文章编号:1001 - 9081(2006)09 - 2145 - 03

基于免疫算法与支持向量机的异常检测方法

周红刚, 杨春德
(重庆邮电大学 计算机科学与技术学院,重庆 400065)
(zhouhg@126.com)

摘 要:在异常检测中,应用支持向量机算法能使检测系统在小样本的条件下具有良好的泛化能力。但支持向量机的参数取值决定了其学习性能和泛化能力,且大量无关或冗余的特征会降低分类的性能。基于此,提出了一种基于免疫算法的支持向量机参数和特征选择联合优化的方法。免疫算法是一种新的有效随机全局优化技术,它具有不易陷入局部最优、解的精度高、收敛速度快等优点。仿真结果表明算法在提高异常检测的检测正确率的同时相应的测试时间也在缩短。

关键词:异常检测; 支持向量机; 泛化能力; 免疫算法; 亲和力
中图分类号: TP393.08 **文献标识码**:A

Anomaly detection approach based on immune algorithm and support vector machine

ZHOU Hong-gang, YANG Chun-de
(*College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing* 400065, *China*)

Abstract: In anomaly detection, utilizing support vector machines can make detection system have good generalization ability in situation of small sample. But appropriate parameters are very crucial to the learning results and generalization ability of support vector machines. And many irrelevant and redundant features degrade the performance of classification. Thus an approach that applied immune algorithm to optimize parameters of SVM(Support Vector Machine) and feature selection was proposed. Immune algorithm is an efficient random global optimization technique. It has nice performances such as avoiding local optimum, high precision solution, and quick convergence. The simulation results show that immune algorithm can improve the detection accuracy and meanwhile shorten the testing time.

Key words: anomaly detection; SVM(Support Vector Machine); generalization performance; immune algorithm; affinity

0 引言

异常检测是入侵检测技术一种。它是指将用户正常的习惯行为特征存储在特征数据库中,然后将用户当前行为特征与特征数据库中的特征进行比较,若两者偏差足够大,则说明发生了异常,这种方法能检测未知的攻击类型。基于统计学习理论的支持向量机(SVM)已被应用于异常检测研究[1]。但支持向量机参数和样本特征的选取好坏直接决定了其学习性能和泛化能力。实质上支持向量机参数和样本特征选择的过程是一个求解组合优化问题的过程。文献[2]提出遗传算法的支持向量机特征选择方法,能够得到相对较优的解,但算法固有缺点易陷入局部最优,且迭代次数较多,耗时较长。文献[3]提出用梯度下降方法来求 L2-SVM 最优参数,文献[4]进一步将此方法用到 L1-SVM。而梯度下降法对初始状态很敏感,若初始点离最优点很远,结果很容易陷入局部最优。因此本文提出一种基于免疫算法(Immune Algorithm, IA)[5]的 SVM 参数和特征选择联合优化的方法。免疫算法不但能够收敛于全局最优,而且收敛速度快。实验表明免疫算法能有效提高基于支持向量机的异常检测的性能。不仅在检测率上有较大的提高,而且测试新样本的时间也有缩短。实验还与标准的遗传算法作了比较,结果表明利用免疫算法优化后的异常检测性能优于利用遗传算法优化后的异常检测的性能。

1 支持向量机参数分析及泛化能力估计

支持向量机是 AT&T Bell 实验室的 Vapnik 等人根据统计学习理论提出的一种新机器学习方法。它的基本思想是根据 Vapnik 提出的结构风险最小化原理,通过最大化分类间隔或边缘尽量提高学习机的泛化能力。设计支持向量机的重要步骤是选择核函数和核参数。而 Vapnik 等人在研究中发现,不同的核函数对支持向量机的影响不大,反而核函数的参数和误差惩罚因子 C 是影响支持向量机性能的关键因素[6]。

由支持向量机原理可知[7],支持向量机的训练是通过求解如下的优化问题:

$$w^2(\alpha) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j) \quad (1)$$

$$t.s. \quad \alpha_i \geqslant 0 \quad \forall i. \text{ and } \sum_{i=1}^{n}\alpha_i y_i = 0$$

其中,$k(x_i,x_j)$ 为核函数。且当样本不可分时,核函数映射为:

$$k(x_i,x_j) \rightarrow k(x_i,x_j) + \frac{1}{c}\delta_{ij} \quad (2)$$

其中,C 为误差惩罚因子,它的作用是在确定的特征空间中调节学习机的置信范围和经验风险的比例。δ_{ij} 为 *Kronecker*δ[8],当 $i = j$ 定义为 1,其余为 0。

收稿日期:2006 - 03 - 16; **修订日期**:2006 - 06 - 11

作者简介:周红刚(1981 -),男,江苏盐城人,硕士研究生,主要研究方向:网络入侵检测、支持向量机; 杨春德,男,副教授,主要研究方向:网络信息安全、统计学习理论.

