

一种支持向量机参数优化的 GA-Powell 算法

王萍萍,王翰虎

(贵州大学 计算机科学与信息学院,贵州 贵阳 550025)

摘要:支持向量机的核心是核函数,选择合适的核函数参数是支持向量机理论研究的重点。文中将遗传算法与 Powell 算法相结合,提出了 GA-Powell 算法来优化核函数的参数。首先利用遗传算法找到一个初始最优解,再利用 Powell 算法在所得解附近进行寻优,反复迭代产生最优解。该算法在保留遗传算法较强的全局搜索能力的同时具有 Powell 算法的较强的局部搜索能力,使得混合算法具有更加精确和快速的收敛性。将该算法应用到银行基金项目的分类实验中取得了良好的结果。

关键词:支持向量机;参数优化;遗传算法;Powell 算法

中图分类号:TP301.6

文献标识码:A

文章编号:1673-629X(2013)02-0015-04

doi:10.3969/j.issn.1673-629X.2013.02.004

A GA-Powell Algorithm for Parameter Optimization of Support Vector Machine

WANG Ping-ping, WANG Han-hu

(College of Computer Science and Information, Guizhou University, Guiyang 550025, China)

Abstract:Support vector machine (SVM) is the core of the kernel function, selecting the proper parameters of kernel function are the focus of support vector machine theory research. Combined GA and Powell algorithm, the GA-Powell algorithm was proposed to search optimal parameters. First, use the genetic algorithm to find a initial solution, and then use the Powell algorithm, eventually produce the optimal solution. This method retained the global search capability of GA algorithm and the good local convergence of Powell, with more accurate and faster convergence. The algorithm is applied to the practice of bank found project classification and has good result.

Key words:SVM; parameter optimization; GA; Powell algorithm

0 引言

支持向量机 (Support Vector Machines, SVM) 由 Vapnik 等人基于统计学习理论的结构风险最小化^[1]原则的基础上提出来的。SVM 可以寻求具有最简的计算复杂性和最佳学习能力的模型,根据有限的样本信息,构成分类,可以最大限度地分割类和类之间的间隔。该算法最终转化为一个二次优化问题。从理论得到全局最优解,解决神经网络方法无法避免的局部极值问题,该算法将实际问题变换到高维特征空间,通过非线性变换到高维特征空间,构造线性判别函数更换原来的空间中的非线性判别函数,这将促进 SVM 的应

用,它巧妙地解决问题的维数,算法的复杂度和样本维数无关。

目前 SVM 研究中还存在一些需要解决的难点。核函数^[2]是支持向量机的核心,对核函数的参数的选择决定了 SVM 的分类性能,但是参数的选择缺乏一定的理论指导。如何根据待解决问题的先验知识和实际样本数据,选择合适的参数是支持向量机理论研究的重点。Keerthis S^[3]和 Chapelle O^[4]等人使用梯度下降法来优化参数,但是梯度下降法对初始点的选择很敏感,若初始点与最优点相差很远,就会使结果容易陷入局部最优;Avci E^[5]和 Huang C^[6]提出了遗传算法的支持向量机参数选择算法,该算法同样会使结果陷入局部最优;Hsu C^[7]等人提出了一种基于 GRID-SEARCH 方法来选择 SVM 的惩罚参数 C 和基于 RBF 核参数的优化方法,但是网格存在计算量大、学习精度低的缺陷;Xiao-yu Fang^[8]等人提出了基于蚁群算法的支持向量机参数优化算法,基于蚁群的方法存在在初始搜索时,蚁群信息素匮乏,求解速度缓慢的缺陷;Moser^[9]

收稿日期:2012-06-21;修回日期:2012-09-27

基金项目:贵阳市 2010 年工业科技攻关项目([2010]筑科工合同字第 28 号);贵州大学 2011 年研究生创新基金资助项目(校研理工[2011039])

作者简介:王萍萍(1987-),女,硕士研究生,主要研究方向为数据库技术与应用系统、数据挖掘;王翰虎,教授,硕士生导师,主要研究方向为数据库应用系统、数据挖掘。

提出了基于 Ho-Kashyap 和 Powell 的参数优化算法,通过 Powell 算法求解容易陷入局部最优,最终的优化结果很大程度上依赖于初始点所在的位置。

