文章编号: 1001-0920(2006)07-0767-04

基于最大熵估计的支持向量机概率建模

张 翔[1,2], 肖小玲[3], 徐光祐[1]
(1. 清华大学 计算机科学系, 北京 100084; 2. 长江大学 地球物理与石油资源学院, 湖北 荆州 434023; 3. 武汉理工大学 计算机科学与技术学院, 武汉 430063)

摘 要: 提出一种基于最大熵估计的支持向量机概率建模方法. 针对传统的支持向量机方法不能提供后验概率的输出问题, 从信息熵的角度采用最大熵估计方法, 直接对支持向量机输出进行后验概率建模. 实验结果表明, 与同类算法相比, 所提出的基于最大熵估计的概率建模方法具有优良的性能.

关键词: 支持向量机; 概率建模; 最大熵估计
中图分类号: TP18 **文献标识码**: A

Probabilistic Outputs for Support Vector Machines Based on the Maximum Entropy Estimation

ZHANG Xiang[1,2], *XIAO Xiao-ling*[3], *XU Guang-you*[1]
(1. School of Computer Science, Tsinghua University, Beijing 100084, China; 2. School of Geophysics and Oil Resources, Yangtze University, Jingzhou 434023, China; 3. School of Computer Science and Technology, Wuhan University of Technology, Wuhan 430063, China. Correspondent: ZHANG Xiang, E-mail: xiang-zhang@tsinghua.edu.cn)

Abstract: A modelling method of probabilistic outputs for support vector machines (SVM) based on the maximum entropy estimation is proposed. To the problem that the standard SVM does not provide probabilities output, the probabilistic outputs for support vector machines is modeled based on the maximum entropy estimation. Experiment results show that the proposed method achieves the better classification effect and the better posterior probability distribution than other methods.

Key words: Support vector machines; Probability modeling; Maximum entropy estimation

1 引 言

在解决样本分类的不确定性问题中, 一般对分类结果采用概率的方式输出. 传统的支持向量机方法(SVM)不能提供后验概率的输出, 在决定样本的分类类别时, 只考虑两种极端情况, 即属于某一类的概率为 1 或为 0. Wahba[1]和 Platt[2]最早在支持向量机方法中引入后验概率, 通过样本的后验概率来确定样本的类别, 使样本在分类时不仅具有定性的解释, 而且具有定量的评价.

后验概率的确定主要有两大类方法: 一类是采用Bayes 框架理论, 先求出各类的类条件概率密度, 再依据Bayes 理论求出其后验概率[3~5]. 文献[5]采用含一个方差参数的高斯函数拟合类条件概率密度的分布, 由贝叶斯方法得到样本的后验概率为 Sigmoid 函数形式. 在大多数情况下, 用含一个方差参数的高斯分布不能准确地拟合类中样本的分布规律. 当采用含方差和均值两个参数的高斯函数时, 由贝叶斯方法得到样本后验概率的函数形式不是单调函数, 违背了后验概率必须满足单调性的条件. 另外, 高斯函数的假设条件在实际过程中也不一定完

收稿日期: 2005-05-25; **修回日期**: 2005-07-19.
基金项目: 国家自然科学基金项目(60273005); 中国博士后科学基金项目(2005038310); 湖北省自然科学基金项目(2004ABA043); 湖北省教育厅科学技术研究重点项目(D200612002).
作者简介: 张翔(1969—), 男, 湖北蕲春人, 副教授, 博士后, 从事模式识别、计算机视觉的研究; 徐光祐(1940—), 男, 上海人, 教授, 博士生导师, 从事计算机视觉等研究.

全满足. 另一类方法不计算类概率密度, 直接拟合后验概率. Vapnik 将后验概率看作余弦函数和的形式[6], Platt 将后验概率看作 Sigmoid 函数的形式[7,8], 然后采用最大似然估计准则, 求出其函数的参数.

