LLE 方法的分类与研究①

屈治礼

(江苏科技大学 计算机科学与工程学院, 镇江 212003)

摘 要: 对于低维数据的分类很常见, 但是对于高维数据的分类却不多, 主要是因为维度太高. 尤其对于分布不均匀的样本集, 传统的局部线性嵌入算法易受到近邻点个数的影响, 为了克服这一问题, 提出改进距离的局部线性嵌入算法. 通过实验表明, 改进距离的局部线性嵌入算法能使原来的样本集尽可能的分布均匀, 从而降低近邻点个数的取值对局部线性嵌入的影响, 在保证分类准确的前提下, 达到了有效缩短时间的目的.

关键词: 局部线性嵌入; 高维数据; 分类

Classification and Research of LLE Method

QU Zhi-Li

(School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang 212003, China)

Abstract: For classification of low-dimensional data is very common, but not for the classification of high-dimensional data, mainly because of too high dimension. In particular, for the uneven distribution of the sample set, the traditional locally linear embedding(LLE) algorithm is vulnerable to the impact of the number of nearest neighbor points, In order to overcome this problem, this paper improves locally linear embedding algorithm by changing the distance. Through the experiments indicates that the improved distance locally linear embedding algorithm can make the original sample set distribute evenly as far as possible, thereby reducing the influence of selection of the number of nearest neighbor points on locally linear embedding, on the premise of ensuring accurate classification, to achieve the purpose of effectively shorten the time.

Key words: locally linear embedding; high-dimensional data; classification

目前, 常用的降维算法[1-4]有很多, 例如主成分分析(Principal Component Analysis, 简称 PCA)、线性判别分析(Linear Discriminant Analysis, 简称 LDA)、多维尺度变换(Multi-Dimensional Scaling, 简称 MDS)、等距特征映射(Isomapetric Mapping, 简称 Isomap)、拉普拉斯特征映射(Laplacian Eigenmaps, 简称 LE)等, 还有相对较新的局部线性嵌入(Locally Linear Embedding, 简称 LLE)[5]. LLE 算法是针对非线性高维数据的一种降维方法, 处理后的低维数据能够保持原有的拓扑关系. 它广泛应用于文本分析、高维维数据的可视化以及生物信息学等众多领域中. LLE 算法具有几何意义直观、有解析的整体最优解且不需迭代、待定参数少、容易执行等很多优势, 因而它的研究和应用价值不言而喻.

在这里, 将作为非线性降维方法代表的 LLE 与常见的线性降维方法进行比较, 对比分类精度; 对于分布不均匀的样本集, 近邻点个数 K 的选取对 LLE 的分类结果影响较大, 采用改进距离的 LLE 算法, 可以有效的解决这一问题.

1 LLE原理

LLE 算法是 Roweis 和 Saul 于 2000 年在 Science 上提出的一种非线性降维方法, 基于使在高维空间中相邻的或相关的两个点映射到低维空间中也同样地相邻或相关的几何思想[6-9], 其核心是保存原流行中的局部几何特征, 以达到高维数据映射到低维全局坐标系中的目的, 主要是将流行上的近邻点映射到低维空间

① 收稿时间:2012-09-19;收到修改稿时间:2012-10-13

的近邻点. LLE 是一种依赖于局部线性的算法, 它认为在局部意义下, 数据的结构是线性的, 或者说局部意义下的点在一个超平面上.

在 LLE 算法中有两个参数需要设置, 近邻点的个数 k 和降维后输出维数 m, 即对观测数据集进行建模所需的最少独立变量的个数, 通常称之为最优嵌入维数, 也称为本征维数[5,10]. k 的选取在算法中是一个关键因素: 如果 k 的取值过大, LLE 就不能体现其局部特性, 这样会导致 LLE 算法趋向于 PCA 算法; 反之取值过小, LLE 就很难保证样本点在低维空间的拓扑结构. 另外, m 的选取也是一个重要因素, m 取值过大将会使降维结果中含有过多噪声; m 取值太小, 致使本来不同的点在低维空间可能会彼此交叠. 具体算法流程可以归结为以下三步:

Step1 计算出每个样本点的 k 个近邻点. 把相对于所求样本点的欧氏距离最近 k 个样本点规定为所求样本点的 k 个近邻点, k 是预先设定的值;

Step2 计算出样本点的局部重建权值矩阵. 这里定义一个成本函数, 其函数形式如下所示:

$$\min \varepsilon(W)=\sum_{i=1}^{N}\left|x_i-\sum_{j=1}^{k}w_j^i x_{ij}\right|^2 \tag{1}$$

