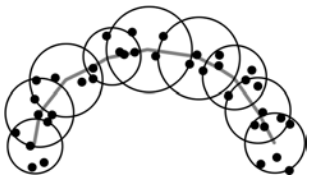
1　最小单元提升

从信号多尺度处理的角度分析, 单个样本与噪声点尺度级别相同, 样本与样本之间的关系容易受噪声的存在而被污染, 仅在样本尺度上难以稳定且真实的获得样本集的分布特性. 如图 3 中真实数据以圆弧曲线特性分布, 受噪声污染后邻近样本之间的关系已经体现不出原有性质, 但通过提高信号空域中的分析尺度, 以若干邻近样本作为最小单元, 那么这些最小单元间的相互关系就能近似体现出原分布的特性.



(a) 受噪声污染前后的邻域关系

(a) Neighborship of data without noise (left) or with noise (right)



(b) 提升分析尺度后得到的单元邻域关系

(b) Neighborship of unit after lifting the scales of analysis

图 3　分析尺度与邻域关系示意

Fig. 3　Schematic diagram of analytical scale and neighborhood relations

根据流形假设, 样本是从 r 维光滑流形 M 上采样得来, 并且采集存在观测噪声. 为获得合理的分析尺度以体现出样本在 M 上的分布特性, 使得点集 $\Delta=\{\boldsymbol{x}_{a_1},\boldsymbol{x}_{a_2},\cdots,\boldsymbol{x}_{a_k}\}$ 可称之为当前样本集 $X=\{\boldsymbol{x}_i|\boldsymbol{x}_i\in\mathbf{R}^d,1\le i\le n\}$ 的一个最小单元, 体现出非线性流形的局部线性特性, 需满足如下四个条件:

1) $a_1,\cdots,a_k\in\{i|i\in\mathbf{Z},1\le i\le n,1<k<n\}$;

2) 存在 $\boldsymbol{x}_{a_c}$, $1\le c\le k$, 使得 $\{\boldsymbol{x}_{a_j}|j\ne c\}\subset N_{\boldsymbol{x}_{a_c}}$ ($N_{\boldsymbol{x}_{a_c}}$ 表示 $\boldsymbol{x}_{a_c}$ 的邻域);

3) Δ 处 M 上各点的切空间均可由 Δ 的前 r 个线性主成分所张空间近似表示;

4) 从 Δ 中剔除若干样本点, Δ 样本所体现出的前 r 个线性主成分近似不变. 同时, 若有若干样本 $\{\boldsymbol{x}_j\}\subset N_{\boldsymbol{x}_{a_c}}$, 将其添加至 Δ, Δ 中样本所体现出的前 r 个线性主成分近似不变.

条件 1) 和 2) 将 Δ 限制在某个样本的邻域内, 使得 Δ 体现的是光滑流形 M 的局部性质. 在条件 3) 下, 若不存在观测噪声, 则 Δ 可认为是 M 与外围观测空间 $\mathbf{R}^d$ 之间满足同胚映射的局部邻域. 对于实际数据集, 观测噪声带来了额外的维度, 然而从信号能量的分布来看, 相对于信号总能量, 额外维度

上的能量所占比例较小, 因此仍可认为同胚映射在条件 3) 下近似满足. 条件 4) 使得 Δ 所体现的 M 的局部性质具有一定稳定性, 保证其不受噪声数据的影响.

基于以上分析, 本文提出最小处理单元提升算法. 提升过程首先建立邻域关系图, 在此基础上进行点集生长, 利用 Grassmann 流形控制生长停止条件, 最终生成满足上述四个线性特性条件的最小单元.

1.1 可视邻域关系图

文献 [2] 提出了可视邻域的概念, 并利用其建立样本之间的邻接关系. 对于样本 $\boldsymbol{x}$, $\boldsymbol{y}$, $N_{\boldsymbol{x}}^k$ 表示 $\boldsymbol{x}$ 的 k 近邻且 $\boldsymbol{y} \in N_{\boldsymbol{x}}^k$, 当不存在这样的一点 $\boldsymbol{z} \in N_{\boldsymbol{x}}^k$, 使得从 $\boldsymbol{z}$ 点指向 $\boldsymbol{x}$ 点的向量 $\boldsymbol{e}_{zx}$ 与从 $\boldsymbol{z}$ 点指向 $\boldsymbol{y}$ 点的向量 $\boldsymbol{e}_{zy}$ 之间的内积 $\langle \boldsymbol{e}_{zx}, \boldsymbol{e}_{zy} \rangle \leq 0$, 则 $\boldsymbol{y}$ 称为 $\boldsymbol{x}$ 的一个可视邻域点. 在此基础上, 若某一样本点存在 k 个可视邻域点, 将其边按长度升序排列得 $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_k\}$, 使用主成分分析 (Principal component analysis, PCA) 内蕴维度估计[20] 计算前 j $(1 < j \leq k)$ 个边 $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_j\}$ 的内蕴维度 d_j, 得到序列 $\{d_1, \cdots, d_k\}$, 遍历该序列, 检查当 $d_j > d_{j-1}$ 时, 是否有 $|\boldsymbol{e}_j| - |\boldsymbol{e}_{j-1}|$ 大于给定阈值, 若大于, 则第 j 个及后续边均被判定为短路边. 与直接使用 k 近邻方法建立邻域关系相比, 此方法可以有效减小短路边出现的概率, 但由于短路边控制条件仍比较松弛, 使得在排除短路边的过程中也剔除了大量有效边, 导致邻域信息的损失. 因此本文对鉴别条件加以补充, 在剔除短路边的同时提高邻域信息的保持度.

观察短路边的形成, 可以发现短路边通常比有效边长, 且短路边的出现会带来以边为分析对象时的局部内蕴维度的突变, 文献 [2] 的鉴别算法即基于此事实. 除此之外, 设在可视邻域选择后, 所建立的邻域关系图为 $G = \langle V, E \rangle$, V 为样本集 $\{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbf{R}^d, 1 \leq i \leq n\}$, E 为连接边集. 对于任一边 $\boldsymbol{e}_j$, 定义其权重 w_j 为图 G 中所有最短路径经过的次数, 可以发现由于短路边承载着本不应连接在一起的两个子集间的最短路径连接, 因此其权重是局部极值点, 本文将其作为短路边的补充判据. 最终短路判据为:

若有

$$w_j \geq w_k, \ \forall \ w_k \in \Omega_1 \cup \Omega_2 \tag{1}$$

且 $\boldsymbol{e}_j$ 同时满足基于内蕴维度检测的短路判据, 则 $\boldsymbol{e}_j$ 可认为是短路边. 补充判据 (1) 中 Ω_1 与 Ω_2 表示连接在 $\boldsymbol{e}_j$ 上的两个节点上边的权重的集合.

1.2 Grassmann 测地距离相似度

对于一个点集, 其线性主分量可以体现出点集在观测空间中的线性结构信息. 设样本点集 Δ 采集自嵌入在 d 维观测空间的 r 维流形, 令矩阵 $F \in \mathbf{R}^{d \times r}$ 由 Δ 的前 r 个线性主分量构成, 则 F 为一列正交阵, 表示点集的线性簇结构特征.

依据微分几何理论, 所有如下等价列正交矩阵的全体构成 Grassmann 流形[21]:

$$G(r, d) = \{q | q = pQ, \forall Q \in SO(r)\} \tag{2}$$

其中, p 和 q 是 $d \times r$ 正交阵, Q 为等价变换阵, $SO(r)$ 为特殊正交群 $\{Q | Q^{\mathrm{T}}Q = I, \det(Q) = 1\}$. Grassmann 流形 $Gr(d, r)$ 是一种解析流形, 其更为直观的解释是: d 维线性空间的所有 r 维子空间的集合, 有关 Grassmann 流形更为详尽的讨论见文献 [21]. 由于 F 是由 Δ 的前 r 个线性主分量构成的, 故其为 $\mathbf{R}^d$ 的某一子空间 $\mathbf{R}^r$ 的基, 而同一线性空间的等价基是满足式 (2) 所定义 Grassmann 流形的一组等价列正交阵, 从而可知, F 与 $Gr(d, r)$ 上的点相互对应. 因此, 通过 Grassmann 流形, 可自然的定义点集线性主分量的相似度度量. 由于 Grassmann 流形为紧致流形[22], 故其上的任意两点都可以通过一条测地线连接, 且其长度为流形上该两点间的最短距离. 因此任意两个点集之间的线性簇结构的相似度可以利用该最短距离来衡量. 为计算此距离, 需引入黎曼指数映射及其逆映射. 设正交阵 Ψ 对应 $G(d, r)$ 上的一点 p, 则有黎曼指数映射 $\exp_p : T_p \rightarrow G(d, r)$:

$$F = \exp_p(x) = \Psi \boldsymbol{V} \cos\Theta + U \sin\Theta \tag{3}$$

可将切空间 T_p 上的一点 x 映射至 $G(d, r)$[23], 其中 $U\Theta V^{\mathrm{T}}$ 是 x 的 THIN-SVD 分解.

