

一种基于局部结构保持的数据降维方法

张 琳

(遵义师范学院 计算机与信息科学学院, 贵州 遵义 563002)

摘 要: 针对传统线性回归方法进行数据降维, 不能很好地利用数据间几何结构的问题, 提出了一种基于局部结构保持的数据降维方法. 利用非负约束下的最小二乘构造关系图, 从而描述局部近邻几何信息. 在将全局结构信息与局部结构信息整合时, 设计了一种新的模型选择方法用于模型参数估计, 极大地减少了计算量. 在公测库 ORL、YaleB、USPS、20Newgroup、Isolet 上的实验结果表明, 该方法用于数据分类时优于 ULDA、OLDA 和 NPE 等方法.

关键词: 数据降维; 几何结构; 回归模型

中图分类号: TP391.41

文献标识码: A

文章编号: 1000-7180(2016)05-0076-04

A Local Structure Preserving Based Data Dimensionality Reduction Approach

ZHANG Lin

(Faculty of Computer and Information Science, Zunyi Normal College, Guizhou 563002, China)

Abstract: The traditional linear regression method used for data dimensionality reduction failed to capture the geometry structure of data. In order to solve this problem, we present a local structure preserving based data dimensionality reduction. A nonnegative least squares method is introduced to construct the graph, which describe the local neighborhood geometry structure information. For the integrating of global structural information and local structure information integration, we also introduce a new model selection method for model parameter estimation, greatly reducing the cost of computation. Experiments on five benchmark datasets (ORL, YaleB, USPS, 20Newgroup and Isolet) show that the proposed approach can achieve better results than ULDA, OLDALDA and NPE.

Key words: data dimensionality reduction; geometric structure; regression model

1 引言

在机器学习、图像识别、机器视觉领域, 数据降维一直是研究的热点, 领域中提出了许多用于降维的技术, 其中 PCA^[1] 和 LDA^[2] 是两种最为流行和常用的方法. PCA 通过寻找一组新的正交基作为投影轴, 这组投影轴能够很好的反映数据的分布趋势. LDA 是监督模式投影轴选择方式, 而 PCA 是非监督模式的. LDA 通过寻找最大化类间距离, 最小化类内距离的方式寻找最优的一组投影正交基.

研究表明, 许多高维数据是嵌入在一个低维流形中的, 数据本身所具有的内在本质维度往往要低很多. 因此, 出现了基于几何驱动的数据降维方法,

通过寻找数据在低维空间下的几何结构来达到降维的目的. 目前已提出了多种在低维流形上寻找几何结构的方法, 包括线性的和非线性的. 例如 Iso-map^[3], 拉普拉斯特征映射^[4] 和局部线性嵌入 (locally linear embedding, LLE)^[5] 是基于非线性降维还原数据的局部几何结构方法. 局部保持投影 (Locality preserving projections, LPP)^[6] 和最近邻局部保持嵌入 (neighborhood preserving embedding, NPE)^[7] 是线性流形学习方法.

大量的研究工作已经就线性回归 (LR) 和众多子空间学习算法 (如 LDA, LPP, NPE) 之间是否存在某种关系展开了讨论. 在文献^[8] 中通过一个特殊的标识矩阵建立了 LDA 与 LR 之间的等价关系. 文

献[9]则提出了另一种基于谱图回归的技术来建立线性回归与子空间学习之间的等价关系。

鉴于上述讨论,我们提出了一种近邻结构保持的岭回归降维技术,在降维过程中整合了全局和近邻结构,通过非负最小二乘来构建图的拉普拉斯矩阵,用于获得局部结构信息。同时设计了一种有效的模式选择算法来估算方法中的最优调和参数。通过分析表明,提出的方法计算成本在一个低维矩阵上即可完成。

2 线性回归和岭回归

首先,简单介绍线性回归(Linear Regression, LR)和岭回归(Ridge Regression, RR)。给定数据集 $X = (x_1, x_2, \dots, x_n)$, 其中 $\{x_i\}_{i=1}^n \in \mathbb{R}^p$, 因变量 $\{y_i\}_{i=1}^n \in \mathbb{R}$, 一般通过最小化如下代价函数来寻找 x^i 和 y^i 之间的线性关系,

$$\sum_i (y_i - w^T x_i)^2 \quad (1)$$

当样本数量 n 比特征向量维度 p 小时,求解会出现病态问题。因此,通常会在 w 上引入一个惩罚项,变为求解如下优化问题,

$$J(w) = \sum_i (y_i - w^T x_i)^2 + \lambda \|w\|_2^2 \quad (2)$$

式中, λ 为正的调和参数,用于平衡估计量的偏差。式(2)通常被称作岭回归,岭回归是一种专用于共线性数据分析的有偏估计回归方法,实质上是一种改良的最小二乘估计法,通过放弃最小二乘法的无偏性,以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法。最优解可计算为 $w = (XX^T + \lambda I)^{-1} Xy$ 。经典RR通常使用单变量标签,为了将RR应用于多分类任务,有必要,需要将RR扩展到多元标签的情况。首先,构建一个 $n \times c$ 的标识矩阵 Y , 其中当 x_j 标记为第 i 类时 $y_{ij} = 1$, 其他情况下 $y_{ij} = 0$ 。针对多元标签情况,计算如下优化式

