A NEW METHOD FOR MODELING PRINCIPAL CURVE

Hao JiSheng[1,2] He Qing[2] Shi Zhongzhi[2]
[1]*College of Computer Science, Yanan University, Shanxi Yanan, 716000,China*

[2]*Key Laboratory of Intelligence Information Processing, Institute of Computing Technology, Chinese Academy of Science, Beijing 100080,China.Email:haojs@ics.ict.ac.cn*

Abstract: Principal curve pass through the middle of a multidimensional data set, to express the distributing shape of the points in the data set, we model principal curve for it. The new method of modeling the complex principal curve, based on B-spline network, is proposed. This method combines the polygonal line algorithm of learning principal curve with B-spline network. At one time, the algorithm finding a bifurcate point of the complex principal curve is presented. Our experimental results on simulate data demonstrate that it is feasible and effective.

Key words: Principal Curve, The Polygonal Line Algorithm, B-spline Network, Bifurcate Point

1. INTRODUCTION

Principal curves were firstly introduced by Hastie and Stuetzle [1], and have been defined as satisfying the self-consistency property. Because complicated mathematics idea was used in describing its elements, then it wasn't noticed in computer science domain. At present, although there are a good many of problems on mathematics in the study of principal curve, yet principal curve approaches has attracted attention owning to its advantages and there are many reports on principal curve application. Actual applications involve the domain of visualization of image, speech recognition, time data analysis, pattern classification, recognition of handwritten digits, pattern clustering, process monitoring and so on[6]-[9]. Principal Curves are the nonlinear generalization of first principal components, and have been defined as smooth one-dimensional curves,

which pass through the middle of a multidimensional data set. At present, there are a good many of the principal curve algorithm proposed and these algorithms merely gained discrete points on principal curve without modeling principal curve. By modeling principal curve, its model can express the nonlinear relation among variables in a data set.

B-spline network is the association memory network composed of three-layer structure [3][5], its structure is illustrated in **Fig.1.** B-spline function in latent layer is used as the basic function. For a random input, in latent layer a few B-spline basic function is active and the network output is a linear combination of these active basic function. Since the support set of the basic function is finite region, the network has the following features: a) the knowledge in the network is locally stored without whole and distributed, learning is local. Therefore the learning from a part in input space isn't influence the learning results in other part .b) the learning algorithm converges quickly. The network is convenient for real time application online. Thereby this kind network draws attention and is applied to the field of controlling, modeling, pattern recognition etc [5].

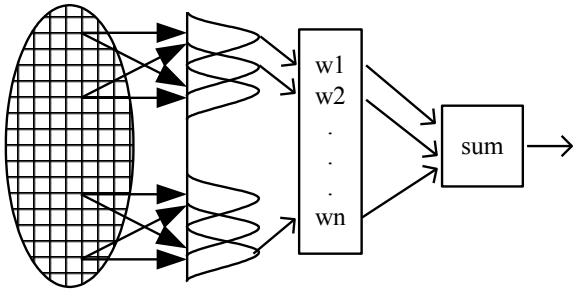

Fig.1. B-spline network structure

In this paper, the method of modeling principal curve is proposed. The kind of principal curve with branches is mainly considered, its form is illustrated in **Fig.3,** at a bifurcate point, the principal curve is partitioned into two branches. In modeling this kind principal curve, a key problem is how a bifurcate point of principal curve is found from a given data set. Thus the problem of modeling principal curve with branches is converted into one of modeling principal curve without branch, therefore, we propose a iterative algorithm for finding a bifurcate point of principal curves (its detail is given in section 3). In the algorithm proposed by us, firstly, the iterative algorithm for finding a bifurcate point of principal curve is used for searching a bifurcate point and the principal curve with branch is partitioned into three branches. Secondly, the polygonal lines algorithm of learning principal curve [3] is respectively applied to three branch and three polygonal lines are

found. Finally, the vertex set of three polygonal lines is regarded as the set of B-spline network training sample and the network is trained. Since B-spline network has the features of a short training time and a fast convergence speed, therefore the approach may quickly create the model of principal curve. To compare with the existent principal curve algorithm, our method may create the model of a smooth principal curve since the basic function of B-spline network is the continuous function.

