基于混合 CatfishPSO-LSSVM 特征选择的入侵检测①

王卫平，唐志煦

(中国科学技术大学 管理学院，合肥 230026)

摘 要：入侵检测系统面临的主要问题是计算量大，特征选择被引入解决这一问题。针对现有方法的缺点，利用改进的粒子群算法来搜索最优特征子集，提出了一种基于混合 CatfishPSO 和最小二乘支持向量机的特征选择方法，利用混合的 CatfishBPSO 和 CatfishPSO 选择特征子集并同步对 LSSVM 的参数进行优化，最后建立了一个基于该特征选择方法的入侵检测模型。在 KDD Cup 99 数据集上进行的实验结果表明该模型的检测性能较高。

关键词：特征选择；粒子群算法；最小二乘支持向量机；入侵检测

Intrusion Detection Based on Hybrid CatfishPSO-LSSVM Feature Selection

WANG Wei-Ping, TANG Zhi-Xu

(School of Management, University of Science and Technology of China, Hefei 230026, China)

Abstract: The main issue of Intrusion detection systems is large computation, feature selection was introduced to solve the problem. According to the shortcomings of existing methods, this paper uses improved Particle Swarm Optimization to search optimal feature subset, proposes a feature selection method based on hybrid CatfishPSO and Least Square Support Vector Machine, uses combined CatfishBPSO and CatfishPSO to select feature subset and optimize the parameters of LSSVM simultaneously, and build a Intrusion detection model based on the feature selection method above. Experiments on KDD Cup 99 show that the model has a good detection performance.

Key words: feature selection; particle swarm optimization; least square support vector machine; intrusion detection

1 引言

随着网络技术的高速发展，网络数据流的规模越来越大，入侵检测系统遇到了很多挑战，这其中一个主要的问题是计算量大、检测速度低，使得入侵检测系统很难在实时处理海量数据。很多研究者运用特征选择来解决这一问题，通过选择重要特征，剔除冗余特征来加快检测速度。

在入侵检测的特征选择中，神经网络[1]、遗传算法[2]等分别用来搜寻最优特征子集，但是这些方法计算消耗大，收敛速度慢。粒子群算法收敛速度快，所需确定参数少，但易陷入局部最优。而在子集评估方面，支持向量机学习性能出色，已成为研究的热点。最小二乘支持向量机作为支持向量机的变体，求解简单，但是其性能仍然受参数影响。

针对上述问题，本文首先对粒子群算法进行了改进，使其能够避免陷入局部最优，然后在此基础上提出了基于混合 CatfishPSO 和最小二乘支持向量机的特征选择方法，在进行特征选择时同步优化最小二乘支持向量机的参数，并用该方法建立了一个入侵检测模型，实验表明该模型检测性能较高。

2 基于混合CatfishPSO-LSSVM特征选择方法的入侵检测模型

2.1 粒子群算法

2.1.1 基本粒子群算法

粒子群算法是由 Dr. Eberhart 和 Dr. Kennedy 提出的一种模拟鸟类社会行为的进化计算技术[3]。在 PSO 中，每一个候选解都可以看作是搜索空间中的一

① 收稿时间:2011-05-19;收到修改稿时间:2011-06-29

个粒子，每个粒子利用其自身的记忆和通过群体整体获得的知识来找出最优解。算法描述如下：

在一个 d 维的搜索空间中，每个粒子可以表示为 $x_i=(x_{i1},x_{i2},...,x_{id})$，第 i 个粒子的速度可以表示为 $v_i=(v_{i1},v_{i2},...,v_{id})$。第 i 个粒子的前一最优位置用 $p_i=(p_{i1},p_{i2},...,p_{id})$ 表示，即 $pBest_i$。粒子群 $pBest_i$ 为 $g=(g_1,g_2,...,g_d)$，即 $gBest$。在每次迭代过程中，粒子的速度和位置根据下面的等式来进行更新：

$$v_i^{n+1}=w\times v_i^n+c_1\times r_1\times\left(pBest_i-x_i^n\right)+c_2\times r_2\times\left(gBest-x_i^n\right) \quad (1)$$

$$x_i^{n+1}=x_i^n+v_i^{n+1} \quad (2)$$

在以上这些等式中，w 是惯性权重，c_1 和 c_2 是学习因子，分别用来表示粒子自身最优和全局最优对粒子速度影响的权重，r_1 和 r_2 是[0，1]内的随机数。粒子的运动速度不能超出预定范围 $[-v_{\max},v_{\max}]$，如果超过，则限定为 $v_{\max}$ 或 $-v_{\max}$。式(1)表示粒子的移动速度的决定因素一共有三个，分别是粒子当前的速度，粒子当前位置和自身最优位置的的偏移量以及粒子当前位置和全局最优位置的偏移量。

