文章编号:1671-4598(2016)09-0215-03 DOI:10.16526/j.cnki.11—4762/tp.2016.09.060 中图分类号:TP181 文献标识码:A

基于自适应遗传算法的SVM参数优化

孟 滔,周新志,雷印杰
(四川大学 电子信息学院,成都 610065)

摘要:针对基于遗传算法支持向量机(SVM)训练时间较长以及分类精度较网格搜索法有所降低等问题,通过重新定义遗传算法参数的寻优范围,提出一种自适应遗传算法;该算法根据网格搜索法得到遗传算法参数的最佳寻优范围,然后遗传算法在这个范围内进行参数的精确寻优,最后得到分类的结果;这样不仅可以有效缩短训练时间,而且拥有更高的分类正确率;通过UCI中的10组经典数据集的实验结果可知,自适应遗传算法较之网格搜索法、常规遗传算法、粒子群算法在训练时间上有较大的提升,并且拥有较高的分类准确率。

关键词:支持向量机;参数优化;遗传算法;网格搜索法

A Parameters Optimization Method for an SVM Based on Adaptive Genetic Algorithm

Meng Tao, Zhou Xinzhi, Lei Yinjie
(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: Aiming at that training needs a long time and classification accuracy worse than grid algorithm which reduced the benefits of support vector machine (SVM) based on genetic algorithm. This paper redefined the optimal range of the parameters of genetic algorithm, and proposed an adaptive genetic algorithm. According to the grid search method to get the best optimization genetic algorithm parameter range, and then the genetic algorithm in this range for accurate optimization of parameters, get the final classification results. It can not only shorten the time effectively, and it has higher classification accuracy. Experimental results on the UCI 10 classical dataset verify that adaptive genetic algorithm is better than the grid search method、traditional genetic algorithm、particle swarm optimization (PSO) algorithm in the aspect of training time, and has higher classification accuracy.

Keywords: support vector machine; parameter optimization; genetic algorithm; grid search method

0 引言

支持向量机(SVM)是Corters和Vapnik等人[1]于1995年提出的一种新的机器学习方法,其根据结构风险最小化原则,最大程度地提高了其泛化能力[2],在解决小样本、非线性和高维模式识别中表现出许多特有的优势,并能够推广应用到函数拟合等其他机器学习问题中。但是支持向量机只是一个二分类器,实际应用中的问题大多数是多分类问题,如何将其推广到多分类,同时提高分类准确度与分类计算速度一直是机器学习领域的研究热点。

SVM进行多分类时,主要包括两个阶段,训练阶段即参数寻优阶段以及利用分类模型进行分类阶段。参数寻优主要包括核函数参数和分类模型参数的寻优,实际使用中核函数选用使用最广泛的径向基核函数,分类模型常选用C-支持向量分类机(C-SVC)支持向量机,其中核参数g和惩罚参数C是决定径向基核函数和C-SVC的关键。g表示样本映射特征空间的复杂程度,C是对错分样本比例和算法复杂度的折衷,这两个参数是可以人为调节的,取值不同,对应的分类器性质以及推广识别率也将有很大差别[3]。这两个阶段中分类阶段耗时很少一般在毫秒级,而参数寻优阶段耗费的时间却远远大于其分类时间,一般是分类阶段的上万倍以上。故找到最佳核参数g和惩罚参数C,是提高分类准确度有效途径。而快速找到这组参数则是提高参数寻优速度的有效途径。针对SVM的这两个参数,国内外部分学者从多个方面提出了解决方法。例如XL Liu等人[4]提出基于网格搜索最佳核参数g和惩罚参数C,该方法通过遍历网格中所有的点来寻找最优解,由于寻找的点多,故得到的分类准确率高,但分类的时间较长。Chen P W等人[5]提出使用遗传算法优化核参数g和惩罚参数C,该方法属于启发式算法,不必遍历网格内所有的参数点,也能找到全局最优解。然而在实际应用中,存在SVM模型参数寻优范围不确定的问题,参数的寻优范围设置过大时,训练的时间过长,参数的寻优范围设置过小时,分类的准确率难以得到保证。Eberhart R等人[6]提出使用粒子群算法来对这两个参数进行寻优,虽然分类的准确率较遗传算法有所提升,但由于复杂度提升了反而训练时间比遗传算法长。

本文优化了现有遗传算法对核参数g和惩罚参数C的寻优范围,首先在大范围中快速找到最佳参数组的粗值,然后通过缩小和放大这个参数组的粗值,得到遗传算法参数的最小寻优范围,这样不仅可以快速得到最佳的参数组,而且对不同的数据集不必每次都重新设定遗传算法参数的寻优范围。并且由于通过网格搜索法得到的是最佳参数的最小值范围,所以得到

