维普资讯 http://www.cqvip.com

一种基于免疫算子的SVM算法*)

刘 芳 梁雪峰
(西安电子科技大学计算机学院 西安710071)

摘 要 SVM是一种基于核函数的机器学习算法,因为它具有良好的推广性和较好的性能,所以成为近些年来大家所关注的热点,但是该算法存在两个问题:一、如何提高SVM的计算精度;二、如何减少计算时间。本文提出一种使用免疫算子的SVM算法,该算法不但能够提高SVM的性能使其更加接近于实际问题,还能避免因问题太复杂使得结果不是最优解的情况。文中最后对样本进行了实验,结果说明了使用免疫算子的方法比经典方法在分类效果上有明显提高。

关键词 支撑矢量机,核函数,疫苗,免疫算子

An Immune Operation Based SVM Algorithm

LIU Fang LIANG Xue-Feng
(Computer School of Xidian University, Xi'an, China 710071)

Abstract In recent years, the researches on SVM focus on two main areas. One is to improve the precision of the SVM algorithm, and another is to improve its speed. In this paper, a new method, which can appropriately tune multiple parameters in the kernel functions of SVM, is proposed. It cannot only improve the algorithm performance and make it approach to the real problem, but also avoid those methods available are too complex, the kernel must be differential and the result may be not optimal. Simulation results for data show the result based on this method is improved much more than normal algorithm's.

Keywords Support vector machines (SVM), Kernel function, Vaccine, Immune operation

1 引言

在20世纪90年代,Vapnik等人通过发展统计学习理论,在其基础上提出了一种新的模式识别的方法——支撑矢量机,它在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势,并能够推广应用到函数拟合、回归等其他机器学习问题中,它正成为神经网络研究之后机器学习领域新的研究热点[1,2]。但是支撑矢量机方法还存在很多问题要进一步研究,如:核函数的选择问题。在实际应用中人们只能从现有的几个核函数中选出一个性能较好的,应用到实际问题中。但是如果问题较复杂,或数据量较大,现有的核函数不能保证能够很好地进行分类,甚者于不能把样本映射到一个能够线性可分的高维空间,这使得很难把传统的SVM应用到复杂的实际问题中。

近来,不少研究者对核函数进行了大量的研究,提出了多种新的核函数并进行了分类[4],然而这些核函数通常较为复杂,参数也较多,对于不同的问题参数要作相应的调整才能使其获得较好的分类效果。Olivier Bousquet 和 Vladimir Vapnik 曾提出使用梯度下降法来求参数[6],但这种方法受到核函数必须可导的限制。本人也曾提出使用遗传算法来求得参数[7],此方法适用于各种核函数,结果也能得到最优解,但迭代的次数较多,耗时较长。

免疫算法(Immune Algorithm)是借鉴了生物学相关内容和知识[3],特别是遗传学方面的理论概念提出的一种智能算法。它是在遗传算法的基础上加入了免疫算子,建立了新的进化理论,提高了算法的整体性能。本文提出了一种使用免疫算子的方法来提取核函数的参数。

2 SVM参数的选择机理

众所周知,决定SVM性能的关键因素是选择一个适当的核函数,然而,在实际应用中只有少数几个较简单的核函数可供选择,如果有一个较好的方法能找出精确合适的函数参数,则复杂的核函数就可以广泛地使用。这种方法还有一个潜在的优点,即:使调整问题的特性成为可能。当然,在对问题特性没有任何先验知识的情况下,我们只能选择球形核(也就是对每一种特性赋予相同的权重),但是实际的数据都包含有其不同的自然特性[6],因此我们希望找到一个合适的核形来对应实际问题。

在模式识别中学习算法通常依赖于参数,类型F的大小和搜索路径也由这些参数控制。为了得到较好的分类效果,我们应最小化算法的误差,误差可由测试一些没有学习过的数据或有理论分析给定的界来估计,但是,误差的估计不是这些参数的显函数,因此通常的做法是遍历所有的参数空间,但是这种遍历的方法在实际应用中是不可取的。

所有常规的搜索方法都是单点搜索,但是这种点对点的搜索结果常常会落入局部最优解,使得SVM不能正确地分类。因此需要一种更有效的方法来进行参数选择。

3 SVM多参数选择的免疫算法

3.1 免疫算法

*)国家自然科学基金(60073053和60133010)资助项目。

维普资讯 http://www.cqvip.com

从理论上分析遗传算法，在迭代过程中保留上一代最佳个体的前提下，它是全局收敛的。但是其两个主要的遗传算子（交叉、变异）都是在一定概率的条件下随机地、没有指导地操作，因此它们在为群体中的个体提供进化机会的同时，也不可避免地导致了退化的可能。另一方面，有些问题有一些自身的特征信息，若能充分地利用这些特征信息对于求解问题有着重要的辅助作用。

