

# 支持向量分类机的参数选择方法研究

向昌盛<sup>1</sup>, 周子英<sup>2</sup>

(1. 湖南农业大学 东方科技学院, 湖南 长沙 410128;

2. 湖南农业大学 资环学院, 湖南 长沙 410128)

**摘 要:**支持向量分类机(Support Vector Classification, SVC)的参数选择一直缺乏一种通用、完善的方法,很大程度上限制了它的应用。为解决 SVC 参数选择的难题,提出了一种基于启发式深度优先搜索(Heuristic Depth-first Search, HDFS)的 SVC 参数自动寻优方法。该方法将 10-fold 交叉验证的最大识别率作为目标,利用 HDFS 算法进行 SVC 参数寻优,减少了 SVC 的训练时间,提高了分类的精度,从而确保了 SVC 参数选择的准确性。将该算法用于 3 个基准数据集的仿真实验,结果表明该方法在保证分类精度前提下,大幅度缩短了训练建模时间,提高了运行效率,具有一定的推广意义。

**关键词:**支持向量分类机;深度优先搜索;核函数;交叉验证

中图分类号:TP181

文献标识码:A

文章编号:1673-629X(2010)09-0094-04

## Parameters Selection Method for Support Vector Classification

XIANG Chang-sheng<sup>1</sup>, ZHOU Zi-ying<sup>2</sup>

(1. Orient Science & Technology College of Hunan Agricultural University, Changsha 410128, China;

2. College of Resources & Environment of Hunan Agricultural University, Changsha 410128, China)

**Abstract:** There have been no perfect algorithms for the selection of the optimal parameters of support vector classification (SVC), therefore, the applications of SVC are limited. In order to get optimal SVM parameter, a parameter selection method for SVC based on heuristic depth-first search (HDFS) is proposed in this paper. In this method, the ten-fold cross-validation recognition rate is used as the classification objection and HDFS is used for parameter selection, which can reduce the train time, improve the precision of SVC, and insure the accuracy of parameter selection. Results on 3 benchmark datasets show that the new method not only can assure the classification precision but also can reduce training time markedly. The new method has certain practical application significance.

**Key words:** support vector classification; depth first search; kernel function; cross-validation

## 0 引言

支持向量机是 Vapnik 等<sup>[1]</sup>提出的一种基于统计学习和结构风险最小化原理的新型机器学习算法,具有好的泛化能力,是一种性能优异的小样本学习机,与神经网络相比,支持向量机的训练算法不存在局部极小和维数灾问题,具有自动设计模型复杂度和泛化能力强等优点,已被成功地用于优化控制、模式识别和金融预测<sup>[2~5]</sup>等领域。虽然支持向量机的应用领域越来越广泛,但其在应用过程中仍有许多疑难问题目前没有得到很好的解决,其中支持向量机算法参数的具体如何选取就是其一。支持向量机是基于统计学习理论的一种机器学习方法,其在学习训练时,支持向量机

算法参数对模型的优劣有着决定性作用,所以说其选择也就是对模型的选择,因为支持向量机训练出来的是每一类模型,都有其相应的一组参数与其对应,模型的参数是唯一确定的。但支持向量机在实际应用过程中存在参数确定没有统一的标准,参数寻优没有规律可循,目前传统的参数寻优方法采用穷尽式搜索法,其耗时相当长。在实际应该过程中,由于具体学习对象的不同,导致学习对象的特征不同,这样模型的寻优每次得到的结果不一样,没有什么永恒的规律。目前最流行的支持向量机参数寻优算法为网格算法,其方法就是首先给定各参数的区间,定义一个步长在区间内穷尽搜索逐个试验,找到使得算法在学习对象上性能最好的参数,参数的寻优不是并行的进行,导致得到的单个最优参数组合在一起不是最优的,比较耗时。现在也有许多学者对支持向量机参数寻优的方式进行了大量的探索<sup>[6~9]</sup>,这些方法大多数都是采用一些智能类算法如遗传算法、模拟退火、蚁群算法等进行启发式

收稿日期:2010-01-02;修回日期:2010-04-12

基金项目:教育部新世纪优秀人才支持计划(NCET-07-0711)

