DOI: 10.13382/j.jemi.2017.03.002

基于 Grassmann 流形的谱聚类分析算法*

谢英红[1,2] 何宇清[1] 王 楠[2]
(1. 天津大学 电子信息工程学院 天津 300072; 2. 沈阳大学 信息工程学院 沈阳 110044)

摘 要:在标准谱聚类分析算法中,基于欧氏空间的度量不能完全反映数据集合复杂的空间分布特性,导致聚类结果不够准确。而使用流形空间能够更准确的描述数据之间的几何结构关系。在基于规范化拉普拉斯矩阵的谱聚类算法基础上,研究 Grassmann 流形的光滑曲面的空间表达方式,应用适合度量数据点之间距离的特性,提出基于 Grassmann 距离度量的改进的谱聚类分析算法,在流形空间上分析待聚类数据点之间的相似性。实验结果表明,该算法不仅能够对分布在相同或不同子空间上的数据进行有效聚类,而且能够对具有复杂几何结构的数据集合进行分析,在流形空间上进行有效聚类。

关键词: 聚类分析; Grassmann 流形; 谱聚类; 距离度量; 流形空间
中图分类号: TP391;TH89 文献标识码: A **国家标准学科分类代码**: 520.20

Spectrum clustering analysis algorithm based on Grassmann manifold

Xie Yinghong[1,2] He Yuqing[1] Wang Nan[2]
(1. School of Electronic Informatin Engineering, Tianjin University, Tianjin 300072, China;
2. School of Information Engineering, Shenyang University, Shenyang 110044, China)

Abstract: In the standard spectrum clustering algorithm, the metric based on Euclidean space cannot represent the complicate space distribution feature of some data set, which might lead to the clustering result inaccuracy. While the geometric relationship between data can be described more precise by manifold space. The special expression on curved surface is researched, the feature which is more fit for measuring the distance between data is applied, and an improved spectrum clustering analysis algorithm based on the distance metric under Graasmann manifold is proposed. The similarity between data is analyzed under manifold space. The experimental results show that the proposed algorithm can cluster data set either belonging the same or different subspace more accurately, furthermore, it can cluster data set with more complicate geometric structure under manifold space efficiently.

Keywords: clustering analysis; Grasmann manifold; spectrum clustering; distance metric; manifold space

1 引 言

聚类分析是现代数据分析的基础。传统的 K 均值聚类和混合模型聚类分析方法,在待分析的数据集合符合假设的模型结构的情况下,能够取得较精确的分类效果[1-2]。但是当数据结构较复杂时,这些方法的分类效果不理想。通常使用谱聚类分析的方法来处理复杂数据聚类问题。典型算法有 Ng 等人提出的基于规范化拉普拉斯矩阵的谱聚类(NJW)算法,也称标准谱聚类算法[3-6]。但是这类传统的谱聚类分析算法[3-9]中,局部流形拓扑结构是基于欧氏距离构建的,可能存在局部流形的拓扑结构置乱现象。

流形学习算法是近些年发展起来的降维方法,它的目的是从高维数据中找到重要的低维结构,流形学习算法在人脸识别、交通标志识别等方面应用广泛[10-15]。考虑到 Grassmann 流形是李群流形中的一种熵流形,不仅具有光滑曲面的空间表达方式,且具有更为适合度量数

收稿日期:2016-04 Received Date: 2016-04
*基金项目:国家自然科学基金(61503274,61603415)、辽宁省博士启动基金(201501090)资助项目

据点之间距离的特性,本文在研究 NJW 算法基础上,提出一种基于 Grassmann 流形的数据聚类分析方法,该方法在 Grassmann 流形空间上比较数据点之间的相似性,可以对分别存在于独立子空间或是子空间不独立情况下的数据结合进行有效聚类。同时能够对流形空间数据集合进行有效聚类。

