基于 Isomap 的流形结构重建方法

孟德宇 徐 晨 徐宗本

(西安交通大学信息与系统科学研究所 西安 710049)

摘 要 已有的流形学习方法仅能建立点对点的降维嵌入,而未建立高维数据流形空间与低维表示空间之间的相互映射.此缺陷已限制了流形学习方法在诸多数据挖掘问题中的进一步应用.针对这一问题,文中提出了两种新型高效的流形结构重建算法:快速算法与稳健算法.其均以经典的 Isomap 方法内在运行机理为出发点,进而推导出高维流形空间与低维表示空间之间双向的显式映射函数关系,基于此函数即可实现流形映射的有效重建.理论分析与实验结果证明,所提算法在计算速度、噪音敏感性、映射表现等方面相对已有方法具有明显优势.

关键词 数据降维;流形学习;等距特征映射;模式分类;特征描述

中图法分类号 TP18 **DOI 号**: 10.3724/SP.J.1016.2010.00545

A New Manifold Reconstruction Method Based on Isomap

MENG De-Yu XU Chen XU Zong-Ben

(*Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an* 710049)

Abstract Most of the existing nonlinear dimensionality reduction methods only realize data embedding from high-dimensional to low-dimensional data spaces but not data mapping between them, which restrict their applications to approximation and prediction tasks. This paper proposes two new data mapping methods, fast method and robust method respectively, which realizes data mapping from data embedding based on the intrinsic executive mechanism of Isomap, one of the most well known nonlinear dimensionality reduction method. It also presents theoretical estimations for the approximation precision and computational complexity of the new methods. Some experiment results on synthetic and real-world data sets are demonstrated, which verifies the feasibility and effectiveness of the new data mapping methods. Particularly, the simulations, which apply the new methods on feature movie description problem and pattern classification problem, are designed. The results further shows the potential usefulness of the new methods.

Keywords dimensionality reduction; manifold learning; Isomap; pattern classification; feature description

1 引 言

近年来,在数据挖掘、人工智能及信息获取等研究领域中,常常需要处理具有高维描述特征的数据.数据的高维特性往往给应用中的数据处理过程带来计算效率低下、维数灾难等问题.从本质上说,实际中高维数据的属性特征之间常存在一定的规律性和

收稿日期:2007-04-18;最终修改稿收到日期:2009-09-19.本课题得到国家"九七三"重点基础研究发展规划项目基金(2007CB311002)、国家自然科学基金(60905003)和重点基金(70531030)资助.**孟德宇**,男,1978 年生,博士,讲师,主要研究方向为非线性降维、模式识别.E-mail: dymeng@mail.xjtu.edu.cn. **徐 晨**,男,1985 年生,博士研究生,主要研究方向为数理统计、信息科学. **徐宗本**,男,1955 年生,博士,教授,博士生导师,主要研究领域为计算智能、信息科学.

相关性,即实际数据经常存在外在(存储的高维数)与内在(本质的低维数)两个维数.若能获得高维数据的本质低维表示,则一方面由于后者的本质性,可对其进行等效处理从而高效挖掘出前者中蕴涵的信息,另一方面由于后者的简单性,可在一定程度上解决高维特征给数据带来的诸多问题(如维数灾难).因此,寻找高维数据的本质低维结构,即数据降维问题,是极为重要的非监督学习问题之一.

传统的降维方法包括主成分分析(Principle Component Analysis, PCA)[1]与高维尺度分析(Multidimensional Scaling, MDS)[2]等方法.当数据位于高维空间中一个低维线性超平面上时,此类方法能够有效对其实行降维.然而,它们不能对具有低维非线性分布结构的高维数据进行有效降维处理,这大大限制了它们的应用范围.近年来,出现了一类新型的非线性降维方法——流形学习降维方法,此类的典型方法包括等距特征映射(Isometric Mapping, Isomap)[3]、局部线性嵌入(Locally Linear Embedding, LLE)[4]与拉普拉斯特征映射(Laplacian Eigenmap)[5]等.此类方法的特点在于其均假设数据本身具有低维的流形形式(即数据空间呈现由少数独立特征共同作用所张成的低维流形形态).相比其它降维方法,流形学习方法具有很多优势:首先,其计算性能对数据的非线性流形结构具有自适应性;其次,只涉及到较少的参数选择问题;另外,基于非常易于理解的模型构造方式,降维后的数据特征具有很好的可解释性.在人脸识别、手写数字辨识、文本归类、轨迹跟踪等方面的成功应用[3-6]也进一步验证了流形学习降维方法的有效性.然而,目前这些方法的有效应用主要体现在聚类与数据可视化等应用领域中,当面对模式分类、回归分析、时间序列分析等需要预测功能的数据挖掘问题时,这些方法便会失效.主要原因是其仅实现了位于高维流形上有限数据集的低维表示(点与点的嵌入),但并未建立高维流形空间与对应低维表示空间之间的相互映射关系(集合与集合的映射),这使其无法获得一个新输入高维(或低维)空间数据在对应低维表示空间(或高维流形空间)的映射表示.此问题已成为限制流形学习方法进一步扩展应用的主要瓶颈[7-9].因此,如何实现流形空间与其对应低维表示空间之间的映射关系,或者说如何基于流形产生的数据完整准确地重建流形结构,是目前流形学习领域亟需解决的主要问题之一.

目前已提出了一些流形重建方法[10].其中最具代表性的包括由 Isomap 方法原理延伸构造的 Landmark Isomap 方法(L-Isomap)[11-12]、由 LLE 方法原理构建的 Extending LLE 方法(E-LLE)[13]、利用 Laplacian Eigenmap 类似原理而提出的 Locality Presearving Projections 方法(LPP)[14]与其非线性核化方法以及其它一些方法[15].这些方法在应用中的表现各有优劣,如 L-Isomap 方法具有鲁棒的流形映射表现,然而对于每个新输入数据此方法需利用图论技术估计其与所有现有数据间的测地距离,因而导致算法计算速度较慢;E-LLE 方法对于无噪音数据表现良好,然而由于 LLE 方法的计算性能对于噪音干扰反应敏感,因此往往导致基于 LLE 方法的 E-LLE 流形映射结果鲁棒性较差;LPP 方法本质为针对非线性问题设计的线性方法,计算速度极快,但由于其本质的线性特征,对于具有非线性结构的数据往往计算失效.

针对以上问题,本文提出了两种全新的高效流形结构重建方法.一种为计算速度较快的快速算法,适用于无噪音非线性结构的流形重建问题;另一种为鲁棒性较高的稳健算法,适用于更大范围的流形重建问题.所提方法的基本构建思想是通过研究 Isomap 运行的内在机理,概括出其本质要求的 3 条假设,在这 3 条假设的前提下,导出了流形空间与其对应低维表示空间之间显式的映射关系函数,进而实现了高效流形重建.分别在标准流形数据集上对新方法进行了无噪音干扰、带噪音干扰、计算速度与对不同程度噪音敏感程度等方面的测试,实验结果证明新方法相对已有方法具有明显优势.另外,我们亦将新方法应用于降维特征动画描述、模式分类等应用中,展示了新方法的潜在应用价值.

