基于随机游走扩散映射的降维算法

薛艳锋[1,2,3] 王三虎[3] 高志娥[3] 高永强[3]

[1](山西大学计算机与信息技术学院 山西 太原 030006)

[2](山西大学复杂系统研究所 山西 太原 030006)

[3](吕梁学院计算机科学与技术系 山西 吕梁 033000)

摘 要 传统数据降维算法分为线性或流形学习降维算法,但在实际应用中很难确定需要哪一类算法。设计一种综合的数据降维算法,以保证它的线性降维效果下限为主成分分析方法且在流形学习降维方面能揭示流形的数据结构。通过对高维数据构造马尔可夫转移矩阵,使越相似的节点转移概率越大,从而发现高维数据降维到低维流形的映射关系。实验结果表明,在人造数据以及真实数据的线性降维中,该算法降维效果与主成分分析算法相当而局部线性嵌入失败;在流形学习降维中,该算法与局部线性嵌入基本相当而主成分分析算法完全失败。

关键词 降维 主成分分析 局部线性嵌入 扩散映射

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2022.03.043

DIMENSION REDUCTION ALGORITHM BASED ON RANDOM WALK DIFFUSION MAPPING

Xue Yanfeng[1,2,3] Wang Sanhu[3] Gao Zhie[3] Gao Yongqiang[3]

[1](*School of Computer and Information Technology, Shanxi University, Taiyuan 030006, Shanxi, China*)

[2](*Institute of Complex Systems, Shanxi University, Taiyuan 030006, Shanxi, China*)

[3](*Department of Computer Science and Technology, Luliang University, Lvliang 033000, Shanxi, China*)

Abstract Traditional data dimensionality reduction algorithms are divided into linear or manifold learning dimensionality reduction algorithms, but in practical applications, it is difficult to determine which kind of algorithm is needed. A comprehensive data dimensionality reduction algorithm is designed to ensure that the lower limit of its linear dimensionality reduction effect is the principal component analysis(PCA) algorithm, and the data structure of manifold can be revealed in the aspect of manifold learning dimensionality reduction. By constructing Markov transition matrix for high-dimensional data, the more similar nodes have greater transition probability, so as to find the mapping relationship between high-dimensional data dimensionality reduction and low-dimensional manifold. The experimental results show that in the linear dimensionality reduction of artificial data and real data, the dimensionality reduction effect of this algorithm is equivalent to that of PCA algorithm, while the locally linear embedding(LLE) failed. In the manifold learning dimensionality reduction, this algorithm is basically equivalent to that of LLE, but PCA algorithm fails completely.

Keywords Dimension reduction Principal component analysis Locally linear embedding Diffusion mapping

0 引 言

降维是通过线性或非线性的映射关系将高维数据转换到低维数据的过程,且该低维数据代表原始高维数据的主要成分,并能描述原始高维数据的空间分布结构。一般情况下,由于降维后的数据更易于被分类、识别、可视化、存储等,故降维在机器学习[1]以及数据可视化[2-3]领域受到越来越多的关注。

现有的降维算法主要分为线性降维和流形学习降维,其中:线性降维仅对于数据维数相对较低且具有全局线性结构的数据有着良好的降维效果,代表算法包

收稿日期:2020-01-01。山西省高等学校科技创新项目(2020L0715);山西省教育厅教学改革项目(J2018195)。**薛艳锋**,讲师,主研领域:机器学习,复杂网络。**王三虎**,教授。**高志娥**,讲师。**高永强**,副教授。

括主成分分析[4-5]、线性判别分析[6-7]、多尺度分析[8]算法等;流形学习降维主要是把高维空间的内在结构或本质特征在低维空间尽量得以保留,代表算法包括局部线性嵌入[9]、核主成分分析[10]、ISOMAP[11]算法等。然而,在实际的科学研究中,需要一种统一的降维算法,使得线性降维效果与线性降维算法相当(本文选择的参照对象为主成分分析算法),同时流形学习降维效果尽可能合理(参照对象为局部线性嵌入)。

为此,本文利用数据点属性之间的欧氏距离定义了数据随机游走的转移概率矩阵 $\boldsymbol{A}$,然后通过归一化矩阵 $\boldsymbol{A}$ 得到马尔可夫转移矩阵 $\boldsymbol{M}$(该矩阵 $\boldsymbol{M}$ 描述数据的离散扩散过程),其次通过该矩阵 $\boldsymbol{M}$ 得到对应的拉普拉斯矩阵 $\boldsymbol{L}$,最后按照该矩阵 $\boldsymbol{L}$ 的特征值升序排列对应的特征向量,按照累积特征值比例,原始数据依次投射到对应特征向量(从第 2 个特征向量开始)上。通过实验结果表明,在线性降维方面,本文算法与主成分分析算法相当,而局部线性嵌入失败;在流形学习降维方面,主成分分析算法失败,而本文算法虽然不及局部线性嵌入,但反映的内在结构一致。

