

一种改进的支持向量机模型优化算法

李利杰¹, 张君华¹, 熊伟清², 张颖³

(1. 宁波城市职业技术学院, 浙江 宁波 315104;

2. 宁波大学 信息学院, 浙江 宁波 315211;

3. 宁波卫生职业技术学院, 浙江 宁波 315104)

摘要:核函数与参数选择即模型优化是影响支持向量机泛化能力的主要因素。为提高支持向量机的泛化能力,文中在最优保存遗传算法的基础上引入学习算子和主成分分析方法,提出一种新的支持向量机模型优化新算法(简称PCA-SLGA)解决支持向量机分类器模型自动优化问题。仿真实验结果表明,与用于支持向量机模型优化的隐马尔可夫、贪心算法、遗传编程等算法相比,PCA-SLGA算法具有快速收敛性和较强的全局搜索能力。实验进一步表明采用混合算法寻找最优核模型是一种可行途径。

关键词:支持向量机;模型选择;主成分分析;自学习;遗传算法

中图分类号:TP18

文献标识码:A

文章编号:1673-629X(2014)12-0114-04

doi:10.3969/j.issn.1673-629X.2014.12.027

An Improved Algorithm for Model Optimization of Support Vector Machine

LI Li-jie¹, ZHANG Jun-hua¹, XIONG Wei-qing², ZHANG Ying³

(1. Ningbo City College of Vocational Technology, Ningbo 315104, China;

2. School of Information, Ningbo University, Ningbo 315211, China;

3. Ningbo College of Health Sciences, Ningbo 315104, China)

Abstract: Model optimization, the choice of kernel functions and its parameters, has a profound impact on the generalization ability of the support vector machine. In this paper, to improve the generalization ability for SVM, a new kernel optimization algorithm (for short PCA-SLGA), based on the elitist of genetic algorithm, which adopts self-learning operator and principle component analysis method, is put forward in order to solve the problem of automatic optimization of VSM classifier model. Compared with the SVMs optimized by hidden Markov, greedy algorithm, genetic programming, the experimental results show that PCA-SLGA converge faster and has stronger global search ability than the algorithms mentioned above. The experiments further indicates that using the hybrid algorithm to optimize the kernel is a promising way to find the optimal kernel model.

Key words: support vector machine; model selection; principle component analysis; self-learning; genetic algorithm

0 引言

支持向量机(Support Vector Machine)是基于统计学习理论的机器学习方法,在解决非线性及高维模式识别中表现出许多特有的优势,并广泛应用于函数拟合、文本和图像识别等机器学习问题中。支持向量机方法是建立在统计学习理论的VC维理论和结构风险最小原理的基础上,根据有限样本信息在模型选择复

杂性和学习能力之间寻求最佳折衷,以求获得最好的推广能力^[1-2]。大量研究表明,支持向量机分类性能与核函数类型、参数与惩罚系数关系紧密,选择合适的核函数及其参数(即模型选择)以获取最优分类结果,是目前支持向量机研究领域的热点。相关研究主要集中在对某种特定形式核函数的参数选择^[3]。选取核函数解决实际问题时常用方法有交叉验证法、正交法、梯

收稿日期:2014-03-28

修回日期:2014-06-30

网络出版时间:2014-10-27

基金项目:国家科技型中小企业技术创新基金(11C26213311798);浙江省医药卫生科技计划项目(2013KYB242);浙江省卫生经济学会资助课题;宁波市自然科学基金(2012A610063);宁波城市职业技术学院科研重点课题(ZZX13035)

作者简介:李利杰(1981-),男,硕士研究生,研究方向为机器学习、智能计算;熊伟清,教授,硕士生导师,研究方向为演化计算、高性能计算。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20141027.1429.007.html>