本文第 2 节将对相关的背景知识进行介绍;第 3 节介绍所提快速与稳健流形结构重建算法的实现步骤及其构建机理,并对其计算复杂度进行理论分析;第 4 节介绍新方法在标准数据集上的仿真实验效果,并展示其在降维特征动画描述、模式分类等方面的应用效果;最后对全文进行总结.

2 Isomap 概述

Isomap 方法是一种利用全局数据信息实现数据降维的流形学习降维方法,其主要思想是利用局部邻域距离对数据点间的全局流形测地线距离进行估计,通过建立原数据间测地距离与降维数据间空间距离的对等关系从而实现数据降维.由于测地距

离一般能够内在地反映数据的本质流形几何特征，Isomap 常可以成功地找到高维数据本质对应的低维嵌入[3,11]. 其主要步骤如下：

(1) 建立邻域图. 定义 V 为原数据集合，E 为连接所有邻域数据对的边集合(一般取 ε-邻域或 k-邻域)，从而建立邻域图 $G=(V,E)$；

(2) 计算测地距离. 计算 V 中任意两节点在邻域图 G 中的最短路径，将此最短路径值作为对应节点间的近似测地距离估计；

(3) 数据嵌入. 将(2)中获得的数据测地距离矩阵作为输入，应用经典的 MDS 方法计算数据最终低维嵌入表示.

本质上，Isomap 方法的有效性首先要求当两节点足够近时，它们之间的距离与其低维嵌入之间的距离近似等同，我们称此要求为局部等距假设；另外还要求所有节点非常稠密地分布在其所处的流形上，我们称此要求为稠密性假设. 这两条假设一方面确保了流形的基本形状能够被数据近似表达，另一方面保证了两数据节点的流形测地距离可由邻域图中其间最短路径近似计算，进而保障了对 Isomap 方法中流形距离计算的合理性. 另外，数据所处的流形被默认为光滑连续，这是由局部等距假设与低维嵌入的连续性导出的自然结论，我们称此为连续性假设. 我们下面将基于这 3 条假设的数学表述形式导出流形结构重建的策略. 在给出相应的表述之前，有必要首先标准化一些记号的意义如下.

记数据所处的高维流形集为 $\Omega_n\subset R^n$，对应的低维表示集为 $\Omega_d\subset R^d$；两集合相互间的映射分别记为 $f:\Omega_d\to\Omega_n$，$g:\Omega_n\to\Omega_d$；初始给定的训练数据集记为 $\{\boldsymbol{x}_i\}_{i=1}^l\subset\Omega_n$，对应的低维表示数据记为 $\{\boldsymbol{y}_i\}_{i=1}^l\subset\Omega_d$，$\boldsymbol{x}_i=f(\boldsymbol{y}_i)$，$\boldsymbol{y}_i=g(\boldsymbol{x}_i)$，$i=1,2,\cdots,l$. 数据间距离均默认取为 2 范数距离. 默认采用的邻域类型为 ε-邻域，记位于 $\boldsymbol{x}_i$ 的 ε-邻域内的数据为 $\{\boldsymbol{x}_{i_j}\}_{j=1}^{l_i}$，其中 l_i 为 $\boldsymbol{x}_i$ 的邻域个数，对应的低维表示记为 $\{\boldsymbol{y}_{i_j}\}_{j=1}^{l_i}$. 记在 Ω_d 上的 ε-邻域为 $O(\boldsymbol{y}_i,\varepsilon)$.

运用这些记号，我们给出 3 条假设的数学表述形式如下.

假设 1(连续性假设，A1).　Ω_d，Ω_n 为有界闭集，$f\in C^2(\Omega_d,\Omega_n)$.

假设 2(局部保距假设，A2).

$$\lim_{\substack{\varepsilon\to 0\\ \boldsymbol{y}_1,\boldsymbol{y}_2\in O(\boldsymbol{y}_i,\varepsilon)}}\frac{\|f(\boldsymbol{y}_1)-f(\boldsymbol{y}_2)\|}{\|\boldsymbol{y}_1-\boldsymbol{y}_2\|}=1;$$

且存在常数 $\gamma\in(0,1)$，使得任意 $\boldsymbol{x}_i$ 的 ε-邻域数据 $\boldsymbol{x}$ 满足 $\left|\dfrac{\|\boldsymbol{x}-\boldsymbol{x}_i\|}{\|g(\boldsymbol{x})-g(\boldsymbol{x}_i)\|}-1\right|<\gamma$.

假设 3(稠密性假设，A3).　$\{\boldsymbol{x}_i\}_{i=1}^l$ 构成 Ω_n 的 ε-覆盖.

容易理解，上面所列的表述 A1，A2 与 A3 分别对应于 Isomap 有效所需满足的 3 个假设要求. 其中特别需要对假设 A2 进行如下合理性说明：假设 A2 的前半部分本质上等价于：对于 Ω_d 上距离足够近的 $\boldsymbol{y}_1,\boldsymbol{y}_2$，$\|f(\boldsymbol{y}_1)-f(\boldsymbol{y}_2)\|=\|\boldsymbol{y}_1-\boldsymbol{y}_2\|$ 成立. 实际上，这是 Isomap 有效的基本条件，在以往关于 Isomap 方法的所有理论研究中，此假设均默认成立[16-19]. 后半部分为此等距条件的一致性推广，其本质含义是对于互为 ε-邻域的 $\boldsymbol{x}$，$\boldsymbol{x}_i$ 两点，其间距离与对应低维数据 $g(\boldsymbol{x})$，$g(\boldsymbol{x}_i)$ 间距离不致偏差过大，以致产生 $\left|\dfrac{\|\boldsymbol{x}-\boldsymbol{x}_i\|}{\|g(\boldsymbol{x})-g(\boldsymbol{x}_i)\|}-1\right|\to\infty$ 的异常情形. 显然，这个是一个较弱的条件，特别在连续性假设的前提下，对于一般的流形映射，此条件均能够保证成立.

在下节中，我们将给出本文的主要结果，主要包括基于 Isomap 的流形结构重建方法的基本构建机理与实现步骤，并对算法进行计算可信度与复杂度分析.

3　主要结果

基于 Isomap 方法所隐含采用的假设 A1～A3，可构造出流形结构的重建方法，特别地，可获得流形集 Ω_n 与低维表示集 Ω_d 之间的相互映射关系解析函数式. 下节中，将对此构建机理进行详细介绍.

