文章编号:1001 - 9081(2007)08 - 1959 - 02

一种对奇异值不敏感的 ISOMAP

魏 莱,王守觉,徐菲菲
(同济大学 计算机科学与技术系,上海 201804)
(weily105@hotmail.com)

摘 要:ISOMAP 是一种经典的非线性降维方法,能够有效地发现高维非线性数据集的低维几何结构,但该算法对奇异值和噪声非常敏感。利用具有鲁棒性的主成分分析(Robust PCA)来探测奇异点,并对奇异点进行适当处理以降低 ISOMAP 对其的敏感程度。所提出的算法直观且易于理解,实验结果也证明它具有较好的鲁棒性,而且在奇异点较多的情况下仍能保持数据的整体结构。

关键词:流形学习;主成分分析;等度规映射
中图分类号: TP181 **文献标志码**:A

Robust ISOMAP insensitive to singular value

WEI Lai, WANG Shou-jue, XU Fei-fei
(*Department of Computer Science and Technology, Tongji University, Shanghai,* 201804)

Abstract: ISOAMP is a classical nonlinear dimensionality reduction algorithm. It is effective to discover the low-dimensional manifold in a high-dimensional data space. But the algorithm is very sensitive to the noises and singular value. Principal Component Analysis with robustness (Robust PCA) was used to detect singular points, and the singularity was also appropriately treated to reduce the ISOMAP's sensitivity to it. The proposed algorithm is intuitive and easy to understand, the results of the experiment prove that it is robust, and can maintain the overall structure of data with more singular points.

Key words: manifold learning; Principal Component Analysis (PCA); ISOMAP

0 引言

从数据中学习和发现其内在规律性是机器学习和多元数据分析的主要目标。面对高数据量、高维数的数据集,我们希望能在保持数据信息足够完整的意义下合理地约简数据,以满足存储和人的感知需要。将高维的数据映射到低维空间中,传统的解决方法包括主成分分析(Principal Component Analysis, PCA)和多维尺度变换(Multidimensional Scaling, MDS)等。但这些方法都是基于数据集具有全局的线性结构这样的假设。在很多情况下,这样的假设都是不合理的。

流形学习即为了解决数据集成高度非线性时,数据的约简问题。流形学习假设数据位于或靠近一个在高维空间中的低维流形[1]。在保持数据整体几何结构的同时将高维数据映射到一个低维嵌入空间。流形学习的方法大体可以分成五种类型[1]:神经网络、主流形、谱分析、变分法和互信息。近年来提出的等度规映射(ISOMAP)[2,3]、局部线性嵌套(LLE)[4]和 Laplacian 特征映射[5],都可以看作谱分析的应用[6]。

虽然 ISOMAP 和 LLE 算法能够很好的恢复数据集内在的几何结构,但是这两种算法对噪声和奇异值都非常敏感,鲁棒性很差。于是学者们提出了一些方法来解决这些问题,如 RLLE[7],Robust Kernel Isomap[8]等。但 RKISOMAP 涉及复杂的数学,且不直观,而使 LLE 算法具有鲁棒性的直观方法也未能在 ISOMAP 上应用。

为此,本文提出了一种处理奇异值的 ISOMAP 算法(R-ISOAMP),利用具有鲁棒性的主成分分析(Robust PCA)来探测奇异值。通过对奇异值和非奇异值不同的处理最大限度地保持数据集内在的结构。算法直观且易于理解,实验结果证明本算法能够有效地恢复具有奇异值的流形结构。

1 ISOAMP 算法及奇异值和噪声对其的影响

1.1 ISOAMP 算法[2]

ISOMAP 从全局入手,通过计算样本空间中任意两点测地线距离来度量两点之间的距离。最后使用经典的多维尺度变换(MDS)处理矩阵得到高维数据集在低维的相应坐标。令 $X = \{x_1, x_2, \cdots, x_N\}$ 是高维样本空间 R^D 的 N 个数据点,假设这些数据点是位于或靠近一个非线性的低维流形 M,设其内在维数 $d \ll D$,$Y = \{y_1, y_2, \cdots, y_N\} \subset R^d$ 是相应的低维空间中的数据点。

