第 33 卷 第 12 期 计 算 机 工 程 2007 年 6 月
Vol.33 No.12 Computer Engineering June 2007
·开发研究与设计技术· 文章编号：1000—3428(2007)12—0283—03 文献标识码：A 中图分类号：TP391

基于 B-样条网络的复杂主曲线建模

郝继升[1,2]，何 清[2]，史忠植[2]
(1. 延安大学计算机学院，延安 716000；2. 中国科学院计算技术研究所智能信息处理重点实验室，北京 100080)

摘 要：提出了一种基于 B-样条网络的复杂主曲线建模的新方法，该方法结合学习主曲线的多边形算法和 B-样条网络来建立主曲线模型，同时提出了用于寻找主曲线分叉点的迭代算法。实验结果表明所提出的方法是简便有效的。
关键词：主曲线；主曲线建模；多边形算法；B-样条网络；分叉点

Complex Principal Curve Modeling Based on B-spline Network

HAO Jisheng[1,2], HE Qing[2], SHI Zhongzhi[2]
(1. College of Computer Science, Yanan University, Yanan 716000;
2. Key Laboratory of Intelligence Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080)

【Abstract】The paper proposes a new method based on B-spline network, modeling complex principal curve. This method combines the polygonal line algorithm of learning principal curve with B-spline network. The iterative algorithm for searching bifurcate point is proposed. Simulation results demonstrate that it is feasible and effective.
【Key words】Principal curve; Principal curve modeling; Polygonal line algorithm; B-spline network; Bifurcate point

1 概述

主曲线的概念最早由Hastie提出[1]，由于他在描述其原理时使用了复杂的数学，当时并没有引起计算机科学界的注意。虽然目前主曲线的研究中仍然存在大量的数学问题，但在 20 世纪 90 年代后期，由于主曲线技术在数据分析领域中的逐步应用，及其本身所具有的多种优点，主曲线技术已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(已有大量应用方面的报道)。目前的应用主要有图像的可视化、数据可听化、语音识别、时间数据分析、模式分类、手写数字识别、模式聚类、过程监控等领域。主曲线是通过数据集“中部”的光滑曲线，它真实反映数据的形态，是数据集合的“骨架”，它是主成分分析的非线性推广。目前已经提出许多主曲线算法，但据目前的文献，主曲线学习算法只能得到近似主曲线上的离散点，没有对主曲线进行建模。而通过建立主曲线模型，可以反映数据集中变量之间的非线性关系。

B样条网络是一种点阵 3 层结构联想记忆网络[2,3]，其结构如图 1 所示， 它在隐层中使用定义在输入空间点阵上的B样条函数作为基函数，对于任一网络输入，隐层中只有少数几个B样条函数激活，而网络输出仅由这几个激活的B样条函数的线性组合形成，由于基函数的支集有限，这种网络有下列特征：(1)知识在网络中是局部存储而不是全局分布式存储，学习也是局部进行的，因而输入空间某一部分的学习不会影响其它部分的学习结果；(2)学习收敛速度快，便于实时在线应用；(3)网络有较强的模糊知识表示能力。因此，这类网络近年来越来越受到人们的重视，并广泛应用于控制、建模和模式识别等领域[3]。

本文提出了一种建立主曲线模型的方法。主要考虑一类带分支的主曲线建模，这种带分支的主曲线的形式如图 2 所示，主曲线在分叉点被分为两个分支。

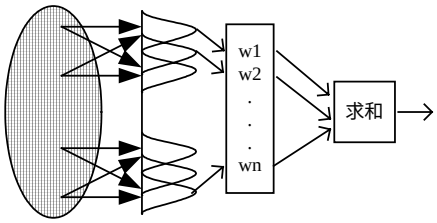

图 1 B 样条网络结构

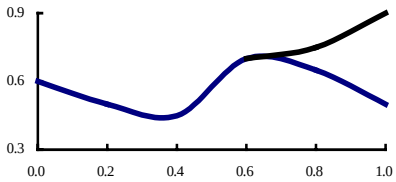

图 2 带分叉点的主曲线

在这类主曲线的建模中，1 个关键的问题是如何根据数据点集来找到主曲线的分叉点，进而可将带分支主曲线的建模问题归结为单支主曲线的建模问题，为此本文提出了一种寻找分叉点的迭代算法，在所提出的复杂主曲线建模算法中，首先利用寻找主曲线分叉点的迭代算法寻找分叉点，进而将带分支的主曲线划分为 3 个单支的主曲线；然后，对每个单