3.1　算法构建机理

在假设 A1 的连续性条件下，容易推导出流形映射 $f:\Omega_d\to\Omega_n$ 能够在任一点 $\boldsymbol{y}_i$ 泰勒展开为

$$f(\boldsymbol{y})=f(\boldsymbol{y}_i)+\boldsymbol{J}(\boldsymbol{y}_i)(\boldsymbol{y}-\boldsymbol{y}_i)+\boldsymbol{R}(\boldsymbol{y},\boldsymbol{y}_i),$$

即

$$f(\boldsymbol{y})=\boldsymbol{x}_i+\boldsymbol{J}(\boldsymbol{y}_i)(\boldsymbol{y}-\boldsymbol{y}_i)+\boldsymbol{R}(\boldsymbol{y},\boldsymbol{y}_i)\qquad(1)$$

其中 $\boldsymbol{J}(\boldsymbol{y}_i)=(\boldsymbol{v}_1(\boldsymbol{y}_i),\cdots,\boldsymbol{v}_d(\boldsymbol{y}_i))\in R^{n\times d}$ 为 f 在 $\boldsymbol{y}_i$ 点的雅各比矩阵，$\boldsymbol{R}(\boldsymbol{y},\boldsymbol{y}_i)$ 为泰勒余项，显然有

$$\lim_{\boldsymbol{y}\to\boldsymbol{y}_i}\frac{\|\boldsymbol{R}(\boldsymbol{y},\boldsymbol{y}_i)\|}{\|\boldsymbol{y}-\boldsymbol{y}_i\|}=0\ \text{与}\ \sup_{\boldsymbol{y}\in\Omega_d}\frac{\|\boldsymbol{R}(\boldsymbol{y},\boldsymbol{y}_i)\|}{\|\boldsymbol{y}-\boldsymbol{y}_i\|}\le B,$$

其中 B 为固定的有界正常数. 利用这些信息，可获得以下定理.

定理 1.　在假设 A1～A2 下，若 $\boldsymbol{y}_i$ 在 Ω_d 中的 ε-邻域具有恒常维数 d，则下式成立

$$\boldsymbol{J}(\boldsymbol{y}_i)^{\mathrm{T}}\boldsymbol{J}(\boldsymbol{y}_i)=\boldsymbol{I}\qquad(2)$$

即列向量$(v_1(y_i),\cdots,v_d(y_i))$为单位向量且相互正交.

证明. 由式(1),易得

$$\|f(y)-f(y_i)\|^2=(y-y_i)^{\mathrm{T}}J(y_i)^{\mathrm{T}}J(y_i)(y-y_i)+2R(y,y_i)^{\mathrm{T}}J(y_i)(y-y_i)+R(y,y_i)^{\mathrm{T}}R(y,y_i),$$

则有

$$\frac{\|f(y)-f(y_i)\|^2}{\|y-y_i\|^2}=\left(\frac{y-y_i}{\|y-y_i\|}\right)^{\mathrm{T}}J(y_i)^{\mathrm{T}}J(y_i)\left(\frac{y-y_i}{\|y-y_i\|}\right)+\frac{2R(y,y_i)^{\mathrm{T}}J(y_i)(y-y_i)+R(y,y_i)^{\mathrm{T}}R(y,y_i)}{\|y-y_i\|^2},$$

等式两边同时 $y\to y_i$,根据假设 A2,可知下式成立

$$1=v_j^{\mathrm{T}}J(y_i)^{\mathrm{T}}J(y_i)v_j \tag{3}$$

由于 y_i的ε-邻域维数为 d,因此可找到 d 个线性无关向量 $v_1,\cdots,v_d$使式(3)成立.

另,由式(1)可得

$$f(y_1)-f(y_i)=J(y_i)(y_1-y_i)+R(y_1,y_i),$$

$$f(y_2)-f(y_i)=J(y_i)(y_2-y_i)+R(y_2,y_i),$$

其中 $y_1=\Delta\cdot v_j$, $y_2=\Delta\cdot v_k$,Δ 为放缩变量,则易知

$$f(y_1)-f(y_2)=J(y_i)(y_1-y_2)+R(y_1,y_i)-R(y_2,y_i),$$

可自然推出

$$\|f(y_1)-f(y_2)\|^2=(y_1-y_2)^{\mathrm{T}}J(y_i)^{\mathrm{T}}J(y_i)(y_1-y_2)+2(R(y_1,y_i)-R(y_2,y_i))^{\mathrm{T}}J(y_i)(y_1-y_2)+\|R(y_1,y_i)-R(y_2,y_i)\|^2,$$

两边同除以 $\|y_1-y_2\|^2$,则有

$$\frac{\|f(y_1)-f(y_2)\|^2}{\|y_1-y_2\|^2}=\left(\frac{y_1-y_2}{\|y_1-y_2\|}\right)^{\mathrm{T}}J(y_i)^{\mathrm{T}}J(y_i)\left(\frac{y_1-y_2}{\|y_1-y_2\|}\right)+\frac{2(R(y_1,y_i)-R(y_2,y_i))^{\mathrm{T}}J(y_i)(y_1-y_2)}{\|y_1-y_2\|^2}+\frac{\|R(y_1,y_i)-R(y_2,y_i)\|^2}{\|y_1-y_2\|^2},$$

等式两边同时使 $\Delta\to 0$,根据假设 A2,可得

$$v_j^{\mathrm{T}}(J(y_i)^{\mathrm{T}}J(y_i)-I)v_k=0 \tag{4}$$

由式(3)、(4),可得

$$V^{\mathrm{T}}(J(y_i)^{\mathrm{T}}J(y_i)-I)V=0,$$

其中 $V=(v_1,\cdots,v_d)$. 显然 V 可逆,则自然可得 $J(y_i)^{\mathrm{T}}J(y_i)=I$. 证毕.

可由此定理获得式(1)的下列转换形式:

$$y=y_i+J(y_i)^{\mathrm{T}}(f(y)-f(y_i))+J(y_i)^{\mathrm{T}}R(y,y_i),$$

即

$$g(x)=y_i+J(y_i)^{\mathrm{T}}(x-x_i)+J(y_i)^{\mathrm{T}}R(y,y_i) \tag{5}$$

将 y_{i_j}(对应于 x_i的ε-邻域数据 x_{i_j} 的低维表示)代入式(1)中,可得

$$X_i=J(y_i)Y_i+R_i \tag{6}$$

其中 $X_i=(x_{i_1}-x_i,\cdots,x_{i_{l_i}}-x_i)\in R^{n\times l_i}$,$Y_i=(y_{i_1}-y_i,\cdots,y_{i_{l_i}}-y_i)\in R^{d\times l_i}$,$R_i=(R(y_{i_1},y_i),\cdots,R(y_{i_{l_i}},y_i))$.

对式(6)两边均乘以 Y_i^{T},可得到

$$X_iY_i^{\mathrm{T}}=J(y_i)Y_iY_i^{\mathrm{T}}+R_iY_i^{\mathrm{T}},$$

当 $Y_iY_i^{\mathrm{T}}$可逆时,成立

$$J(y_i)=X_iY_i^{\mathrm{T}}(Y_iY_i^{\mathrm{T}})^{-1}-R_iY_i^{\mathrm{T}}(Y_iY_i^{\mathrm{T}})^{-1}.$$

为保证上式更普适的可计算性,实际采用以下近似式

$$J(y_i)\approx X_iY_i^{\mathrm{T}}(Y_iY_i^{\mathrm{T}})^{G}-R_iY_i^{\mathrm{T}}(Y_iY_i^{\mathrm{T}})^{G} \tag{7}$$

