文章编号:1003-0530(2016)08-0960-13

# 从高斯过程到高斯过程混合模型:研究与展望

周亚同[1,2] 陈子一[1] 马尽文[1]
(1. 北京大学数学科学学院,数学及其应用教育部重点实验室,北京 100871;
2. 河北工业大学电子信息工程学院,天津 300401)

**摘 要**:高斯过程(GP)模型是核学习方法与贝叶斯推理相结合的典范,现已成为机器学习领域的一个研究热点。作为对 GP 模型的拓展,高斯过程混合(MGP)模型具有更强大的学习能力和适应性。然而,目前关于 GP 和 MGP 模型的研究较为零散,尚缺少系统的分析与总结。本文首先对于 GP 模型的基本原理及其研究进展进行了深入地分析和讨论;然后将 GP 模型拓展至 MGP 模型,从多方面对 MGP 模型的研究现状和进展进行了深入地分析和讨论,并指出未来值得探索的研究方向和应用问题。

**关键词**:高斯过程;高斯过程混合模型;机器学习;回归预测;聚类分析
**中图分类号**:TP18 **文献标识码**:A **DOI**:10.16798/j.issn.1003-0530.2016.08.11

# From Gaussian Processes to the Mixture of Gaussian Processes: A Survey

ZHOU Ya-tong[1,2] CHEN Zi-yi[1] MA Jin-wen[1]
(1. School of Mathematical Science and LMAM, Peking University, Beijing 100871, China;
2. School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300401, China)

**Abstract**: Gaussian process (GP) model is a paradigmatic machine learning model that combines the advantages of both kernel learning method and Bayesian inference mechanism, and thus has become a very popular area in machine learning in recent years. As an extension of the GP model, the Mixture of Gaussian Processes (MGP) fits datasets more effectively and thus it has a better ability of learning and generalization. However, there are only some isolated literatures and reports about the GP and MGP models and no systematic summary on these models. In this paper, we begin to review the GP model and its basic principles and developments on various aspects. We then discuss how to extend the GP model to the MGP model and further review the status and developments of the MGP models, and finally point out some prospective research directions and interesting applications of the MGP model.

**Key words**: Gaussian process; mixture of Gaussian processes; machine learning; regression and prediction; clustering analysis

## 1 引言

高斯过程是最重要的随机过程之一,其任意有限维样本均服从高斯分布。在机器学习领域,高斯过程常被称为高斯过程模型[1-2](Gaussian process model,以下简称 GP 模型)。虽然对于高斯过程的研究已有了悠久的历史,但是将之用于解决机器学习问题的历史却并不久远。近年来 GP 模型在机器学习领域颇受关注,已成为继支持向量机(SVM)之后新的研究热点,主要有以下几方面的原因:

首先,GP 模型被认为是统计机器学习的一个基本框架。许多学习方法比如 SVM、径向基函数(RBF)网络以及卡尔曼滤波器等,均可以看作是 GP 模型在特殊情形下的一种实现[1-4]。第二,GP 模型是核机器学习[5]与贝叶斯推理学习相结合的典范,兼具以上两类学习方式的优势。第三,GP 模型的先

收稿日期:2015-09-30;修回日期:2016-02-22
基金项目:国家自然科学基金(61171138);教育部人文社会科学基金(15YJA630108);中国博士后科学基金(2014M561053)共同资助

验知识导入直观;模型通过矩阵运算表示、简洁高效;模型能产生概率信息[6]。第四,GP 模型具有广阔的应用前景。不仅能解决分类、回归与预测问题,而且还能延展至更广阔的应用领域。

GP 模型尽管有以上诸多优势,但也存在一些不足。主要表现在不能很好地描述多模态(样本可被分为多簇,每簇来自于不同的分布)样本集、模型超参数的学习耗时、难于刻画样本集输出在不同输入区域的波动性差异等。为此,Tresp[7] 在 2000 年首先提出了高斯过程混合模型(Mixture of Gaussian Processes,以下简称为 MGP 模型),很好解决了上述问题且具有更好的灵活性。MGP 模型最初是在专家混合模型框架下发展起来的,但近年来自身发展逐步自成一体,受到机器学习领域的广泛关注。

追溯 GP 模型的历史,其应用于机器学习的时间(1996 年)比 SVM(1992 年)要晚一些。实际上,Rasmussen[8] 首次提出了 GP 模型并将之用于回归预测。经过 20 年的发展,如今有关 GP 模型的研究文献已屡见不鲜。国外有学者建立了关于 GP 模型的网站[9],出版了相关论著[10],并基于不同平台编写了很多模型实现软件[9,11]。国内也有一些文献以 GP 模型为研究对象[12]。与 GP 模型相比,MGP 模型的研究时间还不算太长,相关文献也要少一些。

当前 GP 模型的研究进展主要体现在哪些方面,未来还有哪些问题值得进一步研究? GP 模型为什么要向 MGP 模型拓展? MGP 模型目前的发展状况如何,还面临哪些亟待解决的问题? 上述问题促使我们梳理从 GP 模型发展至 MGP 模型的脉络,评述两类模型各自的研究进展,并展望未来的发展趋势。本文希望以此抛砖引玉,吸引更多学者关注这两类模型,并对相关领域研究者提供参考和借鉴。

## 2　高斯过程模型的基本原理

回归预测和分类是两类基本的机器学习问题,下面分别从这两个角度来介绍模型的原理。对模型原理更详细的阐述可参考文献[1,2,6,10]。

(1)GP 模型用于回归预测

已知 $N$ 个学习样本 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$ 及其对应目标值 $t_1,t_2,\cdots,t_N$,设所求回归函数为 $f(\boldsymbol{x})$,依其预测在新增样本 $\boldsymbol{x}_{N+1}$ 处的函数值为 $f(\boldsymbol{x}_{N+1})$。GP 模型首先假设 $f(\boldsymbol{x})$ 是一个高斯过程;然后将目标向量 $\boldsymbol{t}_N=(t_1,t_2,\cdots,t_N)^{\mathrm{T}}$ 看作是另一个高斯过程 $t(\boldsymbol{x})$ 在 $\boldsymbol{X}_N=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)^{\mathrm{T}}$ 处的实现,且假设 $f(\boldsymbol{x})$ 与 $t(\boldsymbol{x})$ 间满足关系式 $t(\boldsymbol{x})=f(\boldsymbol{x})+\mathrm{e}(\boldsymbol{x})$,其中 $\mathrm{e}(\boldsymbol{x})$ 是一个方差为 $\sigma_\nu^2$ 的高斯白噪声。现将 $f(\boldsymbol{x})$ 表示成 $H$ 个基函数的加权和 $f(\boldsymbol{x})=\sum_{h=1}^{H}w_h\Phi_h(\boldsymbol{x})$,并假定随机向量 $\boldsymbol{w}=(w_1,w_2,\cdots,w_H)^{\mathrm{T}}$ 满足 $\boldsymbol{w}\sim N(0,\sigma_w^2\boldsymbol{I}_H)$。由 $f(\boldsymbol{x})$ 与 $t(\boldsymbol{x})$ 间的关系式可得到 $\boldsymbol{t}_N\sim N(0,\boldsymbol{C}_N)$,其中矩阵 $\boldsymbol{C}_N$ 由下面的协方差函数所决定:

$$C(\boldsymbol{x}_n,\boldsymbol{x}_m)=\sigma_w^2\sum_{h=1}^{H}\Phi_h(\boldsymbol{x}_n)\Phi_h(\boldsymbol{x}_m)+\sigma_\nu^2\delta_{nm}\quad(1)$$

其中当 $n=m$ 时 $\delta_{nm}=1$,否则 $\delta_{nm}=0$。

当给定测试样本 $\boldsymbol{x}_{N+1}$ 时,欲预测其目标值 $t_{N+1}$,此时有

$$P(t_{N+1}\mid\boldsymbol{t}_N)=P(t_{N+1},\boldsymbol{t}_N)/P(\boldsymbol{t}_N)\quad(2)$$