基金项目：国家自然科学基金资助项目(60435010，90604017)；国家“973”计划基金资助项目(2003CB317004)；北京市自然科学基金资助项目(4052025)
作者简介：郝继升(1963－)，男，副教授、硕士，主研方向：神经网络，机器学习，主曲线；何 清，副研究员、博士；史忠植，研究员、博导
收稿日期：2006-10-08 **E-mail：**haojs@ics.ict.ac.cn

支主曲线利用学习主曲线的多边形算法[2]，找到 3 个多边形；最后将多边形的顶点作为B样条网络的训练样本集，训练B样条网络，由于B样条网络训练周期短，收敛速度快，因此利用这种方法可以快速地建立复杂主曲线模型。同已有的学习主曲线算法相比。由于B-样条网络的基函数都是连续函数，因而利用本方法所建立的主曲线是平滑曲线。

2 主曲线定义及学习主曲线的多边形算法

本节给出了几种主曲线的定义[1,4,5]和学习主曲线的多边形算法[5]。

2.1 主曲线定义

定义 1 具有连续概率密度 $h(x)$的数据分布 $D\subset R^d$ 的主曲线 f 是流形 M 中满足自相合性的成员。1 条曲线 $f\in M$ 是自相合的，如果 $E(X|\lambda_f(X)=\lambda)=f(\lambda)$，$\forall\lambda\in I$，其中 I 是实数轴上的闭区间，$M=\{M_f: f\subset F\}$，$M_f=f(D)=\{f(X): X\in D\}$，$F$ 为函数集，对每个 $f\in F$，$f: D\to R^d$。

定义 2 光滑曲线 $f(\lambda)$ 是主曲线当且仅当：

(1) $f(\lambda)$ 不自相交；

(2) $f(\lambda)$ 在任一有界的R^d子集中长度有限；

(3) $f(\lambda)$ 是自相合的，即 $f(\lambda)=E(X|\lambda_f(X)=\lambda)$。

定义 3 对于数据点集X，曲线 f^*称为长度为L的主曲线，如果在所有长度小于或等于 L 的曲线簇上，f^*最小化距离函数 $\Delta(f)$，其中

$$\Delta(f)=E[\Delta(X,f)]=E[\inf_{\lambda}\|X-f(\lambda)\|^2]=E[\|X-f(\lambda_f(X))\|^2]$$

由主曲线的定义可知：主曲线是满足自相合性的光滑曲线，其本质上是嵌入高维欧氏空间的一维流形，曲线上的每个点是所有投影到该点的数据点的条件均值，它能真实反映数据的分布形态。

2.2 学习主曲线的多边形算法

学习主曲线的多边形算法由Kégl B提出[5]。

算法 1 多边形算法

(1)初始化：给定数据点集

$$X_n=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}\subset R^2$$

用 X_n 的最短第一主分量线作为算法迭代的初始多边形，其上包含投影到它上的所有数据点。

(2)投影：根据数据点投影到多边形的顶点或线段而将数据点划分到最近邻区域，数据点集最近邻区域的划分方法如图 3 所示。其中，集合V_i中的任一点关于多边形f的最近邻点是v_i；集合S_i中任一点关于多边形f的最近邻点是线段 s_i 上的一点。

(3)顶点优化：顶点v_i的新位置通过距离平方的均值最小化来决定，所有别的顶点固定不变。

(4)增加新顶点：由上面的投影和优化组成算法的内循环，在内循环执行期间，使优化作用到每一个顶点v_i，$i=1,2,\ldots,k+1$，直到收敛而产生多边形 $f_{k,n}$，这样一个新的顶点被增加。