使用较为常用的径向基核函数

$$k(x_i,x_j) = \exp\left(\gamma \left\| x_i - x_j \right\|^2\right) \tag{3}$$

其中 γ 为核参数,其主要影响样本数据在高维特征空间中分布复杂程度。因此有必要对支持向量机的误差惩罚因子 C 和核函数参数 γ 进行优化,获得更强的泛化能力。

支持向量机在获得分类模型后,需要对模型的泛化能力进行估计。传统的估计方法有交叉验证和 Leave-One-Out 方法,这些方法的优点是几乎是无偏估计,估计相当精确,但其固有的缺点是计算繁琐,效率低下。因此本文使用文献[6]提出的半径—间隔界(Radius-Margin Bound)估计方法。该方法的优点是只对支持向量进行一次训练,得到模型误差的一个上限值。其表示如下:

$$Loo\ Error \leqslant f(u) = \frac{1}{l}R^2 \left\| w \right\|^2 \tag{4}$$

其中,$f(u)$ 为误差上限值。R^2 通过优化下式得到最大值:

$$\begin{aligned} &R^2 = \max_\beta \sum\nolimits_i \beta_i k(x_i,x_j) - \sum\nolimits_{i,j} \beta_i\beta_j k(x_i,x_j) \\ &s.t. \quad \sum\nolimits_i \beta_i = 1 \qquad \beta_i \geqslant 0 \quad \forall i. \end{aligned} \tag{5}$$

2 免疫算法对支持向量机参数和特征联合优化

2.1 免疫算法

免疫算法的思想来源于生物免疫机体的原理。生物的免疫系统是一个高度进化、复杂的系统,它还具有学习、记忆和自适应的调节能力。免疫算法是抽取和反映生物机体免疫系统的特点,结合工程应用而描述的一种计算模型。它模仿生物的免疫过程,具有良好的全局搜索能力和记忆功能。算法流程如图 1 所示。其中抗原对应于优化问题的目标函数,抗体对应于优化问题的解。通过抗原和抗体的亲和力来描述可行解与最优解的逼近程度。对外界抗原的侵入,系统自动产生相应的抗体,通过抗体之间的促进与抑制反应,实现系统对环境的自适应。

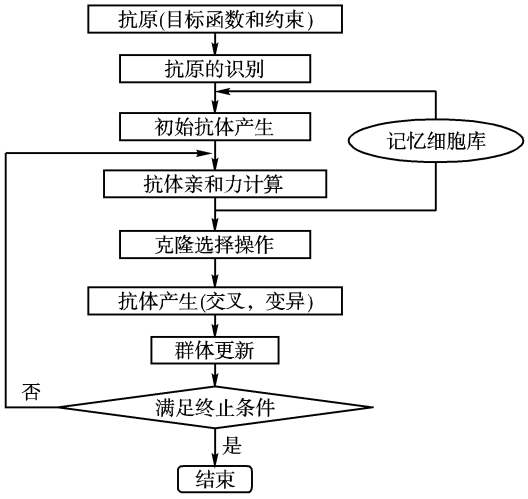

图 1 免疫算法流程

2.2 联合优化算法描述

在免疫优化的支持向量机算法实现中,优化误差上限值 $f(u)$,使训练样本集总误差最小,将其作为抗原。具体定义为:

$$\min f(u) = \frac{1}{l}R^2 \left\| w \right\|^2 \tag{6}$$

第 1 步 对抗体采用二进制编码。每一抗体二进制编码由支持向量机的核函数参数 γ、误差惩罚因子 C 和样本特征三部分组成。在参数选取方面,其二进制编码表示 $P = p_1p_2\cdots p_l$,其中编码的每一位由 1 或 0 组成,l 为参数编码长度。本文设定参数 γ 的取值范围为[0.000 1,2.000 0],精度为 0.000 1,编码长度为 15。参数 C 的取值范围为[1, 200],精度为 0.1,编码长度为 11。因此参数编码 l 总长度为 26。在特征选择方面,若给定特征集 $F = \{f_1,f_2,\cdots,f_d\}$,其中 d 是特征集大小,特征选择可以用一个二进制编码表示,$S = s_1s_2\cdots s_d$。s 中的每一位 1 或 0 分别表示 F 中相应位置的特征是否被选中。则抗体的二进制编码表示为特征和参数的组合 $A = (P,S)$。