度下降法、蚁群优化^[4-5]、动量粒子群优化^[6]、人工蜂群优化^[7]等。

文中在最优保存遗传算法的基础上引入自学习算子和主成分分析方法,提出一种支持向量机模型优化算法(为论述便捷取主成分分析和自学习英文单词缩写,简称 PCA-SLGA)。其核心是首先通过主成分分析方法分析样本分布,确定核函数类型;接着采用自学习遗传算法参数寻优,从而确定最优核模型。实验结果表明 PCA-SLGA 算法能在较少的进化代数内搜索到最优模型,从而提高支持向量机的分类性能。

1 支持向量机

支持向量机定义最优线性超平面,把寻找最优线性超平面转换为求解二次规划问题。

Mercer 定理通过非线性映射把样本空间映射到高维特征空间,从而使用线性方法解决样本空间中的高度非线性问题^[8]。给定训练样本 $\{x_i, y_i\}_{i=1}^m$, x_i 是输入向量, $y_i \in \{-1, 1\}$ 是相应的分类标签,支持向量机寻找一个最优超平面使其分类间隔最大(如图 1 所示)。当训练样本数据集为线性不可分时引入非负松弛变量 $\varepsilon_i, i = 1, 2, \dots, m$ 。分类超平面的最优问题则转化为求解目标函数(1)。

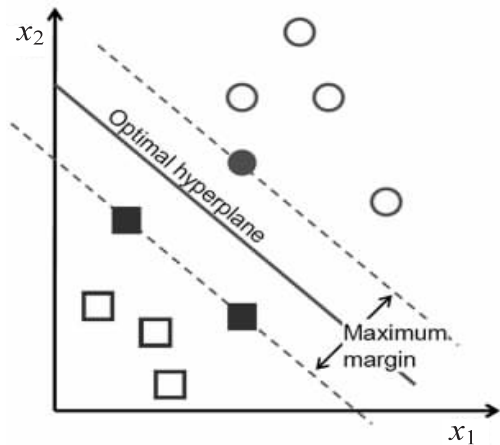


图 1 支持向量机分类示例

$$\begin{aligned} \text{Min}(r, w, b) &= \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \varepsilon_i \\ \text{s. t. } y_i[(w \cdot x_i) + b] &\geq 1 - \varepsilon_i \\ \sum_{i=1}^m y_i \alpha_i &= 0 \\ 0 \leq \alpha_i \leq C, \varepsilon_i \geq 0, i &= 1, 2, \dots, m \end{aligned} \tag{1}$$

其中, $C > 0$ 是一个常数,称为误差惩罚参数,它控制对错分样本惩罚程度; ε_i 是在训练样本线性不可分时引入的非负松弛变量。

通过引入核函数 $K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$, 则目标函数(1)转化为目标函数(2)。

$$Q(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j K(x_i, x_j) \tag{2}$$

而相应的分类决策函数表示为函数(3)。

$$f(x) = \text{sgn}((w^* \cdot x) + b^*) \tag{3}$$

其中, $w^* = \sum_{i=1}^m y_i \alpha_i^* x_i; b^* = y_j - \sum_{i=1}^m y_i \alpha_i^* (x_i, x_j), j \in \{j | 0 < \alpha_j^* < C\}$ 。

由此可见,核函数的种类及其参数、正则化参数的确定对支持向量机的学习性能和泛化能力有非常重要的影响。

2 PCA-SLGA 模型优化算法框架

PCA-SLGA 模型优化算法第一步使用主成分分析技术分析数据分布并选择核函数类型,紧接着采用自学习遗传算法进行参数优化。PCA-SLGA 模型优化算法流程如下所示:

输入:有 m 个数据样本的测试数据集 G ;

