

粒子群特征优选的 SVDD 入侵检测研究

魏振伟^{1,2}, 刘 飞³

(1 郑州航空工业管理学院 航空工程系, 河南 郑州 450015; 2 航空经济发展河南省协同创新中心, 河南 郑州 450046; 3 空军工程大学 防空反导学院, 陕西 西安 710051)

摘 要: 针对入侵检测中样本集维数较高问题, 提出一种基于粒子群算法 (PSO) 优化的支持向量数据描述 (SVDD) 方法, 将其应用于网络异常入侵检测. 该方法采用粒子群算法消除支持向量数据描述中的冗余参数并对数据降维, 并建立 SVDD 超球体模型, 对网络入侵数据进行检测并输出入侵检测结果. 在 KDD CUP'99 的标准检测数据集上进行仿真实验, 结果表明该方法和传统的 SVDD 相比不仅能够有效提高检测率, 而且计算量较小.

关键词: 入侵检测; 支持向量数据描述; 粒子群算法

中图分类号: TP31

文献标识码: A

文章编号: 1000-7180(2016)08-0144-05

Research of Network Intrusion Detection Based on Particle Swarm Optimization and Support Vector Data Description

WEI Zhen-wei^{1,2}, LIU Fei³

(1 Department of Computer Science and Application, Zhengzhou University of Aeronautics, Zhengzhou 450015, China;

2 Collaborative Innovation Center for Aviation Economy Development, Zhengzhou 450046, China;

3 Air-Defense and Anti-Missile Institute, Air Force Engineering University, Xi'an 710051, China)

Abstract: Concerning the data set of high dimensions in intrusion detection, the new algorithm based on support vector data description (SVDD) which optimized by particle swarm optimization (PSO) was proposed. In the improved algorithm, at firstly the particle swarm optimization was used to remove redundant features and reduce the data dimension in support vector data description. And then, the support vector data description built a super sphere model to detect the attacks from internet by analyzing the network connection data. Results from the experiments with the KDD CUP'99 network data indicate that the method of PSO-SVDD is better than traditional one class classifiers algorithm, which can improve the efficiency of intrusion detection and reduce the false detection rate.

Key words: intrusion detection; support vector data description; Particle swarm optimization

1 引言

入侵检测的核心是通过建立正常或异常的行为模式来检测入侵, 从模式识别的角度来看, 其本质是一个二值分类问题^[1]. 支持向量机方法建立在机器学习策略之上, 由于采用了统计理论的经验风险最小化准则, 是一种推广能力较好的大规模数据自学习方法. 文献[1-2]建立了基于 OCSVM 的入侵检测系统, 这些研究充分利用了正样本数据与负样本数

据的差异, 在实际构建入侵检测系统中发挥了一定的作用, 但仍然无法快速地构造分类模型.

Tax 和 Duin^[3]提出了支持向量数据描述方法 (Support Vector Data Description, SVDD). SVDD 方法尝试在数据样本超空间中建立一个所含正常数据尽可能多, 而异常数据尽可能少的最小包围球, 该方法可以应用于入侵异常检测. 为了优化入侵检测模型, 可以通过某种策略, 选择影响分类结果的关键特征, 由此来去除数据冗余属性、降低数据维度及问

题规模,同时提高入侵检测速度和效率。

本文提出一种基于粒子群(Particle Swarm Optimization, PSO)特征优选的 SVDD 入侵检测方法(PSO-SVDD). 该方法采用粒子群算法进行特征选择,把最小化漏报率的核参数作为 PSO 的目标函数,通过支持向量语言描述建立入侵检测模型. 算法验证实验在 KDD CUP'99 的标准入侵检测数据集上进行^[4],实验结果表明该方法提高了检测速度,有效降低了误检率。