其中$(Y_iY_i^{\mathrm{T}})^{G}$表示对 $Y_iY_i^{\mathrm{T}}$求伪逆. 由于

$$\sup_{y\in\Omega_d}\frac{\|R(y,y_i)\|}{\|y-y_i\|}\le B$$

与假设 A1,易知存在 $C>B$ 使下式Ω_d上一致成立:

$$\|R(y,y_i)\|\le C\|y-y_i\|^2 \tag{8}$$

根据假设 A2,可知对任意 x_i的ε-邻域数据 x 成立 $\left|\frac{\|x-x_i\|}{\|y-y_i\|}-1\right|<\gamma$,则有

$$\|y-y_i\|<\frac{1}{1-\gamma}\|x-x_i\|=O(\varepsilon) \tag{9}$$

由式(7)~(9),可以得到

$$\|J(y_i)-X_iY_i^{\mathrm{T}}(Y_iY_i^{\mathrm{T}})^{G}\|\approx\|R_iY_i^{\mathrm{T}}(Y_iY_i^{\mathrm{T}})^{G}\|=O(\varepsilon^2) \tag{10}$$

因此可引入

$$Q_i=X_iY_i^{\mathrm{T}}(Y_iY_i^{\mathrm{T}})^{G} \tag{11}$$

作为 $J(y_i)$的近似. 根据式(1)、(5)与式(11),可得如下近似表达

$$y\approx y_i+Q_i^{\mathrm{T}}(x-x_i)+J(y_i)^{\mathrm{T}}R(y,y_i) \tag{12}$$

与

$$x\approx x_i+Q_i(y-y_i)+R(y,y_i) \tag{13}$$

由于 $\|J(y_i)^{\mathrm{T}}R(y,y_i)\|=O(\varepsilon^2)$与 $\|R(y,y_i)\|=O(\varepsilon^2)$,可进一步近似得

$$y\approx y_i+Q_i^{\mathrm{T}}(x-x_i)$$

与

$$x\approx x_i+Q_i(y-y_i)$$

即

$$g(x)\approx\tilde{g}(x)=y_i+Q_i^{\mathrm{T}}(x-x_i) \tag{14}$$

与

$$f(y)\approx\tilde{f}(y)=x_i+Q_i(y-y_i) \tag{15}$$

则可获得流形映射 f 与 g 函数的近似表达 $\tilde{f}$ 与 $\tilde{g}$. 借助此近似映射函数即可实现流形结构重建.

3.2　流形结构重建的快速算法与稳健算法

基于上节所获的结果(式(14)、(15)),即可直接构造一种快速的流形结构重建方法. 其中包含两个阶段,第 1 阶段对原数据进行预处理,获得流形重建所需的一些必要数据;第 2 阶段实现 Ω_d 与 Ω_n 之间的映射关系. 算法如下.

算法 1.　基于 Isomap 的流形结构重建快速算法.

输入:训练数据集 $\{\boldsymbol{x}_i\}_{i=1}^{l}\subset\Omega_n\subset R^n$;邻域尺寸 ε

输出:任意 $\boldsymbol{x}_0\in\Omega_n$ 的映射 $\boldsymbol{y}\in\Omega_d$ 或任意 $\boldsymbol{y}_0\in\Omega_d$ 的映射 $\boldsymbol{x}\in\Omega_n$

阶段 1(预备部分).

1. 应用 ε-邻域 Isomap 方法计算 $\{\boldsymbol{x}_i\}_{i=1}^{l}$ 的低维表示 $\{\boldsymbol{y}_i\}_{i=1}^{l}$;

2. 构造两矩阵如下:$\boldsymbol{X}_i=(\boldsymbol{x}_{i_1}-\boldsymbol{x}_i,\cdots,\boldsymbol{x}_{i_{l_i}}-\boldsymbol{x}_i)\in R^{n\times l_i}$,$\boldsymbol{Y}_i=(\boldsymbol{y}_{i_1}-\boldsymbol{y}_i,\cdots,\boldsymbol{y}_{i_{l_i}}-\boldsymbol{y}_i)\in R^{d\times l_i}$,其中 $\{\boldsymbol{x}_{i_j}\}_{j=1}^{l_i}$ 为 $\boldsymbol{x}_i$ 的 ε-邻域数据;

3. 计算 $\boldsymbol{Q}_i=\boldsymbol{X}_i\boldsymbol{Y}_i^{\mathrm{T}}(\boldsymbol{Y}_i\boldsymbol{Y}_i^{\mathrm{T}})^G\in R^{n\times d}$,其中 $(\boldsymbol{Y}_i\boldsymbol{Y}_i^{\mathrm{T}})^G$ 为 $\boldsymbol{Y}_i\boldsymbol{Y}_i^{\mathrm{T}}$ 的广义逆.

阶段 2(流形映射).

1. 若输入为 $\boldsymbol{x}_0$,则

1.1. 从 $\{\boldsymbol{x}_i\}_{i=1}^{l}$ 中搜索 $\boldsymbol{x}_0$ 的最近邻 $\boldsymbol{x}_s$;

1.2. 计算 $\boldsymbol{y}=\tilde{g}(\boldsymbol{x}_0)=\boldsymbol{y}_s+\boldsymbol{Q}_s^{\mathrm{T}}(\boldsymbol{x}_0-\boldsymbol{x}_s)$;

1.3. 输出 $\boldsymbol{y}$.

2. 若输入为 $\boldsymbol{y}_0$,则

2.1. 从 $\{\boldsymbol{y}_i\}_{i=1}^{l}$ 中搜索 $\boldsymbol{y}_0$ 的最近邻 $\boldsymbol{y}_s$;

2.2. 计算 $\boldsymbol{x}=\tilde{f}(\boldsymbol{y}_0)=\boldsymbol{x}_s+\boldsymbol{Q}_s(\boldsymbol{y}_0-\boldsymbol{y}_s)$;

2.3. 输出 $\boldsymbol{x}$.

在上述算法的阶段 2 中,分别获得了流形集与其对应低维表示集之间显式的映射函数 $x=\tilde{f}(\boldsymbol{y})$ 与 $y=\tilde{g}(\boldsymbol{x})$,即 $\tilde{f}$ 与 $\tilde{g}$ 近似地反映了产生数据的隐含流形结构.

显然,此方法步骤简单,非常易于执行. 注意到,快速算法仅仅使用了与待预测样本相关的最近邻信息,这个特点一方面使算法具有很快的运算速度,而另一方面却可能导致算法对于噪声干扰的敏感性. 为了加强所提方法的鲁棒性质,我们进一步对其改进实现了流形结构重建的稳健算法. 其与快速算法的主要区别为综合利用了待预测样本的所有 ε 近邻信息,而非仅仅最近邻信息,对流形映射进行构建. 具体地,稳健算法分别将待预测样本的 ε 近邻数据进行流形映射,然后将获得的若干映射向量进行加权平均而求得最终的映射表示向量. 根据连续性假设可推出:数据间距离越近其对应映射应越相似,因此稳健算法中权值可简单设置为待预测样本与其近邻距离的倒数. 基于以上描述,可给出稳健算法的具体步骤. 算法亦分为两个部分,其中第 1 部分与快速算法完全相同,因此仅给出了第 2 部分步骤.

算法 2.　基于 Isomap 的流形结构重建稳健算法.

阶段 2(流形映射).

