主曲线算法初始值选取的分析

王 真 曹立明

(同济大学计算机科学与技术系 上海 200331)

摘 要 主曲线是一种用于数据压缩和特征提取的有效方法,是对主成分分析的非线性推广。由于主曲线与主成分的密切联系,主曲线生成算法通常以第一主成分线做初始值。然而实验发现第一主成分未必是算法初始化的最佳选择。本文将以 HS 算法和多边形算法为例,就初始值的选取对生成主曲线的影响做出分析并通过实验得出结论:HS 算法以原点作初值效果较好,多边形算法应根据数据点集的不同结构选择合适的初值。

关键词 主曲线,主成分,初始值,投影指标

Analysis of Choosing the Initial Value of Principal Curve Algorithm

WANG Zhen CAO Li-Min

(Department of Computer Science and Technology, Tongji University, Shanghai 200331)

Abstract Principal Curves are very useful approaches of feature extraction and data compression, they are nonlinear generalizations of the first linear principal component which can be thought of as 'optimal' linear 1-d summarization of the data. Up to now several algorithms for constructing principal curves have been proposed. They are all initialized by the first principal component for the close relation between principal curves and the principal components. Taking example for the HS and the polygonal line algorithms, this paper analyzes how the initial line affects the principal curves. We conclude the first principal component line is not always the best choice of the initialization step. The experiments show the HS algorithm will produce perfect results if it starts with the origin, and the polygonal line algorithm should choose proper initial value according to different global structure of the data. It proves that local optimization can not always lead to global optimization.

Keywords Principal curves, Principal component, Initial value, Projection indices

1 引言

1984 年,Hastie 和 Stuetzle(简记为 HS)首次提出了主曲线方法[1],它是一种数据压缩和特征提取的有效方法。HS 主曲线被定义为满足自相合性的穿过数据集合"中间"的光滑曲线,其中自相合性是指曲线上每一点都是投影至该点的数据点的均值。基于这个性质,HS 提出了已知概率分布和未知分布下数据点集合的主曲线的实际算法(下文记为 HS 算法)。由于主曲线暗示的广泛应用背景,不断有学者从不同角度对 HS 主曲线加以改善,给出新的诠释并提出新的算法。其中较为有影响的是 Kegl 引入的有长度约束的主曲线概念[4],并提出了相应的多边形算法。

主曲线的提出可以理解为基于线性的主成分分析方法向更精确描述现实世界的非线性分析方法的推广。它与主成分的密切关系体现在两个方面:其一,当主曲线恰好是一条直线时,那么它就是一个主成分;其二,两者都是距离函数的临界点。正是因为主曲线有诸多类似于主成分的性质,是主成分线的自然推广,所以 HS 算法和多边形算法均在起始步中以第一主成分线作初始值,这在理论上似乎是顺理成章的。但是实验表明,提取主曲线的算法中用第一主成分设定初始值在许多情况下未必是最优的。我们产生了几组在一条光滑曲线上加正态分布的噪声的数据点,分别用两种算法提取主曲线,结果发现,对于 HS 算法改用原点作初始值效果比用第一主成分线好得多;对于多边形算法,根据数据集合的整体结构选取初始值才能使得主曲线拟合性更好。本文工作将对主曲线算法的改进起到直接作用。

本文第 2 节简要描述了两种主曲线算法,第 3 节讨论了初始值对算法的影响,并给出实验的具体说明、结果演示以及对实验现象的分析,最后总结全文。

2 主曲线算法

HS 首次提出主曲线的概念之后,由于其暗示的广泛应用前景,不断有学者从不同角度对其进行改善,提出不同的主曲线算法。其中比较有影响的是 HS 算法和多边形算法。本节将对这两种算法做简要描述。

2.1 HS 算法

基于自相合性,HS 分别针对已知概率分布和未知分布的数据点集合提出了生成主曲线的算法[1]。对于已知概率分布的数据点集合,算法如下:

0 步:$j=0$ 时,令 $f^{(0)}(t)$ 为数据集 X 的第一主成分线;

