基于特征选取和样本选择的网络入侵检测①

马世欢, 胡 彬

(河南工业职业技术学院 计算机工程系, 南阳 473000)

摘 要: 为了获得更加理想的网络入侵检测结果, 针对网络入侵特征选取和样本选择问题, 提出一种基于特征选取和样本选择的网络入侵检测模型. 首先提取网络入侵特征, 并进行归一化处理, 然后采用核主成分分析选择入侵特征, 并对样本进行选择, 最后采用极限学习机建立网络入侵检测分类器, 并采用 KDD Cup99 数据集进行仿真实验. 仿真结果表明, 本文模型得到了理想的网络入侵检测结果, 检测率超过 95%以上, 入侵检测效率可以满足网络安全实际应用要求.

关键词: 网络安全; 样本选择; 核主成分分析; 极限学习机

Network Intrusion Detection Based on Features Selecting and Samples Selecting

MA Shi-Huan, HU Bin

(Department of Computer Engineering, Henan Polytechnic Institute, Nanyang 473000, China)

Abstract: In order to obtain a more ideal network intrusion detection results, according to the network intrusion feature selection and sample selection problem, this paper proposes a network intrusion detection model based on features selecting and samples selecting. Firstly, the features of network intrusion are extracted, and normalized, and secondly kernel principal component analysis is used to select intrusion features, and the samples are selection, finally, extreme learning machine is used to set up network intrusion detection classifier, and the simulation experiments are carried out with KDD Cup99 data. The simulation results show that that the proposed model has been better network intrusion detection results, the detection rate is above 95%, the efficiency of intrusion detection can meet the requirements of network security protection.

Key words: network security; samples selection; kernel principal component analysis; extreme learning machine

随着网络应用的日益广泛, 网络安全问题越来越严重, 其中网络入侵是指试图破坏计算机及相关资源的非法行为, 入侵检测是采用一定的技术对入侵行为进行分析和识别[1]. 相对于传统网络安全防范措施, 入侵检测可以发现新型以及变异的入侵行为, 因此网络入侵检测结果的好坏一直人们广泛关注的话题[2].

当前入侵检测划分为两类: 误差检测和异常检测, 由于误差检测只能检测到已经存在的入侵行为, 不能检测新的入侵检测行为或者变异的入侵检测行为, 不能有效保证网络的安全[3]. 异常检测可以发现入侵检测行为或者变异的入侵检测行为, 因此成为当前主要研究方向[4]. 特征选择是网络入侵建模的第一步, 为了提前网络入侵正确率, 一般尽可能多的提取网络行为特征, 导致网络特征规模相当的大, 包含了一些入侵检测无关特征, 以及一些冗余特征, 如果将特征全部输入到分类器进行分类, 那么计算复杂度相当的高, 对入侵检测效率产生不利影响, 而且误检率、漏检率较高, 特征选择就是从特征集选择一些关键特征, 降低特征维数, 减少网络入侵检测的时间[5], 当前特征选择主要包括主成分分析(principal component analysis, PCA)、特征关联分析、群智能优化算法等[6-8], 主成分分析法简单、易实现, 但是其易破坏原始特征的解释性; 特征关联分析方法假设特征与入分类是一种线性映射关系, 难以获得最优特征; 群智能算法将特征选

① 收稿时间:2015-01-30;收到修改稿时间:2015-03-12

择转化为一类优化问题, 通过模拟自然界的生物机制如进化机制, 找到最优特征子集, 取得了不错的应用效果[9]. 网络入侵分类主要采用支持向量机和神经网络进行, 但是神经网络要求“样本”大、成本高, 支持向量机学习速度慢、影响入侵检测的实时性[10].

要建立性能最优的入侵检测模型, 必须选择最合理特征和学习样本, 为了保证网络安全性, 提出一种基于特征选取和样本聚类的网络入侵检测算法(KPCA-ELM), 最后采用仿真实验测试对其性能进行分析.

1 KPCA-ELM的入侵检测原理

基于 KPCA-ELM 的网络入侵检测模型基本原理为: 首先提取网络入侵特征, 并进行归一化处理, 然后采用核主成分分析(kernel principal component analysis, KPCA)[12]选择入侵特征, 并对样本进行选择, 最后采用极限学习(Extreme Learning Machine, ELM)[11]建立网络入侵检测分器, 具体如图 1 所示.