考虑到以上参数优化方法存在的问题,文中提出了将遗传算法与 Powell 算法结合用来优化支持向量机核参数的方法,该算法不但保持了遗传算法的强全局搜索能力,同时具有 Powell 算法的局部搜索优势,在迭代的过程中不断调整全局最优值,避免了遗传算法的早熟收敛,得到了很好的寻优效果。将该算法应用到银行基金客户的分类实验中,实验的结果显示,算法的效率和性能方面都有了很大的提高,且得到了较高的准确率。为银行业务分析提供了参考依据。

1 基于 GA-Powell 的 SVM 参数优化

1.1 遗传算法

遗传算法具有很好的全局搜索能力,在解空间内可以快速地搜索出全局解,避免陷入局部最优解的陷阱,以及利用其固有的内在并行性,可以很容易地进行分布式计算,加快求解速度。但是遗传算法的局部搜索能力差,比较费时,并在进化的后期搜索效率较低。实际应用中,遗传算法容易产生早熟收敛的问题。采用什么样的选择方法使优秀个体被保留,并且保持群体的多样性,一直是遗传算法难以解决的问题^[10]。

遗传算法效法基于自然选择的生物进化,是一种模仿生物进化过程的随机方法。在一开始需要对种群中的个体进行编码,如二进制编码。通过迭代,末代种群中的最优个体经过解码,可以作为问题近似最优解。

遗传算法包括三个基本操作:选择、交叉和变异,算法的参数中交叉概率和变异概率的选择是影响遗传算法行为和性能的关键所在。

算法的参数中交叉概率和变异概率的选择是对遗传算法的行为和性能影响的关键所在。

文中采用的是一种自适应遗传算法,算法的交叉概率和变异概率能够随适应度自动改变。当种群中各个体适应度趋于一致或趋于局部最优时,交叉概率和变异概率会随着二者增加,而当群体适应度比较分散时,就使交叉概率和变异概率减小。同时,保留具有高于群体平均适应度的个体,使其具有较低的交叉概率和变异概率,使该个体得以进入下一代。而低于平均适应度的个体,具有较高的交叉概率和变异概率,使该个体被淘汰掉。自适应的交叉概率和变异概率:

$$p_c(G+1) = p_c(G) - \frac{[p_c(1) - 0.6]}{G_{\max}} \quad (1)$$

$$p_m(G+1) = p_m(G) + \frac{[0.08 - p_m(1)]}{G_{\max}} \quad (2)$$

式中: G 是遗传代数, G_{\max} 是最大遗传代数,

$p_c(1)$ 是第一代交叉概率, $p_m(1)$ 是变异概率, $p_c(G)$ 是第 G 代交叉概率, $p_m(G)$ 是第 G 代的变异概率。

1.2 Powell 算法

Powell 算法既是一种模式搜索法,同时也是一种共轭方向法 (Conjugate Direction Method)。为避免坐标轮换法带来的寻优路径锯齿形转折 (Zigzag), Hook 和 Jeeves 首先提出了模式搜索法,在此基础上, Powell 提出了估计新搜索方向的算法。其基本思想是先给出 n 个正交的初始搜索方向 $S_1^0, S_2^0, \dots, S_n^0$, 第一轮寻优循环先沿这一系列方向做一维极值搜索;完成第一次搜索后,得到设计变量与循环开始前的设计变量之差所指向的方向代替原搜索方向 S_1^0 ;进行第二轮搜索,循环结束后,同样得出循环开始前后的设计变量之差,以此方向代替原搜索方向 S_2^0 ,进行第三轮循环的搜索;……如此不断进行,直到达到终止寻优标准。

Powell 搜索法是一个为无约束最优化问题设计的不使用导数的局部搜索方法。在每一维中使用 Brent 算法迭代搜索,搜索速度比较快,在局部搜索中精度要高于其它的优化算法,到目前为止, Powell 算法是最好的直接搜索算法,学习精度高并且迭代比较简单,局部搜索能力也非常强。但是随着迭代次数的增加,容易陷入局部最小点^[11]。

1.3 GA-Powell 算法

从全局方面考虑,遗传算法与传统的优化方法 (枚举,启发式等)相比较,以生物进化为原型,具有很好的收敛性,在计算精度要求的同时,计算时间少、鲁棒性高等都是它的优点。Powell 算法具有较好的局部收敛性。在遗传算法中引入 Powell 算法可以使算法具有较强的全局搜索能力的同时,也避免陷入局部最小值的缺陷,提高算法的性能^[12]。