相应的, 在以 p 点为中心的内射半径开球内, 存在 $\exp_p$ 的逆映射, 黎曼对数映射 $\log_p : G(d, r) \rightarrow T_p$[23]:

$$x = \log_p(F) = U \tan^{-1}(\Sigma) V^{\mathrm{T}} \tag{4}$$

其中, $U\Sigma V^{\mathrm{T}} = (I - \Psi\Psi^{\mathrm{T}}) F (\Psi^{\mathrm{T}} F)^{-1}$.

对于任意两点 $x \in G(d, r)$, $y \in G(d, r)$, 以 x 为起点, y 为终点的测地线的长度与切空间 T_x 中向量 $\log_x(y)$ 的模相等, 基于此, 对于点集 Δ_1 与 Δ_2, 其特征 F_1 与 F_2 之间的相似性可以通过如下测地距离度量:

$$\begin{aligned} d(F_1, F_2) = & \left\langle \log_{F_1}(F_2), \log_{F_1}(F_2) \right\rangle = \\ & \left\| \log_{F_1}(F_2) \right\|_{F_1} = \\ & \mathrm{tr}\left(\log_{F_1}(F_2) \left(\log_{F_1}(F_2) \right)^{\mathrm{T}} \right) = \\ & \mathrm{tr}\left(U \tan^{-2}(\Sigma) V^{\mathrm{T}} \right) \end{aligned} \tag{5}$$

其中, $U\Sigma V^{\mathrm{T}} = (I - F_1 F_1^{\mathrm{T}}) F_2 (F_1^{\mathrm{T}} F_2)^{-1}$.

1.3 线性簇结构特征变化分析与最小单元生成

最小单元的生成为一点集生长过程, 从种子点开始, 点集按照可视邻域关系逐步向外扩张, 不断纳

入最外层新的数据点, 这一过程保证点集始终满足最小处理单元条件 1) 和 2). 为使得最终生长结果满足条件 3) 和 4), 本文利用点集 Grassmann 测地距离来控制最终子集的大小.

设点集从 p 点开始生长, Δ_t 表示第 t 圈扩张完成后的点集, F_t 为其线性簇结构特征. 对生长过程分析可知, 对于数据集中任意一点处的子集 Δ, F_t 的变化有如下规律: 从第一轮生长扩张开始, 由于点集规模较小, F_t 的变化受单个样本点位置的影响较大, 线性簇结构信息易被噪声淹没, 这使得相邻两次 F_t 和 F_{t+1} 之间的差异较大. 但随着 Δ 规模的扩大, 噪声的相对尺度减少, F_t 受生长过程新纳入的单个样本点的影响消弱, F_t 和 F_{t+1} 之间的差异逐渐缩小, 并在子集规模对条件 3) 和 4) 满足度最高时差异值达到极小值, 在此之后, 规模的逐渐增长会使得样本分布曲率带来的影响无法忽略, 子集无法满足同胚映射, F_t 和 F_{t+1} 之间的差异又逐步扩大. 基于以上分析, 可设计样本集最小单元的提升算法, 如算法 1 所示, 其中 $t_{\min}$ 和 $t_{\max}$ 为子集生长限幅, 用以避免子集规模的过小或过大, $d_{i,j}$ 表示 F_i 和 F_j 之间的 Grassmann 测地距离. 算法从每一个样本点 x_k 开始生长, 寻找生长过程中差异矩阵 D_t 具有最小范数的轮次 t^*, 并将此轮次所获得子集 Δ_{t^*} 作为样本点 x_k 提升所得的最小单元 Θ_k. 最终经过提升, 将获得具有交叠的密集最小单元.

算法 1. 最小单元生成算法

Input. $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$

Repeat for: $k\leftarrow 1$ to n

1) $\Delta_1^k\leftarrow\{\boldsymbol{x}_k\}$

2) 从 Δ_1^k 生长至 $\Delta_{t_{\min}}^k$

3) **Repeat for**: $t\leftarrow(t_{\min}+1)$ to $(t_{\max}-1)$

a) 生成 Δ_t^k

b) 计算 F_{t-1},F_t,F_{t+1}

$$D_t=\begin{bmatrix} d_{t-1,t-1} & d_{t-1,t} & d_{t-1,t+1}\\ d_{t,t-1} & d_{t,t} & d_{t,t+1}\\ d_{t+1,t-1} & d_{t+1,t} & d_{t+1,t+1}\end{bmatrix}$$

$$\xi_t=\|D_t\|_2$$

End repeat

4) $t^*=\arg\min_t|\xi_t-\min\{\xi_t\}|$

5) $\Theta_k=\Delta_{t^*}$, $F_{\Theta_k}=F_{t^*}$

End repeat

Output. $\{\Theta_1,\Theta_2,\cdots,\Theta_n\}$

2　线性结构挖掘的聚类描述

最小单元提升后, 属于同一线性结构内的最小单元具有相似的线性簇结构特征, 而不同线性结构之间的单元线性簇结构特征相异, 基于此可将线性结构的挖掘问题转化为一个聚类问题: 给定提升后的最小单元集合, 通过聚类将其划分为若干子集, 并且要求属于同一线性结构部分的最小单元被划分到相同子集中, 不同子集中的最小单元分属不同的线性结构.

2.1　聚类相似度测度

由于属于同一线性结构内的最小单元的线性簇结构特征相似, 反之则相异, 因此可利用 Grassmann 测地距离相似度来作为线性结构聚类挖掘时的相似度测度. 但 Grassmann 测地距离与观测空间中的空域距离无关, 仅以 Grassmann 测地距离为聚类相似测度可能使在流形上具有相同线性簇结构特征, 但在空域上被其他线性结构间隔开的非连通区域划分进同一线性结构当中, 如图 4 所示, 一维非线性流形 M 上的两个部分 A 和 B 具有相似的线性分布特性, 但被非线性部分 C 隔开, 若仅依靠线性簇特征之间的 Grassmann 测地距离作为相似度进行聚类, 则将形成一个从观测空间来看内聚性不好的簇, 如图 4 右侧所示.

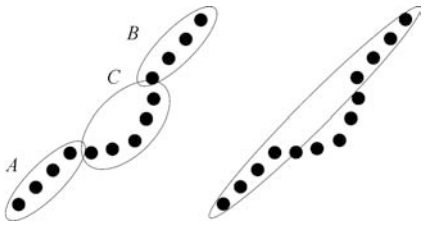


图 4　内部不连通的线性结构

Fig. 4　Internally disconnected linear structure

因此, 为保证挖掘所得线性结构的连通性, 在观测空间定义一种最小单元类测地距离, 设提升后最小处理单元为 $\{\Theta_1,\Theta_2,\cdots,\Theta_n\}$, 根据微分几何中测地线定义, 任意两个最小处理单元 Θ_i 和 Θ_j 之间的类测地距离可定义为

$$g(\Theta_i,\Theta_j)=Q(\eta^*) \tag{6}$$

其中

$$Q(\eta)=\sum_{k,k+1\in\eta}\frac{f(\Theta_k\cap\Theta_{k+1})}{\min(f(\Theta_k),f(\Theta_{k+1}))} \tag{7}$$

$$\eta^*=\arg\min_{\eta}Q(\eta) \tag{8}$$

$$f(\Theta_k\cap\Theta_{k+1})\neq 0 \tag{9}$$

其中, $f(\cdot)$ 表示计算集合的势. 若两个集合交集的势为 0, 则称这两个集合不连通. η 表示 Θ_i 和 Θ_j 之间的某条连通路径, Θ_k 和 Θ_{k+1} 表示在这条连通路径上的两个相邻最小单元, 该类测地距离的求解等价于一最短路径问题, 以每个最小单元作为节点, 连接边存在于连通的两个单元之间, 其权重为集合势的比值, 则 Θ_i 和 Θ_j 之间的类测地距离可通过 Dijkstra 算法[24] 完成求解.