作者简介:向昌盛(1971-),男,湖南怀化人,高级讲师,博士,研究方向为人工智能和模式识别。

搜索。这些算法在参数空间内进行参数寻优时,尽量在保证能够找到最优或次优的参数组合的同时搜索次数尽可能少,来缩短算法学习的训练时间,降低算法整体时间复杂度,但是这些方法都有一个容易陷入局部最优的缺点,每次寻优的参数都不一样,具有随机性,无法满足通用性要求。

针对 SVM 参数寻优问题,本研究提出一种基于启发式深度优先搜索(Heuristic Depth - first Search, HDFS)算法的 SVM 参数寻优方法——HDFS - SVM,利用 HDFS 强大的启发式全局快速寻优能力进行 SVM 参数的优化,对 3 个基准数据集进行仿真试验。仿真结果表明,HDFS - SVM 模型不仅能够找到最优解,并且在 HDFS 进行参数寻优时,SVM 参数之间并行进行,大大加快了寻优效率,缩短了训练时间,降低算法时间复杂度。

## 1 SVM 概述

### 1.1 SVM 分类原理

对于数据分类问题,支持向量机的原理就是寻找一个超平面,尽量使该平面能够满足限制的条件,并且可以把数据集中的点分开,使之尽可能地远离该超平面<sup>[10]</sup>。给定数据集:  $(x^i, y^i), x \in R^n, y \in \{-1, 1\}, i = 1, 2, \dots, l$ , 当  $x_i$  属于正类集合时,输出  $y_i = 1$ , 不然  $y_i = -1$ 。SVM 的目标是根据风险最小的原理,该超平面可以描述为:

$$w^T \Phi(x) + b = 0 \quad (1)$$

其中  $w$  为超平面的法向量,  $b$  为超平面的偏移量。

对线性不可分的问题,为了使得训练集当中的任意数据点向量离超平面最短距离最大的最优超平面,要解决的问题就变成了二次优化问题:

$$\min J(w, \xi) = \frac{1}{2} \|w\|^2 + c \sum_{i=1}^n \xi_i \quad (2)$$

约束条件为:

$$\begin{aligned} y_i(w \cdot \Phi(x_i) + b) &\geq 1 - \xi_i \\ \xi_i &\geq 0, i = 1, 2, \dots, l \end{aligned} \quad (3)$$

其中  $c > 0$  为惩罚参数,其对错分样本的惩罚程度进行控制,  $\xi = (\xi_1, \dots, \xi_l)^T$ 。

对于大样本问题,通过引入 Lagrange 因子将支持向量机处理的分类问题转化为其对偶问题,通过对偶问题来解决这个超平面优化问题,最终得到如下的超平面判别函数:

$$f(x) = \text{sign}(\sum_{i=1}^l \alpha_i y_i (\Phi(x) \cdot \Phi(x_i)) + b) \quad (4)$$

其中  $\alpha_i (i = 1, 2, \dots, s, s \leq l)$  为 Lagrange 因子。

根据泛函的有关理论,只要满足 Mercer 条件的函

数都可以作为支持向量机的核函数,这样就可以通过找到一个满足条件的核函数  $K(x_i, x)$  来代替点积  $(\Phi(x) \cdot \Phi(x_i))$ , 这样 SVM 的判别函数就变成如下形式:

$$f(x) = \text{sign}(\sum_{i=1}^l \alpha_i y_i k(x_i \cdot x) + b) \quad (5)$$

### 1.2 支持向量机的核函数及参数选择

核函数是支持向量机的重要组成部分,具有不同核函数和参数的支持向量机的性能存在很大差异,所以如何根据数据集来构造核函数,通过构造核函数来提高支持向量机的分类能力,这是目前支持向量机技术发展的一个重要研究方向,目前在支持向量机中研究和应用最多的四类核函数有:

(1) 线性核函数:

$$k(x_i, x_j) = x_i \cdot x_j \quad (6)$$

(2) 多项式核函数:

$$k(x_i, x_j) = (x_i \cdot x_j + 1)^d \quad (7)$$

(3) 高斯核函数:

$$k(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2}) \quad (8)$$

(4) sigmodi 核函数:

$$k(x_i, x_j) = \tanh[b(x_i \cdot x_j) + c] \quad (9)$$

不同的核函数,其分类器的性能可能完全不同,但对于某一问题,如何选择核函数的形式还没有一个指导原则,很多研究者进行了各种各样的尝试。很多研究和实验表明<sup>[11,12]</sup>,当缺少过程的先验知识时,选择高斯核函数比其它核函数常能为实际问题提供满意的结果,因此,本研究采用的核参数为高斯核函数。