In the following of this paper, we firstly introduce the definitions of principal curve and the polygonal lines algorithm of learning principal curve in section 2. Then in section 3, we propose the algorithm for searching bifurcate point of principal curve. The new method of modeling principal curves is proposed in section 4. In section 5 our experimental results on simulate data sets are given. Conclusions are provided in the last section.

2. THE DEFINITIONS OF PRINCIPAL CURVES AND THE POLYGONAL LINES ALGORITHM OF LEARNING PRINCIPAL CURVES

In this section, we introduce the definitions of principal curves [1][2][4] and the polygonal lines algorithm of learning principal curves [4].

2.1 Definitions of principal curves

Definition1. The principal curve f of data distributing $D \subset R^d$ with continuous probability density $h(x)$ is a member in manifold M satisfying the self-consistency property. A curve $f \in M$ is the self-consistency if $E(X \mid \lambda_f(X) = \lambda) = f(\lambda), \forall \lambda \in I$, where I is close interval on real number axis, $M = \{M_f : f \subset F\}, M_f = f(D) = \{f(X) : X \in D\}$, F is a function set, to each $f \in F$, $f : D \to R^d$.

Definition2. The smooth curve $f(\lambda)$ is a principal curve if the following hold:

a) $f(\lambda)$ does not intersect itself.

b) $f(\lambda)$ has finite length inside any bounded subset of R^d and.

c) $f(\lambda)$ is self-consistent, i.e. $f(\lambda) = E(X \mid \lambda_f(X) = \lambda)$

Definition3. A curve f^* is called a principal curve of length L for X if f^* minimizes $\Delta(f)$ over all curves of length less than or equal to L. Where $\Delta(f) = E[\Delta(X, f)] = E[\inf \|X - f(\lambda)\|^2] = E[\|X - f(\lambda_f(X))\|^2]$.

According to definition of principal curve, principal curve is a smooth curve of satisfying the self-consistency property. It is essentially low-dimensional manifold embedded in the high-dimensional space. Any point on the curve is the conditional mean of data set over those points of the

space which project to this point, it can factually reflect distributing form of data set

2.2 The polygonal line algorithm of learning principal curve

The polygonal lines algorithm of learning principal curve was proposed by Kégl B [4] , the algorithm is composed of the following steps.

Algorithm1：The Polygonal Line Algorithm；

1. Initialization:

Given a set of data points $X_n = \{(x_1, y_1),(x_2, y_2),\ldots,(x_n, y_n)\} \subset R^2$, the algorithm starts with a straight line segment, the shortest segment of the first principal component line which contains all of the projected data points.

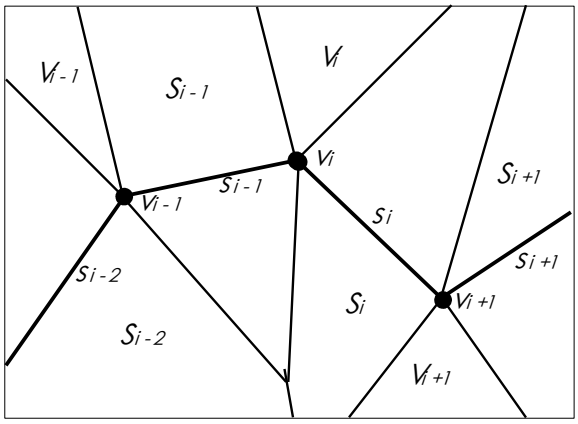

Fig. 2. A nearest-neighbor partition of R^2 induced by the vertices and segments of the polygonal line..

2. The Projection Step：

In the step the data points are partitioned into "nearest neighbor regions" according to which segment or vertex they project and it is illustrated in **Fig.2**. The nearest point of f to any point in the set V_i is the vertex v_i. The nearest point of f to any point in the set S_i is a point of the line segment s_i.

3. The Vertex Optimization Step：

In the step the new position of each vertex v_i is determined in a line search to minimize an objective function that consists an average squared distance term and a curvature penalty. While all other vertices are kept fixed.

4. Adding a New Vertex：

The inner loop consists of a projection step and an optimization step, these two steps are iterated so that the optimization step is applied to each vertex v_i, $i = 1,2,\ldots,k+1$, in a cyclic fashion (so that after v_{k+1}, the procedure starts again with v_1) until convergence is achieved and $f_{k,n}$ is produced. Then, a new vertex is added.