输出:核函数类型及相应参数。

Step1:对于给定数据样本集,使用主成分分析和数据分布估计选择核函数种类。

Step2:给定核函数种类,使用自学习遗传算法对参数进行优化。

Step3:适应度函数值在连续 n 次迭代如果没有改进,则退出,否则转 Step2。

2.1 主成分分析估计核函数类型

支持向量机中应用的核函数类型主要有多项式核函数、径向基核函数等。支持向量机学习性能优劣与核函数机器参数选择有直接联系。核函数与其特征空间是一一对应的,确定核函数就隐含确定映射函数 φ 和特征空间 H 。核函数的改变实际上是隐含改变映射函数,从而改变学习样本特征子空间分布的复杂程度。特征子空间的维数决定此空间构造的线性分类面的最大 VC 维,同时也决定子空间中线性分类面能达到的最小经验误差^[9]。文中利用主成分分析和数据分布估计选择特定核函数类型。如果数据分布接近于圆形选择 Polar 核函数;如果数据分布接近球体或柱形,则相应选择 Sphere 或者 Cylinder 核函数;否则随机选择 Gaussian 核函数或者 Polynomial 核函数。核函数选择算法如下:

Step1:使用主成分分析将测试数据集 G 映射到三维空间 G' 。

Step2:根据三维空间 G' 的样本频率估计数据总体分布。

Step3:根据数据分布从核函数表(见表 1)中确定核函数类型。

表 1 核函数类型

| 核函数名 | 表达式 |
|------------|--|
| RBF | $k_{\text{gau}}(r) = \exp(-\ x-y\ ^2/2\sigma^2)$ |
| Polynomial | $k_{\text{pol}}(d) = (\langle x, y \rangle + c)^d$ |
| Polar | $k_{\text{polar}}(\alpha) = \alpha \tan(x_2/x_1) \cdot \alpha \tan(y_2/y_1) + \ x\ \ y\ $ |
| Sphere | $k_{\text{sp}}(\alpha) = \alpha \cos(x_3/\ x\) \alpha \cos(y_3/\ y\) + \alpha \tan(x_2/x_1) \alpha \tan(y_2/y_1)$ |
| Cylinder | $k_{\text{cy}}(\alpha) = x_3 y_3 + \alpha \tan(x_2/x_1) \cdot \alpha \tan(y_2/y_1) + \sqrt{(x_1^2 + x_2^2) + (y_1^2 + y_2^2)}$ |

2.2 自学习遗传算法优化参数

遗传算法是借鉴生物进化规律演化而来的一种随机搜索优化方法^[10]，已有学者利用该算法来确定支持向量机的参数^[11-13]。遗传学基本原理显示优秀的父代以较大的概率产生优秀的子代。父代通过一定的方式进行学习以提高自身的性能，经复制后会以较大的概率使子代性能提高。

结合这一原理，针对传统遗传算法存在收敛速度慢等问题，文中在最优保存遗传算法的基础上引入自学习算子进行参数寻优。在自学习遗传算法中需搜索和存储优秀模式，方便其他个体进行趋同学习。优秀模式搜索首先找出群体优秀串及其所有优秀位，再根据优秀位确定相应的一个模式。搜索优秀模式过程描述如图 2 所示。

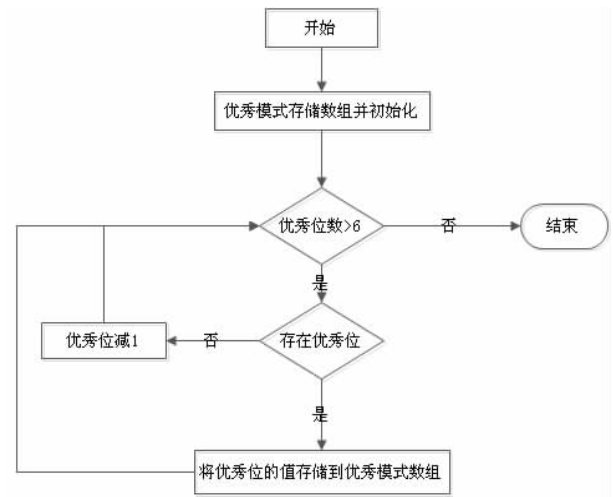


图 2 优秀模式搜索过程

设 x 是群体中的一个性能较差的串， p_i 为学习概率，学习算子定义 S 如下：

$$y_i = \begin{cases} X_i, & i \text{ is random position or } \text{rand} > p_i \\ H_i, & i \text{ is affirmed position or } \text{rand} \leq p_i \end{cases} \quad (4)$$

