深度置信网络结合递归特征添加的网络入侵检测方法

赵 荷[1] 盖 玲[2]

[1](成都东软学院计算机科学与工程系 四川 成都 611844)

[2](上海大学管理学院 上海 200444)

摘 要 针对互联网零日攻击严重威胁网络安全的问题,提出一种深度置信网络(Deep Belief Network,DBN)结合递归特征添加(Recursive Feature Addition,RFA)的网络入侵检测方法。采用深度置信网络对网络入侵特征进行提取,并基于二元组编码技术将长字符串的特征转化为二进制编码;使用递归特征添加方法对影响网络入侵检测性能的主要特征进行选择。为获取更好的入侵检测性能,提出综合考虑检测准确率、检出率和误报率的入侵检测性能评估函数,从而有效改善抵御互联网攻击的能力。实验结果表明,相较于 K-最邻近(K-Nearest neighbor,KNN)算法等传统的入侵检测算法,该算法的检测准确率提升 8% 以上,保证了互联网的安全性。

关键词 深度学习 递归特征添加 网络入侵检测 互联网攻击 网络安全 零日攻击

中图分类号 TP393 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2020.11.049

NETWORK INTRUSION DETECTION METHOD BASED ON DEEP BELIEF NETWORK AND RECURSIVE FEATURE ADDITION

Zhao He[1] Gai Ling[2]

[1](*Department of Computer Science and Engineering, Chengdu Neusoft University, Chengdu 611844, Sichuan, China*)

[2](*School of Management, Shanghai University, Shanghai 200444, China*)

Abstract To solve the problem that zero-day attacks on the Internet which seriously threaten the network security, we propose a network intrusion detection method based on deep belief network(DBN) and recursive feature addition(RFA). The DBN was used to extract the network intrusion features, and the binary encoding technology was used to transform the long-string-features into binary encodings. Then, the RFA was used to select the main features affecting network intrusion detection performance. To obtain better intrusion detection performance, we proposed an intrusion detection performance evaluation function considering detection accuracy, detection rate and false alarm rate, which effectively improve the ability to resist the Internet attacks. The experimental results show that compared with traditional intrusion detection algorithms such as K-nearest neighbor(KNN) algorithm, our algorithm is improved by more than 8%, thus ensuring the security of the Internet.

Keywords Deep learning Recursive feature addition Network intrusion detection Internet attacks Network security Zero-day attacks

0 引 言

在高速发展的信息时代中,网络安全是一个关键性的问题。网络入侵检测系统(Network Intrusion Detection Systems,NIDS)是解决网络安全问题的方案之一。针对不断出现的网络攻击,所有互联网参与者必须考虑构建安全可靠的防御系统。数据挖掘是一种可以与入侵检测一起使用的技术,在描述系统和用户行为的数据特征中用于检测特征模式,以及理想的恶意活动示例[1-3]。由于互联网每天都会产生称为“零日攻击”的新的攻击方式,且事先没有供应商发现或

收稿日期:2019-05-30。四川省科技厅重大科技专项项目(18ZDZX0078)。**赵荷**,副教授,主研领域:无线网络,计算机网络,网络安全。**盖玲**,副教授。

开发出有效的解决方案以应对该威胁。因此,传统的防御手段难以减轻零日攻击带来的损害,需要在零日攻击对网络造成巨大破坏之前抵御这些零日攻击。显然,研究网络入侵检测系统具有很好的现实意义和实用价值[4-5]。

国内外许多专家及学者围绕网络入侵检测系统进行了深入的研究。文献[6]采用类内相关系数和类间相关系数来获得特征的类特定子集,使用类内和类内相关系数分别测量特征的有效性和可靠性,但该方法并没有处理数据稀缺和相互依赖的特征。文献[7]使用基于签名的异常检测方案来检查包头,更准确地提取行为模式。将基于特征的检测系统和基于异常的检测系统相结合,克服前者因无法检测新型攻击而受到的限制以及由此产生的高误报率。但该方法没有考虑使用特征选择去除不相关和冗余的特征。文献[8]设计了一种新的多目标优化方法用于高效的入侵检测。该方法涉及编码提供最佳特征子集的染色体。这些特征可以在以后用于训练组合分类器的各种实例。然而,该模型的缺点是,在计算不同代中的适应度函数时计算成本非常高。文献[9]提出了一种将流量记录作为图像来处理,以使用计算机视觉技术检测拒绝服务攻击的方法。该方法涉及利用多变量相关分析来描述网络流量记录并将其转换为图像,但没有考虑图像可能有一些噪声来自不同的来源,这反过来会产生噪声特征,那些嘈杂的特征可能导致不符合需要的分类结果。