$$J(W) = \sum_i \|Y_i - W^T x_i\|^2 + \lambda \|W\|_F^2 \quad (3)$$

式中, Y_i 是标识矩阵 Y 的行向量,最优解为

$$W^* = (XX^T + \lambda I)^{-1} XY \quad (4)$$

3 局部近邻保持岭回归

3.1 非负约束下的图构造方法

给定一个包含有 n 个样本的数据集 $\{x_i\}_{i=1}^n$, 其中 $x_i \in \mathbb{R}^p$ 。希望每个数据点和它的邻居在流形的一个局部线性面片中。这些面片的局部几何结构可

以通过一组线性系数和邻域点来重构每一个点的方式来表达。重构误差由下面的代价函数来度量:

$$\epsilon(S) = \sum_i \|x_i - \sum_{j \in N_k(x_i)} S_{ij} x_j\| \quad (5)$$

式中, $N_k(x_i)$ 表示 x_i 的近邻个数, S_{ij} 表示 x_j 重构 x_i 的贡献度。如果 x_j 不是 x_i 的近邻点, 则 $S_{ij} = 0$ 。

在许多数据分析任务中,数据点只有非负项。这样由非负系数的表示的线性组合,即具有合理性,也有助于结果的解释。然而,大多数现有的数据依赖图的构造过程并未考虑到非负约束。因此,本文提出了一种非负最小二乘图构造方法。

给定训练数据集 $X = (x_1, x_2, \dots, x_n)$, 数据分为 c 类 $X = (X^{(1)}, X^{(2)}, \dots, X^{(c)})$, 其中 $X^{(j)} =$

$(x_1^{(j)}, x_2^{(j)}, \dots, x_{n_j}^{(j)}) \in \mathbb{R}^{p \times n_j}$ ($n_1 + n_2 + \dots + n_c = n$) 表示第 j 个数据矩阵。第 j 类的任何数据 x_i^j 可以由排除了 $x_i^{(j)}$ 的其他 j 类的数据点的非负线性组合来近似表示。可通过最小化下面的目标函数求解

$$\arg \min_{S_{ij}, j \neq i} \|x_i - \sum_{j \in C_k} S_{ij} x_j\|^2, k = 1, 2, \dots, c \quad (6)$$

$$S_{ij} \geq 0$$

式中, S_{ij} 表示 x_j 重构 x_i 时的贡献度。如果 x_j 不属于 C_k 类, 则 $S_{ij} = 0$ 。

式(6)是典型的非负最小二乘求解问题。许多方法可以求解该问题,大致可分为迭代法、有效集法、其他类方法三种。本文采用有效集法来求解,通过 matlab 中的 lsqnonneg 来求解式(6)。

令 $z_i = W^T x_i$ 为数据点 x_i 在低维空间中的近似描述。 x_i 和 x_j 在原始高维空间中的关系可表示为 $x_i \approx \sum_j S_{ij} x_j$, 为了实现在低维空间这种关系仍然能在 z_i 和 z_j 之间保持,即 $z_i \approx \sum_j S_{ij} z_j$, 我们构造目标函数如下,通过最小化式(7)来实现:

$$\|z_i - \sum_j S_{ij} z_j\|^2 \quad (7)$$

式(7)等价于 $\text{Tr}(Z^T L Z)$, 其中 $L = (I - S)^T (I - S)$, $Z = (z_1, z_2, \dots, z_n)$ 。

3.2 近邻保持岭回归

传统RR没有纳入近邻结构信息,因此无法捕捉到固有的几何结构信息。针对该问题,本文提出了一个邻域保持岭回归方法(Neighborhood Preserving Ridge Regression, NPRR)。该方法可以表示为以下优化问题:

$$W^* = \arg \min_W \|Y - W^T X\|_F^2 + (1 - \lambda) \text{Tr}(W^T W) + \lambda \text{Tr}(W^T X L X^T W) \quad (8)$$

式中, $\|A\|_F$ 表示矩阵 A 的 F 范数, $Tr(A)$ 是矩阵 A 的迹. 右侧优化式对 W 求导, 得到

$$(XX^T + (1 - \lambda)I + XLX^T)W = XY \tag{9}$$

最优解可以容易的计算得到, 如下

$$W^* = [XX^T + \lambda XLX^T + (1 - \lambda)I]^{-1}XY \tag{10}$$

3.3 模型选择

调和参数 λ 在模型中具有重要的作用, 用于权衡全局结构与局部结构的平衡. 参数 λ 通常是在大量的数据集中通过交叉验证来选择. 然而, 当数据维数很高时, XX^T 和 XLX^T 是非常大的, 因此做交叉验证的计算成本是很高的. 为此本文设计了一种有效的模型选择方法用于估计参数 λ 的最优值.

