面向复杂多流形高维数据的 t -SNE 降维方法

边荣正¹⁾, 张鉴²⁾, 周亮³⁾, 蒋鹏⁴⁾, 陈宝权¹⁾, 汪云海^{1)*}

¹⁾ (山东大学计算机科学与技术学院 青岛 266237)

²⁾ (中国科学院计算机网络信息中心 北京 100190)

³⁾ (Scientific Computing and Imaging Institute, University of Utah Salt Lake City 84112)

⁴⁾ (山东大学齐鲁交通学院 济南 250002)

(sdujump@gmail.com, cloudseawang@gmail.com)

摘要: 针对 t -SNE 方法不能很好地区分相互交叉的多个流形的问题, 提出一种可视化降维方法. 在 t -SNE 方法的基础上, 在计算高维概率时考虑欧几里得度量和局部主成分分析以区分不同流形. 然后可直接使用 t -SNE 的梯度求解方法得到降维结果. 最后分别用 3 个人工生成的三维数据集和 2 个通用的机器学习数据集进行实验, 并根据不同流形的区分度和流形内的邻域可信度 2 个指标对降维结果进行量化分析. 结果表明, 该方法在处理有交叉的多流形数据时的效果要明显优于原来的 t -SNE 方法, 并能够较好地保持每个流形的邻域结构.

关键词: 降维方法; 局部主成分分析; 多流形数据; 可视化

中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18806

t -SNE for Complex Multi-Manifold High-Dimensional Data

Bian Rongzheng¹⁾, Zhang Jian²⁾, Zhou Liang³⁾, Jiang Peng⁴⁾, Chen Baoquan¹⁾, and Wang Yunhai^{1)*}

¹⁾ (School of Computer Science and Technology, Shandong University, Qingdao 266237)

²⁾ (Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190)

³⁾ (Scientific Computing and Imaging Institute, University of Utah, Salt Lake City 84112)

⁴⁾ (School of Qilu Transportation, Shandong University, Jinan 250002)

Abstract: To solve the problem that the t -SNE method cannot distinguish multiple manifolds that intersect each other well, a visual dimensionality reduction method is proposed. Based on the t -SNE method, Euclidean metric and local PCA are considered when calculating high-dimensional probability to distinguish different manifolds. Then the t -SNE gradient solution method can be directly used to get the dimensionality reduction result. Finally, three generated data and two real data are used to test proposed method, and quantitatively evaluate the discrimination of different manifolds and the degree of neighborhood preservation within the manifold in the dimensionality reduction results. These results show that proposed method is more useful when processing multi-manifold data, and can keep the neighborhood structure of each manifold well.

Key words: dimensionality reduction; local principal component analysis; multi-manifold; visualization

收稿日期: 2020-11-21; 修回日期: 2021-01-24. 基金项目: 国家自然科学基金(61772315, 61861136012). 边荣正(1995—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为可视化; 张鉴(1972—), 男, 博士, 副研究员, 博士生导师, 主要研究方向为高性能计算、偏微分方程数值算法和并行计算; 周亮(1986—), 男, 博士, 助理教授, 副研究员, 主要研究方向为可视化; 蒋鹏(1986—), 男, 博士, 助理研究员, 主要研究方向为计算机视觉、图像处理、人工智能; 陈宝权(1969—), 男, 博士, 教授, 博士生导师, CCF 会员, 主要研究方向为计算机图形学与可视化; 汪云海(1984—), 男, 博士, 教授, 博士生导师, CCF 会员, 论文通讯作者, 主要研究方向为可视化与人机交互.

近年来, 高维数据的可视化已经成为可视化领域中的一个重要研究方向. 通过对新的高维数据可视化, 可以为研究者后续的数据挖掘与分析提供重要的参考. 当前最常用的高维数据可视化方法之一是先将高维数据降维至二维或三维(通常是二维), 然后用散点图表示. 有研究发现, 虽然人们收集的数据是高维的, 但是往往有一些维度是冗余的, 真正需要的数据可能仅仅是一个本质上拥有较低维度的流形, 它嵌入高维的空间当中^[1]. 这时, 对原高维数据的可视化任务本质上就是对这些流形结构的可视化.