1　算法描述

1.1　局部线性嵌入算法

局部线性嵌入的核心思想是假设数据在较小的局部是线性的,也就是说一个数据可以由它邻域的几个样本来线性表示。具体过程如下:假设有样本 x_1,在该样本的原始邻域中用 K 近邻思想找到其中的 K(超参数)个样本 $x_2,x_3,\cdots,x_{k+1}$ 且可以由它们线性表示:

$$x_1 = w_{1,2}x_2 + w_{1,3}x_3 + \cdots + w_{1,k+1}x_{k+1} \tag{1}$$

式中:$w_{1,2}$、$w_{1,3}$、$w_{1,k+1}$ 为权重系数。降维后,希望 x_1 在低维空间对应的投影 x'_1 和 $x_2,x_3,\cdots,x_{k+1}$ 对应的投影 $x'_2,x'_3,\cdots,x'_{k+1}$ 也尽量保持同样的线性关系,即:

$$x'_1 \approx w_{1,2}x'_2 + w_{1,3}x'_3 + \cdots + w_{1,k+1}x'_{k+1} \tag{2}$$

最后,通过均方差定义损失函数并求其权重系数:

$$\arg\min_W J(W) = \sum_i^N \left| x_i - \sum_j w_{i,j}x_j \right|^2 \tag{3}$$

式中:N 为样本数据点的个数;权重系数 $w_{i,j}$ 表示第 j 个数据点对重建第 i 个数据点的贡献,如果 x_j 不是 x_i 的邻居,则 $w_{i,j}=0$;最后归一化权重系数,使其 $\sum_{j=1}^{K} w_{i,j} = 1$。

1.2　随机游走扩散映射算法

局部线性嵌入的局部线性关系只在样本附近起作用,离样本远的样本对该样本的线性关系没有影响且影响样本点是确定的。随机游走扩散映射的思想是 K 近邻思想的扩展,即所有其他样本都起作用,只是距离较近的样本比距离较远的样本影响更大且影响样本点是随机的。

随机游走扩散映射算法的具体步骤如下:

1) 计算 N 个高维数据点之间的相似性度量,以构造具有元素 $N \times N$ 成对距离的矩阵 $\boldsymbol{D}$,其元素为 $D_{ij} = \|x_i - x_j\|$,其中 $\|\cdot\|$ 为适当距离度量,本文选择欧氏距离。

2) 使用这些距离来定义数据上的随机游走,从点 i 到点 j 的跳跃概率为:

$$A_{ij} = \exp\left(-\frac{D_{ij}^2}{2\varepsilon}\right) \tag{4}$$

式中:ε(>0,为超参数)为软阈值带宽,用于限制 $\sqrt{\varepsilon}$ 邻域内点之间的跳跃。

3) 求对角矩阵 $\boldsymbol{\Sigma}_{ii} = \sum_{j=1}^{N} A_{ij}$,标准化转移矩阵 $\boldsymbol{A}$,得到马尔可夫转移矩阵 $\boldsymbol{M}$:

$$\boldsymbol{M} = \boldsymbol{\Sigma}^{-1}\boldsymbol{A} \tag{5}$$

设原始数据点降维的映射关系为 f,则 $f(i)$($i=1,2,\cdots,N$)为低维空间的坐标点,通过目标函数 $\Phi(f)$ 求其映射关系 f:

$$\arg\min_f \Phi(f) = \sum_{i,j} M_{ij}(f(i) - f(j))^2 \tag{6}$$

4) 令 $\boldsymbol{f} = [f(1)\ \ f(2)\ \cdots f(n)]^{\mathrm{T}}$ 且 $\boldsymbol{f}^{\mathrm{T}}\boldsymbol{f} = 1$,并求其马尔可夫转移矩阵 $\boldsymbol{M}$ 的拉普拉斯矩阵 $\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{M}$,则式(6)可化为如下矩阵形式:

$$\arg\min_f \Phi(f) = 2\boldsymbol{f}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{f} \tag{7}$$