2　Grassmann 流形及其度量

Grassmann 流形 $Gr(k,n)$ 上的点是 $n \times k$ 正交矩阵等价类集合。即:

$$Gr(k,n) = |\boldsymbol{Y}|_{O_k} = \{\boldsymbol{YV}:\boldsymbol{V} \in O_k\} \tag{1}$$

式中:$\boldsymbol{Y}$ 是 $n \times k$ 正交矩阵,$|\boldsymbol{Y}|$ 代表等价关系,$\boldsymbol{V}$ 是 $k \times k$ 正交矩阵。

Grassmann 流形 $Gr(k,n)$ 也可以表示为 n 维向量空间 R^n 中所有的 k 维子空间的集合。Grassmann 流形具有商空间表示形式 $Gr(n,k) = O(n)/(O(n-k) \times O(k))$,可以理解为正交李群 $O(n)$ 中"除去"那些共面的旋转和非共面的旋转剩下的部分。

在流形 M 上赋予度量结构的通常办法是对于每一点 $p \in M$ 的切空间 T_pM 上指定了一个内积$\langle \cdot,\cdot \rangle$,也称黎曼度量。对任一点 $\boldsymbol{p} \in Gr(k,n)$,切空间为:

$$T_pGr(k,n) = \{\boldsymbol{\omega}|\boldsymbol{\omega} = \boldsymbol{p}_\perp \boldsymbol{g},\boldsymbol{g} \in R(n-k,k)\} \tag{2}$$

$\boldsymbol{p}_\perp$ 为 $\boldsymbol{p}$ 的正交补。$Gr(k,n)$ 上的度量定义为:

$$\|\boldsymbol{\omega}\| = Tr(\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\omega}) \tag{2}$$

设 $\gamma:t \Rightarrow \gamma(t)$ 为初始点 $\gamma(0)$、初速度为 $\frac{\mathrm{d}\gamma}{\mathrm{d}t}(0) = \boldsymbol{\omega}$ 的测地线,指数映射 $\mathrm{Exp}_p(\boldsymbol{\omega}) = \gamma(1)$ 给出了测地线的终点:

$$\mathrm{Exp}_p(\boldsymbol{\omega}) = \boldsymbol{pV}\cos(\theta) + \boldsymbol{U}\sin(\theta) \tag{4}$$

式中:$\boldsymbol{U}\theta\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{SVD}(\boldsymbol{\omega})$。相应的逆映射 $\mathrm{Log}_p(\boldsymbol{q}) = \boldsymbol{U}\boldsymbol{\theta}\boldsymbol{V}^{\mathrm{T}}$,这里的 $\theta = \arctan(\boldsymbol{S})$,$\boldsymbol{USV}^{\mathrm{T}} = \boldsymbol{p}_\perp \boldsymbol{p}_\perp^{\mathrm{T}} \boldsymbol{q}(\boldsymbol{p}^{\mathrm{T}}\boldsymbol{q})^{-1}$。因此,Grassmann 流形上两点 (p,q) 的测地距离为:

$$d_c(\boldsymbol{p},\boldsymbol{q}) = \left(\sum_{i=1}^{k}\theta_i^2\right)^{\frac{1}{2}} = \|\theta\|_2 \tag{5}$$

其中 p 和 q 为 Grassmann 流形上的两个点,p 和 q 之间的主角度为 $\theta_1,\cdots,\theta_k$,

计算 Grassmann 流形两点 p 和 q 的距离 $d(p,q)$ 步骤如下。

输入:矩阵 $\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$

输出:p 和 q 之间测地距离

1)计算 $\boldsymbol{Y}_1$ 和 $\boldsymbol{Y}_2$ 的正交基 $\boldsymbol{Q}_1$ 和 $\boldsymbol{Q}_2$,使得 $\boldsymbol{p} = |\boldsymbol{Y}_1| = |\boldsymbol{Q}_1|$,$\boldsymbol{q} = |\boldsymbol{Y}_2| = |\boldsymbol{Q}_2|$;

