文章编号:1671 - 6833(2011)01 - 0075 - 05

SVM 算法的区间自适应 PSO 优化及其应用

王 杰,姜 念,张 毅

(郑州大学 电气工程学院,河南 郑州 450001)

摘 要:核(Kernel)参数选取对支持向量机的推广能力和泛化能力有至关重要的作用,尤其是在对大量数据进行识别分类时,需要占用计算机大量内存,SVM 参数优化速度明显缓慢,从而影响了整个系统性能.针对此问题,笔者提出一种区间自适应粒子群算法来优化 SVM 参数,粒子根据实际情况动态平衡其全局搜索与局部搜索能力,提高了参数优化速度.在入侵检测系统的应用中,与蚁群算法、遗传算法优化 SVM 参数的结果进行比较.实验证明:此方法分类精度提高约 9.7%,响应时间缩短约 40.6% ~56.5%,具有较大的优势.

关键词:支持向量机;自适应;粒子群优化算法;入侵检测

中图分类号:TP181,TP309.1 **文献标志码**:A

0 引言

1963 年,由 Vapnik 领导的 AT&Tbell 实验室研究小组提出了一种新的分类技术——支持向量机(Support Vector Machine, SVM),它是以统计学习理论为基础的一种新型机器学习方法,具有小样本学习能力强、非线性处理能力好、数学理论基础较严密等特点[1].由于支持向量机的内部是基于使结构风险最小化原理的,其泛化能力极强,能有效地避免过学习、局部极小点以及"维数灾"等问题.在模式识别、回归分析、生物信息技术、医学研究以及其它的一些领域得到了广泛的应用[2-5].

SVM 是基于核的学习,其性能主要依赖于核参数的选取,不少学者对参数的选择问题进行了深入的研究,例如,Chapelle 和 Vapnik 的支撑和特征空间的重构方法,Claeskens 的选择信息准则,Debruyne 的影响函数方法,尤其是近几年里流行的进化算法,像遗传算法(Gene Algorithm, GA)、蚁群算法(Ant Colony Algorithm, ACA)、粒子群算法等,与粒子群算法相比:遗传算法[6-7]中,染色体互相共享信息,整个种群的移动是比较均匀地向最优区域移动的,而粒子群算法是单向的信息流动,所有的粒子在大多数情况下可能更快地收敛于最优值,粒子群算法没有使用"适者生存"的概念,没有遗传算法的交叉和变异,形式简单;蚁群算法[7-8]和粒子群算法虽然都同属于仿生群体智能优化算法,但蚁群算法中,每个个体只能感知局部的信息,不能直接利用全局信息,一般需要较长的搜索时间,且容易出现停滞现象,它的收敛性能对初始化参数的设置更加敏感,其复杂度也要高于粒子群算法.

目前,SVM 参数优化速度缓慢的问题突出,尤其是面对大量数据优化时,SVM 训练速度以及参数优化速度缓慢,笔者针对这个问题,根据上面分析,提出了一种区间自适应粒子群算法来优化 SVM 参数,并将其用于入侵检测系统中,实验证明此方法对参数优化速度快且分类精度高.

1 区间自适应粒子群算法

粒子群算法(Particle Swarm Optimization, PSO)[9]是一种新兴的群体智能进化算法,它是由 Kennedy 和 Eberhart 于 1995 年提出的,采用下列公式对每个粒子进行操作:

$$v_{i(d+1)} = wv_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \quad (1)$$

$$x_{i(d+1)} = x_{id} + v_{i(d+1)} \quad (2)$$

收稿日期:2010 - 08 - 10;**修订日期**:2010 - 09 - 27

基金项目:国家自然科学基金资助项目(60974005)/教育部博士点基金资助项目(20094101120008)/河南省杰出人才创新基金资助项目(074200510013)

作者简介:王杰(1959 -),河南周口人,郑州大学教授,博士,博士生导师,研究方向:智能控制与智能计算、信息与计算机网络安全,E-mail:wj@ zzu. edu. cn.

式中:w 是惯性权重因子;c_1,c_2 为学习因子,也称为加速因子;r_1 和 r_2 为[0,1]间的随机数,且 $v \leqslant v_{max}$. 研究发现在算法收敛的情况下,所有的粒子都向最优解的方向飞去,趋于同一化,这就使得在算法后期收敛速度明显变慢,且容易陷入局部最优,所能达到的精度也比较低.