收稿日期:2016-03-30; **修回日期**:2016-04-18。
基金项目:国家"973计划"资助项目(2013CB328903);国家自然科学基金资助项目(61403265)
作者简介:孟 滔(1987-),男,四川成都人,硕士研究生,主要从事智能控制方向的研究。
周新志(1966-),男,四川德阳人,教授,硕士研究生导师,主要从事智能控制方向的研究。

的惩罚参数 C 值较小，具有较好的泛化能力。实验表明该算法比网格搜索法、常规遗传算法、粒子群算法具有更好的训练速度，并且分类精度较常规遗传算法有所提升。

1 支持向量机 SVM

支持向量机是由寻找线性可分情况下的最优分类平面发展而来的[7]，主要思想是建立一个分类超平面作为决策面，使得正例和反例之间的隔离边缘被最大化。其中分类超平面就是求函数：

$$\varphi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(w\cdot w) \tag{1}$$

$s.t\quad yi(w\cdot xi+b)\geqslant 1,\forall i=1,2,\ldots,n$ 的最小值。其中 w 是超平面的法向量，b 是超平面的常数项，xi 为训练样本，yi 为样本的类别。

支持向量机进行多分类时，数据集往往是非线性且不可分的，需要在约束条件中引入松弛因子，在目标函数中加入惩罚参数 C 以及引入核函数将样本映射到一个新的空间，使其在新的空间里可以实现线性可分来解决这一问题。其中，惩罚参数 C 用于控制模型复杂度和逼近误差的折中。若 C 过大，就会引起过学习，影响分类器的泛化能力。

目前实际工程中广泛使用的核函数主要有线性核函数（Linear），多项式核函数（Poly），径向基核函数（Radial Basis Function，RBF），两层感知器核函数（Sigmoid）四类[8]。其中径向基核函数是目前应用最广泛的核函数，文中也将采用此核函数。其形式如下：

$$K(x\cdot xi)=\exp(-g\|x-xi\|^2),g>0 \tag{2}$$

其中，g 为核函数中的重要参数，影响着 SVM 分类算法的复杂程度。

2 自适应遗传算法寻优方法

遗传算法是一种通过模拟自然进化过程搜索最优解的方法，经过数代选择、遗传、变异，最后得到最适应环境的种群[9]。将遗传算法引入到 SVM 参数寻优当中，较之传统的网格搜索法具有更高的分类速度，且当参数的寻优范围设置恰当时，其分类准确度也比较高。

2.1 传统遗传算法

传统的遗传算法对参数的寻优范围往往是基于学者的自身经验进行设定的。收敛速度慢且容易出现种群早熟现象或陷入局部极值点，不能保证收敛到全局最优解。在最大进化代数和总群数相同的情况下，寻优的范围越小得到的结果越好，且寻优的时间也越短。而当寻优的范围足够小但最佳的参数不在这个范围时，得到的结果却不是最佳的。故如何确定这个范围是提高遗传算法寻优的关键问题。并且针对不同的数据集，寻优范围如果是固定在一个范围，得到的结果不一定是最佳的。故能自适应的设置参数寻优的范围是解决此问题的关键。

2.2 自适应遗传算法

通过网格搜索法对 wine 数据（来自 UCI）进行预测，可以得到图 1。由图 1 可知，惩罚参数 C 和核函数参数 g 只有在一定的范围内对应的训练集分类准确率很高，但是绝大部分的范围内，分类准确率都比较低。如能先定位出比较好的参数寻优区间，再进行精确搜索，就能够减少不必要的计算，提高寻优的速度。针对以上问题，提出一种自适应遗传算法。首先，网格搜索法在较大范围内采用较大步距进行粗搜，选择使分类准确率最高的一组 C 和 g。如果参数选择过程中有多组的 C 和 g 对应于最高的验证分类准确率，则选择能够达到最高验证分类准确率中参数 C 最小的这组 C 和 g 做为粗搜的基准值。因为 C 过大时会导致过学习状态发生，即训练集分类准确率很高而测试集分类准确率很低，故选择 C 最小的那组参数做为最佳的选择对象，这样就解决了参数寻优范围的问题。图 2 就是利用该方法后同样通过网格搜索法得到的结果，可以看到参数的寻优范围大大缩小了。

