第33卷 第14期 计 算 机 工 程 2007年7月
Vol.33 No.14 Computer Engineering July 2007
·安全技术· 文章编号：1000—3428(2007)14—0123—02 文献标识码：A 中图分类号：TP393

基于 PSO 算法的神经网络集成入侵检测系统

李朝荣，张 鹰，张安妮
(西华师范大学计算机学院，南充 637002)

摘 要：在系统结构上提出了一种多检测器并行的智能机群入侵检测系统模型，系统中每一个检测器是一个神经网络集成分类检测器，由多个 PC 组成，以提高系统响应速度。采用两次粒子群优化算法选择性集成神经网络集，提高了神经网络集成检测器的预测精度。程序设计采用 PVM 并行方式实现。

关键词：入侵检测系统；神经网络集成；选择性集成；粒子群优化；并行计算

IDS of Neural Network Ensemble Based on PSO Algorithm

LI Chaorong, ZHANG Ying, ZHANG Anni
(College of Computer, China West Normal University, Nanchong 637002)

【Abstract】 A multi-detector IDS model based on computer cluster is put forward. In this system every detector comprises several PCs that can improve the response speed of the system's. In order to improve the prediction and generalization ability of the detector, twice PSO algorithm is proposed to construct neural network ensemble, which is designed in PVM parallel environment.

【Key words】 intrusion detection system(IDS); neural network ensemble; selective ensemble; PSO; parallel computation

为了提高入侵检测系统(IDS)的检测精度与速度，基于智能算法或多智能算法融合的异常检测成为近期的研究热点。目前已有人工神经网络(ANN)、支撑向量机(SVM)[1]、人工遗传(GA)/免疫(AIS)、多智能体(multi-agent)、隐Markov链模型(HMM)等智能检测技术。

异常检测本质上是一个分类问题，神经网络集成[2]能提高单一的神经网络的分类精度。在神经网络集成个体选择上一般用boosting和bagging方法，但这类方法计算量较大，影响IDS的响应速度。近年来出现了GA与PSO算法应用于选择个体网络。

本文采用粒子群优化(particle swarm optimization, PSO)算法[3]选择性集成神经网络[4]，与其它应用PSO算法不同之处在于，在独立地训练一批网络个体后，通过离散PSO(DPSO)算法[5]优选部分差异度较大的神经网络个体，再用连续PSO算法(CPSO)构建神经网络集成。相对于遗传算法，该算法简单易实现、计算量小，加之在程序实现上用并行PSO算法能进一步提高运算速度，满足实时检测要求。

1 多个神经网络集成入侵检测系统设计方案

系统方案如图 1 所示。

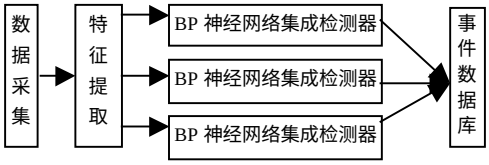

图 1 多个神经网络集成检测器入侵检测系统结构

从网络上采集的数据经主成分分析(PCA)特征提取并格式化为矢量样本数据后，并行分为 3 部分，分别送入 3 个完全一样的神经网络集成检测器并行处理，最后把检测结果提交到事件数据库。神经网络集成采用应用最广泛的BP神经网络组成，每一个神经网络集成检测器是由多台PC构成(也可以是多处理机)，以提高整个系统的运行速度。每台PC上装有PVM(parallel virtual machine)并行编程系统[6]，以便并行处理大量的数据。

2 神经网络集成原理

神经网络集成是用有限个神经网络对同一个问题进行学习，该集成在某个输入下的输出由构成集成的各个体神经网络在同样输入时所得的输出共同决定[2]。组合多个分类器输出结果的集成学习是改善分类精度的重要方法。单一神经网络方法不但分类精度难以达到要求，而且极易陷入局部极小点，神经网络集成能克服上述缺点。

设神经网络集成学习的未知函数是 $f:R^m \to R$，神经网络集成由神经网络 $f_1,f_2,\ldots,f_N$ 集成，各网络分别给定权值 $\omega_i(i=1,2,\ldots,N)$，满足 $\omega_i \geq 0$ 且 $\sum_{i=1}^{N}\omega_i = 1$，输入 $x \in \mathrm{Rm}$ 满足分布 $p(x)$，在 x 输入下目标输出为 $d(x)$，神经网络集成的输出为

$$f(x)=\sum_{i=1}^{N}\omega_i f_i(x) \tag{1}$$

神经网络集成泛化误差为

$$E=\int p(x)(f(x)-d(x))^2\,\mathrm{d}x \tag{2}$$

3 PSO 算法原理

PSO 算法是 Eberhart 与 Kennedy 根据鸟群和鱼群运动行为发明的仿生学优化算法。在 PSO 中每个优化问题的解都是搜索空间中的一只鸟，称之为粒子。所有的粒子都有一个被优化函数决定的适应值(fitness value)与一个决定它们飞行方

