P-ISOMAP: 一种新的对邻域大小不甚敏感的数据可视化算法

邵 超[1,2], 黄厚宽[1], 赵连伟[1]
(1. 北京交通大学计算机与信息技术学院, 北京 100044; 2. 河南财经学院计算机科学系, 河南郑州 450002)

摘 要: ISOMAP算法对邻域大小敏感, 而邻域大小却难以有效选取. 本文根据二阶最小生成树不含有"短路"边的特性提出了能有效删除邻域图中的"短路"边因而对邻域大小不甚敏感的 P-ISOMAP算法. 由于避免了邻域大小难以有效选取的问题, 该算法能更容易地对数据进行可视化, 也获得了一定程度的拓扑稳定性和鲁棒性. 实验结果很好地验证了该算法的有效性.

关键词: ISOMAP; P-ISOMAP; 二阶最小生成树; 成本; 残差
中图分类号: TP181 文献标识码: A 文章编号: 0372-2112 (2006) 08-1497-05

P-ISOMAP: A New ISOMAP-Based Data Visualization Algorithm with Less Sensitivity to the Neighborhood Size

SHAO Chao[1,2], HUANG Hou-kuan[1], ZHAO Lian-wei[1]
(1. School of Computer & IT, Beijing Jiaotong University, Beijing 100044, China;
2. Department of Computer Science, Henan University of Finance and Economics, Zhengzhou, Henan 450002, China)

Abstract The success of ISOMAP depends greatly on choosing a suitable neighborhood size; however, it is still an open problem how to do this effectively. Based on characteristics of the SOMST (Second-Order Minimal Spanning Tree) in which shortcut edges can be avoided, this paper presented a variant of ISOMAP, i.e. P-ISOMAP (Pruned-ISOMAP). P-ISOMAP can prune effectively shortcut edges existed possibly in the neighborhood graph according to their costs over the SOMST, and thus is much less sensitive to the neighborhood size than ISOMAP. Consequently, P-ISOMAP can be applied to data visualization more easily than ISOMAP for the open problem described above can be avoided to a certain extent; in addition, P-ISOMAP can also be more topologically stable and robust than ISOMAP. Finally, the feasibility and effectivity of P-ISOMAP can be verified by experimental results very well.

Key words ISOMAP; P-ISOMAP; SOMST; cost; residual variance

1 引言

作为一种有效的非线性降维技术, ISOMAP算法[1]采用能有效表征数据全局几何结构的测地距离对古典 MDS算法进行了非线性扩展, 从而能很好地对嵌入在高维欧氏空间中的低维非线性流形如 swiss roll数据集等进行可视化.

ISOMAP算法只有一个参数——邻域大小, 并且对它比较敏感; 然而, 如何对它进行有效的选取目前还是一个尚未解决的难题. 众所周知, 一个合适的邻域大小不应在邻域图中引入"短路"(shortcut)边[2]. 为此, 本文提出了 ISOMAP算法的一个变种——P-ISOMAP (Pruned-ISOMAP)算法, 通过有效删除邻域图中的"短路"边能极大地削弱 ISOMAP算法对邻域大小的依赖程度, 从而能更容易地对数据进行可视化, 也获得了一定程度的拓扑稳定性和鲁棒性.

2 ISOMAP算法

在数据的全局几何结构未知 (通常呈非线性)的情况下, 欧氏距离只在很小的邻域内才有意义[3], 因此, 我们需要用这些已知的局部欧氏距离来逼近能有效表征数据全局几何结构的测地距离. ISOMAP算法就是这么做的: 用邻域图来表达数据的邻域结构, 并用邻域图中的最短路径长度来对测地距离进行逼近, 然后运行古典 MDS算法, 在低维可视空间中对数据的全局几何结构进行直观展现, 从而实现了对数据的可视化. ISOMAP算法可简要描述如下[1]:

收稿日期: 2005-10-04; 修回日期: 2006-05-09
基金项目: 国家自然科学基金 (No. 60443003); 北京交通大学基金 (No. 2003SZ003)

(1)对于大数据集,为降低计算量,从中选取 n个代表点以执行下面的操作.选取方法很多,本文采用的是矢量量化方法,因为它能得到更具代表性的数据点,可视化效果会更好[4];

(2)用 K-近邻法(K-nearest neighbors)创建能正确表达数据邻域结构的邻域图,这需要一个合适的邻域大小 K;

(3)运行最短路径算法得到所有数据间的最短路径长度;

(4)将这些最短路径长度作为输入运行古典 MDS算法,将数据重建在一个低维可视空间中.