式中，rand 为 (0,1) 之间的一个均匀分布的随机

数。在学习算子的作用下，串 x 向优秀模式 H 趋同后变为 y ， y 串以概率 p_i 进入群体优秀模式，使得算法搜索性能大大提高。自学习遗传算法描述如下：

步骤 1：初始化种群 P_i 、交叉概率、变异概率；

步骤 2：将训练样本划分成 k 个互不重叠的子样本集；

步骤 3：随机选择一个子样本集作为测试数据集，其余为训练样本集；

步骤 4：根据核函数选定类型，初始化核函数中的参数；

步骤 5：找出群体中的优秀模式 H ；

步骤 6：如果优秀模式 H 存在，则对群体中性能较差的串使用学习算子；

步骤 7：对种群 P_i 作用遗传算子；

步骤 8：计算泛化误差 $\varepsilon = \text{mean} (s_i - \bar{s}_i)^2$ 作为适应度值；

步骤 9：泛化误差在连续 N 次循环中没有提升则结束，否则转步骤 5。

3 实验

文中使用 UCI 数据库中的基准数据集进行测试来检验 PCA-SLGA 所构造的支持向量机在不同分类任务中的性能。在测试过程中，文中将训练样本集分解为 10 个子样本集，停机准则为连续 10 次迭代过程中泛化误差没有提升。所有实验都是在主频为 2.53 GHz，内存为 4 G 的 PC 机中完成。

根据图 3 不难看出随着迭代次数的增加，分类错误率逐渐减小并接近于零，最终趋于稳定。此外，图 3 也直接说明 PCA-SLGA 算法在最优核模型搜索方面具有较强的搜索能力和快速收敛性。

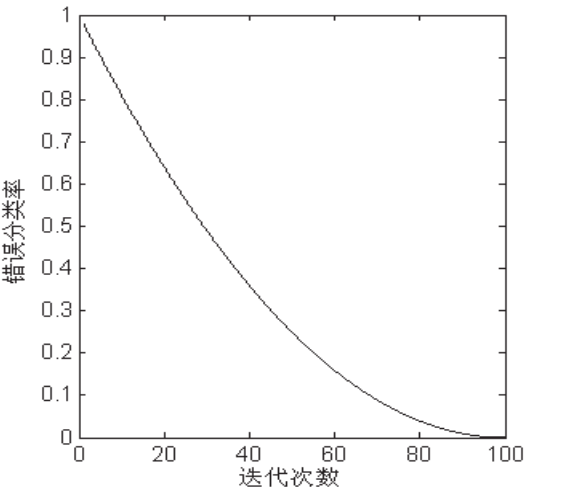


图 3 分类错误率随算法迭代的演变

表 2 显示 PCA-SLGA 与 GS-SVM 优化算法得到的核函数、参数以及分类测试结果。GS-SVM 是采用

贪心算法对支持向量机模型进行优化^[14],其算法复杂度在最坏情况下为 $O(n^2)$ 。

从表 2 中可以看到,在 Iris 和 Heart Disease 测试数据上,PCA-SLGA 算法所构造的支持向量机的错分率远比 GS-SVM 所构造的支持向量机要低。在其他几个测试数据集上,PCA-SLGA 的错分率均为零。

