文章编号: 1001-0920(2017)04-0755-04 DOI: 10.13195/j.kzyjc.2016.0133

基于PLS特征提取的网络异常入侵检测CVM模型

吴丽云[1], 李生林[1], 甘旭升[2†], 王明华[2]

(1. 中国人民解放军后勤工程学院，重庆 401311；2. 空军工程大学 空管领航学院，西安 710051)

摘 要: 为改善网络安全防护水平,提出一种基于偏最小二乘(PLS)法和核向量机(CVM)的组合式异常入侵检测方法. 首先,采用PLS算法提取网络数据的主成分,构建特征集;然后,利用CVM构建特征集的异常入侵检测模型,进而完成异常入侵检测和判定. 仿真实验结果表明,所提出的方法具有CVM的大规模数据快速处理能力,而且检测性能与L1-SVM和L2-SVM大致相当,尤其主成分数为1 538时能保持相对较高的检测水平,验证了将其用于异常入侵检测的有效性和可行性.

关键词: 核向量机；偏最小二乘；特征提取；入侵检测

中图分类号: O235；TN915.08 **文献标志码:** A

Network anomaly intrusion detection CVM model based on PLS feature extraction

WU Li-yun[1], *LI Sheng-lin*[1], *GAN Xu-sheng*[2†], *WANG Ming-hua*[2]

(1. Logistic Engineering University of PLA，Chongqing 401311，China；2. Air Traffic Control and Navigation College，Air Force Engineering University，Xi'an 710051，China)

Abstract: In order to improve the protection level of network security, a combined anomaly intrusion detection method based on the partial least squares(PLS) method and the kernel vector machine(CVM) is proposed. The PLS algorithm is used to extract the principal components of the network data and construct the feature set. Then the CVM is applied to build the anomaly intrusion detection model of the feature set, completing the detection and decision of abnormal intrusion. Simulation results show that the proposed method has the fast processing ability similar with the CVM for the large scale data, and the detection performance is roughly equivalent to that of L1-SVM and L2-SVM, and the detection level is relatively high when the principal component is 1 538, which proves the effectivenes and feasibility of the method in the application of anomaly intrusion detection.

Keywords: core vector machine；partial least square；feature extraction；intrusion detection

0 引 言

随着计算机技术和网络技术的飞速发展,人们的工作、学习和生活越来越依赖网络,随之产生的网络安全问题也突显出来. 而异常入侵检测是改善网络安全的一种策略,目前已经成为热点. 异常入侵检测实质上是个分类问题,即采取一定技术手段将收集到的数据分为正常和异常两类. 为了提高异常入侵检测的效率和效果,近些年来,学者们将诸如免疫算法、粗糙集、神经网络和支持向量机(SVM)等人工智能方法用于网络的异常入侵检测[1-4],其中尤以SVM的检测效果最为高效. 只需收集必要数据,SVM便能根据用户行为和网络活动的特征模式检测出异常入侵行为,此外,其联想功能也能帮助识别新的入侵行为和已知入侵行为的变种. 尽管这样,在实际应用中,SVM也暴露出不足: 1) SVM无法提取特征,而入侵检测所需的属性较多,其对应数据中可能存在噪声,且可能彼此相关,这些会影响模型; 2) 异常入侵检测的数据处理量很大,采用传统SVM会耗费大量资源和时间,无法满足在线检测的实时性要求.

鉴于此,本文提出基于偏最小二乘(PLS)特征提取和核向量机(CVM)的组合异常入侵检测方法,以满足实用性要求. 仿真结果表明,所提出的方法具有CVM的大规模数据快速处理能力,而且检测性能与L1-SVM和L2-SVM大致相当,尤其主成分数为1 538时能够保持相对较高的检测水平,验证了将其应用于异常入侵检测的有效性和可行性.

收稿日期: 2016-01-30；**修回日期:** 2016-07-03.

基金项目: 陕西省自然科学基础研究计划项目(2015JM7364).

作者简介: 吴丽云(1979−), 女, 工程师, 博士生, 从事后勤信息化的研究；李生林(1964−), 男, 教授, 博士生导师, 从事军事后勤与后勤信息化等研究.