如果数据具有良好抽样且位于内在扁平的单一流形之上,那么 ISOMAP算法能否被成功运用就完全依赖于邻域大小的合适与否了.人们通常根据最终映射"质量"的高低来对邻域大小进行合适的选取[5],用残差[1](residual variance)来进行衡量: $1-\hat{\rho}^2_{D_X(K)D_Y}$,其中,$D_X$ 和 D_Y 分别为数据在原数据空间中的测地距离矩阵(由最短路径长度来进行逼近,在给定数据集的情况下为邻域大小 K 的函数)和在低维可视空间中的欧氏距离矩阵,而 $\rho_{D_X(K)D_Y}$ 则为它们之间的线性相关系数.残差越小,映射的"质量"就越高,邻域大小也就越合适.因此,最优邻域大小可定义如下[5]:

$$K_{opt}=\arg\min_K(1-\hat{\rho}^2_{D_X(K)D_Y}) \quad (1)$$

由于残差用到了 ISOMAP算法的运行结果 Y,因此,计算残差需要运行整个 ISOMAP算法.另外,残差只能衡量两个邻域大小的相对合适程度,而不能用来判断某一个邻域大小的合适与否;同时,残差还具有多峰性(multimodality),因此,该方法需要就每一个可能的邻域大小分别计算相应的残差,这是极其耗时的,从而也使邻域大小难以有效选取.

3 P-ISOMAP算法

一个合适的邻域大小应能使相应的最短路径长度对测地距离进行精确逼近,为此,邻域大小应以不引入"短路"边为限.如果能对邻域图中可能存在的"短路"边进行鉴别和删除的话,我们就可以极大地削弱邻域大小的影响,从而可以避免邻域大小难以有效选取的问题.

定义 1 某个数据集的最小生成树(Minimal Spanning Tree,MST)指的是由该数据集生成的完全图(completed graph)上的最小生成树.

定义 2 某条边在最小生成树上的成本(cost)指的是在这个最小生成树上用以连接该边两个端点所需的最少边数.如果这个最小生成树上的所有权重都设定为 1的话,那么这个最少边数也就等于最短路径长度.

众所周知,最小生成树能有效避免"短路"边[6],从而使"短路"边的两个端点在这个最小生成树上相距要比非"短路"边远得多,可以用成本来进行衡量,如图 1(a-b)所示.然而,最小生成树过于稀疏,从而会使某些非"短路"边的成本也很大,难以将其和"短路"边进行区分,如图 1(c-d)所示.

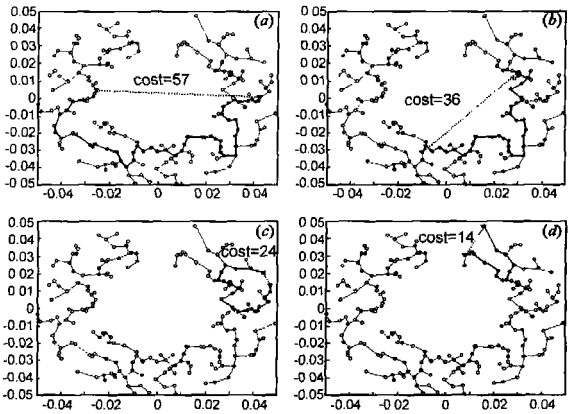


图 1 不同边(用虚线表示)在最小生成树(用细实线表示)上的成本,所用数据集为一带有缺口的环状数据集(共包含 200个随机数据点)

针对最小生成树过于稀疏的问题,我们采用类似于文献[6]中组合最小生成树的方法;为了能同样有效地避免"短路"边,我们仅采用两个最小生成树组合而成的二阶最小生成树(Second-Order Minimal Spanning Tree,SOMST),实现步骤如下所示:(1)在给定数据集的完全图上计算最小生成树;(2)将其从完全图中删除,再次计算最小生成树,合并这两个最小生成树就得到了二阶最小生成树.