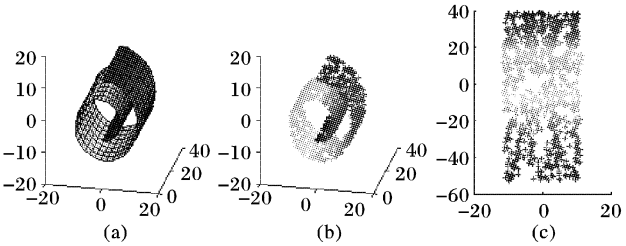


图 1 ISOMAP 应用于 SwissRoll

算法如下:

步骤 1:构造邻接图。任意两点 x_i, x_j 距离为欧式距离 $d_x(i,j)$,邻接关系为 K- 近邻或 ε- 球;

步骤 2:通过计算邻接图上两点间的最短路径来重构流形 M 上两点之间的测地线距离,得到距离矩阵 $\boldsymbol{D}$;

收稿日期:2007 - 02 - 27;**修回日期**:2007 - 04 - 10。
基金项目:国家自然科学基金资助项目(60495019);教育部博士点专项基金资助项目(20060247039)。
作者简介:魏莱(1980 -),男,江苏苏州人,博士研究生,主要研究方向:仿生模式识别、流形学习; 王守觉(1925 -),男,江苏苏州人,院士,主要研究方向:仿生模式识别; 徐菲菲(1983 -),女,江西南昌人,博士研究生;主要研究方向:粗糙集理论、人工智能。

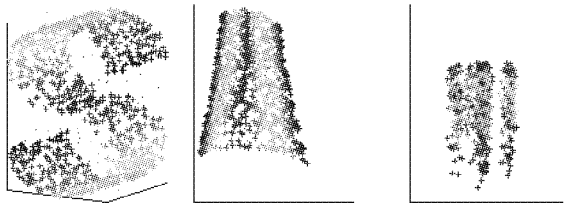
步骤 3:应用 MDS,构建 d 维欧式空间上的嵌入 Y。

通过 ISOMAP 算法高维空间中低维流形的结构能够被发现,图 1 显示 ISOMAP 恢复 3 维空间中 2 维嵌入 SwissRoll 的内在结构。

1.2 奇异值对 ISOMAP 算法的影响

ISOMAP 是一种有效的流形学习算法,能够在数据集高度非线性时发现其内在结构,但它对于奇异值和噪声点非常敏感。与 LLE 相比,奇异值和噪声对 ISOMAP 的影响要大得多。这是因为,ISOMAP 是从全局入手,通过两点间的测地线距离来度量两点之间的距离,而奇异点和噪声的存在直接破坏了邻接图的构造,使通过最短距离算法得到的两点之间距离不能真实反映位于低维流形上数据点的测地线距离,从而使用 MDS 时,不能得到相应的低维嵌入坐标。而 LLE 是从局部考虑,用样本点近邻的线性组合来逼近样本点,噪声和奇异点影响的只是其周围的近邻,所以对整体结构的恢复要好于 ISOMAP。图 2 显示了在具有随机生成的 100 个噪声点的情况下,LLE 和 ISOMAP 分别对 3 维空间中 S-curve 结构的恢复。很明显,ISOMAP 对带有噪音点的流形结构的发现效果要远差于 LLE 算法。

为提高 ISOMAP 的鲁棒性,学者们提出了一些算法如 Robust Kernel Isomap[8] 等,但该算法采用复杂的数学技巧,同时缺乏直观性。也有的学者采用噪声过滤的方法增强算法的鲁棒性。本文借助于具有鲁棒性的 PCA(Robust PCA)提出了一种新的 R-ISOMAP 算法。



(a) S-curve (b) 使用LLE的结果 (c) 使用ISOMAP的结果

图 2 LLE 与 ISOMAP 对 S-curves 结构的恢复,噪声点个数 100

2 具有鲁棒性的主成分分析[7]