1. 若输入为 $\boldsymbol{x}_0$,则

1.1. 从 $\{\boldsymbol{x}_i\}_{i=1}^{l}$ 中搜索 $\boldsymbol{x}_0$ 的 ε 近邻 $\boldsymbol{x}_{s_1},\cdots,\boldsymbol{x}_{s_{l_0}}$;

1.2. 计算 $\tilde{g}_j(\boldsymbol{x}_0)=\boldsymbol{y}_{s_j}+\boldsymbol{Q}_{s_j}^{\mathrm{T}}(\boldsymbol{x}_0-\boldsymbol{x}_{s_j})$,$1\le j\le l_0$;

1.3. 计算 $\boldsymbol{y}=\dfrac{\sum_{j=1}^{l_0}\alpha_j\tilde{g}_j(\boldsymbol{x}_0)}{\sum_{j=1}^{l_0}\alpha_j}$,其中 $\alpha_j=\dfrac{1}{\|\boldsymbol{x}_0-\boldsymbol{x}_{s_j}\|}$,$1\le j\le l_0$;

1.4. 输出 $\boldsymbol{y}$.

2. 若输入为 $\boldsymbol{y}_0$,则

2.1. 从 $\{\boldsymbol{y}_i\}_{i=1}^{l}$ 中搜索 $\boldsymbol{y}_0$ 的 ε 近邻 $\boldsymbol{y}_{s_1},\cdots,\boldsymbol{y}_{s_{l_0}}$;

2.2. 计算 $\tilde{f}_j(\boldsymbol{y}_0)=\boldsymbol{x}_{s_j}+\boldsymbol{Q}_{s_j}(\boldsymbol{y}_0-\boldsymbol{y}_{s_j})$,$1\le j\le l_0$;

2.3. 计算 $\boldsymbol{x}=\dfrac{\sum_{j=1}^{l_0}\beta_j\tilde{f}_j(\boldsymbol{y}_0)}{\sum_{j=1}^{l_0}\beta_j}$,其中 $\beta_j=\dfrac{1}{\|\boldsymbol{y}_0-\boldsymbol{y}_{s_j}\|}$,$1\le j\le l_0$;

2.4. 输出 $\boldsymbol{x}$.

需要说明的是,以上所提快速算法与稳健算法均默认采用 ε-邻域. 实际上,算法若采用 k-邻域,过程完全不变,计算可行性也完全不受影响,实验计算效果亦佳.

3.3　算法复杂度分析

下面我们对所提算法的计算复杂度进行分析,以进一步说明新方法的良好计算性能,并将 L-Isomap、E-LLE 与 LPP 作为对比方法亦进行复杂度分析.

新方法的空间复杂度为 $O(ndl)$. 实际上,在算法运行过程中,我们所有需要存储的量包括 $\boldsymbol{x}_i$,$\boldsymbol{y}_i$,$\boldsymbol{Q}_i$,$i=1,2,\cdots,l$ 与 ε,因此总共需要 $(nl+dl+ndl+1)=O(ndl)$ 单位的存储空间. 显然,其空间复杂度随原数据维数 n,降维表示维数 d 与样本个数 l 增长呈线性增长趋势.

新方法阶段 1 的时间复杂度为 $O(l^3+dl^2\log l+nd^2l)$. 具体地,在阶段 1 中,步骤 1 需要在训练数据上执行 Isomap 方法,时间代价为 $O(l^3+dl^2\log l)$[11];容易计算步骤 2 耗费的计算复杂度为 $O(l^2)$;在步骤 3 中,计算所有 $\boldsymbol{Q}_i$ 所需要的时间复杂度为 $O(nd^2l)$.

快速算法与稳健算法阶段 2 所需的时间复杂度分别为 $O(nl)$与 $O(knl)$，其中 k 为数据 ε 近邻的最大个数. 具体地，步 1 中，两算法步骤 1.1 搜索最近邻数据的时间复杂度最多为 $O(nl)$与 $O(knl)$，其步骤 1.2 的计算复杂度分别为 $O(nd)$与 $O(knd)$，稳健算法步骤 1.3 的复杂度为 $O(k)$；由于在一般情况下 $l>d$ 成立，因此总共耗费的算法复杂度分别为 $O(nl)$与 $O(knl)$. 步 2 中，两算法步骤 2.1 的时间复杂度分别为 $O(dl)$与 $O(kdl)$，步骤 2.2 的时间复杂度为 $O(nd)$与 $O(knd)$，稳健算法步骤 2.3 的复杂度为 $O(k)$，由于 $n>d$ 成立，因此总共耗费的算法复杂度也分别为 $O(nl)$与 $O(knl)$.

注意到，阶段 2 要低于阶段 1 的计算复杂度. 事实上，在算法执行完预备部分之后，所有的 $\boldsymbol{Q}_i$，$i=1,2,\cdots,l$ 得以存储，此时，$\tilde{f}$ 与 $\tilde{g}$ 的形式已完全确定，即流形结构已完全固定. 此时，对任意的一个或多个处于高维流形集或低维表示集中的待预测数据，只需执行流形映射阶段即可获得最终映射结果. 因此，在实际应用快速算法与稳健算法时，数据映射本质上只需耗费 $O(nl)$与 $O(knl)$的时间代价.

下面对 3 种代表性的流形重建方法计算复杂度进行对比分析. L-Isomap 方法首先需计算待预测数据与所有已有数据间的测地距离，时间复杂度为 $O(knl\log l)$，然后需执行 L-MDS 方法，时间代价为 $O(dl)$，因此需要的总时间复杂度为 $O(kdnl\log l)$；其空间复杂度主要为存储训练数据集与其测地距离矩阵，共需约 $O(nl^2)$的存储空间；E-LLE 方法亦包含两个步骤，计算重建权值需耗费的时间代价为 $O(k^3nl)$，对输入数据进行加权映射需要的时间代价为 $O(kd)$，因此总的时间复杂度为 $O(k^3dnl)$；算法需存储权值向量，训练数据集及其降维表示，至少需 $O(kdnl)$的存储空间；LPP 方法本质为线性映射方法，只需通过对待预测数据进行矩阵相乘即可完成数据映射，其时间复杂度与空间复杂度显然均为 $O(nd)$.

通过以上分析结果可知，相比其它非线性流形重建方法，所提算法无论从空间复杂度还是从时间复杂度都具有明显优势，即其具有良好的可扩展性(scalability). 而由于 LPP 方法的简单线性映射方式，其所需的时间代价与空间代价都少于所提算法. 然而，此类线性方法对于非线性流形重建问题一般将会产生明显的计算失效，因此其计算性能一般与所提非线性方法具有明显差距. 以下将通过展示在标准数据集上的一系列数值实验结果对以上分析结论进行进一步验证.

4 实验结果

本节实验包括两个部分，第 1 部分为仿真实验部分，目的是利用标准流形数据集对所提快速与稳健算法进行性能测试，以充分验证其综合的优良性能；第 2 部分为应用实验部分，旨在通过两个应用方向展示所提算法的潜在应用价值. 所有程序均采用 Matlab 7.0 作为运行平台编程实现，计算环境为具有如下配置的个人计算机：中央处理器为 Intel Pentium 1.8GHz，内存为 512MB，操作系统为 Windows XP.