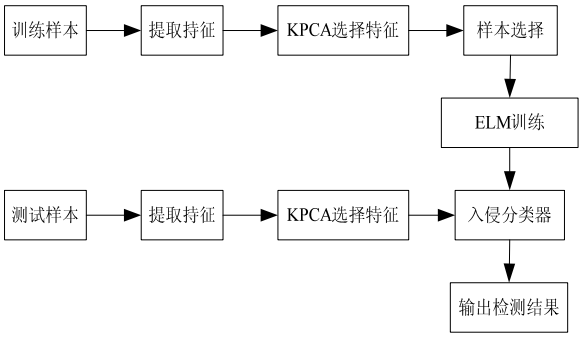

图 1 KPCA-ELM 的网络入侵检测原理

2 KPCA-ELM网络入侵检测模型

2.1 KPCA 选择网络入侵特征

假设输入空间样本集为 $X=(x_1^d,x_2^d,...,x_N^d)$, Φ 为从输入空间到高维特征空间的映射: $X^d \rightarrow F$, 协方差矩阵为:

$$C=\frac{1}{N}\phi(X)\phi(X)^T \tag{1}$$

构造核函数:

$$K=k(x,y)=(\phi(x),\phi(y))=\phi(x)^T\phi(y) \tag{2}$$

求解特征方程:

$$Cv=lv \tag{3}$$

其中, v 和 λ 分别为 C 对应的特征向量和特征值.

特征向量 v 的计算公式如下:

$$v=\sum_{i=1}^{N}\alpha_i\phi(x_i) \tag{4}$$

将公式(3)、(4)带入公式(1)、(2)可得:

$$N\lambda\alpha=K\alpha \tag{5}$$

通过求解公式(5)得到特征向量和特征值, 从而任意样本 x 在特征空间 V 的投影为:

$$(v,\phi(x))=\sum_{i=1}^{N}\alpha_i(\phi(x_i)\cdot\phi(x))=\sum_{i=1}^{N}\alpha_i k(x_i,x) \tag{6}$$

通过 KPCA 可以选择最优的网络入侵检测特征子集, 并对网络入侵样本集进行相应的处理, 得到分类的输入特征.

2.2 样本选择

特征对信噪比(d)的定义为:

$$d=\frac{\mu_1-\mu_2}{\sigma_1+\sigma_2} \tag{7}$$

其中, μ_1和μ_2分别代表特征均值, σ_1和σ_2为该特征标准差[13].

巴氏距离的定义如下:

$$B=\frac{1}{4}\frac{(\mu_1-\mu_1)^2}{(\sigma_1^2+\sigma_2^2)}+\frac{1}{2}\ln\left(\frac{(\sigma_1^2+\sigma_2^2)}{2\sigma_1\sigma_2}\right) \tag{8}$$

由式(8)可知, 巴氏距离既考虑到样本中的均值, 也考虑到样本的方差分布.

样本选择的流程如下:

1) 初始化. 设定所需选择的样本个数为θ, 选择阈值为ε, 初始的样本均值向量为μ_0, 初始样本集为 $S=\{x_i|i=1,2,\ldots,p\}$.

2) 计算样本与μ_0的巴式距离$B_i(i=1,2,\ldots,p)$, 并搜索最小距离B_{min}, 将其所对应的第k个样本记为选中样本x_s. 计算选中样本与其余p-1个样本的欧式距离$B_{ki}(i=1,2,\ldots,p-1)$.

3) 若无任何样本使得$B_{ki}<\varepsilon$, 则退出, 不然重复步骤2).

2.3 网络入侵分类器

对于 N 个任意不相同的样本(x_i,y_i), 其中$x_j=[x_{j1},x_{j2},\ldots,x_{jn}]^T$, $y_i=[y_{i1},y_{i2},\ldots,y_{in}]^T$, , 激励函数为 G(x) 的 ELM 输出可以表示为:

$$f(x_j)=\sum_{i=1}^{L}\beta_i G(a_i\cdot x_j+b_i),a_i\in R^n,\beta_i\in R^m \tag{9}$$

其中, L 为隐神经元数; $\alpha_i=[\alpha_{i1},\alpha_{i2},\ldots,\alpha_{in}]^T$ 是输入到第 i

个隐神经元的输入权重向量; b_i 是第 i 个隐神经元的偏置; $\beta_i=[\beta_{i1}, \beta_{i2},\ldots, \beta_{in}]^T$ 是连接第 i 个隐神经元的输出权值; 激励函数是“Sigmoid”函数[14].

如果 ELM 可以零误差逼近 N 个样本, 则存在 α_i, b_i, β_i 使得下式:

$$f\left(x_j\right)=\sum_{i=1}^{L}\beta_i G\left(a_i\cdot x_j+b_i\right)=y_i,\quad x_j\in R^n, a_i\in R^n, \beta_i\in R^m \tag{10}$$

可以简化为:

$$H\beta=Y \tag{11}$$

式中, H 为隐藏层输出矩阵, 即:

$$H=\begin{bmatrix} G\left(a_1\cdot x_1+b_1\right)\cdots G\left(a_L\cdot x_1+b_L\right)\\ \vdots \quad\cdots\quad \vdots\\ G\left(a_1\cdot x_N+b_1\right)\cdots G\left(a_L\cdot x_N+b_L\right)\end{bmatrix}_{N\times L} \tag{12}$$

由此, 输出层参数 β 可以根据极小 2-范数最小二乘解得:

$$\beta=H^{+}Y \tag{13}$$

式中, H^{+}为 H 的 Moore-Penrose 广义逆.