在过去几十年中, 已有很多优秀的降维方法, 如最简单有效的主成分分析(principal component analysis, PCA)^[2]和多维缩放(multiple dimensional scaling, MDS)^[3], 它们在处理线性流形时有着不错的效果和速度, 但是对于非线性结构的效果十分有限. 而非线性方法, 如等度量映射(isometric mapping, Isomap)^[4]、局部线性嵌入(local linear embedding, LLE)^[5]和 t -分布随机邻域嵌入(t -distributed stochastic neighbor embedding, t -SNE)^[6], 通常能够较好地识别出数据中的非线性流形结构. 但是, 当数据中的多个流形结构存在相互交叉时, 这些方法往往无法准确地识别出原本的流形结构. 为解决此问题, 分别基于 Isomap 和 LLE 提出了针对多流形数据的 MMD-Isomap 方法^[7]和 MM-LLE 方法^[8]. 但是, 它们都是有监督的方法, 需要事先得知数据中的流形划分. 在现实应用中, 更多的情况下, 往往无法事先得知数据中的流形划分. 因此, 设计一种能够识别数据集中多流形结构的无监督降维方法具有十分重要的意义.

虽然流形数据在高维空间中看上去十分复杂, 但是在局部范围内仍然具有欧几里得空间的性质^[1]. 对数据进行 PCA 可以得到数据的主要线性结构特征^[9]. Arias-Castro 等^[10]和 Wang 等^[11]分别基于局部 PCA 改进了传统的谱聚类方法, 证明了局部 PCA 可有效地区分数据集中的相互交叉的流形结构.

因此, 本文将局部 PCA 引入高维空间中样本的距离度量, 对 t -SNE 方法中高维概率的计算部分做了修改. 在本文方法中, 样本之间的距离度量采用的是欧几里得度量与局部 PCA 不相似度的加权和, 然后用此距离代替原 t -SNE 方法中计算高维概率时所用的欧几里得度量. 通过调节合适的权重参数, 该方法既可以较好地地区分数据中的不同流形结构, 同时又能在一定程度上展示每个流形自

身的结构特点. 由于只对 t -SNE 的高维概率部分的计算做了修改, 并不影响迭代求解部分, 因而可以直接使用现有的 t -SNE 求解框架, 也可以直接使用现有的针对 t -SNE 求解的加速方法加快求解速度. 最重要的是不同于 MMD-Isomap 与 MM-LLE, 本文方法是无监督的, 在降维之前不需要得知数据集中类的划分, 因而更适用于数据的预分析可视化.

1 相关工作

本文方法的灵感主要来自 Arias-Castro 等^[10]的谱聚类方法, 把局部 PCA 引入 t -SNE 降维框架中. 本节将主要介绍 t -SNE 算法以及局部 PCA 在多流形谱聚类中的应用.

1.1 t -SNE 降维方法

t -SNE 是目前最受欢迎的可视化降维方法之一. 给定一个数据集 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{M \times n}$, 其中包含 n 个 M 维的点. 假设要将其降维至 m 维, $m \ll M$, 降维结果为 $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbb{R}^{m \times n}$. t -SNE 把样本之间的距离关系转换为概率值表示它们之间的相似性, 并假设样本在高维空间中符合高斯分布, 在低维空间中符合 t -分布. t -SNE 的目标是使这 2 个分布的 KL 散度(Kullback-Leibler divergence)最小, 即

$$\min F = \text{KL}(\mathbf{P} \parallel \mathbf{Q}) = \sum_i \sum_j p_{ij} \text{lb} \frac{p_{ij}}{q_{ij}} \quad (1)$$

其中, $\mathbf{P}$ 与 $\mathbf{Q}$ 均为 $n \times n$ 的矩阵, 其元素分别为 p_{ij} 和 q_{ij} . p_{ij} 为高维空间中 $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 之间的概率, 其计算公式为

$$p_{ij} = \frac{p_{ji} + p_{ilj}}{2n} \quad (2)$$

其中,

$$p_{ji} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2\right)} \quad (3)$$

σ_i 反映了样本 $\mathbf{x}_i$ 周围的样本分布密度, 具体值与用户输入的困惑度(perplexity)参数有关. 困惑度定义为

$$\rho(\mathbf{P}_i) = 2^{H(\mathbf{P}_i)} \quad (4)$$

其中,

$$H(\mathbf{P}_i) = -\sum_j p_{ji} \text{lb} p_{ji} \quad (5)$$

实际计算时, 需要使用二分法不断地尝试不同的

σ_i 取值, 根据式(4)计算的困惑度与用户输入一致.

q_{ij} 为低维空间中的概率, 其计算公式为

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} \quad (6)$$

由于 t -分布长尾分布的特征, t -SNE 会把高维空间中距离较近的点放得更近, 距离较远的点放得更远. 这个特性使 t -SNE 在对一般的多类数据或多流形数据进行可视化时往往能够得到不错的效果, 但是当多流形数据中不同的流形之间存在交叉情况时, t -SNE 的结果往往不能达到预期, 第 3 节给出了这样的例子.

t -SNE 的目标函数不是凸函数, 只能通过梯度方法迭代求解得到某个局部最优解. t -SNE 目标函数的梯度为

$$\frac{\delta F}{\delta y_i} = 4 \sum_j (p_{ij} - q_{ij}) (1 + \|y_i - y_j\|^2)^{-1} (y_i - y_j) \quad (7)$$

使用 t -SNE 方法^[6]的具体步骤如下.

