一种基于高斯隐变量模型的分类算法

王秀美 高新波 李 洁

(西安电子科技大学电子工程学院 西安 710071)

摘 要 高斯过程隐变量模型是近年来新兴的无监督降维方法,它可以找到高维数据的低维流形结构.但是由于高斯过程隐变量模型是无监督的概率降维方法,所以当数据集中的样本有类别标记信息时,高斯过程隐变量模型不能利用这些监督信息,实现分类的任务.为了使高斯过程隐变量模型可以处理分类任务,文中提出了一种监督的高斯过程隐变量模型分类模型.通过最大化后验似然的方法确定观测数据在隐空间的坐标,同时可以完成分类任务.实验结果证明了该模型可以有效地用于分类.

关键词 高斯过程;监督学习;降维;线性判别分析

中图法分类号 TP181

DOI号: 10.3724/SP.J.1016.2012.02661

A New Classification Algorithm Based on Gaussian Process Latent Variable Model

WANG Xiu-Mei GAO Xin-Bo LI Jie

(School of Electronic Engineering, Xidian University, Xi'an 710071)

Abstract Gaussian process latent variable model is a new probabilistic approach for dimensionality reduction. It can obtain a low-dimensional manifold of a data set in an entirely unsupervised way. However, when there is some supervised information in the data set, Gaussian process latent variable model cannot use this information for supervised tasks, e. g., classification and regression learning. For this purpose, a supervised Gaussian process latent variable model for classification is developed. The maximum-a-posterior algorithm is employed to estimate all latent variables position. Compared with the traditional Gaussian process latent variable model, the supervised version of this model shows more advantages in experiments.

Keywords Gaussian process; supervised learning; dimensionality reduction; linear discriminant analysis

1 引 言

“维数灾难”是模式识别和机器学习的常见问题,在人脸识别、动态跟踪、语音识别等实际问题中,高维数据之间往往存在着大量的冗余,如何消除冗

余,寻找数据间的内在联系,从而减少数据处理过程中的计算量,是当前降维技术研究的重点.

经典的降维方法多为线性方法,即寻找高维数据的某个线性子空间,将高维数据投影到这个子空间中,再进行识别或分析,大大降低计算量.主成分分析(Principle Component Analysis, PCA)^[1-2]和线

性判别分析(Linear Discriminant Analysis, LDA)^[3]是最具代表性的两种线性子空间分析方法. PCA 是一种无监督的降维方法,它的主要思想是用较少的综合变量来代替原来较多的变量,同时要求这几个综合变量互不相关,并尽可能多地表示原来数据的能量.它是将多指标化为少数几个综合指标的一种统计分析方法;LDA 是目前常用的有监督的线性降维方法,它利用已知的监督信息进行数据降维,降维过程最大化样本类间散度,同时最小化类内散度.由 PCA 衍生并推广出的概率主成分分析(Probabilistic PCA, PPCA)^[4]是一种从高斯隐变量模型扩展出来的概率方法,通过对高维数据的联合似然最大化,求得投影方向.由于 PCA、PPCA 和 LDA 都是线性降维方法,虽然易于实现,但是对一些结构复杂的数据却无能为力,这是由其本身的线性特性造成的.

核主成分分析(Kernel PCA, KPCA)^[5]和核判别分析(Kernel LDA, KDA)^[6-7]分别是对 PCA 和 LDA 的非线性推广.由于核方法是非线性,多数是不可逆的,即只给出了如何从观测空间到低维隐空间的映射,但对于如何建立从低维到高维数据的映射,没有给出解决方法.

高斯过程隐变量模型(GP-LVM)是一种非线性降维技术^[8-9],是 PPCA 的对偶形式. GP-LVM 克服了线性降维方法的局限,并且利用高斯过程^[10]建立了从低维隐空间到高维空间的映射关系.通过最大化观测数据的联合密度,优化出高维数据在隐变量空间中的坐标位置.然而,GP-LVM 所用到的高斯过程映射是在样本的每一维上单独进行,是无监督的,当数据中有监督信息,如类别标记可以利用时,GP-LVM 不能有效地利用类标,根据样本的类别结构信息,从而建立有效的分类模型.

