基于 GA 和组合核的 SVM 入侵检测算法

陈桂林,王生光,徐静妹,李 雷
(南京邮电大学,江苏 南京 210023)

摘 要:SVM 有着很强的学习能力,已经成为入侵检测的重要算法之一。由于入侵检测原始数据量大,且具有高维性、冗余性等特点,导致传统 SVM 入侵检测算法计算量大、预测时间长。基于此,文中提出一种改进的 SVM 入侵检测算法(KPCA-GA-LC-SVM)。文中利用核主成分分析法(KPCA)进行数据的特征提取,降低数据维数和计算量;使用两个核函数线性加权结合形成的组合核函数代替传统的单一核函数,并通过遗传算法(GA)进行 SVM 核参数及组合核权系数的寻优,来提高 SVM 性能。实验结果表明,文中算法有效地提高了入侵检测的检测精度。

关键词:入侵检测;核主成分分析法;支持向量机;遗传算法
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2015)02-0148-04
doi:10.3969/j.issn.1673-629X.2015.02.034

Intrusion Detection Algorithm of SVM Based on GA and Composed Kernel Function

CHEN Gui-lin, WANG Sheng-guang, XU Jing-mei, LI Lei
(Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: SVM has a strong learning ability, has become one of the most important intrusion detection algorithm. Due to a large amount of raw data in intrusion detection, and with a high dimension, redundancy, etc., result in larger of calculating the volume and the longer of predicted time in the traditional SVM intrusion detection algorithm. Based on this, propose an improved SVM intrusion detection algorithm (KPCA-GA-LC-SVM). In this paper, use Kernel Principal Component Analysis (KPCA) for data feature extraction and reduce the dimensionality of data and computation. Use a combination of kernel functions formed by weighted linear combination of two kernel function instead of the traditional single kernel function, and through genetic algorithm to find the optimization of kernel parameters and the weights of the composed kernel function to improve the performance of SVM. The experimental results show that the improved algorithm can effectively improve the accuracy of intrusion detection.

Key words: IDS; KPCA; SVM; GA

0 引 言

入侵检测系统(Intrusion Detection System, IDS)作为一种主动防御工具,已成为信息安全研究的重要内容[1]。目前主要有基于特征选择、基于贝叶斯推理、基于贝叶斯网络、基于贝叶斯聚类、基于模式预测和基于机器学习的六种异常检测技术[2]。

入侵检测本质上是个分类问题[3]。目前为止,各种技术诸如数据挖掘、机器学习、数据融合和神经网络等[4-5]都在入侵检测工作中取得了可观的成果,而支持向量机(Support Vector Machine, SVM)作为一种新兴的机器学习方法也取得了一定的成绩。传统 SVM 在入侵检测中的应用[6-10],虽说在检测精度以及效率上有所提高,但是入侵检测数据量大、维数多,导致核矩阵的计算量大,同时一些次要因素会影响其最优分类面的选择。因此,传统的 SVM 还存在着应用上的缺陷,需要一种方法来降低维数和缩减数据冗余信息[11];同时 SVM 的分类性能与选取的核函数及核函数参数有着直接关系,因此找到相对最优的核函数[12-13]及参数对提高入侵检测精度来说是至关重要的。

文中先利用核主成分分析法(Kernel Principle Component Analysis, KPCA)对原始数据进行特征提

收稿日期:2014-03-17 **修回日期**:2014-06-23 **网络出版时间**:2014-12-27
基金项目:国家自然科学基金资助项目(61070234,61071167,61373137);国家大学生创新创业训练计划项目(SZDG2013032)
作者简介:陈桂林(1992-),男,研究方向为核方法、机器学习、信息安全;李 雷,教授,博士生导师,研究方向为核方法、机器学习、模糊数理论和智能控制等。
网络出版地址:http://www.cnki.net/kcms/detail/61.1450.TP.20141227.1347.036.html

取,降低数据维数,然后选取由两个核函数线性加权形成的组合核函数作为 SVM 的核函数,并用遗传算法[14-16]来获取最优核参数及组合核函数权系数,由此来进行入侵检测。实验结果表明,文中的方法能更有效地从样本中发现异常数据,进一步提高检测性能及效率。