当多边形包含的线段数k超过阈值 $c(n,\Delta)$ 时算法终止，算法的这一终止条件基于启发式的复杂性度量考虑，由多边形 $f_{k,n}$ 包含的线段数k，数据集大小 n，距离平方的均值 $\Delta_n(f_{k,n})$ 共同决定，算法的外循环由步骤(2)~步骤(4)组成，在每一次的外循环迭代中通过增加一个新的顶点到前一次迭代产生的多边形 $f_{k,n}$ 上而使多边形包含的线段数增加 1，在增加一个新顶点后，多边形所有顶点的位置在内循环中被更新。

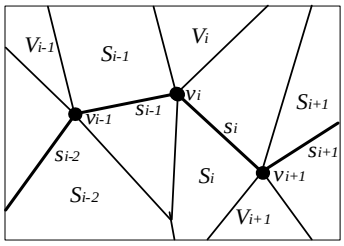

图 3 对平面上点的最近邻区域的划分

3 寻找主曲线分叉点的迭代算法

考虑数据点集 $X_n=\{(x_i,y_i)\}\subset R^2$，其中，$i\in N=\{1,2,\ldots,n\}$。令 $x_{\min}=\min_{i\in N}\{x_i\}$，$x_{\max}=\max_{i\in N}\{x_i\}$，记 $I^{(0)}=[x_{\min},x_{\max}]$，$|I^{(0)}|$表示区间 $I^{(0)}$ 的长度。

下面给出寻找数据点集 X_n 的主曲线分叉点 $(\bar{x},\bar{y})$ 的迭代算法。

算法 2 寻找分叉点算法

{Let $t=0$;

Do {

将区间 $I^{(t)}$ m 等分(m 为正整数)；

得到 m 个小区间记为 I_i，$i=1,2,\ldots,m$；

记 $J_i=\{y_j|\forall(x_j,y_j)\in X_n, x_j\in I_i\}$ $i=1,2,\ldots,m$；

$\bar{y_i}=\frac{1}{n_i}\sum_{y_j\in J_i}y_j$，$i=1,2,\ldots,m$；

$\sigma_i=\frac{1}{n_i}\sum_{y_j\in J_i}(y_j-\bar{y_i})^2$，$i=1,2,\ldots,m$；

其中，n_i 表示 I_i 中所含数据点的个数；

且 $n_1+n_2+\ldots+n_k=n$；

计算 $\{\sigma_{i+1}-\sigma_i\}$，$i=1,2,\ldots,m-1$；

存在 i，使得序列 $\{\sigma_{j+1}-\sigma_j\}$，$j=i,i+1,\ldots,k-1$ 中的每一项都非负且为严格递增序列；

这时数据点集 X_n 的主曲线的分叉点横坐标必然在区间 I_{i+1} 内；

if ($|I_{i+1}|>\varepsilon$) then

{ $t=t+1$；$I^{(t)}=I_{i+1}$；}

(其中，ε 是预先给定的一个小的正数)}

while ($|I_{i+1}|>\varepsilon$)

Let $\bar{x}=\frac{1}{n_{i+1}}\sum_{x_j\in I_{i+1}}x_j$，$\bar{y}=\frac{1}{n_{i+1}}\sum_{y_j\in J_{i+1}}y_j$ }

$(\bar{x},\bar{y})$ 即为数据点集 X_n 的主曲线的分叉点；由于 $|I^{(0)}|=x_{\max}-x_{\min}$，$|I^{(1)}|=\frac{1}{m}|I^{(0)}|$，…，$|I^{(t)}|=\frac{1}{m^t}|I^{(0)}|$，$\lim_{t\to+\infty}|I^{(t)}|=0$，因此上述寻找数据点集主曲线分叉点的迭代算法显然是收敛的，而且当取较大的 m 值时，算法将以很快的速度收敛。

