数据结构与算法(B)2020 年春课程论文

计算界的降维打击——浅谈流形学习与高维数据可视化方法

姓名：田晨霄
学号：1700013239
学院：数学科学学院
课堂：数据结构与算法（B）
指导教师：陈斌

【摘要】 本论文旨在首先通过概述和举例的形式简要回顾流形学习和数据聚类这一类算法的提出及发展的历史，特别的，其间我们会将重点放在介绍 2000 年以后提出的几种基于微分几何常识、线性嵌入以及概率统计思想的有代表性的流行学习算法。最后我们会总结这些降维算法的得与失，优缺点，并通过一些实际的例子，尽可能向读者们阐释这些算法应用在生产生活和科研工作中的场合和它们发挥的重要作用。

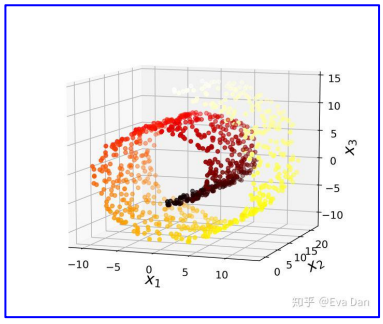
【关键词】 数据聚类 数据降维 流形学习 高维可视化方法

【正文】

一、流形学习算法的起源与发展

在刘慈欣的科幻作品《三体》中，歌者文明使用了二向箔将太阳系压入了二维空间，摧毁了整个地球文明，事实上虽然这种物理上的降维打击在目前看来，还是天方夜谭，遥不可及，但在人类的思维领域，如何将一个高维物体画到或者说铺展到一个平面上——以便于人们直观的去研究的方法，多少年来可以说已是一个老生常谈的话题，最早的方案可以追溯到为大家所熟知的斜二侧画法。

现代的流形学习的起源理论上可以追溯到 18 世纪的微分几何。200 多年来，随着几何学数学理论的研究深入，从原来朴素的平面和立体的欧式几何，以哥廷根学派的高斯等人为代表的数学家发展出了高维的欧式几何以及微分几何的严格的数学体系，并提出了高维空间中微分流形的概念。在此期间，特别的，黎曼还将高斯的思想进一步推广，发展出了黎曼几何和黎曼流形，为后来的广义相对论提供了有力的工具。另一方面，随着近代科学技术的发展，生产力的提高，越来越多的数据亟待处理与分析，数据科学也应运而生，其中一条基本假设便是现实生产生活中的绝大多数数据集都分布在某一个高维空间中的低维流形上。于是，结合以上几个方面，出于一种朴素的直观化思想、微分几何的基本理论和数据科学的基本假设，如何将大规模高维数据集尽可能直观地展现在一个二维平面或三维立体图中就近似的转换为了如何将一个高维空间中薄薄的嵌入的流形铺展到一个二维平面上或三维立体图中的问题(右图是一个流形学习中常用的测试数据集，为一个嵌入在三维空间中的二维流形，官称 *swiss roll*，也就是大家爱吃的甜品瑞士卷，许多包中，例如 Tensorflow 都自带有此数据集）。



至此，在理论发展和时代生产生活背景的驱动下，一大批流形学习与数据降维聚类的算法背后的提出的动机和基本思想就此诞生了。

早在上个世纪 60 年代，一种被称为 MSD 的算法，即一种基于最大化保持高维欧氏距离思想的将高数据集的铺展到二维平面上的算法就被数学家从理论上的提出，往后的几十年间，不少经典的聚类算法例如 PCA 主成分分析，K-均值算法，LE 局部图嵌入算法都被数学家或计算机科学家们提出，并随着计算机的发展得到了实际验证和应用。但真正标志着流形学习黄金时代到来的标志，还要属公元 2000 的时候，麻省理工学院计算机科学与人工智能实验室的 Josh Tenenbaum 教授于《Science》杂志上提出的 Isomap 方法，其实 Isomap 的本质还是 MSD 算法，但是巧妙的是，它把 MSD 算法中的欧式距离改为了微分几何中常用的测地线距离，由此，得到了一种很好的保持高维空间中流形拓扑性质的算法，开启了流形学习的纪元。随后 20 年间，最为代表性的，在线性降维中，有改良了 LE 方法的一种新的局部线性嵌入方法即所谓 LLE 方法；还有另辟蹊径，采用了概率统计中的条件概率代替欧氏距离，结合信息论中的 KL 散度衡量优化降维效果的 SNE 与 t-SNE 算法；以及到 2018 年最新提出的 U-map，它进一步改进了 t-SNE 算法，加快了聚类与流形学习的算法速度。