根据 GP 的假设,上式分子的表达式为:

$$P(t_{N+1},\boldsymbol{t}_N)\propto\exp\left\{-\frac{1}{2}\begin{bmatrix}\boldsymbol{t}_N^{\mathrm{T}} & t_{N+1}\end{bmatrix}\boldsymbol{C}_{N+1}^{-1}\begin{bmatrix}\boldsymbol{t}_N\\ t_{N+1}\end{bmatrix}\right\}\quad(3)$$

现将矩阵 $\boldsymbol{C}_{N+1}$ 分解,并分别记其子块为 $\boldsymbol{k}$、$\boldsymbol{k}^{\mathrm{T}}$ 和 $\kappa$ 等。从而可得目标值 $t_{N+1}$ 的预测分布:

$$P(t_{N+1}\mid\boldsymbol{t}_N)\sim N(\hat{t}_{N+1},\sigma_{\hat{t}_{N+1}}^2)\quad(4)$$

其中 $\hat{\boldsymbol{t}}_{N+1}=\boldsymbol{k}^{\mathrm{T}}\boldsymbol{C}_N^{-1}\boldsymbol{t}_N$,$\sigma_{\hat{t}_{N+1}}^2=\kappa-\boldsymbol{k}^{\mathrm{T}}\boldsymbol{C}_N^{-1}\boldsymbol{k}$。最后令 $f(\boldsymbol{x}_{N+1})=\hat{\boldsymbol{t}}_{N+1}$,即取该预测分布的均值作为 GP 模型在 $\boldsymbol{x}_{N+1}$ 处的预测值。

(2)GP 模型用于二元分类问题

若 GP 模型用于二元分类问题,则式(2)中的条件分布 $P(t_{N+1}\mid\boldsymbol{t}_N)$ 不再是高斯分布。因为 $P(t_{N+1}\mid\boldsymbol{t}_N)$ 也可记为 $P(t_{N+1}\mid\boldsymbol{X}_{N+1},\boldsymbol{t}_N)$,此时其表达式变为:

$$P(t_{N+1}=1\mid\boldsymbol{X}_{N+1},\boldsymbol{t}_N)=\int P[t_{N+1}=1\mid f(\boldsymbol{x}_{N+1})]\,P[f(\boldsymbol{x}_{N+1})\mid\boldsymbol{X}_{N+1},\boldsymbol{t}_N]\,\mathrm{d}f(\boldsymbol{x}_{N+1})\quad(5)$$

上式中的两个分布分别为

$$P[t_{N+1}=1\mid f(\boldsymbol{x}_{N+1})]=\frac{1}{1+\mathrm{e}^{-f(\boldsymbol{x}_{N+1})}}\quad(6)$$

$$P[f(\boldsymbol{x}_{N+1})\mid\boldsymbol{X}_{N+1},\boldsymbol{t}_N]=\int P[f(\boldsymbol{x}_{N+1}),\boldsymbol{f}_N\mid\boldsymbol{X}_{N+1},\boldsymbol{t}_N]\,\mathrm{d}\boldsymbol{f}_N=\int P(\boldsymbol{f}_{N+1}\mid\boldsymbol{X}_{N+1},\boldsymbol{t}_N)\,\mathrm{d}\boldsymbol{f}_N\quad(7)$$

其中 $\boldsymbol{f}_N=[f(\boldsymbol{x}_1),f(\boldsymbol{x}_2),\cdots,f(\boldsymbol{x}_N)]^{\mathrm{T}}$,$P[\boldsymbol{f}_{N+1}\mid\boldsymbol{X}_{N+1},$

$\boldsymbol{t}_N]\ \propto\ P(\boldsymbol{f}_{N+1}\mid\boldsymbol{X}_{N+1})\prod_{n=1}^{N+1}P[t_n\mid f(\boldsymbol{x}_n)]$，$P(\boldsymbol{f}_{N+1}\mid\boldsymbol{X}_{N+1})\propto\exp\left(-\frac{1}{2}\boldsymbol{f}_{N+1}^{\mathrm{T}}\boldsymbol{K}_{N+1}^{-1}\boldsymbol{f}_{N+1}\right)$。

由于式(6)不是高斯分布，将(6)与(7)代入(5)以后，式(5)仍不是高斯分布。因此当 GP 模型用于分类时，数值计算要比回归预测复杂。可用拉普拉斯逼近法、蒙特卡洛法[4,13]、变分法[14]、均值场法[15]等近似计算式(5)中的后验概率。

## 3 高斯过程模型的研究进展与展望

现从以下九个方面综述 GP 模型的研究进展，并对其未来的研究方向进行展望。

(1)GP 模型的协方差函数与模型选择

协方差函数对 GP 模型而言非常重要，它扮演的角色就如核机器学习方法中的核函数。GP 模型的协方差函数大致可以分为平稳协方差和非平稳协方差函数两类。前者如平方指数、分段多项式协方差函数等，后者如点积型协方差函数等。GP 模型通常采用平稳协方差函数，但难于捕捉被表达函数的平滑度随输入的变化。为此需要构建一些新的非平稳协方差函数[16]。另外通过相加、直和、相乘、张量积、嵌入或卷积等手段，根据已有协方差函数也可以构建新的协方差函数。

当用 GP 模型解决实际问题时，首先面临的是如何选择合适的协方差函数，以及如何确定协方差函数中所含待定超参数，可统称之为模型选择。GP 模型最常用的模型选择方法是贝叶斯推理和交叉验证[10]。除此之外也提出了一些其他的方法，例如 Seeger[17]使用变分贝叶斯模型选择法，自动确定协方差函数中的超参数。Schwaighofer[18]曾借助分层贝叶斯框架实现 GP 模型的均值与协方差函数的自学习。Sundararajan 等[19]基于 Geisser 预测概率最大化准则来确定最优超参数等。

(2)GP 模型的改进或扩展

自 GP 模型提出以后，出现了大量改进或扩展算法。例如 Chatzis[20]提出了一种显状态 GP 模型。Soh 等更进一步提出了在线显状态 GP 模型的空时学习策略[21]。Snelson[22]提出了一种弯曲 GP 模型，能处理含非高斯分布噪声的样本。Boyle[23]提出了相关 GP 模型，将高斯过程视为某平滑核与高斯白噪的卷积，从而能解决多输出变量问题。Lawrence[24]和李宏伟[25]等分别通过扩展 GP 模型，使之能实现半监督学习。Pillonetto[26]扩展 GP 模型后能用于在线多任务学习。Gilboa, E. 等提出了 GP 模型的尺度化多维推理学习策略[27]。传统 GP 模型只能用于二元分类，而 Zhao[28]考虑了将模型用于多元分类。Dallaire[29]考虑到当输入训练样本具有不确定性分布时 GP 模型的回归预测等。上述改进或扩展从不同侧面提升了模型性能。

(3)GP 模型的学习曲线和界

所谓学习曲线是指模型的泛化错误随学习样本数变化的曲线，能深入揭示模型泛化性能的内在规律。目前 GP 模型的学习曲线有如下几种研究思路：第一，研究学习曲线的渐近性质，即当样本容量趋向于无穷大时的曲线形态。第二，研究学习曲线的上下界，掌握学习曲线的变化范围。只要上下界足够紧，同样能够准确刻画学习曲线本身。第三，研究学习曲线的逼近。因为通常得到的上下界都比较松，而学习曲线的逼近可能比其上下界更能准确地刻画学习曲线本身。

Williams[30]对 GP 模型的学习曲线进行了深入研究：在假定模型采用平方指数协方差函数的前提下，获得了学习曲线的解析表达式，给出了曲线的单点上界和两点上界公式。Opper[31]通过对协方差函数进行特征值分解，获得了学习曲线的 Opper-Vivarelli 下界和 OU 下界。由于学习曲线的上下界比较松，为此 Sollich[32]转而研究学习曲线的逼近，并比较了各种界与逼近之间的紧致度。Malzahn[33]使用变分法计算相关分割函数，提供了逼近学习曲线的新思路。Kakade 等[34]则给出了 GP 模型在最坏情形下的界等。

(4)GP 模型的数值实现方法

GP 模型的数值实现有两个目标：第一是在无法进行解析计算时，寻找其近似计算方式。第二是当遇到大样本集时加速训练过程。当 GP 模型用于回归预测时，可通过矩阵求逆直接进行数值实现。但 Mackay[6]曾指出直接求逆不仅耗时而且可能使得计算过程不稳定。一种改进措施是用迭代方法逼近矩阵的逆[35]。周杰英等提出用幂级数展开解决 GP 模型中矩阵行列式计算问题。另外 KD 树[36]和