2.2 聚类准则函数

相对于聚类数未知的情况, 在固定聚类数的前提下讨论线性结构的挖掘可以简化问题的求解难度. 设样本集 $X = \{\boldsymbol{x}_i | \boldsymbol{x}_i \in \mathbf{R}^d, 1 \leq i \leq n\}$ 经提升后获得的最小处理单元为 $\{\Theta_1, \Theta_2, \cdots, \Theta_n\}$, 将类测地距离作为对线性结构内部不连通的惩罚引入. 考虑到最小处理单元位于光滑流形上, 其线性簇结构特征的变化也是平滑的, 空域相邻的单元之间具有相似线性簇结构特征的概率更高, 基于此再引入滤波项, 可设计误差平方和准则函数如下:

$$J = \sum_{i=1}^{n}\sum_{j=1}^{k}\omega_{ij} \times \left(d\left(\Theta_i, \bar{\Theta}_j\right) + \alpha g\left(\Theta_i, \bar{\Theta}_j\right) + \beta d\left(\Theta_i, \bar{\Theta}_{\Omega_i}\right)\right) \tag{10}$$

其中, $d(\cdot)$ 和 $g(\cdot)$ 分别为 Grassmann 测地距离和类测地距离, n 和 k 分别表示单元 Θ 的个数和聚类数, $\omega_{ij} = 1$ 表示将第 i 个单元分配给线性结构 j, $\bar{\Theta}_j$ 代表线性结构 j 的聚类中心, Ω_i 为单元 Θ_i 的局部邻域, $\bar{\Theta}_{\Omega_i}$ 表示该邻域内单元在 Karcher 均值意义下的中心. α 和 β 分别为惩罚项与滤波项系数, $\alpha > 0$, $\beta > 0$, ω_{ij} 为分配系数, 有:

$$\omega_{ij} = \begin{cases} 1, & s_i = j \\ 0, & s_i \neq j \end{cases} \tag{11}$$

$s_i = j$ 表示单元 i 属于线性结构 j; $s_i \neq j$ 则反之, 序列 $S = \{s_1, \cdots, s_n\}$ 表示线性结构聚类挖掘最优化问题的解:

$$S^* = \arg\min_S J(S) \tag{12}$$

3 线性结构聚类挖掘的蚁群求解

对于组合优化问题, 无法建立解析模型, 且梯度估计困难, 而与传统算法相比, 蚁群算法作为全局最优解的有效搜索技术, 已被广泛应用于大规模组合优化问题. Shelokar 等提出了一种引入了遗传变异操作的蚁群模型[17], 与传统优化技术相比更适合组合聚类优化, 该算法采用多解并行搜索, 且通过变异操作引入了新信息, 加大了解的搜索范围, 使其具备全局最优的获取能力, 因此本文基于该算法求解线性结构挖掘的优化问题.

Shelokar 蚁群模型是模仿蚂蚁觅食过程的一种仿生算法, 其根据式 (13) 更新信息素阵 T[17]:

$$T_{ij}(t+1) = (1-\varsigma)\, T_{ij}(t) + \sum_{l=1}^{L}\Delta T_{ij}^{l} \tag{13}$$

其中

$$\Delta T_{ij}^{l} = \begin{cases} \dfrac{1}{f_l}, & x_i \text{ 属于簇 } j \\ 0, & \text{否则} \end{cases}$$

由于在上述更新策略中缺乏启发信息, 且线性结构挖掘聚类准则函数中惩罚项 $g\left(\Theta_i, \bar{\Theta}_j\right)$ 的引入会使算法发现位于流形空间球邻域内的线性结构的倾向增大, 这将使得在有限的迭代次数下算法获得最佳线性结构划分的能力受到抑制, 为此本文提出样本曲面复杂度方向估计, 作为搜索时的启发信息, 提高优化所得最优解对应线性结构挖掘的质量.

3.1 点集曲面复杂度最小方向

样本点集的曲面复杂度可由样本点到其最小二乘超平面投影距离在某一方向上体现出的分布方差定义, 如图 5 所示, S_a 和 S_b 分别代表两个不同的样本簇, S_a 更为平坦, 曲面复杂度低, 而 S_b "波折" 程度较大, 曲面复杂度较高, 并且沿 $\boldsymbol{e}_1$ 方向, 曲面复杂度最大. 在此定义下, 一个良好的线性结构应尽可能地保证结构内点的分布平坦, 而分布越 "曲折" 的结构, 其复杂度越高.

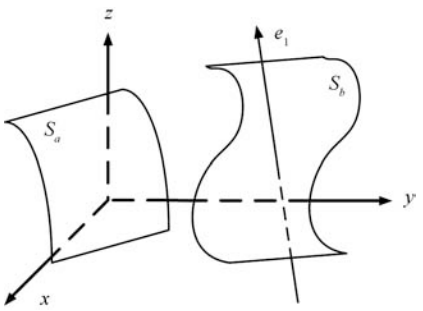


图 5 曲面复杂度示意图

Fig. 5 Schematic diagram of surface complexity

如图 6 所示, 若有一最小单元 Θ_a, $\boldsymbol{e}_a$ 为 Θ_a 的曲面复杂度最小方向, 那么在 $\boldsymbol{e}_a$ 方向上, 如果存在另一最小处理单元 Θ_b, 且 Θ_b 的曲面复杂度最小方向与 $\boldsymbol{e}_a$ 相近, 则将 Θ_a 和 Θ_b 归为同一线性结构内时, 会比将其他单元, 如 Θ_c 和 Θ_a 归为同一线性结构, 使得最终线性结构的曲面平均复杂度更小, 从而获得性质更好的线性结构. 由此可知, 将曲面复杂度最小方向作为启发信息, 可以使得解的搜索过程尽可能地向降低线性结构内曲面复杂度, 增加线性结构内点分布的线性程度的方向进行.

设有点集 $\Delta = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n]^{\mathrm{T}}$, 通过最小二乘拟合得到超平面 $\Gamma = \{\boldsymbol{x} | \boldsymbol{x}^{\mathrm{T}} C + b = 0\}$. 考虑投影方向, 则 Δ 向 Γ 的投影长度阵为

$$H = \frac{\Delta C + bI}{\|C\|} \tag{14}$$

其中, $I = [1, 1, \cdots, 1]^{\mathrm{T}} \in \mathbf{R}^n$.

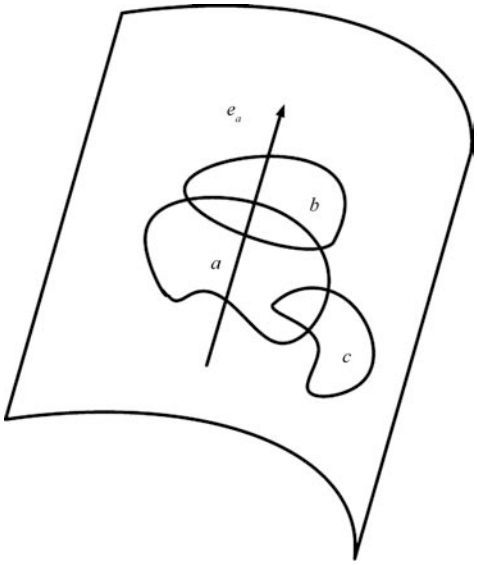


图 6　高线性度聚类示意图

Fig. 6　Schematic diagram of clustering of high linearity

令 $K=[H,H,\cdots,H]_{n\times n}$, 则有投影距离差异阵

$$Z=\left|K-K^{\mathrm{T}}\right|=\begin{bmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{nn} \end{bmatrix} \tag{15}$$

最小曲面复杂度方向可以通过构建相关矩阵并对角化获得求解. 由于 Z 阵为差异阵, 其元素代表了各点向 Γ 投影时投影距离之间的不相似程度, 为构建相关矩阵, 需要将此差异阵转换为相似阵, 即各元素代表各点对之间投影距离的相似程度, 为此可取 $h(x)$ 为单调线性减函数, 且当 $x>0$ 时, $h(x)>0$. 令

$$Z^*=\begin{bmatrix} h(z_{11}) & \cdots & h(z_{1n}) \\ \vdots & \ddots & \vdots \\ h(z_{n1}) & \cdots & h(z_{nn}) \end{bmatrix}=\begin{bmatrix} z_{11}^* & \cdots & z_{1n}^* \\ \vdots & \ddots & \vdots \\ z_{n1}^* & \cdots & z_{nn}^* \end{bmatrix} \tag{16}$$

设 z_{ij}^* 为 Z^* 任一元素, 将其在 $\boldsymbol{x}_i-\boldsymbol{x}_j$ 方向上按 $\boldsymbol{x}_i-\boldsymbol{x}_j$ 模做规范化, 并按矢量 $\boldsymbol{x}_i-\boldsymbol{x}_j$ 与观测空间基 B 之间的相关度正交分解:

$$A_{i,j}^*=\left(\frac{\boldsymbol{x}_i-\boldsymbol{x}_j}{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|}\right)^{\mathrm{T}}\frac{Bz_{ij}^*}{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|} \tag{17}$$

于是相关矩阵可构造为

$$R^*=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{i,j}^{*\mathrm{T}}A_{i,j}^* \tag{18}$$

对 R^* 求取特征分解 $\lambda\boldsymbol{v}=R\boldsymbol{v}$, 则曲面复杂度最小方向 $\boldsymbol{v}_{\min}$ 为最大特征值 $\lambda_{\max}^*$ 对应的特征向量.

