结合半监督核的高斯过程分类

李宏伟 [1] 刘 扬 [1] 卢汉清 [1] 方亦凯 [2]

摘 要 提出了一种半监督算法用于学习高斯过程分类器, 其通过结合非参数的半监督核向分类器提供未标记数据信息. 该算法主要包括以下几个方面: 1) 通过图拉普拉斯的谱分解获得核矩阵, 其联合了标记数据和未标记数据信息; 2) 采用凸最优化方法学习核矩阵特征向量的最优权值, 构建非参数的半监督核; 3) 把半监督核整合到高斯过程模型中, 构建所提出的半监督学习算法. 该算法的主要特点是: 把基于整个数据集的非参数半监督核应用于高斯过程模型, 该模型有着明确的概率描述, 可以方便地对数据之间的不确定性进行建模, 并能够解决复杂的推论问题. 通过实验结果表明, 该算法与其他方法相比具有更高的可靠性.

关键词 高斯过程, 半监督学习, 核方法, 凸最优化
中图分类号 TP391

Gaussian Processes Classification Combined with Semi-supervised Kernels

LI Hong-Wei[1] LIU Yang[1] LU Han-Qing[1] FANG Yi-Kai[2]

Abstract In this paper, we present a semi-supervised algorithm to learn Gaussian process classifiers, which is combined with nonparametric semi-supervised kernels in the presence of unlabeled data. This algorithm mainly includes the following aspects: 1) The spectral decomposition of graph Laplacians is used to obtain kernel matrices incorporating labeled and unlabeled data; 2) The convex optimization method is employed to learn the optimal weights of kernel matrix eigenvectors, which construct the nonparametric semi-supervised kernels; 3) The proposed semi-supervised learning algorithm is obtained by incorporating semi-supervised kernels into Gaussian process model. The main characteristic of the proposed algorithm is that we employ the nonparametric semi-supervised kernels based on the entire dataset into the Gaussian process model, which has an explicit probabilistic interpretation, and can model the uncertainty among the data and solve the complex non-linear inference problems. The effectiveness of the proposed algorithm is demonstrated by the experimental results in comparison with other related works in the literature.

Key words Gaussian processes, semi-supervised learning, kernel method, convex optimization

在机器学习和数据挖掘领域, 半监督学习[1] 愈来愈受到科研人员的关注, 逐步成为当前的研究热点. 半监督学习包括若干研究分支: 半监督分类、半监督回归、半监督聚类、基于图的半监督学习、协同训练以及直推学习等, 本文主要关注的是半监督分类问题. 我们在处理学习问题时所采用的数据集由一系列数据组成, 其中一部分数据被标注, 可由相应的特征向量来描述, 剩余数据未被标注. 传统的分类方法只使用标注数据, 即采用相应的特征向量进行训练和学习. 由于数据信息的多样性, 对某些数据的标注是非常困难、昂贵耗时的, 需要有经验的标注者大量工作和努力, 然而对未标注数据获取会相对比较容易. 相对于传统方法, 半监督学习既考虑了一定的标记数据信息, 又结合了大量的未标记数据信息, 进而建立更好的分类器, 以提高分类性能, 因此半监督学习方法能够有助于解决真实世界中的很多问题.

最近几年, 高斯过程[2−9] 逐渐成为一种倍受关注的监督学习方法, 为核函数的学习提供了一种既具有理论基础, 又可用于实践的概率模型, 并在模型的选择、学习和预测方面提供了一个完整的理论框架. 在监督学习中, 传统的参数模型对简单的数据集可给出方便的描述, 但是对于复杂的数据集, 这类参数模型缺少描述能力, 不能应用在实践工作中. 高斯过程通过一个随机过程支配数据特性, 再采用一个概率分布来描述这个随机过程, 正是这种灵活的非参数特性使得高斯过程能够对预测结果的不确定性提供有效的估计, 可对复杂的数据集进行建模, 并应用于实际问题中去. 高斯过程分类器的目标在于对给定数据点预测其类标的后验概率, 由于此后验概率不受未标记数据的影响, 使得未标记数据并不能影响决策边界的位置. 因而针对高斯过

收稿日期 2008-06-16 收修改稿日期 2008-11-16
Received June 16, 2008; in revised form November 16, 2008
国家高技术研究发展计划 (863 计划) (2006AA01Z315), 国家自然科学基金重点项目 (60833006), 国家自然科学基金 (60605004) 资助
Supported by National High Technology Research and Development Program of China (863 Program) (2006AA01Z315), Key Project of National Natural Science Foundation of China (60833006), and National Natural Science Foundation of China (60605004)
1. 中国科学院自动化研究所模式识别国家重点实验室 北京 100190
2. 诺基亚中国研究中心 北京 100176
1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
2. Nokia Research Center, China, Beijing 100176
DOI: 10.3724/SP.J.1004.2009.00888

程的诸多特性, 我们考虑如何把这种监督学习方法有效地扩展到半监督学习框架中去. 结合未标记数据信息构建性能更好的分类器, 本文认为可以从以下两个方面入手: 1) 选择具有一定特性的似然函数并结合高斯过程先验, 使得后验分布具有类似于支持向机量机 (Support vactor machine, SVM) 的 Margin 特性, 进而利用未标记数据影响其决策面的位置. 文献 [1] 中采用了这一思想, 通过合并高斯过程先验和零类噪声摸型 (Null-category noise model, NCNM), 提出了一种判别式半监督学习方法, NCNM 模型等同于一种 "概率 Margin", 其类似于直推式支持向量机 (Transductive support vector machine, TSVM)[1,10] 方法, 能够将聚类假设合并到高斯过程中. Rogers 在文献 [4] 中采用多项式概率 (Multinomial probit) 似然函数替换 NCNM 模型, 把 Lawrence 的二分类方法扩展到多类问题; 2) 直接从高斯过程先验入手, 修改高斯过程的先验函数, 使先验函数的核函数具有半监督核的特性, 能够合并标记数据信息和未标记数据信息, 进一步提高分类器的性能, 这正是本文所要讨论的方法.