惯性权重 w 的选择对算法性能起着至关重要的作用,较大的惯性权重 w 有利于跳出局部最小值,加强粒子群算法的全局搜索能力,而较小的惯性权重 w 能加速收敛,但容易陷入局部最优,因此,如何选择惯性权重 w 将是这个算法的重点. 在处理大量数据时,由于涉及到的实验数据多,变化跳度大,选取一般线性递减的惯性权重 w 已经不能满足试验的要求,因此,笔者提出一种区间自适应的粒子群优化算法(Interval Adaptive Particle Swarm Optimization,IAPSO). 给出下面三个定义:

定义一:粒子的平均聚焦距离

$$\text{meanDist} = \frac{\sum_{i=1}^{m}\sqrt{\sum_{i=1}^{D}(p_{gd}-x_{id})^2}}{m} \tag{3}$$

定义二:最大聚焦距离

$$\text{maxDist} = \max_{i=1,2,\cdots,m}\left(\sqrt{\sum_{i=1}^{D}(p_{gd}-x_{id})^2}\right) \tag{4}$$

定义三:粒子目前的聚焦距离变化率

$$k = \frac{\text{maxDist} - \text{meanDist}}{\text{maxDist}} \tag{5}$$

式中:m 为粒子的总个数;D 为空间维数;p_{gd} 为粒子群全局最优值.

在粒子群算法中,每迭代一次,我们就计算得到平均聚焦距离和最大聚焦距离,因此得出 k 的大小,据此判断粒子需提高全局搜索能力还是局部搜索能力,进而对惯性权重进行修正. 得到惯性权重由如下公式变化:

$$w' = \begin{cases} \left(a_1 + \dfrac{|r|}{2.0}\right) \times |\ln k|, & |k| > 1 \\ a_1 \times a_2 + \dfrac{|r|}{2.0}, & 0.05 \leqslant |k| \leqslant 1 \\ \left(a_2 + \dfrac{|r|}{2.0}\right) \times \dfrac{1}{|\ln k|}, & |k| < 0.05 \end{cases} \tag{6}$$

式中:$a_1 = 0.3$;$a_2 = 0.2$;r 为[0,1]间均匀分布的随机数.

因此,得到的区间自适应的粒子群优化算法形式为:

$$v_{i(d+1)} = w' v_{id} + c_1 r_1 (p_{id} - x_{id}) + c_2 r_2 (p_{gd} - x_{id}) \tag{7}$$

$$x_{i(d+1)} = x_{id} + v_{i(d+1)} \tag{8}$$

该方法随机地选取 w,使得 w 随着聚焦距离的变化率自适应地进行调整. 根据式(7)这一调节模式,能使得算法较好地适应复杂的实际环境,从而可以更加灵活地调节全局搜索能力与局部搜索能力. 当 k 变化较大时表明粒子的平均聚焦距离和最大聚焦距离相差较大,此时的粒子全局搜索能力较差,故应使粒子尽快地进入全局搜索模式,反之应该提高粒子的局部搜索能力.

2 基于区间自适应 PSO 优化的 SVM 算法

2.1 支持向量机理论

支持向量机最初是在模式分类中提出的,其基本思想是通过非线性变换 $\Phi(\cdot)$ 将输入空间映射到一个高维特征空间,在这个特征空间中,求取最大的分类间隔超平面 $f(x) = w^{\mathrm{T}}\Phi(x) + b$,其中 w 和 b 分别表示这个超平面的权值和阈值. 这样,两类中的数据可以更自由地活动,而产生错误的概率更小.

设训练集 $S = \{(x_k, y_k) \mid k = 1,2,\cdots,N\}$,其中 $x_k \in R^n$,$y_k \in R$ 分别表示输入和输出数据,SVM 构造如下的最小化目标函数及其约束条件:

$$\begin{cases} \min\limits_{\omega,b,\xi} \dfrac{1}{2}\|\omega\|^2 + C\sum\limits_{i=1}^{l}\xi_i \\ s.t. \quad y_i(\omega \cdot \phi(x_i) + b) \geqslant 1, \\ \qquad \xi_i \geqslant 0, i = 1,\cdots,l \end{cases} \tag{9}$$

式中:ω 垂直于超平面;C 为惩罚因子;松弛变量 ξ_i 表示与 x_i 违反约束的程度.