为了利用观测数据的监督信息,文献[11]中提出了一种利用类标信息的方法,用 LDA 的能量函数作为隐空间变量的先验知识,最大化隐变量的后验知识,得到降维后的样本.此方法虽然可以利用监督信息,使得降维后样本在低维空间保持样本类间散度大、类内散度小的性质,但是它没有从根本上解决如何利用类别信息进行分类的问题.本文提出了一种监督的 GP-LVM (SGP-LVM) 分类模型,该模型基于隐变量模型(Latent variable models)建立,利用了隐变量空间条件独立(Conditional independent)的性质^[12].与文献[11]提出的方法不同,SGP-LVM

既考虑观测样本的性质,也考虑已知的类标信息,分别建立了从隐空间到观测样本和类别信息集合的映射,前者达到了降维的目的,后者则实现了分类的任务.在本文的后半部分,在各种数据上的实验结果证明了本文方法的有效性.

本文第 2 节简单介绍 PPCA 和 GP-LVM 的相关工作;第 3 节给出如何利用隐变量过程的条件独立性建立监督的 GP-LVM 分类算法,包括如何进行分类和如何确定超参数和降维后隐变量的位置;第 4 节给出实验结果以证明分类算法的有效性;最后进行总结与讨论.

2 预备知识

2.1 概率主成分分析

PPCA 是 PCA 方法的概率扩展^[2],它建立了一种从低维空间(隐变量空间)到高维空间的线性映射,若 $\mathbf{X}=[\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N]$ 表示需要降维的高维数据, $\mathbf{Z}=[\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N]$ 表示降维后的低维数据,并假设噪声 $\boldsymbol{\eta}_n \in R^D$ 符合独立高斯分布, $\boldsymbol{\eta}_n \sim N(0, \beta^{-1} \mathbf{I})$, β^{-1} 为噪声方差.从低维空间到观测空间的映射可以表示为

$$\mathbf{x}_n = \mathbf{W}\mathbf{z}_n + \boldsymbol{\eta}_n \quad (1)$$

其中线性映射由 $\mathbf{W} \in R^{D \times d}$ 确定,则观测数据点的似然概率为

$$p(\mathbf{x}_n | \mathbf{z}_n, \mathbf{W}, \beta) = N(\mathbf{x}_n | \mathbf{W}\mathbf{z}_n, \beta^{-1} \mathbf{I}) \quad (2)$$

假设隐变量空间中的点是独立同分布的:

$$p(\mathbf{z}_n) = N(\mathbf{z}_n | 0, \mathbf{I}) \quad (3)$$

即通过对隐变量空间中的点积分可以推导出边缘似然,

$$\begin{aligned} P(\mathbf{x}_n | \mathbf{W}, \beta) &= \int p(\mathbf{x}_n | \mathbf{z}_n, \mathbf{W}, \beta^{-1} \mathbf{I}) p(\mathbf{z}_n) d\mathbf{z}_n \\ &= N(\mathbf{x}_n | \mathbf{W}\mathbf{W}^T + \beta^{-1} \mathbf{I}) \end{aligned} \quad (4)$$

和观测数据的联合似然,

$$P(\mathbf{X} | \mathbf{W}, \beta) = \prod_{n=1}^N p(\mathbf{x}_n | \mathbf{W}, \beta) \quad (5)$$

用最大化似然的方法得到投影矩阵 $\mathbf{W}$. 求似然最大化的方法很多,其中 EM^[13]算法是比较常用的一种.

EM 算法是求参数极大似然估计的一种简单实用的学习算法,对应于本文中提到的 PPCA,它包括两个步骤:

E-step. 计算数据的对数似然函数期望. 似然函数为

$$P(\mathbf{X}|\mathbf{W},\beta)=\prod_{n=1}^N p(\mathbf{x}_n|\mathbf{W},\beta) \quad (6)$$

若记对数似然函数为 $\ln p(\mathbf{X};\mathbf{W})$, 对数似然函数可以表示为

$$\ln p(\mathbf{X};\mathbf{W})=\sum_{n=1}^N \ln\{p(\mathbf{x}_n|\mathbf{W},\beta)\} \quad (7)$$

则 E-step 求 $\ln p(\mathbf{X};\mathbf{W})$ 的期望值 $E\{\ln p(\mathbf{X};\mathbf{W})\}$.

