

基于类间距判据的高斯过程分类模型 核参数选择方法研究

周亚同¹, 贾朋朋¹, 刘 龙²

(1 河北工业大学 信息工程学院, 天津 300401; 2 西安理工大学 自动化学院, 陕西 西安 710048)

摘 要: 文中提出了一种基于类间距判据的高斯过程分类(GPC)模型核参数选择方法. 将核参数作为自变量, 类间距作为因变量, 获得类间距随核参数变化的目标函数, 然后采用共轭梯度法求取目标函数极值, 最终获得核参数的最优值. 实验表明, 用 DBTC 作为判据进行核参数选择, 分类正确率与原有参数选择方法基本相当, 但 GPC 模型在进行参数选择时的耗时大幅减少, 因而模型训练速度得到大幅提升.

关键词: 类间距判据; 高斯过程分类模型; 核参数; 共轭梯度法

中图分类号: TP18

文献标识码: A

文章编号: 1000-7180(2012)08-0006-03

Kernel Parameter Selection of Gaussian Process Classification Model Based on DBTC Criterion

ZHOU Ya-tong¹, JIA Peng-peng¹, LIU Long²

(1 School of Information Engineering, Hebei University of Technology, Tianjin 300401, China;

2 School of Automation, Xi'an University of Technology, Xi'an 710048, China)

Abstract: This paper proposes a new kernel parameter selection method of GPC model based on DBTC criterion. After the kernel parameter is used as independent variable, and the DBTC is used as induced variable, we obtain object function that DBTC is varied with kernel parameter. Following that, conjugate gradient method is utilized to calculate the extrema of object function. Finally, the optimal value of kernel parameter is obtained. Experiments illustrates that the proposed method achieved comparable classification accuracy to traditional method. However, the time consuming in parameter section is sharply shortened. Consequently, the training speed of GPC model is improved.

Key words: DBTC criterion; Gaussian process classification model; kernel parameter; conjugate gradient method

1 引言

高斯过程(GP)模型是核机器学习与基于贝叶斯推理的学习完美结合的典范, 兼具以上两种学习方式的优势^[1]. GP 模型先验导入直观; 模型运算简洁高效; 模型的输出能提供概率信息. 当 GP 模型用于分类时, 亦被称为高斯过程分类(GPC)模型, 当其用于回归时被称为高斯过程回归(GPR)模型, 本文

主要研究前者.

近年对 GP 模型的研究逐渐发展. 这些研究大致可分为以下三类: 第一类是提升 GP 模型性能. 比如 Ranganathan 提出了 GP 模型的一种在线学习算法^[2], 李等用一种半监督算法训练模型^[3]. 第二类是将 GP 模型与现有学习方法结合, 比如王等将 GP 模型与粒子滤波相结合^[4]. 第三类是利用 GP 模型捕捉处理对象的动态特性, 比如 Wang 提出了高斯

收稿日期: 2012-01-16; 修回日期: 2012-04-27

基金项目: 国家自然科学基金(60972106); 中国博士后科学基金(20090450750); 陕西省自然科学基金(2009JQ8007); 教育部博士点新任教师基金(20096118120011)

过程动态模型,能捕获人行走时的动力学特征^[5],王等提出了一种基于 GP 模型的动作空间强化学习方法等^[6].

文献[7]中提出用一种基于类间距(DBTC)判据的参数选择方法,并用于支持向量机的核参数选择,获得了良好的效果.本文将该判据用于解决 GPC 模型的参数选择问题.

2 高斯过程分类模型原理及核参数选择

2.1 GPC 模型的基本原理

二元分类问题通过确定一个合适的映射 $f(x)$,使之能够对测试样本 x 进行正确分类, GPC 模型在解决二元分类问题时,首先假设 $f(x)$ 为一个高斯过程,然后通过贝叶斯推理获得合适的 $f(x)$.

具体原理简述如下:例如含有 n 个学习样本的集 $S\{(x_i, y_i) | y_i = \pm 1, i = 1, 2, \dots, n\}$,同时假定类标签 y_i 独立同分布且条件概率(似然函数)为 $p(y/f)$,现假设 $f(x)$ 服从均值为 0,协方差为 K 的高斯分布,即

$$f(x) \sim GP(0, K) \quad (1)$$

则 $p(f/x)$ 是一个多维高斯密度函数:

$$p(f/x) = \frac{1}{(2\pi)^{\frac{1}{2}} |K|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2} f^T K^{-1} f\right\} \quad (2)$$

根据式(1)和(2)能得到联合分布 $p(f, X)$,进而由贝叶斯公式可得到后验概率:

$$p(f/y, X) = \frac{p(y/f) p(f, X)}{p(y/X)} \quad (3)$$

式中 $p(y/X) = \int p(y/f) p(f/x) df$ 为边缘密度函数.