笔者采用径向基函数作为内积函数 $K(x, x_i) = \phi(x)^{\mathrm{T}}\phi(x_i) = \exp\lfloor -\gamma \| x - x_i \|^2 \rfloor$,最优分类面函数的实质是通过求解约束条件下的一个二次规划问题,最后得到的最优分类函数为:$f(x) = \text{sign}\{g(x)\} = \text{sign}\left(\sum\limits_{i=1}^{N}\alpha_i y_i \langle x, x_i\rangle + b^{op}\right)$. (10)

2.2 适应度函数选取

笔者提出的区间自适应 PSO 算法用来优化 SVM 的参数 C 和 γ,采用的粒子群适应度函数定义如下:

$$fitness = w_\alpha / inaccuracy \tag{11}$$

式中:w_α 为分类的精确度权重;inaccuracy 是 SVM 在训练样本集上的错分率,从上式可以看出,当 SVM 分类错误率越低,对应的适应度函数的值越大;SVM 分类错误率越高,对应的适应度函数的值也就越小.

2.3 SVM 参数优化步骤

区间自适应粒子群算法优化 SVM 参数的具

体流程如下：

Step1　初始化粒子种群，设定粒子群的参数，在 R^n 空间中随机产生 n 个粒子，$x_1, x_2, \cdots, x_n$，组成初始种群 $X(t)$，随机产生各个粒子的初始速度 $v_1, v_2, \cdots, v_n$，以及每个粒子的个体最优值 p_{best_i}；

Step2　将随机产生的粒子（即二进制串）对应 SVM 的参数 C 和 γ，然后调用支持向量机算法进行学习和训练，测试样本的错误分类率并记录，计算粒子的适应度；

Step3　对每个粒子，比较当前适应度 $f(x_i)$ 和历史最好适应度 $f(p_{best_i})$：如果 $f(x_i) < f(p_{best_i})$，那么 $p_{best_i} = x_i$；如果粒子种群中所有粒子的当前适应度 $f(x_i) < f(g_{best_i})$，那么 $g_{best_i} = x_i$；

Step4　根据式（7）、（8）更新粒子的速度和位置；

Step5　判断运行是否达到最大迭代次数，如果达到，则输出当前的最优参数 C, γ 以及分类错误率，否则返回 Step2 继续执行，直到满足判决条件为止；

Step6　将所得最优参数带入得到最优分类函数，再对待分类数据进行分类.

3　入侵检测模型建立

入侵检测技术通常有误用检测和异常检测两种，目前大多数入侵检测系统（Intrusion Detection System，IDS）采用的是基于误用的检测模式，即将每一个非法的或者是管理员所不允许的操作定义为一条规则，以一定的数据形式存放在规则库中，当处理数据的时候，将所获取的数据与规则库进行匹配，进而判断网络中的数据操作是否合法，用户再根据这些判断做出相应的报警、处理、反击等操作，以达到保护计算机安全的目的. 如图 1 所示是入侵检测系统架构图：

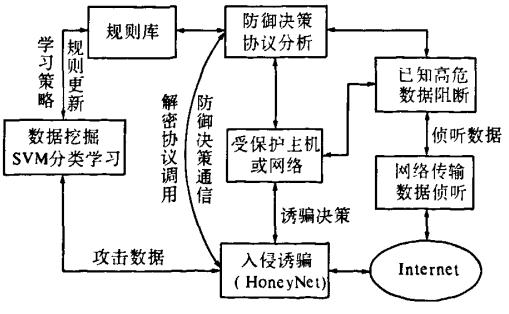


图 1　入侵检测系统构架图

Fig. 1　Framework of an IDS

基于支持向量机的入侵检测系统主要有 3 个部分：审计数据预处理器、支持向量机分类器和决策系统. 其框图如下图 2 所示：

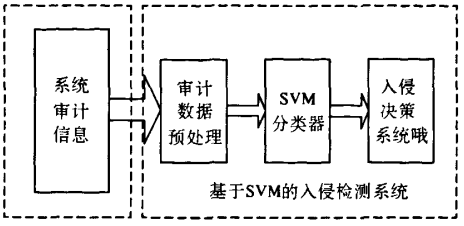


图 2　基于支持向量机的入侵检测系统框图

Fig. 2　Framework of an IDS based on SVM

入侵检测模型主要包括以下几个步骤：

Step1　样本信息的选取与预处理，样本集资料的选取应完备全面，尽量保证所选取的数据不重叠，将样本分为训练样本和测试样本两部分.