第 2 步 产生初始抗体。针对待求问题的特征,若能与记忆细胞库的结构信息相匹配,则由记忆细胞集中的记忆细胞组成初始抗体群,不足部分的抗体随机产生。否则随机产生初始抗体群。本文设抗体规模 s 为 20。

第 3 步 抗体与抗原的亲和力。根据泛化能力估计分析,用误差上限值 $f(u)$ 作为评价标准,因此抗原与抗体的亲和力如下:

$$A_v = 1 - \left\{f(v) \times \left[1 + 0.001\left(\frac{k}{d}\right)\right]\right\} \tag{7}$$

其中,$f(v)$ 为抗体 v 的误差上限值,d 为特征数,k 为抗体 v 所选的特征数。加入 $0.001\left(\frac{k}{d}\right)$ 目的是当误差上限值相同时,取特征数较小的个体。

第 4 步 更新记忆细胞集合。当群体更新后,将与抗原有最大亲和力的抗体添加到抗体记忆细胞库中,当下次碰到类似求解问题时,则在记忆细胞库中直接搜寻问题的记忆抗体,从而提高求解的效率。本文设定记忆细胞规模为 10。

第 5 步 抗体浓度。抗体之间的亲和力描述了抗体的相似度,利用相似度的度量来控制抗体产生的刺激与抑制,从而达到抗体的多样性。

$$B_{vw} = \frac{1}{1 + E(2)} \tag{8}$$

其中 $E(2)$ 表示抗体 V 和抗体 W 的平均信息熵。则抗体 v 浓度表示为:

$$c_v = \frac{1}{N}\sum_{w=1}^{N} ac_{v,w} \tag{9}$$

$$ac_{v,w} = \begin{cases} 1, & B_{v,w} \geqslant Tacl \\ 0, & B_{v,w} < Tacl \end{cases} \tag{10}$$

其中,$Tacl$ 是抗体相似度的阈值。本文设 $Tacl$ 为 0.5。

第 6 步 基于抗体浓度克隆选择。在进化过程中,克隆并进入下一代的抗体是随机选择的,但克隆的概率与亲和力有关。具体地说,那些与抗原亲和力越高,又与其他抗体的亲和力越低的抗体有更大的克隆概率。于是抗体选择的概率 p_v 由抗体和抗原的亲和力 p_{fv} 和浓度抑制概率 p_{dv} 两部分组成,克隆概率如下:

$$\begin{aligned} p_v &= \alpha p_{fv} + (1-\alpha)p_{dv} \\ &= \alpha\left(\frac{A_v}{\sum_{i=1}^{N} A_i}\right) + (1-\alpha)\left(\frac{\exp\left(\frac{c_v}{\beta}\right)}{N}\right) \end{aligned} \tag{9}$$

其中 α 和 β 为加权系数,在[0,1]之间可调。A_v 为抗体与抗原的亲和力,c_v 为抗体浓度。

第 7 步 交叉与变异。不同的抗体之间会交叉变异,从而形成新的抗体。这一过程相当于抗体交换信息。本文采用单点交叉,高斯变异。交叉概率 $p_c = 0.6$,变异概率 $p_m = 0.05$。

第8步　更新抗体集合。为避免陷入局部最优解,要更新抗体集合,每 25 次迭代进行一次,由记忆细胞集中的记忆细胞组成抗体群。

步骤9　判断是否满足终止条件,若满足终止条件,则确定当前种群中的最佳个体作为算法最终寻找到的解,否则转到步骤3。本文设迭代次数为 100。

3　实验分析

3.1　实验数据

本文中所使用的实验数据为 DARPA 为 1999 年的 KDD 竞赛所建立的基本数据。数据集中的 38 种攻击被分为 4 类:Probing、DoS、U2R、R2L。这 41 个特征包括了以下三个方面:TCP/IP 连接的基本属性、数据中域的信息和特定时间内的流量特征。本文对每个连接只区别它是“正常”或是“异常”,也即将各种不同的入侵方式都归类为“异常”,这样可将原来的多值分类问题变为二值分类问题。在样本的选择上,本文采用了 KDDCUP99 10% 的数据集作为基准数据,随机抽取 2 338 个样本作为训练数据,8 615 个与训练样本独立的样本作为测试样本。由于数据中含有非数字型字符,因此需对文件进行处理,使得文件为数字序列,同时对数据进行归一化处理。

3.2　优化检测模型实验

整个系统的检测模型如图 2,分为训练阶段和检测阶段。在训练阶段,免疫算法提交一个抗体,并对其二进制编码解码,得到相应的特征集和参数,并对支持向量进行训练,得到抗原与抗体的亲和力。重复操作此过程,直到亲和力最大,即得到最优的检测模型。在检测阶段,支持向量机应用训练好的模型对输入产生判别。