针对上述问题,本文提出深度学习结合递归特征添加的网络入侵检测系统以提升网络对零日攻击的抵抗性,主要创新点为:

(1) 现有的大多数检测系统中,网络入侵攻击特征由于长字符串特性无法直接采用机器学习。本文方法通过采用深度置信网络与二元组编码技术对网络攻击特征进行有效提取与二进制编码,从而提升入侵检测准确率。

(2) 现有大多数检测方法中,网络入侵攻击决策依据的特征存在相互依赖的问题。本文利用递归特征添加法(Recursive Feature Addition,RFA)进行特征选择,并综合考虑入侵检测的准确率、检出率和误报率,从而提升检测效率与精准率。

1　方法设计

1.1　基于 DBN 与二元组编码的网络攻击特征提取

图 1 为基于 DBN 的网络攻击特征提取算法的架构,其中:输入数据为 I;输出数据为 O(即为网络攻击特征)。为避免深度学习网络函数表达能力过强出现的过拟合情形,采用非监督预训练方式对深度学习网络进行训练,通过逐层学习获取输入数据的主要驱动变量,并利用多层映射单元提取出网络攻击中主要的结构信息。

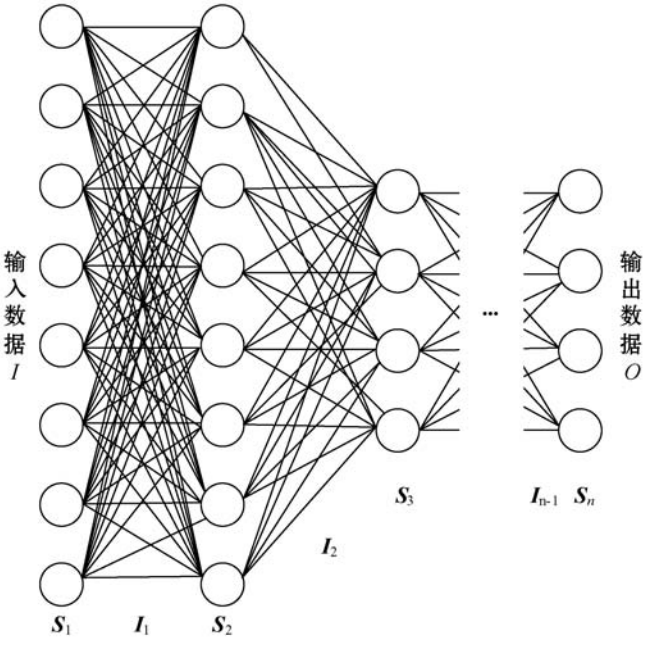


图 1　深度学习算法结构

对隶属于相邻两层(S_i,S_j),$i \neq j$ 的一组神经元(s_{li},s_{mj}),S_i 表示第 i 层特征,S_j 表示第 j 层特征,s_{li} 和 s_{mj} 分别为 S_i 和 S_j 的神经元,定义其能量函数为:

$$E(s_{li},s_{mj};\theta) = -\sum_{l=1}^{L}\sum_{m=1}^{M}\delta_{ij}s_{li}s_{mj} - \sum_{l=1}^{L}\sigma_l s_{li} - \sum_{m=1}^{M}\upsilon_m s_{mj} \tag{1}$$

式中:δ_{ij}、σ_l、υ_m 为权重参数;θ 表示 3 个权重参数的集合,即 $\theta=\{\delta_{ij},\sigma_l,\upsilon_m\}$。

两层神经元间的联合概率分布为:

$$P(s_{li},s_{mj};\theta) = \frac{\exp(-E(s_{li},s_{mj};\theta))}{Z} \tag{2}$$

式中:$Z = \sum_{l}\sum_{m}\exp(-E(s_{li},s_{mj};\theta))$ 为归一化因子。向量$\{s_i\}$的边缘分布或似然函数为:

$$P(s_i;\theta) = \frac{\sum_{s_j}\exp(-E(s_i,s_j;\theta))}{Z} \tag{3}$$

最优的模型参数 θ,可通过最大化训练集上的对数似然函数得到[10]。

进一步地,选择使用二元组技术对这些特征进行编码。其优点在于避免了特征量因字符串过长而不利于在机器学习中直接使用的缺陷。二元组特征编码以构造字典的方式完成对所有特征的编码。

1.2 基于 RFA 的特征选择方法

1.2.1 特征选择方法

相较于传统神经网络特征提取方法,深度学习网络中特征数量往往从数十个增加到数百个。这些特征往往具有高度的冗余性,从而导致最终入侵检测精度的下降[11-13]。针对这一问题,本文提出根据预定标准查找网络攻击特征的一个或多个信息子集的搜索算法,如图 2 所示。