在分类问题中类标签 y 离散取值,因此 GPC 模型的似然函数 $p(y/f)$ 不再服从高斯分布,导致式(3)中后验概率 $p(f/y, X)$ 也不再服从高斯分布,因此只能采用拉普拉斯逼近法等数值计算手段获得其近似值.

2.2 GPC 模型的核参数选择

GPC 模型的性能取决于式(1)中核函数 K 的类型和核函数中可变参数的取值.目前常用的核函数类型有高斯核函数和平方指数核函数等,本文使用前者,其表达式为

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (4)$$

式中核参数为 σ . GPC 模型传统的核参数选择方法是将 σ 视为一个变量,并根据学习样本推导出它的后验分布表达式,再使其后验分布最大化,最后采用

求导方式获得最优值^[1].这种方法能使 GPC 模型获得高分类准确率,因此得到了广泛应用;然而采用这种方法优化核参数时计算复杂度较高.

3 基于类间距判据的 GPC 模型核参数选择

3.1 DBTC 类间距判据的定义

类间距 (Distance Between Two Classes, DBTC) 判据根据两类学习样本均值之间的距离来定义.若设两类学习样本被某未知映射 ϕ 映射至特征空间中并记为:

$$\{\phi(X_{1,i}), i = 1, 2, \dots, n_1\}$$

$$\{\phi(X_{2,i}), i = 1, 2, \dots, n_2\}$$

因为 DBTC 是两类学习样本均值之间的距离,因此其表达式可以写成

$$D(\sigma^2) = \|m_1^e - m_2^e\|^2 \quad (5)$$

式中 n_1 和 n_2 分别是两类样本的个数, m_1^e 和 m_2^e 分别代表两类样本的均值,表达式为

$$m_1^e = \frac{1}{n_1} \sum_{i=1}^{n_1} \phi(x_1, i) \quad (6)$$

$$m_2^e = \frac{1}{n_2} \sum_{j=1}^{n_2} \phi(x_2, j) \quad (7)$$

将式(6)和式(7)代入式(5)后可得:

$$D(\sigma^2) = \|m_1^e - m_2^e\|^2 = \frac{1}{n_1^2} \mathbf{1}^T K_{1,1} \mathbf{1} + \frac{1}{n_2^2} \mathbf{1}^T K_{2,2} \mathbf{1} - \frac{1}{n_1 n_2} \mathbf{1}^T K_{1,2} \mathbf{1} \quad (8)$$

式中 $|K_{1,1}|_{i,j} = k(x_{1,i}, x_{1,j})$, $|K_{2,2}|_{i,j} = k(x_{2,i}, x_{2,j})$, $\mathbf{1} = (1, \dots, 1)^T$, $|K_{1,2}|_{i,j} = k(x_{1,i}, x_{2,j})$. 由于上式已不含未知映射 ϕ , 因此可以计算出距离 D .

3.2 DBTC 判据用于核参数选择的优势

DBTC 不仅能反映两类样本间的分离程度,而且可以反映类内样本分布,因此合适用来作为核参数选择的判据.其具体优势体现在:

(1) DBTC 可同时反映类间分离程度和类内样本分布,因此能在上述两种因素间进行灵活平衡.

(2) 由于 DBTC 只与样本分布有关,因此该判据独立于分类算法本身,可用于不同的场合.

(3) DBTC 判据计算复杂度低,相应的耗时也会较少.

3.3 用 DBTC 判据选择核参数的具体过程

由于 DBTC 是两类样本间距离的度量,因此将其作为判据选择核参数 σ (为计算方便起见通常不采用 σ 而直接采用 σ^2) 时,所选择的核参数应使式(8)中的 D 取最大值.因此可以将 σ^2 作为自变量,将 D 作为因变量,获得 D 随 σ^2 变化的函数 $D(\sigma^2)$,最

后找出函数极大值对应的自变量即可。

根据对 DBTC 的分析可知, $D(\sigma^2)$ 是平滑函数且存在一个全局最大值. 本文用共轭梯度法获取函数极大值对应的自变量. 在数值仿真中为方便起见, 进一步使用 $\ln(\sigma^2)$ 代替变量 σ^2 , 这种技巧使得求解过程更加稳定, 计算量更少.