本文针对采用Bayes 框架理论对支持向量机输出进行概率建模的不足, 从信息熵的角度, 采用最大熵估计方法, 直接对支持向量机输出进行后验概率建模.

2 后验概率模型的确定

支持向量机的标准输出为

$$y = \mathrm{sign}(f(x)), \tag{1}$$

其中

$$f(x) = (w^* \times x) + b^*. \tag{2}$$

在计算过程中需要对样本进行归一化, 即对于离分类面最近的样本点(支持向量), 应满足

$$\| f(x) \| = 1. \tag{3}$$

显然, 对于分类面上的样本点, 有

$$f(x) = 0, \tag{4}$$

对于其他的样本点, 有

$$f(x) = \pm d \| w \|. \tag{5}$$

其中: d 表示样本点 x 到分类面之间的距离, 正负号表示该样本点在分类面的两侧. 则任意样本点 x 到分类面之间的距离为

$$d_x = f(x) / \| w \|, \tag{6}$$

支持向量到分类面之间的距离为

$$d_{sv} = 1 / \| w \|. \tag{7}$$

从支持向量机的分类超平面的几何角度, 如图 1 所示, 可通过样本与最优分类面之间距离的远近, 定量地评价两类分类问题中样本属于所在类程度的大小. 由式(6) 和(7) 可以看出, $f(x)$ 是 d_x 与 d_{sv} 的比率, 可通过支持向量机方法的标准输出 $f(x)$ 来度量样本的后验概率, 因此后验概率模型可看作 $f(x)$ 函数.

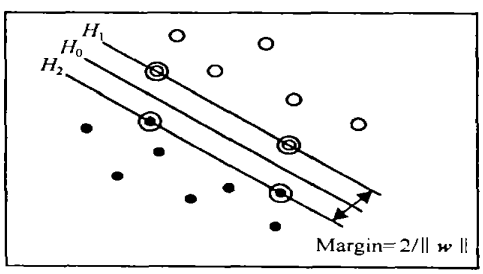

图 1 最优分类面示意

由文献[2] 可知, 要成为概率输出函数必须满足以下条件: 1) 函数的值域为[0, 1]; 2) 函数满足单调性.

依据文献[9] 对几种单调性函数作为概率输出函数的对比分析, 在支持向量机的输出概率建模中, 含两个参数 A 和 B 的 Sigmoid 函数具有灵活的函数形式, 在实际应用中表现出较好的分类精度. 因此, 本文采用含两个参数 A 和 B 的 Sigmoid 函数作为后验概率模型.

在两类分类问题中, 采用含两个参数 A 和 B 的 Sigmoid 函数, 给出支持向量机的概率输出形式

$$p(y = 1 \mid f(x)) = \frac{1}{1 + e^{Af(x)+B}}, \tag{8}$$

$$p(y = -1 \mid f(x)) = 1 - p(y = 1 \mid f(x)). \tag{9}$$

其中: 参数 A 和 B 控制 Sigmoid 函数的形态, $f(x)$ 为支持向量机中样本 x 的标准输出值.

可以看出, 对传统支持向量机进行概率建模后, 通过式(8) 和(9) 确定样本 x 的类别, 概率的大小也提供了样本属于所在类程度的大小. 而在传统的支持向量机方法中, 是通过式(1) 中 $y = 1$ 或 $y = -1$ 来确定样本 x 的类别.

3 基于最大熵估计的模型参数的确定

通过支持向量机方法的标准输出 $f(x)$ 建立 Sigmoid 函数的后验概率模型后, 需要确定概率模型参数 A 和 B. 本文从信息熵的角度, 采用最大熵估计方法确定 Sigmoid 函数中的参数.