2.1.2 离散二进制粒子群算法

上述基本 PSO 是用于实数空间的优化算法，很难用在变量离散的空间上，而特征选择时特征组合的单个分量的状态只能有两种，被选中和不被选中，因此需要应用离散的二进制粒子群算法（Binary Particle Swarm Optimization，BPSO）[4]。BPSO 每个粒子都由二进制变量构成，粒子的速度更新公式仍然是式(1)没有变化，但粒子的速度将转变为概率的变化，即二进制变量取值 1 的几率的变化。故而在 BPSO 中，速度的值必须限定在范围[0.0，1.0]内，为了将原来速度的实数值映射到给定的[0.0，1.0]范围内，引入了 sigmod 函数来处理这个问题。粒子的位置更新公式变为如式(3)所示。这里，$S(v_{id}^{n+1})$ 表示 x_{id}^{n+1} 的概率，$rand()$ 是一个从均匀分布[0，1]中取出来的随机数。为了避免 $S\left(v_{id}^{n+1}\right)$ 变成 0 或者 1，需要预先设定一个常数 $v_{\max}$ 来限制 v_{id}^{n+1} 的范围，即 $v_{id}^{n+1}\in[-v_{\max},v_{\max}]$。

$$\text{If}(rand()<S\left(v_{id}^{n+1}\right))\ \text{then}\ x_{id}^{n+1}=1\ \text{else}\ x_{id}^{n+1}=0 \quad (3)$$

$$S\left(v_{id}^{n+1}\right)=\frac{1}{1+e^{-v_{id}^{n+1}}}$$

2.1.3 CatfishPSO

粒子群算法所存在的一个主要问题就是早熟收敛，这将会导致性能损失和只能得到次优解。为了解决这一问题，当发现粒子陷入局部最优时，将粒子群中适应度最差的 10%原始粒子用 Catfish 粒子代替，Catfish 粒子通过模拟鲶鱼效应，刺激“沙丁鱼”粒子进行更新的搜索，即初始化一个在整个搜索空间上的从极点开始的新搜索，从而能够引导陷入局部最优的粒子朝向搜索空间中有希望的新区域。这种方法简单易行并且不会增加搜索过程的计算复杂度[5]。

2.2 最小二乘支持向量机

支持向量机的原理是，把训练数据集非线性地映射到一个高维的特征空间，随后在特征空间建立一个超平面来分类数据[6]。最小二乘支持向量机（Least Square Support Vector Machine，LSSVM）是标准支持向量机的一种规则变体，在优化阶段只需要解决一个线性方程式而不是像 SVM 一样需要求解二次规划问题，不仅简化了标准支持向量机的求解流程，加快了求解速度，而且能够有效避免支持向量机的局部极值问题[7]。

给定 N 个训练对 $\{(x_i,y_i)\,|\,x_i\in R^n,y_i\in\{-1,+1\}\}_{i=1}^N$，其中 x_i 是输入向量，y_i 是二值类标签。最小二乘支持向量机模型如下：

$$y(x)=\omega^T\varphi(x_i)+b \quad (4)$$

ω 和 b 是模型的参数，φ 是将特征空间映射到高维空间的映射。计算超平面的参数 ω,b,e，LSSVM 分类问题可以转化为求解式(5)的优化问题。

$$\begin{aligned}&\min_{\omega,b,e} j(\omega,b,e)=\frac{1}{2}\omega^T\omega+\gamma\frac{1}{2}\sum_{i=1}^N e_i^2\\&s.t.\ y_i[\omega^T\varphi(x_i)+b]=1-e_i,i=1,\ldots,N\end{aligned} \quad (5)$$

为解决上述优化问题，定义如式(6)的拉格朗日函数

$$\begin{aligned}&L(\omega,b,e;\alpha)=j(\omega,b,e)\\&-\sum_{i=1}^N\alpha_i\{y_i[\omega^T\varphi(x_i)+b]-1+e_i\}\end{aligned} \quad (6)$$