3.2　启发式线性结构蚁群挖掘算法

利用曲面复杂度最小方向估计, 可为 Shelokar 模型信息素矩阵更新引入启发式信息. 设信息素矩阵为 T, 更新时解集为 S_L, S_l 为解集中一解, 对应适应度值为 f_l, Ξ_j^l 表示由 S_l 确定的第 j 个线性结构中最小单元的集合. 对 Ξ_j^l 中任一单元 Θ_i, 设其曲面复杂度最小方向为 $\boldsymbol{e}_i$, 若 $\boldsymbol{e}_i$ 和 Ξ_j^l 中其他单元的并集 $\hat{\Xi}_j^l$ 所对应的曲面最小复杂度方向 $\hat{\boldsymbol{e}}_j^l$ 相近, 并且 Θ_i 位于以 $\hat{\boldsymbol{e}}_j^l$ 为方向, 穿过线性结构中心 $\bar{\Theta}_j$ 的均值点 $\bar{\bar{x}}_j$ 的直线上时, 则将 Θ_i 与 $\hat{\Xi}_j^l$ 归为同一线性结构时不会明显降低结构内数据分布的线性及欧氏同胚程度, 因此遗留在从 Θ_i 到线性结构 j 路径上的信息素应该相对更多, 于是融合启发信息的信息素矩阵 T 的更新策略可设计为

$$T_{ij}(t+1)=(1-\varsigma)T_{ij}(t)+\sum_{l=1}^{L}\Delta T_{ij}^l$$

$$\Delta T_{ij}^l=\begin{cases} Q_l, & \Theta_i \text{ 属于簇 } j \\ 0, & \text{否则} \end{cases} \tag{19}$$

$$Q_l=\frac{1}{f_l}+\gamma\left|\left\langle\boldsymbol{e}_i,\hat{\boldsymbol{e}}_j^l\right\rangle\cdot\left\langle\frac{\bar{x}_i-\bar{\bar{x}}_j}{\|\bar{x}_i-\bar{\bar{x}}_j\|},\hat{\boldsymbol{e}}_j^l\right\rangle\right| \tag{20}$$

其中, $\gamma>0$ 为启发系数, 用以控制启发信息的作用程度, $\bar{x}_j$ 与 $\bar{\bar{x}}_j$ 分别是 Θ_i 和 $\bar{\Theta}_j$ 的欧氏均值.

算法每次迭代过程中需要计算聚类中心. 对于黎曼流形 M 上点的均值, 采用 Karcher 均值的定义[25]:

$$\bar{x}=\arg\min_{x\in M}\sum_{i=1}^{n}d^2(x_i,x) \tag{21}$$

其中, $x_i\in M$, 此均值可通过梯度下降迭代求解[26−27]:

$$\bar{x}^{t+1}=\exp_{\bar{x}^t}\left(\frac{1}{n}\sum_{i=1}^{n}\log_{\bar{x}^t}(x_i)\right) \tag{22}$$

由此可得线性结构 j 的聚类中心的均值线性簇结构特征的求解迭代式:

$$\bar{F}_j^{t+1}=\exp_{\bar{F}_j^t}\left(\frac{1}{m}\sum_{F_i\in\Pi_j}\log_{\bar{F}_j^t}(F_i)\right) \tag{23}$$

其中, Π_j 为 Ξ_j^l 中各最小单元线性簇结构特征的集合, m 为该集合中的元素个数. 对于聚类中心在观测空间中的空域位置, 同样基于 Karcher 均值求取:

$$\bar{\Theta}_j=\arg\min_{\Theta\in\Xi_j}\sum_{\Theta_i\in\Xi_j}g^2(\Theta_i,\Theta) \tag{24}$$

则由式 (5) 得:

$$d\left(\Theta_i,\bar{\Theta}_j\right)=\mathrm{tr}\left(U\tan^{-2}(\Sigma)U^{\mathrm{T}}\right) \tag{25}$$

其中, $U\Sigma V^{\mathrm{T}} = (I - F_i F_i^{\mathrm{T}})\bar{F}_j \left(F_i^{\mathrm{T}}\bar{F}_j\right)^{-1}$. 式 (24) 表示若有一单元 Θ 到线性结构内其他单元的类测地距离之和最小, 则以该单元作为聚类中心在观测空间中的近似位置.

至此, 启发式线性结构蚁群挖掘算法可描述如算法 2 所示, 算法接收最小单元集合 $\{\Theta_1, \Theta_2, \cdots, \Theta_n\}$ 和最大迭代次 c, 进行循环迭代. 每次迭代中首先依据 Θ_i 到聚类 j 的选择概率:

$$p_{ij} = \frac{T_{ij}}{\sum\limits_{u=1}^{k} T_{iu}} \tag{26}$$

构造 $S = [s_1, \cdots, s_P]^{\mathrm{T}}$, 解集中每一个解采用整数编码, 代表一种最小单元聚类, 之后计算解中各个线性结构的线性簇结构特征均值 $\bar{F}_j$ 与空域类测地均值 $\bar{\Theta}_j$, 并依据准则函数计算解的适应度, 之后对解进行变异, 对比变异前后的适应度, 并以高适应度解构成新解集, 最后计算每个线性结构相关的曲面复杂度最小方向, 更新信息素阵, 进行下次迭代.

算法 2. 启发式线性结构蚁群挖掘算法

Input. $\Theta_1, \cdots, \Theta_n$, c

随机初始化 T

Repeat for: $iter \leftarrow 1$ to c

1) 计算 Θ_t 到聚类 j 的选择概率

$$p_{ij} = \frac{T_{ij}}{\sum\limits_{u=1}^{k} T_{iu}}$$

依据选择概率构造解集 $S = [s_1, \cdots, s_p]^{\mathrm{T}}$

2) 针对 S 中各个解计算各聚类中心

$$\overline{F}_j^{t+1} = \exp_{\overline{F}_j^t}\left(\frac{1}{m}\sum_{F_i \in \prod_j}(F_i)\right)$$

$$\overline{\Theta}_j = \arg\min_{\Theta \in \Xi_j} \sum_{\Theta_i \in \Xi_j} g^2(\Theta_i, \Theta)$$

3) 计算每个解的适应度

$$J = \sum_{i=1}^{n}\sum_{j=1}^{k} \omega_{ij} \times \left(d\left(\Theta_i, \bar{\Theta}_j\right) + \alpha g\left(\Theta_i, \bar{\Theta}_j\right) + \beta d\left(\Theta_i, \bar{\Theta}_{\Omega_i}\right)\right)$$

$$d\left(\Theta_i, \bar{\Theta}_j\right) = \mathrm{tr}\left(U \tan^{-2}(\Sigma) U^{\mathrm{T}}\right)$$

$$g\left(\Theta_i, \Theta_j\right) = Q\left(\eta^*\right)$$

取前 L 个最佳适应度解 $S_L = [s_1, \cdots, s_L]^{\mathrm{T}}$

4) 对 S_L 各解进行变异操作, 获得 S'_L

5) 计算 S'_L 中各解的适应度

$$J'_L = [f'_1, \cdots, f'_l, \cdots, f'_L]$$

6) 对于 S'_L 中满足 $f'_l > f_l$ 的解, 使用其将 S_L 中对应解替换, 更新 J_L

7) 更新信息素矩阵 T

$$T_{ij}(t+1) = (1-\varsigma)T_{ij}(t) + \sum_{l=1}^{L}\Delta T_{ij}^{l}$$

$$\Delta T_{ij}^{l} = \begin{cases} Q_l, & \Theta_i \text{ 属于簇 } j \\ 0, & \text{否则} \end{cases}$$

$$Q_l = \frac{1}{f_l} + \gamma\left|\left\langle \boldsymbol{e}_i, \hat{\boldsymbol{e}}_j^{l}\right\rangle \cdot \left\langle \frac{\bar{x}_i - \bar{\bar{x}}_j}{\|\bar{x}_i - \bar{\bar{x}}_j\|}, \hat{\boldsymbol{e}}_j^{l}\right\rangle\right|$$

8) 更新当前适应度最大解 s^*

End repeat

Output. s^*

4 实验与分析

4.1 合成数据集实验分析

为了直观地对算法进行评价, 本文在四种合成数据集上进行了实验, 如图 7 所示, 这四种数据集分别为 Rect-line, Open-D, Swiss-Roll, S-Curve. 其中 Rect-line 和 Open-D 的观测维度为 2D, 内蕴维度为 1D, Swiss-Roll 和 S-Curve 为流形学习中的典型数据集, 其观测维度为 3D, 内蕴维度为 2D, 所有数据集均加入信噪比为 5 db 高斯白噪声. Rect-line 具有 3 个本质线性结构; Open-D 具有两个本质线性结构, 其余部分为光滑变化的非线性结构; Swiss-Roll 无本质线性结构; S-Curve 具有一个本质线性结构.