2)计算 $\boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{Q}_2$ 的奇异值分解 $\boldsymbol{USV}^{\mathrm{T}} = \boldsymbol{SVD}(\boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{Q}_2)$;

3)计算主角度 $\boldsymbol{\theta} = \cos^{-1}\boldsymbol{S}$,$d(p,q) = \sqrt{\sum_{i=1}^{k}\boldsymbol{\theta}_i^2}$。

3　基于 Grassmann 流形的谱聚类分析算法

3.1　改进的谱聚类算法

本文考虑到 NJW 算法使用的谱聚类分析的局部流形拓扑结构是基于欧氏距离构建的,可能存在局部流形的拓扑结构置乱现象。为了能使谱聚类算法对不同结构的数据具有较好的聚类精度,我们在原有谱聚类思想的基础上,提出了一种基于 Grassmann 流形距离度量的谱聚类算法,从而提高聚类的准确性。所述方法按如下步骤进行。

1)输入 n 个数据点 $\{x_i\}_{i=1}^{n}$,待聚类数目 k。

2)基于 Grassmann 流形上两点之间的距离公式 $d_G(p,q) = \left(\sum_{i=1}^{k}\boldsymbol{\theta}_i^2\right)^{1/2} = \|\boldsymbol{\theta}\|_2$,计算数据点之间的距离,构造相似性矩阵 $\boldsymbol{S} \in R^{n\times n}$。

其中 p 和 q 为 Grassmann 流形上的两个点,p 和 q 之间的主角度为 $\theta_1,\cdots,\theta_k$。

3)构造拉普拉斯矩阵 $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2}$,其中 $\boldsymbol{D}$ 为对角矩阵,$D_{ii} = \sum_{j=1}^{n}S_{ij}$。

4)求拉普拉斯矩阵 $\boldsymbol{L}$ 的 k 个最大特征值对应的特征向量 $v_1,v_2,\cdots,v_k$,并且构造矩阵 $\boldsymbol{V} = [v_1,v_2,\cdots,v_k] \in R^{n\times k}$,其中 v_k 为列向量。

5)单位化 $\boldsymbol{V}$ 的行向量,得到矩阵 $\boldsymbol{Y}$,其中 $\boldsymbol{Y}_{ij} = \boldsymbol{V}_{ij}/\left(\sum_{j}\boldsymbol{V}_{ij}^2\right)^{1/2}$。

6)将 $\boldsymbol{Y}$ 的每一行看成是 R^k 空间内的一点,使用 K 均值算法对其进行分类。

7)如果 $\boldsymbol{Y}$ 的第 i 行属于第 j 类,则将原数据点 x_l 也划分到第 j 类输出数据点的划分 $c_1,c_2,\cdots,c_k$。

3.2　核参数确定

同标准的谱聚类算法相同,本文采用高斯核函数做为相似性距离度量。不同的是,本文在更为准确的 Grassmann 流形上进行元素间的距离计算,用于评估相似度。本文所采用的核函数如下:

$$K_{ij} = \exp\left(\frac{-D_{ij}}{2\sigma_1\sigma_2}\right) \tag{6}$$

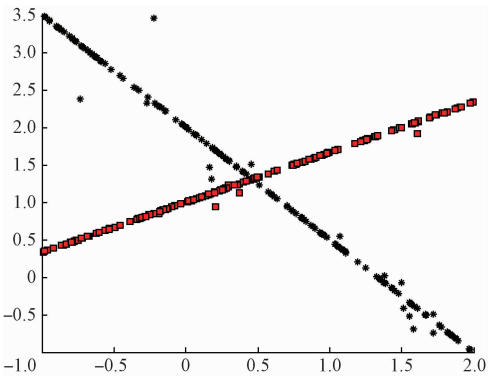
核宽度:

$$\sigma_i = d(x_i,x_{il}) \tag{7}$$

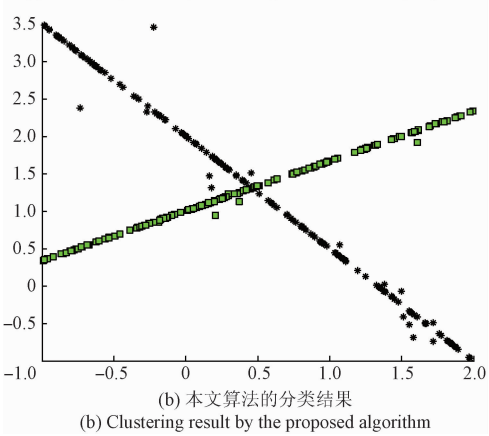
式中:$d(x_i,x_{il})$ 函数由式(5)得到,x_{il} 是 x_i 的第个 l 邻接点。σ_i 随着近邻分布而自适应变化,确保样本集内同类间具有较大的相似度,而不同类的样本具有较小的相似度。

4 实验结果与分析

为验证算法的有效性，将所提算法与标准的谱聚类算法进行比较。第 1 组实验，对于子空间独立情况下的点集进行分类，本例中，将其分为两类，两种算法的分类结果相同，如图 1 所示。



