DOI:10.13718/j.cnki.xsxb.2020.01.010

基于支持向量机方法的网络入侵检测实验研究①

周 飞 菲

郑州升达经贸管理学院 信息工程学院,郑州 451191

摘要:网络信息不断增加和攻击手段日益复杂,给网络安全领域带来了日益严峻的挑战.为了改善网络入侵检测技术现状,提出了一种基于支持向量机和决策集合理论融合的网络入侵检测方法,通过对规则信息、攻击信息、边界信息的准确界定完成检测过程.选取了基于神经网络的入侵检测方法、基于遗传算法的入侵检测方法、基于传统支持向量机的入侵检测方法作为对比算法,在 K-Cup 测试数据集下展开实验研究.实验结果表明,该文提出的方法具有更高的召回率、精确率、查准率和更低的误检率,其性能明显优于其他 3 种方法,可应用于入侵检测领域.

关 键 词:网络入侵检测;支持向量机;攻击信息;召回率;精确率

中图分类号:TP393 **文献标志码**:A **文章编号**:1000-5471(2020)01-0057-05

近 20 年间,互联网技术的飞速发展彻底颠覆了传统生活方式和经济模式,成为人们日常生活和工作中不可或缺的重要组成部分[1-2].通过网络入侵获取企业经营数据和商业机密成为一种重要的犯罪模式[3].如何有效应对网络入侵检测,保护网络安全,已经成为互联网技术领域亟待解决的重要课题[4].所谓入侵检测,就是对试图破坏网络访问规则、试图中断网络连接、试图盗取网络信息的相关攻击行为的检测[5].在抵御入侵检测的早期阶段,信息加密、防火墙、登录认证等方法比较常用[6-7].入侵检测就变成了对规则信息和攻击信息的区分检测,基于机器学习和智能识别的方法广泛应用于入侵检测[8].本文提出一种基于支持向量机的检测方法,以期更加准确地对入侵行为进行判断.

1 理论基础

基于决策树的入侵检测方法,就是将规则信息和攻击信息分别形成信息库,并根据决策树的构建和识别方法形成对 2 种信息的判断,进而有效地检测出入侵信息[9].基于神经网络的入侵检测方法,是将先验数据代入网络中进行训练学习,网络稳定后确定网络参数指标,再对新的数据进行判断确定是否为入侵行为[10].基于遗传算法的入侵检测方法,是根据先验数据计算出父代和子代之间的遗传关系,进而判断新数据是否具有父代入侵行为的相关特征,从而得出是否为入侵行为的结论[11].基于聚类分析的入侵检测方法,是将规则信息和入侵信息分别聚类,根据新信息距离聚类中心的远近判断其是否为攻击行为[12].

支持向量机方法是一类典型的机器学习方法,在学习分类过程中又充分依赖统计学原理和风险最小化原理.支持向量机的方法,可以将要分析的问题按照向量机进行分类,形成明确的 2 个集合.这对于网络入侵检测而言是非常具有针对性的,可以把各种网络访问行为明确地区分为正常行为还是攻击行为.

假设存在一个高维度的特征空间,同时可以在这个空间中构造出一个完成学习过程的函数.借助这个函数,可以不断地进行偏差学习和优化,进而完成整个训练过程.支持向量机方法的突出优点在于:其检测效果直接对应于样本数据,无论是高维度的还是非线性的,检测效果都具有较好的一致性;泛化学习能力强,无论是维度上灾难问题还是过度学习问题,都能被支持向量机较好地解决.

① 收稿日期:2019-04-02
基金项目:2018 年度河南省科技攻关重点研发与推广项目(182102210139).
作者简介:周飞菲(1980-),女,硕士,副教授,主要从事计算机网络、大数据及人工智能研究.

从实现原理上看，支持向量机方法的实现过程就是一个分类过程. 为了区分两类不同性质的样本，它需要构建一个最优超平面，这个平面如果能够保证两类样本之间的距离间隔最大，就达到了最佳分类效果.

在具体操作过程中，如果被分类的样本是线性的，就采用线性分类核函数. 如果被分类的样本是非线性的，就采用非线性分类核函数.

如果在当前维度空间上，无法找到一个最优超平面，就延伸到更高维度的空间上去寻找.

对于入侵检测问题而言，其规则信息和攻击信息的区分是一个典型的二分类问题，与支持向量机的方法具有很好契合性.

从数学意义上讲，支持向量机方法在解决二分类问题时，对一个数据集进行处理，此数据集 $L=\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$，其中 x_i 用于表示输入向量，y_i 用于表示输出向量，这时支持向量机的分类就演化为二次优化问题，为

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N}\alpha_i$$

$$s.t. \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{1}$$

其中，参数 α_i 表示对输入向量 x_i 进行训练时的拉格朗日因子，参数 $K()$ 用于表示支持向量机中的核函数，参数 C 用于表示进行分类过程中的惩罚因子.