我们用 Prim 算法来计算最小生成树,时间复杂度为 $O(n^2)$[7](n 为数据点的个数),该时间复杂度要明显小于最短路径算法和古典 MDS算法的时间复杂度.在二阶最小生成树上对成本的定义和定义 2类似,只是把其中的最小生成树换成了二阶最小生成树.

和最小生成树相比,在大多数情况下,二阶最小生成树同样能很好地避免"短路"边;同时,二阶最小生成树还能在很大程度地避免某些非"短路"边成本过大的问题,如图 2所示.

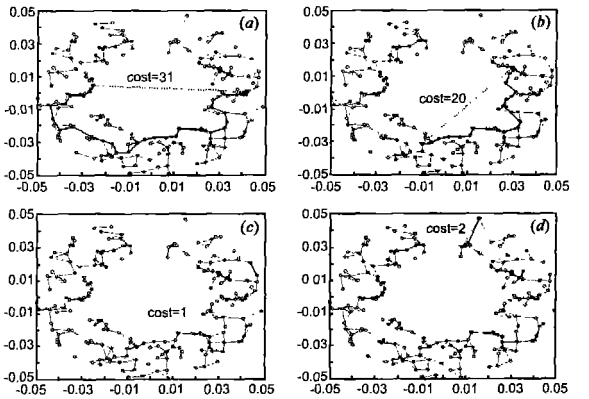


图 2 不同边(同图 1,用虚线表示)在二阶最小生成树(用细实线表示)上的成本,所用数据集同图 1

如上所述,"短路"边的成本明显要比非"短路"边大.如果能找到这样的阈值,使"短路"边的成本都比它大,而非"短路"边的成本都比它小,那么我们就可以删除邻域图中的这些"短路"边了,从而可以更容易地对邻域大小参数

进行指定, 这就是 P-ISOMAP算法:

(1)对于大数据集, 为降低计算量, 用矢量量化方法从中选取 n 个代表点以执行下面的操作;

(2)给定一个邻域大小 K, 用 K-近邻法创建相应的邻域图;

(3)删除邻域图中可能存在的"短路"边.

(a)创建该数据集的二阶最小生成树, 将其中的权重都设定为 1;

(b)在这个二阶最小生成树上运行最短路径算法得到邻域图中所有边的成本;

(c)选取一个合适的阈值 ε(见下一节);

(d)对邻域图中的每一条边, 如果它的成本大于 ε, 就视其为"短路"边, 从邻域图中删除.

(4)在删除了"短路"边的邻域图上运行最短路径算法得到所有数据间的最短路径长度.

(5)将这些最短路径长度作为输入运行古典MDS算法, 将数据重建在一个低维可视空间中.

4　阈值 ε的选取

尽管 P-ISOMAP算法仍然需要指定一个邻域大小K, 但对它的依赖程度已经被极大地削弱了, 从而可以比较容易地对其进行指定. 然而, 为能有效删除邻域图中可能存在的"短路"边, 对阈值 ε的选取至关重要.

由于"短路"边的成本明显要比非"短路"边大得多, 如果邻域图中出现了非"短路"边, 邻域图中的所有成本就可以很明显地分为两组甚至多组, 如图 3所示. 因此, 我们可以按照以下步骤对阈值 ε进行选取:

(1)计算每一个成本对应的边数:

$$s(i)=\sum_{j=1}^{p} I(\mathrm{cost}(j),\ i),\quad i=\mathrm{cost}_{\min},\ \cdots,\ \mathrm{cost}_{\max} \qquad (2)$$

其中, p 为邻域图中边的条数, cost(j)为邻域图中第 j条边的成本, $\mathrm{cost}_{\min}$和 $\mathrm{cost}_{\max}$分别为邻域图中的最小成本和最大成本, 函数 I(x, y)在 x = y 时返回 1, 否则返回 0.

(2)从 $\mathrm{cost}_{\min}$开始沿着成本从小到大的方向进行搜索.