半监督核[11−14] 通过图拉普拉斯的谱分解获得, 其依赖于整个数据集 (包括标记和未标记的样本), 未标记样本能够辅助分类任务, 用于学习更为合理和更为有效的分类器, 例如谱聚类 (Spectral clustering) 方法[12] 采用整合聚类假设的思想, 扩散核 (Diffusion kernels) 方法[13] 采用指数族核函数, 高斯随机域 (Gaussian random field) 方法[14] 采用平滑可逆的核函数, 但这些方法只能定义于有限的样本空间, 不能够在整个样本 (标记和未标记的样本) 空间中定义再生核希尔伯特空间. Sindhwani 在文献 [15] 中提出一种基于图拉普拉斯结构的半监督核, 即再生核希尔伯特空间 (Reproducing kernel Hilbert spaces, RKHS), 其有效地整合了聚类假设和流形假设, 通过对拉普拉斯图顶点的平滑, 使其转化为定义在整个数据空间的再生核希尔伯特空间, 进而结合了未标记样本的几何特性, 并在文献 [16] 中应用于高斯过程中. 但是以上的半监督核方法都属于参数类算法, 此类算法的主要困难在于针对不同的数据库和数据类型如何选择一个合适的谱变换函数族, 因为对于大多数的函数族来说, 参数化的自由度不足以准确地建模数据, 并且在选定核函数类型以后, 又要面临着参数变量的选择和预先假设等一系列繁琐问题. 因此, 在本文中半监督核的设计策略采用一种非参数化方法, 通过核校准 (Kernel alignment) 算法最优化图的谱分解, 获得最优的非参数 RKHS 半监督核形式, 从而避免了参数类算法的核函数类型和模型选择等问题, 并把此核函数整合到高斯过程算法中, 进而把高斯过程算法扩展到半监督学习框架中.

1 高斯过程

1.1 高斯分布和高斯过程

高斯过程[2] 可被视为随机变量的集合, 其中任意有限数量的随机变量均服从高斯分布, 例如一个均值为 μ、协方差矩阵为 Σ 的高斯分布可由下式描述

$$f = (f_1, \cdots, f_n) \backsim N(\mu, \Sigma) \tag{1}$$

其中, f_i 为随机变量. 高斯过程可通过一个随机过程 $f(x)$ 的均值函数 $m(x)$ 和协方差函数 (核函数) $K(x, x')$ 来描述

$$f(x) \backsim GP(m(x), K(x, x')) \tag{2}$$

其中, x_i 为随机变量, 均值函数为: $m(x) = \mathrm{E}[f(x)]$, 协方差函数为: $K(x, x') = \mathrm{E}[(f(x) - m(x))(f(x') - m(x'))]$.

由此可见, 高斯过程是对一个多维高斯分布的泛化, 其通过一个随机过程 $f(x)$ 支配和掌管随机变量的特性, 再采用一个概率分布来描述这个随机过程, 从而定义了一个灵活的非参数概率模型. 总体来说, 高斯过程是描述函数的分布, 其定义是在函数空间.

1.2 高斯过程分类

本文只考虑二分类问题. 假定有数据集 $X = \{X_l, X_u\}$, 其中 $X_l = \{x_1, \cdots, x_l\}$ 为标注数据, 其相应的类标集为 $Y_l = \{y_1, \cdots, y_l\}$, $X_u = \{x_{l+1}, \cdots, x_{l+u}\}$ 为未标记数据. 存在一个潜在函数 $f(x)$ 服从高斯过程: $f(x, \theta) \backsim GP(0, K)$, 其中高斯过程的均值为零, $K \in \mathbf{R}$ 是其协方差函数 (即核函数), θ 为协方差函数 K 的参数. 这个潜在函数 f 便是前面所说的随机过程, 其定义了数据集 X 和类标集 Y 之间的映射关系. 那么相对于潜在函数的类概率可由下式描述

$$p(y = +1|f(x)) = \Phi(f(x)) \tag{3}$$

其中, Φ 函数为 S 型 (Sigmoid) 类函数, 例如逻辑型 (Logistic) 函数或累积高斯 (Cumulative Gaussian) 函数.

由于在给定潜在函数 f 时, 观测数据是相互独立的, 似然函数可以描述为

$$p(y|f) = \prod_{i=1}^{l+u} p(y_i|f_i) = \prod_{i=1}^{l+u} \Phi(y_i f_i) \tag{4}$$

结合高斯过程先验 $p(f)$ 及其似然函数 $p(y|f)$, 可以得到后验分布

$$p(f|y,X,\theta)=\frac{p(y|f)p(f|X,\theta)}{p(y|X,\theta)} \tag{5}$$

给定测试数据点 x_t 时, 潜在函数 $f(x_t)$ 的值为

$$p(f_t|y,X,\theta,x_t)=\int p(f_t|f,X,\theta,x_t)p(f|y,X,\theta)\mathrm{d}f \tag{6}$$

所以给定测试数据点 x_t 的类标预测概率为

$$p(y_t|y,X,\theta,x_t)=\int p(y_t|f_t)p(f_t|y,X,\theta,x_t)\mathrm{d}f_t \tag{7}$$

总体来说, 对于二分类问题高斯过程分类的基本思想为: 首先对潜在函数 $f(x)$ 施以高斯过程先验, 再通过选用适当的似然函数 (例如 Sigmoid 类函数) 来描述此映射关系, 最终实现对测试数据点的类标预测.