根据公式(1)的优化处理，可以进一步获得决策函数，为：

$$D(x) = sign\Big[\sum_{i=1}^{N}\alpha_i^* y_i K(x_i, x) + b^*\Big] \tag{2}$$

其中，参数 $D(x)$ 表示的就是分类结果，如果它的值是1，就表示规则信息，如果它的值是-1，就表示攻击信息.

2 基于支持向量机的入侵检测方法设计

为了使支持向量机方法更加适合于入侵检测，本文进一步引入决策集合理论，构建了一种具体的入侵检测方法，处理过程如图1所示.

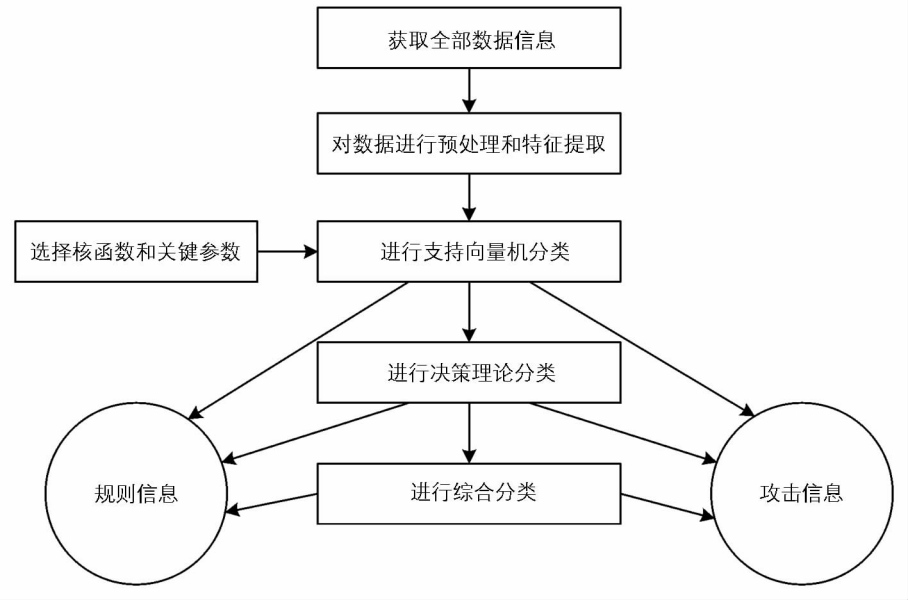

图1 本文设计的基于支持向量机和决策集合理论的入侵检测方法框架

根据图1框架，该方法的具体实现步骤如下：

1) 根据支持向量机方法，计算样本数据对(x_i, y_i)到超平面(w^*, b^*)之间的距离大小，这个距离函数

表达为 $r_i = y_i(w^* x_i + b^*)$. 如果这个距离大小满足 $r_i \geqslant \frac{2}{\| w \|}$，表明待检测信息是规则信息；如果这个距离大小满足 $r_i < \frac{2}{\| w \|}$，将待检测信息纳入不确定集合.

2) 对于不确定集合 Q 中的数据，按照决策集合的理论进一步分类. 这个过程首先需要将集合 Q 中数据所对应的各种属性特征和决策值都看作决策信息表，之后根据属性特征对数据进行等价类确定 $[x]$，再在这个等价类中确定其归属概率 $P(X[x])$. 根据预先设定好的损失函数 λ 以及两个域值 α 和 β，如果 $P(X[x]) \geqslant \alpha$，那么 x 就是规则信息；如果 $\beta < P(X[x]) < \alpha$，那么 x 就是规则信息和攻击信息的边界信息；如果 $P(X[x]) \leqslant \beta$，那么 x 就是攻击信息.

3) 此时仅剩下边界信息集合 N 的数据尚未确定其分类，进一步采用归一化处理，进而计算 $m = \frac{wr_i + (1-w)P(X[x])}{2}$，然后将 m 和实现设置的域值 n 进行对比. 如果 $m \geqslant n$，那么 x 就是规则信息；反之，x 就是攻击信息.

3　实验结果与分析

3.1　实验条件设置

为了验证本文提出的入侵检测方法的效果，在接下来的工作中将展开验证性实验. 实验中，计算机的配置：CPU 为英特尔 i5-6500，其主频大小 3.20 GHz，内存大小 32 GB，操作系统为 windows 8.0，硬盘大小 500 GB，编译平台选择 Matlab.