其中，α_i 为拉格朗日乘子。分别对 ω,b,e_i,α_i 求偏导，得最优化条件：

$$\begin{aligned}&\frac{\partial L}{\partial\omega}=0\rightarrow\omega=\sum_{i=1}^N\alpha_i y_i\varphi(x_i),\\&\frac{\partial L}{\partial b}=0\rightarrow\sum_{i=1}^N\alpha_i y_i=0,\\&\frac{\partial L}{\partial e_i}=0\rightarrow\alpha_i=\Upsilon e_i\\&\frac{\partial L}{\partial\alpha_i}=0\rightarrow y_i[w^T\varphi(x_i)+b]-1+e_i=0\end{aligned} \quad (7)$$

可以改写成求解如下线性问题的解：

$$\left[\begin{array}{ccc|c} I & 0 & 0 & -Z^T \\ 0 & 0 & 0 & -Y^T \\ 0 & 0 & \gamma \mathrm{I} & -I \\ \hline Z & Y & I & 0 \end{array}\right]\left[\begin{array}{c} \omega \\ b \\ e \\ \hline \alpha \end{array}\right]=\left[\begin{array}{c} 0 \\ 0 \\ 0 \\ \hline \vec{1} \end{array}\right] \tag{8}$$

$$Z=\left[\varphi(x_1)^T y_1;...;\varphi(x_N)^T y_N\right],$$

这里 $Y=\left[y_1;...;y_N\right], \vec{1}=[1;...;1],$

$$e=\left[e_1;...;e_N\right], \alpha=\left[\alpha_1;...;\alpha_N\right]$$

它的解由下式给出

$$\left[\begin{array}{c|c} 0 & -Y^T \\ \hline Y & ZZ^T+\gamma^{-1}I \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right]=\left[\begin{array}{c} 0 \\ \hline \vec{1} \end{array}\right] \tag{9}$$

将 Mercer 条件应用在矩阵上，这里

$$\begin{aligned} \Omega_{il} &= y_i y_l \varphi(x_i)^T \varphi(x_l) \\ &= \psi(x_i, x_l) \end{aligned} \tag{10}$$

因此，通过求解式(9)到式(10)的线性问题的集合就可以得到分类器(11)。

$$y=sign[\sum \alpha_i y_i \psi(x, x_i)+b] \tag{11}$$

2.3 基于混合 CatfishPSO-LSSVM 特征选择方法的入侵检测模型

2.3.1 模型概述

模型可以分为两个阶段，如图 1 所示，第一阶段是训练过程，利用混合 CatfishPSO 进行特征选择和优化 LSSVM 的参数，然后 LSSVM 利用得到的特征子集和参数对数据进行分类，用分类精确度作为粒子的适应度，选出最优适应度的粒子，从而得到对应的最优特征子集和参数。第二阶段是检测阶段，测试集预处理后将上阶段得到的参数和最优特征子集所对应的属性值作为 LSSVM 的输入，得到分类结果即能得到检测结果。

2.3.2 粒子编码和适应度函数

为了实施我们的方法，这里 LSSVM 采用 RBF 核函数 $\psi(x, x_i)=\exp(-\frac{\|x-x_i\|^2}{2\sigma^2})$，因为 RBF 核函数只需要定义两个参数 γ 和 σ^2。我们需要同步优化这两个连续型参数和选择特征，因此用 CatfishPSO 来优化上述连续参数，而用 CatfishBPSO 来进行优化特征选择过程，二者组合成混合的 CatfishPSO。因此粒子由 3 部分构成，即离散值的特征掩码，可以用一个长度为特征总数的二进制形式的串来表示粒子，每一个位上的值表示该特征的状态，即某位上的值为{1}时表示该特征是被选中的，而如果该值为{0}时则表示该特征没有被选中。值连续的两个参数，粒子维数为 n_F+2。