输入. 高维数据集 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{M \times n}$, 困惑度 ρ , 迭代次数 T , 学习率 η , 动量 $\alpha(t)$.

输出. 降维结果 $\mathbf{Y}^{(T)} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbb{R}^{m \times n}$.

Step1. 根据 ρ , 使用式(3)计算概率 p_{ji} .

Step2. 根据式(2)计算对称的概率.

Step3. 生成随机的初始化结果 $\mathbf{Y}^{(0)} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n]$.

Step4. for $t = 1$ to T do:

Step4.1. 根据式(4)计算低维的相似度概率;

Step4.2. 根据式(5)计算梯度;

Step4.3. $\mathbf{Y}^{(t)} = \mathbf{Y}^{(t-1)} + \eta \frac{\delta F}{\delta \mathbf{Y}^{(t-1)}} + \alpha(t)(\mathbf{Y}^{(t-1)} - \mathbf{Y}^{(t-2)})$.

但是, 直接使用梯度方法求解速度会比较慢, 一般在应用中可以使用 4 叉树的方法^[12]将求解的时间复杂度从 $O(n^2)$ 降至 $O(n \log n)$, 从而大大提高收敛的速度.

1.2 局部 PCA 在流形聚类中的应用

除了可视化降维之外, 对多流形数据的聚类研究也是一个重要的研究方向. 传统的聚类方法, 如 k -means^[13], DB-SCAN^[14]和谱聚类^[15]等, 不能很好地处理流形交叉的情况. Wang 等^[11]和 Arias-Castro 等^[10]分别以不同的方式在谱聚类算法中引入了局部 PCA, 大大提高了有交叉结构的多流形聚类的正确率, 证明了局部 PCA 在处理交叉流形数据方面具有很好的作用.

Wang 等^[11]基于局部 PCA 中各主成分方向的角度给出样本之间相似度的度量, 并将其作为谱

聚类算法中的相似度, 提出了多流形谱聚类 (spectral multi-manifold clustering, SMMC) 方法. 其相似度的计算公式为

$$W_{ij} = \Delta_{ij} \cdot \left(\prod_{s=1}^d \cos \theta_s(i, j) \right)^\alpha \quad (8)$$

其中, 当且仅当 $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 是邻居时 $\Delta_{ij} = 1$, 否则 $\Delta_{ij} = 0$; 而 $\theta_1(i, j) \geq \dots \geq \theta_d(i, j)$ 为 $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 的局部 PCA 中主成分的角度^[16]; α 为一个用户可以调节的参数.

Arias-Castro 等^[10]在数据集中随机地选择中心点, 这些中心点之间的距离大于事先给定的一个邻居半径 r . 采用

$$W_{ij} = \exp \left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\varepsilon^2} \right) \cdot \exp \left(-\frac{\|\mathbf{V}_i - \mathbf{V}_j\|^2}{\eta^2} \right) \quad (9)$$

计算这些中心点样本之间的相似度. 然后, 将此相似度作为谱聚类算法的相似度, 对这些中心点进行聚类. 每个中心点周围的样本直接划分到这个中心点所在的类中去. 这里, $\mathbf{V}_i$ 为样本 $\mathbf{x}_i$ 的局部 PCA 中的前 d 个特征向量所形成的投影矩阵; ε 和 η 为需要用户设置的参数.

这 2 个方法都是将欧几里得度量与局部 PCA 结合起来, 大幅提高了谱聚类方法处理多流形数据的正确率, 充分证明了局部 PCA 在处理多流形数据的有效性.

因此, 本文借鉴了这些多流形聚类方法的思想, 将局部 PCA 引入 t -SNE 中, 提出了一种针对多流形数据的可视化降维方法.

2 方法设计

在一个多流形数据集中选择 2 个点 $\mathbf{x}_i$ 和 $\mathbf{x}_j$, 可能会出现以下 4 种情况.

(1) $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 欧几里得度量很近, 并且位于同一个流形结构上.

(2) $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 欧几里得度量很近, 但位于不同的流形结构上.

(3) $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 欧几里得度量较远, 并且位于同一个流形结构上.

(4) $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 欧几里得度量较远, 并且位于不同的流形结构上.