2 PSO-SVDD 入侵检测方法

2.1 网络入侵检测工作原理

网络入侵检测首先采集计算机或网络系统的关键节点信息,然后采用适当的检测技术监控检测信息是否包含有恶意攻击行为. 网络入侵检测的本质是对网络攻击模式的识别,由于几乎无法获取多类攻击样本,异常模式也就难以确立. 为解决数据样本严重不平衡带来的分类边界划分难题,只能从已有采样目标类样本中进行学习并形成对该类别的资料描述,然后根据一种相似性度量,设定阈值来判别新样本的类属。

2.2 SVDD 算法原理

支持向量数据描述是由 Tax 和 Duin 在支持向量机理论上提出的一种新的分类方法. 给定数据集 $X = \{x_1, x_2, \dots, x_n\}$, 其中正常数据集为 $P = \{x_1, x_2, \dots, x_p\}$, 异常数据集为 $O = \{x_{p+1}, x_{p+2}, \dots, x_n\}$, 对于 SVDD 所建立的一个最小的包围球(包含所有正类数据), 可将球的中心设为 a , 其半径设为 R . 为使分类效果准确, 球边界应尽可能最小, 即球的半径也应尽可能的小. 另外, 为了使得某些特殊数据能够落于球外, 可设置松弛因子 ξ_1, ξ_2 .

根据以上说明, SVDD 的原问题可以用数学公式描述为

$$\begin{aligned} \min_{R, a, \xi_1, \xi_2} & R^2 + C_1 \sum_{i=1}^p \xi_i + C_2 \sum_{i=1}^n \xi_i \\ \text{s. t. } & \|x_i - a\|^2 \leq R^2 + \xi_i, \xi_i \geq 0, i = 1, 2, \dots, p \\ & \|x_l - a\|^2 \geq R^2 - \xi_l, \xi_l \geq 0, l = p+1, p+2, \dots, n \end{aligned} \quad (1)$$

式中, C_1 为正类的折中参数, C_2 为负类的折中参数. 引入松弛因子和折中参数, 是出于算法实效的考虑, 平衡所建立模型的复杂度和准确性. 为了获取求解问题的最优解, 采用拉格朗日乘法, 构造的拉格朗日函数如下:

$$\begin{aligned} L(R, a, \xi_1, \xi_2, \alpha_i, \alpha_l, \gamma_i, \gamma_l) = & R^2 + C_1 \sum_{i=1}^p \xi_i + \\ & C_2 \sum_{l=p+1}^n \xi_l \sum_{i=1}^p \gamma_i \xi_i - \sum_{l=p+1}^n \gamma_l \xi_l - \\ & \sum_{i=1}^p \alpha_i [R^2 + \xi_i - (x_i - a)^T (x_i - a)] - \\ & \sum_{l=p+1}^n \alpha_l [R^2 + \xi_l - (x_l - a)^T (x_l - a)] \end{aligned} \quad (2)$$

这样, 原问题(1) 可以转换为

$$\begin{aligned} \max_a & \sum_{i=1}^p \alpha_i k(x_i, x_i) - \sum_{l=p+1}^n \alpha_l k(x_l, x_l) - \\ & \sum_{i=1}^p \sum_{j=1}^p \alpha_i \alpha_j k(x_i, x_j) + 2 \sum_{i=1}^p \sum_{l=p+1}^n \alpha_i \alpha_l k(x_i, x_l) - \\ & \sum_{l=p+1}^n \sum_{lm=p+1}^n \alpha_l \alpha_m k(x_l, x_m) \\ \text{s. t. } & \sum_{i=1}^p \alpha_i - \sum_{l=p+1}^n \alpha_l = 1 \end{aligned} \quad (3)$$

$$0 \leq \alpha_i \leq C_1, i = 1, 2, \dots, p$$

$$0 \leq \alpha_l \leq C_2, l = p+1, p+2, \dots, n \quad (3)$$

式中, $k(x_i, x_j)$ 是核函数, 可采用径向基函数, 即 $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$, 对于测试样本, 如果 $\|z - a\|^2 \leq R^2$, 则 z 为正常数据, 反之为异常数据。