假设 $X \in \mathbb{R}^{p \times n}$, 式(10) 可写为

$$W^* = [XX^T + \lambda XLX^T + (1 - \lambda)I]^{-1}XY$$
$$= \frac{1}{1 - \lambda} \left[\frac{1}{1 - \lambda} XX^T + \frac{\lambda}{1 - \lambda} XLX^T + I \right]^{-1}XY$$

令 $\bar{\lambda} = \lambda / (1 - \lambda)$ 并且 $D_{1-\lambda}$ 是一个对角线元素为 $1 / (1 - \lambda)$ 的对角矩阵, 由 $(I + AB)^{-1} = I - A(I + BA)^{-1}B$ 得到

$$W^* = \frac{1}{1 - \lambda} [X(D_{1-\lambda} + \bar{\lambda}L)X^T + I]^{-1}XY$$
$$= \frac{1}{1 - \lambda} (I - X(D_{1-\lambda} + \bar{\lambda}L)^{-1}X^T)XY \tag{11}$$

式中, I 是适当大小的单位矩阵. 当特征向量的维数 p 远大于样本个数 n 时, 矩阵 $I + X^T X (D_{1-\lambda} + \bar{\lambda}L) \in \mathbb{R}^{n \times n}$ 比矩阵 $X^T X + \lambda XLX^T + (1 - \lambda)I \in \mathbb{R}^{p \times p}$ 要小得多, 从而显著的减少了计算成本和内存. 当样本数量比特征数量大很多时, 原计算式也可以执行更加有效的计算. 综上分析, 本文提出的方法只在一个较小的矩阵上进行计算, 因此极大地减少了计算成本.

4 实验结果

实验部分, 本文在 5 个公测数据集上进行分类准确率实验从而评估提出的 NPRR 方法. 模型参数通过交叉验证获得, 然后用最近邻规则去分类测试数据.

4.1 数据集

实验中采用了多种类型的数据集, 如见表 1 所示, 包括数字图像(ORL, YaleB, USPS)、文本和语音等. 按固定大小将数据分为训练集和测试集.

表 1 实验中采用的数据集

数据集	样本量	特征维数	训练集	测试集	类别数
ORL	400	1 024	160	240	40
YaleB	2 414	1 024	1 140	1 274	38
USPS	3 000	256	750	2 250	10
20Newgroup	1 200	8 298	2 40	960	4
Isolet	1 560	617	260	1 300	26

4.2 分类结果

实验中比较了本文提出的能够保持全局和局部结构的 NPRR 方法与只保持全局或者局部的方法, 如 ULDA^[10]、OLDA^[11]、NPE^[7] 的分类结果. 对于每一个数据集, 随机将数据分为训练集和测试集(具体大小见表 1). 表 2 列出了分类效果的对比情况, 为了给出合理的结果, 进行了 30 次的分组实验, 表中给出了分类准确率的均值和标准差.

表 2 NPRR, ULDA, OLDA 和 NPE 分类准确率比较 %

数据	NPRR		ULDA		OLDA		NPE	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差
ORL	93.36	1.73	91.24	1.77	94.40	1.53	88.05	2.04
YaleB	85.33	1.36	57.72	1.25	83.62	1.82	64.50	1.03
USPS	88.90	0.70	85.88	0.76	86.25	0.92	88.78	0.85
20Newgroup	77.77	2.04	77.39	2.04	76.14	2.20	69.98	3.50
Isolet	88.77	0.97	79.62	1.05	88.76	1.01	44.98	4.33

从表 2 可以看出, 本文提出的 NPRR 方法除了在 ORL 数据集略低于 OLDA 方法外, 在其他数据集中表现均优于其他方法. 从实验结果可以看出, 全局和邻域结构的重要性取决于具体应用, 例如在 USPS 库上 NPE 比 ULDA 表现要好, 局部近邻结构在分类中起到了更加重要的作用, 在 ORL、YaleB、20Newsgroups、Isolet 库上, 全局和局部近邻结构彼此互补. 因此, 在 NPRR 中整合全局和局部近邻结构是非常有意义的. 这样可以解释为什么 NPRR 方法在五个公测库上的表现优于 ULDA 和 NPE 方法.