选定核函数后,接下来要解决的问题就是支持向量核函数参数和控制结构风险参数的确定,这是支持向量机中的一个难题,因为到目前为止还没有成熟的理论基础和统一的标准,但是参数的寻优从模型的角度来分析实际上也属于一类学习模型确定过程,一个学习模型性能的好坏是由学习能力和预测能力两个方面决定的,这样就需要在一类模型中确定具体的一个最优模型。目前传统的方法基本上采用试凑的方法来进行,就是在一个给定的参数空间内进行穷尽或网络寻优,按照一定的标准选择一组最优的参数组合。模式识别领域,通常采用 k-折交叉验证方法<sup>[13,14]</sup>(k-fold Cross Validation, k-fold CV)来评估分类模型的泛化性能,但问题是究竟 k 取何值时能够获得最佳的分类准确率,或者说是哪种 k-折交叉验证分类准确率更能客观反映分类模型的泛化性能。对于小样本,许多文献采用留一法(Leave-one-out Cross Validation, LOOCV)来评估支持向量机分类模型,但是 Breiman

等<sup>[15]</sup>为 5 折或 10 折交叉验证方法优于留一法,本研究采用 10 折交叉验证方法。

### 1.3 SVM 模型的评估

分类器算法性能的优劣主要由其分类准确率或回归系数以及泛化能力进行评价,各种参数选择方法优劣的评判依据也是算法性能,一般采用分类器的分类正确率来衡量。由于分类器的泛化性能涉及它在独立测试集上的预测能力,因此,分类器性能评估方法在实际的分类器设计中非常重要。为了验证模型的测试效果,以测试样本集数据的识别率(recognition rate, RR)作为模型评价标准。

识别率定义为:

$$RR = \frac{\text{测试样本集中分类正确的样本数目}}{\text{测试样本集中总的样本数目}} \quad (10)$$

## 2 基于 HDFS 寻优的 SVM 模型

### 2.1 HDFS-SVM 模型原理

支持向量机模型分类精度与惩罚因子  $C$  和核函数均存在一定的关系,为了获取最佳分类性能的 SVM 模型,需要得到最佳的  $C$  和核函数值  $\sigma$ 。搜索最佳参数对  $(C, \sigma)$  是一件非常耗时的工作,如果采取穷举的方式搜索最优值,计算量将十分巨大,甚至无法实现。由于 HDFS 具有启发式、强大全局搜索能力,可以在很短的时间内搜索到全局最优点,本研究采用 HDFS 来进行 SVM 分类模型的参数优化。

### 2.2 HDFS-SVM 算法的具体步骤

HDFS-SVM 参数  $(C, \sigma)$  选择方法的具体步骤描述如下(见图 1):

Step1: 设定参数  $(C, \sigma)$  范围,例如:  $C$  的选择范围  $2^{-10} \sim 2^{15}$ , 步长为 1,  $\sigma$  为  $2^{10} \sim 2^{-15}$ , 步长为 -1;

Step2: 对每一组参数用 LIBSVM 进行预测估计。将训练集平均分成 10 个子集,先随机选择其中 9 个子集作为训练集,得到 1 个判别函数,把没有参加训练的子集作为测试集,用这个判别函数进行预测。这样重复进行 10 次,直到所有的子集都作为测试样本被预测一遍。取 10 次预测所得准确率 RR 的平均值作为最终的准确率 RR;

Step3: 最佳参数和上一轮的最佳参数的 RR 进行比较,如果 RR 的最佳结果比上一轮变小了,则进入 Step4,否则进行 Step5;

Step4: 根据上一轮的参数范围,根据深度优先策略启发式的原则在该参数附近重新定义选择范围。参数范围修改方式:原步数按  $((\text{stepnum} - 2) > 1) + 2$  减少,stepnum 是上阶段的步数,步数降低的同时又保证最小为 2。参数范围上限 = 最优参数 - step, 下限 =