3 仿真实验

3.1 数据来源

数据来自 KDD Cup99 数据集,每条记录中包含 41 个特征, 网络状态其 5 种: normal, Probe, DOS, U2R, R2L, 特征具体如表 1 所示[15].

表 1　41 种网络的特征

编号	特征名	编号	特征名
1	duration	13	num_compromised
2	protocol type	14	root_shell
3	service	15	su_attempted
4	src_byte	16	num_root
5	dst_byte	17	num_file_creations
6	flag	18	num_shells
7	land	19	num_access_shells
8	wrong_fragment	20	num_outbound_cmds
9	urgent	21	is_hot_login
10	hot	…	…
11	num_failed_logins	…	…
12	logged_in	41	dst_host_srv_rerror_rate

由于 KDD Cup99 数据集规模比较大, 为此, 选择部分数据集进行仿真实验, 它们具体如表 2 所示.

表 2　仿真实验的数据样本

状态	训练样本的数目	测试样本的数目
DOS	800	200
U2R	200	50
R2L	100	20
Probe	50	18
Normal	1000	200

在 SVM 训练过程中, 对特征值进行归一化处理, 具体为:

$$x_i^{'}=\frac{x_i}{\max(x_i)} \tag{14}$$

式中, x_i 和 $x_i^{'}$ 分别为原始和归一化后的特征值, $\max(x_i)$为特征 x_i 的最大值.

3.2 结果与分析

为了使KPCA-ELM的结果更具说服力, 选择主成分分析+ELM 的入侵检测模型(PCA-ELM)、核主成分分析+支持向量机的入侵检测模型(KPCA-SVM)、核主成分分析+BP 神经网络的入侵检测模型(KPCA-BPNN)进行对比实验, 所有入侵检测模型均进行 10 次仿真实验, 取它们的平均值进行对比.

3.2.1 网络入侵检测性能对比

PCA-ELM 、 KPCA-SVM 、 KPCA-BPNN 和 KPCA-ELM 的仿真结果如图 2~4 所示, 得到如下结论:

1) 相对于 PCA-ELM, KPCA-ELM 提高了网络入侵检测正确率, 降低了误检率和漏检率, 这主要是由于KPCA 较好的克服了 PCA 只能分析线性数据的缺陷, 改善了网络入侵检测效果.

2) 相对于 KPCA-SVM, KPCA-BPNN, KPCA- ELM 的网络入侵检测性能更优, 这主要是由于 KPCA-ELM 采用 ELM 建立网络入侵分类器, 大幅度降低了网络入侵的误检率和漏检率.