即转化为求特征值问题 $\boldsymbol{L}\boldsymbol{f} = \lambda\boldsymbol{f}$,其中 λ 表示拉普拉斯矩阵 $\boldsymbol{L}$ 的任一特征值。

5) 按照拉普拉斯矩阵 $\boldsymbol{L}$ 的定义以及对称矩阵的性质可知,存在最小的特征值 $\lambda_1 = 0$。按照特征值升序排序 $\lambda_1 \leqslant \cdots \leqslant \lambda_N$,则任一数据 x 映射到 q 维实数空间的坐标为:

$$\Psi(x) = (\lambda_2 f_x^{(\lambda_2)}, \lambda_3 f_x^{(\lambda_3)}, \cdots, \lambda_{q+1} f_x^{(\lambda_{q+1})}) \tag{8}$$

式中:$f_x^{(\lambda_i)}$($i=2,3,\cdots,N$)表示任意数据 x 在拉普拉斯矩阵 $\boldsymbol{L}$ 特征值对应的特征向量上的投影值。

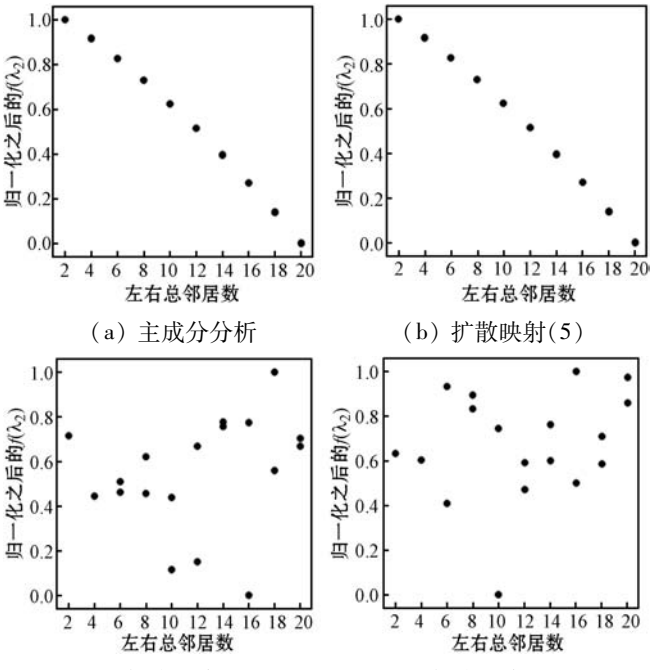
2　实　验

2.1　人造数据(线性降维)

该人造数据集为小世界网络[12-13],通过 Python 库 NetwordX[14] 下面 watts_strogatz_graph(n,k,p) 函数生成,其中:n 表示节点个数;k 表示环状的邻居个数;p 表示每条边的重连概率。本文选择节点个数 n 为 100,k 从 {2,4,6,8,10,12,14,16,18,20} 中等概率随

机选择,p 从$\{10^{-1},10^{-2},10^{-3},10^{-4},10^{-5},10^{-6},10^{-7}\}$中等概率随机选择,重复执行该函数生成 2 000 个小世界网络,每个网络的特征表示为三元组(边密度,开三元组密度,闭三元组密度)[13],最后依据该特征删除重复数据最终得到实验数据 1 317 条。该三元组特征降为 1 维后,为了更好可视化降维效果,需要经过归一化处理且在散点图中以纵坐标轴显示,左右总邻居数以横坐标轴显示。其中各个标题名称括号里的数字为超参数设置,比如图 1(b)括号中数字为扩散映射的软阈值带宽,图 1(c)和图 1(d)括号中数字为局部线性嵌入的最近邻个数。

从图 1 可知,主成分分析与扩散映射降维之后的特征与小世界网络的左右总邻居数有着严格的对应关系,这是由于 watts_strogatz_graph(n,k,p)函数生成的边数 k 是确定的,而重连概率 p 虽然是随机值,但由于有开三元组密度和闭三元组密度对该生成的小世界网络进行描述,故降维之后仍能保持与小世界网络左右总邻居数的严格对应关系,而局部线性嵌入降维之后无法刻画这种严格的对应关系。而且,主成分分析算法与扩散映射的累积方差贡献率分别为 98.95、98.93,定量说明了随机游走扩散映射算法在线性降维方面与主成分分析算法效果相当。