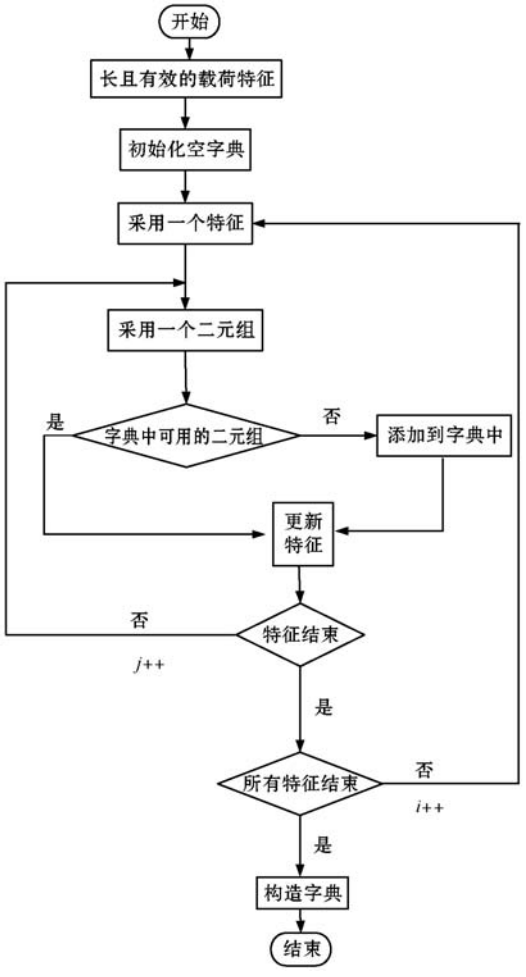
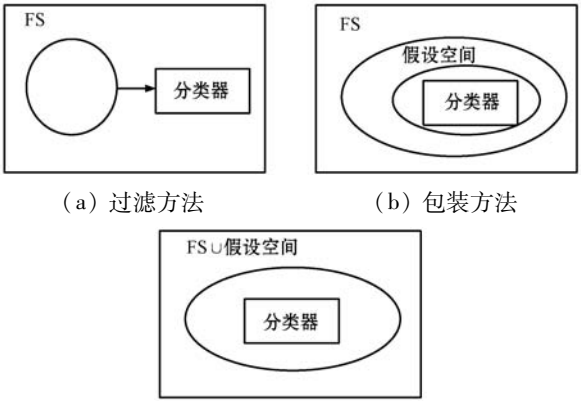


图 2 利用二元组技术提取 ISCX 数据集特征时的字典构建阶段

不妨设 $F=\{F_1,F_2,\cdots,F_n\}$ 是整个特征集;$S=\{F_{\tau(1)},F_{\tau(2)},\cdots,F_{\tau(m)}\}$ $(S\subseteq F)$ 是整个集合中特征的选定子集,其中 $m<n$。特征子集选择的目标是根据一些标准 J 选择表示原始数据的最具信息性的子集 $S_{\text{optimal}}\subseteq F$。但是,原始特征集可能包含一些不相关的特征。当 $p(Y=y\mid F_i=f_i)\neq p(Y=y)$,其中 Y 表示标签或输出时,才认为特征 F_i 是相关的。根据该定义,如果其值可以改变 Y 的预测,则特征 F_i 是相关的。换句话说,Y 有条件地依赖于 F_i。因此,最佳特征集是指数个可能的子集之一,并且比较所有这些子集以找到最好的子集 S_{optimal} 是难以处理的。

此外,由于额外噪声特征的存在,通常会增加训练分类器的难度,故识别噪声特征也是一个关键的预处理步骤,因为在这种噪声数据上构建的分类器的性能将高度依赖于训练数据的质量。换言之,噪声背景下需要确定最优的特征决策边界值,这直接导致特征提取难度更大,当前针对含噪声的特征提取问题而言,可以分为如图 3 所示的三种类型:过滤方法、包装方法和嵌入方法,其中 FS 表示所有特征的集合空间。



(a) 过滤方法 (b) 包装方法

(c) 嵌入方法

图 3 不同特征选择方法

1) 过滤方法:如图 3(a)所示,特征选择与分类过程相互独立,通过检查数据的内在属性来确定特征重要性。一般计算所有特征的相关性得分,去除低评分对应的特征后即为所选择的特征。