由此可见 PCA-SLGA 比 GS-SVM 算法具备更优的全局搜索能力。

表 2 PCA-SLGA,GS-SVM 分类错误率

| 数据集 | 核函数(参数) | PCA-SLGA | GS-SVM |
|---------------|------------------------------|----------|---------|
| WDBC | RBF($r=0.001\ 8,c=13.5$) | 0 | 0.022 8 |
| Heart Disease | Polar($a=18.6,c=15.56$) | 0.007 6 | 0.163 0 |
| Ionosphere | RBF($r=0.53,c=100$) | 0 | 0.060 0 |
| Iris | Sphere($a=16.3,c=163.2$) | 0.002 8 | 0.046 7 |
| Wine | RBF($r=0.008,c=140.5$) | 0 | 0.011 1 |
| Vowel | RBF($r=0.001\ 2,c=808$) | 0 | 0.017 1 |
| Splice | RBF($r=0.006\ 2,c=1\ 088$) | 0 | 0.032 8 |
| glass | RBF($r=19.506,c=30.31$) | 0 | 0.284 6 |

为了进一步测试 PCA-SLGA 的泛化性能,文中进一步将 PCA-SLGA 与 HM-SVM^[15],KGP^[16] 其他两种支持向量机模型优化算法在 Heart Disease, Ionosphere 等基准数据集上进行测试。

其中 HM-SVM 是基于 MOSEK 优化工具箱构造而成的优化算法。PCA-SLGA, HM-SVM, KGP 三种模型优化算法在不同基准数据集上的测试结果如图 4 所示。

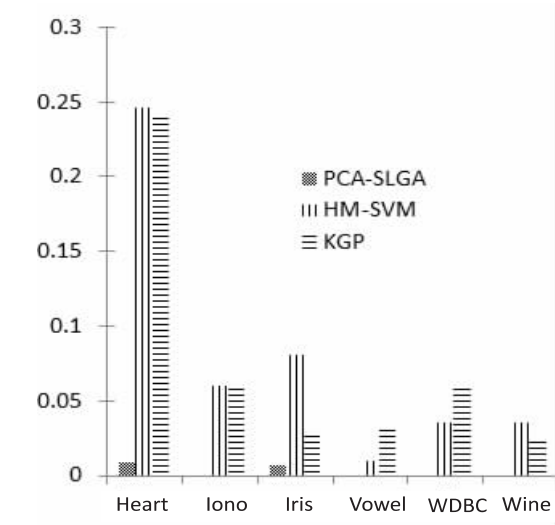


图 4 PCA-SLGA, HM-SVM, KGP 三种算法在 Heart, Ionosphere, Iris, Vowel, WDBC, Wine 基准数据集上的测试结果(分类错误率)

尤其在 Ionosphere (简称 Iono), Vowel, WDBC, Wine 四个基准数据集上,PCA-SLGA 所构造的支持向量机分类错误率为零。这些测试结果进一步证明 PCA

-SLGA 模型优化算法与 HM-SVM, KGP 算法相比具备更强的全局搜索能力,从而提高了支持向量机的分类性能。

4 结束语

支持向量机核模型显著影响其分类性能,在设计支持向量机分类器时通常会因使用不合适的核函数及参数导致分类器无法满足实际应用。文中使用主成分分析和自学习遗传算法构造支持向量机模型优化算法(PCA-SLGA),试图实现核函数模型优化自动化,从而减少人为因素干扰。

实验结果表明,PCA-SLGA 在最优核模型优化方面具有快速收敛和全局搜索能力,从而提高支持向量机的分类性能。

参考文献:

[1] Gonen M, Alpaydin E. Multiple kernel learning algorithm[J]. Journal of Machine Learning Research, 2011, 12(7): 2211-2268.

[2] 马元良,裴生雷. 基于改进遗传算法的 SVM 参数优化研究[J]. 计算机仿真, 2010, 27(8): 150-152.

[3] 杨旭,杨新,熊惠霖. 一种用于支持向量机分类器的组合核优化方法[J]. 上海交通大学学报, 2010, 44(8): 1037-1041.

[4] 张培林,钱林方,曹建军,等. 基于蚁群算法的支持向量机参数优化[J]. 南京理工大学学报(自然科学版), 2009, 33(4): 464-468.