$$E\{\ln p(\mathbf{X};\mathbf{W})\}=-\sum_{n=1}^N \left\{ \frac{D}{2} \ln \beta + \frac{1}{2} \text{tr}(\langle \mathbf{z}_n \mathbf{z}_n^T \rangle) + \frac{1}{2\beta} (\mathbf{x}_n - \boldsymbol{\mu})^T (\mathbf{x}_n - \boldsymbol{\mu}) - \frac{1}{\beta} \langle \mathbf{z}_n \rangle^T \mathbf{W}^T (\mathbf{x}_n - \boldsymbol{\mu}) + \frac{1}{2\beta} \text{tr}(\mathbf{W}^T \mathbf{W} \langle \mathbf{z}_n \mathbf{z}_n^T \rangle) \right\} \quad (8)$$

在上式中 $\boldsymbol{\mu}$ 表示观测变量 $\mathbf{X}$ 的均值并且有

$$\langle \mathbf{z}_n \rangle = \mathbf{M}^{-1} \mathbf{W}^T (\mathbf{x}_n - \boldsymbol{\mu}) \quad (9)$$

$$\langle \mathbf{z}_n \mathbf{z}_n^T \rangle = \beta \mathbf{M}^{-1} + \langle \mathbf{z}_n \rangle \langle \mathbf{z}_n \rangle^T \quad (10)$$

其中 $\mathbf{M} = \mathbf{W}^T \mathbf{W} + \beta \mathbf{I}$.

M-step. 关于投影矩阵 $\mathbf{W}$ 最大化 $E\{\ln p(\mathbf{X};\mathbf{W})\}$, 即关于 $E\{\ln p(\mathbf{X};\mathbf{W})\}$ 对 $\mathbf{W}$ 求导, 得到最优的 $\mathbf{W}$ 值,

$$\tilde{\mathbf{W}} = \left\{ \sum_{n=1}^N (\mathbf{x}_n - \boldsymbol{\mu}) \langle \mathbf{z}_n \rangle^T \right\} \left\{ \sum_{n=1}^N \langle \mathbf{z}_n \mathbf{z}_n^T \rangle \right\}^{-1} \quad (11)$$

通过交替使用 E-step 和 M-step 直至收敛, 收敛性判断可以用相邻两次的迭代得到的 $E\{\ln p(\mathbf{X};\mathbf{W})\}$ 的差值判断. $\|E\{\ln p(\mathbf{X};\mathbf{W})\}_{t+1} - E\{\ln p(\mathbf{X};\mathbf{W})\}_t\| \leq \epsilon$, 则 $E\{\ln p(\mathbf{X};\mathbf{W})\}$ 达到一个极值点, 从而得到投影矩阵 $\mathbf{W}$. 详细步骤请参看文献[4].

2.2 高斯过程隐变量模型

条件独立是隐变量模型的重要性质, 即在隐变量给定的时候, 观测变量的各维之间是独立的^[12].

如图 1 所示, 给出隐变量 $\mathbf{z}$ 时, 观测变量 $\mathbf{x}$ 各维之间条件独立. GP-LVM 正是基于这个性质建立了从隐变量到观测变量的映射关系.

图 1 GP-LVM 中的条件独立性

GP-LVM 的模型是基于图 1 建立, 从所有的隐变量到观测变量的每一维建立映射, 给每个映射一个高斯过程先验, 则建立了一个非线性隐变量模型. 这与 PPCA 不同, PPCA 是假设各个样本独立同分布(i. i. d), 假设降维后的样本服从高斯分布而建立模型.

假设各维度上的映射 f_d 独立同分布, 服从高

斯过程,

$$p(f) = \prod_{d=1}^D p(f_d) = \prod_{d=1}^D N(f_d | 0, \mathbf{K}) \quad (12)$$

则观测数据 d 的似然可表示为

$$p(\mathbf{x}_{:,d}|\mathbf{Z},\beta) = \int p(\mathbf{x}_{:,d}|\mathbf{z}_n, f_d, \beta^{-1} \mathbf{I}) p(f_d) df_d = N(\mathbf{x}_{:,d} | 0, \mathbf{K}) \quad (13)$$