†通讯作者. E-mail: ganxusheng123@163.com

1 入侵检测问题概述

入侵是指在未授权情况下任何试图危及网络资源的完整性、机密性或可用性的故意行为. 入侵检测是指对已经实施、正在实施或试图实施的入侵行为的发现和识别,即收集系统与网络中的诸多关键点信息,并利用一定手段处理这些信息,以此判定是否受到攻击以及是否背离了现有安全策略[5].

入侵检测系统是专门针对网络安全的主动式防护系统,它依据一定的安全策略监视网络系统运行,发现各种入侵的行为、企图或结果,并自动对检测出的入侵特征作出响应,以有效防范非法访问或入侵行为. 该系统一般采用误用入侵检测和异常入侵检测两种处理方法. 前者先将所有可能发生的不利的、不可接受的行为归纳建立一个模型,凡是符合该模型的访问行为将被判定为入侵;后者先构建一个正常访问行为的系统模型,凡是不符合这个模型的访问将被判定为入侵,其原理如图1所示. 本文关注的是异常入侵检测[6-7].

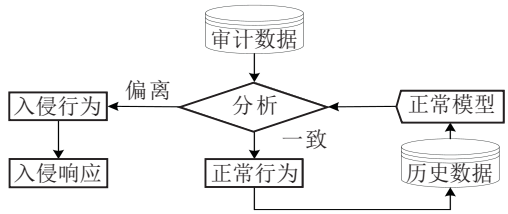


图1 异常入侵检测原理

2 偏最小二乘特征提取

PLS算法在一个统计框架内完成多元线性回归、主成分分析和相关性分析,当用于特征提取时,获取的主成分既能归纳自变量中蕴含的信息,又能诠释因变量,是统计学发展中的一个里程碑[8].

假设有p个自变量$(x_1,x_2,\cdots,x_p)$和q个因变量$(y_1,y_2,\cdots,y_q)$,选取n个样本点,构建数据块$\boldsymbol{X}=[x_1,x_2,\cdots,x_p]_{n\times p}$和$\boldsymbol{Y}=[y_1,y_2,\cdots,y_q]_{n\times q}$. PLS算法分别从$\boldsymbol{X}$和$\boldsymbol{Y}$中获取主成分$t_1$和$u_1$($t_1$和$u_1$分别为$x_1,x_2,\cdots,x_p$和$y_1,y_2,\cdots,y_q$的线性组合),需满足:

1) t_1包含$\boldsymbol{X}$尽可能多的差异信息,u_1包含$\boldsymbol{Y}$尽可能多的差异信息,即$\mathrm{var}(t_1)\to\max$, $\mathrm{var}(u_1)\to\max$,其中$\mathrm{var}(\cdot)$为方差算子;

2) t_1与u_1达到最大相关,即$r(t_1,\ u_1)\to\max$,其中$r(\cdot)$为相关系数算子.

实际上,这两个条件可转化为如下优化问题:

$$\max_{w_1,c_1} w_1^{\mathrm{T}}\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0c_1 : w_1^{\mathrm{T}}w_1=1,\ c_1^{\mathrm{T}}c_1=1. \tag{1}$$

其中:w_1为$\boldsymbol{X}$的第1主轴,c_1为$\boldsymbol{Y}$的第1主轴,$\boldsymbol{E}_0$为$\boldsymbol{X}$的标准化,$\boldsymbol{F}_0$为$\boldsymbol{Y}$的标准化. 通过以上处理,t_1便可以最大程度地表征$\boldsymbol{X}$,并且能最好地解释$\boldsymbol{Y}$.

第1次获取主成分t_1和u_1后,分别对$\boldsymbol{E}_0$和$\boldsymbol{F}_0$进行关于t_1的线性回归,如果此时达到所要求的精度,则计算结束;否则,在t_1解释后的$\boldsymbol{X}$与$\boldsymbol{Y}$剩余信息的基础上,第2次提取主成分t_2. 依此继续下去,直到满足精度要求. 当将PLS用于特征提取时,提取到所需的主成分数$m(m<\mathrm{rank}(\cdot),\mathrm{rank}(\cdot)$用于对矩阵求秩)即可停止运算. PLS算法伪代码如下:

for $i=1$ to m

$w_i=\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0/\|\boldsymbol{E}_0^{\mathrm{T}}\boldsymbol{F}_0\|$