1.3 条件独立性

标准的高斯过程作为一种判别式模型, 其图形描述可由图 1 给出. 图形结构中不同背景色的圆形节点代表了在算法过程中的不同处理方式, 其中灰色节点代表可观测的变量, 黑色节点代表需被最优化处理的变量, 白色节点代表需边际化的变量.

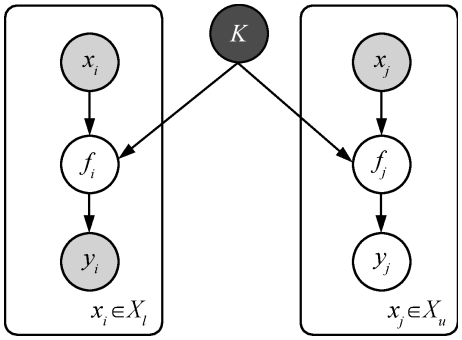


图 1 高斯过程在判别式框架中的图形描述

Fig. 1 The graphical representation of Gaussian processes in the discriminative framework

如图 1 所示, 对于标注数据集 $x_i \in X_l$, 虽然潜在变量 f_i 是不可观测的, 但由于 x_i 和 K 有着共同的延伸节点 y_i, 并且 y_i 是可观测的, 使得 x_i 和 K 相互关联, 也就是说, 核矩阵能够有效捕捉标注数据信息. 对于未标注数据集 $x_j \in X_u$, 由于延伸节点 y_j 是不可观测的, 使得 x_j 和 K 条件独立, 即未标注数据 x_j 不能够影响潜在变量 f 的后验分布, 因此无法影响分类器的决策面位置. 在以下章节中, 我们将提出一种非参数的半监督核, 其能够同时捕捉标注数据和未标注数据的信息, 进而影响分类器的决策面位置.

2 半监督核

2.1 RKHS 半监督核

假设给定一个数据集 $X=\{X_l, X_u\}$, 为了从标注数据中学得一个好的分类器, 可以通过解决如下的正则化问题来实现

$$f=\arg\min\left\{\frac{1}{l}\sum_{i=1}^{l}C(f,x_i,y_i)+\Omega(\|f\|_{\mathcal{H}})\right\} \tag{8}$$

其中, C 为损失函数, 反映了函数 f 对数据的匹配程度; $\|f\|_{\mathcal{H}}$ 为函数 f 在再生核希尔伯特空间中的范数, 反映了在再生核希尔伯特空间中对函数 f 的平滑测量. 依据描述定理 (Representer theorem)[17], 可以获得如下形式的最优解

$$f^*(x)=\sum_{i=1}^{l}\alpha_i K(x,x_i) \tag{9}$$

也就是说, 最优解 f^* 可表示为基于数据 $\{x_i\}$ 核函数的线性组合. 因此在一个无限维的空间中寻求最优解时, 可以通过增加正则项, 把原有问题简化到有限维空间中去.

在半监督学习中, 当我们同时考虑标注数据 $\{x_i,y_i\}_{i=1}^{l}$ 和未标注数据 $\{x_i,y_i\}_{i=l+1}^{l+u}$ 信息时, 原有的最优化问题转化为下式

$$f=\arg\min_{f\in\widetilde{\mathcal{H}}}\left\{\frac{1}{l+u}\sum_{i=1}^{l+u}C(f,x_i,y_i)+\Omega(\|f\|_{\widetilde{\mathcal{H}}})\right\} \tag{10}$$

其中, $\widetilde{\mathcal{H}}$ 为新的再生核希尔伯特空间, 同时包含标注数据和未标注数据. Sindhwani 在文献 [15] 中指出在给定未标注数据时, 我们可以通过对原有再生核希尔伯特空间 $\mathcal{H}$ (对应于标注数据) 进行适当的弯曲, 获得一个更能符合数据分布几何特性的新再生核希尔伯特空间 $\widetilde{\mathcal{H}}$. 假定 $\mathcal{V}$ 为满足半正定内积的线性空间, S 为有界线性算子, 那么把 $\widetilde{\mathcal{H}}$ 定义为由原空间 $\mathcal{H}$ 和修订的内积所构成的函数空间: $\langle f,g\rangle_{\widetilde{\mathcal{H}}}=\langle f,g\rangle_{\mathcal{H}}+\langle Sf,Sg\rangle_{\mathcal{V}}$. 新空间 $\widetilde{\mathcal{H}}$ 仍然为再生核希尔伯特空间, 新的数据依赖范数 $\|f\|_{\widetilde{\mathcal{H}}}$ 使得新空间 $\widetilde{\mathcal{H}}$ 在引入未标记数据后, 相对于原有空间 $\mathcal{H}$ 更加匹配数据分布的几何特性, 与空间 $\widetilde{\mathcal{H}}$ 相对应的新核函数 $\widetilde{K}(x_i,y_j)$ 可通过再生核希尔伯特空间的正交特性和再生特性推论给出[15]

$$\widetilde{K}(x_i,y_j)=K(x_i,y_j)-K_x^{\mathrm{T}}(I+MK)^{-1}MK_x \tag{11}$$

其中, $K_x=[K(x_1,x),\cdots,K(x_{l+u},x)]^{\mathrm{T}}$, M 为对称半正定矩阵, 同时对核函数 $K(x_i,x_j)$ 的选择采用选取不同形式, 可以得到不同参数类型的 RKHS 半监督核, 使得这一方法属于参数类算法, 其主要困难

在于在选择函数类型时, 此类函数的参数化自由度不足以准确地建模数据, 进而引出模型选择问题. 在半监督学习算法中, 模型选择问题主要难点是较小的标注数据集, 通常模型选择算法需要对原本已经较小的标注数据集重新进行子集划分, 用于学习和训练, 往往会出现过学习或模型不准确问题. 因此针对参数类算法的主要缺点, 本文采用基于图的最优化方法学得一种非参数化的 RKHS 半监督核.