其中, $x_{ij}(j=1,2,\cdots,k)$ 为 x_i 的 k 个邻近点, w_j^i 是 x_i 与 x_{ij} 之间的权值, 并且需要满足下式: $\sum_{j=1}^{k}w_j^i=1$. 在计算局部重建权值矩阵 W 的同时应该保证误差函数值取到最小值, 也就是说由样本点的近邻点, 构造出最优 W 矩阵使误差函数值达到最小;

Step3 将所有的样本点映射到低维空间中. 为了使输出数据在低维空间中保持原有的拓扑结构, 这里构造一个损失函数, 映射过程中必须使损失函数值达到最小. 该函数形式如下所示:

$$\min \varepsilon(W)=\sum_{i=1}^{N}\left|y_i-\sum_{j=1}^{k}w_j^i y_{ij}\right|^2 \tag{2}$$

其中, y_i 是 x_i 的输出向量, $y_{ij}(j=1,2,\cdots,k)$ 是 y_i 的 k 个近邻点, 并且要满足如下式(3)和式(4):

$$\sum_{i=1}^{N}y_i=0 \tag{3}$$

$$\frac{1}{N}\sum_{i=1}^{N}y_i y_i^T=I \tag{4}$$

其中, I 是 $m\times m$ 的单位矩阵, 然后求取的最优解 y_i 应使损失函数值达到最小. $w_j^i(j=1,2,\cdots,N)$ 可以存储在 $N\times N$ 的系数矩阵 W 中, 当 x_j 是 x_i 的近邻点时, $W_{i,j}=w_j^i$, 否则 $W_{i,j}=0$. 则损失函数可重定义为下式:

$$\min \varepsilon(Y)=\sum_{i=1}^{N}\sum_{j=1}^{N}M_{i,j}y_i^T y_j \tag{5}$$

其中, M 是 $N\times N$ 的对称矩阵, 其表达式为 $M=(1-W)^T(1-W)$, 由式(5)可知, 要使损失函数值达到最小, 则取 Y 为 M 的最小非零 m 个特征值所对应的特征向量. 在处理过程中, 将 M 的特征值降序排列, 最后一个特征值几乎接近于零, 那么舍去最后一个特征值. 通常取 $1\sim m$ 之间的特征值所对应的特征向量作为最终的输出结果.

以手旋杯原始数据集作为观测数据, 它是由一个视频序列在相等时间间隔中采样得到的, 部分图像如(1)所示. 采用 LLE 算法对 460 幅图像降到 3 维, 如下图(2)所示, 容易发现其本质为嵌入在三维空间中的一维曲线, 水平旋转度是其本征维数, 每个手旋杯的变化则是通过内在一维属性的插值和重构一维模型来完成的.

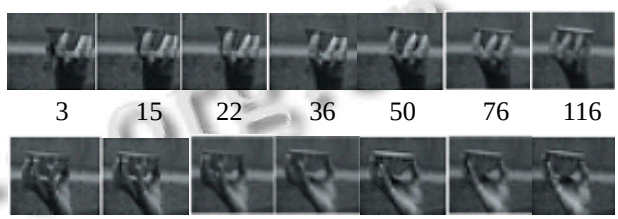

图 1 手旋杯部分图像示例

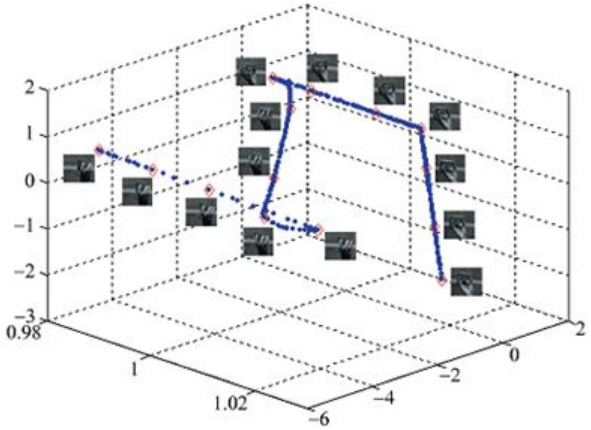

图 2 本征一维曲线

2　LLE与线性降维方法分类精度比较

在测试分类效果的方法上，这里我们采用支持向量机(Support Vector Machine，简称 SVM)，算法实现使用 LIBSVM. 以下的模型选择和参数设置如下：采用高斯核函数(Radial Basis Function，简称 RBF)，即：$T(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{g}\right)$，利用参数 g 和惩罚系数 C 共同寻优，并采用 10 折交叉确认方法，g 的搜索范围 $\{2^{-8},2^{-7},\cdots,2^{7},2^{8}\}$，$C$ 的搜索范围 $\{2^{-5},2^{-4},\cdots,2^{9},2^{10}\}$，如果是多分类的问题，则 LIBSVM 采用的是成对训练方法. 下面选用 Swiss roll 数据集的测试结果如下图所示：