单纯基于欧几里得度量的方法无法很好地处理流形交叉的情况, 主要是无法处理第(1)和第(2)

种情况. 因此, 问题的关键是使这 2 种情况能够区分开. Arias-Castro 等^[10]根据不同流形上的样本的局部 PCA 有较大差别的规律, 将根据欧几里得度量计算的相似度与根据局部 PCA 计算的相似度相乘, 使第(1)种情况的样本间相似度远大于第(2)种情况. 需要注意, 根据指数函数的性质, 式(9)可以改写为欧几里得度量与局部 PCA 加权的形式, 即

$$W_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\varepsilon^2} - \frac{\|\mathbf{V}_i - \mathbf{V}_j\|^2}{\eta^2}\right) \quad (10)$$

其中, ε 与 η 实际上起到了调节权重的作用.

2.1 基于局部 PCA 的距离度量

对于高维空间中的每个样本 $\mathbf{x}_i$, 首先需要得到它的邻居集合 N_{x_i} . N_{x_i} 可以选择距离 $\mathbf{x}_i$ 小于设定的一个半径值 r 的点, 或者直接选择距离 $\mathbf{x}_i$ 最近的 K 个点. N_{x_i} 中点的协方差矩阵为

$$C_i = \frac{1}{|N_{x_i}|} \sum_{\mathbf{x}_j^{(j)} \in N_{x_i}} (\mathbf{x}_i^{(j)} - \boldsymbol{\mu}_i)(\mathbf{x}_i^{(j)} - \boldsymbol{\mu}_i)^T \quad (11)$$

其中, $\boldsymbol{\mu}_i = \frac{1}{|N_{x_i}|} \sum_{\mathbf{x}_k \in N_{x_i}} \mathbf{x}_k$ 为 N_{x_i} 中所有样本的均值.

对 C_i 进行特征值分解可以得到 M 个特征向量 $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M$ 和 M 个特征值 $\lambda_1, \lambda_2, \dots, \lambda_M$, 默认这些特征值已按照从大到小的顺序排序, 即 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_M$. 假设流形的真实维度为 d , 那么只需要用前 d 个特征向量就可以表示 $\mathbf{x}_i$ 处的局部切空间. 令 $\mathbf{U}_i = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_d] \in \mathbb{R}^{M \times d}$ 的 d 列分别为 C_i 的前 d 个特征向量, 则 $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 的局部 PCA 的差别可以表示为

$$S_{ij} = \|\mathbf{U}_i - \mathbf{U}_j\| \quad (12)$$

其中, $\|\cdot\|$ 表示矩阵的谱范数. 但是, 这里 S_{ij} 的值往往会受矩阵 $\mathbf{U}$ 中特征向量符号的影响, 因此, 可以用 $\mathbf{U}_i \mathbf{U}_i^T$ 代替 $\mathbf{U}_i$, 这样实际方法中的 S_{ij} 定义为

$$S_{ij} = \|\mathbf{U}_i \mathbf{U}_i^T - \mathbf{U}_j \mathbf{U}_j^T\| \quad (13)$$

最后, 用样本间欧几里得度量 D_{ij} 与局部 PCA 差别 S_{ij} 的加权和作为本文方法中样本 $\mathbf{x}_i$ 与 $\mathbf{x}_j$ 新的距离度量, 即

$$D_{ij} = \alpha D_{ij} + (1 - \alpha) S_{ij} \quad (14)$$

其中, $\alpha \in [0, 1]$ 为调节欧几里得度量与局部 PCA 不相似度的权重. 为了方便设置权重参数, 避免 $\mathbf{D}$ 和 $\mathbf{S}$ 中数值的取值范围对权重造成影响, 需要事先对矩阵 $\mathbf{D}$ 和 $\mathbf{S}$ 做归一化处理, 即

$$D_{ij} \leftarrow D_{ij} / \max(\mathbf{D}); S_{ij} \leftarrow S_{ij} / \max(\mathbf{S}).$$

2.2 用 t -SNE 框架降维

用式(14)所示的距离度量代替 t -SNE 中计算概率 p_{ij} 所用的欧几里得度量, 即

$$p_{ji} = \frac{\exp(-D_{ij}^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-D_{ik}^2 / 2\sigma_i^2)} \quad (15)$$

然后用式(2)可以求得对称的概率. 需要说明的是, 这里的 σ_i 与 t -SNE 中类似, 代表了在新的距离标度量样本 $\mathbf{x}_i$ 周围的分布情况. 假如有较多的点与 $\mathbf{x}_i$ 的新距离较近, 则 σ_i 会较小, 否则, σ_i 会较大; 即随着 σ_i 的增大, p_{ij} 的熵也会增大. σ_i 的取值需要根据用户输入的困惑度确定.