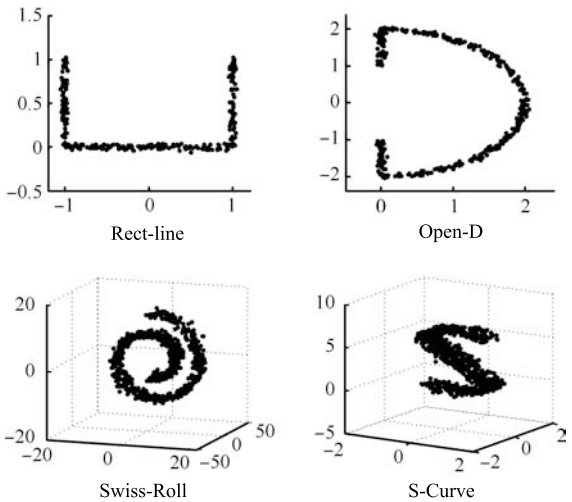


图 7 实验数据集

Fig. 7 Experimental datasets

在线性结构聚类数确定的前提下, 对非线性数据流形的线性结构的理想挖掘应该为: 优先发现本质线性结构, 并将非线性部分拆解为若干最佳的近似线性结构. 在数据集 Rect-line 和 Open-D 上, 本

文对 K-means, Geodesic K-means[18]、有限混合模型 (FMM)[19], 以及本文算法在多种簇划分数 k 下进行了对比实验. 其中 K-means 采用欧氏距离相似测度; Geodesic K-means 采用样本流形上的测地距离估计作为相似测度; FMM 为有限混合线性回归模型, 用以发现 2D 空间下的线状结构; 本文算法所采用的解集规模为 50, 最大迭代次数为 300 次. 图 8 所示第 1～4 列分别对应 K-means、Geodesic K-means、FMM 以及本文算法在 Rect-line 和 Open-D 上的线性结构挖掘结果, 不同的线性结构使用不同的灰度标记.

由图 8 可知, 在 Rect-line 数据集上, 当聚类数 k 取 3 时, FMM 与本文算法均能获得理想的线性结构挖掘结果; 当 k 为 4 时, FMM 与本文算法均将 k 为 3 时的一个线性结构拆分为两个, 但本文算法对原线性结构的拆分具有紧致性, 新结构内的邻域关系与原结构相应位置保持一致, 而在 FMM 下, 新结构内已经不能保持原结构中数据之间的真实邻域关系. FMM 的这一不足也在 Open-D 数据集上被体现. 对于 Open-D 数据集, 在不同 k 下, 本文算法均成功地将两个本质线性结构优先捕获到, 并将数据流形的非线性部分拆分为 k 值约束下的若干近似线性结构; 而 FMM 算法始终将这两个分立的本质线性结构归并为同一结构, 仅作线状回归, 使得结构内无法保持连通. 对于 K-means 算法, 由于其仅依靠数据点之间的欧氏距离衡量相似度, 因此无法捕获流形上的结构信息, 不具备探测线性结构的能力, 如数据集 Rect-line 和 Open-D 上的聚类结果. Geodesic K-means 通过估计样本所在流形上的测地距离来衡量数据相似度, 没有考虑样本的分布情况, 同样无法探测线性同胚簇, 但在 Rect-line 数据集上, 当 k 为 4 时, Geodesic K-means 也获得了理想的聚类结果, 这是因为其倾向于产生测地距离上等规模的聚类簇, 当 $k=4$ 时, 恰好将原数据集按测地距离等分为 4 部分.

由于 FMM 算法在 2 维以上空间未做推广, 因此在数据集 Swiss-Roll 和 S-Curve 上只对 K-means、Geodesic K-means 以及本文算法进行对比. 另外, 由于 3 维及 3 维以上观测空间中, 可以求取曲面复杂度最小方向, 因此, 为考察曲面复杂度最小方向作为搜索启发信息时对聚类质量的提升效果, 实验同时对具有和缺乏此信息时的本文算法结果做了对比.

图 9 所示第 1 行和第 3 行, 从左至右 4 列分别为 K-means、Geodesic K-means 以及本文算法在无曲面复杂度最小方向作为启发信息和有此启发信息时的聚类结果, 其中 Swiss-Roll 上的聚类数为 8, S-Curve 上的聚类数为 5. 由于 Swiss-Roll 与 S-Curve 均与 2D 欧氏空间局部同胚, 因此为方便考察线性同胚簇的产生质量, 除给出观测空间中的聚类结果外, 本文将 Swiss-Roll 与 S-Curve 还原到与之同胚的 2D 欧氏空间中. 如图 10 所示, $A\sim H$ 分

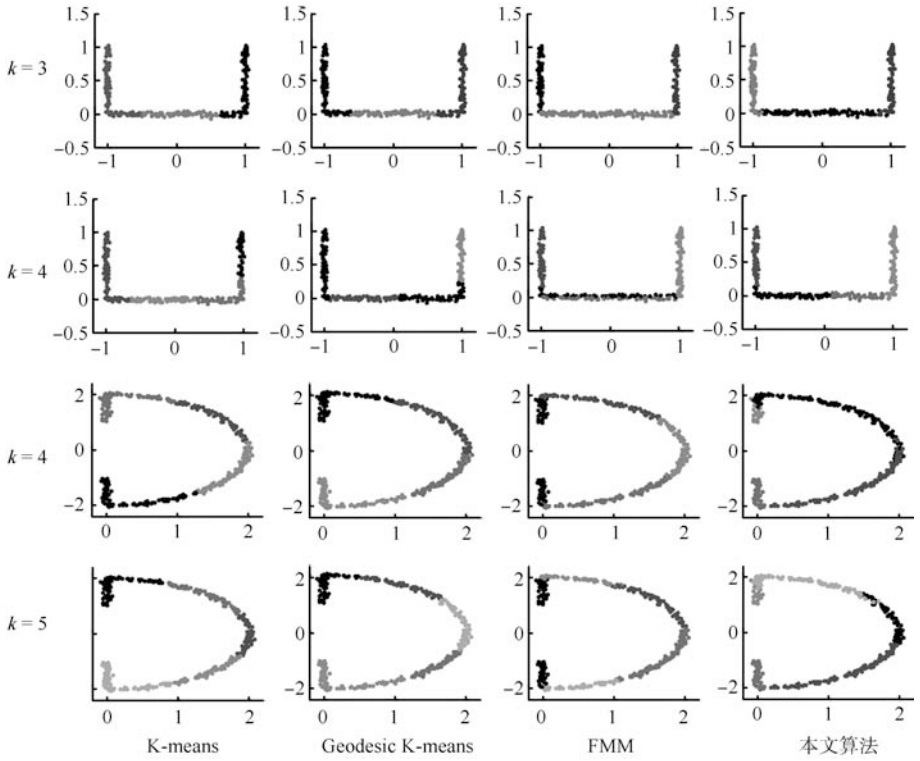


图 8　不同聚类数下 Rect-line 和 Open-D 上的线性结构挖掘结果

Fig. 8　The results of the linear structure mining on Rect-line and Open-D under different clusting numbers