最优参数 + step, step 是上阶段的步长。修改各参数范围后转 Step1, 重新递归运算;

Step5: 退出运算,此时参数  $(C, \sigma)$  为最优参数。

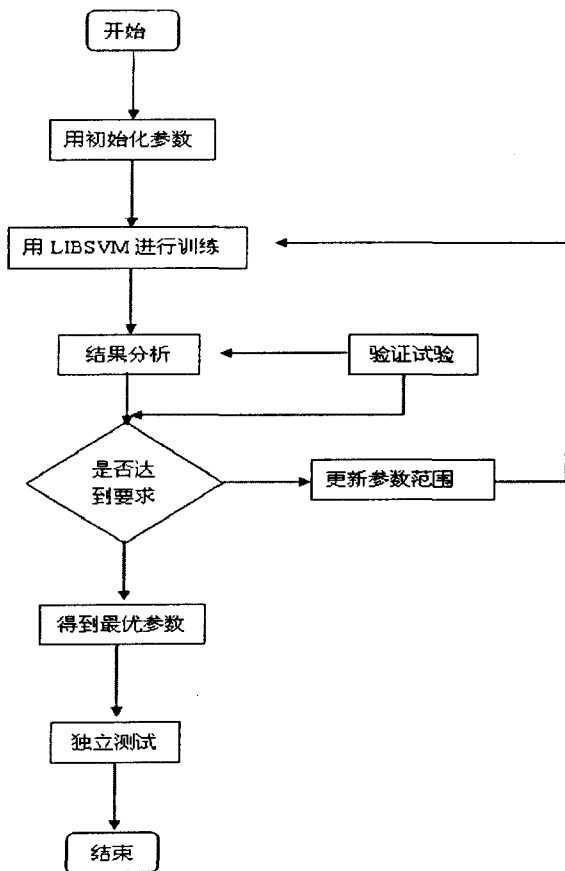


图 1 HDFS-SVM 流程图

## 3 HDFS-SVM 实验仿真

### 3.1 实验环境及数据

在 IBM System x3800 上,利用 LIBSVM 软件工具包和所编制的 HDFS 程序进行实验。所采用的数据集为 <http://www.csie.ntu.edu.tw/~cjlin/LIBSVM-tools/datasets/> 中的 a2a、diabetes、w1a 等 3 个经典分类数据。把上述基准数据划分为训练集和测试集两部分,采用 10 折交叉检验进行测试,这样,既能保证提供足够的数据用来模型的学习,又能够验证模型的效果。参比模型为:LIBSVM 和基于遗传算法的 SVM(GA-SVM)和基于粒子群的 SVM(PSO-SVM)。

### 3.2 实验结果与分析

对 HDFS-SVM、LIBSVM 和 GA-SVM 和 PSO-SVM 进行建模预测,实验结果见表 1 和表 2(其中时间为训练时间,不包括预测时间,单位为秒)。

#### 3.2.1 各种算法的时间复杂度

在给定的实验条件下,当训练的规模达到一定程度时,LIBSVM 的训练就无法完成,如在上述实验条件

下,象对于 w8a(训练样本数为 49749)的大样本数据,LIBSVM 会用上几天的时间去训练,与现代数据处理的要求相差太远,从而导致支持向量机对大样本问题无从下手,成为支持向量机发展的一个瓶颈。对于 w1a 和 a2a 数据集,采用传统的穷尽式搜索即使可以完成,但是由于训练时处理的是一个  $N \times N$  矩阵( $N$  为训练集的样本个数),耗费的代价与样本数的平方成正比,耗费的时间也相当的长。由于 HDFS-SVM 改变了参数搜索策略,从表 1 可知,对于 3 个数据集, HDFS-SVM 预测时间比 LIBSVM 预测花费的时间少几十多倍,相对于 GA-SVM 和 PSO-SVM 速度也提高不少,所以说 HDFS-SVM 大大降低时间复杂,提高了学习速度。

表 1 HDFS-SVM 同 LIBSVM、GA-SVM、  
PSO-SV 训练时间

数据名称	样本总数	训练集	测试集	LIBSVM	GA-SVM	PSO-SVM	HDFS-SVM
a2a	32461	2265	30296	48091	5008	5987	840
diabetes	768	300	468	7650	1112	1727	300
w1a	49749	2477	47272	65887	6012	9989	950