使用 SVDD 工具箱 Dd_tools^[5] 建模, 对 banana 数据集进行分类, 如图 1 所示. 图中“*”表示样本中的正类样本, 即目标点, “+”表示负类样本. 实线为分类边界, 虚线为理想的香蕉数据分类边界。

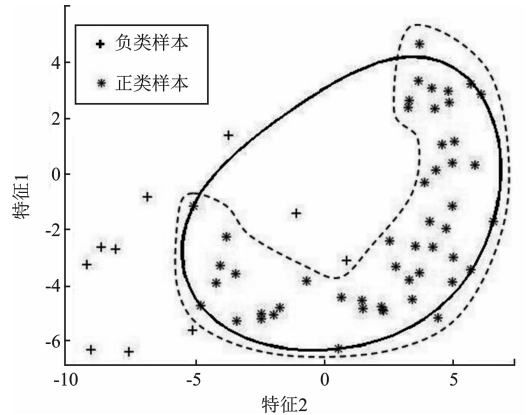


图1 Banana 数据分类

由上述分析可见, SVDD 方法的性能主要取决于 γ, C_1, C_2 等参数. 此外, 数据含有较多的冗余特征, 将会影响 SVDD 的性能. 因此, 为了提高 SVDD 算法性能, 可以从参数选择或特征选择等方面入手, 本文考虑优化特征选择, 将粒子群寻优算法引入到

对入侵检测数据的特征提取中。

2.3 PSO 特征优选的 SVDD 入侵检测

为了提高入侵检测速度,方法之一是通过消除冗余属性来降低数据规模。因此,构建入侵检测模型的第一步是对输入的数据样本进行特征筛选,选择那些影响分类结果的关键特征。可以用于特征优选的基于进化理论的算法很多,其中粒子群算法可在解空间追随最优粒子进行搜索,且较之遗传算法没有交叉和变异,易于实现。使用粒子群算法求解特征子集,可极大地降低搜索代价。

研究中,我们将完整的特征空间作为解空间,不同的特征组合可看作解空间中处于不同位置的粒子,粒子的维数就是特征空间的维数,粒子中各分量的值对应特征空间中各特征分量的状态。例如,某个 6 维粒子的位置向量为 110010,表示第 1、第 2 及第 5 特征是被选中的,其余特征不被选中。

粒子群算法的描述如下:假设粒子群中含有 N 个粒子,可把每个粒子看做 D 维空间中的一个活动点。并且可用 $v_i(t)$, $x_i(t)$, $p_i(t)$ 和 $g_i(t)$ 来表示粒子 i 在 t 时刻的速度、位置、个体最好位置和群体最好位置。由此可得粒子 i 的速度及位置的迭代公式如下:

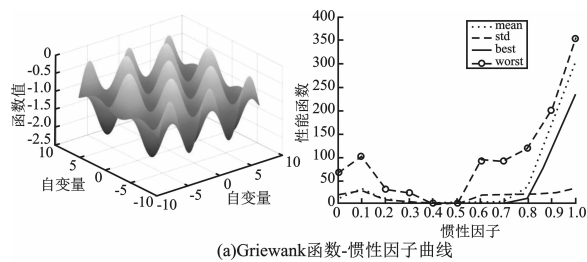
$$v_{id}(t+1) = \omega \cdot v_{id}(t) + c_1 r_1 (p_{id}(t) - x_{id}(t)) + c_2 r_2 (g_d(t) - x_{id}(t))$$

$$x_{id}(t+1) = \begin{cases} 1, r_3 < \text{sig}(x_{id}(t+1)) \\ 0, r_3 \geq \text{sig}(x_{id}(t+1)) \end{cases} \quad (4)$$