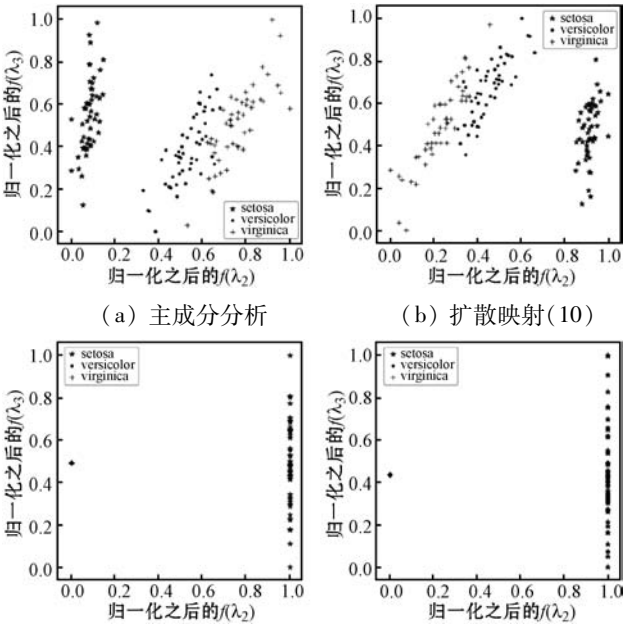
(a) 主成分分析 (b) 扩散映射(5)

(c) 局部线性嵌入(8) (d) 局部线性嵌入(14)

图 1 小世界网络人造数据

2.2 真实数据(线性降维)

出于可视化的考虑,真实数据集为鸢尾花数据集,数据降为 2 维,且每一类数据通过散点图分别以不同的形状显示。如图 2 所示,主成分分析与扩散映射算法把“setosa”类与其他两类明显分开,局部线性嵌入算法也达到同等的分类效果;同时,主成分分析与扩散映射算法基本把剩余两类(versicolor 与 virginica 类)基本分开,而局部线性嵌入把这两类嵌入到二维空间的同一坐标,即分类失败。



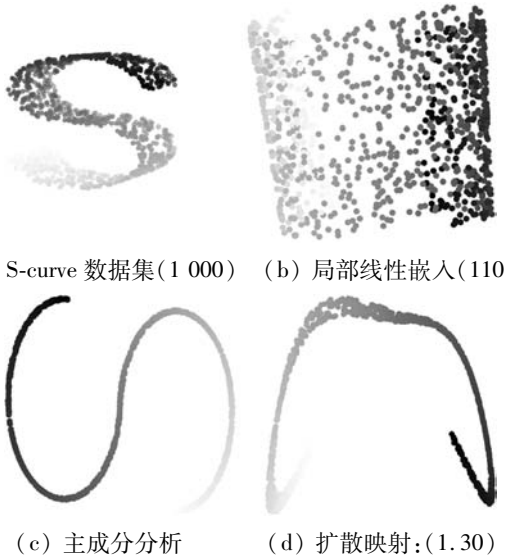
(a) 主成分分析 (b) 扩散映射(10)

(c) 局部线性嵌入(8) (d) 局部线性嵌入(14)

图 2 鸢尾花真实数据集

2.3 人造数据(流形学习降维)

该人造数据集为 S-curve 数据集[15],即流形数据集(一个不闭合的曲面),三维显示如图 3(a)所示(括号中数字为数据点数量),流形曲面具有数据分布比较均匀且比较稠密的特征,流形学习降维就是将流形从高维到低维的映射过程,在该降维过程中,流形的高维特征尽可能在低维空间得以保留。在本文,就相当于把 S-curve 数据集从三维空间投影到二维空间,即把 S-curve 数据集展开到二维空间,展开的过程就是流形学习降维的过程,就像两个人拉开一样(如图 4 所示)。



(a) S-curve 数据集(1 000) (b) 局部线性嵌入(110)

(c) 主成分分析 (d) 扩散映射:(1.30)

