基于自相合性的主曲线的特性分析与研究*)

邵保军　苗夺谦　张红云　唐庆适　王　真
(同济大学计算机科学与工程系　上海200092)

摘　要　主曲线是穿过数据云"中间"的满足自相合性质的光滑曲线，它是线性主成分的非线性推广。在本文中，基于主曲线的自相合条件，构造一个微分方程，通过求解这个微分方程，我们得到了环形均匀分布的主曲线；通过对实验结果的分析与研究，我们得出了一些有趣的结论。

关键词　主曲线，流形，自相合性

Properties Analysis and Research of Principal Curves Based on Self-Consistency

SHAO Bao-Jun　MIAO Duo-Qian　ZHANG Hong-Yun　TANG Qing-Shi　WANG Zhen
(Department of Computer Science and Engineering, Tongji University, Shanghai 200092)

Abstract　A principal curve which satisfies the self-consistency property is a smooth curve passing through the "middle"of a dataset, which can be regard as a nonlinear generalization of linear principal components. In this paper, we construct a differential equation on the basis of the self-consistency condition of principal curves, and by solving this differential equation, we obtain principal curves for uniform densities on annuli; through analyses and studies of the experimentation's findings, we have drawn some interesting conclusions.

Keywords　Principal curves, Manifold, Self-consistency

1　引言

我们现在生活在一个信息化的时代，通信、计算机和网络技术正改变着整个人类和社会。与此同时，随着计算机应用技术特别是数据库技术的迅速发展以及数据库管理系统的广泛应用，人们积累了越来越多的数据，其中包括结构化数据、半结构化数据、分布在网络上的异构型数据等等。因此，数据具有高维、非线性、离散化程度高等特点。如何从高维观测数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的低维结构信息呢?用传统的多元线性分析技术来解决这个问题并不理想。于是在1983年，Stanford 大学统计系的 Hastie T. 以技术报告的形式发表了主曲线[1]的开创性论文"Principal curves and surfaces"。这篇论文指出：主曲线(Principal Curves)是线性主成分的非线性推广，即用光滑的曲线代替线性主成分(Principal Component)来概括数据，这类曲线恰好穿过数据云或者数据集合的"中间"，是一簇流形集合中满足自相合性(Self-consistency)的拟合数据集合的最优曲线。

主曲线的本质是嵌入 Euclid 空间、具有微分结构的一维流形[5](Manifold)，它直观地体现了数据集合内在结构信息的几何形态，实现了满足欧氏距离度量并且支持微分运算的赋范线性子空间之间的同胚转换(Homoeomorphism Transform)，并且在理论上回避了微分流形学在数学上的复杂性，从而可以在欧氏空间上直接建立等价的微分计算模型。在20世纪90年代后期，由于主曲线技术在数据分析(Data Analysis)领域中的逐步应用，及其本身所具有的多种优点，主曲线技术也引起计算机科学家的极大关注，在计算机科学领域的应用发展很快，如图像处理中辨识冰原轮廓[6]、手写字符的主曲线模板化[2]、手写字符的识别与分类[8]、指纹识别[9]、语音识别中对声音数据的约简建模和数据可听化[4]、生态学中寻找种群的有序分布、聚类和分类以及过程监控等。

2　主曲线的特性分析

主曲线在其理论本质上是一种非监督学习方法(Unsupervised Learning Method)，这种学习方法引入坐标邻域和局部坐标的概念，通过微分同胚建立起 Euclid 空间和微分流形(可以看作"局部的 Euclid 空间")这种拓扑结构之间的联系，力求实现信息保持和维数约简之间的平衡关系，因此，主曲线作为一种新的学习方法具有重要的理论研究价值。