$t_i=\boldsymbol{E}_0w_i$

$r_i=\boldsymbol{E}_0^{\mathrm{T}}t_i/\|t_i\|^2$

$p_i=\boldsymbol{E}_0^{\mathrm{T}}t_i/\|t_i\|^2$

$\boldsymbol{E}_0=\boldsymbol{E}_0-t_is_i^{\mathrm{T}}$

$\boldsymbol{F}_0=\boldsymbol{F}_0-t_ir_i^{\mathrm{T}}$

end

通过循环,可提取主成分$\boldsymbol{T}=[t_1,t_2,\cdots,t_m]$,并同时得到$\boldsymbol{W}=[w_1,w_2,\cdots,w_m]$和$\boldsymbol{S}=[s_1,s_2,\cdots,s_m]$. 对于测试样本$\boldsymbol{X}_t$的标准化矩阵$\boldsymbol{E}_{0t}$,其主成分投影矩阵为

$$T_{\mathrm{test}}=E_{0t}W(S^{\mathrm{T}}W)^{-1}T. \tag{2}$$

3 核向量机

CVM是一种基于计算几何学的大数据学习算法,它将以往SVM中对二次规划的求解变换成对最小包含球(MEB)的求解,并利用$(1+\varepsilon)$-近似算法获得最优解. 在快速处理大样本、复杂非线性等问题时具有优势[9-10].

设样本点为$x_1,x_2,\cdots,x_n,x_i\in R^d$,将之由一个非线性函数$\phi$映射到某一特征空间,得到相应映射点$S_\phi=\{\phi(x_1),\phi(x_2),\cdots,\phi(x_n)\}$. 在该特征空间中,MEB可表示为$\boldsymbol{B}(c',R')$,$c'$和$R'$分别为球心和半径. 要寻求包含所有映射点的最小球等价为

$$(c*,R*)=\arg\min_{R,c} R^2:\|c-\phi(x_i)\|\leqslant R^2. \tag{3}$$

由Lagrange法,得到对偶矩阵形式为

$$\max_{\alpha}\alpha^{\mathrm{T}}\mathrm{diag}(K)-\alpha^{\mathrm{T}}K\alpha:\alpha\geqslant 0,\ \alpha^{\mathrm{T}}e=1. \tag{4}$$

其中:$\alpha=[\alpha_1,\alpha_2,\cdots,\alpha_n]^{\mathrm{T}}$为Lagrange乘子,$\boldsymbol{K}_{m\times m}=[K(x_i,x_j)]=[\langle\phi(x_i),\phi(x_j)\rangle]$为核函数矩阵,$\boldsymbol{0}=[0,\cdots,0]^{\mathrm{T}}$,$\boldsymbol{e}=[1,\cdots,1]$. 若式(4)满足条件

$$K(x,x)=\eta, \tag{5}$$

且$\alpha^{\mathrm{T}}e=1,\alpha^{\mathrm{T}}\mathrm{diag}(\boldsymbol{K})=\eta$(常数),则式(4)简化为

$$\max_{\alpha}-\alpha^{\mathrm{T}}K\alpha:\alpha\geqslant 0,\alpha^{\mathrm{T}}e=1. \tag{6}$$

无论何时,只要满足式(5),具有式(4)形式的QP问题,便可视为一个MEB问题,CVM选择的核函数为

$$\kappa(z_i,z_j)=y_iy_jK(x_i,x_j)+y_iy_j+\delta_{ij}/C. \tag{7}$$

由于$\kappa(z,z)=\eta+1+1/C$为常数,满足式(5),式(6)可视为MEB问题. 对于任意一点,有

$$\tilde{\phi}(z_i) = [y_i\phi(x_i)\ \ y_i\ \ \tau_i/\sqrt{C}]^{\mathrm{T}}, \tag{8}$$

其中τ_i为m维向量,除第i个元素为1外,其余皆为0.

CVM使用迭代的$(1+\varepsilon)$-近似算法求取MEB问题的近似最优解. 对于第t次迭代,核心集表示为S_t,球心表示为C_t,半径表示为R_t. 根据预先给出的$\varepsilon > 0$,CVM的$(1+\varepsilon)$-近似算法实现流程如图2所示.