2.2 非参数 RKHS 半监督核

大多数基于图的半监督学习方法都可被视为: 对目标函数施以平滑条件, 其中目标函数反映了数据点的标注关系, 平滑条件 (或平滑函数) 通过核函数作用于数据点间的权重图, 其中最核心的问题是对图拉普拉斯进行谱分解. 针对数据集 $X = \{X_l, X_u\}$ 我们可构建一个图 $G = \{V, E\}$, 其中顶点集 V 代表了整个数据点集, 边界集 E 代表了数据点间的边界权重, 可通过权重矩阵 $W = \{w_{ij}\}$ 来描述各个数据点间的加权情况, 例如当数据点 x_i 和 x_j 不相连时, $w_{ij} = 0$. 设矩阵 $D_{ii} = \sum_j W_{ij}$ 为权重矩阵 W 的对角度矩阵, 图拉普拉斯则被定义为: $L = D - W$, 则图拉普拉斯 L 的谱分解为

$$L = \sum_{i=1}^{l+u} \lambda_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^{\mathrm{T}} \tag{12}$$

其中, λ_i 为图拉普拉斯 L 的特征值, 且 $\lambda_1 \le \cdots \le \lambda_{l+u}$, $\boldsymbol{\phi}_i$ 为图拉普拉斯 L 的标准正交特征向量. 通过对特征值 λ_i 的变换: $\lambda_i = r(\lambda_i)$, 可以获得基于图的半监督核为

$$K = \sum_{i=1}^{l+u} r(\lambda_i) \boldsymbol{\phi}_i \boldsymbol{\phi}_i^{\mathrm{T}}, r(\lambda_i) \ge 0 \tag{13}$$

半监督核矩阵 K 定义了再生核希尔伯特空间: $\|f\|_K^2 = f^{\mathrm{T}} K^{-1} f$. 从半监督学习的角度来看, 平滑函数 f 在权重图中不平滑区域将被惩罚, 也就是说, 惩罚项应该限制频谱中的高频部分而加强低频部分. 由于在谱分解中, 较小的特征值对应的特征向量是平滑的, 代表了数据中较大的聚类结构, 而较大的特征值对应着不平滑的特征向量, 往往被视为噪声. 所以在选取变换函数 $r(\lambda)$ 时, $r(\lambda)$ 本质上翻转了特征值 λ 的次序, 即平滑的特征向量 $\boldsymbol{\phi}_i$ (较小的 λ_i) 对应着半监督核矩阵 K 中较大的特征值. 由于 $r(\lambda)$ 的递减特性, 使得函数在不平滑时将受到较大的惩罚.

对 $r(\lambda)$ 选取不同形式, 可以得到不同参数类型的半监督图核, 但是我们目的是寻求非参数化的 RKHS 半监督核, 因此在选择基于图的半监督核函数时采用如下形式

$$K^* = \sum_{i=1}^{l+u} \alpha_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^{\mathrm{T}}, \quad \alpha_i \ge 0 \tag{14}$$

其中, 对 α_i 的选择应满足序列限制条件: $\alpha_i \ge \alpha_{i+1}$, 进而保证 α_i 的递减特性, 使平滑函数 f 在不平滑区域能够被惩罚, 本文将核校准算法用于最优化框架可学得最优的 α_i. 在获得非参数的半监督图核 K^* 后, 便可得到非参数化的 RKHS 半监督核

$$\widetilde{K}(x_i, y_j) = K^*(x_i, y_j) - K_x^{*\mathrm{T}} (I + LK)^{-1} L K_x^* \tag{15}$$

2.3 核校准

核校准[18−19] 反映了两个不同核函数所产生的映射之间的相互关系, 可用于评估核函数矩阵和训练类标矩阵的相似性. 核校准基本思想在于: 核函数能够提取目标的相关特征用于学习, 由于目标函数未知, 可以假设某一类任务可包含这一目标函数, 并且核函数能够在这类任务中给出好的结果. 本文通过最大化核校准评分 (如下式) 获得最优的半监督核, 用于高斯过程进行分类任务.

$$A(K^*, T) = \frac{\langle K^*, T \rangle_F}{\sqrt{\langle K^*, K^* \rangle_F \langle T, T \rangle_F}} \tag{16}$$

其中矩阵 K^* 为基于整个数据集的半监督核矩阵 K 的子矩阵, 对应于训练数据集; 矩阵 T 为目标矩阵, 反映了训练集中数据点的标注信息, 其中 $T_{ij} = y_i y_j$, 对于二分类问题, 如果 $y_i = y_j$, $T_{ij} = 1$, 否则 $T_{ij} = -1$.

核校准也可被视为一种聚类测量, 其正比类内距离, 反比于类间距离, 并具有如下优良特性: 1) 在核函数训练之前, 只依赖于训练数据, 校准可进行预先计算; 2) 在期望值附近急剧集中, 因此对于不同的数据划分, 其经验值稳定; 3) 具有较好的泛化能力: 如果核函数和类标函数能够较好的匹配, 则存在一个具有较低泛化误差的数据划分.