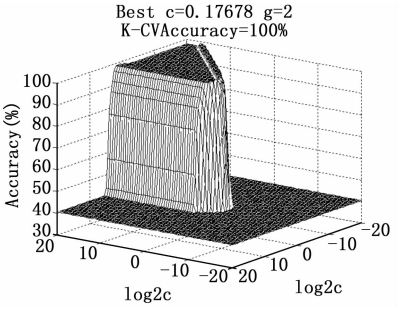

图 1 网格搜索法寻优结果

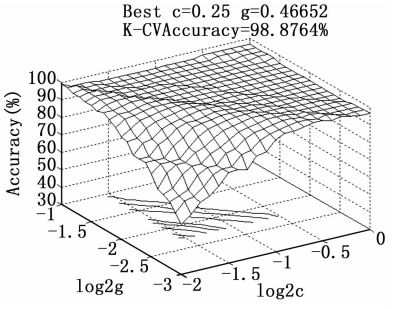

图 2 优化后网格搜索法寻优结果

解决了参数寻优的范围，但如何保证最佳的参数在这个区间内。为此首先来分析参数 C 和径向基核参数 g 在 SVM 中存在的相互影响关系[10]。当 C 值过小时，SVM 预测精度较低，易出现欠学习状态；随着 C 值增大，预测精度逐渐提高，但当 C 超过一定值时又会出现过学习状态；若此时 g 值亦随之增大，则可平衡 C 值产生的影响。同理 g 值过大时，也会出现过学习或欠学习状态。因此，参数 C 和 g 共同作用时，理论上存在一个有效区域，在该区域中存在一对预测结果最佳的参数组合。通过上述可知，这组参数存在着相互制约的关系，即可以认为这组参数不会太大，通过 UCI 上 10 组经典测试得到了验证，一般都小于 100。为了更加有效的找到这组参数，将 C 和 g 参数的粗寻优范围设置的足够大，这样就保证了最佳参数在粗寻优的区间内。

2.3 基于自适应遗传算法的 SVM 参数优化的具体算法实现

自适应遗传算法的步骤如下：

1）载入需要进行测试的数据集包括数据的特征属性和分类类别；

2）随机生成指定数据集的训练数据和预测数据；

3）对训练和预测数据按照式（3）进行归一化处理。

$$f:x\rightarrow y=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{3}$$

其中：$x_{\min}$ 表示此列特征维中最小值；$x_{\max}$ 表示此列特征维中最大值；y 表示 x 被归一化后结果。

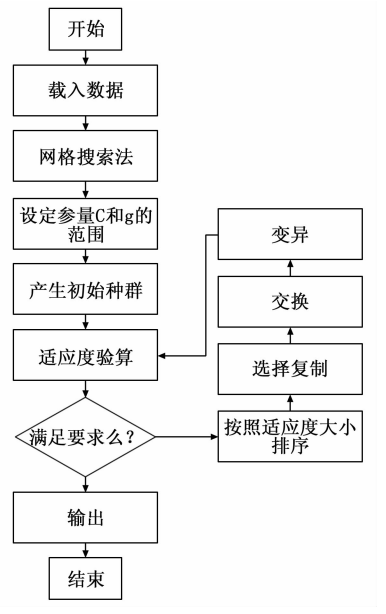

图 3　自适应遗传算法过程图

4）设定网格搜索的变量（C，g）的范围以及搜索的步距。其中 C 的初始范围设置为 $[2^{-5}, 2^{5}]$，g 的初始范围设置为 $[2^{-10}, 2^{10}]$。传统的网格搜索法的步距一般设置为 0.1，而本文方法中采用步距为 2，为传统步距的 20 倍，大大提升了搜索的时间。

5）利用网格搜索法，初步求出变量（C，g）的值即 bestc，bestg。此值作为遗传算法参数 C 和 g 的基准值。

6）设定遗传算法的参量，其中 C 的范围设定为 [0.5 * bestc，2 * bestc]，g 的范围设定为 [0.5 * bestg，2 * bestg]。传统的遗传算法 C 和 g 的寻优范围都设定为 [0，100]。

7）利用遗传算法求取预测数据的分类准确率。

3　仿真结果分析

实验使用的计算机平台为 windows 7（64 位）系统，处理器为 Core（TM）i5，内存为 8 GB。使用的软件仿真平台为 matlab2010b，采用台湾大学林智仁（Chih－Jen Lin）博士等开发设计的 LIBSVM 工具包进行测试。核函数采用用于最广泛的径向基核函数（Radial Basis Function，RBF）。分别采用网格搜索法，传统遗传算法，粒子群算法，自适应遗传算法进行对 UCI 上典型的 10 组数据（见表 1 所示）进行测试。测试结果见表 2 和表 3 所示。