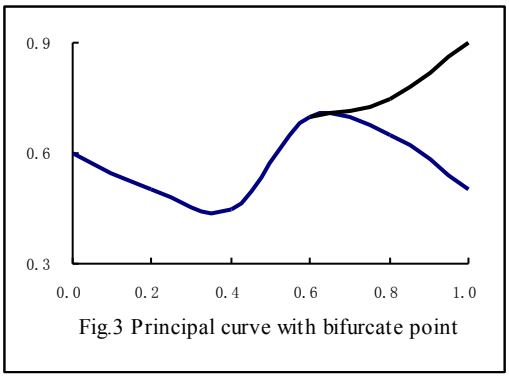

Fig.3 Principal curve with bifurcate point

The algorithm stops when the number k of vertices exceeds a threshold $c(n,\Delta)$. This stopping criterion is based on a heuristic complexity measure, determined by the number of segments k, the number of data points n, and the average squared distance $\Delta_n(f_{k,n})$.

3. THE ALGORITHM FOR FINDING A BIFURCATE POINT OF PRINCIPAL CURVES

In this section, we propose a algorithm of searching a bifurcate point of principal curves.

Consider a set of data points $X_n = \{(x_i, y_i)\} \subset R^2$,where $i \in N = \{1,2,\cdots,n\}$ .Let $x_{\min} = \min_{i \in N}\{x_i\}$, $x_{\max} = \max_{i \in N}\{x_i\}$, $I^{(0)} = [x_{\min}, x_{\max}]$, $\left|I^{(0)}\right|$ denotes the length of interval $I^{(0)}$, (x, y) denotes a bifurcate point for a set of data points X_n.

A algorithm for finding a bifurcate point is given below.

Algorithm2：Algorithm searching bifurcate point;

Input : a set of data points $X_n = \{(x_i, y_i)\} \subset R^2$ $i \in N = \{1,2,\cdots,n\}$

Output: a bifurcate point (x, y) for a set of data points X_n

Process:

{ Let $t = 0$;

Do { the Interval $I^{(t)}$ is parted into m parts, where m is the positive integer.

m small intervals are produced, denoted by I_i, $i = 1,2,\cdots,m$.

Let $J_i = \{y_j \mid \forall (x_j, y_j) \in X_n, x_j \in I_i\}$. $i = 1,2,\cdots,m$;

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

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

n_i denote the number of the data points in I_i.

Where $n_1 + n_2 + \cdots + n_k = n$.

Compute $\{\sigma_{i+1} - \sigma_i\}$, $i = 1,2,\cdots,m-1$;

$\exists i, \ni (\sigma_{j+1} - \sigma_j) \in \{\sigma_{j+1} - \sigma_j\}$ $j = i, i+1, \cdots, k-1$;

$\sigma_{j+1} - \sigma_j \geq 0$, and is increase rigorously;

The abscissa of a bifurcate point $(\overline{x}, \overline{y})$ for X_n is in I_{i+1};

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

{ $t = t + 1$; $I^{(t)} = I_{i+1}$; }

Where ε is the small positive number given

}

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

Let $\overline{x} = \frac{1}{n_{i+1}} \sum_{x_j \in I_{i+1}} x_j$, $\overline{y} = \frac{1}{n_{i+1}} \sum_{y_j \in I_{i+1}} y_j$. $(\overline{x}, \overline{y})$ is a bifurcate point

} The algorithm end.

Since $|I^{(0)}| = x_{\max} - x_{\min}$, $|I^{(1)}| = \frac{1}{m}|I^{(0)}|$, $|I^{(t)}| = \frac{1}{m^t}|I^{(0)}|$, $\lim_{t \to +\infty} |I^{(t)}| = 0$, thus the algorithm is convergence. While m is bigger it converges quickly.

4. A NEW METHOD OF MODELING PRINCIPAL CURVES

In this section, we propose a new method of modeling principal curves, based on the algorithm2 in section 3 for searching a bifurcate point of principal curves and the polygonal line algorithm1 in section 2.2 and B-spline network.

Given a set of data points $X_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subset R^2$.