(3)如果 s(i) = 0, 则此时的成本就作为阈值 ε(如图 3(b-c)中的五星所示), 搜索结束.

(4)如果搜索到 $\mathrm{cost}_{\max}$依然无法满足步骤 3中的条件(即 s(i) = 0)时, 就说明该邻域图不存在"短路"边, 可以将 $\mathrm{cost}_{\max}$作为阈值 ε(如图 3(a)中的五星所示).

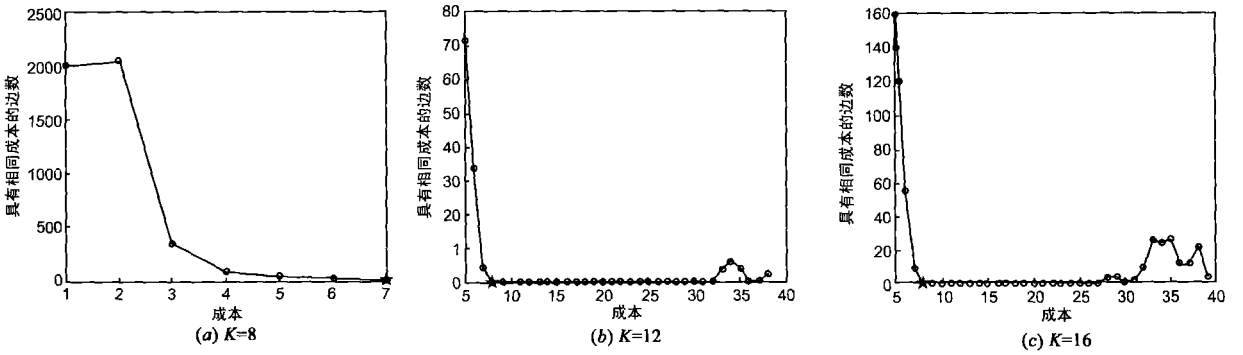


图 3　在 K 取值不同的情况下, 相应邻域图的成本分布, 其中, 五星表示选取出来的阈值 ε. 所用数据集为包含 2000个随机数据点的swiss roll数据集[1]

5　实验结果

这一节, 我们在包含 2000个随机数据点的 swiss roll数据集[1] (如图 4(a)所示, 为降低计算量, 我们采用 MATLAB v6.5工具箱中的 k-均值算法从中选取了 500个代表点)上以不同的邻域大小 K 分别运行 ISOMAP算法和 P-ISOMAP算法, 以验证 P-ISOMAP算法是否能有效删除邻域图中的"短路"边, 进而是否能削弱 ISOMAP算法对邻域大小的依赖程度. 实验结果分别如图 5~7所示.

从图 5~7可以看出, P-ISOMAP算法能有效删除邻域图中的"短路"边(分别如图 6(a)和图 7(a)中的虚线所

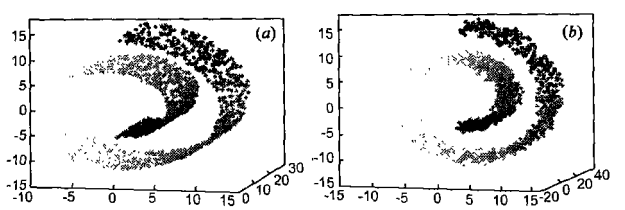


图 4　(a)包含 2000个随机数据点的 swiss roll数据集[1], (b)在如(a)所示的 swiss roll数据集的每一个数据点上都加入一服从正态分布的噪音, 该正态分布的期望为 0, 标准差为数据集在各维上跨度最小值的 2%[2]

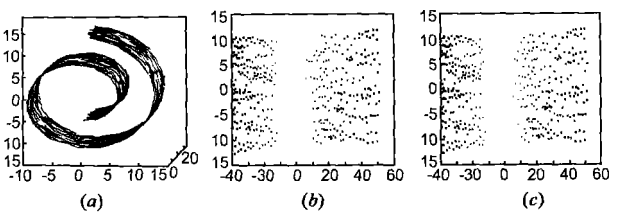


图 5　在 swiss roll数据集上以 K=8分别运行这两个算法, 得到的邻域图如(a)所示, ISOMAP算法和 P-ISOMAP算法的运行结果分别如(b)和(c)所示

示), 从而削弱了 ISOMAP算法对邻域大小的依赖程度: 在 K = 12和 K = 16时, P-ISOMAP算法依然能很好地对数据的全局几何结构进行直观展现, 而 ISOMAP算法则会造成一定程度的扭曲, 分别如图 6(b – c)和 7(b – c)所示.