2.4 凸最优化

本文采用一种非参数最优化的算法最优化式 (16) 中的核校准评分以及序列约束条件, 从而获得最优的半监督核矩阵. 其中序列限制条件为: $\alpha_i \ge \alpha_{i+1}$, 该条件反映了图拉普拉斯的先验知识, 即增加平滑函数的权重, 同时惩罚不平滑函数. 这样便构成了一个非线性凸优化问题, 其中目标函数为

$$\max_{K^*} [A(K^*, T)] \tag{17}$$

s. t.

$$K^* = \sum_{i=1}^{l+u} \alpha_i K_i \tag{18}$$

$$\mathrm{tr}(K^*) = 1 \tag{19}$$

$$\alpha_i \geq 0 \tag{20}$$

$$\alpha_i \geq \alpha_{i+1},\ \ i = 1, \cdots, l+u \tag{21}$$

其中, $K_i = \boldsymbol{\phi}_i \boldsymbol{\phi}_i^{\mathrm{T}}$, $\boldsymbol{\phi}_i$ 为图拉普拉斯的特征向量. 式 (19) 保证了核校准的尺度不变性, 式 (20) 保证了半监督核 K 为半正定矩阵. 从式 (17) 可以发现在凸最优化算法中, 最大化目标函数 $A(K^*, T)$ 等同于最大化其分子 $\langle K^*, T\rangle_F$, 由于目标矩阵 T 的元素 T_{ij} 在二分类问题中取 $+1$ 或 -1, 结合式 (18) 不能发现目标函数是 α_i 的线性函数, 这使得在最优化过程中达到最优解的迭代次数可与线性规划相比, 使得该算法能够应用于大数据库. 在获得最优的核函数 K^* 后, 代入到式 (15) 中便可获得非参数化的 RKHS 半监督核 $\widetilde{K}$, 并应用于高斯过程框架中进行分类任务.

3 结合半监督核的高斯过程分类

3.1 半监督高斯分类算法

在获得最优的半监督核矩阵后, 使其同高斯过程相结合, 便可形成结合半监督核的高斯过程分类算法, 此算法的图形结构如图 2 所示. 其中, X 表示数据集, Y 表示类标集, f 为服从高斯过程分布的潜在函数, $\widetilde{K}$ 为通过学习获得的非参数化 RKHS 半监督核矩阵.

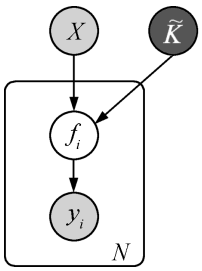


图 2 结合半监督核的高斯过程图形结构

Fig. 2 The graphical structure of Gaussian processes combined with semi-supervised kernels

此算法的基本流程为:

1) 首先对给定的训练数据集进行特征提取, 获得标准的图拉普拉斯 L;

2) 对标准的图拉普拉斯 L 进行谱分解, 获得相应的次序标准正交特征向量 $\boldsymbol{\phi}_i$;

3) 利用凸最优化算法计算出最优的半监督核 K^*, 并代入式 (15) 中获得非参数化 RKHS 半监督核 $\widetilde{K}$;

4) 将计算出的最优半监督核矩阵 $\widetilde{K}$ 导入式 (7), 替换潜在函数 $f \sim GP(0, K)$ 的核函数 K, 利用高斯过程分类算法便可计算出测试数据点的类标预测概率. 因此结合式 (4), (5) 和 (7) 变换为以下形式

$$p(f|X, \widetilde{K}) = \frac{N(f|0, \widetilde{K})}{p(y|X, \widetilde{K})} \prod_{i=1}^{l+u} \Phi(y_i f_i) \tag{22}$$

$$p(y_t|y, X, \widetilde{K}, x_t) = \int p(y_t|f_t) p(f_t|y, X, \widetilde{K}, x_t) \mathrm{d}f_t \tag{23}$$

观察式 (4) 可发现由于似然函数 $p(y|f)$ 采用的是 Sigmoid 类函数, 即便先验分布 $p(f|0, \widetilde{K})$ 为高斯函数, 也使得式 (22) 中的后验分布 $p(f|X, \widetilde{K})$ 以及式 (23) 中的预测分布 $p(y_t|X, \widetilde{K}, x_t)$ 无法通过解析求解. 因此在进行高斯过程分类算法时的核心思想是: 把非高斯类的真实后验分布 $p(f|X, \widetilde{K})$ 通过一个高斯类的近似后验分布 $q(f|X, \widetilde{K}) = N(f|m, \Sigma)$ 来代替, 并通过此近似后验分布给出测试数据的近似预测分布. 把近似高斯后验分布代到式 (6) 中, 可以得到潜在函数 f 在测试数据点 x_t 的近似后验 $q(f|y, X, \widetilde{K}, x_t) = N(f_t|\mu_t, \sigma_t^2)$, 其中均值和方差如下式所示

$$\mu_t = k_t^{\mathrm{T}} \widetilde{K}^{-1} m \tag{24}$$

$$\sigma_t^2 = k(x_t, x_t) - k_t^{\mathrm{T}} (\widetilde{K}^{-1} - \widetilde{K}^{-1} \Sigma \widetilde{K}^{-1}) k_t \tag{25}$$

其中, k_t 为测试数据 x_t 与训练数据集 X 的先验协方差函数. 例如对于采用 Probit 似然函数, 测试数据点 x_t 的近似预测分布可由下式计算

$$q(y_t = +1|y, X, \widetilde{K}, x_t) = \int \Phi(f_t) N(f_t|\mu_t, \sigma_t^2) \mathrm{d}f_t = \Phi\left(\frac{\mu_t}{\sqrt{1+\sigma_t^2}}\right) \tag{26}$$

在本文中, 近似高斯后验分布 $q(f|X, \widetilde{K}) = N(f|m, \Sigma)$ 的参数 m 和 Σ 采用期望传播 (Expectation propagation, EP)[20] 迭代算法求得. EP 算法的主要思想是在每次迭代过程中通过一个区域高斯似然函数来近似潜在函数 f 的非标准高斯后验, 迭代过程中采用一种相对熵 (Kullback-Leibler divergence) 算法[21] 更新参数. 虽然在每次迭代中的似然函数都是区域性的, 但由于潜在变量都经过先验分布进行耦合, 使得 EP 算法所形成的最终后验近似分布是全局性的.