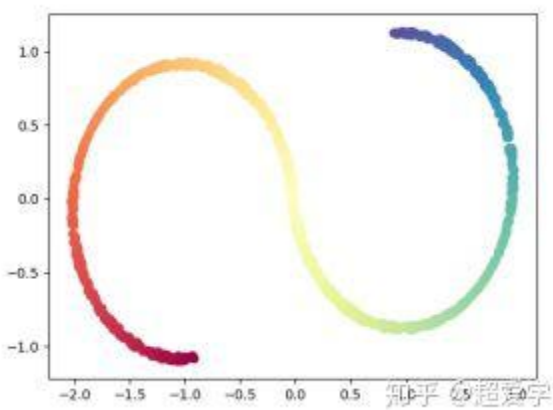
二、流形学习算法与一般数据聚类算法的联系和区别

从严格意义上讲，高维流形学习和高维数据聚类虽同属于高维数据可视化方法，但两者事实上有要求上的不同，高维数据聚类只要求把性质相似的高维数据点放置在一起，而流形学习则更进一步的提出了，要把高维流形尽可能的保持其拓扑性质的展开到一个平面上，也就说它不仅仅要求把相似的数据聚类那么简单，还需要按照高维空间中流形的几何或者说拓扑关系，让降维后的数据尽可能展示高维流形的本来面貌。当然两者之间也没有绝对的界限，因为降维总是要损失信息的，无论哪种算法都不可能百分之百的保全高维流形的拓扑性质，而且不同算法对不同流形的降维效果也是大相径庭，不过，判别一个算法是普通的数据聚类还是流形学习，关键还是看它的原理有没有去有意识的保一定拓扑性质，还有它的效果有没有达到保拓扑的要求。典型的流形学习算法就是 Isomap，即测地线方法，原理上是微分几何的常识，功能上也确实效果很好，而典型的聚类算法则是 PCA 主成分分析，它一般而言，起到的效果侧重点在聚类，但保持拓扑性质的功能很一般。

三、流形学习算法举例——的 Isomap 算法、SNE 与 t-SNE 算法，以及局部线性嵌入（LLE）简介

1．Isomap

Isomap 的主要适用范围是对于给定的高维流形，欲找到其对应的低维嵌入。它的核心目标是要使得高维流形上数据点间的近邻结构在低维嵌入中得以保持。正如前面所提到的，Isomap 事实上是发展了

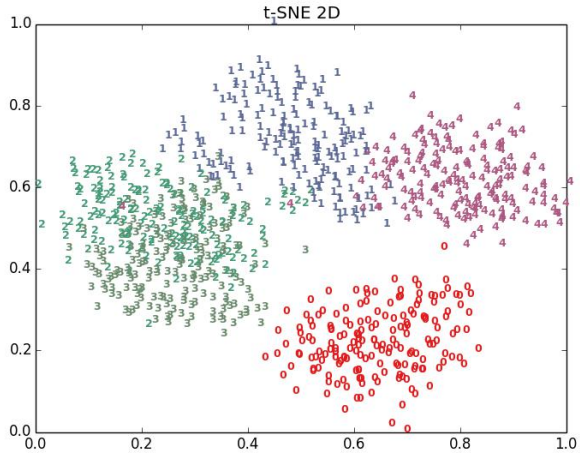


MDS(Multidimensional Scaling)算法，或者说以它为计算工具，唯一的不同之处在于计算高维流形上数据点间距离时，它没有用传统的欧式距离，而是采用微分几何常识中的测地线距离（或称为曲线距离）。另外，它在估计测地线距离时巧妙地采用了一种应用了实际的输入数据估计其测地线距离的算法（即图论中的最小路径逼近法)。这种算法保持拓扑的效果很好，如上图所展示的那样，它把一个嵌入在三维空间中的瑞士卷很好的展开在了二维平面上。