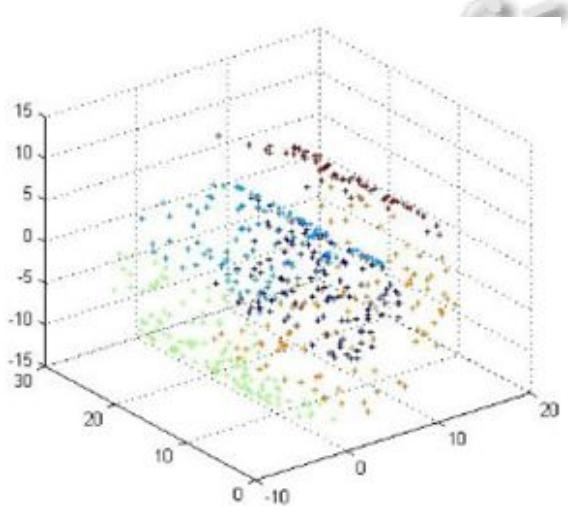

图 3　Swiss roll 数据集

这里 Swiss roll 数据集采集 500 个样本，维数为 3，分类数目为 5. 原始数据在 SVM 分类器上的精度是 99.6%，如下表所示：

表 1　分类精度比较

精度	PCA	LDA	LLE
1 维	66%	71.6%	26.8%
2 维	98.2%	99.8%	27.4%
3 维	99.6%	99.6%	-
…	…	…	…

由上表分析可知，LLE 降维处理之后 SVM 分类器的性能大幅下降，这应该是 LLE 方法本身的问题，可能是距离度量不适合，局部线性嵌入采用的是欧氏距离，或许采用测地线距离能提高分类效果；也有可能是邻域选择问题，尝试不同的 K 值或许可以改善分类器的分类效果.

3　改进距离度量的局部线性嵌入算法

由于分布不均匀的样本集，会导致近邻点个数 K 的选取对 LLE 的分类结果影响较大，所需分类时间也相应较长；改进的 LLE 的思想，就是在保证分类正确的前提下，减少近邻点个数 K 的选取，进而减少分类时间. 在这里，我们不妨引入 Conformal-Isomap[11,12] 中的一种度量距离的方法：

$$d_{ij}=\|x_i-x_j\|\Big/\sqrt{T(i)T(j)} \tag{6}$$

其中，$\|x_i-x_j\|$ 表示 x_i 和 x_j 之间的欧式距离，$T(i)$，$T(j)$ 分别表示 x_i 到它的 K 个近邻之间的距离的平均值和 x_j 到它的 K 个近邻之间的距离的平均值. 我们采用这个新的距离来寻找样本集中的每个样本 x_i 的 K 个近邻，然后按照经典的 LLE 算法计算权重和嵌入空间中的样本.

这种新的距离使得处于样本分布较密集区域的样本之间的距离增大，而使得处于样本分布较稀疏的区域的样本之间的距离缩小，这样的话使样本集的整体分布趋于均匀化，从而降低 K 的取值对 LLE 的实验结果的影响. 那么改进后的 LLE 算法可归纳如下：

步骤 1.　计算出每个样本点的 K 个近邻点，把相对于所求样本点的式(6)距离最近 K 个样本点规定为所求样本点的 K 个近邻点；

步骤 2.　计算出该样本点的局部重建权值矩阵；

步骤 3.　由该样本点的局部重建权值矩阵和其近邻点计算出该样本点的输出值.