具有鲁棒性的主成分分析(Robust PCA)算法主要依靠在局部领域采用带权重的 PCA 技巧来自动的探测出奇异值。具体描述可以参考文献[7],本文在这里简述一下其方法。

同样令 $X = \{x_1, x_2, \cdots, x_N\}$ 是高维样本空间 R^D 的 N 个数据点,对每个样本点 x_i 用 K-近邻法确定其近邻,设为 $\{x_{i1}, x_{i2}, \cdots, x_{iK}\}$。如果 x_i 位于一个低位流形上,那么它的每个近邻可以近似地看作位于流形的一个小切片上。设低维流形的内在维数为 d,那么通过标准的 PCA,x_i 近邻点 x_{ij} 的 d 维空间嵌入坐标可以写成 $z_j = B^t(x_{ij} - m)$,m 为样本空间 R^D 中的一个点,$B = [b_1, b_2, \cdots, b_d]_{D\times d}$,其中 b_i 互为正交向量。这样 x_{ij} 可以这样估计,$\hat{x}_{ij} = m + Bz_j = m + BB^t(x_{ij} - m)$。令 $\varepsilon_j = x_{ij} - \hat{x}_{ij}$,根据 ε_j 的大小来确定每个近邻的权值和重新计算 m 和 B,直到 m 和 B 变化不在过大。m 通过下式调整:$m = \sum_{j=1}^{K} a_j x_{ij} \Big/ \sum_{j=1}^{K} a_j$,$B$ 通过计算 $S = \frac{1}{K}\sum_{j=1}^{K} a_j (x_{ij} - m)(x_{ij} - m)^t$ 的 d 个最大特征向量来调整。其中 a_j 由 $\|\varepsilon_j\|$ 确定,本文中这样计算 a_j:

$$a_j = \begin{cases} 1, & \|\varepsilon_j\| \leqslant c \\ \dfrac{c}{\|\varepsilon_j\|}, & \|\varepsilon_j\| > c \end{cases}, \text{而 } c = \frac{1}{2K}\sum_{j=1}^{K} \|\varepsilon_j\| 。$$

算法如下:

步骤 1:通过标准 PCA,求出 m 和 B 的初值,记为 $m^{(0)}$,$B^{(0)}$,令 $t = 0$;

步骤 2:DO

1) $t = t + 1$;

2) 计算:

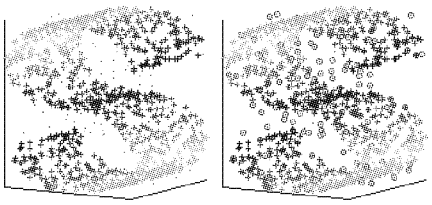
$\varepsilon_j^{(t-1)} = x_{ij} - \hat{x}_{ij} = x_{ij} - m^{(t-1)} - B^{(t-1)}(B^{(t-1)})^t(x_{ij} - m^{(t-1)})$;$1 \leqslant j \leqslant K$

3) 计算 $a_j^{(t-1)} = a(\|\varepsilon_j^{(i-1)}\|)$;

4) 由 $a_j^{(t-1)}$,重新计算 $m^{(t)}$,$B^{(t)}$;

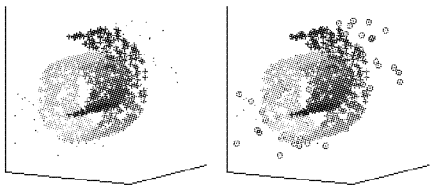
Until m 和 B 变化不再过大。

算法对每个局部领域计算后,每一个点都会得到相应的权值,我们看到 $\|\varepsilon_j\|$ 越大,a_j 会越小,记录权值。算法对数据集中点遍历后,将该点记录的所有权值相加,若结果小于某一阈值 t,即确定其为奇异点。图 3、4 分别显示 RPCA 区分奇异点和位于 S-curve、SwissRoll 的数据点的情况。



(a) (b)

图 3 RPCA 应用于 S-curve,噪音点个数 100,$K = 12$,$t = 0.5$



(a) (b)