通过观察 t -SNE 的梯度计算公式(式(7)), 可以发现对 p_{ij} 计算方式的修改并不影响其他参数, 因而可以继续使用 t -SNE 的求解框架进行降维计算.

本文方法的主要步骤如下.

输入. 高维数据集 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{M \times n}$, 权重参数 α , 邻域半径 r 或邻居数 K , 流形维度 d , 困惑度 ρ .

输出. 降维结果 $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \in \mathbb{R}^{m \times n}$.

Step1. 对每个样本 $\mathbf{x}_i$, 根据 r 或 $K3$ 计算其邻居集合 N_{x_i} .

Step2. 根据式(11)计算 N_{x_i} 的协方差矩阵 C_i , 对 C_i 特征值分解, 得到前 d 个特征向量组成的矩阵 $\mathbf{U}_i$.

Step3. 计算每对样本之间的欧几里得度量和局部 PCA 不相似度, 并根据式(14)计算得到新的距离度量.

Step4. 基于新的距离度量计算 t -SNE 中的高维概率, 调用 t -SNE 的求解方法计算最终的降维结果 $\mathbf{Y}$.

2.3 讨论

2.3.1 算法中的参数

式(14)中的 α 是本文方法中最重要的参数. 通过实验发现, 当 α 值较大时, 每个流形的样本更加倾向于展开, 但是当 α 取值过大时, 不同的流形不容易区分. 特别地, 当 α 接近于 1 时, 本文方法所得结果将接近于 t -SNE 方法, 而 α 取值较小时, 流形的结构往往不能充分展开. 因此, 一般情况下, 如果降维结果中各个流形的区分不够明显, 需要减小 α 的值; 如果希望每个流形的结构能够更加充分地展开, 则往往需要适当增大 α 的值.

式(11)中计算局部协方差矩阵时, 需要确定邻域内的样本. 可以选择与样本距离小于给定半径 r 或直接选择距离该样本最近的 K 个点. 具体的参数大小可以参考 LLE 算法的设置.

式(13)中矩阵 $\mathbf{U}$ 的特征向量个数的确定, 需要

用户输入一个流形的估计维度 d . 但是现实应用中往往无法事先得知 d , 当前一个可行的方案是根据局部 PCA 中特征值的大小, 满足前 d 个特征值的和占所有特征值的和比重超过某个阈值, 如 75%.

2.3.2 算法的时间复杂度

假设输入的高维数据包含 n 个 M 维的样本. 在计算局部 PCA 时, 首先要寻找每个点的邻居, 如果使用效率较高的 kD Tree 方法^[17]或 Ball-Tree 方法^[18], 其时间复杂度为 $O(M \log n)$. 求解所有样本的局部 PCA 的时间复杂度为 $O(nM^3)$, 计算所有样本对的局部 PCA 不相似度的时间复杂度为 $O(n^2M^3)$. 最后, 使用 t -SNE 求解框架迭代计算降维结果的时间复杂度为 $O(n^2)$, 通过使用 4 叉树的方法^[12]可以使其降至 $O(n \log n)$. 由于计算局部 PCA 和局部 PCA 不相似度所需的时间远高于计算欧几里得度量的时间, 因此, 与 t -SNE 方法相比, 本文方法运行效率较低. 如表 1 所示, 这里的 t -SNE 是调用的机器学习库 `scikit-learn`^①中的实现版本, 该版本是用 4 叉树方法^[12]优化过的.

表 1 本文方法与 t -SNE 时间对比

数据集	样本数	维数	时间/s	
			本文	t -SNE ^[6]
瑞士卷与平面交叉	1 520	3	12.45	7.83
美元符号数据	5 000	3	89.15	30.15
2 个抛物面交叉	6 000	3	124.69	38.44
MNIST-digit 7-9	2 037	50	30.23	14.53
COIL-10	720	50	5.78	3.95

2.3.3 与聚类方法比较

本文方法与 Arias-Castro 等^[10]的基于局部 PCA 的谱聚类方法思路是相似的, 在衡量 2 个样本之间的相似度时, 同时考虑了局部 PCA 与欧几里得度量. 但是, 这 2 种方法也有诸多不同之处. 首先, 2 种方法的目标不同, 后者是一个聚类方法, 其目标是将每个流形的样本聚集成一个簇, 而不必过于关心每个簇内部样本之间的位置关系. 而本文方法是一个用于可视化的降维方法, 在保证能够区分不同流形的同时, 还要适当表现每个流形自身的结构. 所以, 在 Arias-Castro 等^[10]的聚类方法中, 可以先对部分点聚类, 然后再将其余的点归入距离其最近的簇中, 而本文的降维方法需要对所有的点同时降维. 另外, 在实现细节方面, 本文方法与 Arias-Castro 等^[10]的方法本质上都是将局部 PCA

与欧几里得度量的加权和作为新的样本间距离度量, 并进一步转换为指数形式的相似度或概率. 但是在转换为指数形式时, 本文采用了原 t -SNE 方法中的方案, 式(15)中每个点都根据其周围的样本分布情况分配了不同的 σ 值, 因而能够更好地反映流形结构, 而在 Arias-Castro 等^[10]的聚类方法中没有相应的机制.