4 建立复杂主曲线模型的方法

本文所提出的基于 B-样条网络和寻找数据点集主曲线分叉点算法的建立复杂主曲线模型方法基本思路是：对于给

定数据点集 $X_n=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}\subset R^2$。

(1)利用算法 2，找到其分叉点 $(\bar{x},\bar{y})$，记

$$X^{(1)}=\{(x_i,y_i)\,|\,i\in N,(x_i,y_i)\in X_n,x_i<\bar{x}\}$$

$$X^{(21)}=\{(x_i,y_i)\,|\,i\in N,(x_i,y_i)\in X_n,x_i\geq\bar{x},\ y_i\geq\bar{y}\}$$

$$X^{(22)}=\{(x_i,y_i)\,|\,i\in N,(x_i,y_i)\in X_n,\quad x_i\geq\bar{x},y_i<\bar{y}\}$$

$$X^{(2)}=X^{(21)}\cup X^{(22)}$$

显然分叉点 $(\bar{x},\bar{y})$ 的横坐标 $\bar{x}$ 将数据点集 X_n 划分为 2 个不相交的子集 $X^{(1)}$ 和 $X^{(2)}$，即 $X_n=X^{(1)}\cup X^{(2)}$，且 $X^{(1)}\cap X^{(2)}=\phi$；分叉点 $(\bar{x},\bar{y})$ 的纵坐标 $\bar{y}$ 将 X_n 的子集 $X^{(2)}$ 划分为 2 个不相交的子集 $X^{(21)}$ 和 $X^{(22)}$，即 $X^{(2)}=X^{(21)}\cup X^{(22)}$，且 $X^{(21)}\cap X^{(22)}=\phi$。分叉点 $(\bar{x},\bar{y})$ 将数据点集 X_n 划分为 3 个互不相交的子集 $X^{(1)}$、$X^{(21)}$ 和 $X^{(22)}$，即 $X_n=X^{(1)}\cup X^{(21)}\cup X^{(22)}$，且 $X^{(1)}\cap X^{(21)}\cap X^{(22)}=\phi$。显然，分叉点 $(\bar{x},\bar{y})$ 将数据点集 X_n 划分成的 3 个互不相交的子集 $X^{(1)}$、$X^{(21)}$ 和 $X^{(22)}$ 分别对应于其主曲线的 3 个分支。

(2)利用学习主曲线的多边形算法，分别去求对应于数据点集 X_n 的 3 个互不相交的子集 $X^{(1)}$、$X^{(21)}$ 和 $X^{(22)}$ 相应的多边形。对于每个子集，可以得到由k(对于不同的子集有不同的k值)个线段和k+1 个顶点组成的多边形 $f_{k,n}$。

(3)把这 k+1 个顶点作为 B-样条网络训练样本集，训练 B-样条网络，从而可以对数据点集：

$$X_n=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}\subset R^2$$

建立其主曲线模型。在 B-样条网络的训练中，对于分叉点之后的对应于子集 $X^{(21)}$ 和 $X^{(22)}$ 的 2 个多边形顶点，采用分别训练的方式，训练的结果被分别保存。