3.2 算法复杂度

针对标准监督高斯分类算法和本文所提出的结合非参数半监督核高斯分类算法而言, 算法复杂度可从预测阶段和学习阶段两个方面进行阐述: 1) 在预测阶段: 获得所需的核函数 K 后, 对于每个给定的测试数据 x_t, 两种方法的算法复杂度均为 $\mathrm{O}(N^2)$, 其中 N 表示训练样本数; 2) 在学习阶段: 对于标准监督高斯方法, 由于需要计算 $N \times N$ 核矩阵的逆, 使得算法的复杂度为 $\mathrm{O}(N^3)$, 但本文采用 EP 方法进行模型选择时 (学习核函数的超参数 θ), 通过对核矩阵 K 进行乔里期基 (Cholesky) 分解, 使得算法的复杂度为 $N^3/6$; 对于半监督高斯方法, 由于目标

函数是关于最优参数 α_i 的线性函数, 结合限制条件 $\alpha_i \geq 0$, 使得最优化问题变为二次方程约束二次规划问题, 其所需迭代次数可以同线性规划相比拟, 其算法复杂度一般为 $\mathrm{O}(N^2)$, 无需像标准监督高斯算法那样通过 $\mathrm{O}(N^3)$ 次计算来进行模型选择, 更适合应用于大样本、高维数的分类问题. 如果采用一些加速技术, 例如采用并行计算的主分解技术[22], 可将复杂度进一步降低为 $\mathrm{O}(N \log N)$, 这将在进一步的工作中实践和验证.

而且, 标准的高斯算法在进行模型选择 (学习核函数的参数) 之前, 要针对不同的数据库和数据类型选择不同的核函数类型, 往往越是复杂的数据, 待学习的核函数参数也就越多 (为了反映数据间的复杂关系), 算法效率也就越差; 本文提出的半监督高斯算法无需考虑核函数类型, 只需针对不同数据库学习其特征向量的最优权值, 其有着更高的算法效率.

4　实验

4.1　实验数据

为了验证非参数半监督核的高斯过程分类算法的有效性, 我们分别进行了两组实验. 在第一组实验中采用了四个数据库[11], 数据库信息如表 1 所示. One vs Two 和 Odd vs Even 取自 Cedar Buffalo 二类数字数据库, 用于手写数字识别任务. 数据库 One vs Two 用于分类手写数字 "1" 和 "2", 包含 2 200 个数据. 数据库 Odd vs Even 用于分类手写偶数数字 "$0, 2, 4, 6, 8$" 和奇数数字 "$1, 3, 5, 7, 9$", 包含 4 000 个数据. Pc vs Mac 和 Baseball vs Hockey 取自 20 个新闻组数据库, 用于文档分类任务. 数据库 Pc vs Mac 包含 1 943 个数据, 数据库 Baseball vs Hockey 包含 1 993 个数据. 实验在数据特征提取方面, 数据库 One vs Two 和 Odd vs Even 对原始特征采用欧式距离, 获得欧氏距离最近邻 (Euclidean 10-nearest-neighbor, 10NN) 未加权图用于谱分解; 数据库 Pc vs Mac 和 Baseball vs Hockey 对词频逆文本频率向量采用余弦相似性测量, 获得余弦相似度最近邻 (Cosine similarity 10-nearest-neighbor, 10NN) 未加权图用于谱分解.

表 1　第一组实验中的数据库信息

Table 1　The information of database on Experiment 1

数据库	样本数	类别个数	特征提取
One vs Two	2 200	2	Euclidean 10 NN
Odd vs Even	4 000	2	Euclidean 10 NN
Pc vs Mac	1 943	2	Cosine similarity 10 NN
Baseball vs Hockey	1 993	2	Cosine similarity 10 NN

在第二组实验中采用了三个数据库[15], 数据库信息如表 2 所示. G50c 为人工数据库, 其数据由标准正态分布以等概率生成; Coil20 包含 20 类物体的 32×32 的灰度图像, 并且包含视角的变换; Uspst 取自 USPS 数据库, 用于手写数字识别任务.

表 2　第二组实验中的数据库信息

Table 2　The information of database on Experiment 2

数据库	类别个数	样本数	标记样本数	特征维数
G50c	2	550	50	50
Coil20	20	1 140	40	1 024
Uspst	10	2 007	50	256

4.2　实验结果与分析

在第一组实验中, 对每一个数据库分别选取 5 个大小不同的标注数据集, 剩余数据用于测试. 对每个标注数据集的给定规模, 随机抽取 20 次训练数据集, 分别进行最优化核校准评分, 学习最优的半监督核矩阵以及高斯过程分类, 最后通过平均获得半监督核的高斯过程分类算法在各数据库中的平均正确率. 为了验证算法的有效性, 实验中对两种算法进行比较, 一种是结合非参数半监督核的高斯过程分类算法 (即本文所提出的算法), 另一种为结合标准监督核 RBF (Radial basis function) 的高斯过程分类算法, 相当于监督学习算法, 其中 RBF 核采用如下形式

$$K(x, x') = \theta_1 \exp\left(-\frac{(x - x')^{\mathrm{T}}(x - x')}{2\theta_2^2}\right) \tag{27}$$

其中, θ_1 和 θ_2 为 RBF 核函数的参数, 通过最大化下式中的边际似然函数 $q(y|X, \theta)$

$$q(y|X, \theta) = \int p(y|f) q(f|X, \theta) \mathrm{d}f \tag{28}$$

即计算 $\dfrac{\partial q(y|X, \theta)}{\partial \theta_i}$, 求出最优的 RBF 核参数 θ_i, 将最优参数导入式 (6) 和 (7) 中便可实现结合标准监督核 RBF 的高斯过程对测试数据的分类.