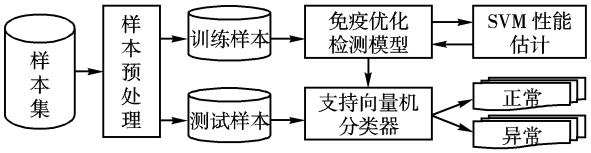

图2　基于免疫算法与支持向量机的异常检测模型

本文分别利用免疫算法(IA)、遗传算法(GA)进行优化选择。为达到实验的公平,遗传算法的参数与免疫算法参数尽量保持一致。遗传算法使用标准的 GAOT 遗传工具箱,采用二进制编码,群大小 s =20,采用轮盘赌选择,交叉率 p_c = 0.6,变异率 p_m =0.05,终此条件为迭代 100 次。

3.3　异常检测结果及分析

经免疫优化后,异常检测系统的样本特征只有 24 维,而经遗传优化后样本特征有 27 维,两种优化算法都去除了一些对检测贡献不是很大的特征。由表 1 可以看出在基于支持向量机的异常检测系统中,经优化后系统的检测时间明显少于未经优化的系统。且当未经优化的系统参数采用系统默认值,其检测正确率明显低于经优化后的系统。

表1　异常检测结果对比表

	免疫算法优化	遗传算法优化	未优化
最优 C	86	90	0.3
最优 γ	0.000 1	0.000 2	0.5
最优特征数	24	27	41
检测时间/s	25.43	27.61	48.71
检测正确率(%)	96.57	94.25	84.60

图 3 为免疫算法和遗传算法的寻优过程的对比图。整个搜索过程中,每代运行的结果,免疫算法都优于遗传算法,且免疫算法在第 41 代就已达到最优点,而遗传算法到第 92 代才达到最优。在运行 100 代后,免疫算法的最优点要高于遗传算法近 2 个百分点。这表明免疫算法相对遗传算法,不仅有更快的收敛速度,而且有更好的全局收敛性。

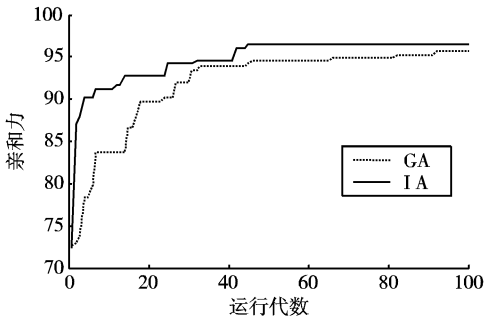

图3　IA 与 GA 寻优结果比较

4　结语

本文提出利用免疫算法对支持向量机参数和样本特征进行联合优化选择,免疫算法相对遗传算法,能够有效的解决局部搜索能力和全局搜索能力的矛盾、有效维持种群多样性、较好地防止早熟,找到比遗传算法更优的解,且寻优时间比遗传算法短。实验使用半径—间隔界估计方法作为异常检测性能的评估标准,它能够有效地减少训练的复杂度。因此,利用本文方法不仅可以构造出最优的检测模型,还可以选择出最优的样本特征。实验结果表明,经免疫算法优化后的基于支持向量机的异常检测系统不仅在检测的正确率上有较大的提高,而且在测试新样本的速率上也有提高,满足异常检测的实时性和高检测率的要求。

参考文献:

[1] 饶鲜, 董春曦, 杨绍全. 基于支持向量机的入侵检测系统[J]. 软件学报, 2003,14(4):798 - 803.

[2] FRONHLICH H, CHAPELLE O. Feature selection for support vector machines by means of genetic algorithms[A]. Proceedings of the 15th IEEE international conference on tools with artificial intelligence[C],2003. 142 - 148.

[3] KEERTHI S. Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms[J]. IEEE transactions on neural network, 2002, 13(5):1225 - 1229.

[4] CHUNG KM, KAO WC, SUN T. Radius margin bounds for support vector machines with the RBF kernel[J]. Neural Computation, 2003, 15(11):2643 - 2681.

[5] MORI K, TSKUKLYAMA M, FUKUDA T. Immune algorithm with searching diversity and it's application to resource allocation problem [J]. Transactions of the Institute of Electrical Engineers of Japan, 1993,113(10):872 - 878

[6] VAPNIK V, Statistical Learning Theory[M]. New York: John Wiley and Sons, 1998.

[7] CORTES C, VAPNIK V. Support vector networks[J]. Machine Learning, 1995, 20(1):273 - 297.

[8] CHAPELLE O, VAPNIK V, BOUSQET O. Choosing Multiple Parameters for Support Vector Machines [J]. Machine Learning, 2002, 46(1):131 - 159.