4.1 仿真实验测试结果

仿真实验采用的流形数据是经典的 Swiss roll 流形数据集. Swiss Roll 流形(如图 1(a)所示)分布在 3 维空间上，具有 2 维本质结构. 在流形学习研究

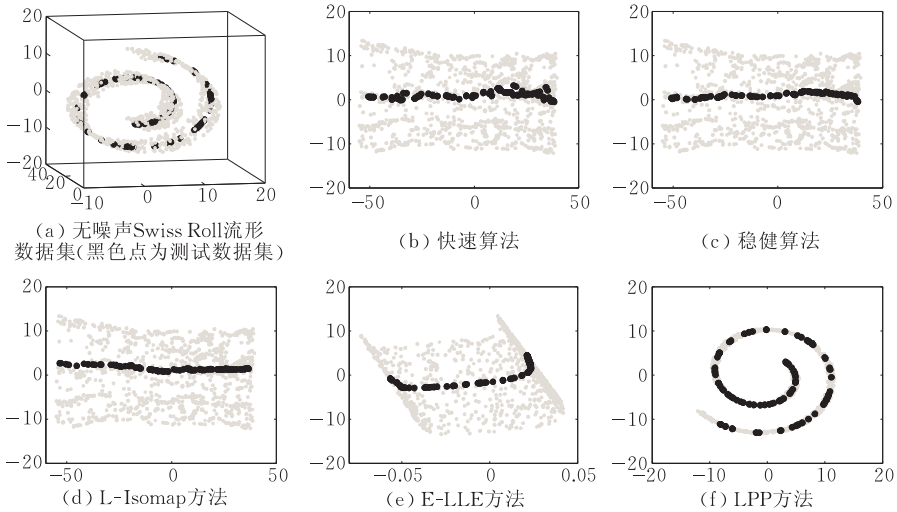


图 1 无噪声数据集上的映射结果

中，此类流形是公认的对检测流形学习方法有效性具有最优良直观展示性能的流形结构，因而此类型的数据集也是目前被最广泛采用的标准 Benchmark 数据集[3-5,11,13]. 以下分别展示所提算法在无噪音干扰与带噪音干扰情况下的 Swiss Roll 流形数据集上的计算效果，并对其进行计算速度与噪音敏感程度测试. 将 L-Isomap、E-LLE、LPP 作为对比方法来对新算法性能进行比较说明.

4.1.1　无噪声数据测试

本组实验采用的训练数据为由 Swiss Roll 型流形分布随机产生的规模为 1000 的无噪声数据集，测试数据集为沿流形上的一条直线随机产生的规模为 100 的流形数据集，如图 1(a)所示. 分别利用所提快速算法与稳健算法、L-Isomap 方法、E-LLE 方法、LPP 方法对上述训练数据进行学习，并利用所得的流形映射对测试数据向其对应低维表示空间进行映射，计算结果分别如图 1(b)～(f)所示. 其中，每种方法均在不同参数下进行多次计算，图 1 展示的为各方法所得的最佳映射结果.

由此图易观察到，LPP 线性方法计算结果失效，并未恢复出流形本质的映射结构. 而其它 4 种非线性方法均具有良好的表现. 其中特别对于所提稳健算法与 L-Isomap 方法，其映射效果相对更佳：其均近似获得了测试数据集本质的连续直线形态.

4.1.2　带噪声数据测试

本组实验采用的训练数据为规模为 1000 的 Swiss Roll 流形数据集，其中所有数据均附加了幅度为[−0.2,0.2]的随机噪声；测试数据集则仍为上述实验中应用的位于流形直线上规模为 100 的无噪声数据集，如图 2(a)所示. 仍采用同以上实验的 5 种流形重建方法对测试数据集计算其映射表示，计算结果如图 2(b)～(f)所示.

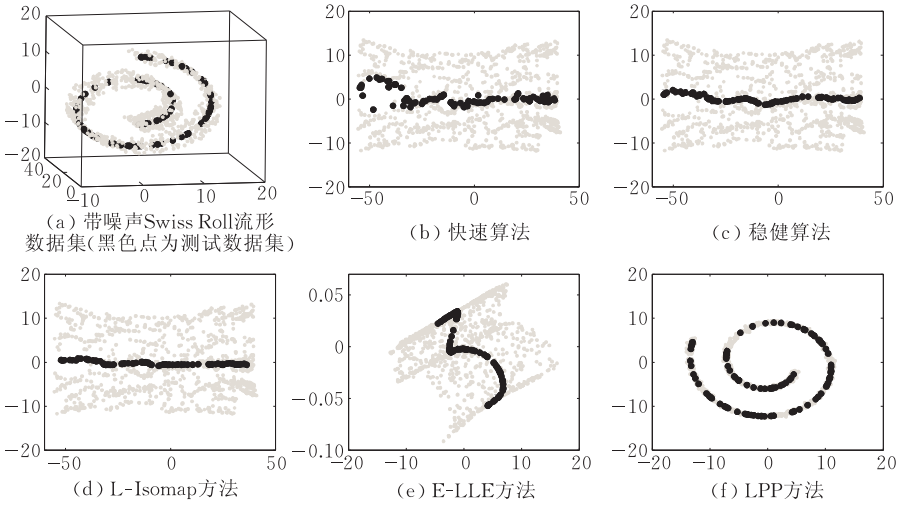


图 2　带噪声数据集上的映射结果

易从图 2 中观察出如下事实：LPP 方法显然计算失效；所提快速算法的映射结果则明显受到噪声干扰的负面影响，沿其本质所在直线发生了较大的浮动偏差；E-LLE 方法所获映射亦未准确地反映出预测数据本质的直线形态，具有较严重的扭曲变形问题. 相对来说，所提稳健算法与 L-Isomap 方法具有明显的计算优势，均良好地保持了测试数据在降维表示空间的直线形态.

4.1.3　计算速度测试

在上述的无噪音与带噪音干扰的测试实验中，我们分别记录了 5 种方法对测试数据进行数据映射过程所耗费的计算时间，如表 1 所示. 可显然观察到，所提快速方法与 LPP 方法具有最为快速的计算速度，所提稳健算法的计算过程亦较为快速. 两种所提的新型方法均明显改善了 L-Isomap 方法与 E-LLE 方法的计算效率. 这与 3.4 节中的计算复杂度分析结果完全一致.

表 1　5 种流形重建方法在无噪音与带噪音情形下的测试速度比较

方法	时间/s	
	无噪音数据	带噪音数据
快速算法	0.2360	0.2140
稳健算法	8.0910	8.3150
L-Isomap	67.7670	58.1360
E-LLE	28.6270	28.6630
LPP	0.0102	0.0082

4.1.4　噪音敏感程度考察

本组实验旨在考察在噪音干扰情形下所提稳健算法相对快速算法的鲁棒性改进程度. 实验方法如

下:首先产生规模为1000的无噪声Swiss Roll流形数据集,然后分别对其附加幅度由[−0.1,0.1]至[−1,1]的随机噪声,从而构成11组含不同程度噪声干扰的流形训练数据集(包括无噪声数据集).对应的测试数据集均取为相同的位于流形直线上规模为100的无噪声数据集.然后采用如下的方法对两算法的噪音敏感程度进行量化计算:首先在各个训练数据集下运行算法获得重建流形映射;然后通过阶段2的步1部分程序将所有测试数据分别从高维流形空间映入低维表示空间;再通过阶段2的步2部分程序将其反映回高维流形空间;计算映回向量与对应原测试向量间的偏差,将其均值作为算法在对应噪音干扰下的敏感度度量.实验结果如图3所示.