免疫算法实践免疫的概念及其理论应用于遗传算法，在保留原算法优良特性的前提下，力图有选择、有目的地利用待求问题中的一些特征信息或知识来抑制其迭代过程中出现的退化现象[3]。免疫算法的核心在于免疫算子的构造，免疫算子是在原有遗传算法的框架基础上引入的一个新的算子。

实际操作过程中，首先对所求解的问题（即抗原，Antigen）进行具体分析，从中提取最基本的特征信息（即疫苗，Vaccine）；其次，对特征信息进行处理，将其转化为求解问题的一种方案（即抗体，Antibody）；最后，将此方案以适当的形式装化为免疫算子以实施具体的操作。

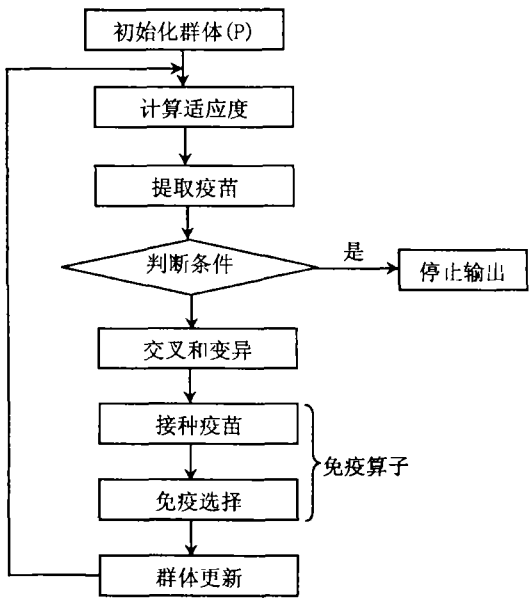
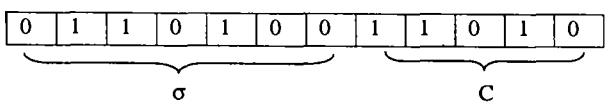


图1　算法流程图

3.2　基于免疫算子提取 SVM 参数

本文为了便于说明，使用较简单的 RBF 核的 SVM。RBF 核 SVM 的参数有两个，即：σ，C，现使用两组测试数据。这两组数据在使用 RBF 核的 SVM 进行分类时，若使用较常用的参数 σ[3，6]，C[7，11]，最好的结果也总有两个错分点，这个结果说明了两个可能：1. RBF 核的 SVM 不可全部正确分开此数据；2. 没有找到最优的 RBF 核 SVM 的参数，使其不能全部正确分类。此处我们使用免疫算子来提取 SVM 的参数。其执行步骤如下：

1. 编码　对于 RBF 核 SVM 的参数 σ，C，它们可能的取值范围分别是[1，10]，[1，20]。现给定其精度分别为0.1、1，所以编码长度分别为7，5，总长度为12。



2. 适应度函数　该算法的适应度可用分类边界的大小或正确分类的个数作为适应度。在本文中经过试验得出正确分类的个数更加能够体现适应度，所以仿真试验中使用的是正确分类的个数。

3. 交叉和变异算子　对于交叉操作的设计，我们采用两点交叉，其位置随机确定。变异算子是保证群体都言行的一个关键部分。在迭代初期，应增大变异概率来保证多样性，但在迭代的后期应减小变异概率，来保证算法的收敛。所以我们采用了随着代数增加而概率减小的方法，具体关系为：

$$p=C_2\frac{\exp(-\frac{t}{2}C_1)}{m*\sqrt{n}}$$

其中 P 是变异概率，t 是代数，m 是群体个数，n 是个体长度，C_1，C_2是调整参数。

4. 免疫算子　由于在迭代的过程中不断产生更优解，我们就把更优解的前几位作为疫苗，为遗传操作后的某些个体进行疫苗注射，接下来作免疫检测。若适应度提高了，就保留注射过疫苗的个体。反之，则说明在交叉和变异过程中出现了严重退化现象，这时保留父代。免疫选择主要使用退火选择，即：子代个体以退火选择计算其被选中的概率。