4　实验应用与分析

根据上述对 LLE 算法中的距离改进，现将其中的关键部分程序列出如下：

```
//对矩阵中的每列进行求和运算
Y = sum(X.^2, 1);
//计算空间中的任两点距离
distance_graph= repmat(Y, N, 1)+repmat(Y, 1, N)-2*X'*X;
//对计算的距离进行排序
[sorted_ distance, index] = sort(distance_graph);
//查找 K 个近邻点
temp = sorted_ distance (2:(1+K), ;);
vec = sum(temp, 1);
vec_mat = vec*vec';
```

```
vec_mat = vec_mat';
vec_mat = sqrtm(vec_mat);
//对改进的距离进行计算
distance_graph = sorted./vec_mat;
[sorted_ distance, index] = sort(distance_graph);
//求出最后的 K 个近邻点
neighborhoods = index(2:(1+K), ;);
```

以手写体数字图像分类分析为例, 采用 USPS 数据集[13], 数字("0"到"9")的图像经预处理后变成分辨率16×16 (D=256), 并把灰度值量化为 256 阶. 预处理后的图像作为改进的 LLE 的输入数据, 经改进的 LLE 降维后在前两维坐标中显示如图 4.

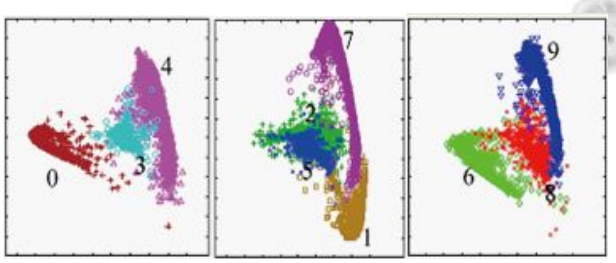

图 4 改进的 LLE 将手写体数字图像降到二维的效果

实验过程中采用 6000 个样本作为训练数据, 2000 个样本作为测试数据, 在改变 K 的情况下, 比较改进 LLE 算法与传统 LLE 算法的分类所需时间. 以不同的近邻数 K 为基础, 对比效果如下表所示:

表 2 传统 LLE 与改进 LLE 分类时间对比

分类所需时间	传统 LLE	改进 LLE
近邻数 K=60	4000s	1600s
近邻数 K=65	4400s	1650s
近邻数 K=70	5000s	1700s
近邻数 K=80	7600s	2000s
近邻数 K=85	8400s	2200s
近邻数 K=90	9600s	2600s
近邻数 K=100	11160s	2800s
…	…	…

另外, 改进距离的 LLE 算法在 K=50 时可以进行运算, 并且保持大约 96%的准确率; 而传统的 LLE 算法在此情况下却无法进行运算. 在保证准确率的前提下, 可以适当通过减小 K 来降低分类时间, 这样有利于进一步接近实时性的需求.

5 结束语与展望

"局部线性, 全局非线性"是 LLE 最为显著的特点, 而"全局非线性"是 LLE 之所以被看作是非线性降维方法的最主要原因. 改进的 LLE 使样本尽量的均匀化, 在一定程度上确实有效的缩短了分类时间, 有一定的实用性. LLE 假设中要求流行上的点分布均匀且稠密采样, 所学习的流行只能是非闭合的, 算法使用欧氏距离在高维空间中构造的局部邻域未必能够真实地反映流行的内在结构, 当两个卷曲状曲面间距离比较小时, 重构过程可能导致不同曲面的点进入同一个局部邻域, 造成流行结构在重构过程的扭曲[5]. 利用某些线性降维方法可剔除一些没用的噪声[14], 但是对于维数越大的数据, 它所需的开销就越大, 花费的时间也就越长. 基于此, 下一步将研究线性降维方法和非线性降维方法的结合以达到效率的提高和一定程度上消除噪声的目的.

参考文献

1 Hotelling H. Analysis of a complex statistical variable into principal components. Journal of Educational Psychology, 1933,24:417–441.

2 Tenenbaum JB, Silva V, Landford JC. A global geometric framework for nonlinear dimensionality reduction. Science, 2000,290(22):2319–2323.

3 Scholkopf B, Smola AJ, Muller KR. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998,10(5):1299–1319.

4 Zhang ZY, Zha HY. Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. SIAM Journal of Scientific Computing, 2004,26(1):313–338.

5 Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding.Science,2000,290(5500):2323–2326.

6 Roptevao K,Okuno P. Incremental locally linear embedding. Pattern Recognition,2005,38(10):1764 –1767.

7 罗四维,赵连伟.基于谱图理论的流行学习算法.计算机研究与发展,2006,43(7):1173–1179.

8 Saul L, Roweis S. Think Globally, Fit locally: unsupervised learning of low dimensional manifolds. Journal of Machine Learning Reserrch, 2003,2:119–155.