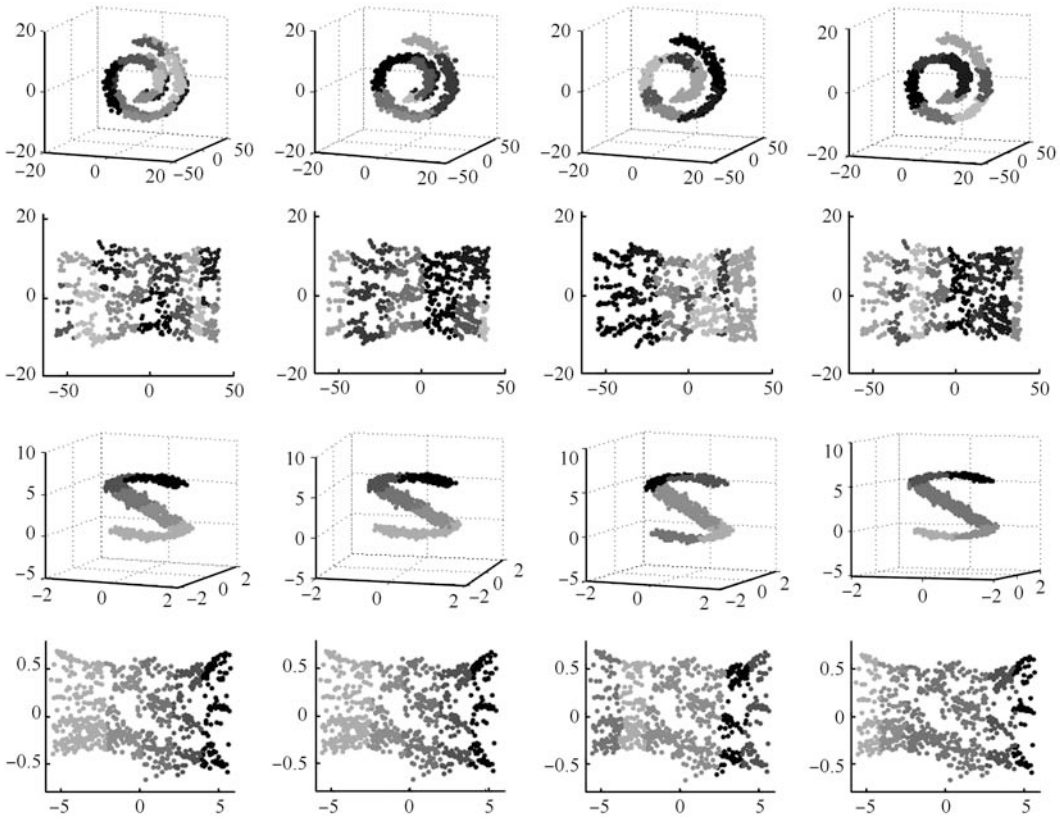


图 9 Swiss-Roll 和 S-Curve 上的线性结构挖掘结果

Fig. 9 The results of the linear structure mining on Swiss-Roll and S-Curve

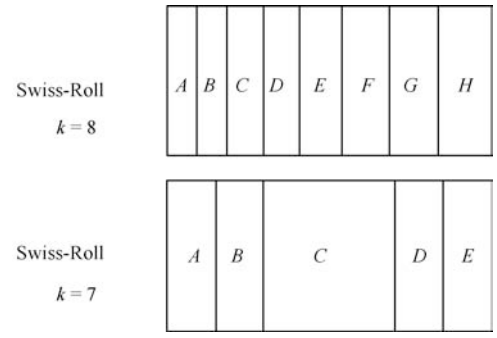


图 10 理想线性结构在 2D 欧氏空间中的映射

Fig. 10 The mapping of ideal linear structure in the 2D Euclidean space

别代表在线性结构聚类挖掘的理想结果下, 原数据集的不同线性结构在 2D 欧氏空间中的映射, 其中竖直方向代表原观测空间中的 y 轴方向, 由于 Swiss-Roll 和 S-Curve 均在观测空间 y 轴方向上体现出最佳的线性特性, 故在利用 2D 映射作为聚类效果的直观验证方法时, 理想聚类情况下 2D 映射竖直方向上数据点所属线性结构的类别应保持不变. 对于 Swiss-Roll, 簇 A 对应观测空间中最中心位置处的线性结构, H 对应数据卷最外部的线性结构, 而对于 S-Curve, 簇 A、B、D 所对应的线性结构规模相等, 约为簇 C 线性结构的一半. 对于实际聚类结果, 本文采用 ISOMAP 算法[7] 将其映射至 2D 欧氏空间, 如图 9 中第 2 行与第 4 行, 通过将其与图 10 理想聚类映射对比, 可直观地体现出各算法的优劣.

在 Swiss-Roll 数据集上, K-means 类算法产生的线性结构在结构内无法连通, 并且在 y 轴方向上不具备线性结构类别的一致性, 对应在观测空间中体现为近似线性结构的非线性程度较大; Geodesic K-means 的表现优于 K-means, 近似线性结构内保持连通, 但在 y 轴方向体现出的结构类别一致性仍较差, 即所挖掘的近似线性结构线性程度较低; 本文算法在没有曲面复杂度最小方向作为启发信息时, 所得解对应的线性结构质量高于 Geodesic K-means, 但其线性程度仍需提高, 从图 9 中第 2 行第 3 列子图中可以看到, 中间仍有两不同簇分布在 y 轴方向上, 这意味着其他线性结构的非线性程度被抬高. 这一问题在使用曲面复杂度最小方向作为启发信息时得到了很好的解决, 聚类结果接近理想聚类挖掘划分结果. 在 S-Curve 数据集上, K-means 与 Geodesic K-means 聚类结果相似, 在 y 轴方向上均有良好的线性结构类别一致性, 这主要是因为 S-Curve 在 y 轴方向上的分布在 $-0.5 \sim 0.5$ 区间范围, 相比其他维度较小. 但两种算法所得结构在观测

空间的线性程度仍然较差, 从图 9 中第 3 行第 1 列和第 2 列两个子图可以发现, 有一个线性结构同时包含了 S 型第一个折角处上下两部分的数据点, 而这些数据点在理想聚类中分属不同的线性结构. 除此之外, 斜坡部分为一本质线性结构, 在聚类数比较小的情况下, 算法应优先拆分此本质线性结构之外的非线性部分, 但 K-means 与 Geodesic K-means 均将此本质线性结构拆分为两部分. 本文算法在启发信息有无时均获得了质量较高的线性结构聚类挖掘结果, 而在有曲面复杂度最小方向作为启发信息时获得的结果更接近理想的线性结构挖掘结果.

表 1 所示为获得理想聚类结果下各个数据集上的蚁群参数设置情况. 由表 1 可知, 本文算法对蚂蚁个数与信息素保留系数并不敏感, 对于 Open-D、Swiss-Roll、以及 S-Curve 三个数据集, 蚂蚁个数设置为 50 及以上时均能收敛至理想聚类结果, 而信息素保留系数只要设置在 0.5～0.9 区间内即可, 并且由于 Rect-line 数据集相对简单, 获得理想聚类结果时需要的蚂蚁个数下限更低. 蚁群聚类目标函数中涉及两个权系数 α 和 β, 分别代表类测地相似度与 Grassmann 测地相似度的重要程度, 由表 1 所示实验所设参数可知, 相对蚂蚁个数与信息素保留系数, 算法对 α 表现敏感. α 的经验取值区间为 0 ～ 20, 但对于 Open-D 和 Swiss-Roll 数据集, 其能获得理想聚类结果的参数变化区间长度仅为 2. β 取值区间的上限虽然有所波动, 但相比 α, 仍然体现出良好的区间一致性.

4.2　实际数据集实验分析

一般认为, 人脸图像采集自流形空间, 因此除合成数据集外, 本文在 LLE 实际人脸数据集[28] 上进行了算法实验. LLE 人脸数据集是对同一对象在相同光照条件、不同表情和姿态变化下采集获得, 用于表情与姿态变化下的人脸流形建模与分析, 共包含 1 965 个样本, 每个人脸样本均为分辨率为 28 像素 × 20 像素的单通道图像. 本文选取包含了正脸、左向侧脸、右向侧脸三种姿态变化, 以及普通表情、嘴部动作 (撅嘴及吐舌)、少量眼部动作 (闭眼或挤眉) 三类面部表情变化, 共 600 幅样本构成实验数据集, 如图 11 所示. 实验对本文算法以及 K-means 聚类算法在不同聚类数下进行了对比分析, 其中本文算法蚁群参数设置为: 蚂蚁个数 150, 信息素保留系数 0.8, 目标函数权系数 $\alpha = 5.0$, 目标函数权参数 $\beta = 1.0$.