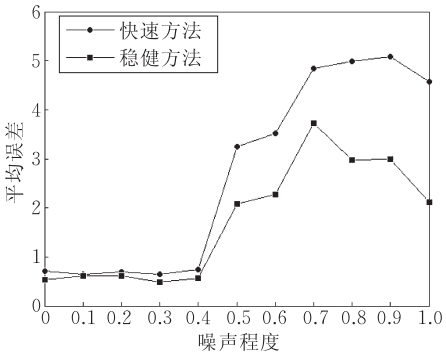


图3 快速算法与稳健算法在不同噪声程度下所得的计算误差变化趋势.

由图3易观察到,稳健算法相对快速算法在不同程度的噪声干扰情形下均表现得更为稳健.特别地,当噪声干扰的程度逐渐增大时,前者相对后者的优势更加明显,这表明了稳健算法对于快速算法在鲁棒性上的显著改进.

综上所述,可以得出如下结论:在无噪音情形下,LPP线性方法计算失效,而各非线性流形重建方法均具有良好的计算性能,其中所提快速算法与稳健算法在计算速度方面具有明显优势;在带噪音情形下,LPP方法亦失效,而对于各非线性方法,所提出的稳健算法与L-Isomap方法的表现更为良好,而稳健算法在计算效率方面具有显著优势.

4.2 应用实例

所提快速与稳健算法分别近似建立了高维流形空间与低维嵌入空间之间的映射关系,所获的双向映射关系均扩展了流形学习方法的应用范畴,分别具有其潜在的应用价值.如低维至高维的映射可应用于信息恢复、图像重建等领域;高维至低维的映射可应用于各种有监督学习问题等领域.我们下面将展示两种映射分别在维数特征动画描述问题与模式分类问题上的应用实验效果.

应用1. 维数特征动画描述.

在以往流形学习方法降维中,维数特征描述的方法一般采用首先计算数据的低维嵌入,然后在沿某一维变化的直线附近依次寻找距离较近的低维嵌入,通过观察其对应于原数据的变化规律来对维数特征进行解释与描述,或采用直接观察的手段也是常用的降维特征描述最常用的方法之一.利用所提算法,可针对图像数据发展出一种更为生动的维数特征描述方法:利用动画来解释维数的特征意义.更具体地,首先通过对原图像数据执行阶段1程序,然后在低维空间沿某一维变化的直线依次产生低维数据序列,将此数据序列映射回原流形空间后可产生图像序列,将此序列按帧顺次播放形成动画,则可生动地反映出此维的特征意义.下面展示所提算法(采用快速算法)在人脸图像数据上的应用效果.

采用的人脸数据为流形学习领域中被多次应用的Isomap标准数据集[3,11].首先对数据运行阶段1程序,获得数据的低维(3维)嵌入表示(如图4).构造三条直线,分别沿第1,2,3维变化(另两维固定为其数据嵌入集合的中点位置).在每条直线上按序列选取14个点,利用所提算法将其映射至原空间获得对应高维数据(图像数据),如图5所示.将其按帧依次播放,可获得维数特征描述的动画演示.

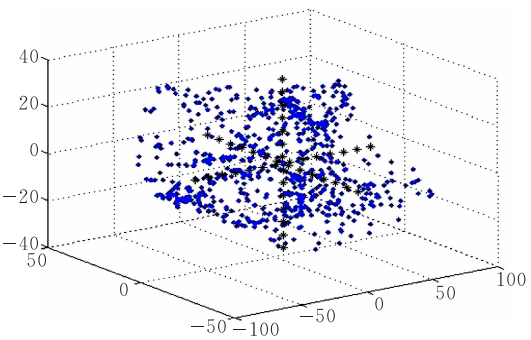


图4 3维数据的嵌入表示(·点为应用Isomap方法获得的人脸数据三维表示集,* 点为用以构造维数特征描述动画的低维数据位置)

从图5可以看出,降维的第1维特征为人脸由右到左,第2维为由下而上,第3维为光照由左至右.3维特征均得以生动而准确地演示.

应用2. 模式分类.

将流形学习方法有效应用于有监督学习问题是模式识别领域的重要研究方向之一.其关键的难点在于一般的流形学习方法缺乏预测功能,即无法判

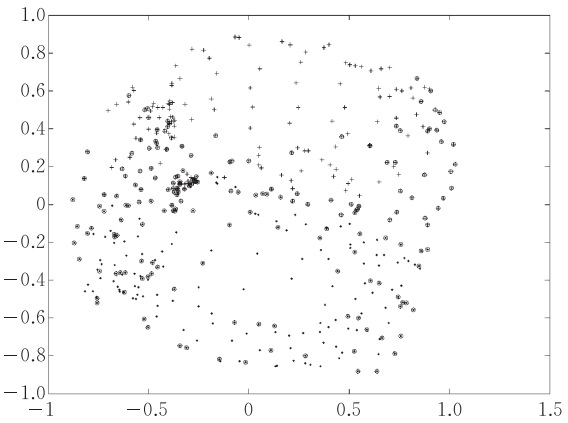
图 5　人脸数据的维数特征动画描述帧演示

断一个新输入样本的降维表示. 应用本文的流形结构重建算法，可将流形学习方法的应用拓展至各种有监督学习问题中. 以下介绍其针对流形数据模式分类问题的实验结果.

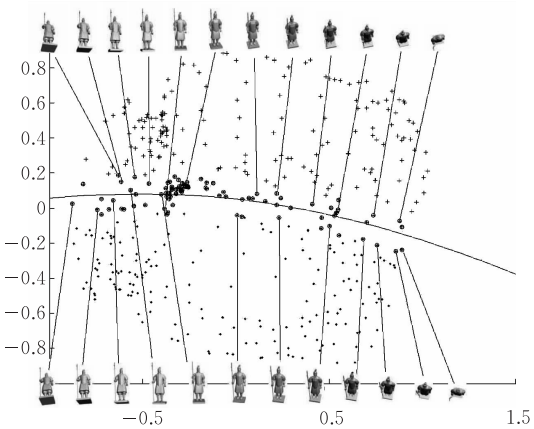
所采用的数据为由 434 个兵马俑图像数据(40×100＝4000 维图像)构成的分类训练数据集与 50 个兵马俑图像构成的测试数据集，其中兵马俑左倾图像定义为正类数据，右倾图像定义为负类数据. 分别利用 SVM 在原高维训练数据与流形学习方法(Isomap方法)获得的对应低维表示数据上进行训练，可获得相应的位于高维与低维空间上的两个分类决策函数. 方便起见，记后者方法为 RSVM 方法. 分别将测试数据在高维与低维空间上进行分类预测，特别地，对于后者需首先利用本文所提方法将测试数据映射入低维空间再进行预测. 测试结果如表 2 所示，其相应获得的支撑向量位置信息如图 6 所示.