4 实验结果

为了验证本文核参数选择算法的实际应用效果, 使用标准测试样本集 banana 进行分类实验. banana 是一个典型的二元分类样本集, 学习样本个数为 400, 测试样本个数为 4 900, 无论从样本量大小还是分类复杂度等来看, 都适合用来做算法验证. GPC 模型的数值实现采用 Williams 等提供的 GPML 工具包, 该工具包在 Matlab 环境下开发, 使用传统核参数选择方法(即先推导出后验分布表达式, 再使后验最大化)^[1].

如上文所言, 基于 DBTC 判据的核参数选择方法只与待分类样本分布有关, 独立于分类算法本身. 因此使用 DBTC 判据对 banana 进行核参数选择, 如图 1 所示. 由图可看出 $\ln(\sigma^2)$ 取值为 -0.5 时对应的 $D(\sigma^2)$ 最大, 由此求出最优的 $\sigma = 0.7788$, 该值即为所求最优核参数.

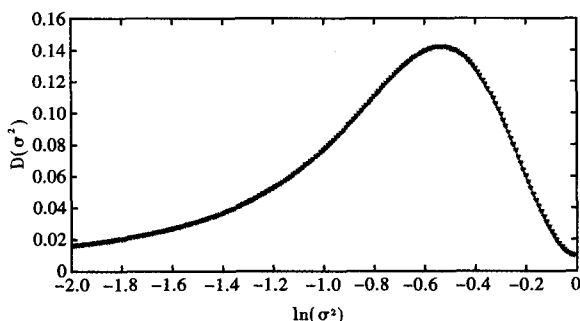


图 1 根据 DBTC 判据选择 banana 的最优核参数

将所得最优核参数代入 GPC 模型并对 banana 样本集进行分类. 将分类结果与原工具包中分类结果对比. 其中图 2 与图 3 为对 banana 集中的学习样本分类后的结果.

由于样本量很大, 单纯从上述分类图中并不能很清楚看到分类效果的差别. 为了更清楚展示两种参数选择方法在分类效果上的差别, 选用“分类正确率”和“核参数选择耗时”两个指标, 对分类结果进行定量比较. 所用测试样本为 4900 个, 仿真微机配置为 2GB 内存, AMD 2GHz 处理器, 比较结果如表 1

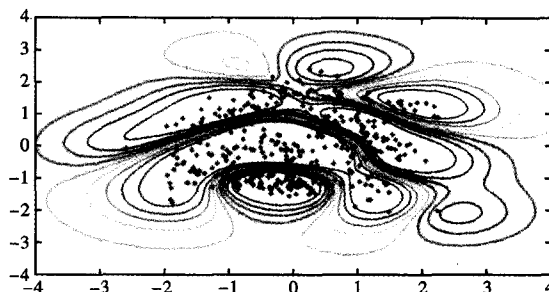


图 2 学习样本分类图(传统方法)

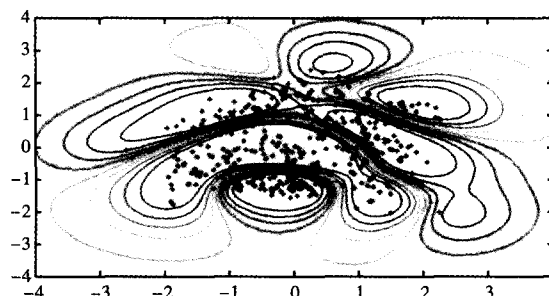


图 3 学习样本分类图(本文方法)

所示. 其中对测试样本类别的判断规则如下: 若其属于 +1 类的概率大于 0.5, 则将其判定为 +1 类, 否则将其判定为 -1 类.