2.1　主曲线及其自相合性

主曲线是线性主成分分析技术的非线性推广，它的研究目标和任务是在一个函数族中，寻求满足给定目标函数的最优函数解来表示数据集合，它的定义如下：

定义1　主曲线：如果光滑曲线 Γ 满足

(1)Γ 自身不相交；

(2)在任何 R^n 的边界子集内部，Γ 的长度有限；

(3)Γ 是自相合的，形式化为：

$$E(X|\lambda_f(X)=s)=f(s) \qquad (2.1.1)$$

则称 Γ 是数据分布的一条主曲线。

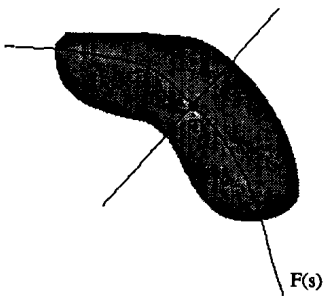

图1　主曲线上的每个点是投影至该点的所有数据点的条件均值

由条件(3)可知：主曲线上的每个点是所有投影至该点的

*)本文得到国家自然科学基金项目(No.60175016)资助.邵保军　硕士研究生，研究方向为计算机软件与理论，主曲线。

数据点的条件均值(图1),这一特性称为主曲线的自相合性。所谓自相合性,就是将数据集合的所有数据点按照某一度量投影到一个初始向量(一般是第一主成分线),引入坐标邻域的概念建立起数据集合的数据点与初始向量上的数据点的一一对应关系,然后求取投影到初始向量上某一点邻域内所有数据点的条件均值,从而得到新的曲线,并且根据约束条件,反复迭代,直到满足某一设定的阈值为止,这样的最终曲线就是满足自相合性的主曲线。

目前,对于主曲线性质的研究在国内还是一项空白,至今没有相关的文献发表。1995年,Washington Seattle 大学统计系的 Werner Stuetzle 发表了一篇有关主曲线性质的论文 Geometric Properties of Principal Curves in the Plane,这篇论文提出主曲线恰好是一个微分方程的解,根据这一思想,我们紧扣自相合性这一主曲线的根本特性,首先进行公式的逻辑推导,最终获得一个微分方程,然后基于这个微分方程进行实验,对于这一微分方程的解(即主曲线)进行验证分析,并提出我们的结论。

2.2 自相合性作为一个曲率条件

自相合性是研究分析主曲线其它特性的基础,它直接或间接地决定了主曲线的其他特性。我们首先引入法向坐标[3](Normal Coordinates)的概念来解释自相合性为一个曲率条件。

定义2 由公式:

$$\omega_\Gamma(s,v)=f(s)+vN(s) \tag{2.2.1}$$

定义 Γ 的法向坐标映射为:

$$\omega_\Gamma:\Lambda\times R\to R^2$$

并由公式:

$$\mu_\Gamma(X)=(\lambda(X),\langle X-f(\lambda(X)),N(\lambda(X))\rangle) \tag{2.2.2}$$

定义 Γ 的法向坐标转换映射为:

$$\mu_\Gamma:\Omega\to\Lambda\times R$$

其中,Ω 是不含模糊点的紧支撑,则我们把 $\mu_\Gamma(X)$ 的元素 (s,v) 称为主曲线上点 X 的法向坐标。

设 Γ 为具有连续概率密度 $\varphi(x)$ 的数据分布的一条正则主曲线,根据法向坐标中的自相合条件,可以推导出 Γ 的曲率和 $\varphi(x)$ 的某些条件矩之间的关系。基于主曲线 Γ 的自相合性质,利用条件期望的定义,在所有的可测量区间 $\Lambda\subset\Gamma$ 上,可以得到公式:

$$\int_{\mu_\Gamma^{-1}(A)} x\varphi(x)\mathrm{d}x=\int_{\mu_\Gamma^{-1}(A)}\mu_\Gamma(x)\varphi(x)\mathrm{d}x \tag{2.2.3}$$