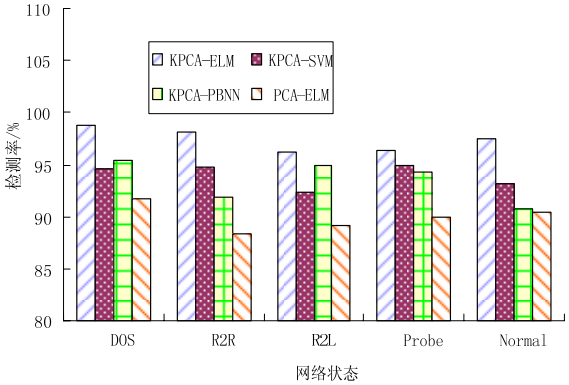

图 2　检测率对比

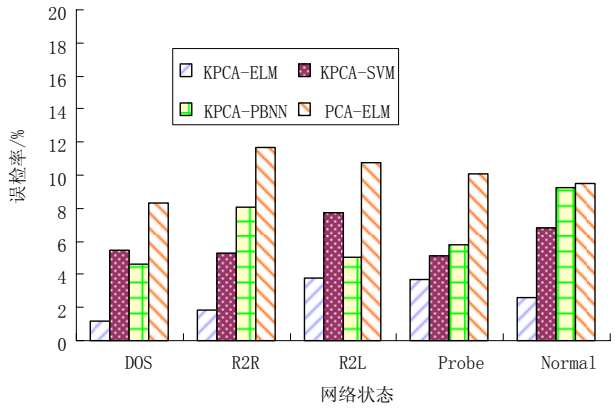

图 3 误检率对比

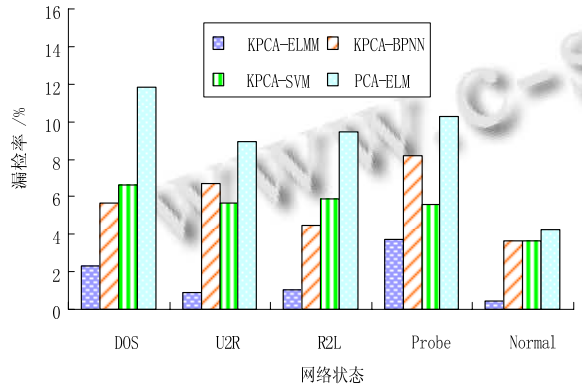

图 4 漏检率对比

3.2.2 执行效率对比

PCA-ELM、KPCA-SVM、KPCA-BPNN 和 KPCA-ELM 的平均执行时间如图 5 所示. 从图 5 可知, 相对于 PCA-ELM、KPCA-SVM、KPCA-BPNN, KPCA-ELM 缩短了网络入侵的平均执行时间, 提高了网络入侵检测效率.

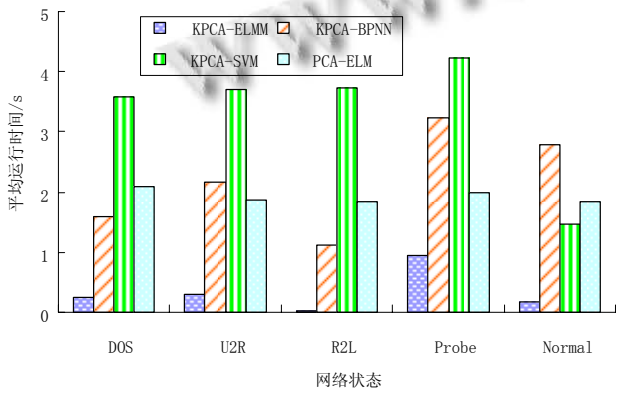

图 5 平均执行时间对比

4 结语

在网络入侵检测建模过程中, 特征和训练样本的选择网络检测效果具有重要的影响, 为此, 提出一种特征选取和样本选择的网络入侵检测模型, 仿真结果表明, KPCA-ELM 提高了网络入侵检测的正确率, 降低了误检率、漏检率, 而且提高了网络入侵检测的效率, 可以较好的保证网络安全.

参考文献

1 唐正军,李建华.入侵检测技术.北京:清华大学出版社,2004.
2 Denning DE. An intrusion detection model. IEEE Trans. on Software Engineering, 2010, 13(2): 222–232.
3 张昊,陶然,李志勇,蔡镇河.基于 KNN 算法及禁忌搜索算法的特征选择方法在入侵检测中的应用研究.电子学报告, 2009,37(7):1628–1630.
4 闫新娟,谭敏生,严亚周,吕明娥.基于隐马尔科夫模型和神经网络的入侵检测研究.计算机应用与软件,2012, 29(2):294–297.
5 龚明朗,许榕生.一种改进的 PSO 算法在网格入侵检测系统中的研究.计算机应用与软件,2011,28(3):274–278.
6 汪厚祥,聂凯,罗志伟.面向入侵检测的基于 IMGA 和 MKSVM 的特征选择算法.计算机科学,2012,39(7): 96–100.
7 曾凡培.粗集神经网络在网络入侵中的应用研究.计算机仿真,2011,28(7):161–164.
8 陈友,沈华伟,李洋.一种高效的面向入侵检测系统的特征选择算法.计算机学报,2007,30(8):1398–1408.
9 彭义春,牛熠,胡琦伟.基于 IRBF 的入侵检测系统的研究. 计算机应用与软件,2013,30(9):187–190.
10 栾庆林,卢辉斌.自适应遗传算法优化神经网络的入侵检测研究.计算机工程与设计,2008,29(12):3022–3024.
11 Xia M, Zhang YC, Weng LG, et al. Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. Knowledge-Based Systems, 2012, 36: 253–259.
12 高海华,杨辉华,王行愚.基于 PCA 和 KPCA 特征抽取的 SVM 网络入侵检测方法.华东理工大学学报(自然科学版), 2006,32(3):321–326.
13 张晓惠,林柏钢.基于特征选择和多分类支持向量机的异常检测.通信学报,2009,30(10A):68–73.
14 Huang GB, Zhou HM, Ding XJ, et al. Extreme learning machine for regression and multiclass classification. IEEE Trans. on Systems Man and Cybernetics, Part B: Cybernetics, 2012, 42(2): 513–529.
15 张新有,曾华燊,贾磊.入侵检测数据集 KDD CUP99 研究. 计算机工程与设计,2010,31(22):4809–4816.