表 1　10 组经典数据集

数据名称	训练规模	测试规模	特征维数	分类数目
iris	100	50	4	3
wine	89	89	13	3
glass	150	64	9	6
fourclass	400	462	2	2
svmguide4	200	100	10	3
vowel	300	228	10	11
german	500	500	24	2
segment	1500	810	19	7
page－blocks	3000	2473	10	5
satimage	3000	1435	36	6

从时间效率方面，从表 2 可以明显看出，传统的网格搜索法耗时最长，因为它是遍历大范围的所有点进行寻优，且在该范围内大多数点都不是最佳的点，故其寻优时间最长。而传统的遗传算法由于是固定参数的搜索范围，往往范围过大，导致故其搜索的时间也较长。本文提出的自适应遗传算法耗时最少，因为其参数的寻优范围大大缩小了，特别是在数据集越加复杂时，训练的速度更加明显。

从分类精度方面，从表 3 可以看出，网格搜索法在数据集较简单的情况下得到的分类准确率往往较高，而当数据集越加复杂时自适应遗传算法的分类准确率较高，且往往比常规遗传算法高，因为常规遗传算法是基于经验设置的参数寻优范围，在相同的情况下搜索的不够精细，往往搜索到的 C 值过大，即容易导致过学习状态的发生，而采用本文提出的自适应遗传算法因为从网格搜索法得到的参数范围的值较小，可以有效的减小寻优得到的参数 C，从而改善这种情况，即对分类的准确率有所提升。

表 2　预测结果对比

数据名称	网格搜索法		常规遗传算法		粒子群算法		自适应遗传算法	
	准确率/%	训练时间/s	准确率/%	训练时间/s	准确率/%	训练时间/s	准确率/%	训练时间/s
iris	98	68.5804	98	1.6833	98	2.998	100	0.8875
wine	96.6292	90.6902	95.5056	3.2784	95.5056	3.8964	95.5056	0.7696
glass	73.4375	249.9086	75	7.3312	76.5625	11.446	76.5625	2.8647
fourclass	100	508.0341	100	3.8653	100	19.186	100	3.1565
svmguide4	73	445.7502	67	7.5879	67	17.6752	69	5.346
vowel	96.9298	1090.8849	97.3684	37.1449	97.807	41.5873	98.6842	10.274
german	72.8	3187.1217	71.6	33.9676	73.4	106.66	74	22.339
segment		＞3600	97.037	189.9301	96.1728	469.9445	97.4074	80.59
page－blocks		＞3600	89.81	365.6175	89.6886	383.1883	89.6886	156.744
satimage		＞3600	90.5923	1291.8063		＞3600	91.0105	687.8736

（下转第 223 页）

[A]. IEEE International Conference on Fuzz Systems [C]. Jeju Island, Korea: IEEE, 2009: 660 - 665.
[4] An K, Chen Q J. Passive dynamic model for walking down stairs [A]. 25th Control and Decision Conference (CCDC) [C]. Guiyang, GuiZhou, China: IEEE, 2013: 3166 - 3172.
[5] 付根平，杨宜民，陈建平，等. 基于 ZMP 误差校正的仿人机器人步行控制 [J]. 机器人，2013，35 (1)：39 - 44.
[6] 李　诚，张奇志，周亚丽. 半被动双足机器人控制系统设计 [J]. 计算及测量与控制，2015，23 (11)：3651 - 3653.
[7] Lim I S, Kwon O, Park J H. Gait optimization of biped robots based on human motion analysis [J]. Robotics and Autonomous Systems, 2014, 62: 229 - 240.
[8] Paola C F, Ruben A, Salim G, et al. Gait event detection during stair walking using a rate gyroscope [J]. Sensor, 2014, 14: 5470 - 5485.
[9] Yuan Q L, Chen I M, Lee S P. SLAC: 3D localization of human based on kinetic human movement capture [A]. IEEE International Conference on Robotics and Automation [C]. Shanghai, China: IEEE, 2011: 848 - 853.
[10] Matthew J P, Zhao H H, Aaron D A. Motion primitives for human－inspired bipedal robotic locomotion: walking and stair climbing [A]. IEEE International Conference on Robotics and Automation [C]. Saint Paul, Minnesota, USA: IEEE, 2012: 543 - 549.
[11] Park C S, Ha T S, Kim J H, et al. Trajectory generation and control for a biped robot walking upstairs [J]. International Journal of Control, Automation and Systems, 2010, 8 (2): 339 - 351.
[12] 柯显信，龚振邦，吴家麒. 双足机器人上楼梯步态规划的复现性要求 [J]. 应用科学学报，2003，21 (1)：63 - 67.
[13] GB/T 15759－1995，人体模板设计和使用要求 [S]. 北京：中国标准出版社，1996.
[14] GB/T 17245－2004，成年人人体惯性参数 [S]. 北京：中国标准出版社，2004.
[15] Bradley S M, Hernandez C R. Geriatric assistive devices [J]. American Family Physician, 2011, 84 (4): 405 - 411.
[16] 邱启祥，王小敏，刘　浩. 助行设备在美国老年人群中防跌倒作用的现状分析 [J]. 中国康复理论与实践，2014，20 (1)：85 - 87.
[17] 朱卫娟，王　彤. 不同步行辅助器对健康人负重影响的分析 [J]. 中国伤残医学，2010，18 (5)：109 - 111.