1 步:(投影步)对所有 $x\in R^d$,设置 $t_{f^{(j)}}(x)=\max\{t:\|x-f(t)\|=\min_\tau\|x-f(\tau)\|\}$

2 步:(期望步)令 $f^{(j+1)}(t)=E[X|t_{f^{(j)}}(X)=t]$;

3 步:如果 $1-\frac{\Delta(f^{(j+1)})}{\Delta(f^{(j)})}$ 小于某个阈值,则停止;否则,令 $j=j+1$,返回第 1 步。

针对未知分布的数据点集合生成主曲线,算法略有调整。我们后面的讨论主要基于这一种情况,所以以下着重介绍 HS 提出的调整处理。

假设给定点集 $X_n=\{x_1,x_2,\cdots,x_n\}\subset R^d$,将数据点投影到依弧长参数化的曲线 f 上,投影点为 $f(t_1),f(t_2),\cdots,f(t_n)$,那么按照弧长参数 t_i 由小到大的顺序就可将数据点 x_i 进行排序。假设算法产生的所有曲线是由 n 个顶点顺次连接构成的多边形曲线,且以弧长作参数(即投影指标:从始点到投影点的沿曲线距离),n 个顶点即投影点 $f(t_i)(i=1,\cdots,n)$,其中 t_i 递归的定义为

王 真 博士生,主要研究领域为模式识别,容错计算;曹立明 博士生导师,主要研究领域为模式识别,人工智能等。

$t_1=0$;

$t_i=t_{i-1}+\|f(t_i)-f(t_{i-1})\|$,$i=2,\cdots,n$;

在第 0 步,$f^{(0)}(t)$是数据集 X_n 的第一主成分线。投影步通过将数据点投影到 $f^{(j)}(t)$来计算新的投影指标 $t_i^{(j)}$,$(i=1,\cdots,n)$。第一次达到这一步时,便将数据点投影到第一主成分线设置投影指标。第 j 次迭代后,当前曲线被描述为顶点是 $f^{(j)}(t_1^{(j-1)}),\cdots,f^{(j)}(t_n^{(j-1)})$的多边形曲线。期望步中首先按前述将数据点 x_i 升序重排,再通过求均值重新生成曲线。如此迭代直至收敛。

2.2 多边形算法

多边形算法[4]是继 HS 算法之后又一个提取主曲线的算法,它生成有长度约束的主曲线,其在应用范围、算法复杂性和准确率上都比 HS 算法有明显改善。多边形算法主要就主曲线的存在性问题在附加约束条件后给予解决。下面首先简要介绍一下此算法:

0 步:(初始化)$k=0$ 时,$f_{0,n}$为数据集 X 的第一主成分线上包含所有投影点的长度最短的一段;

1 步:(投影)将数据点根据其投影点的位置划分到不同区域 $V_1,\cdots,V_{k+1}$和 $S_1,\cdots,S_k$

$V_i=\{x\in X_n|\Delta(x,v_i)=\Delta(x,f),$

$\Delta(x,v_i)<\Delta(x,v_m),m=1,\cdots,i-1\}$

$S_i=\{x\in X_n|x\notin V_i,\Delta(x,s_i)=\Delta(x,f),$

$\Delta(x,s_i)<\Delta(x,s_m),m=1,\cdots,i-1\}$;

2 步:(顶点优化)依次调整顶点 v_i 的位置使 $E(V_i)=\Delta_n(V_i)+\lambda_p P(V_i)$达到最小;

3 步:若 $\Delta(f_{k,n})<\Delta(f_{k-1,n})$则收敛,再判断 k,若 $k>c(n,\Delta)$则终止,否则加入新结点返回第 1 步。