由于已经假设 Γ 为正则性主曲线,利用法向坐标,公式(2.2.3)可以改写为:

$$\int_{N(s)} v\varphi(f(s)+vN(s))\frac{\partial(x,y)}{\partial(s,v)}\mathrm{d}v=0 \tag{2.2.4}$$

再利用法向坐标映射的雅可比行列式:

$$\frac{\partial(x,y)}{\partial(s,v)}=\left|\frac{\partial\omega_\Gamma(s,v)}{\partial s}\times\frac{\partial\omega_\Gamma(s,v)}{\partial v}\right|$$
$$=|(1-vk(s))T(s)\times N(s)|=1-vk(s) \tag{2.2.5}$$

代入(2.2.4)得到:

$$\int_{N(S)} v\varphi(f(s)+vN(s))dv-k(s)\int_{N(s)} v^2\varphi(f(s)+vN(s))\mathrm{d}v=0 \tag{2.2.6}$$

对于所有的 $s\in\Lambda$,令:$\mu_\perp(s)=\dfrac{\int_{N(S)} v\varphi(f(s)+vN(s))\mathrm{d}v}{\int_{N(S)}\varphi(f(s)+vN(s))\mathrm{d}v}$ 表示法向概率密度的均值,并且令 $\sigma_\perp^2(s)$ 表示方差,则公式(2.2.6)的形式可以改写为:

$$k(s)=\frac{\mu_\perp(s)}{\mu^2{}_\perp(s)+\sigma^2{}_\perp(s)} \tag{2.2.7}$$

由公式(2.2.7)可以看出:主曲线 Γ 的曲率与主曲线的法向概率 $k(s)$ 密度的第一、第二阶矩联系起来,并且在特殊情况下,如果概率密度 $\varphi(x)$ 是均匀的,那么与 $f(s)$ 法向均值重合,就不会存在曲率 $k(s)$。

2.3 建立微分方程组

为了便于建立微分方程,特引入参数 θ,θ 是点 $t\in\Gamma$ 处的切向量与水平坐标轴正向的夹角。下面从法向矩的角度,写出具有法向连续概率密度的均匀分布的均值和方差为:

$$\mu_\perp(x,\theta)=\frac{\mu_1(x,\theta)}{\mu_0(x,\theta)}$$

$$\sigma_\perp^2(x,\theta)=\frac{\mu_2(x,\theta)}{\mu_0(x,\theta)}-\mu_\perp^2(x,\theta)$$

从而,得到主曲线方程的一阶微分方程组:

$$\begin{cases}\dfrac{\mathrm{d}x}{\mathrm{d}s}=\cos(\theta)i+\sin(\theta)i\\ \dfrac{\mathrm{d}\theta}{\mathrm{d}s}-\dfrac{\mu_1(x,\theta)}{\mu_2(x,\theta)}=\dfrac{\mu_\perp(s)}{\mu_\perp^2(s)+\sigma_\perp^2(s)}\end{cases} \tag{2.2.8}$$

从以上的推导过程可以看出,方程组(2.2.8)的任意解均满足主曲线的自相合条件,故而得到下面的定理:

定理1 如果一条正则曲线是而且只是方程组(2.2.8)的一个解,那么它就是具有连续概率密度 $\varphi(x)$ 的均匀分布的一条主曲线。

3 实验分析

在我们的实验当中,主要研究了服从概率密度 $\Omega_{R_1,R_2}=\{(r,\theta):R_1\leqslant r\leqslant R_2\}$ 的环形均匀分布的主曲线。根据公式(2.2.8),推导出圆形主曲线的半径为:

$$r=\frac{2(R_1^2+R_1R_2+R_2^2)}{3(R_1+R_2)} \tag{3.1}$$

由公式(3.1)可以看出,圆形主曲线的半径只与环形分布的内外半径的取值有关,并且通过计算发现,主曲线的某些性质直接决定于内外半径的比值。下面,我们首先选取一组实验数据来说明主曲线的一些性质。取环形均匀分布区域的外半径 $R_1=1$,内半径 $R_2=0.5$,并且给定初始条件 $r(0)=0.670$,$r'(0)=0$。经过计算,得到的实验结果为:圆形主曲线(半径 $r\approx0.778$)和周期为 $T=\pi/4$ 的振荡主曲线($0.670<r<0.882$);如果把初始条件改为 $r(0)<0.670$,$r'(0)=0$,除了圆形主曲线外,得到周期略大于 $\pi/2$ 的振荡主曲线,并且振荡主曲线不封闭;如果进一步把初始条件该为 $r(0)>0.670$,$r'(0)=0$,得到周期略小于 $\pi/2$ 的振荡主曲线,并且振荡主曲线同样不封闭(图2)。

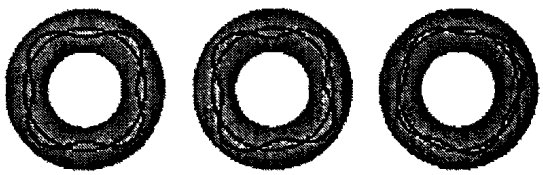

图2 微分方程的解

从以上实验结果可以看出,环形均匀分布的主曲线是具有周期性的,并且周期 T 依赖于初始化值 r_0。这一点可以利用数值分析方法加以解释,由于我们是利用计算机求解方程组(2.2.8)的数值解,然而我们知道,在多数情况下找出微分

(下转第66页)

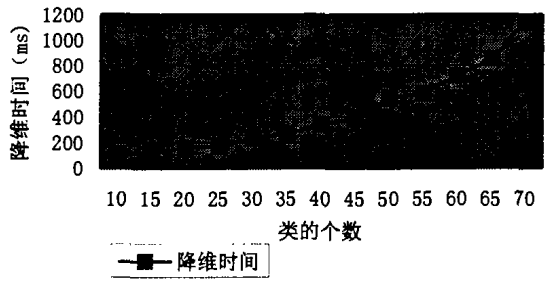

图3

另外，核矩阵的形式比较多[3]，本文使用高斯核 $K(x,y)=\exp(-\|x-y\|^2/\sigma^2)$，通过实验分析 $\sigma=1000$比较合适。

下图是 PCA 方法、KPCA 方法、基于聚类的 KPCA 方法的对比。降维时间如图4；效果趋势如图5。

方法	时间(ms)
PCA	47
KPCA	2200
基于聚类的 KPCA	510

图4

从图5可以看到，基于聚类的 KPCA 方法，大大提高了 KPCA 计算核矩阵的速度，而且随着图像个数的增多，几乎不影响原来的检索精度。

结论 本文首先分析了 PCA 和 KPCA 的不足，着重讲述了 KPCA 针对样本点计算带来的复杂度分析，介绍了基于聚类的 KPCA 方法，该方法不是对整个样本点做主成分分析，而是对代表样本点每类的均值向量做主成分分析，大大缩减了核矩阵的阶数，通过试验验证了算法的优越性。未来继续研究的工作是其他核分析方法。

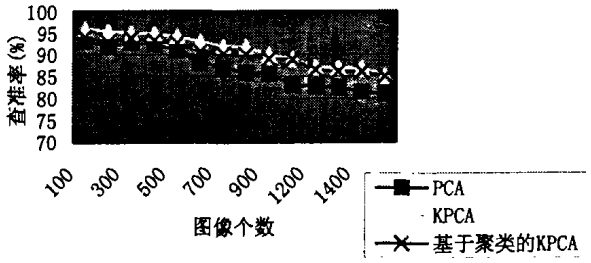

图5

参 考 文 献

1 Carreira-Perpian M A. A Review of Dimension Reduction Techniques: [Technical Report CS-96-09]. Department of Computer Science, University of Sheffield, 1997