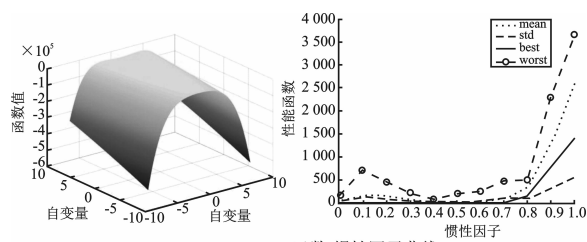
式中, $i = 1, 2, \dots, N$ (N 表示群体大小); $d = 1, 2, \dots, D$ (D 表示粒子编码中各分量的维数), r_1, r_2, r_3 为 0 到 1 的随机数; c_1, c_2 为学习因子; ω 为惯性因子,是一个非负数; $\text{sig}(\cdot)$ 为 sigmoid 函数,通常取 $\text{sig}(x) = 1/(1 + \exp(-x))$ 。

惯性因子 ω 是粒子群算法的一个重要参数,选择恰当的惯性因子,对于提升粒子群算法的性能意义重大。^[6] 当 ω 较大时,粒子群算法能够较好地进行全局搜索,有利于算法探索;当它较小时,粒子群算法能够较好地进行局部搜索,有利于算法开发。为了评价算法的性能,本文采用 Griewank 函数、Rosenbrock 函数对粒子群算法的性能和参数进行分析进行测试。

本文在种群规模等于 20, 最大迭代次数为 2 000, $c_1 = c_2 = 2$, 问题维数为 30 的情况下,惯性因子在 0 到 1 之间间隔 0.1 取 11 个值,对这两种性能函数分别进行 25 次实验,取性能函数的均值、方差、最差期望、最好期望进行分析,结果如图 2 所示。



(a) Griewank 函数-惯性因子曲线



(b) Rosenbrock 函数-惯性因子曲线

图 2 惯性因子与性能函数的曲线

图 2 中左图分别为性能函数的三维显示,右图为惯性因子与性能函数统计量的关系曲线。通过实验分析可知,当惯性因子 ω 取值在 0.4~0.5 时,其性能函数的结果要比取其他值是至少好两个数量级,而且随着性能函数的均值的减小,相应的标准差,最好期望,最差期望也相应减小。

不同的适应度函数,其粒子的搜索方向也不同,粒子群算法通过设置目标函数,来在搜索过程中确定具有最强分类能力的特征组合。本文使用距离准则作为判据,选择特征子集时考虑类间距离最大且类内各样本间的距离最小,具体如下:

$$\text{fitness}_i = S_b - S_w \quad (5)$$

式中, S_b 为类间欧氏距离, S_w 为类内样本距离。

基于上述理论, PSO-SVDD 入侵检测方法的流程图如图 3 所示。

3 实验验证

3.1 数据来源

为了验证算法有效性,采用 KDD CUP'99^[7] 实验数据集展开研究和分析。该数据集是 1998 年 MIT Lincoln 实验室与美国国防部高级规划署合作开展入侵检测系统评估所建设的,一共收集了 9 周的 TCPdump 资料,含约 700 万条连接记录,所有的数据记录已经进行了初步的处理和特征提取。数据样本中攻击记录数量较多,攻击类型大致有 36 种,可以分为 4 大类: DoS——拒绝服务攻击, R2L——来自远程的非授权访问, U2R——非法获得超级用户

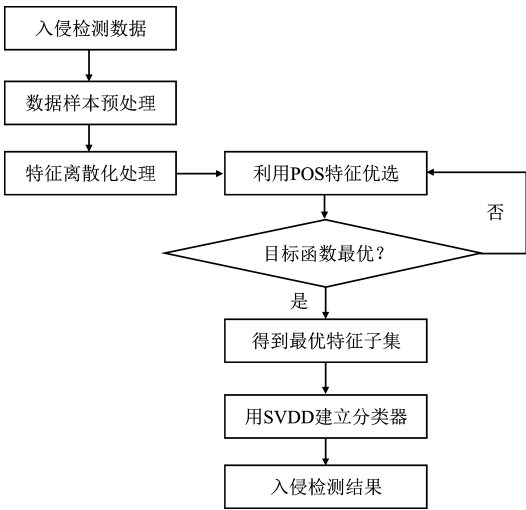


图 3 PSO-SVDD 入侵检测方法的流程

权限,Probe——测和扫描。每条连接记录有网络包基本信息、各应用的描述信息及两秒窗口统计信息等 41 个特征。目前,KDD CUP’99 数据集已经成为研究和评估入侵检测的标准数据集。