图 4 RPCA 应用于 SwissRoll,噪音点个数 100,$K = 12$,$t = 0.5$

3 R-ISOAMP 算法

令 $X = \{x_1, x_2, \cdots, x_N\}$ 是高维样本空间 R^D 的 N 个数据点,近邻选择用 K-近邻法,设低维流形的内在维数为 d,$Y = \{y_1, y_2, \cdots, y_N\} \subset R^d$ 是相应的低维空间中的数据点。

R-ISOMAP 算法如下:

步骤 1 对输入数据点用 RPCA 处理,得到非奇异点集 P 和奇异点集 O;

步骤 2

1) 对 $x_i(\in X)$,IF $x_i \in P$ THEN 选择其在 P 中的 K 个近邻,距离为欧式距离,构造邻接图;

2) 对 $x_i(\in X)$,IF $x_i \in O$ THEN 选择其与 P 中最近点,仅使这两点连通;

3) 通过在上两步所得邻接图上使用最短路径算法,得到距离矩阵 $\boldsymbol{D}$;

步骤 3:应用 MDS,构建 d 维欧式空间的嵌入 Y。

我们知道,ISOMAP 算法通过度量流形上两点之间测地线的距离来计算两点之间距离,对于非奇异点可以设想它位于流形上,所以它们之间的处理方法可以与原来一样,寻找通过其在非奇异点集中的 K 个近邻来构建邻接图。而对于奇异点,我们只连接与它最接近的非奇异点,这样,任意点与一奇异点之间的距离,仍需要通过寻找流形上该点投影(最接近的非奇异点)和奇异点在流形上的投影之间测地线来获取,再通过最小路径算法来计算它们之间的距离,这样即可以保持整体数据集可能的潜在结构,又保留了奇异点的信息。

(下转第 1963 页)

数$\tilde{C}$;然后对 RBF 核的 SVM,固定$\tilde{C}$,对满足 $\log\sigma^2 = \log C - \log\tilde{C}$ 要求的(C,σ)组合,用于训练SVM,估算其推广识别率,与最高推广识别率对应的(C,σ)组合就是最优参数。对应于最优识别率,混沌优化确定的模型参数为 RBF 核函数的参数σ为2,惩罚因子C为4。应用混沌优化、网格法和双线性法所得到的对比结果如表 1 所示。为了充分说明网格法和双线性法的不确定性,表 1 中网格法和双线性法给出了两种不同的参数组合。

表 1　不同方法决定的 SVM 对样本识别性能比较

方法	模型参数	支持向量数	识别率/%	训练的 SVM 个数
混沌优化	(4,2)	32	84.23	50
网格法	(4,0.5)	61	83.20	100
	(2,1)	61	84.50	100
双线性法	(2,1)	61	83.22	60
	(1,1)	62	84.60	60

从以上实验可以看出,基于混沌优化的 SVM 具有最少的支持向量数目,证明了该 SVM 模型参数的最优性。此外,混沌优化在减少训练 SVM 个数的情况下,基本可以达到与网格法和双线性法相当的识别率。

4　结语

为了进一步提高笔迹鉴别问题的识别率,本文提出了基于混沌优化的 SVM 模型参数(C,σ)的自动选择方法。采用基于混沌优化自动选择参数,不但确保了 SVM 具有最优的泛化性能,而且减少了 SVM 的训练数目。与网格法和双线性法的模型参数选择相比,在保持同样识别率的情况下,基于混沌优化的 SVM 能有效减少支持向量数。本文提出的方法实现了 SVM 模型参数的自动选择,为快速设计高性能的实用 SVM 提供了一种简便、高效、通用的方法,基于 SVM 的笔迹识别结果也证明了本文所提出方法的有效性。

参考文献:

[1] CHAPELLE O, VAPNIK V, BOUSQUET O, *et al*. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46(1/3): 131 - 159.

[2] CHERKASSKY V, MA Y Q. A practical selection of SVM parameters and noise estimation for SVM regression[J]. Neural Networks, 2004, 17(1): 113 - 126.