(a) 标准谱聚类算法分类结果
(a) Clustering result bystandard spectrum clustering algorithm

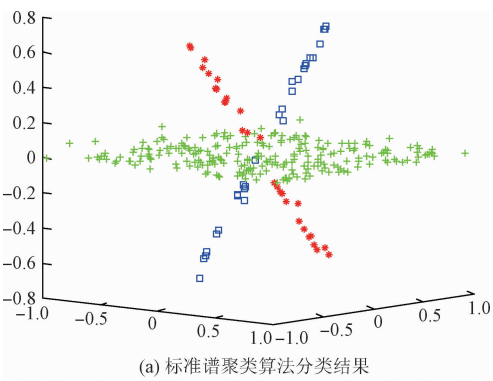


(b) 本文算法的分类结果
(b) Clustering result by the proposed algorithm

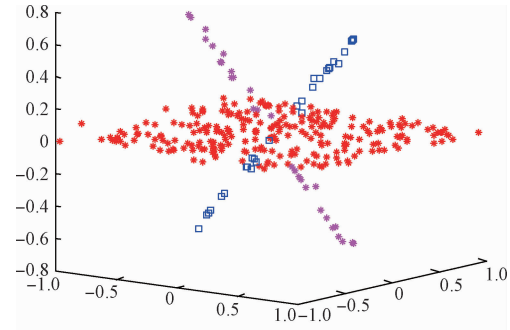
图 1 独立子空间数据集分类结果

Fig. 1 Clustering result for data set in independent subspace

第 2 组实验，选取不独立子空间的样本集，验证算法的有效性。手动设定分类数为 3 类，两种分类算法的结果相同，都能够对样本集正确分类。结果如图 2 所示。



(a) 标准谱聚类算法分类结果
(a) Clustering result by standard spectrum clustering algorithm

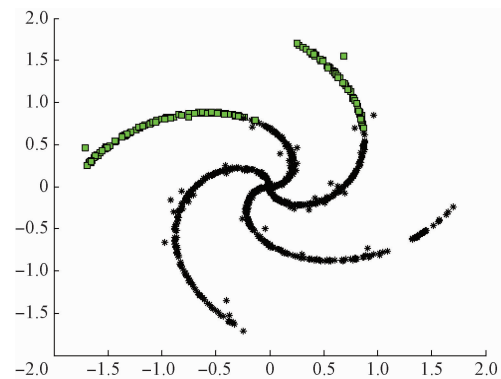


(b) 标准谱聚类算法分类结果
(b) Clustering result by standard spectrum clustering algorithm

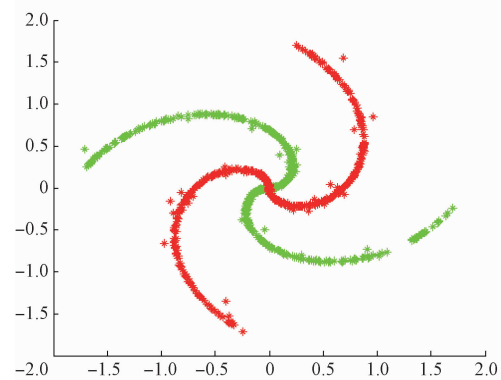
图 2 不独立子空间数据集聚类结果

Fig. 2 Clustering result for data set in dependent subspace

第 3 组实验，验证算法对流形空间数据分类的有效性，手动确定分类数为 2 类。两种方法的实验结果如图 3 所示。可以看出，标准的谱聚类算法分类失误，而本文所提算法，能够对流形空间数据进行准确分类。



(a) 标准谱聚类算法分类结果
(a) Clustering result by standard spectrum clustering algorithm



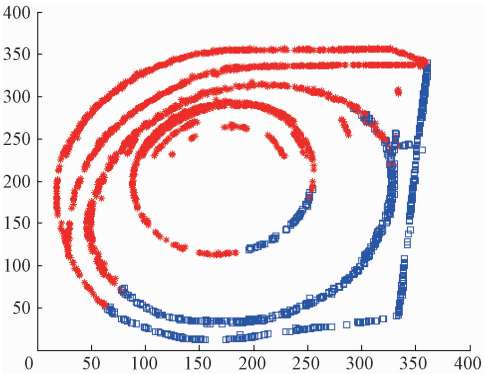
(b) 本文算法的分类结果
(b) Clustering result by the proposed algorithm

图 3 流形空间数据集聚类结果

Fig. 3 Clustering result for data set in manifold space

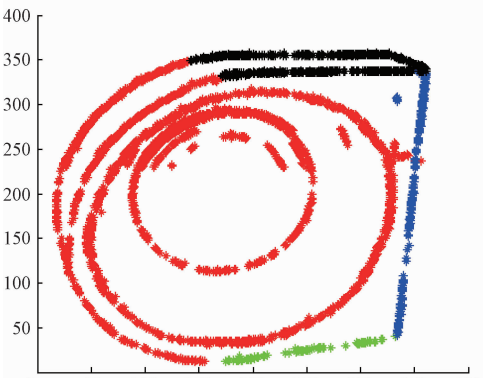
第 4 组实验，验证算法对于复杂流形空间数据集分类的有效性。本实验中，分类数据由算法自行判断。两种算法的分类结果如图 4 所示。可以看出，标准的谱聚

类算法将该流形数据集合分为 2 类。本文算法分为 4 类。但是从图中的分类结果可以判断,本文的分类算法具有较高的准确性。



(a) 标准谱聚类算法分类结果

(a) Clustering result by standard spectrum clustering algorithm



(b) 本文算法的分类结果

(b) Clustering result by the proposed algorithm

图 4　流形空间数据集聚类分类结果

Fig. 4　Clustering result for data set in manifold space

综上所述,该跟踪方法不仅能够对分布在不同子空间上的数据进行有效聚类,而且能够对具有复杂几何结构的数据集合进行分析,在流形空间上进行有效聚类。

5　结　论

在标准谱聚类分析算法中,基于欧氏空间的度量不能完全反映数据聚类复杂的空间分布特性,导致聚类结果不够准确。而使用流形空间能够更准确的描述数据之间的几何结构关系。考虑到 Grassmann 流形是李群流形中的一种熵流形,不仅具有光滑曲面的空间表达方式,且具有更为适合度量数据点之间距离的特性,可以使聚类结果更加准确,本文提出基于 Grassmann 距离度量的改进的谱聚类分析算法,能够更准确地计算数据点间的距离,提高聚类的准确性。可以对分别存在于独立子空间或是子空间不独立情况下的数据结合进行有效聚类。能够对流形空间数据集合进行有效聚类。

参考文献

[1] JAIN A K. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666.