第一组实验结果由表 3 ∼ 6 (见下页) 给出, 其中包括 Zhu 在文献 [11] 中的实验结果 (分别为每个表中的第三列和第七列), 以及我们采用经典半监督算法中的扩散核方法 (每个表中的第四列) 和高斯随机域方法 (每个表中的第五列) 进行的实验结果. 这三种方法都使用半监督核, 采用标准的 SVM 作为分类器. 从实验结果不难发现, 本文提出的半监督核的高斯过程分类算法在大多数样本集下都优于其他算法, 体现了高斯过程模型在执行分类任务时的可靠性和优越性. 实验结果还包括了半监督核和监督核两种算法的比较 (每个表中的第六列和第七列). 结果显示半监督核通过引入未标记数据信息, 使得分类算法性能得到提升, 充分显示了本文所提出的结合半监督核的高斯过程分类算法的有效性.

表 3 One vs Two 数据集的平均正确率 (%)

Table 3 The average accuracies on the One vs Two datasets (%)

训练集大小	半监督核				监督核 (RBF)	
	本文方法	文献 [11] 方法	扩散核方法	高斯随机域方法	本文方法	文献 [11] 方法
10	96.1	96.2	67.2	65.3	85.1	78.7
20	97.3	96.4	81.4	85.6	91.6	90.4
30	98.3	98.2	92.5	93.4	94.9	93.6
40	98.7	98.3	95.6	94.7	95.6	94.0
50	98.9	98.4	96.9	95.9	97.1	96.1

表 4 Odd vs Even 数据集的平均正确率 (%)

Table 4 The average accuracies on the Odd vs Even datasets (%)

训练集大小	半监督核				监督核 (RBF)	
	本文方法	文献 [11] 方法	扩散核方法	高斯随机域方法	本文方法	文献 [11] 方法
10	69.7	69.6	62.7	60.1	69.3	65.0
30	83.7	82.4	80.4	76.3	79.6	77.7
50	88.2	87.6	85.2	82.6	83.6	81.8
70	90.1	89.2	87.3	85.3	86.4	84.4
90	92.6	91.5	90.1	89.5	87.2	86.1

表 5 Pc vs Mac 数据集的平均正确率 (%)

Table 5 The average accuracies on the Pc vs Mac datasets (%)

训练集大小	半监督核				监督核 (RBF)	
	本文方法	文献 [11] 方法	扩散核方法	高斯随机域方法	本文方法	文献 [11] 方法
10	87.8	87.0	55.2	53.4	54.9	51.6
30	90.9	90.3	75.7	72.6	65.4	62.6
50	92.2	91.3	79.3	77.2	70.2	67.8
70	93.4	91.5	83.5	80.7	76.8	74.7
90	93.6	91.5	85.8	83.2	81.2	79.0

表 6 Baseball vs Hockey 数据集的平均正确率 (%)

Table 6 The average accuracies on the Baseball vs Hockey datasets (%)

训练集大小	半监督核				监督核 (RBF)	
	本文方法	文献 [11] 方法	扩散核方法	高斯随机域方法	本文方法	文献 [11] 方法
10	96.2	95.7	61.2	59.4	60.2	53.6
30	98.2	98.0	69.9	68.7	72.6	69.3
50	98.6	97.9	77.5	79.6	80.4	77.7
70	98.8	97.9	84.1	85.3	85.5	83.9
90	99.1	98.0	90.4	92.6	89.7	88.5

表 7 第二组实验结果的平均正确率 (%)

Table 7 The average accuracies of the results on Experiment 2 (%)

数据库	本文方法	扩散核法	高斯随机域法	SVM[15]	LapSVM[15]
G50c	94.6	88.9	85.2	90.3	95.0
Coil20	87.8	72.1	73.5	77.4	85.4
Uspst	84.2	74.2	71.6	77.9	82.3

第二组实验结果由表 7 给出, 表中第三列和第四列分别采用经典半监督核的扩散核方法和高斯随机域方法, 表中第五列和第六列采用了标准的 RKHS 核, 分别通过 SVM 和拉普拉斯支持向量机 (Laplacian SVM, LapSVM) 方法执行分类任务, 其中 SVM 和 LapSVM 都采用了 4 折交叉验证, 即把每一个数据集都分为 4 个相等的子集, 其中三个子集构成训练数据集, 剩余一个子集用于测试; 表中第二列为本文算法, 即采用本文所提出的非参数化 RKHS 半监督核, 并把此半监督核整合到高斯过程框架中去, 执行半监督学习任务. 从实验结果可发现, 本文提出的结合半监督核高斯过程算法要明显优扩散核算法和高斯随机域算法; 对于标准的 RKHS 核, 除在 G50c 数据库下的结果略低于文献 [15], 总体来说与文献 [15] 中的方法具有可比较性, 体现了结合非参数化 RKHS 半监督核高斯过程分类算法的可靠性和优越性.