3 实验分析

为了验证本文方法的有效性, 分别使用合成数据和真实数据进行了实验分析, 并给出了量化评价. 本文实验均是在 64 位 Windows10 操作系统, Intel i5-2.90 GHz CPU, 16 GB 内存的环境下进行的, 程序实现所用的语言是 Python 3.7.

3.1 合成数据

在流形学习研究中有几个常用的经典人造数据集, 如瑞士卷和 S 形曲线. 这些数据集都是二维的流形嵌入三维的空间当中, 由于人类可以直接感知三维和二维空间, 因而通过比较不同降维方法处理这些数据集的效果, 可以比较直观地了解方法的有效性. 本文基于这些经典的数据合成了一些有交叉结构的数据, 验证本文方法的有效性. 在图中用不同类型的颜色表示不同的流形, 在同一流形中用渐变的颜色标记位置关系, 使同一流形上距离相近的点具有相似的颜色. 好的降维结果应该能够使不同的流形区分开, 并且同一流形内相邻的点仍然相邻.

3.1.1 瑞士卷与平面交叉数据

图 1a 所示的数据是由一个平面(图中红-黄色的点)和瑞士卷交叉组成的. 其中红色-黄色的平面共有 475 个点, 蓝色的瑞士卷共有 1 045 个点. 图 1b~图 1d 分别是邻居数为 10 的 Isomap 和 LLE 以及困惑度为 30.0 时 t -SNE 的降维结果, 可以看到, 这些方法不能区分开平面和瑞士卷. 图 1e~图 1h 是本文方法在相同困惑度下、 α 取不同值时, 选择 10 个点作为邻域的降维结果. 可以看到, 当 $\alpha=0.9$ 时, 本文方法能够很好地区分开 2 个流形, 并较好地反映了每个流形自身的结构, 唯一不太完美的地方是瑞士卷的右下方部分断开了一些.

3.1.2 美元符号数据

如图 2a 所示, 美元符号数据是由一个包含 3 000 个点的 S 形曲面和一个包含 2 000 个点的平面

① <https://scikit-learn.org/>

组成相互交叉组成的. 图 2b 是当困惑度取 30.0 时用 t -SNE 降维得到的可视化结果. 可以看到图中 2 部分样本各自被分成几簇并混合. 而图 2c 为相同困惑度下本文方法的结果, 其中 $\alpha=0.9$, 每个点的邻域内样本的个数为 10. 本文方法比较明显地区分出了这 2 个流形结构, 并且相似颜色的点的降维结果也基本是相邻的.

3.1.3 2 个抛物面

图 3a 是由 2 个抛物面交叉形成的数据, 每个抛物面有 3000 个点. 图 3b 是当困惑度取 30.0 时, 用 t -SNE 降维所得的可视化结果, 其中红色的样本与蓝色的样本几乎完全混合. 图 3c 是本文方法取得的结果, 这里设置 $\alpha=0.7$, 每个样本选择距离其最近的 10 个点作为其邻域, 困惑度同样取 30.0. 可以看到, 其划分效果要明显好于原来的 t -SNE 方法.

3.2 真实数据

3.2.1 MNIST digit 数据

MNIST digit 数据集^[19]为手写数字图片的集合, 每幅图片尺寸为 28×28 像素, 所以每个样本有 784 个维度. 这里本文选择了比较相似的数字 7 和 9 进行测试, 其中包括 1028 幅数字 7 的图片和 1009 幅数字 9 的图片. 当前 t -SNE 方法已经能够比较好地区分这些数字^[6], 为了测试这些流形存在交叉时的情况, 对这 2 类数据做了平移, 每类数据各减去其类的均值, 这样 2 类数据必然都分布在坐标原点周围, 即 2 类数据必然存在交叉. 由于是对每个类分别进行整体的平移, 因此每个类内部样本之间的相互位置关系不会发生变化, 即每个类内部的结构不会发生变化. 图 4a 所示的是平移前后手写数字图片的变化, 其中前 2 行是平移之前的

图 1 不同方法对瑞士卷与平面交叉数据的降维结果

图 2 不同方法对美元符号数据降维结果

图 3 不同方法对 2 个抛物面交叉数据降维结果

图片, 后 2 行是平移之后的图片. 在正式降维之前首先使用 PCA 方法对平移后的样本做预处理, 使其降维至 50 维. 图 4b 是困惑度取 30.0 时 t -SNE 方法的结果, 而图 4c 是在相同困惑度下, $\alpha=0.3$ 时本文方法的结果. 可以看到本文方法能够更好地区分开这 2 个相似的数据.