图 3 S-curve 人造数据集

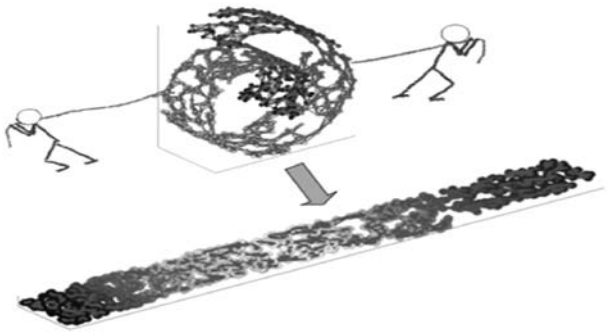


图 4　流形学习降维示意图

由图 3(b)可知,局部线性嵌入降维基本成功,即成功展开到二维平面。虽然左右两端没有展开,即在二维平面内有数据点重叠现象,但是设想从左右两边观察 S-curve 数据集的话,确实最下面与最上面互相重叠,且把三维空间下面的数据点映射到二维空间的左面,上面的数据点映射到二维空间的右面符合人的直观认识。由图 3(c)可知,主成分分析算法流形学习降维失败,虽然“S”型轮廓从 3 维空间可以前后观察得到,但流形学习降维的目的是展开该数据集,且主成分分析算法把上面的数据集映射到二维空间左面,下面的数据集映射到二维空间右面不符合人的直观认识。图 3(d)为扩散映射的降维效果,从展开效果看,不及局部线性嵌入,但对比局部线性嵌入算法,扩散映射降维算法左右两端对称的 V 字结构与原始 S-curve 数据集从三维空间左右观察有数据点重叠一致,且二维空间展开效果与人的直观认识一致。最后,三个算法就流形学习降维的效果比较如表 1 所示。

表 1　流形学习降维效果比较

算法	展开效果	直观认识
局部线性嵌入	基本全面展开	符合
扩散映射	未展开,但能部分说明数据点重叠现象	符合
主成分分析	未展开	不符合

3　结　语

本文算法在线性降维效果方面,与主成分分析算法相当,局部线性嵌入完全失败;而在流形学习降维方面,对标局部线性嵌入,主成分分析算法在展开效果与直观认识上全面失败,而扩散映射在展开效果上虽然不理想,但与左右两端数据点重叠结果相一致,且符合直观认识。

今后的研究方向是改进扩散映射算法,使它在流形学习降维效果有所提升,比如设计随机游走的路径以及步长,设计更合理的距离度量以及转移概率等。

参　考　文　献

[1] Sachdev K, Gupta M K. Predicting drug target interactions using dimensionality reduction with ensemble learning[C]// International Conference on Robotics and Intelligent Control, 2019.

[2] Sun G, Zhang S, Zhang Y, et al. Effective dimensionality reduction for visualizing neural dynamics by laplacian eigenmaps[J]. Neural Computation, 2019, 31(7): 1356 – 1379.

[3] Lorenzo A D, Medvet E, Tušar T, et al. An analysis of dimensionality reduction techniques for visualizing evolution [C]//Genetic and Evolutionary Computation Conference Companion, 2019.

[4] Guo L, Wu P, Lou S, et al. A multi-feature extraction technique based on principal component analysis for nonlinear dynamic process monitoring[J]. Journal of Process Control, 2020, 85: 159 – 172.

[5] Alver A, Kazan Z. Prediction of full-scale filtration plant performance using artificial neural networks based on principal component analysis [J]. Separation and Purification Technology, 2020, 230: 115868.

[6] 杨悦,顾晓瑜. 基于线性判别分析的室内声源定位方法[J]. 计算机技术与发展,2017,27(6):187 – 190,194.

[7] 杨茜. 基于 Fisher 线性判别分析的情景感知推荐方法[J]. 计算机工程与设计,2018,39(3):848 – 853.

[8] Leonid B, Daniel C O. Outlier detection for robust multi-dimensional scaling[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(9): 2273 – 2279.

[9] 邱建荣,罗汉. 改进的局部线性嵌入算法及其应用[J]. 计算机工程与应用,2020,56(3):176 – 179.

[10] Xu D, Wang Y, Peng P, et al. Kernel PCA for road traffic data nonlinear feature extraction[J]. IET Intelligent Transport Systems, 2019, 13(8): 1291 – 1298.

[11] Liu Q, Cai Y, Jiang H, et al. Traffic state prediction using ISOMAP manifold learning[J]. Physica A: Statistical Mechanics and Its Applications, 2018, 506: 532 – 541.

[12] Wills P, Meyer F G. Metrics for graph comparison: A practitioner's guide[EB]. arXiv: 1904.07414, 2019.

[13] Kattis A A, Holiday A, Stoica A A, et al. Modeling epidemics on adaptively evolving networks: A data-mining perspective[J]. Virulence, 2016, 7(2): 153 – 162.

[14] Hagberg A A, Schult D A, Swart P J. Exploring network structure, dynamics, and function using NetworkX[C]//7th Python in Science Conference, 2008.

[15] S-curve 数据集[DB/OL]. [2019 – 12 – 24]. https://scikit – learn.org/stable/datasets/index.html.