3 初始值的选取

3.1 初始值对 HS 算法的影响

我们在实现 HS 算法时,在第 0 步将初始值不固定为第一主成分线而设为可调整的量,在期望步对排好序的数据点依次取 dn 个求一次期望,然后顺次连接均值点生成每一次迭代的新曲线。也就是说程序中可以根据需要调整初始值和 dn 来观察实验结果。我们的实验数据主要由附加模型 $Y=f(X)+e$ 定义的集合,其中 $e\sim N(0,\sigma^2)$,也就是在光滑曲线上加正态分布的噪声生成数据点。用 HS 算法对同一组数据点首先选取合适的参数 dn,再固定 dn 设定不同初始值提取主曲线,比较实验结果,发现初始值设为第一主成分线并不总是最优的(见图 1)。尽管数据点集合形态不同会出现不同的情况,但以原点做初始值几种数据均能产生理想的结果。图 1 中三组数据点分别是以圆、抛物线、正弦线为原始曲线加白噪声生成的,每一组中三个图是在相同数据集合上选取原点、任一直线、第一主成分线作初始值生成了主曲线。

由图 1 可以看出,HS 算法中初始值设定第一主成分线并不总是最好的,用原点效果总能理想(实验可知原点也适用于其他类型数据点)。我们进一步对实验结果进行分析。实际上,原点做初始值可以保证在第一次迭代中数据点的排序跟录入的原始顺序保持一致。因为第一次达到投影步时,将数据点投影到原点,其投影指标均为零,那么在进行期望步之前要做的升序重排也就是按照原始顺序,以此顺序每 dn 个点求一次均值,使得第一次得到的多边形曲线符合数据点的本来结构,这样有利于增强主曲线的拟合性。

然而,初始值为第一主成分线却不总能保证数据点的排序合理。例如,对于第二组抛物线上加噪声的情形,其第一主成分线为抛物线的对称轴。如图 2,第一次迭代中,数据点 x_i 先向直线 l 上投影,按照投影指标由小到大的顺序将 x_i 排序。假设期望步中令 dn=3,则按 $t(x_1),t(x_2),t(x_3)\cdots$的顺序 x_1,x_2,x_3 在一组求期望,均值点落在点 A 上。如此进行下去,经过有限次迭代生成的主曲线便会出现图 1(b3)的情形。

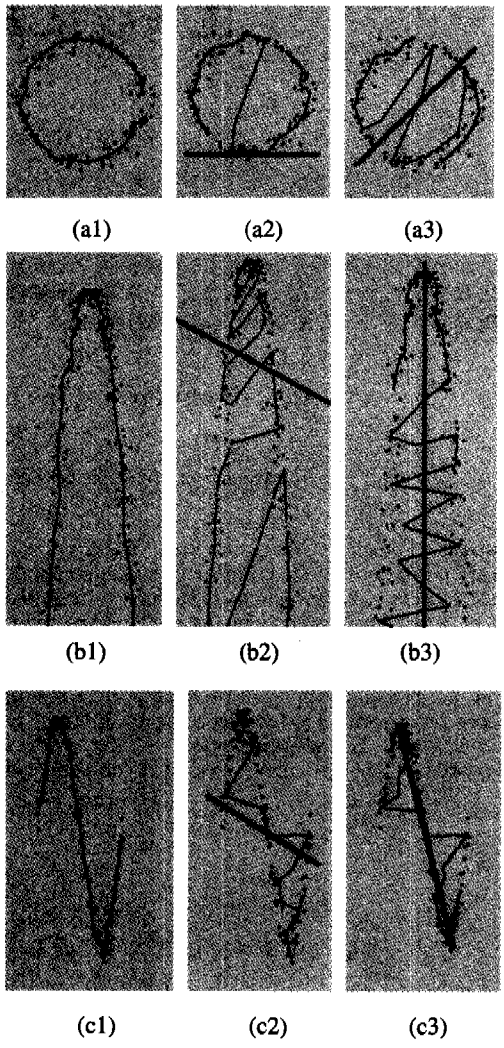

图 1 以上三组数据点分别是在圆、抛物线、正弦曲线上加正态分布噪声产生的,每一组(例如 a1,a2,a3)是对同样数据点用 HS 算法提取主曲线。每一组第一个图初始值为原点,后两个的为图中直线,其中第三图为第一主成分线。