表 2 HDFS-SVM 同 LIBSVM、GA-SVM、  
PSO-SV 精度(RR%) 结果

数据名称	样本总数	训练集	测试集	LIBSVM	GA-SVM	PSO-SVM	HDFS-SVM
a2a	32461	2265	30296	82.8393	84.427	82.8393	84.4567
diabetes	768	300	468	76.823	79.9145	78.8462	80.7692
w1a	49749	2477	47272	97.039	97.3705	97.7217	98.7583

3.2.2 各种算法的分类精度比较

从表 2 可以看出,对于 a2a 数据集,LIBSVM 和 PSO-SVM 的 RR 相差不大,HDFS-SVM 和 GA-SVM 优于两者,其中 HDFS-SVM 的效果最好;而对于 w1a 和 diabetes 数据集,HDFS-SVM 精度要明显高于其它算法,所以 HDFS-SVM 的分类精度不仅没有因算法时间复杂度减小、训练速度提高而降低,反而提高了,这表明 HDFS-SVM 的参数寻优策略是合理的、可行的。所以,无论从学习能力和时间复杂度综合来看, HDFS-SVM 是个不错的选择,体现了其优势。

4 结束语

SVM 是一种具有严格理论基础的机器学习方法,但在实际应用过程中存在参数确定没有统一的标准、传统的参数寻优方法采用穷尽式搜索法耗时的难题,从而影响了 SVM 的应用范围。为此本研究通过引入 HDFS 搜索算法机制来解决 SVM 参数寻优这一难题,并用 3 个基标数据集对该算法进行了仿真实验。

结果表明 HDFS 搜索算并行的对  $C$  和  $\sigma$  2 个参数同时进行寻优,通过判别函数来确定最终获取的最优参数组合,得到参数组合是使分类准确率达到最优,同时在寻优过程中各组参数相互耦合、并行进行,大大缩短了 SVM 参数的时间,为 SVM 在实际应用中解决参数寻优的难题提供了一种十分好的解决方案,使 SVM 的应用范围进一步拓宽。

参考文献:

[1] 张学工.关于统计学习理论与支持矢量机[J].自动化学报,2000,26(1):32-34.

[2] Vladimir C, Yun Q M. Practical selection of SVM parameters and noise estimation for SVM regression[J]. Neural Networks,2004,17(1):113-126.

[3] Burges C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery,1998,2(2):121-127.

[4] Sanchez A D. Advanced support vector machines and kernel methods[J]. Neuro Computing, 2003, 55(1):5-20.

[5] Francis E H Tay, Cao L J. Application of Support Vector Machines in Financial Time Series Forecasting[J]. The International Journal of Management Science,2001,29:309-317.

[6] 田 鹏. 改进粒子群算法在支持向量机训练中的应用[J]. 自动化技术与应用,2009,28(3):5-8.

[7] Chapelle O, Vapnik V. Choosing multiple parameters for support vector machines[J]. Machine Learning,2002,46(1):131-160.

[8] 周红刚,杨春德.基于免疫算法与支持向量机的异常检测方法[J].计算机应用,2006,26(9):2145-2147.

[9] 李良敏,温广瑞,王生昌.基于遗传算法的回归型支持向量机参数选择法[J].计算机工程与应用,2008,44(7):23-26.

[10] 邓乃扬,田英杰.数据挖掘中的新方法—支持向量机[M].北京:科学出版社,2004.

[11] 冯兴杰,魏 新,黄亚楼.基于支持向量机的旅客吞吐量预测研究[J].计算机工程,2005,31(14):172-173.

[12] 杨俊燕,张优云,赵荣珍.支持向量机在机械设备震动信号趋势预测中的应用[J].西安交通大学学报,2005,39(9):950-953.

[13] Wang W, Xu Z, Lu W, et al. Determination of the spread parameter in the Gaussian kernel for classification and regression[J]. Neurocomputing,2003,55:643-663.

[14] Blum A L, Langley P. Selection of relevant features and examples in machine learning[J]. Artificial Intelligence,1997,97:245-271.

[15] Breiman L, Spector P. Submodel selection and evaluation in regression: the X-random case[J]. International Statistical Review,1992,60(3):291-319.