$$p(x_i)=\frac{e^{\frac{f(x_i)}{T_k}}}{\sum_{i=1}^{m}e^{\frac{f(x_i)}{T_k}}}$$

其中 $f(x_i)$为个体适应度，$\{T_k\}$是趋近于0的温度控制序列。

5. 判断条件　可以使用两种判断条件，一是固定的迭代次数。二是个体都能达到某一较高的适应度。本文中使用的是第二种方法。

4　仿真试验

在仿真试验中，我们对 RBF 核的 SVM 进行参数搜索。这里分别用本人曾提出的遗传方法和本文提出的免疫方法作比较。结果如下：

表1　RBF 核的 SVM 参数选择结果

	代数	时间	结果	错分数
遗传方法	16～22代	51～62秒	σ=6.5，C=1	1
免疫方法	4～5代	10～15秒	σ=6.6，C=1	1

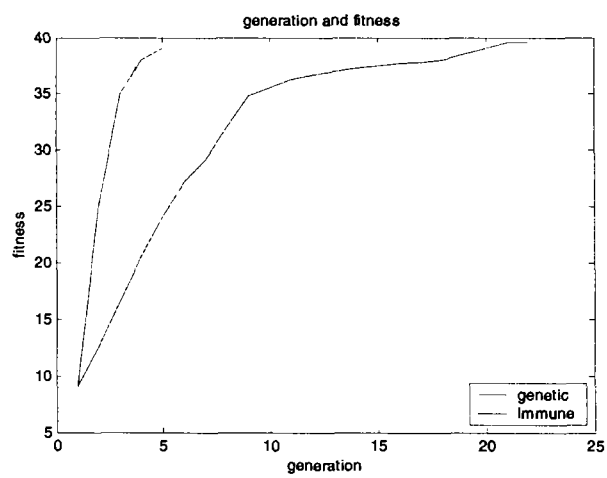


图2　遗传和免疫方法适应度的对比

我们可从结果中看到使用免疫算子的方法可提高算法的效率，但是因为问题较简单，效果不是很明显，所以我们又对复杂核做了一次试验。这里我们使用的核矩阵[8,9]为：

（下转第119页）

维普资讯 http://www.cqvip.com

5. 对每一个 $c \in MC_k$
   计算 $SIG(\{c\},RED,D)=\gamma(RED\cup\{c\},D)-\gamma(RED,D)$;
6. $MC_k=0$? 如果是,转(9)
7. 选取 SIG 最大的 c_m,$MC_k=MC_k\backslash\{c_m\}$;
   $RED=RED\cup\{c_m\}$
8 如果 $SIG(c_m,RED,D)=0$,找到一个属性约简,转(10);否则转(6)
9. $k=\max(j)$? 如果是,转(10),
   否则,$k=k+1$,转(5)
10. 输出 RED,算法终止。

算法性能分析:算法第一步的时间耗费是 $O(|C|)$,算法第二步最坏时间复杂度是 $O(|C|n)$,算法第三步时间复杂度是 $O(|C|)$,算法第四步最坏时间复杂度为 $O(|C|^2)$,整个算法时间复杂度为 $O(|C|^2+|C|n)$。

4 实例

以表1为例来验证算法。选取域值 $\beta=1/3$。由表1可知,$M(C)=0$,$M(a)=M(d)=1$,$M(b)=3$,则首先选取 c,但 c 不满足条件(*),就继续选取 a 或者 d,由于 $SIG(a,\{c\},D)=1/3$,$SIG(d,\{c\},D)=0$,因此优先选取 a,$\{a,c\}$还是不满足条件(*),于是选取 d,条件满足,算法中止。于是得到一个约简 $\{a,c,d\}$。b 是可省略的。

5 与传统算法的比较

5.1 算法性能的比较

属性约简算法由于涉及的数据量比较大,算法的时间开销和空间开销也比较大。因此算法的时间复杂度应尽量控制在多项式范围内,而且次数越低越好。在经典算法中,许多算法一般控制在 $O(n^2)$或者 $O(n^3)$这个数量级上。本论文提出的算法大约是 $O(|C|^2+|C|n)$,依然在多项式范围内,是可以接受的。

5.2 适用范围比较

传统算法都是基于经典模型,以等价关系为基础的。它们在处理信息系统的时候有一定的局限性,基于改进模型的算法则给予了较大的灵活性。它实际上是经典模型的扩展,当 $\beta=1$ 时,就变成传统的等价关系;$\beta=0$ 时就变成 M. Kryszkiewcz 提出的容差关系。它借鉴了量化容差关系的思想,使算法既可以处理完备信息系统,也可以处理不完备信息系统,只需要改变 β 值即可,而不用更换算法,因此具有较大的灵活性。