[5] 刘春波,王鲜芳,潘丰. 基于蚁群优化算法的支持向量机参数选择及仿真[J]. 中南大学学报(自然科学版), 2008, 39(6): 1309-1313.

[6] 王佳,徐蔚鸿. 基于动量粒子群的混合核 SVM 参数优化方法[J]. 计算机应用, 2011, 31(2): 501-503.

[7] 于明,艾月乔. 基于人工蜂群算法的支持向量机参数优化及应用[J]. 光电子·激光, 2012, 23(2): 374-378.

[8] Lu Zhao, Sun Jing. Non-Mercer hybrid kernel for linear programming support vector regression in nonlinear systems identification[J]. Applied Soft Computing, 2009, 9(1): 94-99.

[9] Adankon M M, Cheriet M. Model selection for the LS-SVM. Application to handwriting recognition[J]. Pattern Recognition, 2009, 42(12): 3264-3270.

[10] 万源,童恒庆,朱映映. 基于遗传算法的多核支持向量机的参数优化[J]. 武汉大学学报(理学版), 2012, 58(3): 255-259.

[11] Wu Chih-Hung, Tzeng Gwo-Hshiung, Lin Rong-Ho. A novel hybrid genetic algorithm for kernel function and parameter optimization in support vector regression[J]. Expert Systems with Applications, 2009, 36(3): 4725-4735.

[12] 李良敏,温广瑞,王生昌. 基于遗传算法的回归型支持向量

(下转第 123 页)

6 结束语

针对多透视拼接采集点密集和视角固定的限制,文中有针对性地引入了虚拟变换和同步校正与配准的两步走策略,从而创新地提出视点稀疏且可旋转的多透视拼接算法。实验结果表明文中算法可以取得较满意的拼接效果。笔者认为文中算法较好地发展了图像拼接理论,具有较高的应用价值。

目前,文中算法还受到角度和遮挡变化等的影响,Seitz 和 Dyer^[19]或许提供了一种解决办法,因此将在接下来的研究工作中尝试,以期实现更加完美的多透视拼接。

参考文献:

[1] Peleg S, Herman J. Panorama mosaic by manifold projection [C]//Proceedings of the IEEE computer society conference on computer vision and pattern recognition. San Juan, Puerto Rico:IEEE,1997:338-343.

[2] 方贤勇. 图像拼接技术研究[D]. 杭州:浙江大学,2005.

[3] Zheng Jiangyu. Digital route panoramas[J]. IEEE Multimedia,2003,10(3):57-67.

[4] Zhu Zhigang, Hanson A R, Schultz H, et al. Stereo mosaics from a moving video camera for environmental monitoring [C]//Proc of first international workshop on digital and computational video. Tampa, Florida, USA:[s. n.],1999:45-54.

[5] Molina E, Zhu Zhigang, Taylor C N. A layered approach for fast multi-view stereo panorama generation[C]//Proc of the IEEE international symposium on multimedia. California, USA:IEEE,2011:589-594.

[6] Szeliski R. Image alignment and stitching;a tutorial[J]. Foundations and Trends in Computer Graphics and Vision,2006,2(1):1-104.

[7] Huang Yuwen, Chen C Y, Tsai C H, et al. Survey on block matching motion estimation algorithms and architectures with new results[J]. Journal of VLSI Signal Processing,2006,42(3):297-320.

[8] Zitova B, Flusser J. Image registration methods;a survey[J]. Image and Vision Computing,2003,21(11):977-1000.

[9] Cho S H, Chung Y K, Lee J Y. Automatic image mosaic system using image feature detection and Taylor series[C]//Proceed-

ings of 7th digital image computing: techniques and applications. Biennial, Australian:[s. n.],2003:549-560.