[3] KEETHI S, LIN C-J. Asymptotic behavior of support vector machines with Gaussian kernel[J]. Neural Computation, 2003, 15(7): 1667 - 1689.

[4] FRÖNHL ICH H, CHAPELLE O, SCHÖLKOPF B. Feature selection for support vector machines by means of genetic algorithms [C]// Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03), November 3 - 5, 2003. [S.l.]: IEEE Computer Society, 2003, 142 - 148.

[5] KEERTHIS S. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms[J]. IEEE Transactions on Neural Network, 2002, 13(5): 1225 - 1229.

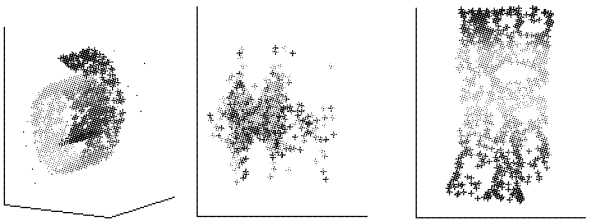
[6] ZHANG H D, HE Y Y. Comparative study of chaotic neural networks with different models of chaotic noise[C]// Proceedings of First International Conference on Natural Computation (ICNC2005), LNCS 3610, Changsha, China. Berlin: Springer, 2005: 273 - 282.

[7] 朱勇,谭铁牛. 基于笔迹的身份鉴别[J]. 自动化学报, 2001, 27(2): 229 - 234.

[8] WESTON J, WATKINS C. Multi - class support vector machines [R]. Royal Holloway: University of London, Department of Computer Science, 1998.

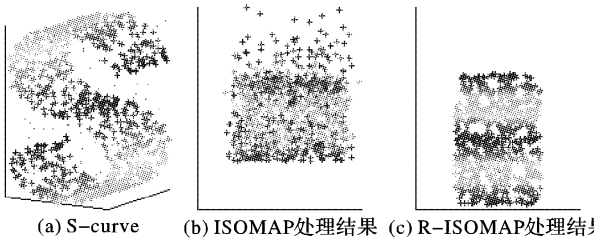
(上接第 1960 页)

图 5、6 分别显示了 ISOMAP 和 R-ISOMAP 对含有噪音点的 SwissRoll 和 S-curve 处理结果。



(a) SwissRoll　(b) ISOMAP处理结果　(c) R-ISOMAP处理结果

图 5　噪音点个数为 50 时三种算法的处理结果



(a) S-curve　(b) ISOMAP处理结果　(c) R-ISOMAP处理结果

图 6　噪音点个数为 100 时三种算法的处理结果

4　结语

ISOMAP 是经典的非线性降维算法,但该算法对奇异值和噪音非常的敏感,本文提出的R-ISOMAP算法,首先利用具有鲁棒性的 PCA 来探测奇异点,对奇异点进行简单处理后,可以大大增强 ISOMAP 算法的鲁棒性,而且算法直观,通过试验显示在奇异点较多的情况下,仍能保持数据整体结构。但是该算法速度较慢,提高算法的速度将是今后的研究方向。

参考文献:

[1] 张军平. 流形学习与应用[D]. 北京: 中国科学院自动化所, 2003.

[2] TENENBAUM J B, De SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319 - 2323.

[3] BALASUBRAMANIAN M, SCHWARTZ E L. The ISOMAP algorithm and topological stability, Science, 2002, 295(5552): 7a.

[4] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323 - 2326.

[5] BELKIN M, NIYOGI P. Laplacian eigenmaps and spectral techniques for embedding and clustering[G]// DIETTERICH T G, BECKER S, GHAHRAMANI Z. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002, 14: 585 - 591.

[6] 罗四维,赵连伟. 基于谱图理论的流形学习[J]. 计算机研究与发展, 2006, 43(7): 1173 - 1179.

[7] CHANG H, YEUNG D-Y. Robust locally linear embedding[J]. Pattern Recognition, 2006, 39(6): 1053 - 1065.

[8] CHOI H, CHOI S. Robust kernel ISOMAP[J]. Pattern Recognition, 2007, 40(3): 853 - 862.