2) 包装方法:如图 3(b)所示,该方法采用机器学习确定特征子集的空间范围,且分类器的构造过程与特征选择过程紧密结合,但机器学习过程与后续的分类器构造与特征选择过程独立。

3) 嵌入方法:如图 3(c)所示,该方法在机器学习过程中执行特征选择步骤,其优势在于提升算法执行效率,且机器学习和特征选择部分不能分开。

1.2.2 递归特征添加方法(RFA)

考虑到嵌入式特征选择方法存在的明显优势,本文提出基于递归特征添加方法的特征选择算法。其基本原理为,通过根据计算出的剩余特征的排序系数(Ranking Coefficient,RC),一次性地向该集合添加一个特征,来初始化要用于所选特征的一组空特征。

将针对支持向量(其基本上代表训练示例的小子集)计算决策函数 $D(x)$ 的权重 w_i。支持向量是最接近决策边界的训练示例,并提供类之间的最大间隔。对 RFA 中的特征进行排序取决于权重大小作为排序系数,通过在成本函数的最大变化时添加一个特征来执行。

需要最小化的 SVM 的成本函数为:

$$\boldsymbol{J}=(1/2)\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\alpha}-\boldsymbol{\alpha}_1 \tag{4}$$

式中:$\boldsymbol{\alpha}_1$ 是 n 维向量;$\boldsymbol{H}$ 是可以计算的矩阵:

$$\boldsymbol{H}=\boldsymbol{y}_h\boldsymbol{y}_k K(x_h,x_k) \tag{5}$$

式中:x_h 和 x_k 是训练样例;K 是用于测量训练样例 x_h 和 x_k 之间的相似性的核函数,且有 $h,k=1,2,\cdots,N$,N 为需要选择的特征数量;$\boldsymbol{y}$ 是类标签的向量。此算法中使用 RBF 核函数,可以计算为:

$$K(x_h,x_k)=\exp(-\gamma\|x_h-x_k\|^2) \tag{6}$$

式中:γ 是常数,通常选择为总特征数量的倒数;为了计算由于添加一个特征 $\boldsymbol{i}$ 而导致的成本函数的变化,需要重新计算 $\boldsymbol{H}$ 矩阵,不妨记为 $\boldsymbol{H}(+\boldsymbol{i})$,其中符号 $(+\boldsymbol{i})$ 对应于添加特征 $\boldsymbol{i}$。这一过程中需要同步更新计算 $K(x_h(+\boldsymbol{i}),x_k(+\boldsymbol{i}))$。最终的排序系数 RC 计算为:

$$RC=(1/2)\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{\alpha}-(1/2)\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{H}(+\boldsymbol{i})\boldsymbol{\alpha} \tag{7}$$

算法第一次迭代时,由于还没有选择特征,此时的排名系数 RC 仅有第一项。随后迭代地执行该算法以执行递归特征加法(RFA),从而对应于取最大值的 $RC(\boldsymbol{i})$ 的特征被不断添加到排序特征列表,算法的结果将是从最重要到最不重要的特征的排序列表。

以图 4 所示的 4 特征排序问题对所提算法进行详细阐述,嵌入式前向特征选择有四个特征,选择顺序为实线(2,4,3,1),其中:1 表示存在特征;0 表示缺失。

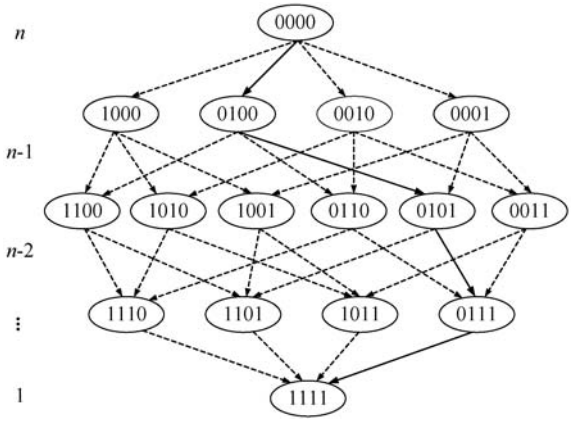


图 4　4 特征排序图

图 4 中,每个特征位都表示为二进制值(0 或 1)。选择此特征后,其位置将为 1,否则为 0。RFA 方法以空特征集(0 0 0 0)开始。示例中,特征 2 被选择为其他特征中最相关的特征,因此首先选择它。算法继续进行,直到它根据排名系数对所有特征进行排名。实线表示方法遵循的路径,而虚线表示当前案例中的所有可能情况。该示例的最终排名分别为(2,4,3,1)。