由于各维之间独立, 则观测数据的似然可以表示为数据各维度似然的乘积

$$P(\mathbf{X}|\mathbf{Z},\beta) = \prod_{d=1}^D N(\mathbf{x}_{:,d}|\mathbf{Z},\beta) = \prod_{d=1}^D \frac{1}{(2\pi)^{N/2} |\mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2} \mathbf{x}_{:,d}^T \mathbf{K}^{-1} \mathbf{x}_{:,d}\right) \quad (14)$$

上式可以表示为矩阵的形式:

$$P(\mathbf{X}|\mathbf{Z},\beta) = \frac{1}{(2\pi)^{N/2} |\mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{K}^{-1} \mathbf{X} \mathbf{X}^T)\right) \quad (15)$$

其中, $\mathbf{K}$ 是协方差函数矩阵, 或者称为核函数矩阵. 若用线性的核函数, 则可以定义为 $\mathbf{K} = \mathbf{X} \mathbf{X}^T + \beta^{-1} \mathbf{I}$. 若选用非线性核函数则可以定义为

$$k(\mathbf{z}_i, \mathbf{z}_j) = \theta_{\text{rbf}} \exp\left(-\frac{\gamma}{2} (\mathbf{z}_i - \mathbf{z}_j)^T (\mathbf{z}_i - \mathbf{z}_j)\right) + \theta_{\text{bias}} + \theta_{\text{white}} \delta_{ij} \quad (16)$$

$k(\mathbf{z}_i, \mathbf{z}_j)$ 为矩阵 $\mathbf{K}$ 的第 i 行第 j 列对应的元素; δ_{ij} 为 Kronecker delta 函数. 这时, 从隐空间到高维空间的映射是一个非线性映射的高斯过程, 而且对于观测数据, 每一维是独立的. 联合似然式(9)可进一步简化为

$$P(\mathbf{X}|\mathbf{Z},\beta) = \frac{1}{(2\pi)^{N/2} |\mathbf{K}|^{1/2}} \exp\left(-\frac{1}{2} \text{tr}(\mathbf{K}^{-1} \mathbf{X} \mathbf{X}^T)\right) \quad (17)$$

3 监督的 GP-LVM 分类算法

GP-LVM 算法可以有效地找到高维数据(观测数据)在低维空间中上的流形, 但是由于它是一种无监督的机器学习方法, 所以当我们的观测数据中有类标信息时, 它不能利用这些类标信息, 实现分类目标. 针对这个问题, 我们将监督信息加入 GP-LVM 算法中, 建立一种有监督的 GP-LVM 分类模型. 这种有监督的模型是建立在隐变量模型的条件独立性质之上的, 我们参考文献[14-15]中的思路建立分类模型.

下面, 我们将给出监督的 GP-LVM 模型, 我们

用 $\mathbf{L}=[\mathbf{l}_1, \dots, \mathbf{l}_N]^T$ 表示输出矩阵, 其中 $\mathbf{l}_i \in R^C$, 或者表示分类中观测数据的类标, 例如二值分类时, $\mathbf{l}_i \in \{+1, -1\}$. 我们用 $(\mathbf{X}, \mathbf{Y})$ 表示观测数据, 则可以建立以下映射关系:

$$\begin{cases} \mathbf{x} = f_1(\mathbf{z}, \Theta_1) + \epsilon_x \\ \mathbf{l} = f_2(\mathbf{z}, \Theta_2) + \epsilon_L \end{cases} \quad (18)$$

式(18)中, 函数 f_1 表示从隐空间(低维空间) $\mathbf{Z}$ 到高维空间 $\mathbf{X}$ 的映射关系. 若 f_1 为高斯过程, 则可以用 Θ_1 表示映射的超参数. 同样, 函数 f_2 表示从 $\mathbf{Z}$ 到 $\mathbf{L}$ 的映射, 映射的超参数用 Θ_2 表示. 模型中的噪声分别为 ϵ_x 和 ϵ_L , 假设为零均值的高斯分布, 即 $\epsilon_x \sim N(0, \sigma_x^2 \mathbf{I})$, $\epsilon_L \sim N(0, \sigma_L^2 \mathbf{I})$, 其中 σ_x^2 和 σ_L^2 为噪声方差.