Nyström[37]等也用于加速 GP 模型训练。

当 GP 模型用于分类时,因为模型中的后验概率不再服从高斯分布,因此计算过程比用于回归预测时要复杂一些。目前有四种数值实现方法:第一种是用拉普拉斯逼近法得到后验概率的近似解;第二种是用蒙特卡洛法对后验概率进行抽样后逼近[4,13];第三种是用变分法获得后验概率的上界与下界,进而求出后验概率的逼近[14];第四种是借助均值场方程近似计算后验概率[15]。另外 Chalupka 等[38]详述了关于 GP 模型的更多数值实现方法。

GP 模型的稀疏化也是一种减少运算量的重要策略。所谓稀疏化是指用尽可能少的代表性样本代替原始训练样本集,并用于模型的学习及预测,这些代表性样本称为支持点[39]。支持点的求解非常关键,通常可以在某种准则下用贪婪算法逐一从原样本集中选取[39]。Snelson 等[40]将支持点作为待估参数,和协方差参数一起通过最大似然估计求解。Quiñonero-Candela 等[39]系统总结了常见的 GP 模型稀疏化策略并将其纳入统一框架进行比较。

(5)GP 模型与一些统计学习模型间的联系

研究 GP 模型与一些统计学习模型间的联系,不仅有助于深入认识 GP 模型,而且能建立起与既有模型体系间的关联。限于篇幅本文仅讨论 GP 模型和 RBF 网络以及 SVM 间的联系,与更多学习模型如相关向量机、正则化模型、最小二乘分类器间的联系可参见 Rasmussen 等[10]的第 6 章。另外 GP 模型和卡尔曼滤波器、Kriging 方法和自回归滑动平均模型等也有密切联系[6]。

Mackay[6]分析了 GP 模型与 RBF 网络间的联系:当 GP 模型用于回归预测时,若采用某种特殊的指数协方差函数,相当于是用高斯径向基及其平移构成的一组基对被回归函数进行线性表示,此时它等价于一个 RBF 网络。Sollich[41]则指出了 GP 模型与 SVM 间的关系:SVM 可以理解成在显著度框架下,某推断问题采用了 GP 模型的先验和合适的似然函数后得到的最大后验概率解。另外 Gestel[42]讨论了 GP 模型与最小二乘 SVM 间的联系。

(6)GP 模型用于强化学习和用于捕捉动态特性

强化学习是一种重要的学习策略,其本质是对变化的环境进行感知和适应。在强化学习领域,解决连续空间的表示问题主要有离散化或者价值函数逼近等,后者实际上是将强化学习转化为价值函数或回报函数的回归估计。Rasmussen[43]曾将 GP 模型用于强化学习,实际上是用模型去逼近价值函数。与价值函数逼近不同,王雪松等[12]在连续状态空间中把强化学习转化为一个二元分类问题,然后借助 GP 模型的分类能力获得强化学习策略。时序差分学习是强化学习中的一种重要技术,Engel[44]曾提出一种高斯过程时序差分学习框架,并进一步考虑状态转移的随机性和行动选择问题。Ko[45]给出了一个 GP 模型用于强化学习的应用实例。

经典 GP 模型无论是用于回归预测还是分类识别,处理对象都是静态样本集。但模型通过与强化学习结合,能够捕捉处理对象的动态特征。例如 Wang[46]提出了高斯过程动态模型(GPDM),能捕获人行走时的动力学特征。吕培[47]则提出了一种基于 GPDM 的人体节奏运动合成方法。Deisenroth[48]提出了一种高斯过程动态规划(GPDP)模型,能在连续状态空间中逼近值函数。Amoto 等[49]则提出了一种新的基于 GP 模型的增强学习策略。

(7)高斯过程隐变量模型

高斯过程隐变量模型(GPLVM)是 GP 模型的另一种拓展,可以捕捉对象运动的非线性特征[50]。GPLVM 从原理上看与概率主成分分析(PPCA)模型具有某种对偶性。Lawrence[51]详细讨论了 PPCA、核主成分分析和 GPLVM 间的关系,给出了 GPLVM 非线性化的途径,讨论了 GPLVM 如何处理大样本集以及样本属性缺失等问题。Lawrence 等[52]紧接着将 GPLVM 扩展至更为复杂的层次结构,并用于人体运动数据可视化。另外,GPLVM 虽能发掘嵌入在高维空间中的低维流型,但只能用于非监督学习。为此 Urtasun 等[53]提出了判别型 GPLVM,通过在低维流型上施加判别型先验,使之能用于监督学习。

(8)GP 模型在各个领域内的应用

学习模型的应用与理论发展相辅相成,良好的应用效果会使模型获得更多关注,从而反过来促进理论发展。GP 模型主要用在涉及分类、回归及预测等场合。例如 Cheng 等[54]以日本女性 JAFFE 表情库为样本集,将 GP 模型用于人脸表情识别。有些学者则将 GP 模型用于非平稳时间序列预测,但计

算效率大为降低[55]。近年来 GP 模型的应用面逐步拓宽,如用于超光谱图像分类[56],天线输入特征建模[57]、风能预测[58]、音乐风格及情感识别[59]等。

(9)GP 模型未来研究展望

GP 模型欲在理论上更加完善,同时取得更好的应用效果,还存在着很多问题值得深入研究。我们认为 GP 模型有以下六个值得探索的研究方向。

第一,将不变性等先验知识融入 GP 模型,给协方差函数的设计或选择提供依据。例如在模式分类中存在着一些不变性先验:待分类目标经平移、旋转或加粗等变换后其类别保持不变。若能在设计协方差函数时考虑上述先验,会显著提升 GP 模型的分类能力。第二,能处理具有特殊形式的输入和输出变量。在当前的 GP 模型中,每一个输入变量均是普通的向量,对应着唯一的输出目标变量。但在有些应用场合,例如对字符串或者其他多媒体数据进行特征提取后,所获取的输入变量具有结构化特征。此时必须改进 GP 模型,使之能适应上述应用场合。第三,寻找 GP 模型学习曲线的更一般形式以及更紧的上下界。目前在推导 GP 模型的学习曲线时,假定的协方差函数非常特殊,因此推导出的学习曲线不具备一般性;另外在推导上下界的过程中进行了各种近似,导致界不够紧。因此在上述两方面均值得深入探索。

第四,GP 模型在处理海量样本时的快速计算问题。尽管借助各种数值计算方法以及稀疏化策略,GP 模型已经可以处理大规模学习样本集。但当处理海量样本时 GP 模型的学习仍面临困难。此时需要从模型原理和数值计算技巧两方面出发加以解决。第五,丰富 GP 模型用于强化学习的途径,增强模型捕捉动态特性的能力。第六,将 GP 模型用于无监督学习。GP 模型本身是一个有监督学习模型,但可通过拓展使之能解决样本降维等无监督学习问题,前文提及的 GPLVM 就是一个典型例子,其实质是对嵌入在高维样本空间中的低维流型进行建模。未来还需提升 GP 模型的建模能力,解决更多的无监督学习问题。

## 4 从 GP 模型向 GPM 模型的拓展

虽然 GP 模型具有诸多优势,但也存在一些局限:第一,单 GP 模型在学习样本数为 $N$ 的样本集时,其计算复杂度为 $O(N^3)$,不适用于大样本数据学习[60]。第二,单 GP 模型难于精确描述多模态样本集[60]。第三,单 GP 模型难以刻画样本输出在不同输入区域的波动性差异[61]。为解决上述问题,Tresp[7]于 2000 年首次提出 MGP 模型,随后人们在其基础上进行了各种改进。基本思路都是采用“分而治之”策略:将样本集分组,每组样本服从一个高斯过程。这样每个高斯过程只涉及小规模协方差矩阵求解,不仅减少了运算量,而且能自动适应多模态样本集。例如图 1 中的 Toydata 样本集由具有 4 种模态的样本组构成,故用由 4 个 GP 构成的混合模型描述显然比单个 GP 模型更准确[62-65]。