GA-Powell 算法的核心思想是:首先使用遗传算法在全局范围内搜索最优解,得到一个较优的参数,再通过 Powell 算法在这个较优解附近进行局部搜索,若搜索结果没有达到指定的结束条件,则再次进行全局搜索,依次循环,直到最后找到最优参数。支持向量机的核函数选择 RBF 核函数。利用 GA-Powell 算法对核函数进行参数优化,评价函数为方均误差,用来对核函数参数进行评价选择。

GA-Powell 算法步骤:

Step 1:初始化 SVM 参数并编码,随机产生遗传算法的初始群体,群体数量选取 100。支持向量机中的参数称为抗体。本算法中的抗体以实数进行编码。采用 RBF 核函数,此时支持向量机中有 2 个参数,那么抗体编码由 2 部分组成,分别存放 c, σ 。

Step 2:设定遗传算法的参数,将初始代数设为 G

= 1, 初始交叉概率 $p_c(1) = 0.8$, 初始变异概率 $p_m(1) = 0.008$, 最大遗传代数 $G_{\max} = 100$ 。

Step 3: 将群体解码, 然后送入 SVM 进行训练, 并且计算个体的适应度, 文中的适应度函数采用分类器与参考模型的方均误差 (MSE), 即 $\min f^2 = \frac{1}{n} \sum_{i=1}^n (y - f(x))^2$ 。确定最优个体 x_{best} , 给出 σ, c 取值;

Step 4: 根据最优保存策略, 保存最优个体 x_{best} , 标记适应度 (误差最大) 最差个体 x_{bad} 。

Step 5: 以当前最好群体位置 $x^{(0)} = x_{\text{best}}$ 为初始点, 作为 Powell 搜索的初始值, 并给定 n 个线性无关的初始向量组: $\{p^0, p^1, \dots, p^n\}_0$ 及精度 $\varepsilon > 0$, 置 $k = 0$;

Step 6: 令 $y^0 = x^k$, 依次沿 $\{p^0, p^1, \dots, p^n\}_k$ 中的方向进行一维搜索:

$$\begin{cases} y^j = y^{j-1} + t_{j-1} p^{j-1} \\ f(y^{j-1} + t_{j-1} p^{j-1}) = \min_{t \geq 0} (f(y^{j-1} + t p^{j-1})) ; \\ j = 1, 2, \dots, n; \end{cases}$$

Step 7: 令 $p^n = y^n - y^0$, 若 $\|p^n\| \leq \varepsilon$, 停止迭代, 输出 $x^* = y^n$, 转 Step 11, 否则转 Step 8;

Step 8: 求出 m , 使得

$$f(y^{(m-1)}) - f(y^{(m)}) = \max_{1 \leq j \leq n} \{f(y^{(j-1)}) - f(y^{(j)})\}$$

若下式成立

$$f(y^0) - 2f(y^n) + f(2y^n - y^0) < 2[f(y^{m-1}) - 2f(y^m)] \text{ 转 Step 9, 否则转 Step 10;}$$

Step 9: 求解 $\min_{t \geq 0} (f(y^{(m)} + t p^{(m)})$, 设最优解 t_n , 令

$$x^{k+1} = y^n + t_n p^n, \text{ 同时令 } \{p^0, p^1, \dots, p^n\}_{k+1} = \{p^0, p^1, \dots, p^{m-1}, p^{m+1}, \dots, p^{n-1}, p^n\}, \text{ 置 } k = k + 1, \text{ 转 Step 6;}$$

Step 10: 令 $x^{k+1} = y^n$, 置 $k = k + 1$, 转 Step 6;

Step 11: 进行优化计算求出最优解 x^* , 并令 $x_{\text{best}} = x^*$;

Step 12: 进行遗传操作, 赌轮选择法作为选择算子, 单点交叉作为交叉算子, 通过遗传算子操作最后产生新群体, 并用 Step11 的最优个体 x_{best} 替换代号为 x_{bad} 的新个体, 并最终产生新的群体;

Step13: 检查算法是否满足终止条件, 当连续几代最优个体的适应度相等时可以认为种群不能再进化, 将其作为算法终止的条件之一, 将最大遗传代数也作为算法的终止条件之一, 当满足上述两个条件中的任何一个时, 算法自动终止, 输出最终解 x_{best} , 否则, 转 Step 3 再进行训练, 直到满足算法终止条件。