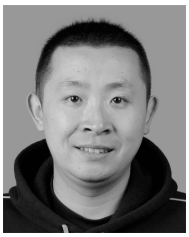
5 结论与展望

半监督学习逐渐成为近年来被广泛关注的研究热点. 本文提出并评价了一种结合半监督核的高斯过程分类算法, 由非参数的半监督核引入未标注数据信息, 并应用于高斯过程模型中执行分类任务, 未标记数据信息的引入提高了分类器的性能, 实验结果同时验证了该算法的可靠性和有效性. 由于本文算法主要考虑的是二分类问题, 未来的工作将集中在如何把该算法推广到多类目标的分类问题上, 例

如采用多项式概率似然函数; 同时考虑到标准的高斯过程模型对大样本集的计算复杂度较高, 如何改进高斯过程模型, 降低计算复杂度, 使其能够适合大样本集, 也将是我们下一步的研究重点.

References

1 Chapelle O, Scholkopf B, Zien A. *Semi-Supervised Learning*. Cambridge: MIT Press, 2006
2 Rasmussen C E, Christopher K I W. *Gaussian Processes for Machine Learning*. Cambridge: MIT Press, 2006
3 Rasmussen C E. Advances in Gaussian processes. In: Proceedings of the Neural Information Processing Systems Conference. Cambridge, USA: MIT Press, 2006. 1−55
4 Rogers S, Girolami M. Multi-class semi-supervised learning with the E-truncated multinomial probit Gaussian process. In: Proceedings of the Gaussian Processes in Practice Workshop. Berlin, German: Springer, 2007. 17−32
5 Lawrence N D. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *The Journal of Machine Learning Research*, 2005, **6**: 1783−1816
6 Lawrence N D, Candela J Q. Local distance preservation in the GP-LVM through back constraints. In: Proceedings of the 23rd International Conference in Machine Learning. Pittsburgh, Pennsylvania: ACM, 2006. 513−520
7 Lawrence N D. Learning for larger datasets with the Gaussian process latent variable model. In: Proceedings of the 11th International Workshop on Artificial Intelligence and Statistics. San Juan, Puerto Rico: Morgan Kaufmann Publishers, 2007. 1−8
8 Lawrence N D, Moore A J. Hierarchical Gaussian process latent variable models. In: Proceedings of the 24th International Conference in Machine Learning. Corvalis, Oregon: ACM, 2007. 481−488
9 Urtasun R, Darrell T. Discriminative Gaussian process latent variable model for classification. In: Proceedings of the 24th International Conference in Machine Learning. Corvalis, Oregon: ACM, 2007. 927−934
10 Joachims T. Transductive inference for text classification using support vector machines. In: Proceedings of the 16th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 1999. 200−209
11 Zhu X J, Kandola J, Ghahramani Z, Laffert J. Nonparametric transforms of graph kernels for semi-supervised learning. In: Proceedings of the Conference on Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2005. 1641−1648
12 Chapelle O, Weston J, Scholkopf B. Cluster kernels for semi-supervised learning. In: Proceedings of the Conference on Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2002. 585−592
13 Kondor R I, Lafferty J D. Diffusion kernels on graphs and other discrete input spaces. In: Proceedings of the 19th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers, 2002. 315−322
14 Zhu X J, Ghahramani Z, Lafferty J D. Semi-supervised learning using Gaussian fields and harmonic functions. In: Proceedings of the 20th International Conference on Machine Learning. Washington D. C., USA: MIT Press, 2003. 912−919
15 Sindhwani V, Niyogi P, Belkin M. Beyond the point cloud: from transductive to semi-supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany: MIT Press, 2005. 824−831
16 Sindhwani V, Chu W, Keerthi S S. Semi-supervised Gaussian processes classifiers. In: Proceedings of the International Joint Conference on Artificial Intelligence. San Francisco, USA: Morgan Kaufmann Publishers, 2007. 1059−1064
17 Kimeldorf G S, Wahba G. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 1971, **33**(1): 82−95
18 Cristianini N, Shawe-Taylor J, Elisseeff A, Kandola J. On kernel-target alignment. In: Proceedings of the Conference on Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2002. 367−373
19 Lanckriet G R G, Cristianini N, Bartlett P, Ghaoui L E, Jordan M I. Learning the kernel matrix with semidefinite programming. *The Journal of Machine Learning Research*, 2004, **5**: 27−72
20 Minka T P. A Family of Algorithms for Approximate Bayesian Inference [Ph. D. dissertation], Massachusetts Institute of Technology, USA, 2001
21 Kullback S. *Information Theory and Statistics*. New York: Dover, 1959
22 Smith B, Bjrstad P, Gropp W. *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*. Cambridge: Cambridge University Press, 1996



李宏伟 中国科学院自动化研究所博士研究生. 主要研究方向为图像与视频处理, 机器学习. 本文通信作者.
E-mail: hwli@nlpr.ia.ac.cn
(**LI Hong-Wei** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interest covers image and video processing, and machine learning. Corresponding author of this paper.)

刘 扬 中国科学院自动化研究所博士研究生. 主要研究方向为图像与视频处理, 机器学习.
E-mail: liuyang@nlpr.ia.ac.cn
(**LIU Yang** Ph. D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interest covers image and video processing, and machine learning.)

卢汉清 中国科学院自动化研究所研究员. 主要研究方向为图像与视频处理, 多媒体信息检索.
E-mail: luhq@nlpr.ia.ac.cn
(**LU Han-Qing** Professor at the Institute of Automation, Chinese Academy of Sciences. His research interest covers image and video processing, and multimedia technique.)

方亦凯 诺基亚中国研究中心研究员. 2008 年获得中国科学院自动化研究所模式识别国家重点实验室博士学位. 主要研究方向为图像处理和模式识别.
E-mail: ykfang@gmail.com
(**FANG Yi-Kai** Researcher at Nokia Research Center, China. He received his Ph. D. degree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science in 2008. His research interests covers image processing and pattern recognition.)