熵的概念来源于香农信息理论[10], 它描述系统的不确定性, 是关于事件的不确定因素的度量方法. 一个概率分布 $P = (p_1, p_2, \cdots, p_n)$ 的熵定义为

$$H(P) = E(-\log P) = -\sum_{i=1}^{n} p_i \log(p_i). \tag{10}$$

当各信号出现的概率相等时, 式(10) 的熵达到最大. 此时, 系统给每个输出提供最大可能的平均信息量.

设训练样本集为 (f_i, y_i), $i = 1, 2, \cdots, n$ (为了方便将 $f(x_i)$ 记为 f_i). 其中: $y_i \in \{-1, 1\}$, $y_i = 1$ 类的样本数为 N_+, $y_i = -1$ 类的样本数为 N_-. 为避免采用小数据集拟合 Sigmoid 函数时出现过拟合现象, 最简单的方法是在原始数据集中加入高斯噪声[6]. 在重新构造的训练样本集中, 正样本的支持向量机输出值为 $f(x_i)$, 对应的目标值 $t_i = 1 - \epsilon_+$, 而没有令 $t_i = 1$, 这里假设了同一值 $f(x_i)$, 还存在一些对应的负样本; 同理, 负样本对应的目标值 $t_i = \epsilon_-$. 利用 Bayes 后验概率估计 $\epsilon_+ = \frac{1}{N_+ + 2}$ 和 $\epsilon_- = \frac{1}{N_- + 2}$, 即得到重新定义的一组新的训练样本 (f_i, t_i), $i = 1, 2, \cdots, n$. 其中

$$t_i = \begin{cases} \dfrac{N_+ + 1}{N_+ + 2}, & y_i = 1; \\ \dfrac{1}{N_- + 2}, & y_i = -1. \end{cases} \tag{11}$$

通过基于最大香农信息熵对支持向量机后验概率的估计, 即通过对新训练样本集的香农信息熵的最大化, 求取 Sigmoid 函数的参数A 和B. 将求解参数A 和B 写为向量形式, 用向量Z 表示为 $Z = (A, B)^{\mathrm{T}}$, 即最大化下式:

$$\max_{Z=(A,B)^{\mathrm{T}}} F(Z). \tag{12}$$

其中

$$F(Z) = -\sum_{i=1}^{n} t_i p_i \log(p_i) + (1 - t_i)(1 - p_i)\log(1 - p_i), \tag{13}$$

$$p_i = \frac{1}{1 + \mathrm{e}^{Af_i + B}}. \tag{14}$$

采用具有逆向线性搜索特点的牛顿迭代方法[8], 对式(12) 求解参数A 和B. 即将式(12) 转换为迭代求解

$$H(Z) + \sigma I = -\nabla F(Z). \tag{15}$$

其中: $H(Z)$ 为 $F(Z)$ 的 Hessian 矩阵, $\nabla F(Z)$ 为 $F(Z)$ 的梯度

$$\nabla F(Z) = \begin{bmatrix} \sum_i [t_i(1 + \log(p_i)) - (1 - t_i) \times (1 + \log(1 - p_i))]p_i(1 - p_i)f_i \\ \sum_i [t_i(1 + \log(p_i)) - (1 - t_i) \times (1 + \log(1 - p_i))]p_i(1 - p_i) \end{bmatrix}, \tag{16}$$

$$H(Z) = \nabla^2 F(Z) = \begin{bmatrix} H_{AA} & H_{AB} \\ H_{BA} & H_{BB} \end{bmatrix}. \tag{17}$$

其中

$$H_{AA} = \sum_i [(1 - p_i)t_i + p_i(1 - t_i) + [t_i(1 + \log(p_i)) - (1 - t_i)(1 + \log(1 - p_i))](1 - 2p_i)](1 - p_i)p_i f_i^2,$$

$$H_{BB} = \sum_i [(1 - p_i)t_i + p_i(1 - t_i) + [t_i(1 + \log(p_i)) - (1 - t_i)(1 + \log(1 - p_i))](1 - 2p_i)](1 - p_i)p_i,$$

$$H_{AB} = H_{BA} = \sum_i [(1 - p_i)t_i + p_i(1 - t_i) + [t_i(1 + \log(p_i)) - (1 - t_i)(1 + \log(1 - p_i))](1 - 2p_i)](1 - p_i)p_i f_i.$$