3.2.2 COIL 数据

COIL 数据集^①是研究流形学习时经常用到的一个数据集, 其中包含从 72 个不同角度拍摄的 20 个物体的照片, 相邻角度之间相差 5° , 因而每个物体的图片具有一个流形的结构. 每幅图片的尺寸为 128×128 像素, 所以这个数据集有 16384 个维度. 本文选择了后 10 个物体的图片进行实验, 与 MNIST 一样, 当前 t -SNE 方法已经能够比较好地区分出 COIL 中的不同的物体^[6]. 为了测试这些流形结构发生交叉时的效果, 同样对每类数据减去其类的均值, 进行了整体的平移, 并用 PCA 方法将其降维到了 50 维. 图 5a 显示的是平移前后的图片效果, 其中, 前 2 行是平移之前的图片, 后 2 行是平移之后的图片. 图 5b 所示的是困惑度为 30.0 时 t -SNE 的降维结果, 可以看到中间区域有很多类的数据混合在了一起, 区分度较低. 图 5c 是在相同困惑度下, $\alpha=0.3$ 时本文方法的结果, 各类样本被较好地区分了出来.

3.3 量化评价

为了进一步从数值上度量算法的效果, 本文

引入了 2 个评价指数来分别对降维结果中不同流形的区分度和每个流形自身结构的保持度进行量化评价.

3.3.1 不同流形的区分度

本文希望在降维结果中, 属于同一流形结构上的点聚集在一起, 而属于不同流形结构上的点尽可能分开. 因此, 可以借鉴线性判别式分析方法^[20]中类间散度与类内散度比值的形式对流形区分度进行量化. 在本文中流形区分度的计算方式为

$$R = \frac{\sum_{i=1}^L |X_i| (u_i - u)^T (u_i - u)}{\sum_{i=1}^L \sum_{x \in X_i} (x - u_i)^T (x - u_i)}$$

其中, L 为数据集中流形的个数; X_i 为第 i 个流形

所包含样本的集合; $u_i = \frac{1}{|X_i|} \sum_{x \in X_i} x$ 和 $u = \frac{1}{|X|} \sum_{x \in X} x$

分别为 X_i 和 X 中样本的均值. 对于同一个数据集, 区分度的值越大, 表示不同流形的区分效果越好, 反之则表示区分效果不明显. 表 2 给出了本文前面所用数据集降维结果的区分度值, 可以看到本文方法结果的区分度要明显远大于原 t -SNE 方法.

3.3.2 每个流形结构的保持程度

除了使不同流形的样本得以较好的区分之外, 本文还希望在降维结果中, 每个流形自身的结构不要遭到太严重的破坏, 即降维之后每个流形内部的局部邻域关系能够得到较好的保持. 在第 3.1

图 4 不同方法对 MNIST-digit 7-9 数据的降维结果

图 5 不同方法对 COIL-10 数据的降维结果

① <https://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php>

表 2 本文方法与 t -SNE 的区分度对比

数据集	t -SNE ^[6]	本文
瑞士卷与平面交叉	0.013	0.285
美元符号数据	0.001	0.459
2 个抛物面交叉	0.003	0.752
MNIST-digit 7-9	0.110	0.476
COIL-10	1.282	33.339

节中, 用渐变的颜色展示了三维数据降维结果中的邻域结构保持情况. 但是, 对于第 3.2 节中所用的更高维度的真实数据, 无法像三维数据那样直观地看到其流行结构的原貌. 因此, 需要另外用数值的方法来进行评价. 本文用邻域可信度^[21] (trustworthiness) 对这一信息进行度量. 邻域可信度表示降维前后某个样本邻域信息的相似程度, 该值越大, 表明降维结果中邻域关系保持得越好. 给定一个样本 $\mathbf{x} \in X_i$ 及其降维结果 $\mathbf{y}$, 在计算其邻域可信度时, 需要用到其 k -近邻信息. 由于本文的目的是衡量每个流形内邻域的保持情况, 因而与 Espadoto 等^[21]在整个数据集中寻找 $\mathbf{x}$ 的 k -近邻不同, 本文只考虑其所在流形 X_i 上的 k -近邻. 表 3 所示为一些数据集样本的平均邻域可信度. 可以看到, 本文方法的平均邻域可信度只比原 t -SNE 方法稍低, 大致持平, 即本文方法比较好地保持了每个流形自身的局部邻域关系.