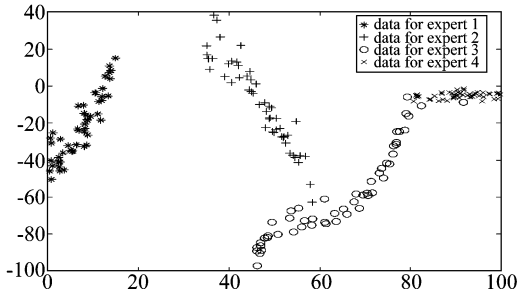


图 1 具有多模态特征的 Toydata 样本集

Fig. 1 The Toydata sample set with multimodality

## 5 高斯过程混合模型的基本原理

MGP 模型最初从混合专家模型(Mixture of Experts,以下简称 ME)发展而来[7]。在最初的 MGP 模型中,每个专家(又称分量)包含三个 GP 模型,分别用于刻画门限函数、输出的均值和方差,训练较费时[7]。后续提出的 MGP 模型为了提升训练效率,大多用单个 GP 表示每个专家。为便于比较 ME 和 MGP 模型,现将二者的边际似然函数表示如下:

$$P(\boldsymbol{t}_N \mid \boldsymbol{X}_N, \boldsymbol{\theta}, \boldsymbol{Z}_N) = \prod_i P(t_i \mid x_i, \theta_{z_i}) \tag{8}$$

$$P(\boldsymbol{t}_N \mid \boldsymbol{X}_N, \boldsymbol{\theta}, \boldsymbol{Z}_N) = \prod_j P(\{t_i : z_i = j\} \mid \{x_i : z_i = j\}, \theta_j) \tag{9}$$

其中 $\boldsymbol{\theta} = \{\theta_j\}_{j=1}^M$ 表示各专家所包含的参数,$\boldsymbol{Z}_N = \{z_i\}_{i=1}^N \subseteq \{1,2,\cdots,M\}^N$ 表示各样本所属的专家编号。$P(\{t_i : z_i = j\} \mid \{x_i : z_i = j\}, \theta_j)$ 表示 MGP 的第 $j$ 个专家的输出服从高斯分布,即

$$P(\{t_i : z_i = j\} \mid \{x_i : z_i = j\}, \theta_j) \sim N[\boldsymbol{\mu}(\{x_i : z_i = j\} \mid \theta_j), \boldsymbol{Q}(\{x_i : z_i = j\}, \{x_i : z_i = j\} \mid \theta_j) + \sigma_{v,j}^2 \boldsymbol{I}_{N_j}] \tag{10}$$

其中$\boldsymbol{\mu}(\{x_i:z_i=j\}\mid\theta_j)$和$\boldsymbol{Q}(\{x_i:z_i=j\},\{x_i:z_i=j\}\mid\theta_j)$分别是第$j$个高斯过程分量的均值和协方差矩阵。通常假定均值$\boldsymbol{\mu}=0$,这样式(10)就退化为单GP模型。

ME和MGP的共同点是:样本均被划分至各专家,且专家间相互独立。二者的差异是:MGP模型的每个专家均由各自对应的GP描述,因而样本间不再像ME那样相互独立,这种混合能同时融合GP和ME的优势。关于ME模型的更多介绍参见综述[66]。

## 6　高斯过程混合模型的研究进展与展望

本节主要阐述MGP模型的常见类型、学习算法、输出预测规则等,并展望未来的发展方向。若无特别说明,以下MGP模型只针对回归预测问题。

### 6.1　MGP模型的常见类型

MGP模型虽形式多样,但大多符合通式(9)和(10),根据模型混合方式的不同可分为判别模型[7,61,67-81]和生成模型[60,62-65,82-88]两类。前者将输入作为已知常量,后者将输入当作随机变量,让每个专家随机生成输入和输出。生成模型与判别模型相比有如下优势[62]:由于输入具有先验分布,根据贝叶斯公式,生成模型可以由输出反推输入,从而可应对某些样本集输入特征缺失的问题;且输出协方差矩阵随输入的变化规律会更清晰。

MGP判别模型和生成模型的表达式分别为[62]

$$P(\boldsymbol{t}_N\mid\boldsymbol{X}_N,\boldsymbol{\mu})=\sum_{\mathbf{Z}_N}\prod_i\Pr(z_i\mid x_i,\theta^g)\prod_j P(\{t_k:z_k=j\}\mid\{x_k:z_k=j\},\theta_j^{\mathrm{GP}}) \quad (11)$$

$$P(\boldsymbol{X}_N,\boldsymbol{t}_N\mid\boldsymbol{\theta})=\sum_{\mathbf{Z}_N}\prod_i\Pr(z_i\mid\theta^g)\prod_j P(\{t_k:z_k=j\}\mid\{x_k:z_k=j\},\theta_j^{\mathrm{GP}})P(\{x_k:z_k=j\}\mid\theta_j^x) \quad (12)$$

上式中,第$j$个专家的输出分布$P(\{t_i:z_i=j\}\mid\{x_i:z_t=j\},\theta_j^{\mathrm{GP}})$大多由式(10)描述,但Yuan等[60]和Sun等[82]为了消除样本输出值间的相关性,用核线性回归模型近似描述各专家输出分布。门限函数$\Pr(z_i\mid x_i,\theta^g)$和$\Pr(z_i\mid\theta^g)$多采用混合比例系数,这些系数有些未给定先验分布[63-65,73,75,84,86-88],有些服从Dirichlet分布[60,67-69,78,85,89],还有些通过Dirichlet过程或Pitman-Yor过程生成[62,70,74,77,79,82-83]。生成模型中第$j$个专家的输入分布$P(\{x_i:z_t=j\}\mid\theta_j^x)$多相互独立且服从高斯分布[62-65,82-88],而Yuan等[60]采用高斯混合分布,输入假设更精确,但运算量也更大。

除了判别模型和生成模型,还有少数MGP模型并不服从式(9)。例如Wang和Khardon[67]采用多任务学习框架,每个任务对应的输出是两个GP的叠加;Kapoor等[75]中的各个GP用于描述不同的输入分量而非不同组样本;Fox和Dunson[90]将输入空间递归地细分为若干子区域,每个子区域由一个GP刻画等。除了混合方式和门限函数等不同以外,MGP模型的多样性还体现在参数的先验分布上面。有些模型不对参数施加先验[63-65,67-68,71-72,79,86-88],有些对均值参数施加高斯先验[60-62,83,85],有些对协方差矩阵的逆施加Wishart先验[60-62,69,82-83,85],有些施加Gamma先验[60-62,69,74,82-83]等。

### 6.2　MGP模型的学习算法

目前MGP模型的主流学习算法有三类:马尔科夫链蒙特卡洛算法(以下简称MCMC)、变分贝叶斯算法(简称VB)、期望最大算法(简称EM),以下逐一介绍。

对于有先验分布的参数,求解其后验分布是学习算法的重要目标。而MGP模型较为复杂,后验分布通常没有解析式,需要利用近似算法,如MCMC算法和VB算法。其中MCMC算法是在给定训练样本的前提下,从某些预设先验分布的隐参量的后验概率中进行序贯抽样[61-62,68-69,77-79,83,85,89]。根据后验概率的具体形式,可灵活选用合适的抽样策略,如M-H抽样(含Gibbs抽样)、混合MCMC抽样等。该方法理论上精确度高[60]。然而由于MGP模型往往参数较多,MCMC用于学习MGP模型时收敛很慢[60,71,82],且收敛条件难于判断[60]。因此有些学者尝试用VB算法逼近后验概率[70,74-75],假设随机隐参量在给定训练数据后条件独立,然而这样近似较粗略,特别当参数的相关性较强时[86]。

EM算法被公认为学习混合模型的有效算法,不需要先验分布且理论上等价于最大似然估计。由于MGP模型中样本输出值相互关联,Q函数计算的时间复杂度很大,目前用于学习MGP模型的EM算法都不可避免地引入了一些近似策略。根据近似策略的差异,又可以将这些EM算法细分为四种:变分EM算法、启发式EM算法、硬分类EM算法、MCMC-EM算法。