基于 B-样条网络的复杂主曲线建模方法的流程如图 4 所示。

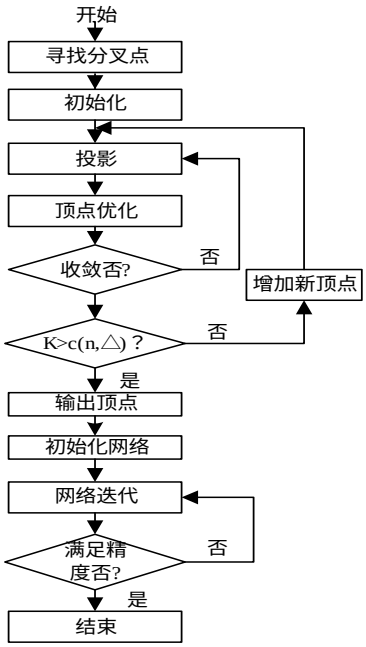

图 4 主曲线建模方法流程

利用本文所提出的基于 B-样条网络的复杂主曲线建模方法做了实验。图 5 和图 6 给出了利用本文所提出的方法建立的主曲线模型，其中的数据点是从 $y=\sin x,0\leq x\leq\frac{3}{2}\pi$ 和 $y=-\sin x,\ \pi\leq x\leq\frac{3}{2}\pi$ 上分别任意选取的 300 个和 100 个数据点，分别增加了独立分布的高斯噪音 $\varepsilon_i\sim N(0,0.1)$。

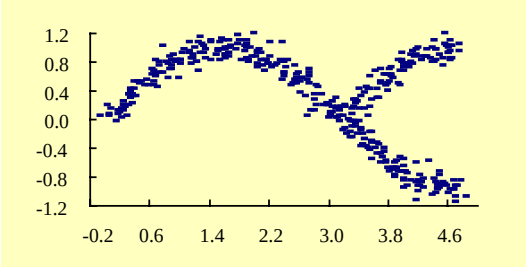

图 5 400 个原始数据点

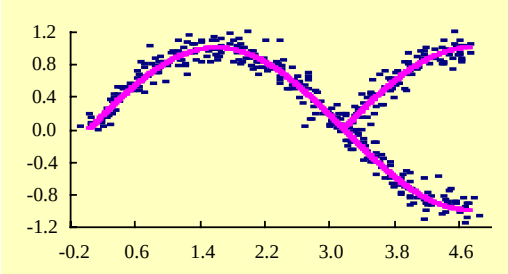

图 6 主曲线

由以上的实验的结果，可以看出，对于给定的数据点集，结合学习主曲线的多边形算法和 B-样条网络的新的主曲线建模方法是，首先利用算法 2 寻找数据点集 X_n 的分叉点 $(\bar{x},\bar{y})$，分叉点将数据点集 X_n 划分为 3 个互不相交的子集 $X^{(1)}$、$X^{(21)}$ 和 $X^{(22)}$，再对这 3 个子集分别利用算法 1，通过逐步增加多边形顶点，优化顶点，使得多边形的边数逐步增加，用多边形去逼近数据点集的主曲线，得到这 3 个子集的 3 个多边形。利用多边形的顶点作为新的数据点集，训练 B-样条网络，进而建立主曲线模型，得到一条平滑的曲线，这是一种简便而且有效的方法，而目前的主曲线学习算法只能得到近似主曲线上的离散点，没有对主曲线进行建模。

5 结论

对于给定的数据点集，本文所提出的基于 B-样条网络的复杂主曲线建模方法，可以为其建立平滑的主曲线模型。同已有的学习主曲线算法相比，本方法可以为给定的数据点集建立主曲线模型，这在一定的意义上是对目前的主曲线学习算法的改进和完善；由于 B-样条网络的基函数都是连续函数，因此而利用本方法所建立的主曲线是平滑曲线；另外，由于 B-样条网络知识的局部存储、训练周期短收敛速度快等特点，使得本方法能够快速地建立数据集的主曲线模型。

实验结果表明本方法，对于数据点集的主曲线建模是简便而且有效的，具有一定的应用前景，可应用于建模与过程控制等领域。

参考文献

1 Hastie T. Principal Curves and Surfaces[R]. Laboratory for Computational Statistics, Stanford University, 1984.

2 Moody J. Fastlearning in Multi-resolution Hierarchies[M]//Advances in Neural Information Processing System 1. Morgan Kaufman, 1989.

3 Brown M, Harris C. Neurofuzzy Adaptive Modeling and Control[M]. Prentice Hall, 1994: 89-100.

4 Hastie T, Stuetzle W. Principal Curves[J]. Journal of the American Statistical Association, 1989, 84(406): 502-516.

5 Kégl B, Krzyzak A, Linder T, et al. Learning and Design of Principal Curves[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(3): 281-297.