(上接第 217 页)

表 3　预测结果对比

数据名称	网格搜索法		常规遗传算法		粒子群算法		自适应遗传算法	
	C 值	准确率/%	C 值	准确率/%	C 值	准确率/%	C 值	准确率/%
iris	5.6569	98	42.1542	98	8.3945	98	4.3576	100
wine	1.1487	96.6292	71.53	95.5056	35.4669	95.5056	13.9654	95.5056
glass	194.0117	73.4375	85.0329	75	6.8443	76.5625	15.9507	76.5625
fourclass	0.7071	100	85.1189	100	0.861	100	2.1744	100
svmguide4	776.0469	73	87.0764	67	100	67	59.068	69
vowel	2.4623	96.9298	24.4508	97.3684	2.8849	97.807	7.1139	98.6842
german	238.8564	72.8	82.9427	71.6	2.3701	73.4	3.658	74
segment			80.0284	97.037	9.1393	96.1728	49.422	97.4074
page－blocks			96.299	89.81	0.1	89.6886	3.2325	89.6886
satimage			7.3151	90.5923			4.8064	91.0105

4　结论

针对常规遗传算法存在 SVM 模型参数搜索范围不确定，导致参数搜索时间过长，分类准确率较网格搜索法有所降低等问题，基于核参数 g 和惩罚参数 C 之间的相互关系，通过网格搜索法确定最佳参数的最小寻优范围，有效地帮助常规遗传算法避免陷入局部最优解，为遗传算法的参数范围设置，提供了有效的方法，保证了搜索的效率，并改善了基于常规遗传算法得到的惩罚参数 C 过大，导致的分类准确率较低的问题。最后通过网格搜索法、常规遗传算法、粒子群算法以及自适应遗传算法对比实验。表明该算法具有更快的寻优速度，能较好的解决常规遗传算法参数设置不确定性导致的寻优较慢以及分类精度较网格搜索法有所降低等问题。

参考文献：

[1] Cortes C, Vapnic V. Support－vector networks [J]. Machine Learning. 1995, 20 (3): 273－297.
[2] Vapnik V, Levin E, Cun Y L. Measuring The VC－dimension of a learning machines [J]. Neural Computation, 1994, 6 (5): 851 - 876.
[3] 王　鹏，朱小燕. 基于 RBF 核的 SVM 的模型选择及其应用 [J]. 计算机工程与应用，2003，39 (24)：72 - 73.
[4] Liu Xianglou, Jia Dongxu, Li Hui. Research on Kernel parameter optimization of support vector machine in speaker recognition [J]. Science Technology And Engineering, 2010, 10 (7): 1669 - 1673.
[5] Chen P W, Wang J Y, Lee H. Model selection of SVMs using GA approach [A]. Proc of 2004 IEEE Int Joint Conf on Neural Networks [C]. Piscataway, USA, 2004: 2035 - 2040.
[6] Eberhart R, Kenney J. A new optimizer using particle swarm theory [A]. Proc of the sixth Internation Symposium on Micro Machine and Human Science [C]. Piscataway, USA, 1995: 39 - 43.
[7] Cevikalp H. New clustering algorithms for the support vector machine based hierarchical classification [J]. Pattern Recognition Letters, 2010, 31: 1285.
[8] 杨志民，刘广利. 不确定性支持向量机－算法及应用 [M]. 北京：科技出版社，2012.
[9] 朱庆生，程　柯. 一种基于累积适应度遗传算法的 SVM 多分类决策树 [J]. 计算机应用研究，2016，33 (1)：64－67.
[10] Keerthiss, Lin C J. Asymptotic behaviors of support vector machines with Gaussian kernel [J]. Neural Computation, 2003, 15 (7): 1667 - 1689.