这里存在一个问题,原始数据点的录入顺序如果是杂乱无章的怎么办?事实上,真正将主曲线算法结合应用背景进行特征提取或数据压缩时,数据文件的读入是可以控制的,我们完全可以令其以可视化的顺序读入。

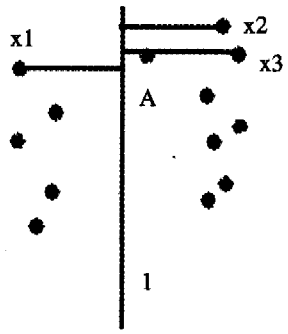

图 2 x_i 为数据点,l 为初始值,点 A 为均值点

3.2 初始值对多边形算法的影响

实验表明,在多数情况下多边形算法提取的主曲线效果

优于 HS 算法，但此算法的初始化步也会影响结果，对有些数据点以第一主成分做初始值不能得到理想主曲线。如图 3，对于在正弦上为加正态分布噪声的数据点，(1)是用第一主成分做初值提取的主曲线，(2)换做 s 形折线做初值，显然，后者提取的主曲线才能更好地拟合其本来结构。

对于多边形算法，当处理结构比较复杂的数据类型时，往往第一主成分线不能达到较好的效果，需要调整初始值，使其基本体现数据点的整体结构。

结论 通过对上述两种算法进行分析，我们可以得出这样的结论：局部最优未必得到全局最优。

由于主曲线与主成分有密切联系，是主成分分析的非线性推广，在提取主曲线的算法中往往以第一主成分线做初始值，这在理论上被认为是合理的选择。但是，我们通过实验，比较分析得出不同的结论。对于 HS 算法把初始值换作原点会得到比原算法更好的结果。而多边形算法要针对具体数据点的结构特点适当调整初始值。

参 考 文 献

1 Hastie T. Principal curves and surfaces:[PhD thesis]. Standford University, 1984

2 Hastie T, Stuetzle W. Principal curves. Journal of the American Statistical Association, 1989, 84: 502～516

3 Zhang Junping, Wang Jue. An Overview of Principal Curves. Chinese Journal of Computers, 2003, 26(2): 129～146

4 Kegl B. Principal curves: Learning, design, and applications: [Ph D dissertation]. Concordia University, Canada, 1999

5 Delicado P. Another look at principal curves and surfaces. Journal of Multivariate Analysis, 2001, 77(1): 84～116

6 Duchamp T, Stuetzle W. Geometric properties of principal curves in the plane. Robust Statistics, Data Analysis, and Computer Intensive Mehtods. New York: Springer Press, 1996, 109: 135～152

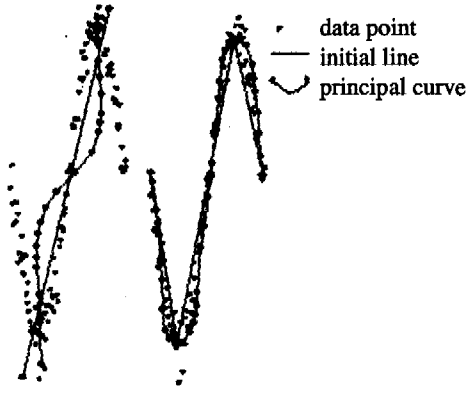

图 3 对同一组数据点(灰色点)用多边形算法提取的主曲线(光滑曲线)。(1)中初始值为第一主成分(图中直线)，(2)中初始值为图中折线。

(上接第 210 页)

表 6 Wine 数据集各有效性指标所确定的最佳聚类数随 m 的变化，$\beta=10$

m	V_{PC}	V_{PE}	V_{OS}	V_{KXB}	V_{LL}	V_{KFS}	V_{KK}	V_{KCWB}	V_{KB}	V_{KSV}
1.5	2	2	8	3	2	8	3	2	3	2
1.7	2	2	8	3	2	8	3	2	3	2
1.9	2	2	8	3	2	8	3	3	7	2
2.1	2	2	8	3	2	8	3	3	3	2
2.3	2	2	8	3	2	8	3	3	3	2
2.5	2	2	8	3	2	8	3	2	4	2