2. SNE 与 t-SNE 算法

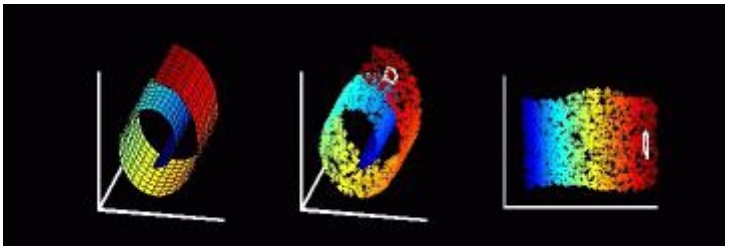
虽然 Isomap 取得了很好的保拓扑效果，但它最大的缺点在于，当面对高维数据和大规模数据时，时间复杂度平方级增长，在这种情况下它很难得到实际的应用，于是在此，有人巧妙地利用概率统计中的正态分布与 t 分布分别提出了效率更高的 SNE 算法和 t-SNE 算法。

事实上， t-SNE 算法是由 SNE 衍生出的一种算法。SNE 最早出现在 2002 年，它改变了 MDS 和 ISOMAP 中基于距离不变的思想，而是巧妙地以降维前后两个图之间内在的条件分布概率的 KL 散度为驱动的损失函数，进行优化和嵌入。只不过，SNE 将高维和低维中的样本分布都看作正态分布，而 t-SNE 则将样本分布当做 T 分布考虑，t-SNE 这样做的好处是利用了 T 分布的密度函数峰值区间短小，故可以让距离大的簇之间距离拉大，从而解决了 SNE 降维后拥挤问题。当然，这两者在功能上如右边降维后的效果图所展示的那样，它们更加的侧重于聚类，而非保持拓扑。



3.LLE 算法

除了 Isomap、SNE 和 t-SNE，还有一种发展了相对古老的线性降维的 LE 方法而成 LLE 算法，即所谓局部线性嵌入。LLE 基于的假设是数据在较小的局部是线性的——这当然是合理地，因为高维流形局部都同构于一小块欧式空间，用线性代数的语言说，即是某一个数据可以由它邻域中的几个样本来线性表示。比如我们有一个样本 x1,我们在它的原始高维邻域里用 K-近邻思想找到和它最近的三个样本 x2,x3,x4. 然后我们假设 x1 可以由 x2,x3,x4 线性表示，即存在 w1,w2,w3 使得 x1 可以写成 x1w1+x2w2+x3w3, 然后在平面上依葫芦画瓢，使得前后线性关系的权重系数 w1，w2，w3 尽量不变或者最



小改变的，这样就在考虑局部点的同时也由于较远的点对系数没有太大影响，不用去做过多的考虑，大大降低了降维的复杂程度和计算复杂度，上图为针对一个2000 样本点的 swiss-roll 数据集的 LLE 降维。

四．流形学习在生产生活与科研中的应用

例如在分析预测股票的行走趋势时，我们往往会采集上百个各种不同的特征，这对于时间序列预测的计算复杂度是一个很大的考验，这个时候往往就需要使用 PCA（主成分分析）去除掉一些不是很重要的特征，留下几十个关键特征降低时间复杂度。

事实上，不仅是股票预测等金融行业的应用，在其他涉及大规模数据采集和分析的科研领域，流形学习也有着广泛的应用，例如在细胞学习领域，研究蛋白质和 DNA 的关系，还有亲代与子代细胞间的遗传关系，往往就会采集上万维的向量，上万个细胞的数据，这个时候如何直观高效率的准确的实现降维就显得非常重要，让生物研究者，能一目了然的看出，这个数据与那个数据间有一些联系.例如，这个 DNA 与那个蛋白质有一定关系，亦或这个细胞是另一个细胞的子代细胞,从而更易于得出科学的结论。

【参考文献】

【1】 J. B. Tenenbaum, V. de Silva, J. C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science 290, (2000), 2319－2323.

【2】中科院计算所,《流形学习专题》

【3】浙江大学，何晓飞，《机器学习的几何观点》

【4】Mikhail Belkin and Partha Niyogi, Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering, Advances in Neural Information Processing Systems 14, 2001, p. 586－691, MIT Press

【5】Sam T. Roweis & Lawrence K. Saul, Locally Linear Embedding,

【6】S. T. Roweis and L. K. Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding, Science Vol 290, 22 December 2000, 2323－2326.