其中变分EM算法的E步通过VB算法近似求

解隐参量的后验概率,M 步则通过最大化 Q 函数[60,71,82]或变分自由能[67]求解模型超参数。然而,变分算法的独立性假设较为粗略,例如我们曾尝试编写 Sun 和 Xu[82]的变分 EM 算法,其预测误差往往较大且对参数初始值非常敏感。

启发式 EM 算法的近似主要体现在 M 步的参数学习上。其中 Tresp[7]的核参数值事先给定而未经过学习,Stachniss 和 Plagemann[91]则从与训练样本无关的概率分布中对参数值进行抽样,二者均没有充分利用训练样本的信息;Yang 和 Ma[63-64]、Schiegg 等[76]将 M 步的 Q 函数近似为基于留一交叉验证的概率分解形式,但这种近似需要对每个样本求解其输出值的预测分布,运算量较大。

硬分类 EM 算法的 E 步是按最大后验概率准则,直接将样本分配给对应的高斯过程分量,然后 M 步根据分配结果分别训练每个高斯过程分量。目前的硬分类算法各具特色,如 Yang 和 Ma[63-64]提出的硬分类 EM 算法由上述启发式软分类 EM 算法直接改造而成,精度较高但计算量较大;Nguyen 和 Bonilla[71]的 E 步用变分法对后验概率进行近似,再按最大后验概率准则分配样本,速度快而变分法的独立性假设较为粗略;Yu 和 Chen[84]的 E 步则在最大后验概率大于某一阈值时才将样本分配至对应的高斯过程分量,计算简单但用于分配的后验概率之条件只有输入没有输出,理论依据不完备;Chen 等[65,86]精确推导了硬分类 EM 算法,除硬分类外没用到其他近似策略,实验表明精度高且对中小规模的数据集耗时较少,但对大数据训练速度慢。MCMC-EM 算法由 Wu 等[86]2015 年提出,它将 Chen 等[65]E 步的硬分类策略由最大化后验概率改为 MCMC,提升预测精度的同时运算量增大。

除了以上三类主流学习算法外,MGP 模型还存在其他学习算法,这里简单举例:(1)拉普拉斯逼近算法:Dong[72]用单 GP 模型去近似 MGP 模型,虽然做法简单,但这种近似只保留一、二阶矩的信息,在理论上比较粗糙。(2)一次性硬分类:Tuong 等[80]、Liu[81]等根据马氏距离到聚类中心最小的准则直接对样本分组,然后对各组分别训练一个 GP 模型。这样的马氏距离准则缺少概率依据,较难保证来自同一个 GP 的样本点被分配到同一组。(3)期望传播算法(EP)和变分法联合使用:Kapoor[75]将 MGP 模型用于分类,类别的条件概率用 EP 算法去近似,隐变量的后验概率用变分法近似,该模型不是通常意义下的 MGP(见第 5 节),兹不赘述。

### 6.3 MGP 模型的预测规则

MGP 模型经过学习以后即可用于回归预测,其预测规则大致有如下四种:

(1)加权预测:类似于 ME 模型的预测规则[66],多数文献是用全概率公式预测,即将测试样本放入各高斯过程分量分别预测,然后将这些预测分布或预测值按测试样本分配到各分量(专家)的后验概率加权平均[67,74-76,86,91]。该预测方法理论上十分精确,适用于解析式易求情形,但和某些模型、学习算法匹配后计算量很大,需要下述近似策略。(2)硬分类预测:若模型用硬分类 EM 算法学习,则学习结束后训练样本所属专家均已知,因此可按最大后验概率准则将测试样本分配给相应的专家,再基于单个 GP 模型的预测规则即可获得预测结果[65,67,71,84,86,88]。Nguyen 和 Bonilla[71,86]指出此时硬分类预测通常比加权预测更精准,因为在硬分类 EM 学习过程中,每个专家只与它接收的样本相关,因此对不属于该专家的样本,预测效果通常欠佳。(3)MCMC 抽样预测:若模型用 MCMC 学习,会获得隐参数的大量抽样值。对其中每组抽样值均易求得输出值的预测分布,这些预测分布的均值可近似为真实预测分布[68-69,83,86]。Meeds 和 Osindero[62]的做法则不同,是将预测值作为待估参数,在学习阶段就对其进行 MCMC 抽样,最后直接将这些抽样值取平均。(4)变分参数均值预测:Yuan 等[60]和 Sun 等[82]采用变分 EM 学习算法,预测时随机隐参量的估计值近似为它的期望值。当隐参量的估计值确定后,预测分布就很容易求解。

### 6.4 MGP 模型的研究近况和未来展望

近年来研究 MGP 模型的文献数量明显增长,且从模型、算法的理论改进到应用研究都颇为丰富,涉及几乎各种模型类型、学习算法、预测算法。如 2014 年 Chen 等[64]通过删除不必要参数精简 MGP 模型,进而减少近似策略精确推导该精简模型的学习算法——硬分类 EM 算法。在此基础上,Chen 等[86]通过稀疏化策略发展了此模型[64];Zhao 等[88]为此算法[64]提供模型选择准则;Wu 等[86]将 E 步的硬分类策略由最大化后验概率改为 MCMC,提升了

预测精度。但目前为止,MGP 领域还有些问题尚待进一步探索,包括:自动模型选择、模型超参数的初始值确定和稀疏化策略等,以下逐一阐述。

(1)自动模型选择问题

MGP 模型的模型选择问题,是指根据样本集特点选择合适的高斯过程分量(专家)的个数 $K$。虽然该问题很重要,但目前很多文献都没有讨论,而是事先直接指定 $K$[7,61,63-65,67-69,71,91]。Huang 等[73]提出用 Akaike 信息准则或贝叶斯信息准则进行模型选择,Zhao 等[88]提出了同步平衡准则,虽均比事先指定 $K$ 要准确的多,但需要针对一系列 $K$ 值逐一学习模型,运算量较大。若实现自动模型选择,即在学习过程中自动调整 $K$ 值至合适值,则能在大幅降低运算量的同时增加模型的自适应能力。

如采用 Dirichlet 过程或 Pitman-Yor 过程当作 MGP 模型的门限函数,则混合成分的个数没有上限,因而这样的模型具有自动模型选择的潜能。然而相关实验结果显示,这种方法得到的专家个数往往偏离真实的专家个数,例如 Rasmussen 等[92]用 Dirichlet 过程门限函数,在 Motorcycle 样本集上多次重复实验,专家个数 $K$ 取过很多值且概率分布很不集中。并且,模型选择的效果需要在 MGP 模型合成的数据上检验,而目前的相关文献都没按这个规范进行实验。

Shi 等[93]设想用逆跳 MCMC 算法可实现自动模型选择,2015 年 Qiang 和 Ma[89]实现了这一设想,在合成数据上选出了正确的 K 值。然而所采用的模型都假定样本分组事先已知,比一般的 MGP 模型学习起来较为容易。但在实际中这样的分组往往未知。因此,如何对一般的 MGP 模型实现自动模型选择,是 MGP 模型未来重要的研究方向之一。

(2)模型超参数的初始值确定问题

MGP 模型的超参数,比如协方差函数和先验分布中的待定参数等,其初始值对学习和预测结果均有较大影响。然而,当前多数文献没有明确阐述如何确定合适的初始值。有些文献尝试将训练样本聚类,在每类样本上通过学习获得一个高斯过程分量,从而得到超参数初始值[60,65]。然而初始聚类结果对预测精度和运行时间的影响仍然很大。另外若聚类后某类样本偏多,高斯过程分量的学习运算量仍然较大,且聚类个数的选择问题尚未很好解决。因此如何既快又好确定超参数初始值,是 MGP 模型未来的另一个重要研究方向。

(3)用于大样本数据的稀疏化策略

就象单个 GP 模型一样,MGP 模型也可采用稀疏逼近策略。考虑到训练样本较多时,MGP 的某些高斯过程分量仍然可能包含较多样本,因而有些 MGP 模型用稀疏高斯过程来描述每个分量,从而大幅提高学习效率[60,67,82,91]。例如 Nguyen 和 Bonilla[71]采用基于稀疏化策略的 MGP 模型,在 3 小时以内对 1 万个训练样本进行学习,并对 5 千个测试样本作出预测。当前处于大数据时代,稀疏化策略必将成为 MGP 模型的重要研究方向之一。