Step2　将样本数据进行归一化处理，使得输入的样本数据在[0,1]之间，归一化公式

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

x_{min} 和 x_{max} 分别为数据的最小值和最大值；

Step3　SVM 模型建立，在 SVM 模型中，输入数据经过属性约简后的样本信息，采用区间自适应粒子群优化算法迭代寻找支持向量机中的最优参数 C 和 γ，从而得到训练好的 SVM 识别模型；

Step4　识别输出，用训练好的识别模型对入侵检测数据的入侵行为进行识别分类，并输出结果.

4　实验仿真与分析

4.1　实验数据

笔者基于 MATLAB 语言编写的算法，实验中训练和测试的数据来源于麻省理工学院林肯实验室（MIT Lincoln Laboratory）在 1999 年发布的评价测试数据集[10]. 此数据集包括 24 种攻击，可以分为：Normal（正常连接）、DOS（拒绝服务攻击）、R2L（Remote to Local 攻击）、U2R（User to Root 攻击）和 Probing（探测攻击）.

4.2　实验环境与算法参数选取

笔者采用的仿真环境为方正 PC 机，内存为 512 MB，MATLAB 仿真软件.

实验中，粒子群算法的参数选取如下表 1 所示：

表1 区间自适应粒子群算法中粒子的参数设计

Tab.1 Particle swarm parameterdesign in adaptive PSD algorithm

参数	取值	参数	取值
学习因子 c_1	2	最大迭代次数	200
学习因子 c_2	2	步长	500
种群大小	40	判断误差 e	0.000 1
粒子最大速度	1		

4.3 实验结果与比较

自适应区间粒子群算法运用训练数据优化SVM得到的最优参数为 $C=4.625, \gamma=0.053$.

为了检验IAPSO的优越性,分别用区间自适应粒子群算法(IAPSO)、蚁群算法(ACA)和遗传算法(GA)对SVM的核参数进行优化,然后将训练后的SVM用于入侵检测系统中,检测网络数据的攻击行为.三次试验对网络入侵数据检测及分类结果如下表2.

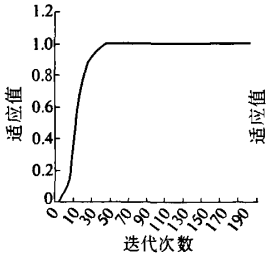
表2 算法优化SVM参数的网络入侵检测分类结果比较

Tab.2 Comparison of classification results of IDS with SVM optincised parameters

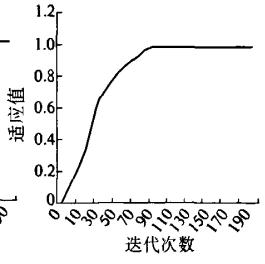
检测结果	IAPSO-SVM	GA-SVM	ACA-SVM
检测精度/%	97.325	96.610	88.732
误报率/%	1.136	1.338	3.315
漏报率/%	1.539	2.052	7.953
平均检测时间/s	32.6	74.8	54.8

从实验结果可以看出,区间自适应粒子群算法优化SVM核参数后,用于入侵检测模型中,对网络入侵数据的检测精度等各项性能指标都要优于其他两种算法优化后的效果.具体是:在检测的精度上,IAPSO比GA优化SVM后的检测精度略高,比ACA优化SVM后的检测精度要高出9.7%左右,IAPSO-SVM的误报率和漏报率都最低;在检测时间上,IAPSO-SVM模型所用的时间最短,其次是ACA-SVM模型,而GA-SVM模型需要的时间最长,这与遗传算法复杂有很大的关系.