上面建立了含噪声的概率模型, 接下来推导观测数据 $(\mathbf{X}, \mathbf{L})$ 的似然 $p(\mathbf{X}, \mathbf{L} | \mathbf{Z})$.

$$p(\mathbf{X}, \mathbf{L} | \mathbf{Z}) = \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{l}_n | \mathbf{z}_n) \quad (19)$$

由于在隐变量模型中, 观测数据 $(\mathbf{X}, \mathbf{Y})$ 的各维之间独立, 即 $p(\mathbf{x}, \mathbf{l} | \mathbf{z}) = p(\mathbf{x} | \mathbf{z}) p(\mathbf{l} | \mathbf{z})$, 所以可以得到观测数据的联合似然为

$$\begin{aligned} p(\mathbf{X}, \mathbf{L} | \mathbf{Z}) &= \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{l}_n | \mathbf{z}_n) \\ &= \prod_{n=1}^N p(\mathbf{x}_n | \mathbf{z}_n) p(\mathbf{l}_n | \mathbf{z}_n) \end{aligned} \quad (20)$$

此时, 我们利用最大似然原则就可以优化得到隐变量 $\mathbf{Z}$ 的值. 先对式(15)求对数可以得到

$$\mathbf{L} \triangleq \ln p(\mathbf{X}, \mathbf{L} | \mathbf{Z}) = \ln p(\mathbf{X} | \mathbf{Z}) + \ln p(\mathbf{L} | \mathbf{Z}) \quad (21)$$

可以看到等式第 1 项为

$$\ln p(\mathbf{X} | \mathbf{Z}) = -\frac{DN}{2} \ln 2\pi - \frac{D}{2} \ln |\mathbf{K}| - \frac{1}{2} \text{tr}(\mathbf{K}^{-1} \mathbf{X} \mathbf{X}^T) \quad (22)$$

即为模型 GP-LVM 中的似然, 是属于原来无监督的部分. 式(16)右边的第 2 项为监督项, 如果我们给映射 f_2 高斯过程先验, 则当 $\mathbf{l}_i \in R^C$ 时, 似然可以表示成

$$\ln p(\mathbf{L} | \mathbf{Z}) = -\frac{N}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \text{tr}(\mathbf{L}^T \Sigma^{-1} \mathbf{L}) \quad (23)$$

因此, 用尺度共轭梯度方法^[16]最大化观测数据的联合似然, 可以得到隐变量在低维空间中的位置和映射的参数. 首先用对数似然对隐变量 $\mathbf{Z}$ 求梯度,

$$\frac{\partial \mathbf{L}}{\partial \mathbf{Z}} = \frac{\partial \ln p(\mathbf{X} | \mathbf{Z})}{\partial \mathbf{K}} \cdot \frac{\partial \mathbf{K}}{\partial \mathbf{Z}} + \frac{\partial \ln p(\mathbf{L} | \mathbf{Z})}{\partial \Sigma} \cdot \frac{\partial \Sigma}{\partial \mathbf{Z}} \quad (24)$$

其中,

$$\frac{\partial \ln p(\mathbf{X} | \mathbf{Z})}{\partial \mathbf{K}} = \mathbf{K}^{-1} \mathbf{X} \mathbf{X}^T \mathbf{K}^{-1} - D \mathbf{K}^{-1} \quad (25)$$

$$\frac{\partial \ln p(\mathbf{L} | \mathbf{Z})}{\partial \Sigma} = \Sigma^{-1} \mathbf{L} \mathbf{L}^T \Sigma^{-1} - D \Sigma^{-1} \quad (26)$$

根据选用的核函数形式不同, $\frac{\partial \mathbf{K}}{\partial \mathbf{Z}}$ 和 $\frac{\partial \Sigma}{\partial \mathbf{Z}}$ 得到不

同的值. 例如若选用线性的核函数, 则 $\frac{\partial \mathbf{K}}{\partial \mathbf{Z}} = \frac{\partial \Sigma}{\partial \mathbf{Z}} = \mathbf{Z}$.