图 11　LLE 人脸样本

Fig. 11　The samples of LLE dataset

图 12 所示为两种算法在挖掘数为 3 时的聚类结果, 其中位于上方的 3 个子图为本文算法线性结构挖掘结果, 下方 3 个子图为 K-means 算法的聚类结果. 可以看出, 本文算法将人脸数据集划分为三个线性结构簇, 总体上可以分为正面人脸、左向侧脸以及右向侧脸三个类别, K-means 算法也获得了类似的划分结果. 图 11 所示为两种算法在挖掘数为 4 时的结果, 上方 4 个子图为本文算法结果, 下方 4 个子图为 K-means 算法结果. 本文算法的 4 个结果簇分别为正面人脸、右向无嘴部动作侧脸、左向无嘴部动作侧脸、以及嘴部动作, 而 K-means 算法的 4 个簇分别为正面姿态、正面姿态、右向侧脸、以及左向侧脸.

在 LLE 人脸数据集中, 眼部动作带来的维度变化较为微弱, 而嘴部动作所产生样本变化较大, 因此所有无嘴部动作的正面姿态样本不会体现出剧烈的分布变化, 在挖掘数为 3 时, 本文算法所得结果将其划分在同一线性结构簇中, K-means 算法也将其归为同一簇中. 而当挖掘数为 4 时, 正面人脸簇在本文

表 1　获得理想聚类结果的蚁群参数

Table 1　Parameters of ant colony optimization with acceptable clustering results

数据集	蚂蚁个数	信息素保留系数	目标函数权参数 α	目标函数权参数 β
Rect-line	> 30	0.5～0.9	1.0～14.0	0.1～10.0
Open-D	> 50	0.5～0.9	3.0～5.0	0.1～3.5
Swiss-Roll	> 50	0.5～0.9	11.0～13.0	0.2～6.5
S-Curve	> 50	0.5～0.9	5.0～12.0	0.2～7.0

算法的结果中基本保持不变, 但 K-means 算法将该簇拆分为 2 个, 从 K-means 算法结果可以看出, 拆分开的两簇并没有显著性的表情或姿态的差别. 另外, 在人脸流形上, 嘴部动作与姿态变化一样, 将使流形产生扭曲. 在挖掘数为 3 时, 本文算法与 K-means 算法一样, 并未检测到此变化. 可以认为, 相对于姿态变化, 嘴部动作带来的表情变化所产生的流形扭曲程度相对较小, 在线性簇划分过程中, 较小的挖掘数下本文算法倾向于捕获较大的分布变化. 通过增加挖掘数, 在挖掘数为 4 时, 本文算法将嘴部动作作为单独一簇划分出来, 成功捕获到了这一变化, 而 K-means 算法却仍将嘴部动作与人脸侧向姿态的变化混在一起. 由上述实验结果及分析可知, 本文算法在揭示流形变化方面体现出了 K-means 算法所不具备的新特性.

5 结论

本文针对在观测空间呈非线性分布, 且内蕴维度低于观测维度的数据流形的线性结构聚类挖掘问题作了研究, 提出了基于蚁群算法与 Grassmann 流形的聚类求解算法. 通过将聚类处理的最小单元从单一样本点提升为子集, 使得噪声的影响降低, 并利用 Grassmann 流形测地距离度量聚类单元的相似度, 同时设计了单元类测地距离保证聚类所得簇的连通性, 定义了曲面复杂度最小方向以提高蚁群算法所产生线性同胚簇划分的质量. 在多个数据集上的实验表明了算法在非线性数据流形线性结构聚类挖掘方面的可行性. 在算法对于一般数据流形的推广性方面, 由于均匀密集的采样有助于算法获得更佳的结果, 故当数据集采样非均匀, 或者样本数量远小于观测空间维度, 使得在观测空间中的数据分布过于稀疏时, 算法性能则可能下降, 因此根据实际数据集的特点对算法做出改进, 使得算法能够应对高度非均匀采样以及数据稀疏性的问题, 并利用线性结构挖掘结果构建非线性流形的显式表达, 从而实现对未知流形结构数据的建模, 是进一步工作的重点.

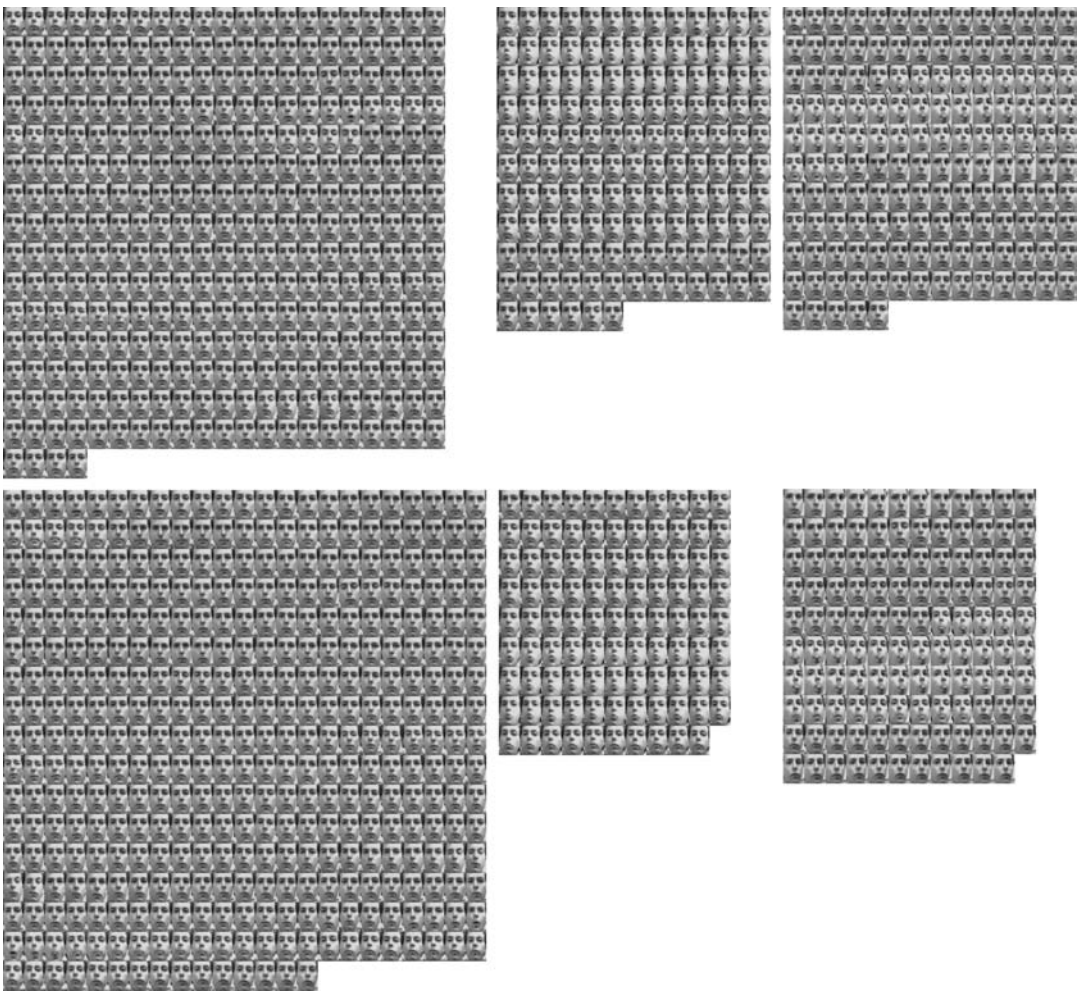


图 12 挖掘数为 3 时本文算法和 K-means 在 LLE 人脸数据集上的实验结果

Fig. 12 The result of the proposed method and K-means on LLE dataset when the clustering number is 3