1　KPCA-GA-LC-SVM 入侵检测算法

1.1　KPCA 算法特征提取

每条入侵检测数据可抽象成矢量 $\boldsymbol{x}_i = (x_i^1, x_i^2, \cdots, x_i^N)$,其中 x_i^N 为 x_i 的第 N 维数据,KPCA 特征提取步骤如下:

(1)将数据通过非线性变换 $\varphi(x_i)$ 映射到高维特征空间,计算可得核矩阵:经过映射后的训练样本的协方差矩为:

$$\bar{\boldsymbol{K}} = \sum_{i=1}^{M} \varphi(x_i) \cdot \varphi(x_i)^{\mathrm{T}} / M \tag{1}$$

其中, M 为训练样本数。

定义核矩阵 $\boldsymbol{K}$:

$$K_{ij} = [\varphi(x_i) \cdot \varphi(x_j)], i,j = 1,2,\cdots,M \tag{2}$$

(2)数据中心化:如果 $\sum_{i=1}^{M} \varphi(x_i) \neq 0$,则进行中心化处理,用 $\boldsymbol{K}^* = \boldsymbol{K} - \boldsymbol{AK} - \boldsymbol{KA} + \boldsymbol{AKA}$ 替换上述的 $\boldsymbol{K}$,式中 $A_{ij} = \frac{1}{M}$。

(3)计算核矩阵的特征值和特征向量: $\bar{\boldsymbol{K}}$ 的特征值 λ 和特征向量 $\boldsymbol{\nu}$ 满足:

$$\lambda \boldsymbol{\nu} = \bar{\boldsymbol{K}} \boldsymbol{\nu} \tag{3}$$

将式(2)代入式(3)可得:

$$M\lambda \boldsymbol{\alpha} = \boldsymbol{K}\boldsymbol{\alpha} \tag{4}$$

求解式(4)得到大于零的特征值 $\lambda_1, \lambda_2, \cdots, \lambda_p$ 和对应的特征向量 $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_p$,令

$$V^k = \sum_{i=1}^{M} \alpha_i^k \varphi(x_i)$$

(4)提取主成分:计算数据 $\varphi(x_i)$ 在特征向量空间 V^k 上的投影:

$$V^k \cdot \varphi(x) = \sum_{i=1}^{M} \alpha_i^k \varphi(x_i) \varphi(x)$$

其中, $\varphi(x_i) \cdot \varphi(x)$ 可以利用核函数技巧计算,核函数是使得 $K_{ij} = \varphi(x_i) \cdot \varphi(x_j)$ 成立的一类函数。令

$$g_k(x) = V^k \cdot \varphi(x) = \sum_{i}^{M} \alpha_i^k K(x_i, x)$$

其中, K 为核函数; $g_k(x)$ 成为对应于 $\varphi(x)$ 的第 k 个非线性主元分量,将所有投影值作为样本的特征矢量。

$$g_k(x_i) = [g_1(x_i), g_2(x_i), \cdots, g_p(x_i)]$$

只要 $\boldsymbol{K}$ 的前 r 个最大的特征值满足: $\sum_{j=1}^{r} \lambda_j / \sum_{j=1}^{p} \lambda_j \geqslant 85\%$,即可从上面 p 个特征分量中抽取 r 个,从而样本特征矢量从抽取为:

$$\boldsymbol{x}_i' = [g_1(x_i), g_2(x_i), \cdots, g_r(x_i)]$$

1.2　基于组合核函数的 SVM

用经 KPCA 提取的网络数据支持向量机的输入源来训练支持向量机,得到支持向量同时优化向量参数。若将 KPCA 抽取的样本特征送入 SVM 分类器中训练,得到的支持向量记为 $\boldsymbol{x}_i'(i = 1,2,\cdots,N_{\mathrm{SV}})$,则将最后的判别函数写为:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i', x) + b\right) \tag{5}$$

由于 SVM 的性能主要由核函数及核函数的参数决定,故文中对式(5)中的核函数 K 进行优化,即将两个核函数进行线性加权形成组合核函数,文中只对三个常用核函数进行组合。常用的核函数有:

多项式核函数(Poly):

$$K(x, x_i) = [(\boldsymbol{x} \cdot \boldsymbol{x}_i) + 1]^q$$

径向基核函数(RBF):

$$K(x, x_i) = \exp(-\| \boldsymbol{x} - \boldsymbol{x}_i \|^2 / \sigma^2)$$

Sigmoid 核函数(Sigmoid):

$$K(x, x_i) = \tanh[\upsilon(\boldsymbol{x} \cdot \boldsymbol{x}_i) + c]$$

形成的组合核函数为:

Poly+RBF:

$$K(x, x_i) = \alpha [(\boldsymbol{x} \cdot \boldsymbol{x}_i) + l]^q + (1 - \alpha) \exp(-\| \boldsymbol{x} - \boldsymbol{x}_i \|^2 / \sigma^2)$$

Poly+Sigmoid:

$$K(x, x_i) = \alpha [(\boldsymbol{x} \cdot \boldsymbol{x}_i) + l]^q + (1 - \alpha) \tanh[\upsilon(\boldsymbol{x} \cdot \boldsymbol{x}_i) + c]$$

RBF+Sigmoid:

$$K(x, x_i) = \alpha \cdot \exp(-\| \boldsymbol{x} - \boldsymbol{x}_i \|^2 / \sigma^2) + (1 - \alpha) \tanh[\upsilon(\boldsymbol{x} \cdot \boldsymbol{x}_i) + c]$$

式中,权系数 $\alpha (0 \leqslant \alpha \leqslant 1)$ 的作用是调节两种核函数的共同作用。

1.3　GA 对核参数及组合核权系数优化

通过对 SVM 预测原理分析可知,当核函数确定后,其预测性能主要由核函数参数、惩罚因子 C 以及加权系数 α 决定。如何选择一个最优的参数是当前支持向量机应用过程中的一个难题。文中采用遗传算法来进行参数寻优。遗传算法是一种通过模拟自然进化过程搜索最优解的方法,其优化 SVM 核参数即组合核函数权系数集体过程如下:

(1)对问题的遗传表达:进行编码;

(2)初始化:确定种群的规模和终止准则;随机生成 N 个个体作为初始种群 $X(0)$;置进化代数 $t = 0$;

(3)个体评价:对当前群体 $X(t)$ 中的个体 x_i 用“适应值”评价解得适应程度;

(4)种群进化:改变后代组成的各种遗传算子,包括交叉和变异等,产生下一代种群 $X(t+1)$;

(5)终止检验:如果 $X(t+1)$ 满足终止准则,则输出 $X(t+1)$ 中具有最大适应度的个体作为最优解,终止计算;否则令 $t = t+1$,转步骤(3)。

1.4 KPCA-GA-LC-SVM 模型

文中的 IDS 包含八个模块,如图 1 所示。

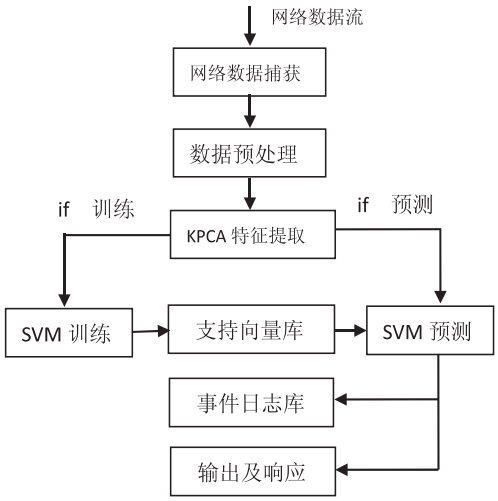

图 1 KPCA-GA-LC-SVM 模型

该模型的处理流程为:

(1)首先由网络数据捕获模块捕获网络数据中的数据流,并提取出每条网络连接的特征信息;

(2)将网络连接的特征信息由数据预处理模块进行数据预处理:将网络连接数据中的字符变量转化为数字变量;得到原始数据的输入向量形式;

(3)对原始输入向量利用 KPCA 进行特征提取,得到 SVM 的输入向量形式;

(4)如果处于 SVM 训练状态,则训练 SVM,并将训练后的结果即若干个支持向量存入 SVM 支持向量库;