[2] 谢颖. MeanShift 和聚类算法的服装图像分割[J]. 电子测量技术,2013,36(8):53-60.
XIE Y. Segmantation method combined with MeanShift and K-mean clustering algorithm for clothing image[J]. Electronic Measurement Technology, 2013, 36(8): 53-60.

[3] CHI Y, SONG X D. On evolutionary spectral clustering[J]. ACM Transactions on Knowledge Discovery from Data, 2009,3(4):17-47.

[4] 王玲,薄列峰,焦李成. 密度敏感的半监督谱聚类[J]. 软件学报,2007,18(10):1214-1225.
WANG L, BO L W, JIAO L CH. Density-sensitive semi-supervised spectral clustering[J]. Journal of Software, 2007, 18(10): 2412- 2422.

[5] 周文刚, 陈雷霆,董仕. 基于谱聚类的网络流量分类识别算法[J]. 电子测量与仪器学报,2013,27(12): 1114-1119.
ZHOU W G, CHEN L T, DONG SH. Network traffic classification algorithm based on spectral clustering[J]. Journal of Electronic Measurement and Instrument, 2013,27(12): 1114-1119.

[6] CHANG H, YEUNG D Y. Robust path-based spectral clustering[J]. Pattern Recognition, 2008, 41(1): 191-203.

[7] VIDAL R. Subspace clustering[J]. IEEE Signal Processing Magazine, 28(2):52-68, 2011.

[8] ELHAMIFAR E, VIDLA R. Sparse subspace clustering, algorithm, theory, and applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013,35(11):2765-2781.

[9] WANG Y, JIANG Y, WU Y, et al. Spectral clustering on multiple manifolds[J]. IEEE Transactions on Neural Networks, 2011,22(7):1149-1161.

[10] 刘云鹏, 李广伟, 史泽林, 基于 Grassmann 流形的仿射不变形状识别[J]. 自动化学报, 2012, 38(2): 248-258.
LIU Y P, LI G W, SH Z L. Projective registration algorithm based on Grassmann manifold[J]. Acta Automatica Sinica, 2009,35(11): 1378-1386.

[11] ELHAMIFAR E, VIDAL R. Sparse manifold clustering and embedding[C]. Proceedings of the 25th Annual Conference on Neural Information Processing Systems, Sierra Nevada,2011: 55-63.

[12] 谢英红,吴成东. 基于投影群和协方差流形双重建模

的目标跟踪[J]. 仪器仪表学报, 2014, 35(2): 374-379.
XIE Y H, WU CH D. Object tracking with dual modeling based on projection group and covariance manifold[J]. Chinese Journal of Scientific Instrument, 2014, 35(2): 374-379.

[13] GUO Q, WU C D, FENG Y, et al. Conjugate gradient algorithm for efficient covariance tracking with jensen-bregman logdet metric[J]. IET Computer Vision, 2015, 9(6): 814-820.

[14] 施晓东,刘格. 一种光学遥感图像海陆分割方法[J]. 国外电子测量技术, 2014, 33(11): 29-32.
SI X D, LIU G. Method of water and land segemantation in optical remote sensing images[J]. Forgein Electronic Measurement Technology, 2014, 33(11): 29-32.

[15] LU J, TAN Y P, WANG G. Discriminative multi-manifold analysis for face recognition from a single training sample per person[C]. IEEE International Conference on Computer Vision, 2011: 1943-1950.

作者简介

谢英红,1976年出生,2014年获得东北大学博士学位,现为沈阳大学副教授,在天津大学博士后工作站从事研究工作。主要研究方向为视频图像处理、模式识别等。
E-mail: xieyinghong@163.com

Xie Yinghong born in 1976, received Ph. D. from Northeastern University in 2014. Now she is an associate professor in Shenyang University, and works at the postdoctoral workstation at Tianjin University. Her main research interests include video image processing, and pattern recognition.