2　实　验

2.1　评估指标

为衡量特征添加前后对网络入侵检测效果的影响,采用检测精度和 F 度量值作为指标。其中检测精度计算公式为:

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

式中:TP、TN、FP 和 FN 分别为 True Positive、True Negative、False Positive 和 False Negative 值。可知分类器检测精度是正确分类的示例与示例总数的百分比。另一方面,F 度量值是检测精度和召回率的调和平均值,公式为:

$$F\text{-}measure=\frac{2\times TP}{(2\times TP)+FP+FN} \tag{9}$$

此外,引入额外三个指标以评估所提入侵检测方法的有效性:

(1) 检测率(Detection Rate,DR)。任何入侵检测系统的检测率根据以下公式得到,表示正确检测到的攻击占攻击总数的百分比:

$$DR=\frac{TP}{TP+FN} \tag{10}$$

(2) 误报率(False Alarm Rate,FAR)。FAR 表示根据以下公式将正确实例的百分比错误地分类为对正常实例总数的攻击:

$$FAR=\frac{FP}{FP+FN} \tag{11}$$

(3) 综合性能。除了前面提到的指标之外,提出了一个综合指标,以便将三个指标(准确度、检测率和 FAR)合并在一起来比较:

$$Combined=\left\{\frac{ACC+DR}{2}\right\}-FAR \tag{12}$$

该度量的结果将是 -1 和 1 之间的实数值,即 $Combined\in[-1,1]$,其中:-1 对应于最差的整体系统性能;1 对应于最佳的整体系统性能;而 0 对应于 50% 的整体系统性能。在目前的形式中,该公式给予所有三个指标(准确度、检测率和误报率)相等的权重。但在其他场景下,可通过赋予不同权重值以实现对不同性能的侧重关注程度。

2.2　评估指标不同对入侵检测方法性能的影响

为了阐明所提出的组合度量的重要性,假设有一个数据集,其中 200 个实例分为 100 个普通实例和 100 个攻击实例。现在讨论具有相同精度 50% 的三种不同场景。

脚本 1　$\boldsymbol{Confusionmatrix}=\begin{bmatrix}1 & 99\\ 1 & 99\end{bmatrix}$

脚本 2　$\boldsymbol{Confusionmatrix}=\begin{bmatrix}50 & 50\\ 50 & 50\end{bmatrix}$

脚本 3　$\boldsymbol{Confusionmatrix}=\begin{bmatrix}70 & 30\\ 70 & 30\end{bmatrix}$

对于上述三种不同的混淆矩阵,可以计算出性能指标,如表 1 所示。

表 1 具有相同准确性的三种不同方案的性能指标

脚本	准确性	DR	FAR	组合
脚本 1	0.5	1.0	0.95	-0.25
脚本 2	0.5	0.5	0.45	0.10
脚本 3	0.5	0.3	0.30	0.20

所有上述场景具有相同的准确度值。但是,这三种情况具有不同的检测率和误报率。这使得很难确定哪个系统在产生三种情景的三个系统中表现最佳。但是,如果计算提出的组合度量,可以得出结论,生成第三个场景的系统是其他系统中最好的系统。

同样,可能存在具有相同检测率、不同精度和误报率的不同系统,如下面三种情况所示,它们都具有相同的检测率 50%。表 2 为脚本 4 - 脚本 6 所对应的性能指标。

脚本 4 $Confusionmatrix = \begin{bmatrix} 1 & 99 \\ 50 & 50 \end{bmatrix}$

脚本 5 $Confusionmatrix = \begin{bmatrix} 99 & 1 \\ 50 & 50 \end{bmatrix}$

脚本 6 $Confusionmatrix = \begin{bmatrix} 50 & 50 \\ 50 & 50 \end{bmatrix}$

表 2 具有相同检测率的三种不同方案的性能指标

脚本	准确性	DR	FAR	组合
脚本 4	0.30	0.50	1.00	-0.63
脚本 5	0.75	0.50	0.02	0.62
脚本 6	0.50	0.50	0.50	0

从表 2 可得,具有最大组合度量的最佳系统是生成第二个场景的系统。同样地,可能存在不同的系统,这些系统产生具有不同精度和检测率的相等误报率,如下面三种情况所示,它们都具有相同的误报率 50%。表 3 为对应的性能指标。通过考虑组合性能可以得出结论,与第一个场景相对应的系统是最好的系统,因为它的组合度量是 0.3,这是其他系统中最高的。