### 6.5　高斯过程泛函回归模型(GPFR)及其混合

上文提及的 MGP 模型都用于普通样本学习,Shi 和 Wang[94]于 2007 年提出 GPFR 模型,用于函数型样本集的学习。该类样本集可分为多组,每组均看成一个对象或一条函数曲线的记录,每组记录包括对象或函数特征与各时刻的输入和输出。其实这样的分组模型也出现在 Shi 等[68-69]、Qiang 和 Ma[89]等文献中,只是描述每一组分布的是 GP 模型而非 GPFR 模型。由此看出 GPFR 对样本集的格式要求较高,因此只适用于某些特定的应用问题。

GPFR 模型假设输入是时间的函数,输出的均值通常表示为一些常见基函数如 B 样条的线性组合,输出减去均值后的残差由 GP 模型描述。GPFR 模型的学习分为两步:第一步是用 B 样条插值求均值;第二步是由输出值减去第一步所得的均值估计获得残差,然后用 GP 模型拟合残差。GPFR 模型的预测分为两种情形:若测试样本与某些训练样本在同一组,则 GP 模型预测值与均值函数估计值之和即为最终预测结果;若测试样本所在组未知,则假定它被等概率分配到各组中,此时将测试样本分别放入各组预测,预测值取平均即为最终结果。

为了更好地反映各对象和各输入区域间的差异,Shi 和 Wang[93]紧接着提出了 GPFR 混合模型,其中每组样本被分配给一个 GPFR,分配时所用的混合比例系数由 logistic 模型决定,且各组样本间相互独立,因而输出的似然函数较易解析表示,可采用精确 EM 算法求解。GPFR 混合模型的预测规则类似于单个 GPFR,也分为上述两种情形。2011 年,Shi 等[95]又提出了混合效应 GPFR 模型,由参数化

的混合效应模型与 GPFR 模型融合而成。

## 7 高斯过程模型、高斯混合模型、高斯过程混合模型间的联系

高斯混合模型(GMM)是一种经典的混合模型[96],GMM 用多个高斯分布的和来描述学习样本的分布。GP、GMM、MGP 等三种模型间具有很强的层次关系,现结合图 2 进行描述。图中的 IMGP 是高斯过程无限混合模型的简称,它是 MGP 模型的推广,即采用 Dirichlet 过程或 Pitman-Yor 过程作为门限函数,参与混合的分量个数可以无限。类似地,IGMM 是高斯无限混合模型的简称,它是 GMM 模型的推广。从图中可见,MGP 模型本身又是由有限个 GP 模型混合而成,而 GMM 模型是由有限个单高斯分布混合而成。从图中还可见,IMGP 在 IGMM 的基础上引入了"过程"的概念,相当于在 IGMM 的基础上加入了时间维度。MGP 与 GMM 之间、GP 与单高斯分布之间也有类似关系。

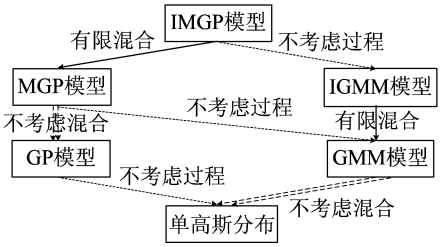


图 2 GP、MGP 与 GMM 等三种学习模型间的层次关系,图中箭头表示退化方向

Fig. 2 The hierarchical relationship of GP, MGP and GMM models, where the arrows denote the direction of degeneration

反过来若只考虑有限个分量的混合,则 IMGP 和 IGMM 分别退化为特例 MGP 和 GMM。更进一步,若不考虑混合即分量个数等于 1 时,MGP 和 GMM 又分别退化为其特例 GP 和单高斯分布。另外如不考虑过程因素,则 IMGP、MGP 和 GP 等三个模型分别退化为其特例 IGMM、GMM、单高斯分布。由此看出 GP、MGP 和 IMGP 模型具有较好的包容性,可看作统计学习研究的基本框架之一。

## 8 结论

本文从九个方面评述了 GP 模型的研究进展,并指出了未来值得探索的六个研究方向。同时梳理了 GP 模型向 MGP 模型拓展的脉络及思路,总结了 MGP 模型常见的混合类型、学习算法以及输出预测规则等,并展望了未来的发展方向。最后展示了 GP、MGP 与 GMM 模型三者间的紧密联系。纵观 GP 和 MGP 模型国内外的整体研究状况,尽管相关研究一直在持续进行,并在实用化方面取得了可喜进展,但在研究深度和广度方面都还有很大的拓展潜力。本文提及的一些亟待解决的问题和值得关注的方向,十分值得进一步深入研究。我们相信,GP 和 MGP 模型凭借着其优良的特性,一定会吸引更多的学者关注与研究,同时也期待这两类模型在更多领域获得成功应用。

### 参考文献

[1] Williams C K I. Prediction with Gaussian processes: From linear regression to linear prediction and beyond[M]. M. I. Jordan. Learning in Graphical Models. [s. l.]: Netherlands: Springer Science & Business Media, 1998:599-621.

[2] Williams C K I, Barber D. Bayesian classification with Gaussian processes[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1998, 20(12): 1342-1351.

[3] Sollich P. Bayesian methods for support vector machines: Evidence and predictive class probabilities[J]. Machine Learning, 2002, 46(1-3): 21-52.

[4] Neal R M. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification[EB/OL]. http://arxiv.org/pdf/physics/9701026v2.pdf, 1997/2015-05-22.

[5] 周亚同, 张太镒, 刘海员. 基于核的机器学习方法及其在多用户检测中的应用[J]. 通信学报, 2005, 26(7): 96-108.
Zhou Y T, Zhang T Y, Liu H Y. Kernel-based machine learning method and the applications to multi-user detection: a survey[J]. Journal on Communications, 2005, 26(7): 96-108. (in Chinese)

[6] MacKay D J C. Introduction to Gaussian processes[J]. NATO ASI Series F Computer and Systems Sciences, 1998, 168: 133-166.

[7] Tresp V. Mixtures of Gaussian processes[C]//Advances in Neural Information Processing Systems 13. Cambridge: MIT Press, 2000: 654-660.

[8] Rasmussen C E. Evaluation of Gaussian processes and other methods for non-linear regression[D]. Toronto: University of Toronto, Department of Computer Science, 1996.

[9] Rasmussen C E. The Gaussian Processes Web Site[EB/

OL]. http://www.gaussianprocess.org, 2011-02-23/2015-05-22.

[10] Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning[M]. Cambridge:MIT Press,2006:1-248.

[11] Seeger M. Gaussian processes for machine learning[J]. International Journal of Neural Systems, 2004, 14(2): 69-106.

[12] 王雪松, 张依阳, 程玉虎. 基于高斯过程分类器的连续空间强化学习[J]. 电子学报, 2009, 37(6): 1153-1158.
Wang X S, Zhang Y Y, Cheng Y H. Reinforcement Learning for Continuous Spaces Based on Gaussian Process Classifier[J]. Acta Electronica Sinica, 2009, 37(6): 1153-1158. (in Chinese)

[13] Barber D, Williams C K I. Gaussian processes for Bayesian classification via hybrid Monte Carlo[J]. Advances in Neural Information Processing Systems, 1996, 9: 340-346.

[14] Gibbs M N, MacKay D J C. Variational Gaussian process classifiers[J]. IEEE Transactions on Neural Networks, 2000, 11(6): 1458-1464.

[15] Opper M, Winther O. Gaussian processes for classification: Mean-field algorithms[J]. Neural Computation, 2000, 12(11): 2655-2684.

[16] Paciorek C, Schervish M. Nonstationary covariance functions for Gaussian process regression[J]. Advances in Neural Information Processing Systems, 2003, 16: 273-280.

[17] Seeger M. Bayesian model selection for support vector machines, Gaussian processes and other kernel classifiers [C]// Proceedings of the 13th Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2000 (EPFL-CONF-161324): 603-609.

[18] Schwaighofer A,Tresp V,Yu K. Learning Gaussian process kernels via hierarchical Bayes[C]// Advances in Neural Information Processing Systems 17. Cambridge:MIT Press, 2004:1209-1216.

[19] Sundararajan S,Keerthi S. Predictive approaches for choosing hyperparameters in Gaussian processes[J]. Neural Computation,2001,13(5):1103-1118.