3.2 实验方案设计

考虑到 KDD CUP’99 原始数据集较大,按照大部分研究者所采用的常规做法,本实验方案也选择 kddcup. data_10. percent 作为模型的训练数据。该数据集中已有 494 021 条记录被标号成正确的攻击类型,其中 Normal 有 97 278 条记录,Probe 有 4 107 条记录,DoS 有 391458 条记录,R2L 有 1 126 条记录,U2R 有 52 条记录。本文所提出的 PSO-SVDD 入侵检测方法属于异常检测,其中所用到的 SVDD 方法属于单类分类器。所以从 kddcup. data_10. percent 中抽取数据记录经预处理后作为训练数据集。

实验中,按以下步骤对数据进行预处理。

(1)去除重复样本,合并属性值相同的数据。

(2)由于 SVDD 只能对数字型数据进行处理,所以需对具有字符型属性的数据进行转换,对数量级相差较大的数据进行规范化,将所有特征规范到 [0,1] 范围内。字符型特征“服务类型”可能的取值有几十个之多,将其归为 http、smtp、ftp、ftp_data、others 5 种,然后按上述方法处理。

(3)数据规范化处理

为防止那些初时值取值域较大的属性和取值域较小的属性相比,其权重过大,对一些连续的属性进行了规范化处理,其方法如式(6)所示。

$$x'_{i,j} = \left| \frac{x_{i,j} - \min_j(x_{i,j})}{\max_j(x_{i,j}) - \min_j(x_{i,j})} \right| \tag{6}$$

经上述步骤处理后,训练数据集的构成如表 1 所示,考虑到 kddcup. data_10. percent 数据集中 U2R 仅有 52 条记录,为扩大样本集,从 kddcup. data. corrected 数据集中抽取 396 条数据作为训练数据的补充。

表 1 训练数据集的构成

类型	Normal	DoS	R2L	U2R	Probe
Normal	564	4860	360	20	290
DoS	3680	680	180	20	290
R2L	3680	1300	450	20	290
U2R	3680	1300	180	396	290
Probe	3680	1300	180	20	570

实验中,粒子群参数粒子数设为 40,最大迭代次数 N 设为 100, $c_1 = c_2 = 2$,惯性因子设置为 0.45,SVDD 算法中核函数 $k(x_i, x_j)$ 采用径向基函数,即 $k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ 。测试数据也是从 kddcup. data. corrected 数据集中获得,一共抽取 5 000 条记录作为测试数据集。

3.3 实验结果分析

实验平台采用 PC 机配置如下:CPU 为 Pentium Dual-core 3.2GHz,内存为 2.00 GB,硬盘为 500 GB,所采用操作系统为 Window XP。在 Matlab 2008a 中实现本文算法。实验中,Normal、DoS、R2L、U2R、Probe 等 5 种不同分类器各自所选择的特征如表 2 所示。

表 2 不同分类器特征选择结果

Normal	DoS	R2L	U2R	Probe
1		3	3	3
3	3	5	5	5
5	5	6	14	27
6	23	22	23	40
23	38	33	32	41
25		37	33	