结论 本文针对核空间中聚类算法的有效性评价问题，在简要介绍 KFCM 算法的基础上，借助核函数与核空间中内积的计算关系，将六个著名的有效性指标 V_{XB}，V_{FS}，V_K，V_{SV}，V_{CWB}，V_B 推广到核 Hilbert 空间，得到其核化表示。同时，通过数值实验测试，比较了这些核化指标随模糊指数 m、高斯核宽度 β 的依赖关系。结果表明，在所考察的指标中，Xie-Beni 指标及其改进的 Kwon 指标的核化形式具有最好的性能和稳健性，可优先作为 KFCM 聚类算法的有效性指标。

参 考 文 献

1 张莉，周伟达，焦李成. 核聚类算法. 计算机学报，2002，25(6)：587～590

2 伍忠东，高新波，谢维信. 基于核方法的模糊聚类算法. 西安电子科技大学学报(自然科学版)，2004，31(4)：533～537

3 Girolami M. Mercer Kernel-Based Clustering in Feature Space. IEEE Trans. Neural Networks, 2002, 13(3): 780～784

4 Li K, Liu Y S. KFCSA: A Novel Clustering Algorithm for High-Dimension Data. FSKD 2005, LNAI 3613, Springer-Verlag, Berlin Heidelberg, 2005. 531～536

5 Kim D-W, Lee K Y, Lee D, et al. Evaluation of the Performance of Clustering Algorithms in Kernel-Induced Feature Space. Pattern Recognition, 2005, 38: 607～611

6 Chen S C, Zhang D Q. Robust Image Segmentation Using FCM with Spatial Constraints Based on New Kernel-Induced Distance Measure. IEEE Trans. SMC PART B: CYBERNETICS, 2004, 34(4): 1907～1916

7 Xu R, Wunsch II D C. Survey of Clustering Algorithms. IEEE Trans. Neural Networks, 2005, 16(3): 645～678

8 Halkidi M, Batistakis Y, Vazirgiannis M. Cluster Validity Methods: Part I. Available at: http://citeseer. ist. psu. edu/534869. html

9 Bezdek J C. Numerical Taxonomy with Fuzzy Sets. J. Math. Biol., 1974, 1: 57～71

10 Bezdek J C. Cluster Validity with Fuzzy Sets. J. of Cybernetics, 1974, 3: 58～72

11 Kim D-W, Lee K H, Lee D. On Cluster Validity Index for Estimation of the Optimal Number of Fuzzy Clusters. Pattern Recognition, 2004, 37: 2009～2025

12 Xie X L, Beni G. A Validity Method for Fuzzy Clustering. IEEE Trans. PAMI, 1991, 13(8): 841～847

13 Fukuyama Y, Sugeno M. A New Method of Choosing the Number of Clusters for the Fuzzy C-Means Method. In: Proc. of the fifth Fuzzy Systems Symposium, 1989. 247～250

14 Kwon S H. Cluster Validity Index for Fuzzy Clustering. Electronics Letters, 1998, 34(22): 2176～2177

15 Rezaee M R, Lelieveldt B P F, Reiber J H C. A New Cluster Validity for the Fuzzy C-Mean. Pattern Recognition Letters, 1998, 19: 237～246

16 Boudraa A-O. Dynamic Estimation of Number of Clusters in Data Sets. Electronics Letters, 1999, 35(19): 1606～1608

17 Kim D J, Park Y W, Park D J. A Novel Validity Index for Determination of the Optimal Number of Clusters. IEICE Trans. Inform. Syst. D-E, 2001, 84(2): 281～285

18 Pal N R, Bezdek J C. On Clustering for the Fuzzy C-Means Model. IEEE Trans. Fuzzy Systems, 1995, 3(3): 370～379

19 UCI Repository of Machine Learning Databases[DB/OL]. Available at: ftp://ftp. ics. uci. edu/pub/machine-learning-database