[20] Chatzis S P,Demiris Y. Echo state Gaussian process[J]. Neural Networks, IEEE Transactions on,2011,22(9): 1435-1445.

[21] Soh, H, Demiris, Y. Spatio-Temporal Learning with the Online Finite and Infinite Echo-State Gaussian Processes [J]. IEEE Trans. Neural Networks and Learning Systems, 2015, 26(3): 522-536.

[22] Snelson E, Rasmussen C E, Ghahramani Z. Warped Gaussian processes[J]. Advances in Neural Information Processing Systems,2003,16:337-344.

[23] Boyle P, Frean M. Dependent Gaussian processes[J]. Advances in Neural Information Processing Systems, 2004, 17: 217-224.

[24] Lawrence N D, Jordan M I. Semi-supervised learning via Gaussian processes[C]// Advances in Neural Information Processing Systems 17. Cambridge: MIT Press, 2004: 753-760.

[25] 李宏伟, 刘扬, 卢汉清, 等. 结合半监督核的高斯过程分类[J]. 自动化学报, 2009, 35(7): 888-895.
Li H W, Liu Y, Lu H Q, et al. Gaussian Processes Classification Combined with Semi-supervised Kernels [J]. Acta Automatica Sinica, 2009, 35(7): 888-895. (in Chinese)

[26] Pillonetto G, Dinuzzo F, De Nicolao G. Bayesian online multitask learning of Gaussian processes[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(2): 193-205.

[27] Gilboa, E, Saatci, Y, Cunningham, J.P. Scaling Multidimensional Inference for Structured Gaussian Processes [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2015, 37(2): 424-436.

[28] Zhao X,Cheung L W K. Multiclass Kernel-Imbedded Gaussian Processes for Microarray Data Analysis[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB),2011,8(4):1041-1053.

[29] Dallaire P, Besse C, Chaib-Draa B. An approximate inference with Gaussian process to latent functions from uncertain data[J]. Neurocomputing, 2011, 74(11): 1945-1955.

[30] Williams C K I, Vivarelli F. Upper and lower bounds on the learning curve for Gaussian processes[J]. Machine Learning, 2000, 40(1): 77-102.

[31] Vivarelli F,Opper M. General bounds on Bayes errors for regression with Gaussian processes[J]. Advances in Neural Information Processing Systems,1999,11:302-308.

[32] Sollich P,Halees A. Learning curves for Gaussian process regression: Approximations and bounds[J]. Neural Computation,2002,14(6):1393-1428.

[33] Opper M,Malzahn D. Learning curves for Gaussian Processes regression: A framework for good approximations[J]. Advances in Neural Information Processing Systems 14, 2001,14:273-279.

[34] Kakade S, Seeger M, Foster D. Worst-case bounds for Gaussian process models[C]// Proc. of the 18th Annual Conference on Neural Information Processing Systems.

Cambridge: MIT Press, 2005: 619-626. (EPFL-CONF-161315).

[35] Gibbs M. Bayesian Gaussian processes for classification and regression [D]. Cambridge: University of Cambridge, Department of Physics, 1997.

[36] Shen Y, Ng A, Seeger M. Fast Gaussian process regression using kd-trees[C]//Proceedings of the 18th Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2005: 1225-1232.

[37] Williams C K I, Rasmussen C E, Scwaighofer A, et al. Observations on the Nyström method for Gaussian process prediction[R]. London: University of Edinburgh and University College London, 2002: 1-9.

[38] Chalupka K, Williams C K I, Murray I. A framework for evaluating approximation methods for Gaussian process regression[J]. The Journal of Machine Learning Research, 2013, 14(1): 333-350.

[39] Quiñonero-Candela J, Rasmussen C E. A unifying view of sparse approximate Gaussian process regression[J]. The Journal of Machine Learning Research, 2005, 6: 1939-1959.

[40] Snelson E, Ghahramani Z. Sparse Gaussian Processes using pseudo-inputs[C]//Advances in Neural Information Processing Systems 18. Cambridge: MIT Press, 2005: 1257-1264.

[41] Sollich P. Probabilistic Methods for Support Vector Machines[C]//Advances in Neural Information Processing Systems 12. Cambridge: MIT Press, 1999:349-355.

[42] Van Gestel T, Suykens J A K, Lanckriet G, et al. Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis[J]. Neural Computation, 2002, 14(5): 1115-1147.

[43] Rasmussen C E, Kuss M. Gaussian processes in reinforcement learning[C]//Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2003: 751-759.

[44] Engel Y, Mannor S, Meir R. Reinforcement learning with Gaussian processes[C]//Proceedings of the 22nd International Conference on Machine Learning. New York: ACM, 2005: 201-208.

[45] Ko J, Klein D J, Fox D, et al. Gaussian processes and reinforcement learning for identification and control of an autonomous blimp[C]//Robotics and Automation, Jinan Shandong China: 2007 IEEE International Conference on. IEEE, 2007: 742-747.

[46] Wang J M, Fleet D J, Hertzmann A. Gaussian process dynamical models for human motion[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2008, 30(2): 283-298.

[47] 吕培, 张明敏, 徐明亮, 等. 基于高斯过程动态模型的人体节奏运动合成[J]. 中国图象图形学报, 2011, 16(8): 1511-1515.
Lv P, Zhang M M, Xu M L, et al. Rhythmical motion synthesis based on Gaussian process dynamical model [J]. Journal of Image and Graphics, 2011, 16(8): 1511-1515. (in Chinese)

[48] Deisenroth M P, Rasmussen C E, Peters J. Gaussian process dynamic programming [J]. Neurocomputing, 2009, 72(7): 1508-1524.

[49] Amoto C, Chowdhary, G, Liu M, et al. Off-policy reinforcement learning with Gaussian processes[J]. IEEE/CAA Journal of Automatica Sinica, 2014, 1(3): 227-238.

[50] Gao X, Wang X, Tao D, et al. Supervised Gaussian process latent variable model for dimensionality reduction[J]. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2011, 41(2): 425-434.

[51] Lawrence N. Probabilistic non-linear principal component analysis with Gaussian process latent variable models [J]. The Journal of Machine Learning Research, 2005, 6: 1783-1816.

[52] Lawrence N D, Moore A J. Hierarchical Gaussian process latent variable models[C]//Proceedings of the 24th International Conference on Machine Learning. Corvallis, OR, USA: ACM, 2007: 481-488.

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

[54] Cheng F, Yu J, Xiong H. Facial expression recognition in JAFFE dataset based on Gaussian process classification [J]. Neural Networks, IEEE Transactions on, 2010, 21(10): 1685-1690.

[55] Brahim-Belhouari S, Bermak A. Gaussian process for nonstationary time series prediction[J]. Computational Statistics & Data Analysis, 2004, 47(4): 705-712.

[56] Sun S, Zhong P, Xiao H, et al. Active Learning With Gaussian Process Classifier for Hyperspectral Image Classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(4): 1746-1760.

[57] Jacobs, J. P, Koziel, S. Two-Stage Framework for Efficient Gaussian Process Modeling of Antenna Input Characteristics [J]. IEEE Transactions on Antennas and Propagation, 2014, 62(2): 706-713.

[58] Chen N, Qian Z, Nabney I T, et al. Wind Power Forecasts Using Gaussian Processes and Numerical Weather Prediction[J]. IEEE Transactions on Power System, 2014, 29(2): 656-665.

[59] Markov K, Matsui T. Music Genre and Emotion Recognition Using Gaussian Processes[J]. IEEE Access, 2014, 2: 688-697.

[60] Yuan C,Neubauer C. Variational mixture of Gaussian process experts[C]// Advances in Neural Information Processing Systems 21. Cambridge:MIT Press,2008:1897-1904.

[61] Gramacy R B,Lee H K H. Bayesian treed Gaussian process models with an application to computer modeling[J]. Journal of the American Statistical Association, 2008, 103(483):1119-1130.

[62] Meeds E, Osindero S. An alternative infinite mixture of Gaussian process experts[C]// Advances in Neural Information Processing Systems 18. Cambridge: MIT Press, 2005: 883-890.