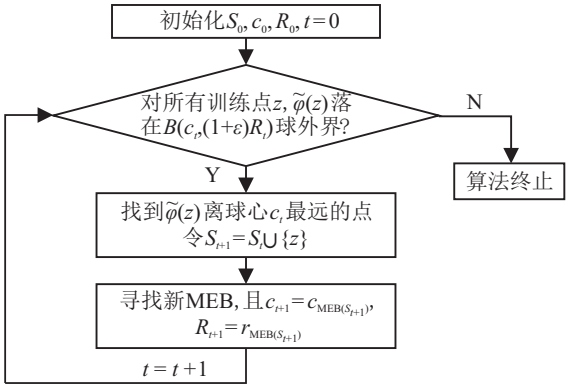


图 2　$(1+\varepsilon)$近似算法实现流程

通过以上过程,加入到核心集的点称为核心向量,得到的MEB如图3所示.

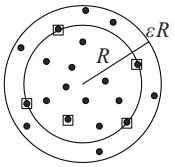


图 3　最小包含球(MEB)

方块集合为一个核心集(Core Set,CS),内圆为方块集合的MEB,外圆为内圆的$(1+\varepsilon)$延拓,包围住所有的点. 尽管过程较简单,但相对于SVM中的二次规划问题,计算复杂性大大降低.

4　PLS-CVM的异常入侵检测模型

PLS-CVM算法实际上是一种组合式分类器. 采用PLS-CVM构建入侵检测模型原理如图4所示,具体实现步骤如下.

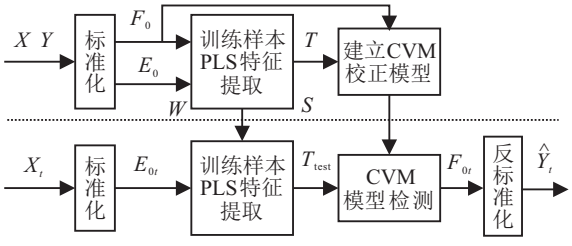


图 4　PLS-CVM入侵检测建模原理

Step 1: 提取网络连接与审计数据特征.

1) 将训练用网络连接与审计数据块$\boldsymbol{X}$标准化为$\boldsymbol{E}_0$, $\boldsymbol{Y}$标准化为$\boldsymbol{F}_0$;

2) 通过PLS算法获取$\boldsymbol{E}_0$与$\boldsymbol{F}_0$的主成分, 计算出向量$\boldsymbol{T}$、$\boldsymbol{W}$和$\boldsymbol{S}$;

3) 将测试用数据块$\boldsymbol{X}_t$标准化为$\boldsymbol{E}_{0t}$, 并由式(2)得到其主成分投影$\boldsymbol{T}_{\mathrm{test}}$.

Step 2: 建立异常入侵检测CVM模型.

将$\boldsymbol{T}$作为输入,$\boldsymbol{Y}$作为输出,选取合适参数,训练异常入侵检测CVM模型.

Step 3: 测试异常入侵检测CVM模型.

将$\boldsymbol{T}_{\mathrm{test}}$输入所构建的CVM异常入侵检测模型,对输出值作反标准化处理, 完成模型测试.

5　实验仿真

实验环境: Pentium IV 2.4 GHz CPU, 2 GB DDR内存, 80 GB+7 200转硬盘; Windows XP操作系统. 仿真中,L1-SVM和L2-SVM采用LIBSVM软件,CVM采用LIBCVM软件,涉及的核函数均采用径向基函数.

5.1　实验方案设计

为验证PLS-CVM算法在解决异常入侵检测问题上的优越性,实验数据选择入侵检测领域广泛采用的KDD99数据集的子集Kddcup. data_10_percent.

实验前,将随机得到的样本数据的正常记录标记为1,入侵记录标记为0,对离散型和符号型的属性值作数字化处理. 如对于协议类型属性: tcp表示为1, icmp表示为2, udp表示为3. 此外,为防止取值范围较大属性的影响强于其他属性,并兼顾计算速度,需预先对样本数据在[0,1]区间作标准化处理.