2 Maeda E, Murase H. Multi-category classification by kernel based nonlinear subspace method. In: Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 99), IEEE Press, 1999

3 Scholkopf B, Smola A J, Muller Klaus-Robert. Kernel Principal Component Analysis [M]. Advances in Kerrel Methods, MIT Press, 1998. 327～352

4 Vapnik V N. Statistical learning theory. AT&T Research, London University, 1998

5 Smola A J, Scholkopf B. Sparse greedy matrix approximation for machine learning. In: Proc. of ICML'00, Bochum, Germany, Morgan Kaufmann, 2000. 911～918

6 Williams C, Seeger M. Using the Nystrom method to speed up kernel machines. Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2001, 13

7 Tipping M. Sparse kernel principal component analysis. In Advances in Neural Information Processing Systems 13, MIT Press, Cambridge, MA, 2001. 633～639

8 边肇祺，张学工，等. 模式识别. 清华大学出版社

（上接第63页）

方程解的解析表达是极其困难的，因此，只能采用一种近似的数值方法，这种方法在假定方程组(2.2.8)的解存在并且满足正则性，同时还必须假定精确解连续依赖于初始值，这样，根据输入初始值的不同，在计算过程中沿着搜索方向就会陷入不同的局部最小，从而得到不同周期的振荡曲线。

我们再分析圆形主曲线的半径 $r=(R_1+R_2)/2$。由于环形概率密度的分布是对称的，理论上必然得到圆形主曲线，并且在直觉上，所得到的圆形主曲线的半径 $r=(R_1+R_2)/2$。但是，我们所得到的圆形主曲线的半径并不等于环形分布的内外半径的平均值，而是 $r>(R_1+R_2)/2$。由于我们所建立的微分方程是基于主曲线的自相合性，而根据自相合性，生成曲线的外侧区域内的数据点多于内侧区域内的数据点，这样经过计算得到的实际圆形主曲线必然向外偏离半径为 $r=(R_1+R_2)/2$的圆。

由图 2还可以看到，环形均匀分布的主曲线不止一条，并且它们是相交的。这与线性主成分十分相似，数据分布一般也有很多主曲线，并且这些主曲线的相交性类似于线性主成分的正交。

我们选取不同的内外半径的比值和圆形主曲线的半径初始值 r_0，通过大量计算，概括出以下结论：环形均匀分布的主曲线不唯一且相交，并且主曲线的数目随内外半径比值的增加而增加，而周期 T 随比值的增加而减小。

结束语 我们所做的工作是对主曲线的某些性质进行初步的分析和研究，研究的数据分布仅针对特殊的均匀分布，对于平凡分布甚至是未知分布的主曲线性质的研究还有待进一步的研究，并且研究的空间仅限于二维空间，有待于向高维空间发展。

参 考 文 献

1 Hastie T. Principal curves and surfaces. Laboratory for computational Statistics: [Technical Report 11]. Stanford University, Dept. of Statistics, 1984

2 Kégl B. Principal curves: learning, design, and applications: [Ph. D. Thesis]. Concordia University, Canada, 1999

3 Duchamp T, Stuetzle W. Extremal properties of principle curves in the plane. Annals of Statistics, 1996, 24(4): 1511～1520

4 Hermann T, Meinicke P, Ritter H. Principal curve sonification. In: Proc. of Intl. Conf. on Auditory Display, 2000. 81～86

5 Seung H S, Lee D D. The manifold ways of perception? Science, 2000, 12: 2268～2269

6 Banfield J D, Raftery A E. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. Journal of the American Statistical Association, 1992, 87(417): 7～16

7 张军平，王珏. 主曲线研究综述. [J] 计算机学报，2003，26(2)：129～146

8 张红云，苗夺谦. 基于主曲线的相似字符模糊分类方法. 模式识别与人工智能，2004，4

9 唐庆适，苗夺谦. 基于主曲线的指纹细节特征提取方法. 计算机科学，2005，32(1)