另外, 在 ORL 库上的实验 OLDA 方法优于 NPRR, 一部分原因是因为 OLDA 方法中的辨别特征向量是正交的, 对数据噪音有一定的鲁棒性. OLDA 方法可以捕捉到更加有用的全局结构信息. 这表明在特定应用如 ORL 这类数据中, 全局结构包含更多有用的信息, 加入局部结构反而会降低分类准确率. 在 20Newsgroup 库上, NPRR 和 OLDA 方法表现相当, 表明在有些类数据上, 是否加入局部结构对结果影像不大.

5 结束语

本文提出了一种局部近邻结构保持的岭回归降维方法. 利用非负最小二乘法来构造图结构, 非负约束条件可以更加合理的解释模型结果, 近邻结构是通过图形的正则化来保持. 我们还设计了一种有效的模型选择方法用于最优参数估计, 使得计算仅需要在一个较小的矩阵上即可完成. 在几类公测库上的实验结果表明, 提出的方法在分类精度方面相比于同类方法有所提高.

参考文献:

- [1] Jolliffe I. Principal component analysis[M]. New York: John Wiley & Sons, 2002.
- [2] Fukunaga K. Introduction to statistical pattern recognition[M]. New York: Academic press, 2013.
- [3] Tenenbaum J B, De Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.
- [4] Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering[C] // Proceedings of NIPS01. Vancouver, 2001: 585-591.
- [5] Roweis S T, Saul L K. Nonlinear dimensionality re-

duction by locally linear embedding[J]. Science, 2000, 290(5500): 2323-2326.

- [6] Niyogi X. Locality preserving projections[C] // Proceedings of NIPS03. Vancouver, 2003: 153-160.
- [7] He X, Cai D, Yan S, et al. Neighborhood preserving embedding[C] // Proceedings of ICCV'05. Beijing, 2005: 1208-1213.
- [8] Ye J. Least squares linear discriminant analysis[C] // Proceedings of ICML07. Corvallis, 2007: 1087-1093.
- [9] Cai D, He X, Han J. Spectral regression: A unified approach for sparse subspace learning[C] // Proceedings of ICDM07, Omaha, 2007: 73-82.
- [10] Jin Z, Yang J Y, Hu Z S, et al. Face recognition based on the uncorrelated discriminant transformation [J]. Pattern recognition, 2001, 34(7): 1405-1416.
- [11] Ye J, Xiong T. Computational and theoretical analysis of null space and orthogonal linear discriminant analysis[J]. The Journal of Machine Learning Research, 2006(7): 1183-1204.

作者简介:

张琳女, (1985-), 硕士, 讲师. 研究方向为数据挖掘技术、计算机教育. E-mail: zyzhanglin@163.com

(上接第 75 页)

该方法能校准 Nyquist 频率内的所有输入信号. 本文通过 MATLAB 搭建一个 8 好 Bit 的五通道模型仿真结果显示校准效果非常明显. 且该校准算法可以应用于任意通道数 TIADC 的通道间误差校准.

参考文献:

- [1] 吴光林. 多通道时间交叉 ADC 校准技术研究及实现 [D]. 南京: 东南大学, 2006.
- [1] Black W C, Hodges D A. Time interleaved converter arrays[J]. IEEE Journal of Solid-State Circuits, 1980, 15(6): 1022-1029.
- [3] 叶凡. 多通道时间交织模数转换器的校正与集成电路实现方法研究[D]. 上海: 复旦大学, 2010.
- [4] Lee Z M, Wang C Y, Wu T J. A CMOS 15 bit 125 MS/s time-interleaved ADC with digital background calibration [J]. IEEE Journal of Solid-State Circuits, 2007, 42(10): 2149-2160.
- [5] Hegong Wei, Peng Zhang, Datta Sahoo, et al. An 8 Bit 4 GS/s 120 mW CMOS ADC [J]. IEEE Journal of Sol-

id-State Circuits, 2014, 49(8): 1-4.

- [6] Jamal S M, Fu D H, Hurst P J, et al. A 10 b 120 Msample/s time-interleaved analog-to-digital converter with digital background calibration[J]. IEEE Journal of Solid-State Circuits, 2002, 37(12): 1618-1627.
- [7] 王韩. 6 位 2GSps 时间交织 ADC 设计[D]. 哈尔滨: 哈尔滨工业大学, 2007
- [8] Wang ChungYi, Wu JiehTsorng. A multiphase timing-skew calibration technique using Zero-Crossing Detection[J]. IEEE Journal of Solid-State Circuits, 2009, 56(6): 1102-1114.

作者简介:

王晓蕾女, (1978-), 博士, 副研究员, 硕士生导师. 研究为填混合集成电路设计.

王玉莹女, (1989-), 硕士研究生. 研究方向为混合集成电路设计.

陈红梅(通讯作者)女, (1986-), 博士, 讲师, 硕士生导师. 研究方向为高速数模混合集成电路设计.

E-mail: hmchen@hfut.edu.cn.