脚本 7 $Confusionmatrix = \begin{bmatrix} 50 & 50 \\ 10 & 90 \end{bmatrix}$

脚本 8 $Confusionmatrix = \begin{bmatrix} 50 & 50 \\ 90 & 10 \end{bmatrix}$

脚本 9 $Confusionmatrix = \begin{bmatrix} 50 & 50 \\ 50 & 50 \end{bmatrix}$

表 3 具有相同误报率的三种不同场景的性能指标

脚本	准确性	DR	FAR	组合
脚本 7	0.8	1.0	0.5	0.3
脚本 8	0.2	0.2	0.5	-0.3
脚本 9	0.4	0.6	0.5	0.1

综上,当单独拍摄时,在测量入侵检测系统的性能时,准确度、检测率和误报率都不足以表达。但是提出的组合指标可以整合上述三个指标给出的信息,以便更加彻底地衡量入侵检测效果,正如从上述情景中注意到的那样。通过使用所提出的组合度量,可以选择具有最佳的最大准确度、最大检测率和最小误报率的系统。如前所述,可以根据应用修改提出的组合公式以测量其性能。这可以通过根据它们对该应用的重要性给予三个分量(准确度、检测率和误报率)不同的权重来执行。

2.3 特征提取和数据集准备

为进一步评估不同入侵检测方法的检测性能,采用基准数据集 ISCX 进行评估,其由不同类型的特征组成:数字,分类,日期时间和字符串。通常,网络流量信息由上述类型混合表示,但是蕴含的特征通常由长字符串值表示,这使得在机器学习中难以处理,为此采用二元组编码技术进行问题化简。

不失一般性,考虑将网络入侵特征转换为双字节表示的示例。三个网络特征具有不同的长字符串:“B7z2”,“Vud3j”和“z2nB7”。依据图 2 所示的字典生成过程,得到由 9 个单词组成的字典(即二元组):B7|7z|z2|Vu|ud|d3|3j|2n|nB。如表 4 所示,三个网络特征的二元组表示转化为具有“0”和“1”的二进制编码。

表 4 示例中三个网络入侵特征的二元组表示

原始负载	B7	7z	z2	Vu	ud	d3	3j	2n	nB
B7z2	1	1	1	0	0	0	0	0	0
Vud3j	0	0	0	1	1	1	1	0	0
z2nB7	1	0	1	0	0	0	0	1	1

为了准备用于特征选择的结果数据集,使用快速过滤器选择算法对特征进行预排序步骤。由于当前的特征数量很大,并且在这种情况下特征选择可能非常耗费时间,因此从原始特征中获取 350 个特征的子集。产生的 350 个特征分别用于生成大小为 25、50、100 和 500 的数据集。为了模拟“零日攻击”,使用不同数量的示例来监视每种数据集大小的特征选择行为,从特征提取到使用特征选择算法对特征进行排名的所有步

骤如图 5 所示。

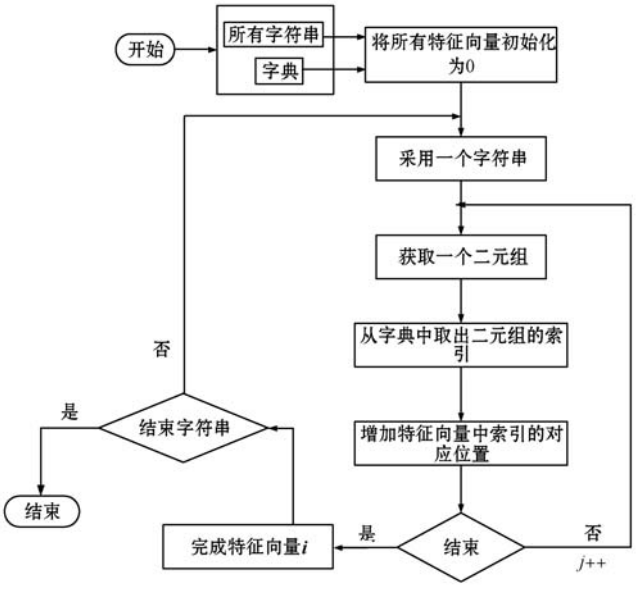


图 5　利用二元组技术提取 ISCX 数据集的特征向量

2.4　在 ISCX 数据集上应用 RFA 的结果

为了观察包含网络入侵特征添加对提高检测精度的影响,在所有 ISCX 数据集上测量网络入侵特征添加前后的分类精度和 F-度量值,如表 5 所示。

表 5　在 ISCX 数据集上应用 RFA 前后的性能指标　%

数据集大小	无二元组特征[12]		RFA	
	ACC	F-度量	ACC	F-度量
25 样本	66	64	74.5	79.6
50 样本	75	76	86.7	89.6
100 样本	81	85	89.8	91.8
500 样本	87	86	91.6	93.9