表 2　兵马俑数据分类实验效果比较

算法名称	错分个数	支撑向量个数
SVM	1	187
RSVM	2	95



(a) 利用SVM所获的支撑向量



(b) 利用RSVM所获的支撑向量

图 6　Isomap 方法获得的兵马俑数据的 2 维表示集((a)中带圈的 * 数据为 SVM 所获的支撑向量；(b)中带圈的 * 数据为 RSVM 所获的支撑向量，实线为 RSVM 获得的分类决策面)

由表 2 可看出，RSVM 方法与 SVM 方法均具有良好的分类预测能力. 而易注意到，RSVM 方法的支撑向量个数要远小于对应 SVM 方法所获得的支撑向量个数，这说明前者所得的分类决策具有更简单的形式. 这一点可从图 6 中清晰地观察到：RSVM获得的支撑向量均为位于流形分类面附近的数据，即左右倾相对较不明显的兵马俑数据，真正反映了分类决策的边缘信息，因此所获的分类决策更加本质；而 SVM 方法在原数据集上获得的支撑向量分布较为分散，并未很好地反映分类决策的实质边缘，因此所获的分类决策函数形式复杂且本质性较差.

在模式分类问题上的以上成功应用进一步验证了所提算法由高维向低维映射程序的有效性.

5　结　论

本文基于 Isomap 方法本质应用的连续性、局部等距性与稠密性等内在原理，建立了合理的高维流形空间与对应低维表示空间之间的显式映射关系函数. 基于此理论结果，实现了两种新型有效的流形结构重建方法：快速算法与稳健算法. 理论分析与仿真实验结果验证了所提算法相比已有方法具有明显的应用优势：首先，所提算法显著地提高了已有非线性流形结构重建方法的计算效率；其次，所提算法实现了高维流形空间与低维表示空间的双向映射关

系,拓宽了已有方法(其仅实现了高维至低维的映射关系)的应用范畴;另外,在数据集存在噪音干扰时,所提稳健算法仍具有良好的鲁棒表现,在一定程度上改善了其它方法的计算性能.新方法在实际应用领域(如图像重建、数据本质特征描述、模式分类、聚类与回归等领域)中具有极大的潜在应用价值,已在文中的实际实验结果中得以初步体现.

仍然需要继续探索的问题包括:针对小样本数据增强方法使用有效性;与核技巧结合发展新型流形结构重建方法等.

参 考 文 献

[1] Jolliffe I T. Principal Component Analysis. New York: Springer-Verlag, 1989

[2] Cox T, Cox M. Multidimensional Scaling. London: Chapman & Hall, 1994

[3] Tenenbaum J B, Silva V D, Langford J C. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290: 2319-2322

[4] Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000, 290: 2323-2326

[5] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation, 2003, 15: 1373-1396

[6] Hu Zhao-Hua, Fan Xin, Liang De-Qun, Song Yao-Liang. Trajectory tracking and recognition using bi-directional nonlinear learning. Chinese Journal of Computers, 2007, 30(8): 1389-1397(in Chinese)
(胡昭华, 樊鑫, 梁德群, 宋耀良. 基于双向非线性学习的轨迹跟踪和识别. 计算机学报, 2007, 30(8): 1389-1397)

[7] Cottrell G W. New life for neural networks. Science, 2006, 313: 454-455

[8] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504-507

[9] Zhan De-Chuan, Zhou Zhi-Hua. A manifold learning-based multi-instance regression algorithm. Chinese Journal of Computers, 2006, 29(11): 1948-1955(in Chinese)
(詹德川, 周志华. 基于流形学习的多示例回归算法. 计算机学报, 2006, 29(11): 1948-1955)

[10] Bengio Y, Paiement J F, Vincent P, Delalleau O. Out-of-sample extensions for LLE, Isomap, MDS, Eigenmaps, and spectral clustering//Proceedings of the Advances in Neural Information Processing Systems. Whistler, Canada, 2004: 16

[11] Silva V D, Tenenbaum J B. Global versus local methods in nonlinear dimensionality reduction. Neural Information Processing Systems, 2003, 15: 705-712

[12] Silva V D, Tenenbaum J B. Sparse multidimensional scaling using landmark points. Stanford University, Stanford, CA, USA: Technical Report, 2004

[13] Saul L K, Roweis S T. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research, 2003, 4: 119-155

[14] He X, Niyogi P. Locality preserving projections//Proceedings of the Advances in Neural Information Processing Systems. Cambridge, UK, 2003: 16

[15] Zhang J P, Stan Z L. Adaptive nonlinear auto-associative modeling through manifold learning//Ho T B, Cheung D, Liu H eds. PAKDD. LNAI 3518. Berlin Heidelberg: Springer-Verlag, 2005: 599-604

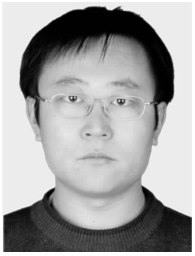
[16] Donoho D L, Grimes C E. When does Isomap recover the natural parameterization of families of articulated images? Department of Statistics, Stanford University, Stanford, CA, USA: Technical Report 2002-27, 2002

[17] Donoho D L, Grimes C E. Hessian eigenmaps: Locally linear embedding techniques for highdimensional data. National Academy of Arts and Sciences, 2003, 100: 5591-5596

[18] Zha H, Zhang Z. Isometric embedding and continuum Isomap//Proceedings of the 20th International Conference on Machine Learning. Washington, USA, 2003: 864-871

[19] Bernstein M, Silva V D, Langford J C, Tenenbaum J B. Graph approximations to geodesics on embedded manifolds. Stanford University, Stanford, CA, USA: Technical Report, 2000

[20] Vapnik V N. The Nature of Statistical Learning Theory. New York: Spinger-Verlag, 1995



MENG De-Yu, born in 1978, Ph.D., lecturer. His current research interests include manifold learning, feature extraction, compressed sensing, and sparse dimensionality reduction techniques.

XU Chen, born in 1985, M. S.. His research interests include statistical data mining and machine learning.

XU Zong-Ben, born in 1955, Ph. D., professor, Ph. D. supervisor. His research interests include neural networks, computational intelligence, compressed sensing, data mining, and machine learning.

Background

The paper mainly focuses on one of the significant research directions in data mining and machine learning, nonlinear dimensionality reduction, also called manifold learning. In the recent decade, this is one of the hottest research issues in the data mining area and has been paid a lot of attentions by the related researchers. However, limitations still remained in the current researches. One of the most typical is the manifold reconstruction issue, also named as the out-of-sample issues. Currently, some techniques have been proposed to solve the issue. Nevertheless, the efficiency and the robustness of these proposed techniques are still far from satisfactory. This paper proposes two methods for this issue, respectively called the fast method and the robust method for manifold reconstruction. The two methods significantly improve the capability of the existing manifold reconstruction methods, especially on efficiency and robustness.

This research was supported by the National Natural Science Foundation of China under contracts 60905003, 70531030 and the National Basic Research Program (973 Program) of China (2007CB311002). The projects mainly focus on the fundamental and technical issues of the information discovery from the huge data with complex intrinsic structures. The investigation of the research group is mainly on the dimensionality reduction techniques of data with very high dimensionality.