入侵检测的数据样本，选择网络安全领域比较常用的 K-Cup 测试数据集. 在这个数据集中，每个信息都带有标记，以表明其是攻击信息还是规则信息. 同时，这个数据集中的攻击信息还分为 4 个类型，分别是 DOS 型攻击信息，其意义为拒绝服务攻击信息；R2L 型攻击信息，其意义为未对远程主机进行访问授权的攻击信息；U2R 型攻击信息，其意义为未对本地主机进行访问授权的攻击信息；PROBING 型攻击信息，其意义为不断对主机端口进行监听和扫描处理的攻击信息.

3.2　实验结果分析

为了便于对入侵检测结果的量化分析，首先来设置 4 个参数，分别是参数 TP、参数 TN、参数 FP、参数 FN. 这里，参数 TP 代表的是实际结果和检测结果都是攻击信息，参数 TN 代表的是实际结果和检测结果都是规则信息，参数 FP 代表的是实际结果是规则信息、检测结果是攻击信息，参数 FN 代表的是实际结果是攻击信息、检测结果是规则信息.

根据上述 4 个参数，可以进一步得到 4 个具体的入侵检测性能指标：指标 DR，用于表示召回率；指标 FAR，用于表示误检率；指标 ACC，用于表示检测精确率；指标 PR，用于表示查准率.

4 个指标和 4 个参数之间的对应关系分别如下：

召回率指标的计算：$DR = TP/(TP+FN)$

误检率指标的计算：$FAR = FP/(FP+TN)$

检测精确率指标的计算：$ACC = (TP+TN)/(TP+TN+FP+FN)$

查准率指标的计算：$PR = TP/(TP+FP)$

选择基于神经网络的入侵检测方法、基于遗传算法的入侵检测方法、基于传统支持向量机的入侵检测方法作为本文方法的对比方法，对 4.1 节中的样本数据进行入侵检测实验，实验中得到 4 个参数和 4 个指标的结果表 1、表 2 所示.

表 1　4 种方法进行入侵检测实验得到的 4 个参数对比　　bit

	TP	*TN*	*FP*	*FN*
基于神经网络的入侵检测方法(BP)	242 283	28 537	13 941	15 239
基于遗传算法的入侵检测方法(GA)	232 981	34 820	15 928	16 271
基于传统支持向量机的入侵检测方法(SVM)	227 653	35 511	16 292	20 544
本文提出的入侵检测方法(OURS)	263 108	27 835	4 531	4 526

表2 4种方法进行入侵检测实验得到的4个指标对比 %

	DR	*FAR*	*ACC*	*PR*
基于神经网络的入侵检测方法(BP)	94.082 4	32.819 3	90.273 3	94.559 1
基于遗传算法的入侵检测方法(GA)	93.472 1	31.386 5	89.267	93.600 9
基于传统支持向量机的入侵检测方法(SVM)	91.722 7	31.449 9	87.721 3	93.321 4
本文提出的入侵检测方法(OURS)	98.308 9	13.999 3	96.981	98.307

从表2中的数据结果可以看出：

1) 对于召回率指标 *DR*，本文提出的入侵检测方法达到了98.31%，高出排名第2的基于神经网络入侵检测方法(BP)近5个百分点.

2) 对于误检率指标 *FAR*，本文提出的入侵检测方法最低，仅为13.99%，而其余3种方法的误检率都超过了30%.

3) 对于检测精确率指标 *ACC*，本文提出的入侵检测方法达到了96.98%，高出排名第2的基于神经网络的入侵检测方法(BP)近7个百分点.

4) 对于查准率指标 *PR*，本文提出的入侵检测方法达到98.31%，而其余3种方法的查准率指标都不超过95%.

上述4组指标可以明显看出，本文提出的入侵检测方法明显优于其它3种方法，从而证实了其有效性. 表2中数据的可视化结果如图2所示.