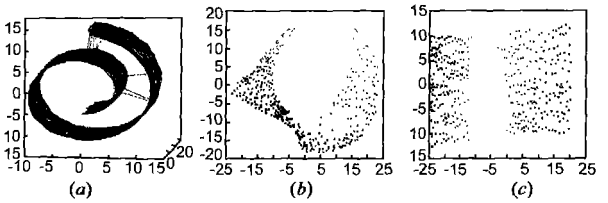


图 6 在 swiss roll数据集上以 K=12分别运行这两个算法, 得到的邻域图如 (a)所示 (虚线表示被 P-ISOMAP算法删除的"短路"边), ISOMAP算法和 P-ISOMAP算法的运行结果分别如 (b)和 (c)所示

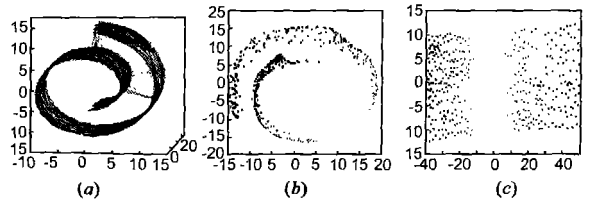


图 7 在 swiss roll数据集上以 K=16分别运行这两个算法, 得到的邻域图如 (a)所示 (虚线表示被 P-ISOMAP算法删除的"短路"边), ISOMAP算法和 P-ISOMAP算法的运行结果分别如 (b)和 (c)所示

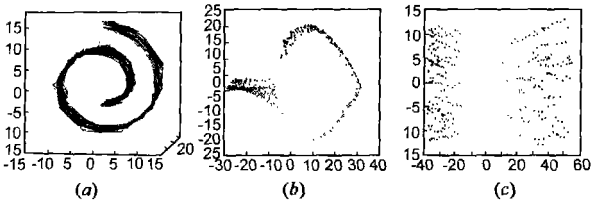


图 8 在带有噪音的 swiss roll数据集上以 K=8分别运行这两个算法, 得到的邻域图如 (a)所示 (虚线表示被P-ISOMAP算法删除的"短路"边), ISOMAP算法和 P-ISOMAP算法的运行结果分别如 (b)和 (c)所示

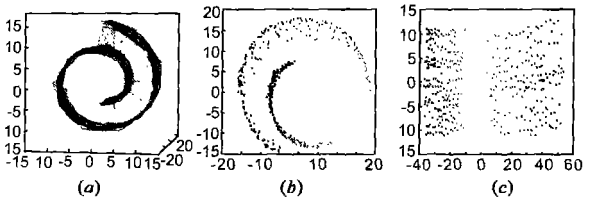


图 9 在带有噪音的 swiss roll数据集上以 K=12分别运行这两个算法, 得到的邻域图如 (a)所示 (虚线表示被 P-ISOMAP算法删除的"短路"边), ISOMAP算法和 P-ISOMAP算法的运行结果分别如 (b)和 (c)所示

众所周知, 二阶最小生成树和 ISOMAP算法对噪音都比较敏感, 为验证 P-ISOMAP算法是否能缓解这一问题, 我们在带有噪音的 swiss roll数据集[2] (如图 4(b)所示, 我们同样从中选取了 500个代表点)上以同样的邻域大小再次运行这两个算法, 实验结果分别如图 8~ 10所示. 从中我们可以看出, P-ISOMAP算法依然能有效删除邻域图中的"短路"边 (分别如图 8(a)、图 9(a)和图 10(a)中的虚线所示), 进而依然能很好地对数据的全局几何结构进行直观展现, 从而获得了一定程度的拓扑稳定性和鲁棒性.