为衡量本文参数选择优化的有效性,将传统的 SVDD 作为参比模型,其检测效果如表 3 所示,表中的检测准确率和误报率的计算方法如前所述。从表 3 可看出采用粒子群算法进行特征提取的 SVDD 算法比原算法的检测率要高,误报率低于原算法。在对 R2L 的入侵检测时效果不太理想,其原因在于 R2L 入侵通过伪装成为合法用户,这使得它的特征和正常数据相比基本类似,因此检测效果不理想。

表 3 检测结果对比表 %

样本类型	SVDD		PSO-SVDD	
	检测准确率	误报率	检测准确率	误报率
Normal	87. 24	0. 23	92. 30	0. 14
DoS	86. 56	0. 15	95. 14	0. 07
R2L	78. 09	0. 26	82. 47	0. 16
U2R	85. 74	0. 17	89. 53	0. 09
Probe	81. 27	0. 29	93. 61	0. 13

为比较计算量大小,表 4 给出了两种算法的时间对比情况,从结果可看出本文算法所需的训练和测试时间大大减少,这是由于采用 PSO 特征优选后,参与入侵检测的数据量大大减少,因此计算量较小.

表 4 检测速度对比 s

样本类型	SVDD		PSO-SVDD	
	训练	测试	训练	测试
Normal	51	23	25	11
DoS	52	26.	27	12
R2L	57	22	22	9
U2R	56	24	21	14
Probe	54	24	20	10

4 结束语

本文从入侵检测特征提取和分类模型建立两方面,提出了基于粒子群优化的支持向量描述算法的入侵检测方法,该方法降低了冗余数据对分类器性

能的影响. 通过特征提取降低特征数据的维数,降低了计算的复杂程度,节省了计算时间且提高了分类器的性能. 在以后的工作中,将进一步研究算法的适应性和实时性.

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):798-802.

[2] 刘宇,朱随江,刘宝旭. 采用改进 PTSVM 的入侵检测研究[J]. 计算机工程与应用,2012,48(5):1-3,74.

[3] Tax D M J, Duin R P W. Support vector data description[J]. Machine Learning,2004,54(1):45-66.

[4] Mahbod Tavallace, Ebrahim Bagheri, Wei Lu, et al. A Detailed Analysis of the KDD CUP 99 Data Set[C]// Proceeding of the Setonal IEEE International Conference on Computational Intelligence for Security and Defense Applications. [s. l.]: IEEE,2009:1-10.

[5] Dd_tools (Version 2. 0. 0 by July 2013[EB/OL]. [2015-10-27]. http://prlab.tudelft.nl/david-tax/dd_tools.html, 2013.

[6] 金欣磊,马龙华,吴铁军,等. 基于随机过程的 PSO 收敛性分析[J]. 自动化学报,2007,33(12):1263-1268.

[7] 梁海军. 人工蜂群优化支持向量机算法在网络安全中的应用[J]. 微电子学与计算机,2013,30(2):95-98.

作者简介:

魏振伟 男,(1981-),硕士,工程师. 研究方向为检测与诊断,模式识别. E-mail:ygszzia@126. com

刘 飞 男,(1981-),博士研究生. 研究方向为计算机系统安全.

(上接第 143 页)

[9] 王健. 椭圆曲线加密体质的双有限域算法及其硬件实现[D]. 北京:北京大学,2008.

[10] 陈传鹏,覃中平. 改进的素数域椭圆曲线密码处理器[J]. 武汉大学学报(工学版),2011,44(1):124-127.

[11] Urbano-Molano F A, Trujillo-Olaya V, Velasco-Medina J. Design of an Elliptic Curve Cryptoprocessor using Optimal Normal Basis over GF(2²³³)[C]//Circuits and Systems(LASCAS). Cusco: IEEE,2013:1-4.

作者简介:

陈俊杰 男,(1991-),硕士. 研究方向为计算机系统结构与 VLSI. E-mail:chenjj519@163. com.

孟李林 男,(1962-),教授. 研究方向为计算机系统结构,嵌入式系统.

袁 阳 男,(1990-),硕士. 研究方向为集成电路系统设计.