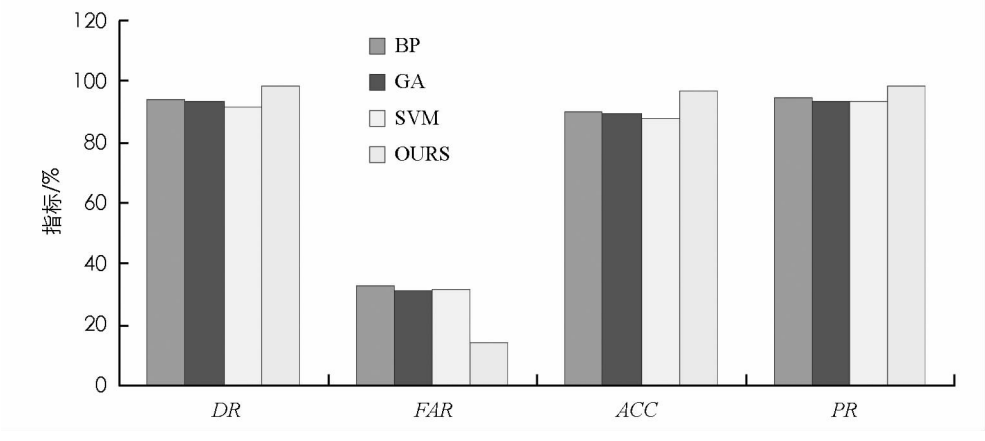

图2 表2中数据的柱状图显示效果

4 结 语

支持向量机方法是一种典型的二分类方法，这对于入侵检测中规则信息和攻击信息的明确划分具有良好的对应性. 基于这种考虑，本文将支持向量机和决策集合理论结合起来，构建了一种新的网络入侵检测方法，并给出了详细的算法流程和实验验证过程. 在实验过程中，选择了网络安全领域常见的K-Cup测试数据集，并以基于神经网络的入侵检测方法、基于遗传算法的入侵检测方法、基于传统支持向量机的入侵检测方法作为对比算法. 实验结果表明，本文提出的方法在召回率、误检率、精确率、查准率等方面均优于其它3种方法，适合应用于网络安全领域的入侵检测.

参考文献：

[1] SEDJELMACI H, SENOUCI S M, ABU-RGHEFF M A. An Efficient and Lightweight Intrusion Detection Mechanism for Service-Oriented Vehicular Networks [J]. IEEE Internet of Things Journal, 2014, 1(6): 570-577.

[2] 海小娟. 计算机网络安全入侵检测系统的设计与应用研究 [J]. 自动化与仪器仪表, 2017(10): 142-143.

[3] 刘胜珍, 李田英. 网络信息安全保护技术——评《网络安全中的信息隐藏与入侵检测技术探究》[J]. 高教发展与评估, 2017, 33(3): 133.

[4] AL-YASEEN W L, OTHMAN Z A, NAZRI M Z A. Multi-Level Hybrid Support Vector Machine and Extreme Learning Machine Based on Modified K-Means for Intrusion Detection System [J]. Expert Systems with Applications, 2017,

67：296-303.
[5] 涂宇飞，金柏杉. 机场入侵检测与计算机网络安全问题的探讨 [J]. 网络安全技术与应用，2017(11)：143-144.
[6] LIU X X，ZHU P D，ZHANG Y，et al. A Collaborative Intrusion Detection Mechanism Against False Data Injection Attack in Advanced Metering Infrastructure [J]. IEEE Transactions on Smart Grid，2015，6(5)：2435-2443.
[7] 陈泽强. 防火墙与入侵检测技术联动的急救网络安全探讨 [J]. 无线互联科技，2018(6)：39-40.
[8] HODO E，BELLEKENS X，HAMILTON A，et al. Threat Analysis of IoT Networks Using Artificial Neural Network Intrusion Detection System [J]. Tetrahedron Letters，2017，42(39)：6865-6867.
[9] 江　峰，王春平，曾惠芬. 基于相对决策熵的决策树算法及其在入侵检测中的应用 [J]. 计算机科学，2012，39(4)：223-226.
[10] 罗俊松. 基于神经网络的 BP 算法研究及在网络入侵检测中的应用 [J]. 现代电子技术，2017，40(11)：91-94.
[11] BOSTANI H，SHEIKHAN M. Modification of Supervised OPF-based Intrusion Detection Systems Using Unsupervised Learning and Social Network Concept [J]. Pattern Recognition，2017，62：56-72.
[12] LEU F Y，TSAI K L，HSIAO Y T，et al. An Internal Intrusion Detection and Protection System by Using Data Mining and Forensic Techniques [J]. IEEE Systems Journal，2017，11(2)：427-438.

Support Vector Machine Method for Network Intrusion Detection

ZHOU Fei-fei

College of Information Engineering，Zhengzhou Shengda University Of Economics，Business and Management，Zhengzhou 451191，China

Abstract：With the increasing of network information and the increasing complexity of attack means，the network security field is facing more and more severe challenges. In order to improve the current situation of network intrusion detection technology，a method of network intrusion detection based on the fusion of support vector machine and decision set theory has been proposed. The detection process is completed by defining the rule information，attack information and boundary information accurately. The intrusion detection methods based on neural network，genetic algorithm and traditional support vector machine have been selected as comparison algorithms，and experiments been carried out under K-Cup test data set. The experimental results show that the proposed method has higher recall rate，accuracy rate，precision rate and lower false detection rate，and its performance is obviously superior to the other three methods，which can be applied in the field of intrusion detection.

Key words：network intrusion detection；support vector machine；attack information；recall rate；accuracy rate

责任编辑　夏　娟