优化后得到误差惩罚参数 c, σ 。将处理过的训练数据集和得到的最优参数输入决策函数

$$f(x) = \text{sgn} \left[\sum_{i=1}^l y_i \alpha_i K(x_i, x) + b \right]$$

其中 $K(x_i, x_j) = \exp(-\sigma \|x_i - x_j\|)^2$ 进行训练, 最后得到分类结果。

2 实验结果与分析

从上海农商银行数据仓库中选取数据, 首先从购买基金的客户数据中提取 300 条数据, 从没有购买基金的客户中提取 300 条数据, 选取开户时长 (月)、年龄、性别、最近半年活期账户数、最近半年定期账户数等 10 个维度。将这 600 条数据分为两组: 一组为训练集, 共 300 条, 其中基金数据 150 条; 另一组为测试集, 其中基金数据 150 条。通过对第一组的数据进行训练, 并建立一个训练模型; 第二组是测试集, 用于检测训练模型的分类能力^[13]。

把测试数据集输入到分类模型进行预测。根据所得的结果判别测试样本属于哪一类, 这样就可以通过模型判断哪些客户会成为购买基金的潜在客户, 为业务部门的决策提供有利的科学依据。

在实验中分别采用遗传算法、Powell 算法和 GA-Powell 算法进行参数优化并带入支持向量机进行分类预测。程序基于 Matlab2010a 平台, 机器配置为 Intel i3(2.27GHz) 处理器, Windows XP SP3, 2G RAM。三种实验的结果如图所示, 图 1 为采用遗传算法优化参数的分类预测效果图, 图 2 为采用 Powell 算法优化参数的分类预测效果图, 图 3 为采用 GA-Powell 优化参数的分类预测效果图。其中圆圈坐标点表示测试数据集的数据所属分类, 星形坐标点表示预测的测试数据集的数据所属分类, 当星形坐标和圆圈坐标重合, 即表示预测正确, 相反则预测错误。

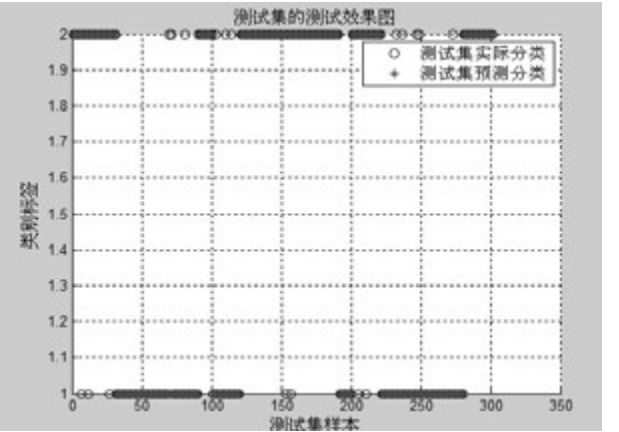


图 1 采用遗传算法优化参数的分类预测效果图

表 1 三种优化方法的性能比较

算法	学习时间/min	准确率
遗传算法	21.12	94.0199%
Powell 算法	19.67	93.0233%
GA-Powell 算法	18.13	97.6744%