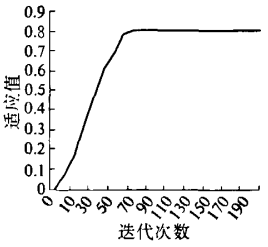
为了更细致的观察3种算法的性能,下面如图3所示绘制了3种算法在优化SVM时的适应值曲线图,实验数据和测试数据都相同.如下所示:



(a)IAPSO-SVM模型适应值曲线



(b)GA-SVM模型适应值曲线



(c)ACA-SVM模型适应值曲线

图3 不同算法在优化SVM时的适应值

Fig.3 Fit Values of SVM optimization of different alogrithms

从上面图3-5看出,IAPSO算法在经过经过40次迭代后,其适应值达到1;GA算法在经过90次迭代后,其适应值达到1;ACA算法在经过70次迭代后,其适应值达到0.8.我们可以再次印证IAPSO优化SVM参数速度更快,而且蚁群算法在优化SVM时,适应值只达到0.8而停滞,这说明蚁群算法陷入了局部最优中,不能达到最优效果.

5 结论

利用区间自适应粒子算法优化SVM后,在网络入侵检测系统中,它对入侵行为的检测精度要高于遗传算法和蚁群算法,约9.7%,检测时间也缩短40.6%~56.5%,而且蚁群算法在优化过程中陷入了局部最优,这与其每个个体只能感知局部的信息,不能直接利用全局信息,一般需要较长的搜索时间,且容易出现停滞现象有很大的关系,遗传算法由于其具有交叉和变异等复杂行为,在优化过程中所消耗的时间要远远高于其他两种算法.从试验中我们可以得出,此种区间自适应粒子群优化算法是可行且高效的.

在网络入侵系统中,操作对象都是网络中的实时数据流,且数据庞大,SVM在实时操作方面的性能明显有缺陷,仍需要进一步的研究,目前尝试着将神经网络(Neural Network,NN)与支持向量机相结合的方法运用到入侵检测系统中,充分利用NN的自主学习能力,这将会是一个很好的解决方法.

参考文献

[1] VAPNIK V N. The nature of statistical learning thory [M]. New York: Springer Verlag, 1995.

[2] 张学工.关于统计学习理论与支持向量机[J].自动化学报,2000,26(1):32-42.

[3] 袁小艳,刘爱伦.基于PSO算法的支持向量机核参数选择问题研究[J].控制理论与应用,2007,26

(5):5 - 8.

[4] 夏克文,董瑶,杜红斌. 基于改进 PSO 算法的 LS - SVM 油层识别模型[J]. 控制与决策,2007,22(12):1385 - 1389.

[5] 李涛,吕勇哉,陈鹏. 基于 SVM 和 PSO 的新型非线性模型预测控制[J]. 控制工程,2008(15):102 - 105.

[6] 刘胜,李妍妍. 自适应 GJ - SVM 参数选择算法研究[J]. 哈尔滨工程大学学报,2007,28(4):398 - 402.

[7] 刘志硕. 基于自适应蚁群算法车辆路径问题研究[J]. 控制与决策,2005,20(5):522 - 526.

[8] 曲倩倩,曲仕茹,温凯歌. 混合遗传算法求解配送车辆调度问题[J]. 计算机工程与应用,2008,44(15):205 - 207.

[9] KENNEDY J, EBERHART R. Particle swarm optimization [C]. Perth: IEEE Piscataway, 1995: 1942 - 1948.

[10] KDD cup 99 Intrusion detection data set (http://kdd.ics.uci.edu/databases/kddcup99/kddcup.data_10_percent.gz).

SVM Algorithm Based on Interval Adaptive PSO and Its Application

WANG Jie, JIANG Nian, ZHANG Yi

(School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: The parameters of the kernel functions are very improtant for the SVMs (Support Vector Machines) in the generalization ability, especially when we use the SVM for data classification with a great deal of data, it will need so much computer memory that the speed of parameter optimization will be decreased. For this problem, this paper presents a method that uses an interval adaptive particle swarm optimization to optimize the parameters of the SVMs. Then we apply this method to the intrusion detection systems, and compare it with the Ant Colony Algorithm and the Genetic Algorithm. The experimental results show that this method improves the classification accuracy by 9.7%, and the response time is shortened by 40.6% ~ 56.5%. That proves this method is workable.

Key words: support vector machine; adaptive; particle swarm optimization; intrusion detection system