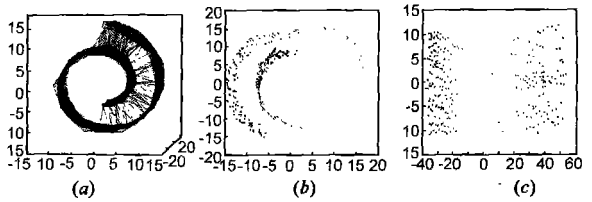


图 10 在带有噪音的 swiss roll数据集上以 K=16分别运行这两个算法, 得到的邻域图如 (a)所示 (虚线表示被 P-ISOMAP算法删除的"短路"边), ISOMAP算法和 P-ISOMAP算法的运行结果分别如(b)和(c)所示

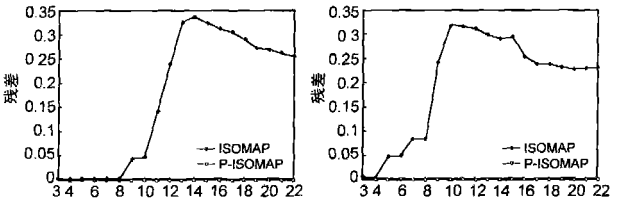


图 11 这两个算法在不同数据集上的残差比较

由于残差可用来衡量邻域大小的相对合适程度. 因此, 我们又计算了这两个算法在以上两个数据集上的残差, 如图 11所示, 从中可以进一步验证以上结论.

6 小结

针对 ISOMAP算法对邻域大小比较敏感而邻域大小在实际中却难以有效选取的问题, 本文根据二阶最小生成树在大多数情况下都不存在"短路"边的这一特性提出了 P-ISOMAP算法. 该算法能有效删除邻域图中的"短路"边, 从而对邻域大小不再像 ISOMAP算法那样敏感, 能更容易地对数据进行可视化, 也获得了一定程度的拓扑稳定性和鲁棒性.

P-ISOMAP算法能成功运行的依据是二阶最小生成树具有不存在"短路"边的特性, 由于这一特性在大多数情况下都能成立, 从而使该算法具有良好的推广性, 在其它低维非线性流形如 S形数据集上也有很好的表现.

参考文献:

[1] Tenenbaum J B, et al. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(22): 2319– 2323.

[2] Balasubramanian M, et al. The isomap algorithm and topological stability[J]. Science, 2002, 295(5552): 7a– 7.

[3] Saxena A, et al. Non-linear dimensionality reduction by locally linear isomaps[A]. Pal N R, Kasabov N, Mudi R K, Pal S, Parui S K, Proc of the 11th International Conference on Neural Information Processing[C]. Calcutta: Springer-

Verlag, 2004. 1038- 1043.

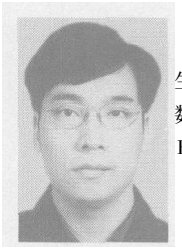
[4] Lee J A, et al. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis[J]. Neurocomputing, 2004, 57: 49- 76.

[5] Kouropteva O, et al. Selection of the optimal parameter value for the locally linear embedding algorithm[A]. Wang L, Halgamuge S K, Yao X. Proc of the 1st International Conference on Fuzzy Systems and Knowledge Discovery: Computational Intelligence for the E-Age[C]. Singapore, 2002. 359- 363.

[6] Carreira-Perpiñán M Á, et al. Proximity graphs for clustering and manifold learning[A]. Saul L K, Weiss Y, Bottou L. Advances in Neural Information Processing Systems 17[C]. Vancouver: MIT Press, 2004. 225- 232.

[7] 严蔚敏, 等. 数据结构(第二版)[M]. 北京: 清华大学出版社, 1992. 171- 193.

作者简介:



邵　超　男, 1977年生于河南渑池, 博士研究生, 主要研究领域包括人工神经网络、机器学习、数据可视化和数据挖掘等.

E-mail: shaochao051227@gmail.com

黄厚宽　男, 1940年生于四川遂宁, 教授, 博士生导师, 主要研究领域包括人工智能、机器学习、数据仓库、数据挖掘、决策支持系统和多智能体系统等.