The basic idea of the algorithm is: **Firstly**, the algorithm2 in section 3 is used for finding the bifurcate point and the bifurcate point $(\overline{x}, \overline{y})$ is found.

Let $X^{(1)} = \{(x_i, y_i) \mid i \in N, (x_i, y_i) \in X_n, x_i < \overline{x}\}$

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

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

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

Distinctly X_n is partitioned into $X^{(1)}$ and $X^{(2)}$ making use of the abscissa $\overline{x}$ for bifurcate point $(\overline{x}, \overline{y})$, where $X_n = X^{(1)} \cup X^{(2)}$, $X^{(1)} \cap X^{(2)} = \phi$; $X^{(2)}$ is partitioned into $X^{(21)}$ and $X^{(22)}$ making use of the ordinate $\overline{y}$ for bifurcate point $(\overline{x}, \overline{y})$. thus X_n is partitioned into $X^{(1)}$, $X^{(21)}$ and $X^{(22)}$,where $X_n = X^{(1)} \cup X^{(21)} \cup X^{(22)}$, $X^{(1)} \cap X^{(21)} \cap X^{(22)} = \phi$. $X^{(1)}$, $X^{(21)}$ and $X^{(22)}$ respectively correspond to the three braches of the principal curves for X_n.

Secondly, by using the polygonal line algorithm1 in section 2.2 three polygonal line is found corresponding to three subset of X_n: $X^{(1)}$, $X^{(21)}$ and

$X^{(22)}$. For each subset, a polygonal $f_{k,n}$ line with k line segments and $k+1$ vertices are gained, different subset has different k value.

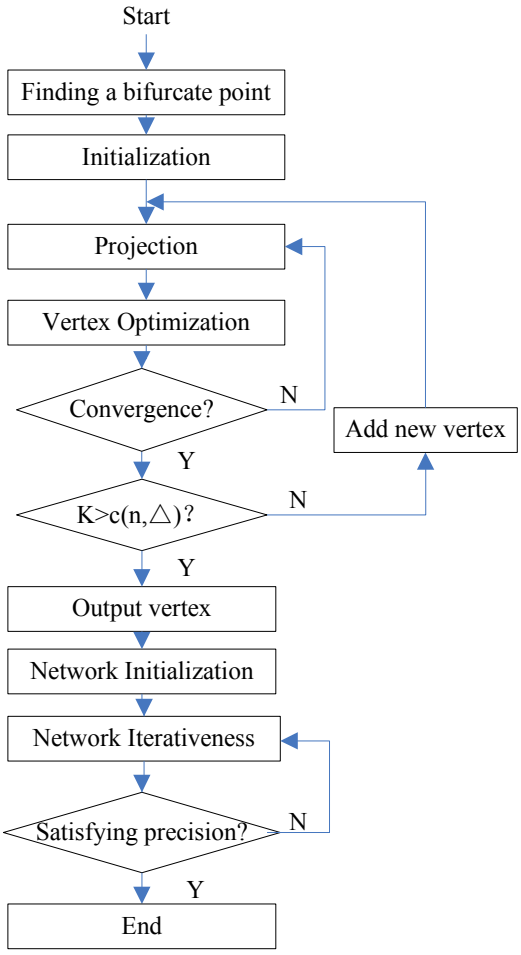

Fig.4.The flow chart of the approach

Finally, $k+1$ vertices of every polygonal line are regarded as the training sample for B-spline network and it is trained. The principal curve for a set of data points $X_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \subset R^2$ is created. While B-spline network is trained, the points in $X^{(21)}$ and $X^{(22)}$ are respectively trained and the training results are respectively stored.

The flow chart of the approach, proposed by us, is given in **Fig.4.**

5. THE EXPERIMENT RESULTS ON SIMULATE DATA SET

To test the algorithm presented above, we conducted experiments on simulate data sets. Our experiment results on simulate data set are given in **Fig.5** and **Fig.6.** We can see that the principal curves constructed with the proposed algorithm have approximated to the origin continuous functions.

In the **Fig.5** consider the $y = \sin x, 0 \le x \le 2\pi$, randomly select 400 points, and add independent Gaussian noise $\varepsilon_i \sim N(0,0.1)$.