(5)如果处于 SVM 预测状态,则由 SVM 预测模块对输入向量进行预测,预测后将结果存入事件日志库,并根据设置执行相应的响应操作。

2 实验及结果分析

2.1 数据来源

文中实验数据来自 KDD CPU99 数据集,该数据集包含正常数据与异常数据。每条数据包含 41 个字段,其中 34 个为连续变量,7 个为符号变量。

由于数据集 KDD CPU99 过于庞大,随机选取了其中 54 个具有代表性的数据当作训练集,随机抽样选取 12 175 个数据作为测试集。

2.2 数据预处理

实验在 Matlab7.0 和 SVM 工具箱的平台下实现。通过数据预处理将网络连接记录中的字符变量转化为数字变量,所有字符属性对应一张数值字典。对字符变量进行转化之后,再将数据每条记录一行的格式存储在 *.txt 文件中,供实验过程中使用。

数据预处理后,利用 KPCA 提取特征分量,此时 KPCA 的核函数选取为 RBF 核函数。计算样本核矩阵各特征值的贡献比可知,其最大的五个特征值对应的贡献比为 87.682% > 85%,所以只选取前面的 5 个特征分量,从而原始数据可以从 41 维降低到 5 维。

经 KPCA 特征提取后得到的特征分量,为了防止不同的度量标准造成数值较小的特征被掩盖,文中利用 SVM 工具箱再进行归一化处理,对每一个属性进行标准化处理,使得到的支持向量机的输入源的各维数值特征统一到[0,1]范围内。

2.3 结果分析

实验使用检测精度来定量描述入侵检测系统的性能,定义如下:

$$\text{检测精度} = \frac{\text{正确分类的样本数}}{\text{总样本数}} \times 100\%$$

文中先采用 KPCA 进行特征提取,然后选取由两个核函数线性加权形成的组合核函数作为 SVM 的核函数,并用遗传算法来获取最优核参数及组合核函数权系数,以此来进行入侵检测。为了验证文中算法的优越性,做了一系列的对比实验,其中惩罚因子 C 均取无穷大。基于单一核函数的实验结果如表 1 和表 2 所示。

表 1 SVM 算法检测精度 %

检测算法	SVM	
	无 GA	GA
Poly	94.98	97.43
RBF	89.33	94.76
Sigmoid	87.93	95.54

由表 1 实验结果可知,SVM 算法在入侵检测上的直接应用已有较高的检测精度,而使用遗传算法来进行核函数参数的寻优之后,对于三个常用核函数来说,SVM 的检测精度都有了不同程度的提高,说明遗传算法可优化 SVM 的性能。

表 2 KPCA-SVM 算法检测精度 %

检测算法	KPCA-SVM	
	无 GA	GA
Poly	98.29	98.36
RBF	92.98	95.35
Sigmoid	92.91	96.21

由表 2 的实验结果与表 1 对比可知,使用 KPCA

的 KPCA-SVM 入侵检测算法可有效提高 SVM 的检测精度,使 SVM 性能得到改善;通过 GA 来优化参数的 KPCA-GA-SVM 入侵检测算法相对 KPCA-SVM 算法,其检测精度也有所提高。

综合表 1 和表 2 的实验结果,分析可得,对于 SVM 使用单一核函数时,KPCA 算法和遗传算法均可改善其性能,提高入侵检测精度,而且两者结合同时使用的 KPCA-GA-SVM 算法可进一步提高入侵检测精度。

由于核函数的选取对 SVM 的性能影响很大,故文中选用组合核函数来优化 SVM 算法,基于组合核函数的实验结果如表 3 和表 4 所示。

表 3　LC-SVM 算法检测精度　%

检测算法	LC-SVM	
	无 GA	GA
Poly+RBF	98.76	98.96
Poly+Sigmoid	96.52	97.35
RBF+Sigmoid	98.50	98.71

从表 3 的实验结果和表 1 对比可以看出,使用组合核函数比使用单一核函数的入侵检测精度高,可见组合核函数更符合数据的特性;通过遗传算法对核参数及组合核函数权系数进行寻优的 GA-LC-SVM 算法,入侵检测精度有了进一步的提高,再次验证了 GA 算法和 SVM 结合的优越性。

表 4　KPCA-LC-SVM 算法检测精度　%

检测算法	KPCA-LC-SVM	
	无 GA	GA
Poly+RBF	99.04	99.25
Poly+Sigmoid	97.32	98.65
RBF+Sigmoid	98.73	99.42