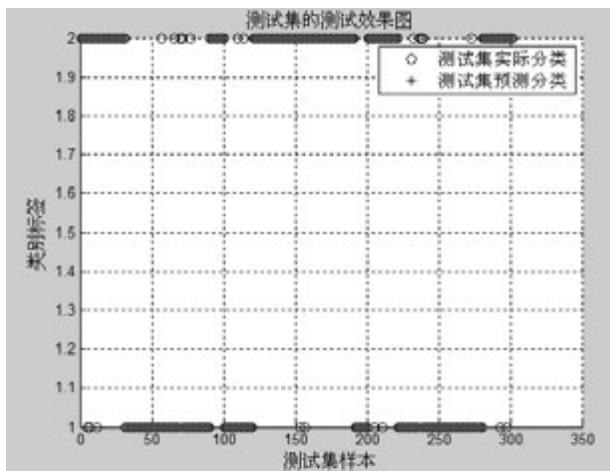


图 2 采用 Powell 算法优化参数的分类预测效果图

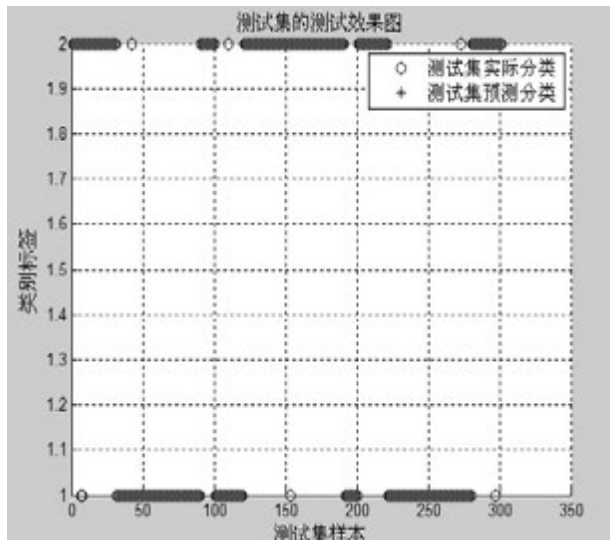


图 3 采用 GA-Powell 优化参数的分类预测效果图

三种算法的性能对比结果如表 1 所示,可以看出遗传算法的执行时间比 Powell 算法长,但是学习精度高于 Powell 算法,GA-Powell 算法在学习时间和精度方面都优于遗传算法和 Powell 算法,该算法通过将遗传算法和 Powell 算法的优点结合起来,在进行全局最优搜索的同时进行局部最优搜索,提高了算法的搜索效率。在保证遗传算法快速收敛的前提下不影响其较强的全局搜索能力。总之,将遗传算法和 Powell 算法结合起来,避免其缺点,使 GA-Powell 算法具有优于另外两种算法的分类性能。

3 结束语

文中提出的 GA-Powell 算法将遗传算法和 Powell 算法结合起来,用来优化支持向量机的参数。算法不

仅具有较强的全局搜索能力,同时增强其局部搜索能力,避免了遗传算法局部搜索能力差、Powell 算法容易陷入局部最优的可能。将该算法运用到银行基金业务客户分类实验中,具有较高的准确率与性能优势,实验结果表明,GA-Powell 算法是一种高效可行的 SVM 参数优化算法。

参考文献:

- [1] 白 鹏,张喜斌,张 斌,等.支持向量机理论与工程应用实例[M].西安:西安电子科技大学出版社,2008.
- [2] Müller K R, Mika S, Ratsch G, et al. An Introduction to Kernel Based Learning Algorithms[J]. IEEE Transactions on Neural Networks, 2001, 12(2): 181-202.
- [3] Keerthi I S S. Efficient tuning of SVM hyper parameters using ra2 dius/margin bound and iterative algorithms[J]. IEEE Trans on Neural Network, 2002, 13(5): 1225-1229.
- [4] Chapelle O, Vapnik V, Bousquet O, et al. Choosing multiple parameters for support vector machines[J]. Machine Learning, 2002, 46(1-3): 131-159.
- [5] Avci E. Selecting of the optimal feature subset and kernel parameters in digital modulation classification by using hybrid genetic algorithm support vector machines: HGASVM[J]. Expert Systems with Applications, 2009, 36(2): 1391-1402.
- [6] Huang C L, Wang C J. A GA-based feature selection and parameters optimization for support vector machines[J]. Expert Systems with Applications, 2006, 31(2): 231-240.
- [7] Hsu C W, Chang C C, Lin C J. A practical guide to support vector classification[EB/OL]. 2003. <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
- [8] Fang Xiaoyu, Bai Tao. Share Price Prediction Using Wavelet Transform and Ant Colony Algorithm for Parameters Optimization in SVM[C]//GCIS '09. [s. l.]: [s. n.], 2009: 288-292.
- [9] Moser G, Serpico S B. A Markovian generalization of support vector machines for contextual supervised classification of hyperspectral images[C]//WHISPERS. [s. l.]: [s. n.], 2010: 1-4.
- [10] 计智伟,吴耿锋,胡 珉.基于自适应遗传算法和 SVM 的特征选择[J].计算机工程,2009,35(14):200-202.
- [11] 刘叶青,刘三阳,谷明涛. Powell 算法在线性支持向量机中的应用[J].计算机工程,2011,37(12):161-163.
- [12] Chen P W, Wang Jung-Ying, Lee Hahn-Ming. Model selection of SVMs using GA approach[C]//Proceedings of 2004 IEEE International Joint Conference on Neural Networks. Piscataway: IEEE, 2004: 2035-2040.

一种支持向量机参数优化的GA-Powell算法

作者: [王萍萍](#), [王翰虎](#)
作者单位: [贵州大学 计算机科学与信息学院, 贵州 贵阳550025](#)
刊名: [计算机技术与发展](#)
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2013 (2)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjtz201302006.aspx