4　实验结果及分析

采用 Heart-scale 数据, 对支持向量机进行概率输出的实验. 实验数据为两类分类问题, 总样本数为 270. 其中正样本数为 120, 负样本数为 150, 数据特征维数为 13. 采用交叉验证法对训练和拟合数据集进行选择, 将原始数据集分为 3 等份, 其中任意 2 份用于训练支持向量机分类器, 余下的 1 份作为拟合后验概率函数的数据集. 重复 3 次训练和拟合, 将求出的 3 次A 和B 参数的平均值作为最终参数A 和B 的值.

采用基于最大熵估计方法对支持向量机进行概率建模实验, 将得到的后验概率与 Platt 的概率建模得到的后验概率进行对比, 两种建模得到的后验概率曲线如图 2 所示. 利用获得的后验概率对 Heart-scale 数据进行分类预测, 采用支持向量机硬输出方法和两种概率建模方法预测的错误率如表 1 所示.

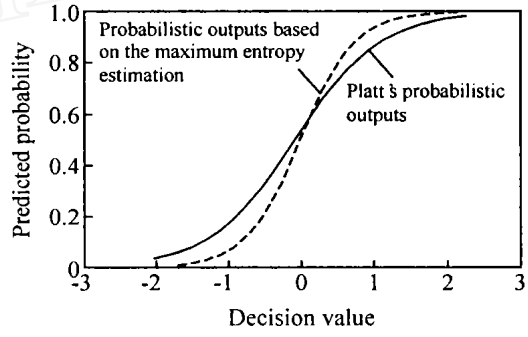

图 2　基于最大熵估计建模和 Platt 概率建模两种后验概率分布

表 1　3 种输出方法预测错误率对比

方　法	分类错误率/%	参　数
标准支持向量机	18.15	
Platt's 概率建模	14.8	$A = -1.706\,948$ $B = -0.165\,390$
最大熵估计的概率建模	13.7	$A = -2.665\,771$ $B = -0.739\,645$

由表 1 可以看出, 采用支持向量机硬输出方法, Platt 的概率建模和基于最大熵估计的概率建模分类预测的错误率分别为 18.5%, 14.8% 和 13.7%, 可见采用支持向量机概率建模输出方法提高了支持向量机的分类性能. 与 Platt 的概率建模方法相比, 本文提出的基于最大熵估计的概率建模方法具有优良的性能.

5　结　论

传统的支持向量机方法不能提供后验概率的输出. 本文针对采用 Bayes 框架理论对支持向量机输出进行概率建模的不足, 从信息熵的角度, 采用最大熵估计方法, 直接对支持向量机输出进行后验概率建模. 该方法不仅使支持向量机的分类精度得到了

提高, 而且提供了样本属于所在类中的可信程度 实验结果表明, 与 Platt 的概率建模方法相比, 本文提出的基于最大熵估计的概率建模方法具有优良的性能 在下一步工作中, 作者将对支持向量机方法中多类分类问题进行概率建模的研究

参考文献(References)

[1] Wahba G. Support Vector Machines, Reproducing Kernel Hilbert Spaces and the Randomized GACV [A]. *Advances in Kernel Methods Support Vector Learning* [C]. Massachusetts: MIT Press, 1999: 69-88.

[2] Platt J C. Probabilities for Support Vector Machines [A]. *Advances in Large Margin Classifiers* [C]. Massachusetts: MIT Press, 2000: 61-74.

[3] Sollich P. Bayesian Methods for Support Vector Machines: Evidence and Predictive Class Probabilities [J]. Machine Learning, 2002, 46: 21-52.