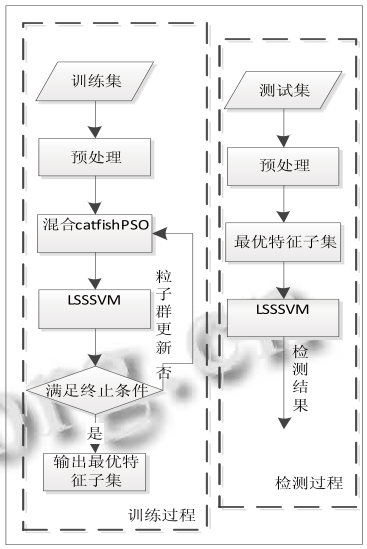

图 1 基于混合 CatfishPSO-LSSVM 特征选择方法的入侵检测模型

这里我们采用如式的分类精确度作为适应度函数。分类精确度表示被正确分类样本的百分比，如式(12)，其中 cc 和 uc 分别表示被正确分类的样本数和被错误分类的样本数。

$$fitness=\frac{cc}{cc+uc}\times 100\% \tag{12}$$

2.3.3 特征选择过程

特征选择的过程描述如下：

Step1 随机初始化粒子群，设置粒子数量 P，粒子的速度范围，最大迭代次数 N。计算每个粒子的适应度函数，初始化 $pBest_i$ 和 $gBest$，初始化迭代次数 n=0。

Step2 for i=1 to P

If ($fitness(x_i) > fitness(pBest_i)$) then $pBest_i = x_i$

If （ $fitness(x_i) > fitness(gBest)$ ）then $gBest = x_i$

If （粒子群陷入局部最优）then

将粒子群按照适应度从高到低进行排序

For i =0.9P to P

If rand number >0.5 then

For d = to 粒子维数

Catfish 粒子的位置 中离散部分值为 1，连续部分值为其极大值

Else

For d =1 to 粒子维数

Catfish 粒子的位置 中离散部分值为 0，连续部分值为其极小值

Step 3 For i=1 to P

For d=1 to D

粒子根据式(1)更新速度，粒子中离散部分根据式(3)更新位置，连续部分根据式(2)更新位置

n=n+1

Step4 For i=1 to P

If(未达到最大迭代次数或停止条件)

返回 Step 2

Else

停止迭代，输出 *gBest* 和 *gBest* 的适应度

3 实验及结果分析

3.1 数据集及其预处理

实验采用KDD Cup 99数据来进行实验，KDD Cup 99 是一个有名的入侵评估数据集，数据集包含大约 500 万条连接记录[8]。数据集中的每条记录包括 41 个连续的数值类型或者名义类型的特征，外加一个类标签。各特征编号与类型如表所示。KDD Cup 99 包含 24 种攻击类型，可以分为 4 大类：Probe，Denial of Service（DOS），User to Root（U2R）以及 Remote to User（R2L）。数据预处理时，首先将这里的只有名字值的特征通过简单地将类属性替换成成数值的方法转换成数值特征。如名义特征 Protocol 包括 tcp、udp、icmp，就将值 tcp 换成 1，udp 换成 2，icmp 换成 3，其他的也是相同的方法处理。然后用每个属性的值分别除以该属性的最大值，从而将其转化为[0，1]之间的值。

在数据源的选择上，由于 KDD Cup 99 数据量太大，这里我们随机从“10% data of KDD Cup 99 trainning set”中抽取了 5 类各 6472 条记录作为训练集，6803 条记录作为测试集，训练集样本的分布如表 1 所示：

3.2 实验方案和参数选择

由于 LSSVM 是一种二分类方法，而数据中包含 5 种类型，即一种正常行为，和 4 种攻击行为，而且各种类型的重要特征并不一致，因此这里采用 5 个 LSSVM 分类器来进行测试。正常类分类器将数据分为正常和非正常（包括所有类型的攻击），Probe 分类器将数据分为 Probe 攻击和非 Probe 数据（包括正常和其他三种类型的攻击），此外还有另外三种攻击类型的分类器。

表 1　训练集集数据样本

类型	Normal	DOS	Probe	R2L	U2R
Normal	2000	3790	300	350	32
Probe	1300	3390	1450	300	32
DOS	1410	4340	340	350	32
U2R	600	3330	300	242	2000
R2L	1000	4240	200	1000	32

模型中各参数分别设置为 $w=1$，$c_1=c_2=2$，N=100，P=30，特征掩码部分对应的 $v_{\max}=6$，γ 对应的 $v_{\max}=200$，σ^2 对应的 $v_{\max}=50$。