In the **Fig.6** consider the $y = \sin x, 0 \le x \le \frac{3}{2}\pi$ and the $y = -\sin x, \pi \le x \le \frac{3}{2}\pi$, randomly select 300 points and 100 points, and add independent Gaussian noise $\varepsilon_i \sim N(0,0.1)$.

Based on above experiment results, we can see the approach of modeling principal curve, combining B-spline network with the polygonal line learning algorithm, may construct a smooth curve model and is a feasible and effective method. The presented learning principal curve algorithm merely gained discrete points on approximate principal curve without modeling principal curve.

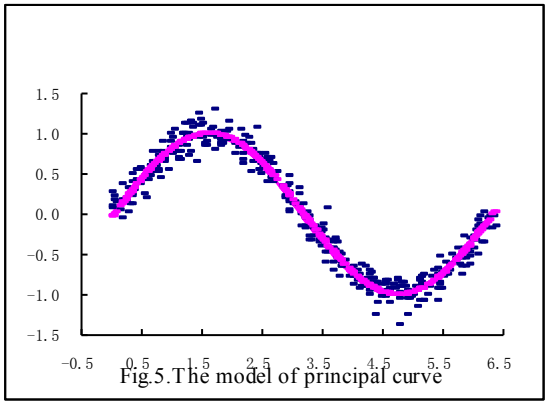

Fig.5.The model of principal curve

6. CONCLUSION

For the data points set given, the proposed approach in the paper, based on B-spline network, can model a smooth principal curve of it. To compare with the presented algorithm of principal curve, **firstly,** our method may construct the model of principal curve for a data points set and this is a improvement and perfection on the presented algorithm of principal curve in some sense;

Secondly, since the basic function of B-spline network is continuous function the principal curve constructed using our proposed method is a smooth curve. **Finally,** because the knowledge in B-spline network is locally stored and the network has the features of a short training time and a fast convergence speed, therefore the approach may quickly create the model of principal curve.

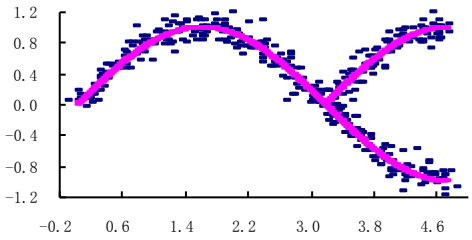

Fig.6 The model of the principal curve with bifurcate point

Our experiment results on simulate data sets demonstrate that the proposed method, based on B-spline network and the polygonal line algorithm of learning principal curve is feasible and effective for modeling principal curve of the data points set given. This method is applied to the fields of modeling and the process controlling and so on.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No.60435010, 90604017), the 973 Project (No.2003CB17004), the Natural Science Foundation of Beijing (No. 4052025).

REFERENCES

1. Hastie T. Principal Curves and surfaces. Laboratory for Computational Statistics, Stanford University, Department of Statistics : Technical Report 11, 1984.
2. Hastie T and Stuetzle W. Principal Curves. Journal of the American Statistical Association. 1989,84: 502-516.
3. Moody J. Fastlearning in multi-resolution hierarchies. Advances in Neural information Processing System, vol.1, 1989 : 29-39.
4. Kégl B, Krzyzak A, Linder T and Zeger K. Learning and design of principal curves. IEEE Trans. On Pattern Analysis and Machine Intelligence. 2000, 22 (3): 281-297.

5. Martin Brown, Chris Harris. Neurofuzzy adaptive modeling and control. Prentice Hall International (UK) Limited, 1994: 89-100.
6. Stanford D. and Raftery A.E. Finding Curvilinear Features in Spatial Point Patterns: Principal Curve Clustering with Noise. IEEE Trans. on Pattern Analysis and Machine Intelligence. 2000,22(6): 601-609.
7. Kegl B and Krzyzak A. Piecewise linear skeletonization using principal curves. IEEE Trans on Pattern Analysis and Machine Intelligence 2002,24(1): 59-74.
8. Hemann T, Meinicke P, and Ritter H. Principal curve sonification. International Conference on Auditory Display 2000: 81-86,
9. Einbeck J, Tutz G, and Evers L. Exploring Multivariate Data Structures with Local Principal Curves". In: C. Weihs and W. Gaul (Eds.): Classification - The Ubiquitous Challenge, Springer, Heidelberg2005: 256-263.