表 5 的第 2 列和第 3 列表示不添加网络入侵特征时的检测性能;第 4 列和第 5 列表示在利用二元组编码技术添加网络入侵特征之后,从分类器获得的最大性能。可以看出,与没有二元组特征的相同数据集上分类器的性能相比,本文算法的检测精度提升 8% 以上。

表 6 为采用 RFA 方法后相关评估指标随数据集规模的变化情形。

表 6　在 ISCX 数据集上应用 RFA 后的所有性能指标　%

数据集大小	RFA				
	ACC	F-度量	DR	FAR	组合
25 样本	79.6	77.5	71.9	11.1	62.5
50 样本	87.6	89.6	87.9	8.5	78.2
100 样本	87.9	88.7	85.8	1.3	80.6
500 样本	91.8	93.4	88.2	2.8	86.8

可以看出,大多数单一指标随着数据集大小的增加而提高,而 FAR 指标则不严格遵守随数据集规模扩大而下降的趋势。然而,对于组合评估指标而言,由初始 25 个样本所对应的指标 62.5% 逐步提升至 500 个样本所对应的指标(86.5%),与大部分单一性指标的变化趋势相同。这表明采用组合性能指标能够较好地反映入侵检测的性能,且随着数据集样本规模的扩大,其评估结果越好。

需要指出的是,文献[12]提出的检测方法未使用 RFA 添加网络入侵特征,故在此作为本文算法的对比算法,表 7 为相应的评估性能。为直观分析本文方法的优异性,相比文献[12]方法,在不同样本数情况下,本文方法的评估指标提升情况如图 6 所示。

表 7　在 ISCX 数据集上的所有性能指标(文献[12])　%

数据集大小	ACC	F-度量	DR	FAR	组合
25 样本	68.3	72.3	65.5	12.7	58.2
50 样本	81.3	82.9	82.6	10.9	73.8
100 样本	84.5	79.6	78.6	2.6	76.4
500 样本	88.6	86.7	82.5	5.3	81.6

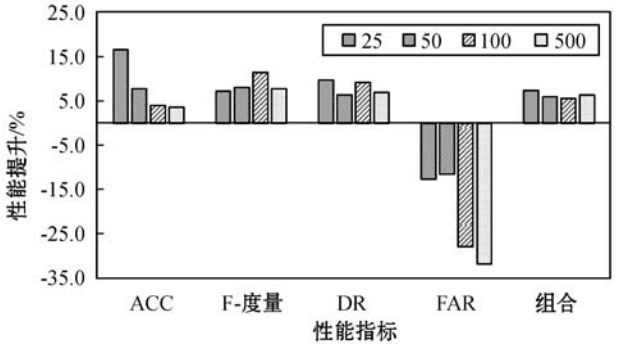


图 6　与文献[12]相比本文方法评估指标提升百分比

根据图 6,相比于传统的未考虑特征选择的方法(文献[12]),在不同数据集样本规模下,本文方法的检测准确率、F-度量值、检出率和综合性能指标均有所提升,而误报率低于未考虑特征选择的方法。例如,当数据样本数量为 25 时,与文献[12]方法相比,本文方法的准确率、F-度量值、检出率和综合性能指标分别提升了 16.5%、7.2%、9.8%、7.4%,而误报率则下降了 12.6%;而其余数据样本规模下,本文方法的准确率、F-度量值、检出率和综合性能指标则至少提升了 3.6%、7.7%、6.4% 和 5.5%。因此,算例结果表明,本文方法在提升检测精度的同时,由于综合考虑了网络入侵的特征,进一步降低了将正常情形分类为入侵事件以及将入侵事件误判为正常情形的风险,因而所提方法的检测性能更优。

2.5 方法对比结果分析

为验证所提深度置信网络结合递归特征添加的网络入侵检测方法的有效性,采用 KNN 算法[14]、决策树算法[15]、Adaboost 算法[16]、K-means 算法[17]、SVM 算法[18]等主流入侵检测算法进行对比。对比指标为检测精度,如式(8)所示。相应的检测结果如图 7 所示。