9 He X, Niyogip. Locality preserving projections.Address in Neural information Processing Systems. Cambridge: MIT

(下转第 50 页)

4 实验结果与分析

野外文物环境恶劣、地形复杂，为了测试监测系统的连通性与准确性，在实验室的外部模拟一个保护区域进行测试. 将传感器节点分布在模拟的保护区域内，ZigBee 网关和监控计算机放置在实验室，传感器节点与 ZigBee 网关之间相隔在 50m 左右. 启动监测系统后，网络成功地将采集的数据传输到监控平台上，数据采集软件界面如图 10 所示. 从监测的数据可知，系统在实验环境下数据传输可靠、准确.

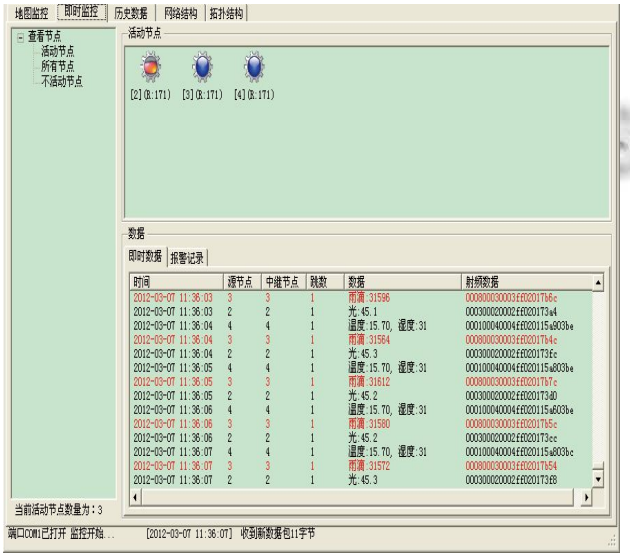

图 10 信息采集界面

5 结论

本文设计的基于 Z-Stack 协议栈的环境监测系统，能支持多点多区域监测，监测区域的划分取决于野外文物分布的地理位置，并且节点只需部署一次就可以进行长期的监测工作，功耗低、实时性高. 在使用的过程中，还可以根据实际需求扩展传感器模块用来监测野外文物所处环境的其他参数(如大气二氧化碳的浓度等)，或者可利用外部能量(如太阳能、风能等)来增加网络的使用寿命，实现智能无线监测系统的能源自给、低功耗和自组网.

参考文献

1 孙利民,李建中,陈渝,朱红松.无线传感器网络.北京:清华大学出版社,2005.4–14.

2 何文德,杨凤年,刘光灿.无线传感器网络在文物保护中的应用.计算机技术与自动化系统,2007,26(2):99–03.

3 DatesheetCC2530.Texas Instruments Inc.2009-2011.

4 DatesheetSHT1x.5thed,http://www.Seneirion.com,2011.

5 李兴法,尹冠飞.数字式加速度传感器 ADXL345 的原理及应用.黑龙江科技信息,2010,36:2–24.

6 张永梅,杨冲,马礼,王凯峰.一种低功耗的无线传感器网络节点设计方法.计算机工程,2012,38(3):71–73.

7 狄飞,张莉君.基于ZigBee无线传感器网络的森林环境监测系统.福建农林大学学报(自然科学版),2011,40(4):435–438.

8 匡兴红,邵惠鹤.无线传感器网络网关的研究.计算机工程,2007,339(6):228–230.

9 欧杰锋,刘兴华.基于 CDMA 模块的无线传感器网络网关的实现.计算机工程,2007,33(1):115–124.

10 章伟聪,俞新武,李忠成.基于 CC2530 及 ZigBee 协议栈设计无线传感器节点.计算机系统应用,2011,20(7):184–187.

11 李琳,高军伟.基于 ZigBee2006 协议栈的分布式温度采集系统的设计.青岛大学学报(工程技术版),2011,26(3):37–40.

12 吴建华,罗鑫,苏瑾.基于 GIS 的公安视频监控指挥管理系统.测绘通报,2011,11:67–70.

(上接第 17 页)

Press,2003:291–299.

10 Verleyse M. Learning high-dimension data. In: Ablameyko S, et al. eds. Limitations and Future Trends in Neural Computation. Amsterdam, The Netherlands: IOS Press, 2003: 141–162.

11 Saul LK, Roweis ST. An Introduction to Locally Linear Embedding(draft version). Pattern Recognition, 2001.

12 de Silva V, Tenenbaum JB. Global Versus Local Methods in Nonlinear Dimensionality Reduction 2002.

13 Hull JJ. A database for handwritten text recognition research. IEEE Trans. on PAMI,1994,16(5):550–554.

14 尹峻松,肖建,周宗潭.非线性流行学习方法的分析与应用.自然科学进展,2007,17(8):1015–1023.