具体方案为: 从Kddcup.data_10_percent中随机抽取5 000, 10 000, 20 000, 40 000个样本,组成4个训练集,同时,对应于各训练集分别构建包含10 000个随机样本的测试集. 实验中,首先对各训练集分别建立基于所有41个属性特征值的L1-SVM、L2-SVM和CVM的检测模型,然后使用PLS算法对各训练集分别提取15个主成分,建立基于相应主成分的CVM检测模型,并采用测试集进行检验. 模型性能的评价指标如下: 支持向量(SV)、CPU时间(单位秒)、检测精度(DA)、检测率(DR)和误报率(FPR). 其中

DA = 正确分类的样本个数/样本总数,

DR = 检测出的异常样本数/异常样本总数,

FPR = 误报为异常的正常样本数/正常样本总数.

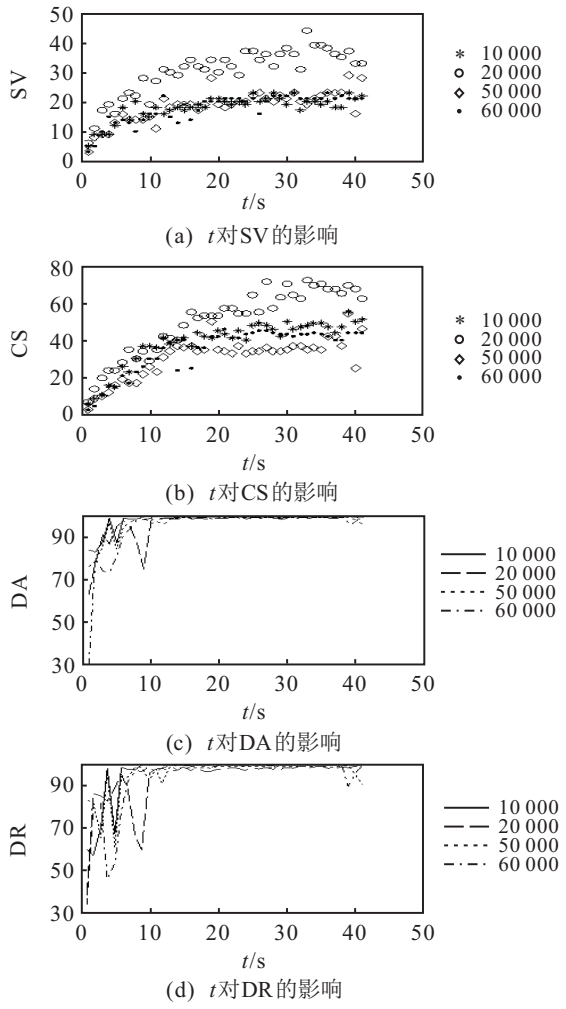
5.2　实验结果分析

表1为不同规模训练集下建立的L1-SVM、L2-SVM、CVM和PLS-CVM异常入侵检测模型性能的实验测试结果对比. 图5为不同规模训练集下主成分数t对PLS-CVM入侵检测模型性能的影响.

仿真结果表明,在检测精度、检测率和误报率方面,CVM模型相对L1-SVM和L2-SVM模型并无优势,有时甚至较差,但在建模CPU时间上优势明显,相差几个数量级,所需支持向量也远少于L1-SVM和L2-SVM模型. 但相同条件下,PLS-CVM模型不仅继承了CVM处理大规模数据的速度优势,而且得到了与L1-SVM和L2-SVM大致相当的检测精度、检测率

表 1 不同规模训练集下的实验结果对比

训练集规模	模型	SV个数	CPU时间/s	DA/%	DR/%	FPR/%
5 000	LI-SVM	65	19.06	99.69	99.49	0.42
	L2-SVM	216	17.63	99.68	99.48	0.33
	CVM	22	0.53	97.49	94.76	2.26
	PLS-CVM	22	0.67	99.69	99.31	0.34
10 000	LI-SVM	51	14.05	99.69	99.49	0.40
	L2-SVM	80	18.28	99.75	99.63	0.35
	CVM	20	1.22	98.64	98.45	1.21
	PLS-CVM	16	1.72	99.52	99.27	0.37
20 000	L1-SVM	125	150.86	99.84	99.66	0.32
	L2-SVM	1 769	186.64	99.85	99.70	0.27
	cvm	19	0.98	98.99	97.84	1.30
	PLS-CUM	25	2.31	1 000.00	1 000.00	0.00
40 000	L1-SVM	185	1 033.81	99.91	99.77	0.21
	L2-SVM	1 666	3 565.17	99.85	99.70	0.19
	CVM	27	1.19	80.04	50.32	39.27
	PLS-CVM	21	2.72	90.72	99.51	0.26