参考文献

1 曾黄麟. 粗集理论极其应用——关于数据推理的新方法. 重庆:重庆大学出版社,1988
2 王国胤. Rough 集理论与知识获取. 西安:西安交通大学出版社,2001
3 张文修,等. 粗糙集理论与方法. 北京:科学出版社,2001
4 Liang Jiye, Xu Zongben. Uncertainty Measures of Roughness of Knowledge and Rough Sets in Incomplete Information Systems. In:IEEE Proc. of the 3rd World Congress on Intelligent and Automation,2000. 526～2629
5 赵卫东,吴明赞. 不完全信息下的粗集拓展. 计算机科学,2002(9. 专刊)
6 Pawlak Z. Rough set theory and its applications to data analysis [J]. Cybernetics and System,1998,29(7):661～688
7 石峰,娄臻亮,张永清. 一种改进的粗糙集属性约简启发式算法. 上海交通大学学报,2002,26(4):478～481

(上接第110页)

$k=\sum_{i=1}^{n}a_ik_i$,其中 k_1为线性核函数 $k_1(\chi_1,\chi_2)=\chi_1^T,\chi_2$,$k_2$为多项式核函数 $k_2(\chi_1,\chi_2)=(1+\chi_1^T\chi_2)^d$,$k_3$为高斯核函数 $k_3(\chi_1,\chi_2)=\exp(-0.5(\chi_1-\chi_2)^T(\chi_1-\chi_2)/\sigma)$,$a_i$ 为各个核函数的权重。这时的 SVM 的参数有六个,分别为:a_1,a_2,a_3,d,σ,C。这时的计算结果如下:

表2 线性组合核的 SVM 参数选择结果

	代数	时间	结果	错分数
遗传方法	116～132代	351～423秒	$a_1=0.18,a_2=0.23,$ $a_3=0.59,\sigma=0.5,C=1$	0
免疫方法	14～23代	37～49秒	$a_1=0.19,a_2=0.22,$ $a_3=0.59,\sigma=0.5,C=1$	0

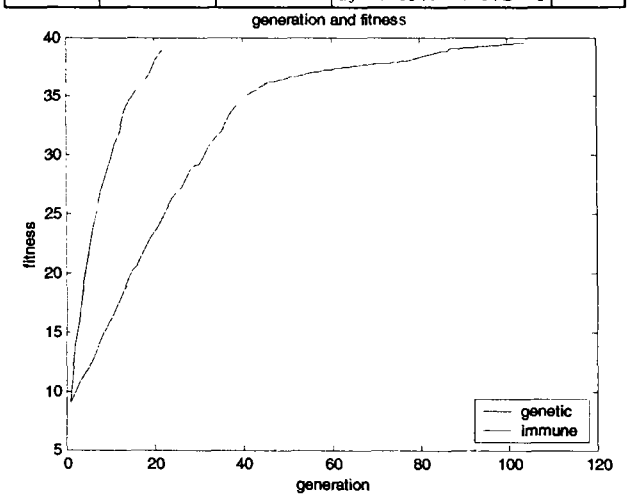


图3 遗传和免疫方法适应度的对比

从这个结果可以看出,对于复杂问题,免疫方法比遗传方法有较大的优势。

结论 从实验结果可以看出,免疫算法在遗传操作的过程中能够自动地找出局部的最优解作为疫苗注射给子代的个体,使得搜索空间在算法的中后部分大大减小,问题也就能够以较快的速度收敛,但又保持了种群的多样性,不会使结果落入局部最优解,较好地解决了复杂核函数的多参数选择的问题。

参考文献

1 边肇祺,张学工. 模式识别. 清华大学出版社
2 Vapnik V N. The Nature of Statistical Learning Theory. New York: Springer- Verlag
3 王磊. 免疫进化计算及应用:[西安电子科技大学博士论文]
4 Genton M G. Classes of Kernels for Machine Learning: A Statistics Perspective. Department of Statistics North Carolina State University
5 Christopher J.Burges C. A Tutorial on Support Vector Machines for Pattern Recognition. Microsoft Research
6 Chapelle O.Vapnik V. Choosing Multiple Parameters for Support Vector Machines. AT&T Research Labs
7 Liang Xue-feng, Liu Fang. Choosing Multiple Parameters for SVM Based on Genetic Algorithm. In: Intl. Conf. on Signal Processing Proc.
8 刘芳,梁雪峰. 一种基于线性组合核的 SVM 算法. 计算机科学
9 程云鹏,张凯院. 矩阵论. 西北工业大学出版社