其次用对数似然对参数 θ 求梯度

$$\frac{\partial \mathbf{L}}{\partial \theta_1} = \frac{\partial \ln p(\mathbf{X} | \mathbf{Z})}{\partial \mathbf{K}} \cdot \frac{\partial \mathbf{K}}{\partial \theta_1} \quad (27)$$

$$\frac{\partial \mathbf{L}}{\partial \theta_2} = \frac{\partial \ln p(\mathbf{L} | \mathbf{Z})}{\partial \Sigma} \cdot \frac{\partial \Sigma}{\partial \theta_2} \quad (28)$$

对隐变量 $\mathbf{Z}$ 和参数 θ 交替优化直至收敛. 上面给出的是处理分类的情况, 其推导是在给出类标 $\mathbf{L}=[\mathbf{l}_1, \dots, \mathbf{l}_N]^T \in R^{N \times C}$ 的基础上得出的. 若给出的监督信息是回归的输出值, 则可以直接将回归值替代类标用于给出的模型, 该模型可以直接用于回归.

4 实验结果与分析

本文算法在很多数据上进行了实验, 以验证算法的有效性. 实验分为两部分: 第 1 部分验证了本算法作为分类器的有效性, 第 2 部分验证了本算法作为回归算法时的有效性. 第 1 部分又分为两组: 第 1 组给出了本算法对 USPS 手写数字数据降维的结果, 第 2 组实验给出了本算法在人脸数据上的识别结果.

4.1 数据降到 2D 空间的结果对比

实验选取了 USPS 数据库中的手写数字 (Handwritten Digits)^[8] 识别的“3”和“5”, 选取准则是这两个数字相对其它数字较难识别.

USPS 的手写数字数据是由 0~9 的 10 个数字组成, 即数据共分为 10 类, 其中每一类数据由大概 500 个样本组成, 每个样本的维数是 256. 实验是在数字“3”和“5”上进行, 对每类样本选取 400 个, 共有 800 个样本, 分别用 GP-LVM 算法和监督的 GP-LVM 算法降到 2 维, 图 2 给出两种方法降维的结果, 其中“3”的样本降维后用十字表示, “5”的样本降维后用圆圈表示.

图 2 表示我们将 USPS 数据用 GP-LVM 和 SGP-LVM 降至 2 维的结果, 来比较证明我们提出算法的有效性. 从左边的图 2(a) 看出, 图像的上方圆点较集中的地方有很多“+”也在其中, 这表明 GP-LVM 的结果不能有效地将两类数据分开, 两种

图 2 将 USPS 数据降为 2 维空间结果显示

类型存在交叠. 图 2(b)表示的是 SGP-LVM 降维的结果,从图中可以看出,该算法不仅可以将两类数据有效分开,而且类间距较大,各自集中的区域很少有不同类型的点出现,这表示当用 SGP-LVM 将该类数据降至 2 维时,可以有效地将两类数据分开.

4.2 SGP-LVM、GP-LVM 和 LDA 分类结果比较

在本小节中我们对 SGP-LVM、GP-LVM 和 LDA 的分类效果进行验证,在下面的实验中,我们选用 YaleB(The extended Yale Face Database B)和 ORL(Olivetti Research Laboratory)人脸数据^[15]. SGP-LVM 可以对数据直接分类,而对于 GP-LVM 和 LDA 这两种方法,我们先用它们降维,然后用最紧邻(1-NN)进行分类.对于 LDA,我们对数据先用 PCA 处理,再用 LDA+KNN 分类,这样做可以避免 LDA 的欠采样问题.

YaleB 数据包含 38 个人的 2414 幅人脸图像,每个人有 64 幅接近正面的不同光照下的人脸图像.每幅图片被裁减成 32×32 大小,即每个样本有 1024 维的特征来刻画.在实验中,对每个人随机选择 30 幅作为训练图像,余下的 38 幅图用作测试. ORL 数据库是由 40 个人的人脸图像组成,每个人有 10 幅图像,分别在不同时刻、光照、表情(睁眼或者闭眼,微笑或者不笑)和不同面部细节(戴眼镜或者不戴眼镜)等情况下得到的.所有图像都是 32×32 像素大小.在分类实验中,我们随机选取每个人的 5 幅图像用作训练,剩余 5 幅用作测试.