(a) t对SV的影响

(b) t对CS的影响

(c) t对DA的影响

(d) t对DR的影响

图 5 主成分数对PLS-CVM模型性能的影响

和误报率,所需支持向量也与CVM算法较为接近.这说明,采用经PLS算法降维、去噪和消除相关性处理后的特征构建的CVM模型, 入侵检测性能有了显著提高.由图5可见,在不同规模训练集基础上得到的PLS-CVM模型的检测精度和检测率,在主成分个数为15~38时能够保持相对较高的水平,而主成分数为1~14和39~41时则因提取的有用信息不足或引入噪声成分而受到一定影响.此外,PLS-CVM采用不同主成分数建模所需的支持向量个数和核心向量个数都较少,这对大规模入侵检测数据的处理也非常有利.

6 结 论

本文针对网络安全问题,提出了PLS-CVM异常入侵组合检测方法.PLS算法能够减少变量维数,去除噪声污染,消除变量间多重相关性,CVM在快速处理大样本数据方面具有优势,将两者有机结合,优势互补,以期解决异常入侵检测的大样本快速建模问题.仿真表明,当主成分个数处于15~38之间时,PLS-CVM继承了单纯CVM的快速处理数据优势,简化了建模难度,检测效率也较单纯CVM有所改善,从而为异常入侵检测问题提供了新的解决途径.

参考文献(References)

[1] 冯翔,马美怡,赵天玲,等.基于复合免疫算法的入侵检测系统[J].计算机科学, 2014, 41(12): 43-47.
(Feng X, Ma M Y, Zhao T L, et al. Intrusion detection system based on hybrid immune algorithm[J]. Computer Science, 2014, 41(12): 43-47.)

[2] 杨波, 王欣, 杜佳. 基于粗糙集的自适应网络入侵检测方法[J]. 计算机应用与软件, 2014, 31(11): 318-320.
(Yang B, Wang X, Du J. Adaptive network intrusion detection method based on rough set[J]. Computer Application and Software, 2014, 31(11): 318-320.)

[3] 王亚, 熊焰, 龚旭东, 等. 基于混沌PSO算法优化RBF网络入侵检测模型[J]. 计算机工程与应用, 2013, 49(10): 84-87.
(Wang Y, Xiong Y, Gong X D, et al. Based on chaos PSO algorithm optimize RBF network intrusion detection[J]. Computer Engineering and Applications, 2013, 49(10): 84-87.)

[4] 孙敏, 徐彩霞, 高阳. 基于FWKN-SVM的Android异常入侵检测的研究[J]. 计算机科学, 2015, 42(4): 116-118.
(Sun X, Xu C X, Gao Y. Research of android abnormal intrusion detection based on feature-weighted K-nearest-neighbor SVM[J]. Computer Science, 2015, 42(4): 116-118.)

[5] Bace R. Intrusion detection[M]. New York: Macmillan Technical Publishing, 2000.

[6] Verwoerd T, Hunt R. Intrusion detection techniques and approaches[J]. Computer Communications, 2002, 25(15): 1356-1365.

[7] Endorf C, Schultz E, Mellander J. Intrusion detection & prevention[M]. New York: Mc Graw-Hill, 2004.

[8] Barker M, Rayens W. Partial least squares for discrimination[J]. J of Chemometrics, 2003, 17: 166-173.

[9] Tsang I W, Kwok J T, Cheung P M. Core vector machines: Fast SVM training on very large data sets[J]. J of Machine Learning Research, 2005, 6: 363-392.

[10] Chu C S, Tsang I W, Kwok J T. Scaling up support vector data description by using core sets[C]. Proc of IEEE Int Joint Conf on Neural Networks. Hong Kong, 2004: 425-430.

(责任编辑: 郑晓蕾)