基金项目：四川省教育厅重点科学科研基金资助项目(2005A109)；西华师范大学校立科研基金资助项目(05A009)
作者简介：李朝荣(1976－)，男，硕士研究生，主研方向：计算机网络及安全，计算智能；张 鹰，副教授；张安妮，硕士研究生
收稿日期：2006-09-25 **E-mail：**licr_yibin@yahoo.com.cn

向和距离的速度。然后粒子追随当前最优粒子在解空间搜索。

设在一个N维的目标搜索空间中，有M个粒子组成的一个种群，第i个粒子在N维空间里的位置表示为矢量$X_i=(x_{i1},x_{i2},...,x_{iD})$，即第$i$个粒的在$N$维空间中的位置是$X_i$，飞行的速度表示为矢量$V_i=(v_{i1},v_{i2},...,v_{iD})$。至此，每个粒子找到的最好位置表示为$P_i=(p_{i1},p_{i2},...,p_{iD})$，整个粒子群找到的最好位置表示为$P_g=(p_{g1},p_{g2},...,p_{gD})$，又是在$p_i$中的最优值。每次迭代，每个粒子的当前速度$V_i$与当前位置$X_i$更新如下：

$$v_{id}=\omega v_{id}+c_1*rand()(P_{id}-x_{id})+c_2*rand()*(p_{gd}-x_{id}) \quad (3)$$

$$x_{id}=x_{id}+v_{id} \quad (4)$$

其中，$w\geq 0$称为惯性权重；c_1、c_2称为学习因子；$rand()\in[0,1]$是随机函数。

以上 PSO 适用于连续空间，即 CPSO，可以变形为离散二进制形式的 PSO 算法。离散 PSO(DPSO)粒子的位置只能取 0 或 1，取代式(3)与式(4)的更新式为

$$v_{id}=v_{id}+rand()*(p_{id}-x_{id})+rand()*(p_{gd}-x_{id}) \quad (5)$$

$$x_{id}=\begin{cases}1 & \rho_i<sig(v_i)\\ 0 & \rho_i>sig(v_i)\end{cases} \quad (6)$$

其中，p_i是取值为 0 或 1 的随机向量；$sig()$是sigmoid函数。

从中可以看出，由于 PSO 算法简单，没有太多参数，因此已应用于函数优化、神经网络训练、模糊控制等领域。

4 神经网络集成检测器

4.1 DPSO 算法选择个体子集方案

设已训练出的N个神经网络的集合为$F=\{f_1,f_2,...,f_N\}$，从集合$\{f_1,f_2,...,f_N\}$中取有$M(1<M\leq N)$个个体的子集$Sbest$，要求$Sbest$集合中所有差异度最大，即体集成后具有最好的泛化能力。定义粒子编码为$S_i(1\leq i\leq M)$，S_i为F的一个子集，且

$$S_i=\{S_{i1},S_{i2},...,S_{iM}\},\ S_{id}=\begin{cases}1 & f_k\in S_i\\ 0 & f_k\notin S_i\end{cases} \quad (7)$$

式(7)表示粒子群中的每一个粒子对应于$\{f_1,f_2,...,f_N\}$的一个子集为S_i，而且粒子向量的每一分量只能取 0 或取 1，1 表示个体将被选入二次集成，0 表示淘汰。如果用神经网络集成在校验集上的泛化误差E^V为优化目标函数，见式(8)，这样第 1 次选择集成就转换为在N维$\{0,1\}$空间中用选择最优粒子的最优化问题。此问题恰好可以用DPSO算法解决。

$$J=E^V=\sum_{f_i,f_j\in S}C^V\times M^{-2} \quad (8)$$

其中，$C^V=\sum_{x\in V}(f_i(x)-d(x))(f_j(x)-d(x))\times|V|^{-1}$；$M$ 是集合 S 的长度；V 为校验集；$|V|$是集合 V 的长度。下面给出算法描述：

```
double *DPSO(F[N])
{   Initialize all the variables;
 while(Jg>ε&& t < MaxIter)
{ for(int i=0;i<M;i++)
      { compute J[i] according function (8) and samples;
          if(J[i]<Jbest[i])
{ Jbest[i]=J[i]; for(int j=0;j<D;j++)pbest[i][j]=x[i][j];}
if(Jbest[i]<Jgbest)
{ Jgbest=Jbest[i]; for(int j=0;j<D;j++)pgbest[j]=x[i][j];} }
     for(int i=0;i<M;i++)
for (int j=0;j<D;j++)    Update v[i][j] and x[i][j]   according (5) and (6);
      t++;}
Return pgbest[j];   }
```