基于上面两组数据的实验结果如表 1 所示,最左边的列中的 d 值表示样本降维后的维数,余下 3 列,从左至右分别表示 LDA、GP-LVM 和 SGP-LVM 在不同维数上的分类结果.对 YaleB 数据,我们分别降至 5、15、30 和 37 维比较分类结果;对 ORL 数据,则降维至 5、10、15 和 30 维来比较 3 个方法的分类结果.

表 1 基于数据 YaleB 和 ORL 分类错误率

YaleB	不同算法的分类错误率		
	LDA+KNN	GP-LVM+KNN	SGP-LVM
$d=5$	0.5205+0.0085	0.6813+0.0190	0.5373+0.0240
$d=15$	0.2332+0.0114	0.5788+0.0326	0.4831+0.0113
$d=30$	0.1495+0.0064	0.1341+0.0065	0.1218+0.0525
$d=37$	0.1272+0.0098	0.1221+0.0090	0.1165+0.0667
ORL	LDA+KNN	GP-LVM+KNN	SGP-LVM
$d=5$	0.3500+0.0184	0.3540+0.0429	0.3440+0.0363
$d=10$	0.1950+0.0137	0.1740+0.0277	0.1770+0.0332
$d=15$	0.1720+0.0231	0.1570+0.0271	0.1453+0.0266
$d=30$	0.1250+0.0262	0.1240+0.0323	0.1174+0.0469

从以上两组实验结果可以看出,SGP-LVM 和 GP-LVM 能得到比 LDA 较好的分类结果.对于第 1 组数据 YaleB,虽然在维数降至 $d=5$ 和 $d=15$ 时 LDA 得到的分类效果比 SGP-LVM 和 GP-LVM 都要好,但随着维数的增加,SGP-LVM 的分类效果明显得到提高.特别是对 ORL 数据,与 LDA 和 GP-LVM 相比较,SGP-LVM 的分类效果更加明显.尤其当数据降维的维数高于 15 以后,SGP-LVM 就能得到非常好的分类结果.这主要是因为相比于 LDA 和 GP-LVM,SGP-LVM 非线性和监督信息的利用取得了很好的效果. GP-LVM 虽然可以对高维数据降维得到低维流形结构,但是这种降维流形不能区别类别信息. SGP-LVM 正是对其进行了改进,使得 GP-LVM 这种有效的降维方法可以用于分类,以上的实验也证明了本文方法的有效性.

4.3 回归数据处理结果

前两组实验验证了本文提出的算法在处理分类数据时的有效性,下面的实验则说明该算法在处理回归问题时同样可以取得较好的效果,实验数据 Housing 和 Concrete Compressive Strength 来自 UCI 数据库. Housing 是由 506 个样本组成,每个样本有 13 个特征来描述,即样本的维数为 13. 第 2 个

Concrete Compressive Strength 数据包含 1030 个样本,特征维数为 8. 这两个数据都要回归到 1 维输出上,输出数据与输入的高维特征间有很强的非线性. 利用本文提出的算法可以找到高维特征潜在的流形结构,然后建立低维的流形到输出值之间的关系. 在实验中我们也比较了经典非线性回归模型(高斯过程回归, GPR)^[17].

由于本文提出的方法是先试图消除次要或者无关因素对数据的干扰,找到数据潜在的流形结构,所以需要高维数据降维,然后从低维空间再映射到

输出值上. 对于 Housing 数据,我们选 $d=4, 6, 8$, 对于 Concrete Compressive Strength 数据,我们选用 $d=2, 3, 5$. 从图 3 给出的实验结果可以看出,本文的方法在各个维度上回归性能上优于 GPR 方法,尤其是当训练数据集较少的时候. 这是因为我们提出的算法能通过降维的方法去除一些无关或者干扰信息,找到数据内在的非线性结构,在回归的时候就可以得到较为精确的结果. 而 GPR 方法没有去冗余信息的过程,因此模型的精度受到了影响.

图 3 本文算法与 GPR 回归性能比较

5 结 论

本文提出了一种基于高斯过程隐变量模型的非线性分类方法,该方法利用先验的类别信息,得到隐变量的后验概率. 然后通过后验概率,得到高维数据点在隐空间的位置,同时优化出模型中映射的超参数的值,建立分类模型. 同时该方法仍然保持了高斯过程隐变量模型的降维优势,即可以有效地找到高维数据的低维流形结构. 实验结果证明了本文方法的有效性. 由于我们利用的类别信息得到隐变量的后验,所以这种方法具有开放性,我们可以利用其它的监督信息替代这里的类别信息,例如回归模型的输出值;而若我们已知一些半监督的信息,那么该方法可以很容易扩展到半监督模型.