[63] Yang Y, Ma J. An efficient EM approach to parameter learning of the mixture of gaussian processes[C]// Advances in Neural Networks-ISNN 2011. Berlin Heidelberg: Springer, 2011: 165-174.

[64] 杨燕. 专家混合系统的 EM 算法研究[D]. 北京: 北京大学数学科学学院, 2011.
Yang Y. Study of the EM algorithms for the Mixture of Experts Architecture[D]. Beijing: Peking University. School of Mathematical Sciences, 2011. (in Chinese)

[65] Chen Z, Ma J, Zhou Y. A Precise Hard-Cut EM Algorithm for Mixtures of Gaussian Processes[C]// Intelligent Computing Methodologies. Switzerland: Springer International Publishing, 2014: 68-75.

[66] Yuksel S E,Wilson J N,Gader P D. Twenty years of mixture of experts[J]. Neural Networks and Learning Systems,IEEE Transactions on,2012,23(8):1177-1193.

[67] Wang Y,Khardon R. Sparse Gaussian Processes for multi-task learning[C]// Machine Learning and Knowledge Discovery in Databases. Berlin Heidelberg: Springer, 2012:711-727.

[68] Shi J Q, Murray-Smith R, Titterington D M. Bayesian regression and classification using mixtures of Gaussian processes[J]. International Journal of Adaptive Control and Signal Processing, 2003, 17(2): 149-161.

[69] Shi J Q, Murray-Smith R, Titterington D M. Hierarchical Gaussian process mixtures for regression[J]. Statistics and Computing, 2005, 15(1): 31-41.

[70] Ross J, Dy J. Nonparametric mixture of Gaussian processes with constraints[C]// Proceedings of the 30th International Conference on Machine Learning (ICML-13). [s. l.]: [s. n.], 2013: 1346-1354.

[71] Nguyen T, Bonilla E. Fast allocation of Gaussian process experts[C]// Proceedings of the 31st International Conference on Machine Learning (ICML-14). [s. l.]: [s. n.], 2014: 145-153.

[72] Lu Z. The Laplace Approximation of Gaussian Process Mixture [EB/OL]. http: // snowbird. djvuzone. org/2007/abstracts/144. pdf, 2007/2015-05-22.

[73] Huang M, Li R, Wang H, et al. Estimating Mixture of Gaussian Processes by Kernel Smoothing[J]. Journal of Business & Economic Statistics, 2014, 32(2): 259-270.

[74] Platanios E A, Chatzis S P. Mixture Gaussian Process Conditional Heteroscedasticity[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(5): 888-900.

[75] Kapoor A, Ahn H, Picard R W. Mixture of Gaussian processes for combining multiple modalities[C]// Multiple Classifier Systems. Berlin Heidelberg:Springer,2005:86-96.

[76] Schiegg M, Neumann M, Kersting K. Markov Logic Mixtures of Gaussian Processes: Towards Machines Reading Regression Data[C]// Proceedings of the 15th International Conference on Artificial Intelligence and Statistics. [s. l.]:[s. n.], 2012: 1002-1011.

[77] Wei H, Lu W, Zhu P, et al. Camera control for learning nonlinear target dynamics via Bayesian nonparametric Dirichlet-process Gaussian-process (DP-GP) models[C]// Intelligent Robots and Systems (IROS 2014). Chicago, IL: 2014 IEEE/RSJ International Conference on. IEEE, 2014: 95-102.

[78] Hernández S, Sallis P. Distributed Minimum Temperature Prediction Using Mixtures of Gaussian Processes[C]// Environmental Software Systems. Infrastructures, Services and Applications. [s. l.]: Springer International Publishing, 2015: 484-491.

[79] Ouyang R, Low K H, Chen J, et al. Multi-robot active sensing of non-stationary Gaussian process-based environmental phenomena[C]// Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems. [s. l.]: International Foundation for Autonomous Agents and Multi-agent Systems, 2014: 573-580.

[80] Nguyen-Tuong D,Seeger M,Peters J. Model learning with local Gaussian Process regression[J]. Advanced Robotics,2009,23(15):2015-2034.

[81] Liu Z, Zhou L, Leung H, et al. Kinect Posture Recon-

struction based on a Local Mixture of Gaussian Process Models[J]. IEEE Transactions on Visualization and Computer Graphics, 2015, PP(99).

[82] Sun S, Xu X. Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2011, 12(2): 466-475.

[83] Sun S. Infinite mixtures of multivariate Gaussian processes[C]// International Conference on Machine Learning and Cybernetics. Tianjin: 2013 IEEE International Conference on. IEEE, 2013: 1011-1016.

[84] Yu J, Chen K, Rashid M M. A Bayesian model averaging based multi-kernel Gaussian process regression framework for nonlinear state estimation and quality prediction of multiphase batch processes with transient dynamics and uncertainty[J]. Chemical Engineering Science, 2013, 93(19): 96-109.

[85] Ohishi Y, Mochihashi D, Kameoka H, et al. Mixture of Gaussian process experts for predicting sung melodic contour with expressive dynamic fluctuations[C]// Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: 2014 IEEE International Conference on. IEEE, 2014: 3714-3718.

[86] Wu D, Chen Z, Ma J. An MCMC based EM algorithm for mixtures of Gaussian processes[C]// Advances in Neural Networks-ISNN 2015. Berlin Heidelberg: Springer, 2015: 327-334.

[87] Chen Z, Ma J. The Hard-Cut EM Algorithm for Mixture of Sparse Gaussian Processes[C]// Intelligent Computing Methodologies. Switzerland: Springer International Publishing, 2015:13-24.

[88] Zhao L, Chen Z, Ma J. An Effective Model Selection Criterion for Mixtures of Gaussian Processes[C]// Advances in Neural Networks-ISNN 2015. Berlin Heidelberg: Springer, 2015: 345-354.

[89] Qiang Z, Ma J. Automatic Model Selection of the Mixtures of Gaussian Processes for Regression[C]// Advances in Neural Networks-ISNN 2015. Berlin Heidelberg: Springer, 2015: 335-344.

[90] Fox E B, Dunson D B. Multiresolution Gaussian Processes[C]// Advances in Neural Information Processing Systems 25. Cambridge: MIT Press, 2012: 737-745.

[91] Stachniss C, Plagemann C, Lilienthal A J, et al. Gas Distribution Modeling using Sparse Gaussian Process Mixture Models[C]// Robotics: Science and Systems. Cambridge: MIT Press, 2008: 310-317.

[92] Rasmussen C E, Ghahramani Z. Infinite mixtures of Gaussian process experts[C]// Advances in Neural Information Processing Systems 14. Cambridge: MIT Press, 2001: 881-888.

[93] Shi J Q, Wang B. Curve prediction and clustering with mixtures of Gaussian process functional regression models[J]. Statistics and Computing, 2008, 18(3): 267-283.

[94] Shi J Q, Wang B, Murray-Smith R, et al. Gaussian process functional regression modeling for batch data[J]. Biometrics, 2007, 63(3): 714-723.

[95] Shi J Q, Wang B, Will E J, et al. Mixed-effects Gaussian process functional regression models with application to dose-response curve prediction[J]. Statistics in medicine, 2012, 31(26): 3165-3177.

[96] Ma J, Liu J. The BYY annealing learning algorithm for Gaussian mixture with automated model selection[J]. Pattern Recognition, 2007, 40(7): 2029-2037.

## 作者简介

**周亚同** 男,1973 年生,湖北人,工学博士,河北工业大学电子信息工程学院教授,2013 年 9 月 ~ 2014 年 6 月在北京大学数学科学学院访问。主要研究方向为机器学习与模式识别。

E-mail:zyt@ hebut. edu. cn

**陈子一** 男,1990 年生,江西人,2015 年毕业于北京大学数学科学学院,获理学硕士学位,现为美国康奈尔大学统计系博士研究生。从事统计学习及量子力学的应用研究。

E-mail:kazy90@ 126. com

**马尽文** 男,1962 年生,陕西人,1992 年毕业于南开大学数学系,获理学博士学位。现为北京大学数学科学学院信息科学系主任、教授、博士生导师,中国电子学会信号处理分会常务委员,中国工业与应用数学学会理事。主要从事智能信息处理、神经计算、模式识别、生物信息学等方面的研究。

E-mail:jwma@ math. pku. edu. cn