4.2 CPSO 算法构建神经网络集成及程序的并行实现

在 4.1 节中已经取得了子集S为$\{s_1,s_2,...,s_m\}$，本节子集集成主要任务是确定各集成个体的权重向量ω。把第 2 节中的式(1)代入式(2)，并设分布$p(x)$为均匀分布，可得

$$E=\frac{1}{M}\int(\sum_{i=1}^{N}\omega_i f_i(x)-d(x))^2dx \quad (9)$$

为使神经网络集成的性能最优，则应使其泛化误差最小：

$$\begin{cases}\min E\\ \sum_{i=1}^{N}\omega_i=1\end{cases} \quad (10)$$

式(10)是一个带约束的最优化问题，为便于计算将其转化为不带约束的优化问题，目标函数右面加上一惩罚 PE，$PE=\sigma(\sum_{i=1}^{N}\omega_i-1)^2$，最终得

$$E'=\frac{1}{M}\int(\sum_{i=1}^{N}\omega_i f_i(x)-d(x))^2dx+PE \quad (11)$$

这样问题转化为求关于ω的函数E'的最小值。把E'作为适应度函数，把权重向量ω作为粒子的位置，用连续 PSO 可以求解此问题(实际计算时以上各式的积分号都采用累积和运算)。以下给出并行 PSO 算法的 PVM master 程序描述：

```
 #include   "pvm3.h"
ParellelPSOEnsemble()
 {initialize all the variables
pvm_spawn(f[1],(char**)0,0,"",1,&childtid1[1]);
       …
pvm_spawn(f[N],(char**)0,0,"",1,&childtid1[1]);
//分配任务到各点计算机
pvm_barrier();        //同步所有进程
pvm_recv(chiletid[1],1); pvm_upkdouble(F[1],1,1);
    …
pvm_recv(chiletid[N],N);pvm_upkdouble(F[N],1,1);
//从各子进程接收适应值 f[N].
S[M]=DPSO(F[N]);        //用 DPSO 从 F[N]选取最优子集 S[M]
ω=CPSO(S[M]);           //用 CPSO 把集合 S[M]里的个体集成
Get output of ensemble according equation (1);
pvm_exit();}
```

5 实验结果

实验环境：CPU 奔腾 4 2.8GHz，内存 256MB，操作系统 RedHat Linux 9.0。每 5 台 PC 组成一个集成检测器，每个 PC 上用 2 个进程运行 2 个 BP 神经网络(有条件也可以一台一个 BP)，这样一个集成有 BP10 个网络。BP 网络采用 2 个隐层 41-60-15-5 结构，输入层 41 对应于 41 个特征，输出层 5 对应于 4 个攻击类型与 1 个正常(normal)类型。神经网络集成的输入向量归一化到[0,1]之间，集成个体网络(单个 BP)连接权值矩阵取[-1,1]之间的随机数。PSO 算法的粒子群个数设为 20，训练神经网络用 KDD Cup 1999 Data 10%训练集训练。

实验分为 2 次：第 1 次是用 KDD 1999 Data 10%测试集进行测试，结果为“normal”类识别率为 99.25%，其它 4 种攻击 DOS、R2L、U2L、Probing 识别率分别为 94.30%，90.08%，96.02%与 98.73%。第 2 次选取典型的 4 类攻击工具分别用 land、teardrop、(DOS)mayday-linux.c、guess_passwd(R2L)、c-marbles.c、rootkit(U2L)、netcat、ipsweep(Probing)各攻击 10 次，正确检测次数分别为 9、10、9、10。而且系统学习稳定以后可以立即检测出正在被攻击。

6 总结

本模型是基于多个检测器的机群并行系统，用神经网络集成作为系统的异常检测器，提高了系统的预测能力。算法上采用具有内在并行性质的 PSO 算法，试验结果证明了系统的响应速度快。随着硬件性价比的提高，构建更大的并行集成智能检测系统也是确实可行的。

(下转第 127 页)