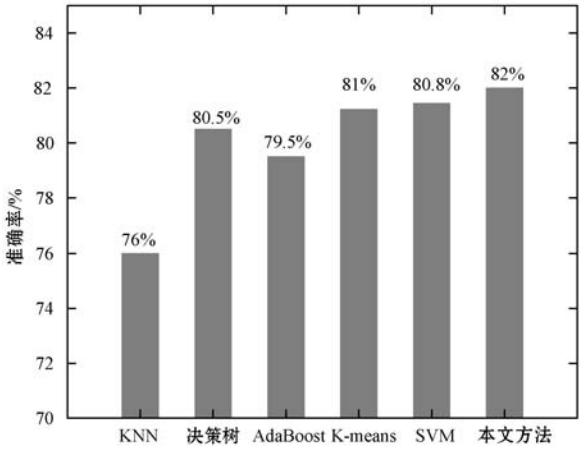


图 7 几种算法对网络入侵准确率检测结果

可以看出,相较于传统入侵检测算法,本文算法的准确度最高,且准确度最多提升 7.89%。这是因为本文算法首先采用深度置信网络对网络入侵特征进行提取,并基于二元组编码与 RFA 算法相结合的方法对主要特征量进行选择,因此对未知网络入侵攻击的检测准确率更高,网络安全防御能力更强。

3 结 语

为实现对网络“零日攻击”的有效判别,在综合考虑检测准确率、检出率和误报率的基础上,提出一种深度学习结合递归特征添加的网络入侵检测方法。同一数据集下的算例实验表明,相较于传统不考虑网络入侵特征的方法,本文算法的检测精度得到有效提升。此外,与传统入侵检测方法相比,本文算法同样展现出更高的检测能力,从而保证对未知网络攻击手段的检出能力,保障互联网安全。

未来的研究方向是使用集成分类器方法进行入侵检测,以提高检测性能,并利用 Android 僵尸网络数据集,来研究这些技术在检测 Android 系统恶意软件中的行为。

参 考 文 献

[1] 吕雪峰,谢耀滨. 一种基于状态迁移图的工业控制系统异常检测方法[J]. 自动化学报,2018,44(9):128 – 137.

[2] 曹科研,栾方军,孙焕良,等. 不确定数据基于密度的局部异常点检测[J]. 计算机学报,2017,40(10):2231 – 2244.

[3] 王晓程,刘恩德,谢小权. 攻击分类研究与分布式网络入侵检测系统[J]. 计算机研究与发展,2017,38(6):254 – 261.

[4] 陈科,李之棠. 网络入侵检测系统和防火墙集成的框架模型[J]. 计算机工程与科学,2018,23(2):26 – 28.

[5] Aghdam MH, Kabiri P. Feature selection for intrusion detection system using ant colony optimization[J]. International Journal of Network Security, 2016,18(3):420 – 432.

[6] Sultana N, Chilamkurti N, Wei P, et al. Survey on SDN based network intrusion detection system using machine learning approaches[J]. Peer-to-Peer Networking and Applications, 2018, 12(2):22 – 31.

[7] Beniwal S, Arora J. Classification and feature selection techniques in data mining[J]. International Journal of Engineering Research Technology, 2018, 28(6):11 – 19.

[8] Chang C, Lin C. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology, 2011, 2(3):27 – 32.

[9] 李信满,赵大哲,赵宏,等. 基于应用的高速网络入侵检测系统研究[J]. 通信学报,2015,23(9):150 – 156.

[10] 张承智,肖先勇,郑子萱. 基于实值深度置信网络的用户侧窃电行为检测[J]. 电网技术,2019,43(3):1083 – 1091.

[11] Heidarian Z, Movahedinia N, Moghim N, et al. Intrusion detection based on normal traffic specifications[J]. International Journal of Computer Network and Information Security, 2015, 7(9):32 – 36.

[12] 邓贵仕,刘金峰. 基于免疫原理的网络入侵检测系统的研究[J]. 计算机应用研究,2017,21(9):139 – 141.

[13] 黄伟庆,丁昶,崔越,等. 基于恶意读写器发现的 RFID 空口入侵检测技术[J]. 软件学报,2018,29(7):100 – 114.

[14] Peng H, Sun Z, Zhao X, et al. A detection method for anomaly flow in software defined network[J]. IEEE Access, 2018, 6(10): 27809 – 27817.

[15] 黄金超,马颖华,齐开悦,等. 一种基于集成学习的入侵检测算法[J]. 上海交通大学学报,2018,52(10):1382 – 1387.

[16] 董超,周刚,刘玉娇,等. 基于改进的 Adaboost 算法在网络入侵检测中的应用[J]. 四川大学学报(自然科学版),2015,52(6):1225 – 1229.

[17] 肖苗苗,魏本征,尹义龙. 基于 BFOA 和 K-means 的复合入侵检测算法[J]. 山东大学学报(工学版),2018,48(3):119 – 123.

[18] Saleh A I, Talaat F M, Labib L M. A hybrid intrusion detection system(HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers[J]. Artificial Intelligence Review, 2019,51(3):403 – 443.