从表 4 的实验结果和表 3 对比可知,使用 KPCA 特征提取同时选用组合核函数的 KPCA-LC-SVM 算法,入侵检测精度相对 LC-SVM 算法有所提高。综合表 2 的结果可知,不管是使用单一核函数还是组合核函数的 SVM 算法,KPCA 都能有效地提高其检测精度;在 KPCA-LC-SVM 的基础上,利用遗传算法对核参数及组合核函数权系数寻优的 KPCA-GA-LC-SVM 算法,也就是文中的方法,其入侵检测精度有了进一步的提高,最高可达 99.42%。可见文中方法有效地结合了 KPCA、GA 以及组合核函数的优势,较大地提高了入侵检测精度,具有强大的优势和应用前景。

3　结束语

文中提出一种基于 GA 和组合核函数的 SVM 入侵检测模型。大量实验结果表明,利用 KPCA 进行特征提取,可有效降低数据的维数及冗余,降低次要因素对入侵检测结果的影响,同时文中 SVM 内部使用线性加权形成的组合核函数,并通过遗传算法对核函数参数及组合核函数权系数寻优的方法,可有效提高 SVM 的性能,改善入侵检测精度,是入侵检测算法改进的一种有效尝试。

参考文献:

[1] 鲜永菊. 入侵检测[M]. 西安:西安电子科技大学出版社,2009.

[2] 蒋建春,马恒太,任党恩,等. 网络安全入侵检测:研究综述[J]. 软件学报,2000,11(11):1460-1466.

[3] 杨义先,钮心忻. 入侵检测理论与技术[M]. 北京:高等教育出版社,2006.

[4] Panda M, Abraham A, Patra M R. A hybrid intelligent approach or network intrusion detection[J]. Procedia Engineering,2012,30:1-9.

[5] Yusufovna S F. Integrating intrusion detection system and data mining[C]//Proc of international symposium on ubiquitous multimedia computing. Hobart:IEEE,2008:256-259.

[6] Yi Yang, Wu Jiansheng, Xu Wei. Incremental SVM based on reserved set for network intrusion detection[J]. Expert Systems with Applications,2011,38(6):7698-7707.

[7] Scherer P, Vicher M, Drazdilova P, et al. Using SVM and clustering algorithms in IDS systems[C]//Proc of Dateso'11. [s. l.]:[s. n.],2011:108-119.

[8] Mohammed M N, Sulaiman N. Intrusion detection system based on SVM for WLAN[J]. Procedia Technology,2012,1:313-317.

[9] 柏海滨,李　俊. 基于支持向量机的入侵检测系统的研究[J]. 计算机技术与发展,2008,18(4):137-139.

[10] 张　琨,曹宏鑫,严　悍,等. 支持向量机在网络异常入侵检测中的应用[J]. 计算机应用研究,2006,23(5):98-100.

[11] 陈　武,梁　刚,杨　进. 一种改进的 SVM 算法在入侵检测中的应用[J]. 计算机安全,2013(6):2-7.

[12] 颜根廷,马广富,肖余之. 一种混合核函数支持向量机算法[J]. 哈尔滨工业大学学报,2007,39(11):1704-1706.

[13] 杨冬云,李数函. 支持向量机核函数的构造方法研究与分[J]. 高师理科学刊,2010,30(2):23-26.

[14] 李　锋,汤宝平,刘文艺. 遗传算法优化最小二乘支持向量机的故障诊断[J]. 重庆大学学报:自然科学版,2010,33(12):14-20.

[15] 孙　浩,陶　亮. 基于佳点集遗传算法的支持向量机的参数选择[J]. 计算机技术与发展,2009,19(8):86-88.

[16] 汪世义. 基于优化支持向量机的网络入侵检测技术研究[J]. 计算机技术与发展,2009,19(7):177-179.

基于GA和组合核的SVM入侵检测算法

作者： 陈桂林， 王生光， 徐静妹， 李雷， CHEN Gui-lin， WANG Sheng-guang， XU Jing-mei ， LI Lei

作者单位： 南京邮电大学,江苏 南京,210023

刊名： 计算机技术与发展 ISTIC

英文刊名： Computer Technology and Development

年，卷(期)： 2015(2)