表 3 本文方法与 t -SNE 的邻域可信度对比

数据集	t -SNE ^[6]	本文
瑞士卷与平面交叉	0.960	0.954
美元符号数据	0.971	0.956
2 个抛物面交叉	0.969	0.929
MNIST-digit 7-9	1.000	1.000
COIL-10	1.000	1.000

通过对不同流形的区分度以及每个流形结构的保持度的量化, 可以看到, 本文方法能够在较好地保持样本邻域的情况下, 显著提高不同流形的区分度.

4 结 语

本文针对当前降维方法无法有效处理有交叉的多流形数据的问题, 提出了一种无监督的多流形数据的可视化降维方法. 首先, 将样本之间的欧几里得度量与局部 PCA 差异相结合, 提出了一种样本之间的距离度量. 然后, 用这个距离度量代替

原 t -SNE 算法中计算高维空间中样本之间概率所用的欧几里得度量. 最后, 使用现有的 t -SNE 求解框架进行降维计算. 实验证明, 本文方法在处理有存在交叉的多流形数据时, 对不同流形的区分效果明显优于原 t -SNE 方法. 但是, 由于计算每个样本的局部 PCA 以及样本间局部 PCA 的差异时都需要进行矩阵的特征值分解, 本文方法的效率不及 t -SNE. 在接下来的工作中, 需要继续研究如何对算法进行优化, 以提高算法的效率.

参考文献(References):

- [1] Zhou Zhihua. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 225-246(in Chinese)
(周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 225-246)
- [2] Hotelling H. Analysis of a complex of statistical variables into principal components[J]. Journal of Educational Psychology, 1933, 24(6): 417-441
- [3] Torgerson W S. Multidimensional scaling of similarity[J]. Psychometrika, 1965, 30(4): 379-393
- [4] Tenenbaum J B, de Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 29(5500): 2319-2323
- [5] Saul L K, Roweis S T. Think globally, fit locally: unsupervised learning of low dimensional manifolds[J]. Journal of Machine Learning Research, 2003, 4: 119-155
- [6] van der Maaten L, Hinton G. Visualizing data using t -SNE[J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605
- [7] Yang B, Xiang M, Zhang Y P. Multi-manifold discriminant Isomap for visualization and classification[J]. Pattern Recognition, 2016, 55: 215-230
- [8] Hettiarachchi R, Peters J F. Multi-manifold LLE learning in pattern recognition[J]. Pattern Recognition, 2015, 48(9): 2947-2960
- [9] Bishop C M. Pattern recognition and machine learning[M]. Heidelberg: Springer, 2006: 559-603
- [10] Arias-Castro E, Lerman G, Zhang T. Spectral clustering based on local PCA[J]. Journal of Machine Learning Research, 2017, 18(9): 1-57
- [11] Wang Y, Jiang Y, Wu Y, et al. Spectral clustering on multiple manifolds[J]. IEEE Transactions on Neural Networks, 2011, 22(7): 1149-1161
- [12] van der Maaten L. Accelerating t -SNE using tree-based algorithms[J]. Journal of Machine Learning Research, 2014, 15(93): 3221-3245
- [13] Lloyd S. Least squares quantization in PCM[J]. IEEE Transactions on Information Theory, 1982, 28: 129-137
- [14] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C] //Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 1996: 226-231

- [15] von Luxburg U. A tutorial on spectral clustering[J]. *Statistics and Computing*, 2007, 17(4): 395-416
- [16] Stewart G W, Sun J G. *Matrix perturbation theory (computer science and scientific computing)*[M]. Boston: Academic Press Inc., 1990
- [17] Samet H. *The design analysis of spatial data structures*[M]. New York: Addison-Wesley Publishing Company, 1989: 66-80
- [18] Omohundro S M. *Five balltree construction algorithms*[R]. Berkeley: International Computer Science Institute, 1989
- [19] LeCun Y, Bottou L, Bengio Y, *et al.* Gradient-based learning applied to document recognition[J]. *Proceedings of the IEEE*, 1998, 86(11): 2278-2324
- [20] Fisher R A. The use of multiple measurements in taxonomic problems[J]. *Annals of Eugenics*, 1936, 7(2): 179-188
- [21] Espadoto M, Martins R M, Kerren A, *et al.* Towards a quantitative survey of dimension reduction techniques[J]. *IEEE Transactions on Visualization and Computer Graphics*, 2021, 27(3): 2153-2173