[10] Fang Xianyong, Luo Bin, Zhao Haifeng, et al. New multi-resolution image stitching with local and global alignment[J]. IET Computer Vision,2010,4(4):231-246.

[11] Tang Hao, Zhu Zhigang. Content-based 3-D mosaics for representing videos of dynamic urban scenes[J]. IEEE Transactions on Circuits and Systems for Video Technology,2012,22(2):295-308.

[12] Peleg S, Rousso B, Rav-Acha A, et al. Mosaicing on adaptive manifolds[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(10):1144-1154.

[13] Fang Siyuan, Neill C. Multi-perspective panoramas of long scenes[C]//Proceedings of the IEEE international conference on multimedia and expo. Melbourne:IEEE,2012:91-96.

[14] Fang Siyuan, Neill C. Visualization of long scenes from dense image sequences using perspective composition[J]. Communication in Computer and Information Science,2013,359:67-81.

[15] Zoghalmi I, Faugeras O, Deriche R. Using geometric corners to build a 2D mosaic from a set of image [C]//Proceedings of the IEEE computer society conference on computer vision and pattern recognition. San Juan, Puerto Rico:IEEE,1997:420-425.

[16] Capel D, Zisseman A. Automated mosaicing with super-resolution zoom [C]//Proceedings of the IEEE computer society conference on computer vision and pattern recognition. Santa Barbara, CA, USA:IEEE,1998:885-891.

[17] Mclauchlan P F, Jaenicke A. Image mosaicing using sequential bundle adjustment[J]. Image and Vision Computing,2002,20(9-10):751-759.

[18] Lallier E, Farooq M. A real time pixel-level based image fusion via adaptive weight averaging [C]//Proceedings of the third international conference on information fusion. [s. l.]:[s. n.],2000.

[19] Seitz S M, Dyer C R. Towards image-based scene representation using view morphing[C]//Proceedings of the 13th international conference on pattern recognition. Vienna, Austria:[s. n.],1996:84-89.

(上接第 117 页)

机参数选择法[J]. 计算机工程与应用,2008,44(7):23-26.

[13] 唐 静,胡云安,肖支才. 基于遗传算法的电路故障诊断超参数优化算法框架[J]. 计算机工程与应用,2012,48(3):13-16.

[14] Bo Liefeng, Wang Ling, Jiao Licheng. Training hard-margin support vector machines using greedy stagewise algorithm[J]. IEEE Transactions on Neural Networks,2008,19(8):1446-

1455.

[15] Finley T, Joachims T. Training structural SVMs when exact inference is intractable [C]//Proceedings of the 25th international conference on machine learning. NY:ACM,2008:3-10.

[16] Sullivan K M, Luck S. Evolving kernels for support vector machine classification [C]//Proceedings of the 9th annual conference on genetic and evolutionary computation. London:ACM,2007:1702-1707.

一种改进的支持向量机模型优化算法


作者:

[李利杰](#), [张君华](#), [熊伟清](#), [张颖](#), [LI Li-jie](#), [ZHANG Jun-hua](#), [XIONG Wei-qing](#), [ZHANG Ying](#)

作者单位:

[李利杰, 张君华, LI Li-jie, ZHANG Jun-hua \(宁波城市职业技术学院, 浙江 宁波, 315104\), 熊伟清, XIONG Wei-qing \(宁波大学 信息学院, 浙江 宁波, 315211\), 张颖, ZHANG Ying \(宁波卫生职业技术学院, 浙江 宁波, 315104\)](#)

刊名:

[计算机技术与发展](#) 

英文刊名:

[Computer Technology and Development](#)

年, 卷(期):

2014(12)

引用本文格式: [李利杰](#), [张君华](#), [熊伟清](#), [张颖](#), [LI Li-jie](#), [ZHANG Jun-hua](#), [XIONG Wei-qing](#), [ZHANG Ying](#) [一种改进的支持向量机模型优化算法](#) [期刊论文] - [计算机技术与发展](#) 2014(12)