3.3 结果分析

对于 5 种不同的分类器，选择的特征分别为：

Normal：1、3、5、6、23、35

DOS： 3、5、23、38

Probe：3、5、27、40、41

R2L： 3、5、6、22、33、37

U2R： 3、5、14、23、24、32、33

表 2　检测速度比较表

分类	方法	精确度	建模时间（s）	测试时间（s）
Normal	选择特征	99.80%	25	11
	全部特征	99.75%	53	20
Probe	选择特征	99.83%	19	8
	全部特征	99.85%	53	20
DOS	选择特征	99.00%	35	13
	全部特征	98.93%	54	20
U2R	选择特征	99.84%	5	4
	全部特征	99.01%	54	21
R2L	选择特征	99.91%	23	10
	全部特征	99.46%	53	20

检测结果及比较如下列表中所示，从表 2 中可以看出，相对于利用全部 41 个特征进行检测，混合 CatfishPSO-LSSVM 方法无论是建模时间还是测试时间都有了大幅度的降低，说明模型的检测速度较快。而从表 3 中可以看出，混合 CatfishPSO-LSSVM 方法

对每个分类都能选择出很好的特征，在 Normal、Probe 和 R2L 三类上，检测精确度全面优于文献[9]方法，而在另外两类，差距也很小。

表 3 检测精确度与其他方法的比较

类型 / 方法	Normal	Probe	DOS	U2R	R2L
CatfishPSO-LSSVM	99.80%	99.83%	99.00%	99.84%	99.91%
SVM[9]	99.55%	99.70%	99.25%	99.87%	99.78%

4 结语

本文首先对粒子群算法进行了改进，使其能够避免陷入局部最优，利用其收敛速度快的特点，用二进制粒子群算法作为特征选择的子集搜索算法，用粒子群算法对 LSSVM 的参数进行同步优化，形成了基于混合 CatfishPSO 和最小二乘支持向量机的特征选择方法，，并建立了一个基于该方法的入侵检测模型，实验表明该模型的检测速度和精度都比较高。

参考文献

1 Sung AH, Mukkamala S. Identifying important features for intrusion detection using support vector machines and neural networks. Proc. of the 2003 International Symposium on Applications and the Internet Technology. IEEE Computer Society Press, 2003: 209−216.

2 Stein G, Chen B, Wu AS, Hua KA. Decision tree classifier for network intrusion detection with GA-based feature selection, Proc. of the 43rd ACM Southeast Regional Conference. Kennesaw, Georgia: 2005,2:136−141.

3 Kennedy J, Eberhart RC. Particle swarm optimization. Proc. of the IEEE International Conference on Neutral Networks. .Perth: 1995: 1942−1948.

4 Kennedy J, Eberhart RC. A discrete binary version of the particle swarm algorithm. Proc. of the IEEE International Conference on Systems, Man, and Cybernetics. Washington: 1997: 4104−4109.

5 Chuang LY, Tsai SW, Yang CH. Catfish Particle Swarm Optimization. 2008 IEEE Swarm Intelligence Symposium. St. Louis: 2008: 1−5.

6 Vapnik V. Statistical Learning Theory. New York:Wiley, 1998.

7 Suykens JAK, Vandewalle J. Least Squares Support Vector Machine Classifiers.Neural Processing Letters.1999,9(3):293−300.

8 KDD Cup 99 Datasets http://kdd.ics.uci.edu/databases/kddc−up99/kddcup99.html:1999.

9 Mukkamala S, Sung A, Abraham A. Intrusion detection using an ensemble of Intelligent paradigms. Journal of Network and Computer Applications, 2005,28(2):167−182.

(上接第 171 页)

报,2005,16(9):1568−1576.

4 Yoshinari K. A human motion estimation method using 32successive video frames. Proc. of International Conference on Virtual Systems and Multimedia. Gifu: IEEE Publisher, 1996: 135−140.

5 Haritaoglu I. Real-time surveillance of people and their activities. IEEE Trans. on PAMI, 2000,22(8):809.

6 Linda G.计算机视觉.赵清杰等译.北京:机械工业出版社, 2005.220−221.

7 李征.基于核心密度估计的动态目标分割改进模型.四川大学学报(自然科学版),2006,43(5):1007−1013.

8 Cucchiara R, Grana C, Piccardi M, et al. Improving shadow suppression in omoving object detection with HSV color information. IEEE Transportation Systems Conference Proc. Oakland, USA: IEEE Publisher, 2001.