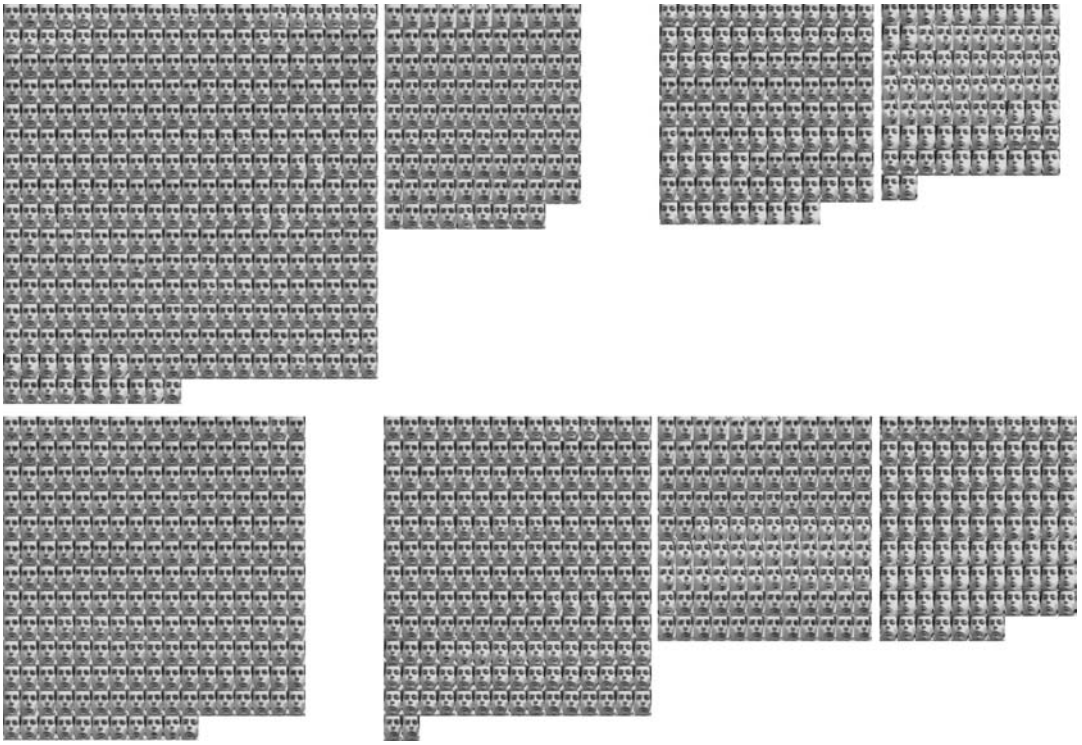


图 13 挖掘数为 4 时本文算法和 K-means 在 LLE 人脸数据集上的实验结果

Fig. 13 The result of the proposed method and K-means on LLE dataset when the clustering number is 4

References

1 Seung H S, Lee D D. The manifold ways of perception. *Science*, 2000, **290**(5500): 2268−2269

2 Lin T, Zha H B. Riemannian manifold learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(5): 796−809

3 Tenenbaum J B, Silva V D, Langford J C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 2000, **290**(5500): 2319−2323

4 Yan De-Qin, Liu Sheng-Lan, Li Yan-Yan. An embedding dimension reduction algorithm based on sparse analysis. *Acta Automatica Sinica*, 2011, **37**(11): 1306−1312
(闫德勤, 刘胜蓝, 李燕燕. 一种基于稀疏嵌入分析的降维方法. 自动化学报, 2011, **37**(11): 1306−1312)

5 Li Guang-Wei, Liu Yun-Peng, Yin Jian, Shi Ze-Lin. Planar object recognition based on riemannian manifold. *Acta Automatica Sinica*, 2010, **36**(4): 466−474
(李广伟, 刘云鹏, 尹健, 史泽林. 基于黎曼流形的平面目标识别. 自动化学报, 2010, **36**(4): 466−474)

6 Liu Sheng-Lan, Yan De-Qin. A new global embedding algorithm. *Acta Automatica Sinica*, 2011, **37**(7): 828−835
(刘胜蓝, 闫德勤. 一种新的全局嵌入降维算法. 自动化学报, 2011, **37**(7): 828−835)

7 Weinberger K Q, Sha F, Saul L K. Learning a kernel matrix for nonlinear dimensionality reduction. In: Proceedings of 21st International Conference on Machine Learning. Banff, Canada: Association for Computing Machinery, 2004. 839−846

8 Zhang S W, Lei Y K. Modified locally linear discriminant embedding for plant leaf recognition. *Neurocomputing*, 2011, **74**(14−15): 2284−2290

9 Yang W K, Sun C Y, Zhang L. A multi-manifold discriminant analysis method for image feature extraction. *Pattern Recognition*, 2011, **44**(8): 1648−1657

10 Zhang J P, Wang X D, Krger U, Wang F Y. Principal curve algorithms for partitioning high-dimensional data spaces. *IEEE Transactions on Neural Networks*, 2011, **22**(3): 367−380

11 Zhang J P, Huang H, Wang J. Manifold learning for visualizing and analyzing high-dimensional data. *IEEE Intelligent Systems*, 2010, **25**(4): 54−61

12 Xiang S M, Nie F P, Zhang C S, Zhang C X. Nonlinear dimensionality reduction with local spline embedding. *IEEE Transactions on Knowledge and Data Engineering*, 2009, **21**(9): 1285−1298

13 Xiang S M, Nie F P, Pan C H, Zhang C S. Regression reformulations of LLE and LTSA with locally linear transformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 2011, **41**(5): 1250−1262

14 Chang H, Yeung D Y. Robust locally linear embedding. *Pattern Recognition*, 2006, **39**(6): 1053−1065

15 Wang J, Zhang Z, Zha H. Adaptive manifold learning. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2005, **17**: 1473−1480

16 Li G G, Wang Z Z, Wang X M, Ni Q S, Qiang B. Linear manifold clustering for high dimensional data based on line manifold searching and fusing. *Journal of Central South University of Technology*, 2010, **17**(5): 1058−1069

17 Shelokar P S, Jayaraman V K, Kulkarni B D. An ant colony approach for clustering. *Analytica Chimica Acta*, 2004, **509**(2): 187−195

18 Asgharbeygi N, Maleki A. Geodesic k-means clustering. In: Proceedings of 19th International Conference on Pattern Recognition. Tamp, USA: IEEE, 2008. 3450−3453

19 Ma Jiang-Hong, Ge Yong. The finite mixture model and its EM algorithm for line-type image patterns. *Chinese Journal of Computers*, 2007, **30**(2): 288−296
(马江洪, 葛咏. 图像线状模式的有限混合模型及其 EM 算法. 计算机学报, 2007, **30**(2): 288−296)

20 Camastra F. Data dimensionality estimation methods: a survey. *Pattern Recognition*, 2003, **36**(12): 2945−2954

21 Edelman A, Arias T A, Smith S T. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 1998, **20**(2): 303−353

22 Armstrong M A. *Basic Topology*. New York: Springer-Verlag, 1997. 43−51

23 Amsallem D, Farhat C. Interpolation method for adapting reduced-order models and application to aeroelasticity. *American Institute of Aeronautics and Astronautics Journal*, 2008, **46**(7): 1803−1813

24 Misra J. A walk over the shortest path: Dijkstra′s algorithm viewed as fixed-point computation. *Information Processing Letters*, 2001, **77**(2−4): 197−200

25 Karcher H. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 1977, **30**(5): 509−541

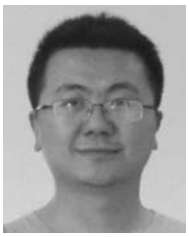
26 Pennec X, Fillard P, Ayache N. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 2006, **66**(1): 41−66

27 Tuzel O, Porikli F, Meer P. Pedestrian detection via classification on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, **30**(10): 1713−1727

28 Roweis S T. LLE face dataset [Online], http://www.cs.nyu.edu/~roweis/data/frey_rawface.mat, December 19, 2011

王 力 东北大学信息科学与工程学院博士研究生. 2009 年获东北大学模式识别与智能系统专业硕士学位. 主要研究方向为模式识别与图像处理. 本文通信作者. E-mail: wl1986_ren_ren@163.com
(**WANG Li** Ph. D. candidate at the School of Information Science and Engineering, Northeastern University. He received his master degree from Northeastern University in 2009. His research interest covers pattern recognition and image processing. Corresponding author of this paper.)

吴成东 东北大学教授. 主要研究方向为图像处理, 模式识别, 无线传感器网络. E-mail: wuchengdong@ise.neu.edu.cn
(**WU Cheng-Dong** Professor at Northeastern University. His research interest covers image processing, pattern recognition, and wireless sensor networks.)



陈东岳 东北大学副教授. 主要研究方向为模式识别, 计算机视觉.
E-mail: chendongyue@ise.neu.edu.cn
(**CHEN Dong-Yue** Associate professor at Northeastern University. His research interest covers pattern recognition and computer vision.)

李孟歆 沈阳建筑大学教授. 主要研究方向为数据挖掘与模式识别.
E-mail: limengxinf1972@yahoo.com.cn
(**LI Meng-Xin** Professor at Shenyang Jianzhu University. Her research interest covers data mining and pattern recognition.)

陈 莉 沈阳建筑大学教授. 主要研究方向为智能计算与建筑智能化.
E-mail: wuhaoli@online.ln.cn
(**CHEN Li** Professor at Shenyang Jianzhu University. Her research interest covers intelligent computing and intelligent building.)