参 考 文 献

- [1] Basilevsky A T. Statistical Factor Analysis and Related Methods: Theory and Applications. New York: Wiley, 2004
- [2] Roweis S. EM algorithms for PCA and SPCA//Proceedings of the International Conference on Advances in Neural Information Processing Systems. Cambridge, MA, USA, 1997:

626-632

- [3] Duda R O, Hart P E, Stork D G. Pattern Classification. 2nd Edition. Hoboken, NJ: Wiley-Interscience, 2000
- [4] Tipping M E, Bishop C M. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B, 1999, 61(3): 611-622
- [5] Schölkopf B, Smola A, Müller K-R. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 1998, 10(5): 1299-1319
- [6] Liu Qing-Shan, Huang Rui, Lu Hang-Qin, Ma Song-De. Face recognition using kernel based fisher discriminant analysis//Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition. Washington DC, USA, 2002: 187-191
- [7] Mika S, Ratsch G, Weston J, Schölkopf B, Müller K-R. Fisher discriminant analysis with kernels//Proceedings of the IEEE Neural Networks for Signal Processing Workshop IX. Madison, WI, USA, 1999: 41-48
- [8] Lawrence N D. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. Journal of Machine Learning Research, 2005, 6(11): 1783-1816
- [9] Lawrence N D. Gaussian process models for visualization for high dimensional data//Proceedings of the International Conference on Advances in Neural Information Processing Systems. Cambridge, MA, USA, 2004: 329-336

[10]

Rasmussen C E, Williams C K I. Gaussian Process for Machine Learning. Cambridge, MA: The MIT Press, 2006

[11]

Urtasun R, Darrell T. Discriminant Gaussian process latent variable model for classification//Proceedings of the 24th International Conference on Machine Learning. Corvallis, OR, USA, 2007: 927-934

[12]

Dawid A P. Conditional independence in statistical theory. Journal of the Royal Statistical Society, Series B, 1979, 41(1): 1-31

[13]

Rubin D B, Thayer D T. EM algorithm for ML factor analysis. Psychometrika, 1982, 47(1): 69-76

[14]

Bartholomew D J. Latent Variable Models and Factor Analysis. London: Charles Griffin & Co. Ltd, 1987

[15]

Yu Shi-Peng et al. Supervised probabilistic principal component analysis//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2006). Philadelphia, PA, USA, 2006: 464-473

[16]

Moller M F. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks, 1993, 6(4): 525-533

[17]

Williams C K I, Rasmussen C E. Gaussian processes for regression//Touretzky D S, Mozer M C, Hasselmo M E eds. Advances in Neural Information Processing Systems. British Columbia, Canada, 1996: 598-604

WANG Xiu-Mei, born in 1978, Ph. D., associate professor. Her research interests mainly involve nonparametric statistical models and machine learning.

GAO Xin-Bo, born in 1972, professor, Ph. D. supervisor. His research interests include computer vision, pattern recognition.

LI Jie, born in 1972, professor, Ph. D. supervisor. Her research interests include pattern recognition, data mining.

Background

In pattern recognition and data mining, many interested objects are naturally stored or hidden in very high-dimensional space for better descriptions. However, these descriptions will lead to great troubles for data analysis. And the high dimensions would lead to “curse of dimensionality” in data analysis and processing. Therefore dimensionality reduction plays more and more important role in machine learning and pattern recognition.

To handle with the small sample size and noisy problem, probabilistic latent variable models, which establish mappings from the low-dimensional latent variables space to the

high-dimensional observed space, was developed as a popular framework used in dimensionality reduction. In this paper, a supervised latent variable model for classification is developed. The MAP algorithm is employed to estimate all latent variables position. This research was supported partially by the National Natural Science Foundation of China under Grant Nos.61125204, 61172146 and 61100158, the Ph. D. Programs Foundation of Ministry of Education of China under Grant No.20090203110002, and the Fundamental Research Funds for the Central Universities (K50511020002).