表 1 两种核参数选择方法分类效果对比

核参数选择	发生分类错误的测试样本个数	核参数选择耗时(秒)
本文方法	484	20
传统方法	492	240

根据表 1 中“发生分类错误的测试样本个数”可分别统计出 GPC 模型在两种参数选择情形下的分类正确率: 若采用本文方法分类正确率为 90.1%, 若采用传统方法分类正确率为 89.9%. 由此可见两种方法分类正确率基本相当, 但本文方法耗时大幅减少. 由于参数选择是 GPC 模型训练中的重要一步, 因此本文方法能大幅提升模型训练速度.

5 结束语

本文提出了一种基于 DBTC 判据的 GPC 模型核参数选择方法. 实际分类实验表明该方法与传统 GPC 模型核参数选择方法相比具有一定优势. 二者分类准确率基本相当, 但文中方法进行核参数选择时的耗时大幅减少, 因而模型训练速度得到大幅提升, 尤为适合处理大样本数据集.

(下转第 14 页)

- on rough set and AP clustering algorithm[J]. Journal of Zhejiang University-SCIENCE C, 2012, 13(2):131-138.
- [2] Huang G B, Zhu Q Y, Siew C K. Extreme learning machine: theory and applications[J]. Neurocomputing, 2006, 70(1/3):489-501.
- [3] 张文修, 仇国芳. 基于粗糙集的不确定决策[M]. 北京: 清华大学出版社, 2005.
- [4] Pawlak Z, Wong S K M, Ziarko W. Rough Sets: Probabilistic versus deterministic approach[J]. International Journal of Man-Machine Studies, 1988, 29(1):81-95.
- [5] Duntsch I, Gediga G. Uncertainty measures of rough set prediction[J]. Artificial Intelligence, 1998, 106(1):106-109.
- [6] Wierman M J. Measuring uncertainty in rough set theory[J]. International Journal of General Systems, 1999, 28(4):283-283.
- [7] 程林凤, 胡建华. 矩阵论[M]. 徐州: 中国矿业大学出版社, 2009.
- [8] Guang-Bin Huang, Chee Kheong Siew. Extreme learning machine: RBF network case [C] // Automation, Robotics and Vision Conference. Control, 2004 (2):1029-1036.
- [9] Qin Yu Zhu, Qin A K, Suganthan P N, et al. Rapid and brief communication evolutionary extreme learning machine[J]. Pattern Recognition, 2005 (38): 1759-1763.
- [10] Guang-Bin Huang, Xiaojian Ding, Hongming Zhou. Optimization method based extreme learning machine for classification[J]. Neurocomputing, 2010(74):155-163.
- [11] Pierre Courrieu. Fast computation of moore-penrose inverse matrices[J]. Neural Information Processing, 2005, 8(2):25-29.

作者简介:

马刚 男, (1986—), 硕士. 研究方向为机器学习、数据挖掘、神经计算.

丁世飞 男, (1963—), 博士后, 教授, 博士生导师. 研究方向为人工智能、机器学习、数据挖掘、粒度计算等.

史忠植 男, (1941—), 研究员, 博士生导师. 研究方向为智能科学、知识工程、分布智能、机器学习、神经计算、认知科学.

(上接第 8 页)

参考文献:

- [1] Rasmussen C E, Williams C K I. Gaussian processes for machine learning[M]. Boston: the MIT Press, 2006.
- [2] Ranganathan A, Yang Ming-Hsuan, Ho J. Online sparse Gaussian process regression and its applications [J]. IEEE Trans. Image Processing, 2011, 20(2): 391-404.
- [3] 李宏伟, 刘扬, 卢汉清. 结合半监督核的高斯过程分类[J]. 自动化学报, 2009, 35(7): 888-895.
- [4] 王磊, 邹北骥, 彭小宁, 等. 基于高斯过程的表情动作单元跟踪技术[J]. 电子学报, 2007, 35(11): 2087-2091.
- [5] Wang J M, Fleet D J, Hertzmann A. Gaussian process dynamical models for human motion[J]. IEEE Trans. PAMI, 2008, 30(2): 283-298.
- [6] 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.
- [7] Jiancheng Sun, Chongxun Zheng, Xiaohe Li, et al. Analysis of the distance between two classes for tuning

SVM hyperparameters[J]. IEEE transactions on neural networks, 2010, 21(2):305-318.

作者简介:



周亚同 男, (1973—), 博士, 副教授. 研究方向为机器学习、模式识别.



贾朋朋 男, 硕士研究生. 研究方向为机器学习、模式识别.

刘龙 男, 副教授. 研究方向为视频图像处理.