[4] Kwok J T Y. Moderating the Outputs of Support Vector Machine Classifiers[J]. *IEEE Trans on Neural Networks*, 1999, 10(5): 1018-1031.

[5] Hastie T, Tibshirani R. Classification by Pairwise Coupling [J]. *The Annals of Statistics*, 1998, 26(1): 451-471.

[6] Vapnik V. *Statistical Learning Theory* [M]. New York: Wiley, 1998.

[7] 张文生, 王珏, 戴国忠. 支持向量机中引入后验概率的理论和方法研究[J]. *计算机研究与发展*, 2002, 39(4): 392-397.
(Zhang W S, Wang J, Dai G Z. Study of Theory and Method Introducing Posteriori Probability into Support Vector Machines [J]. *J of Computer Research and Development*, 2002, 39(4): 392-397.)

[8] Lin H T, Lin C J, Weng R C. *A Note on Platt's Probabilistic Outputs for Support Vector Machines*[R]. National Taiwan University, Taibei, 2003.

[9] 张翔. *支持向量机及其在医学图像分割中的应用*[D]. 武汉: 华中科技大学, 2004.
(Zhang X. *Support Vector Machine and Its Applications in Medical Image Segmentation*[D]. Wuhan: Huazhong University of Science and Technology, 2004.)

[10] Theodoridis S, Koutroumbas K. *Pattern Recognition* [M]. Amsterdam: Elsevier Press, 2003.

(上接第 766 页)

5 结 论

本文分析了用蚁群算法求解 TSP 问题的收敛性 得到的结论是: 若全局最优解属于严格趋近的封闭路径, 当设定 $q_0 = 1$ 时, 算法一定能快速收敛于最优解; 当设定 $q_0 < 1$ 时, 会使收敛速度变慢 若全局最优解不属于严格趋近的封闭路径, 当设定 $q_0 = 1$ 时, 算法不能收敛到最优解; 当设定 $0 < q_0 < 1$ 时, 算法能够收敛到最优解, 但收敛时间将很长 在一定范围内加大优化路径上的信息素, 会加快收敛 根据这一结论, 在算法设计时, 应估计或分析优化解的性质, 据此合理地设定 q_0, 并在一定范围内加大全局优化路径上的信息素, 以利于提高收敛速度

通过本文的分析可以看出, 影响蚁群算法收敛速度的两大因素是信息素和启发函数 选择合适的启发函数对提高收敛速度至关重要, 启发信息的值过大, 将抑制信息素的作用; 启发信息的值过小, 则会影响收敛速度 通过分析证明, 采用最近邻选择策略可以提高收敛速度

参考文献(References)

[1] Colorni A, Dorigo M, Maniezzo V. Distributed Optimization by Ant Colonies [A]. *Proc of ECAL'91 European Conf on Artificial Life* [C]. Paris: Elsevier Publishing, 1991: 134-144.

[2] Thomas Stutzle, Marco Dorigo. A Short Convergence Proof for a Class of Ant Colony Optimization Algorithms [J]. *IEEE Trans on Evolutionary Computation*, 2002, 6(4): 358-365.

[3] Amr Badr, Ahmed Fahmy. A Proof of Convergence for Ant Algorithms [J]. *Information Sciences*, 2004, 160: 267-279.

[4] Walter J. Gutjahr. A Graph-based Ant System and Its Convergence [J]. *Future Generation Computer Systems*, 2000, 16: 873-888.

[5] Walter J Gutjahr. ACO Algorithms with Guaranteed Convergence to the Optimal Solution [J]. *Information Processing Letters*, 2002, 82: 145-153.

[6] 朱庆保, 杨志军. 基于变异和动态信息素更新策略的蚁群算法[J]. *软件学报*, 2004, 15(2): 185-192.
(Zhu Q B, Yang Z J. Ant Colony Optimization Algorithm Based on Mutation and Dynamic Pheromone Updating[J]. *J of Software*, 2004, 15(2): 185-192.)